INFORMATIK
für Mathematiker und Physiker

Eine Einführung in C++

Skript zur Vorlesung 252-0847-00
Herbstsemester 2013, ETH Zürich[1]

Bernd Gärtner Michael Hoffmann

[1]This book originates from a set of lecture notes written by Joachim Giesen and the authors in 2005. It is also influenced by discussions with Felix Friedrich.

Contents

Contents 3

1 Introduction 9
- 1.1 Why learn programming? 10
- 1.2 How to run a program 15
  - 1.2.1 Editor 15
  - 1.2.2 Compiler 15
  - 1.2.3 Computer 17
  - 1.2.4 Operating system 18
  - 1.2.5 Platform 19
  - 1.2.6 Details 19

2 Foundations 21
- 2.1 A first C++ program 22
  - 2.1.1 Syntax and semantics 23
  - 2.1.2 Comments and layout 25
  - 2.1.3 Include directives 25
  - 2.1.4 The main function 26
  - 2.1.5 Values and effects 27
  - 2.1.6 Types and functionality 27
  - 2.1.7 Literals 27
  - 2.1.8 Variables 27
  - 2.1.9 Constants (I promised myself) 28
  - 2.1.10 Identifiers and names 29
  - 2.1.11 Objects 30
  - 2.1.12 Expressions 30
  - 2.1.13 Lvalues and rvalues 31
  - 2.1.14 Operators 31
  - 2.1.15 Statements 35
  - 2.1.16 The first program revisited 36
  - 2.1.17 Details 38
  - 2.1.18 Goals 40
  - 2.1.19 Exercises 41

2.1.20 Challenges . . . 44
2.2 Integers . . . 46
2.2.1 Associativity and precedence of operators . . . 47
2.2.2 Expression trees . . . 48
2.2.3 Evaluating expressions . . . 49
2.2.4 Arithmetic operators on the type int . . . 50
2.2.5 Value range . . . 54
2.2.6 The type unsigned int . . . 55
2.2.7 Mixed expressions and conversions . . . 56
2.2.8 Binary representation . . . 57
2.2.9 Integral types . . . 59
2.2.10 Details . . . 60
2.2.11 Goals . . . 64
2.2.12 Exercises . . . 65
2.2.13 Challenges . . . 67
2.3 Booleans . . . 68
2.3.1 Boolean functions . . . 68
2.3.2 The type bool . . . 70
2.3.3 Short circuit evaluation . . . 73
2.3.4 Details . . . 73
2.3.5 Goals . . . 75
2.3.6 Exercises . . . 75
2.3.7 Challenges . . . 77
2.4 Control statements . . . 80
2.4.1 Selection: if– and if-else statements . . . 80
2.4.2 Iteration: for statements . . . 81
2.4.3 Blocks and scope . . . 85
2.4.4 Iteration: while statements . . . 90
2.4.5 Iteration: do statements . . . 92
2.4.6 Jump statements . . . 93
2.4.7 Equivalence of iteration statements . . . 95
2.4.8 Choosing the “right” iteration statements . . . 97
2.4.9 Details . . . 98
2.4.10 Goals . . . 102
2.4.11 Exercises . . . 103
2.4.12 Challenges . . . 107
2.5 Floating point numbers . . . 109
2.5.1 The types float and double . . . 110
2.5.2 Mixed expressions, conversions, and promotions . . . 112
2.5.3 Explicit conversions . . . 113
2.5.4 Value range . . . 114
2.5.5 Floating point number systems . . . 115
2.5.6 The IEEE standard 754 . . . 120

2.5.7 Computing with floating point numbers . . . 121
2.5.8 Details . . . 125
2.5.9 Goals . . . 127
2.5.10 Exercises . . . 128
2.5.11 Challenges . . . 132
2.6 A first C++ function . . . 134
2.6.1 Pre- and postconditions . . . 135
2.6.2 Function definitions . . . 137
2.6.3 Function calls . . . 138
2.6.4 The type void . . . 139
2.6.5 Functions and scope . . . 140
2.6.6 Procedural programming . . . 142
2.6.7 Modularization . . . 144
2.6.8 Using library functions . . . 150
2.6.9 Details . . . 152
2.6.10 Goals . . . 152
2.6.11 Exercises . . . 153
2.6.12 Challenges . . . 157
2.7 Arrays . . . 160
2.7.1 Static arrays . . . 160
2.7.2 Initializing static arrays . . . 162
2.7.3 Random access to elements . . . 162
2.7.4 Static arrays are primitive . . . 163
2.7.5 Vectors . . . 164
2.7.6 Strings . . . 166
2.7.7 Multidimensional arrays . . . 170
2.7.8 Details . . . 179
2.7.9 Goals . . . 180
2.7.10 Exercises . . . 180
2.7.11 Challenges . . . 182
2.8 Pointers, Algorithms, Iterators and Containers . . . 185
2.8.1 Arrays and functions . . . 185
2.8.2 Pointer types and functionality . . . 186
2.8.3 Array-to-pointer conversion . . . 187
2.8.4 Objects and addresses . . . 188
2.8.5 Pointer arithmetic . . . 190
2.8.6 Traversing arrays . . . 193
2.8.7 Const pointers . . . 195
2.8.8 Algorithms . . . 196
2.8.9 Vector iterators . . . 199
2.8.10 Containers and iterators . . . 201
2.8.11 Details . . . 204
2.8.12 Goals . . . 205

2.8.13 Exercises . . . 206
2.9 Recursion . . . 213
2.9.1 A warm-up . . . 213
2.9.2 The call stack . . . 214
2.9.3 Basic practice . . . 215
2.9.4 Recursion versus iteration . . . 217
2.9.5 Primitive recursion . . . 218
2.9.6 Sorting . . . 220
2.9.7 Lindenmayer systems . . . 229
2.9.8 Details . . . 235
2.9.9 Goals . . . 236
2.9.10 Exercises . . . 236
2.9.11 Challenges . . . 242

3 Classes **245**
3.1 Structs . . . 246
3.1.1 Struct definitions. . . . 249
3.1.2 Structs and scope . . . 250
3.1.3 Member access . . . 251
3.1.4 Initialization and assignment . . . 251
3.1.5 User-defined operators . . . 253
3.1.6 Details . . . 255
3.1.7 Goals . . . 256
3.1.8 Exercises . . . 257
3.1.9 Challenges . . . 260
3.2 Reference Types . . . 262
3.2.1 The trouble with `operator+=` . . . 262
3.2.2 Reference Types: Definition . . . 263
3.2.3 Call by value and call by reference . . . 264
3.2.4 Return by value and return by reference . . . 265
3.2.5 More user-defined operators . . . 266
3.2.6 Const references . . . 269
3.2.7 Const-types as return types. . . . 272
3.2.8 Goals . . . 273
3.2.9 Exercises . . . 273
3.2.10 Challenges . . . 276
3.3 Class types . . . 278
3.3.1 Encapsulation . . . 278
3.3.2 Public and private . . . 281
3.3.3 Member functions . . . 282
3.3.4 Constructors . . . 284
3.3.5 Default constructor . . . 285
3.3.6 User-defined conversions . . . 286

3.3.7 Member operators . . . 287
3.3.8 Nested types . . . 288
3.3.9 Class definitions . . . 289
3.3.10 Random numbers . . . 290
3.3.11 Details . . . 296
3.3.12 Goals . . . 297
3.3.13 Exercises . . . 297
3.3.14 Challenges . . . 299
3.4 Dynamic Types . . . 302
3.4.1 Lists . . . 304
3.4.2 Dynamic memory allocation . . . 305
3.4.3 List nodes . . . 308
3.4.4 Data members and default constructor . . . 310
3.4.5 Inserting an element at the front . . . 311
3.4.6 Outputting a list . . . 312
3.4.7 Appending an element at the end . . . 313
3.4.8 The copy constructor . . . 314
3.4.9 The assignment operator . . . 316
3.4.10 The destructor . . . 318
3.4.11 Definition of a dynamic type . . . 320
3.4.12 More list functionality . . . 320
3.4.13 Standard lists . . . 322
3.4.14 Details . . . 322
3.4.15 Goals . . . 323
3.4.16 Exercises . . . 323

A C++ Operators 325

B Solutions 327
B.1 A first C++ program . . . 327
B.2 Integers . . . 334
B.3 Booleans . . . 343
B.4 Control statements . . . 349
B.5 Floating point numbers . . . 362
B.6 A first C++ function . . . 377
B.7 Arrays and pointers . . . 391
B.8 Pointers, Algorithms, Iterators and Containers . . . 399
B.9 Recursion . . . 409
B.10 Structs . . . 429
B.11 Reference Types . . . 439
B.12 Classes . . . 450
B.13 Dynamic Types . . . 461

Index **464**

Chapter 1

Introduction

1.1 Why learn programming?

You can tell I'm educated, I studied at the Sorbonne
Doctored in mathematics, I could have been a don
I can program a computer, choose the perfect time
If you've got the inclination, I have got the crime

Pet Shop Boys, Opportunities (1986)

This section explains what a computer program is, and why it is important for you not only to use *computer programs, but also to* write *them.*

When people apply for a job these days, their resume typically contains a section called *computer skills*. Items listed there might include *Lotus Notes*, *Excel*, or *PhotoShop*. These are the names of *application programs*, programs that have been written by certain people (in the above cases, at Microsoft corporation) to be used by other people (for example, a sales representative).

The *computer skills* section might also list items like *HTML*, *Java*, or *C++*. These are the names of *programming languages*, languages used to instruct, or program, a computer. Using a programming language, *you* can write the programs that will subsequently be used by others, or by yourself.

A computer program is a list of instructions to be automatically processed by a computer. The computer itself is stupid—all the intelligence comes from the program. In this sense, a program for the computer is like a cookbook recipe for someone who cannot cook: even with very limited skills, impressive results can be obtained, through a step-by-step instruction.

Most people simply use programs, just like they use cookbooks. A sales representative, for example, needs application programs as tools for his work. The fact that you are reading this lets us believe that you potentially belong to the category of people who also need to write programs.

There are many reasons for writing programs. Some employer might pay for it, some bachelor course might require it, but ultimately, there is a deeper reason behind it that we plan to explain next. The upshot is that nowadays, you cannot be a serious engineer, let alone a serious scientist, without at least some basic programming skills. Even in less serious contexts, we can recommend to learn programming, because it can bring about a lot of fun and satisfaction.

In the twentieth century, computers have revolutionized the way science and engineering are done. To be more concrete, we will underpin this with an example from mathematics. You probably don't expect math to be mentioned first in connection with computers; indeed, many mathematicians still use paper and pencil on a daily basis. But *what* they write down has changed. Before computers were available, it was often

necessary to write down actual numbers, and to perform calculations with them by hand. This happened not so much in writing proofs for new theorems, but in the process of *finding* these theorems. This process often requires to go over many concrete examples, or counterexamples, in order to see certain patterns, or to discover that some statement is false. The computer has tremendously accelerated this process by taking over the routine work. When you look at a mathematician's notepad today, you still find greek letters and all kinds of strange symbols, but most likely no numbers larger than ten.

There is one topic that nicely illustrates the situation, and this is the search for *Mersenne primes*. In 1644, the French monk and mathematician Marin Mersenne established the following claim.

> **Mersenne's Conjecture.** *The numbers of the form* $2^n - 1$ *are prime numbers for* $n = 2, 3, 5, 7, 13, 17, 19, 31, 67, 127, 257$*, but for no other number* $n < 257$*.*

Mersenne corresponded with many of the leading mathematicians at that time, so his conjecture became widely known. Up to $n = 7$, you can verify it while you read this, and in 1644, the conjecture was already verified up to $n = 19$.

It took more than hundred years until the next exponent on Mersenne's list could be verified. In a letter to Bernoulli published in 1772, Leonhard Euler proved that $2^{31} - 1 = 2147483647$ is a prime number. But in 1876, another hundred years later, Mersenne posthumously received a heavy blow. Edouard Lucas proved that $2^{67} - 1 = 147573952589676412927$ is *not* a prime number (Lucas showed his passion for large numbers also when he invented the *Tower of Hanoi* puzzle). Lucas's proof does not work the way you would expect: it does *not* exhibit a prime factor of $2^{67} - 1$ (the most direct way of proving that a number is not prime), but it uses a clever indirect argument invented by Lucas in the same year. The factorization of $2^{67} - 1$ remained unknown for another 25 years.

In 1903, Frank Nelson Cole was scheduled to give a lecture to the American Mathematical Society, whose title was 'On the Factorization of Large Numbers'. Cole went to the blackboard, and without saying a single word, he first wrote down a calculation to obtain $2^{67} - 1$ by repeated multiplication with two. He finally had the number

```
147573952589676412927
```

on the blackboard. Then he wrote down another (much more interesting) calculation for the product of two numbers.

```
761838257287 x 193707721
------------------------
     761838257287
     6856544315583
      2285514771861
       5332867801009
```

```
        5332867801009
         5332867801009
          1523676514574
            761838257287
---------------------
147573952589676412927
```

Cole had proved that $2^{67} - 1 = 761838257287 \cdot 193707721$, making the result of Lucas believable to everybody: $2^{67} - 1$ is not a prime number! He received standing ovations for this accomplishment and later admitted that he had worked on finding these factors every Sunday for the last three years.

Today, you can start a computer algebra program on your computer (a popular one is `Maple`), type in

```
ifactor(2^67-1);
```

and within less than a second get the output

```
(761838257287)(193707721)
```

To summarize: hundred years ago, a brilliant mathematician needed three years to come up with a result that much less brilliant people (we are not talking about you) could get in less than a second today, using a computer and the right program. This seems disturbing at first sight, and thinking about the precious time of his life Cole devoted to the problem, you may even feel sorry for him. You shouldn't; rather, the story has three important lessons in store.

Tool skills. Lesson one is that Cole's calculations were extremely difficult, given the tools he had (paper, pencil, and probably very good mental arithmetic). Given the tools *you* have (the computer and a computer algebra program called `Maple`), Cole's calculations are easy routine. We are sure that Cole would feel sorry for anyone using these new tools only to reproduce some hundred-year old calculation. Useful new tools lead to new possibilities and challenges. On the one hand, this *allows* you to do more than you could do before; on the other hand it also *forces* you to do more if you want to keep up with the developments. Whatever you do, nowadays you must acquire and maintain at least some basic knowledge of computers and application programs.

Problem skills. Lesson two is that tool skills alone would not have helped Cole to factor $2^{67} - 1$. Cole also was a good mathematician who knew a lot of theory he could use to save calculations. This is the reason why he "only" needed three years.

Even nowadays, computers and application programs are not everything. Factoring $2^{67} - 1$ is easy because this is a small number by today's standards. But factoring large numbers (2^{1000} is considered large today; in a couple of years, it might be 2^{2000}) is still a very difficult problem for which no efficient solutions are known. The problem of

factoring large numbers is the most prominent problem for which most people must actually hope that *no* efficient solution will ever be found. The reason is that many cryptosystems that are in use (think of secure internet banking) are purely based on the practical impossibility of factoring large numbers. Therefore, the worst scenario would be that the "bad guys" discover first how to factor large numbers efficiently.

There are many other problems that are as far from a solution as they were in pre-computer days. Coming back to Mersenne, we still cannot characterize the exponents n for which the number $2^n - 1$ is a prime number. We don't even know whether there are infinitely many such Mersenne primes. If you plan to make a contribution here, you should not buy a faster computer with the latest version of `Maple`, but study math. Even in the case of problems for which computers can really contribute to (or actually find) the solution, you typically need to have a deep understanding of the problem in order to know *how* to use the computer. If you want to become an engineer or a scientist, you must acquire and maintain a profound knowledge about the problems you will be dealing with. This fact was true hundred years ago, and it is still true—computers have not yet learned to solve interesting problems by themselves.

Programming Skills. Lesson three is one that Cole did not live to see: nowadays, problem-specific knowledge can be turned into problem-specific computer programs. That way, the state of the art concerning Mersenne primes has advanced quite far. It turned out that Mersenne had made five mistakes: $n = 67$ and $n = 257$ in Mersenne's list do not lead to prime numbers; on the other hand, Mersenne had "forgotten" the exponents $n = 61, 89$ and 107.

As of September 2013, we know 48 Mersenne primes, the largest of which has an exponent of $n = 57,885,161$ (see the GIMPS project at `www.mersenne.org`). But don't believe that this one was found with off-the-shelf programs.

Problems occurring in the daily life of an engineer or a scientist are often not easy to solve, even with a computer and standard software at hand. In order to attack them, you need *tool skills* for the routine calculations, and *problem skills* to understand and extract the aspects of the problem that can in principal be solved by a computer. But in the end, you need *programming skills* to actually do it.

The art of computer programming. To conclude this section, let us be honest: for many people (including the authors of this book), the process of writing programs has some very non-utilitarian aspects as well. We have mentioned two of them before: fun and satisfaction. We could add mathematical beauty and ego boost. In one way or another, every passionate programmer feels at least a little bit like an artist.

The prime advocator of this view on programming is Donald E. Knuth. He is the author of a monumental and seminal series of seven books entitled *The Art of Computer Programming*. Starting with Volume I in 1968, three of the seven volumes are published by now. Drafts of Volume IV circulate since 2005, and the planned release date of Volume V is 2015 (it should be added that Knuth was born in 1938, and on his webpage `http:`

//www-cs-faculty.stanford.edu/~knuth/taocp.html, he at least implicitly mentions the possibility that Volumes VI and VII will not be written anymore).

Let Knuth have the final say here (a quote from the beginning of Volume I):

> *The process of preparing programs for a digital computer is especially attractive, not only because it can be economically and scientifically rewarding, but also because it can be an aesthetic experience much like composing poetry or music.*

1.2 How to run a program

In Paris they just simply opened their eyes and stared when we spoke to them in French! We never did succeed in making those idiots understand their own language.

Mark Twain, The Innocents Abroad (1869)

This section explains what it really means to "write a program", and how you enable the computer to run it. For this, we describe the ingredients involved in the process: the editor, *the* compiler, *the* computer *itself, and the* operating system. *Computer, compiler and operating system together form the* platform *on which you are writing programs.*

1.2.1 Editor

Writing a program is not so different from writing a letter. One composes a text, that is, a (hopefully) meaningful sequence of characters. Usually, there are certain conventions on how such a text is structured, and the purpose of the text is to transport information.

What has been said so far applies to both letters and programs. But when writing a program, there is another aspect that has to be taken into account: A program has to be "read" by a computer, meaning that it must be available to the computer in electronic form. In the future, we might be able to orally dictate the program to the computer, but nowadays, the common way is to use a keyboard and simply type it in. An *editor* is an application program that allows you to display, modify, and electronically store such typed-in text. The use of editors is not restricted to programming, of course. With some still existing romantic exceptions, even letters are composed using editors such as *Word*.

1.2.2 Compiler

Making a program available to the computer in electronic form is usually not enough. The *machine language* a computer can understand directly is very primitive and quite different from natural languages.

Writing the programs in machine language is no viable alternative, since that would require to break the program into a large number of primitive instructions that the computer can understand. This is like telling your friend to come over for dinner by telling her which muscles to move in order to get to your place.

Moreover, machine languages vary considerably between different computers. That is, in order to use a program written for one specific computer A on a different computer B, one first has to translate the program from the machine language of A to the

Figure 1: *A compiler translates the sourcecode into an executable program.*

machine language of B. This process, called *porting*, can be very cumbersome if the machine languages of A and B are substantially different. Also, porting can only be done with a detailed knowledge of the peculiarities of the involved computers. But this type of knowledge is not generally worthwhile to acquire, as it is tied to one very specific computer. As soon as this computer is replaced by another one, major parts of such computer-specific knowledge become worthless and have to be rebuilt from scratch.

To reduce this undesirable entanglement of computers and programs, and to allow us to write programs in less primitive language, (high-level) programming languages have been developed. These are standardized languages that form a kind of compromise between natural languages and machine language. Indeed, the use of the word "compromise" is justified because there are two conflicting goals: On the one hand, we would like to write programs in a language that is as close to natural language as possible. On the other hand, we have to make the computers understand the programming language as well; this task is obviously much easier if the programming language is close to machine language.

What does it mean "to make the computers understand the programming language"? In the end, any program has to be translated into machine language. The process of this translation is called *compilation*. Now you will probably ask: "Where is the benefit of this whole programming language concept? In order to do the translation I still have to know all these computer-specific details, don't I?" Right. If you would have to translate the program yourself. The key is: You are not supposed to translate it yourself. Instead, let a program do it for you. Such a program is referred to as a *compiler*; it translates a given program in a programming language, the *sourcecode*, into a program in machine language, the *executable*. See Figure 1 for an illustration. The picture of the executable is somewhat inappropriate, since it does not show what the computer gets to see after compilation, but rather what *you* might see when you (accidentally) load the executable into the editor. The main point that we are trying to make here is that the executable is not human-readable.

In summary: The big benefit of (high-level) programming languages is that they abstract from the capabilities of specific computers. Programs written in a high-level language can be run on all kinds of computers, as long as a compiler for the language is available on the particular computer.

1.2.3 Computer

If you are not interested in writing compilers, it is not necessary to understand in detail how a computer works. But there are some basic principles behind the design of most computers that *are* important to understand. These principles form the *von Neumann architecture*, and they are important, since almost all programming languages are tailored to the von Neumann architecture.

Every computer with von Neumann architecture has a *random access memory* (RAM, or simply main memory), and a *central processing unit* (CPU, or simply processor). The main memory stores the program to be run, but also data that the program requires as input, and data that the program produces as output. The processor is the "brain" of the computer: it executes the program, meaning that it carries out the sequence of instructions prescribed by the program in machine language.

Main memory. You can think of the computer's main memory as a long row of switches, each of them being either on or off. During program execution, switches are flipped. At any time, the memory content—the current positions of all switches—defines the *program state*. The program state completely determines what happens next. Conceptually, we also consider user input and program output as part of the program state, even though the corresponding "switches" might be in the user's brain, or on printed paper.

Since modern computers are capable of flipping several switches at the same time, consecutive switches are grouped into *memory cells*. The positions of all switches in the cell define the *content* of the cell; in more abstract terms, the switches are called *bits*, each of them capable of storing one of the numbers $\{0, 1\}$. The memory cells are usually called *bytes* and represent the smallest segments of memory that can individually be manipulated. In this sense, you can interpret the content of a memory cell as a binary number with, for example, 8 digits.

The computer may be able to manipulate more than 8 bits simultaneously; a typical value is 32 bits, or 4 bytes. We also say that we have a 32-bit machine, or a 32-*bit system*.

Each memory cell is uniquely identified by its *address*. You can think of the address simply as the position of the memory cell in the list of all memory cells.

To look up bit values, or to flip bits within a specific memory cell, the cell has to be *accessed* through its address. Think of a robot arm with 8 fingers that you can tell to move to memory cell number 17.

The term *random access* refers to a physical property of the computer's memory:

the time it takes to access a cell (to "move to its bits") is the same for all cells; in particular, it does *not* depend on the address of the cell. When you think in terms of the robot arm analogy, it becomes clear that random access cannot be taken for granted. It is not necessary to discuss the physical means by which random access is realized; the important point here is that random access frees us from thinking about where to store a data item in order to access it efficiently.

Processor. You can think of the computer's processor as a box that is able to load and then execute the machine language instructions of a program in order. The processor has some memory cells of its own, called registers, and it can transfer data from the computer's main memory to its registers, and vice versa. The register contents are also part of the program state. Most importantly, the processor can perform a fixed set of simple operations (like adding or subtracting register contents), directly corresponding to the machine language instructions. This is where the functionality of the whole program comes from in the end. Even very complicated and useful programs can be put together from a simple set of machine language instructions.

A single instruction acts like a mathematical function: given the current program state, a valid instruction generates a new and well-defined next program state. This implies that every sequence of instructions, and in particular the whole program has a determined behavior, depending on the initial program state.

1.2.4 Operating system

We have seen that in order to write a program and run it, you first have to start an editor, type in the program, then call the compiler to translate the program into machine language, and finally tell the computer to execute it. In all this "starting", "calling" and "telling", you rely on the computer's *operating system* (OS), a program so basic that you may not even perceive it as a program. Popular operating systems are *Windows*, *Unix*, *Linux*, and *Mac OS*.

For example, you may start the editor by clicking on some icon, or by using a *command shell* (a text window for typing in commands; under Unix and Linux, this is still the default working mode).In both cases, the operating system makes sure that the editor program is loaded into the main memory, and that the processor starts executing it. Similarly, when you store your written program, the operating system allocates space for it on the hard disk and associates it with the file name you have provided.

A computer without operating system is like a car without tires, and most computers you can buy come with a pre-installed operating system. It is important to understand, though, that the operating system is not inextricably tied to the computer: you can take your "Windows PC" and reinstall it under Linux.

1.2.5 Platform

The computer, its operating system and the compiler are together referred to as the *platform* on which you are writing your programs. The editor is not part of the platform, since it does not influence the behavior of the program.

In an ideal world, there is no need for you to know the platform when you are writing programs in a high-level programming language. Recall that the plan is to delegate the platform-specific aspects to the compiler. A typical such platform-specific aspect is the number of bits that can be manipulated together. This is mostly 32 these days, but for some computers it is 64, and for very primitive computers (like they are used in smart cards, say), it can be less than 32.

When you are using or relying on machine-oriented features of the programming language, platform-specific behavior might be the result. Many high-level programming languages have such low-level features to facilitate the translation into *efficient* machine language.

Your goal should always be to write *platform-independent* code, since otherwise, it may be very difficult to get your program to run on another computer, even if you have a compiler for that computer. This implies that certain features should be avoided, even though it might seem advantageous to use them on a specific platform.

1.2.6 Details

Von Neumann's idea of a common memory for the program *and* the data seems obvious from today's point of view, but the earliest computers like Konrad Zuse's Z3 didn't work that way. In the Z3, for example, the memory for the program was a punch tape, decoupled from the input and output device, and from the main memory.

An interesting feature of the von Neumann architecture is that it allows self-modifying programs. These are popular among the designers of computer viruses, for example.

The von Neumann architecture with its two levels of memory (main memory and processor registers) is an idealized model, and we are implicitly working under this model throughout the course.

The reality looks more complicated. Modern computers also have a *cache*, logically belonging to the main memory, but allowing much faster access to memory cells (at the price of a more elaborate and expensive design). The idea is that frequently needed data are stored in the cache to speed up the program.

While caching is certainly a good thing, it makes the life of a programmer more difficult: you can no longer rely on the fact that access time to data is independent from where they are stored. In fact, to get the full performance benefit that caching can offer, the programmer has to make sure that data are accessed in a *cache-coherent* way. Doing this, however, requires some computer-specific knowledge about the cache, knowledge we were originally trying to avoid by using high-level programming languages. Luckily, we can often ignore this issue and (successfully) rely on the automatic cache management being offered. There is also a theoretical model for so-called *cache-oblivious* algorithms,

in which the algorithm does not know the parameters of the cache. Algorithms which are efficient under this model, are (in a certain sense) efficient for every concrete cache size.

In real-life applications, we also observe the phenomenon that the data to be processed are too large to fit into the computer's main memory. Operating systems can automatically deal with this by logically extending the main memory to the hard disk. However, the *swapping* that takes place when hard disk data to be accessed are transferred to the main memory incurs a severe performance penalty, much worse than poor cache usage. In this situation, it is often useless to rely on the automatic mechanisms provided by the operating systems, and the programmer is challenged to come up with *input/output efficient* programs.

Even when we extend the von Neumann architecture to include several layers of memory, there are computers that don't fit in. Most notably, there are *parallel computers* with more than one processor. In fact, even consumer PCs have more than one processor these days. Writing efficient programs for such a computer is a task entirely different from programming for the von Neumann architecture. To take full advantage of the parallelism, programs have to be decomposed into independent parts, each of which is then run by one of the processors. In many cases, this is not at all a straightforward task, and specialized programming languages have to be used.

A recent successful alternative to parallel computers are networks of single-processor computers. You can even call this a computer architecture. Finally, there are *quantum computers* that are based on completely different physical principles than the von Neumann architecture. "Real" quantum computers cannot be built yet, but as a theoretical model, quantum computers exist, and algorithms are already being developed in this promising model of computation.

Chapter 2

Foundations

2.1 A first C++ program

The basic tool for the manipulation of reality is the manipulation of words. If you can control the meaning of words, you can control the people who must use the words.

Philip K. Dick, How to Build a Universe That Doesn't Fall Apart Two Days Later (1978)

This section presents a first complete C++ program and introduces the syntactical *and* semantical *terms necessary to understand all its parts.*

Here is our first C++ program. It asks for a number a as input and outputs its eighth power a^8. If you have never seen a C++ program before, even this short one might look scary, since it contains a lot of strange-looking symbols and words that are not found in natural language. On the other hand, this is good news: as short as it is, this program already contains many important features of the C++ language. Once we have gone through them in this section, this program (and even other, bigger programs) won't look scary anymore.

```
// Program: power8.cpp
// Raise a number to the eighth power.

#include <iostream>

int main()
{
  // input
  std::cout << "Compute a^8 for a =? ";
  int a;
  std::cin >> a;

  // computation
  int b = a * a; // b = a^2
  b = b * b;     // b = a^4

  // output b * b, i.e., a^8
  std::cout << a << "^8 = " << b * b << ".\n";
  return 0;
}
```

Program 2.1: *../progs/lecture/power8.cpp*

If you compile this program on your computer and then run the executable file produced by the compiler, you find the following line on the standard output. Typically, the standard output is attached to some window on your computer screen.

```
Compute a^8 for a =?
```

You can now enter an integer, e.g. 2, using the keyboard. After pressing ENTER, the output on your screen reads as follows.

```
Compute a^8 for a =? 2
2^8 = 256.
```

Before discussing the program power8.cpp in detail, let us go over it once quickly. The lines starting with two slashes // are *comments*; they document the program such that it can easily be understood by a (human) reader. Line 4 contains an *include-directive*; in this case, it indicates that the program uses the input/output library `iostream`. The *main function* which is the heart of every C++ program spans lines 6–20. This function is called by the operating system when the program is started; it ends with a *return statement* in line 19. The value 0 is returned to the operating system, which by convention signals that the program terminated successfully.

The main function is divided into three parts. First, in lines 8–11 the input number is read. Line 9 outputs a message to the user that tells her which kind of input the program expects. In line 10 a *variable* `a` is declared that acts as a placeholder to store the input number. The keyword `int` indicates that `a` is an integer. In line 11, finally, the variable `a` receives its value from the input.

Then in lines 13–15 the actual computation takes place. In line 14, a new variable `b` is declared which acts as a placeholder to store the result of the computation. The variable `b` is initialized to the product `a * a`. Line 15 computes the product `b * b`, that is, a^4 and stores this result again in `b`.

The third part in lines 17–18 provides the program output. Part of it is the computation of the product `b * b`, that is, a^8.

2.1.1 Syntax and semantics.

In order to understand the program power8.cpp in detail, and more importantly, to write programs yourself later, you need to know the rules according to which programs are written. These rules form the *syntax* of C++. You further need to know how to interpret a program ("what does the program do?"), and this is determined by the *semantics* of C++. Even a program that is well-formed according to the C++ syntax may be *invalid* from a semantical point of view. A *valid* program is one that is syntactically *and* semantically correct.

It's the same with natural language: grammar tells you what sentences are, but the interpretation of a sentence (in particular, whether it makes sense at all) requires a concept of meaning.

When a program is invalid, the compiler may output an error message, and this will definitely happen when the program contains *syntax errors*, violations of the syntactical

rules. A program that is semantically invalid may compile without errors, but we are not allowed to make any assumptions about its behavior; the program *could* run fine, for example if the semantical error in question has no consequences on a particular platform. On other platforms, the program may behave strangely, or crash. Even on the same platform, it might work sometimes, but fail at other times. We say that the program's behavior is *undefined*. Clearly, one should avoid writing programs that exhibit undefined behavior.

The syntax of C++ is specified formally in a mathematical language. The description of the semantics is less strict; it rather resembles the text of a law, and as such it suffers from omissions and possible misinterpretations. The official law of C++ covering both syntax and semantics, is the ISO/IEC standard 14882 from 2011. [1]

While such a formal specification is indispensable (otherwise, how should a compiler know whether your program text is actually a C++ program, and what it is supposed to do?), it is not suitable for learning C++. Throughout this book, we explain the relevant syntactical and semantical terms in natural language and by example. For the sake of readability, we will often not strictly distinguish between syntactical and semantical terms: some terms are most naturally introduced as having both syntactical and semantical aspects, and it depends on the context which aspect is relevant.

Unspecified and implementation defined behavior. Sometimes, even valid programs behave differently on different platforms; this is one of the more ugly aspects of C++ that we'd prefer to sweep under the rug. Unfortunately, we can't ignore the issue completely, since it occasionally pops up in "real life".

There are two kinds of platform-dependent behavior. The nicer one is called *implementation defined* behavior.

Whenever the C++ standard calls some aspect of the language "implementation defined", you can expect your platform to contain documentation that fully specifies the aspect. The typical example for such an an implementation defined aspect is the number of bits that can be manipulated at once, see Section 1.2.3. In case of implementation defined aspects and resulting behavior, the C++ standard and the platform together completely determine the actual behavior.

The less nice kind is called *unspecified behavior*, coming from some unspecified aspect of the language. Here you can rely on a well-defined and usually *small* set of possible specifications, but the platform is not required to contain a full specification of the aspect. A typical example for such an unspecified aspect is the evaluation order of operands within an expression, see Section 2.1.12.

In writing programs, unspecified aspects cannot always be avoided, but usually, some care ensures that no unspecified or even undefined behavior results.

[1] This book is based on the previous version of the document from 1998, and as far as the book content is concerned, the differences between the previous and the new version are minor. A future version of this book will also address some of the new features of C++ that are officially available since 2011.

2.1.2 Comments and layout

Every good program contains comments, for example

```
// Program: power8.cpp
// Raise a number to the eighth power.
```

A comment starts with two slashes // and continues until the end of the line. Comments do not provide any functionality, meaning that the program would do exactly the same without them. Why is a program without comments bad, then? We do *not* only write programs for the compiler to translate them into executables, but we also write them for other people (including ourselves) to read, modify, correct or extend them.

Without comments, the latter tasks become very tedious when the program is not completely trivial. Trust us: Even you will not be able to understand your own programs after a couple of weeks, without comments. There is no *standard* way of writing comments, but we will follow some common-sense guidelines. One of them is that every program—even if it is very simple—should start with one or more lines of comments that mention the program's name and say what it does. In our case, the above two lines fully suffice.

Another key feature of a readable program is its layout; consider the version of power8.cpp shown in Program 2.2. We have removed comments, and all "unnecessary" layout elements like spaces, line breaks, blank lines, and indentations.

```
1 #include <iostream>
2 int main(){std::cout<<"Compute a^8 for a =? ";
3 int a;std::cin>>a;int b=a*a;b=b*b;std::cout<<
4 a<<"^8 = "<<b*b<<".\n";return 0;}
```

Program 2.2: *../progs/lecture/power8_condensed.cpp*

The compiler is completely ignorant about these changes, but a person reading the program will find this condensed version quite difficult to understand. The purpose of a good layout is to visualize the program structure. This for example means that logical blocks of the program should be separated by blank lines, or that one line of sourcecode should be responsible for only one thing. *Indentation*, like power8.cpp has it between the pair of curly braces, is another indispensable ingredient of good layout, although you will only later be able to fully appreciate this.

Typically, collaborative software projects have layout guidelines, making sure that everybody in the project can easily read everybody else's code. At the level of the simple programs discussed in this book, such formal guidelines are not necessary; we simply adhere to standard guidelines that have proven to work well in practice, and that are being used in almost every other book on C++ as well.

2.1.3 Include directives

Every useful program contains one or more include *directives*, such as

```
#include <iostream>
```

Usually, these appear at the very beginning of the program. `#include` directives are needed since, in C++, many important features are not part of the core language. Instead, they are implemented in the so-called *standard library* which is part of every C++ implementation. A *library* is a logical unit used to group certain functionality and to provide it to the user in a succinct form. In fact, the standard library consists of several libraries one of which is the input/output library.

A library presents its functionality to the user in the form of one or several *headers*. Each such header contains information that is needed by the compiler. In order to use a certain feature from a library, one has to include the corresponding header into the program by means of an `#include` directive. In `power8.cpp`, we want to use input and output which are (maybe surprisingly) not part of the core language. The corresponding header of the standard library is called `iostream`.

A well-designed C++ library puts its functionality into a *namespace*. The namespace of the standard library is called `std`. Then, in order to access a feature from the library, we have to *qualify* its name with the namespace, like in `std::cin` (this is the feature that allows us to read input from the keyboard). This mechanism helps to avoid *name clashes* in which different features accidentally get the same name. At the same time, explicit qualification increases the readability of a program, as it is immediately apparent from which library a given feature comes. A name that is not qualified is called *unqualified* and usually corresponds to a feature defined in our own program.

2.1.4 The main function

Every C++ program must have a `main` function. The shortest program reads as follows.

```
int main() { return 0; }
```

This program does nothing. The `main` function is called by the operating system when you tell it to run the program; but why is it a *function*, and what is "`return 0;`" supposed to mean? Just like a mathematical function, the `main` function can have arguments given to it upon execution of the program, and the computations within the curly braces yield a function value that is given back (or returned) to the operating system. In our case, we have written a `main` function that does not expect any arguments (this is indicated by the empty brackets `()` behind `main`) and whose return value is the integer 0. The fact that the return value must be an integer is indicated by the word `int` before `main`. By convention, `return 0` tells the operating system that the program has run successfully (or that we don't care whether it has), while any other value explicitly signals failure.

In a strict mathematical sense, the `main` function of `power8.cpp` is utterly boring. The whole functionality of the program comes from the *effect* of the function. This effect is to read a number from the standard input and write its eighth power to the standard output. The fact that functions can have effects sets C++ apart from many *functional programming languages*.

2.1.5 Values and effects

The value and effect of a function are determined by the C++ semantics. Merely knowing the syntactical rules of writing functions does not tell us anything about values and effects. In this sense, value and effect are purely semantical terms.

For example, we have to *know* that in C++, the character `0` is interpreted as the integer 0 (although this is not difficult to guess). It is also important to understand that value and effect depend on the concrete program state in which the function is called.

2.1.6 Types and functionality

The word `int` is the name of a C++ type. This type is used since the program `power8.cpp` deals with integers. In mathematics, integers are modeled by the ring $(\mathbb{Z}, +, \cdot)$. This algebraic structure defines the integers in terms of their value range (the set $\mathbb{Z}$), and in terms of their functionality (addition and multiplication). In C++, integers can be modeled by the type `int`. Like a "mathematical type", a C++ type has a *name*, a *value range*, and *functionality*, defining what we can do with it. When we refer to a type, we will do so by its name. Note that the name is a syntactical aspect of the type, while value range and functionality are of semantical nature.

Conveniently, C++ contains a number of *fundamental types* (sometimes called built-in types) for typical applications. The type `int` is one of them. The major difference to the "mathematical type" $(\mathbb{Z}, +, \cdot)$ is that `int` has a finite value range only.

2.1.7 Literals

A literal represents a constant value of some type. For example, in line 19 of the program `power8.cpp`, `0` is a literal of type `int`, representing the value 0. For each fundamental type, it is separately defined how its literals look like, and what their values are. A literal can be seen as the syntactical counterpart of a value: it makes the value "visible" in the program.

2.1.8 Variables

The line

```
int a;
```

contains a *declaration* of a *variable*. A variable represents a not necessarily constant value of some type. The variable has a *name*, a *type*, a *value*, and an *address* (typically in the computer's main memory; you can think of the address simply as the position of the variable in the main memory). The purpose of the address is to store and look up the value under this address. The reason for calling such an entity a *variable* is that its value can be changed by modifying the memory content at the corresponding address. In contrast, the name and type remain fixed.

When we refer to a variable, we will do so by its name. The declaration `int a` *defines* a variable with the following characteristics.

name	type	value	address
a	`int`	undefined	chosen by compiler/OS

You might wonder why this is called a *definition* of a, even though it does not define the value of a. But recall that this value depends on the program state, and that the definition fully specifies *how* the value is obtained: look it up at a's address. Saying that a variable *has* a value is therefore somewhat imprecise, but we'll stick to it, just like mathematicians talk about *function value* when they actually mean the value obtained by evaluating the function with concrete arguments. We even go one step further with our sloppiness: if a has value 2, for example, we also say that "a is 2". This is the way that programmers usually talk about variables and their values. We will get to know mechanisms for assigning and changing values of variables in Section 2.1.14.

In C++, it is good general practice to define a variable immediately before it is used for the first time. This practice improves the readability of your programs.

2.1.9 Constants (I promised myself)

Imagine that you wake up on a Sunday morning with a raging hangover, and you promise yourself that you will never touch alcohol again. But if there is no one to check whether you keep that promise, chances are that next Sunday you wake up with the next hangover.

In the ethical language of C++, you can give promises that you must keep. Here is an example of the most basic kind of promise. In writing

```
const int speed_of_light = 299792458;
```

you define a variable `speed_of_light` whose value is $299,792,458$ (m/s). The keyword `const` in front of the declaration is the promise that this variable will never change its value throughout the program. Such a variable is called a *constant*. The type of `speed_of_light` is `const int`, the *const-qualified* version of `int`.

The compiler will check whether you as the programmer keep your promise. Any subsequent lines of code that attempt to store a value under the address of `speed_of_light` will generate an error message. For example, if your program contains the two lines

```
const int speed_of_light = 299792458;
...
speed_of_light = 300000000; // let's make it a bit easier
```

the compiler will remind you of your promise not to change the value of `speed_of_light`. The const promise implies that declarations as in

```
const int speed_of_light;
```

make no sense and are actually forbidden in the main program, as they would yield constants of eternally undefined value.

You don't *have* to give a `const` promise, and since it seems like extra effort at first sight, why should you do it? Because it helps you (and others that may further develop your code in the future) to avoid programming errors. If you define a variable (such as `speed_of_light`) with the intention of keeping its value fixed once and for all, then you should use `const` to tell the compiler about your intention. If you don't, it may (accidentally) happen that a 'constant' (such as the speed of light, or the VAT rate) gets modified; the resulting erroneous behavior of the program may be very hard to discover and to track down.

The whole point of high-level programming languages is to make the programmer's life easier; the compiler is our friend and can help us to avoid many time-consuming errors. The `const` mechanism is like a check digit: by providing additional redundant data (the `const` keyword), we make sure that inconsistencies in the whole data set (the program) are automatically detected.

Also, consistently using `const` makes the program more readable, an aspect whose importance we have already stressed in Section 2.1.2. In reading

```
const int speed_of_light = 299792458;
```

you immediately know that the program is working with a fixed `speed_of_light` throughout. We therefore advocate the following

Const Guideline: Whenever you define a variable in a program, think about whether its value is supposed to change or not. In the latter case, use the const keyword to turn the variable into a constant.

In existing code (of "real programmers"), this guideline is often not followed, and in the context of very short programs, it may even be perceived as pedantic. The authors prefer being pedantic to risking unnecessary programming errors. A program that follows the **Const Guideline** is called *const-correct*.

2.1.10 Identifiers and names

The name of a variable must be an *identifier*, according to the following definition, and it must be different from certain *reserved* names like `int`.

Definition 1 *An* **identifier** *is a sequence of characters composed of the 52 letters* `a...z` *and* `A...Z`*, the 10 digits* `0...9`*, and the underscore (_). The first character has to be a letter.*

A C++ program may also contain other names, for example the *qualified* names `std::cin` and `std::cout`. The C++ syntax specifies what a name is, while the C++ semantics tells us what the respective name refers to in a given context.

2.1.11 Objects

An object is a part of the computer's main memory that is used by the program to store a value. An object has an address, a type, and a value of its type (determined by the memory content at the object's address).

With this definition, a variable can be considered as a named object, but we may also have unnamed objects. Although we can't show an example for an unnamed object at this point, we can argue that unnamed objects are important.

In fact, if you want to write interesting programs, it is absolutely necessary to work with objects that are not named by variables. This can be seen by the following simple thought experiment: suppose that you have written a program that stores a sequence of integers to be read from a file (for example, to sort them afterwards). Now you look at your program and count the number of variables that it contains. Say this number is 31. But in these 31 variables, you can store no more than 31 integers. If your program is of any practical use, it can certainly store a sequence of 32 integers, but then there must be at least one integer that cannot be stored under a variable name.

2.1.12 Expressions

In the program `power8.cpp`, three character sequences stand out, because they look familiar and are chiefly responsible for the functionality of the program: these are the character sequences `a * a` in line 14 and `b * b` in lines 15 and 18.

An expression represents a computation involving other expressions. More precisely, an expression is either a *primary expression*, for example a literal or a name, or it is a *composite expression*. A composite expression is obtained by combining expressions through certain operations, or by putting a pair of parentheses `()` around an expression.

The expression `a * a` is an *arithmetic* expression, involving *numeric* variables (actually, the *names* of the variables, but for the sake of readability, we suppress this subtlety) and the multiplication operator, just like we know it from mathematics. According to our above definition, `a * a` is a composite expression, built from the multiplication operator and the two primary expressions `a` and `a`.

According to the above definition, an expression is a syntactical entity, but it has semantical aspects as well: every expression has a type, a value of this type, and possibly an effect. The type is fixed, but the value and the effect only materialize when the expression gets *evaluated*, meaning that the computation it represents is carried out. Evaluating an expression is the most frequent activity going on while a C++ program is executed; the evaluation computes the value of the expression and carries out its effect (if any).

Type and value of a primary expression are determined by its defining literal, or by type and value of the entity behind its defining name. Primary expressions have no effect. Type, value and effect of a composite expression are determined by the involved operation, depending on the values and effects of the involved sub-expressions. Putting parentheses `()` around an expression yields an expression with the same type, value and

effect.

The expression `a * a`, for example, is of type `int`, and not unexpectedly, its value is the square of the value of `a`. The expression has no effect. The expression `b = b * b`, built from the *assignment operator* and the two expressions `b` and `b * b`, has the same type and value as `b * b`, but it has an additional effect: it assigns the square of `b` back to `b`. (This is a shorthand for the correct, but somewhat clumsy formulation that the new value of `b` is set to the square of the old value of `b`.)

We say that an expression is *evaluated* rather than *executed*, because many expressions do not have an effect, so that their functionality is associated with the value only. Even for expressions with effect, some books use the term *side effect* to emphasize that the important thing is the value. The C++ entities chiefly responsible for effects are the *statements* to which we get below.

We want to remark that the *only* way of accessing an expression's value is to evaluate it, and this also carries out its effect. You cannot get the value without the effect.

2.1.13 Lvalues and rvalues

An lvalue is an expression that has an address. In the program `power8.cpp`, the variable `b` is an lvalue, and its address is the address of the variable `b`.

The value of an lvalue is defined as the value of the object at its address. An lvalue can therefore be viewed as the syntactical counterpart of an object: it gives the object a (temporary) name and makes it "visible" within a C++ program. We also say that the lvalue *refers* to the object at its address.

In particular, every variable is an lvalue. But lvalues provide a means for accessing and changing object values, even without having a corresponding variable. As we will see in Section 2.1.14 below, the expression `std::cout << a` "hidden" in line 18 is such an lvalue.

Every expression that is not an lvalue is an rvalue. For example, literals are rvalues: there is no address associated with the `int`-literal `0`, say. Putting a pair of parenthesis around an lvalue yields an lvalue, and similarly for rvalues.

The terms l*value* and r*value* already indicate that we think about them not so much in terms of expressions, but rather in terms of their values. We will often identify an lvalue with the object it refers to, and an rvalue simply with its value.

2.1.14 Operators

Line 14 of `power8.cpp`, for example, features the binary multiplication operator `*`.

Like a function, an operator expects arguments (here also called *operands*) of specified types, from which it computes a *return value* of a specified type, according to its *functionality*. In addition, these computations may have an effect.

This was the semantical view; on the syntactical level, the operands as well as the composite expression (built from the operator and its operands, see Section 2.1.12), are

expressions; the operator specifies for each of them whether it is an lvalue or an rvalue. If the composite expression is an lvalue, the operator is said to return the object referred to by the lvalue. If the composite expression is an rvalue, the operator simply returns its value.

The number of operands is called the *arity* of the operator. Most operators have arity 1 (unary operators) or 2 (binary operators).

Whenever an rvalue is expected as an operand, it is also possible to provide an lvalue. In this case, the lvalue will simply be interpreted as an rvalue, meaning that its address is only used to look up the value, but *not* to change it. This is known as *lvalue-to-rvalue conversion.* In stating that an operand must be an rvalue, the operator therefore guarantees that the operand's value remains unchanged; by expecting an lvalue, the operator explicitly signals its intention to change the value.

Evaluation of composite expressions. When a composite expression involving an operator gets evaluated, the operands are evaluated first (recall that this also carries out the effects of the operands, if any). Based on the resulting values, the operator computes the value of the composite expression. The latter computations may have additional effects, and all effects together form the effect of the composite expression.

The order in which the operands of a composite expression are evaluated is (with rare exceptions) unspecified, see also Section 2.1.1.

Therefore, if the effect of one operand influences values or effects of other operands, value and effect of the composite expression may depend on the evaluation order. The consequence is that value and effect of the composite expression may be unspecified as well.

Since the compiler is not required to issue a warning in such cases, it is the responsibility of the programmer to avoid any expression whose value or effect depends on the evaluation order of operands.

Operator specifics. What is it that sets operators apart from functions? On the one hand, there is only a finite number of possible operator *tokens* such as $*$ or $=$. Many of these tokens directly correspond to well-known mathematical operator symbols indicating the functionality of the operator. Unfortunately, the token = corresponds to mathematical assignment $:=$, and not to mathematical equality $=$, a constant source of confusion for beginners.

On the other hand, and most conveniently, operator calls do not have to obey the usual function call notation, like in `f(x,y)`. After all, we want to write `a * a` in a program, and not `*(a,a)`. In summary, operators let us write more natural and more readable code.

Four different operators (all of them binary) occur in `power8.cpp`, namely the multiplication operator $*$, the assignment operator =, the input operator >>, and the output operator <<. Let us discuss them in turn.

Multiplication operator. The multiplication operator `*` expects two rvalue operands of some type T, and it returns the product of its two operands as an rvalue. The multiplication operator has no effect on its own.

Assignment operator. The assignment operator = expects an lvalue of some type T as its first operand, and an rvalue of the same type as its second operand. It assigns the value of the second operand to the first operand and returns the first operand as an lvalue. In our program `power8.cpp`, the expression `b = b * b` therefore sets the value of `b` to the square of its previous value, and then returns `b`.

In fact, the letter "l" in the term lvalue stands for the fact that the expression may appear on the *left* hand side of an assignment. Similarly, the term rvalue signals an expression that may appear only on the *right* hand side of an assignment.

Input Operator. In `power8.cpp`, the composite expression `std::cin >> a` in line 11 sets the variable `a` to the next value from the *standard input*, usually the keyboard.

In general, the input operator >> expects as its first operand an lvalue referring to an *input stream*. The second operand is an lvalue of some type T. The operator sets the second operand to the next value read from the input stream and returns the stream as an lvalue.

An input stream represents the state of some input device. We think of this device as producing a continuous stream of data that can be tapped to provide input on demand. Under this point of view, the state of the stream corresponds to the sequence of data not yet read. In setting the value of its second operand, the input operator removes one data item from the stream to reflect the fact that this item has now been read. For this, it is important that the stream comes as an lvalue. Conceptually, an input stream is also considered part of the program state.

How much of the data is read as one item, and how exactly it is interpreted as a value of type T highly depends on the type T of the second operand. For now, it is enough to know that this interpretation is readily defined for the type `int` and for the other fundamental types that we will encounter in the following sections.

In C++, the lvalue `std::cin` refers to the variable `cin` defined in the input/output library, and this variable corresponds to the standard input stream.

It is up to the program's caller to fill the standard input stream with data. For example, suppose that the program was started from a command shell. Then usually, while the program is running, all input to the command shell is forwarded to the program's standard input stream. It is also possible to redirect a program's standard input stream to read data from a file instead.

The fact that the input operator returns the input stream is not accidental, as it allows to build expressions involving chains of input operations, such as `std::cin >> x >> y`. We will discuss this mechanism in detail for the output operator below.

Output Operator. In `power8.cpp`, the composite expression `std::cout << a` in line 18 writes the value of `a` to the *standard output*, usually the computer screen.

In general, the output operator << expects as its first operand an lvalue referring to an *output stream*. The second operand is an rvalue of some type T. The operator writes the value of the second operand to the output stream and returns the output stream as an lvalue.

An output stream represents the state of some output device. We think of this device as storing the continuous stream of output data that is generated by the program. In writing to the stream, the output operator therefore changes the stream state, and this makes it necessary to provide the stream as an lvalue. Conceptually, an output stream is also considered part of the program state.

It depends on the type T in which format the second operand's value is written to the stream; for the type `int` and the other fundamental types, this format is readily defined.

C++ defines a standard output stream `std::cout` and a *standard error* stream `std::cerr` in the input/output library.

It is up to the program's caller to process these output streams. For example, suppose that the program was started from a command shell. Then usually, while the program is running, both standard output stream and standard error stream are forwarded to the command shell. But it is also possible to redirect one or both of these streams to write to a file instead. This can be useful to separate regular output (sent to `std::cout`) from error output (sent to `std::cerr`).

As indicated above for input streams, it is possible to output several values through one expression, as in

```
std::cout << a << "^8 = " << b * b << ".\n"
```

Maybe this looks a bit strange, because there is more than one << operator token and more than two operands; but in mathematics, we also write $a + b + c$ as a shortcut for either $(a + b) + c$ or $a + (b + c)$; because addition is associative, we don't even have to specify which variant we intend.

In C++, such shortcuts are also allowed in order to avoid cluttering up the code with parentheses. But C++ operators are in general not associative, so we have to know the 'logical parentheses' in order to understand the meaning of the shortcut.

The operators >> and << are *left-associative*, meaning that the above expression is logically parenthesized as follows.

```
(((std::cout << a) << "^8 = ") << b * b) << ".\n"
```

Recall that the innermost expression `std::cout << a` is an lvalue referring to the standard output stream. Hence, this expression serves as a legal first operand for the next outer composite expression `(std::cout << a) << "^8 = "` and so on. The full expression therefore outputs the values of *all* expressions occurring after some <<, from left to right. The rightmost of these expressions ends with `\n` that causes a line break.

Warning: In using such nested output expressions, it is very easy to make mistakes

based on false assumptions about the evaluation order. As an example, consider the expression

```
std::cout << (a = 10) << ", " << a
```

where the variable a initially has value 5, say. You might expect that this outputs `10, 10`. But it is equally conceivable that the output is `10, 5`. The latter happens if the *right* operand a of the outermost composite expression

```
((std::cout << (a = 10)) << ", ") << a
```

is evaluated first, and the former happens if the *left* operand (whose effect includes changing the value of a to 10) is evaluated first. On the platform of the authors, the output is `10, 5`.

2.1.15 Statements

A statement is a basic building block of a C++ program, and it usually has an effect. The effect depends on the program state and materializes when the statement is *executed*. As with expressions, we say that a statement *does* something. A statement usually ends with a semicolon and represents a "step" of the program. Statements are executed in top-to-bottom order. The shortest possible statement is the *null statement* consisting only of the semicolon; it has no effect. In a typical program, most statements evaluate one or several expressions.

A statement is not restricted to one line of sourcecode; on the contrary, readability often requires to break up statements into several lines of code. The compiler ignores these line breaks, as long as we do not put them at unreasonable places like in the middle of a name.

In `power8.cpp`, there are three kinds of statements.

Expression statement. Appending a semicolon to an expression leads to an expression statement. It evaluates the expression but does not make use of its value. This is a frequent form of statements, and in our small program, the statement

```
b = b * b;
```

as well as all statements starting with `std::cin` or `std::cout` are expression statements.

Declaration statement. Such a statement introduces a new *name* into the program. This can be the name of a variable of a given type, like in the declaration statements

```
int a;
```

and

```
int b = a * a;
```

A declaration statement consist of a *declaration* and a concluding semicolon. In our program `power8.cpp`, we deal with *variable* declarations; they can be of the form

T x

or

T x = expr

where *T* is a type, *x* is the name of the new variable, and *expr* is an rvalue. A variable declaration is not an expression; for example, it can occur at specific places only. But when it occurs, it behaves like an expression in the sense that a declaration also has an effect and a value. Its effect is to allocate memory for the new variable at some address, and to *initialize* it with the value of *expr*, if present. Its value is the resulting value of the new variable. The declaration is said to *define* the variable.

As in the case of expression statements, a declaration statement carries out the effect of the declaration and ignores its value.

Return statement. Such a statement is of the form

`return` *expression*;

where *expression* is an rvalue. It only occurs within a function. The return statement evaluates *expression*, finishes the function's computations, and puts *expression*'s value at some (temporary) address that the caller of the function can access. Abstracting from these technicalities, we simply say that the statement *returns expression* to the caller.

We have seen only one example so far: the statement `return 0;` returns the value 0 (formally, the literal `0` of value 0) to the operating system which has called the `main` function of our program.

Figure 2 summarizes the syntactical and semantical terms that we have introduced, along with their relations. The figure emphasizes the central role that expressions play in C++.

2.1.16 The first program revisited

If you run the executable file resulting from Program 2.1 for a couple of input values, you will quickly notice that something is weird. For example, on the platform of the authors, the following happens:

```
Compute a^8 for a =? 15
15^8 = -1732076671.
```

Obviously, the eighth power of a positive number cannot be negative, so what is going on? We will discuss this in detail in the next section, but the short explanation is that the type `int` can only deal with numbers up to a certain size. If the mathematical

Figure 2: *Syntactical and semantical terms appearing in our first program* `power8.cpp`. *Purely semantical terms appear in white, purely syntactical terms in dark gray. Mixed terms are drawn in light gray. Solid arrows* A → B *are to be read as "*A *has* B*", while solid lines in the expression part mean that the upper term is more general than the lower one.*

result of a computation exceeds this size, the C++ result will necessarily differ from the mathematical result.

This sounds like bad news; after all, 15^8 is not *such* a big number, and if we cannot even compute with numbers of this size, what can we compute at all? The good news is that the problem is easy to fix. The authors have implemented a type called `ifmp::integer` that is capable of dealing with integers of *arbitrary* size (up to the memory limits, of course). Using this type is very easy, and this is one of the key strengths of C++: we simply have to replace `int` by `ifmp::integer` in our program and in addition include a library that contains the definition of the new type. Here is the accordingly changed program.

```
1 // Program: power8_exact.cpp
```

```
// Raise a number to the eighth power,
// using integers of arbitrary size

#include <iostream>
#include <IFMP/integer.h>

int main()
{
  // input
  std::cout << "Compute a^8 for a =? ";
  ifmp::integer a;
  std::cin >> a;

  // computation
  ifmp::integer b = a * a; // b = a^2
  b = b * b;               // b = a^4

  // output b * b, i.e., a^8
  std::cout << a << "^8 = " << b * b << ".\n";
  return 0;
}
```

Program 2.3: *../progs/lecture/power8_exact.cpp*

Using the above program, you *can* compute the correct value of 15^8:

```
Compute a^8 for a =? 15
15^8 = 2562890625.
```

But also much larger values will work (if you happen to be interested in them):

```
Compute a^8 for a =? 1234567
1234567^8 = 5396563761318393964062660689603780554533710504641.
```

We will not discuss the type `ifmp::integer` any further in this book, and there's no need for it, since it just works like `int` (except that it does not have the size limitations of `int`). But whenever you need (in an exercise or challenge) larger numbers, you are free to use the type `ifmp::integer` instead of `int`.

2.1.17 Details

Commenting. There is a way of writing comments that are not limited to one line of code. Any text enclosed by /* (start of comment) and */ (end of comment) is ignored by the compiler. The initial comment of our program power8.cpp could also have been written as

```
/*
  Program: power8.cpp
```

```
  Raise a number to the power 8.
*/
```

This mechanism may seem useful for longer comments spanning several lines of code, but the problem is that you do not immediately recognize a line in the middle of such a construction as a comment: you always have to look for the enclosing /* and */ to be sure.

Sometimes, /* and */ are used for very short comments within lines of code, like in

```
c = a * /* don't divide! */ b;
```

For readability reasons, we do not advocate this kind of comment, either.

Identifiers starting with an underscore. Occasionally, real-life C++ code contains "identifiers" starting with the underscore character _, although this is not allowed according to Definition 1. The truth is that *the programmer* is not allowed to use such "identifiers"; they are reserved for internal use by the compiler. Compilers should issue at least a warning, when they discover such a badly formed "identifier", but often they just let it pass.

The main function. The main function is an exceptional function in several ways. One particular specialty is that the return statement can be omitted. A main function without a return statement at the end behaves precisely as if it would end with return 0;. This definition has been made for historical reasons mostly; it is an anomaly compared to other functions (which will be discussed later). Therefore, we stick to the explicit return statement and ask you to do the same.

Using directives. It is possible to avoid all std:: prefixes through one additional line of code, a using *directive*. In case of power8.cpp, this would look like in Program 2.4.

```
1  // Program: power8_using.cpp
2  // Raise a number to the eighth power.
3
4  #include <iostream>
5
6  using namespace std;
7
8  int main()
9  {
10   // input
11   cout << "Compute x^8 for x =? ";
12   int a;
13   cin >> a;
14
15   // computation
```

```
16    int b = a * a; // b = a^2
17    b = b * b;     // b = a^4
18
19    // output b * b, i.e., a^8
20    cout << a << "^8 = " << b * b << ".\n";
21    return 0;
22 }
```

Program 2.4: *../progs/lecture/power8_using.cpp*

The using directive is a declaration statement of the form

using namespace X;

It allows us to use all features from namespace X without qualifying them through the prefix X::. This mechanism seems quite helpful at first sight, but it has severe drawbacks that prevent us from using (let alone advocating) it in this book.

Let's start with the major drawback. Namespaces may have a large number of features (in particular, the namespace std has), with a large number of names. cin and cout are two such names from the namespace std. It is very difficult (and also not desirable) to know all these names. On the other hand, it *would* be good to know them in order to avoid conflicts with the names *we* introduce. For example, if we define a variable named cout somewhere in Program 2.4, we are asking for trouble: when we later use the expression cout, it is not clear whether it refers to the standard output stream, or to our variable. We can easily avoid the variable name cout, of course, but we may accidentally introduce another name that also appears in the namespace std. The unfortunate consequence is that in some expression of our program, this name might *not* refer to the feature we introduced, but to a feature of the same name from the standard library. We may end up silently using a feature from the standard library that we don't even know and that we never intended to use. The resulting strange behavior of our program can be very difficult to track down.

In the original program power8.cpp, introducing a name cout (or any other name also appearing in namespace std) does not cause any harm: without the std:: qualification, it can never "accidentally" refer to something from the standard library.

Here is the second drawback of using directives. A large program contains many names, and in order to keep track of, it is desirable that the name "tells" us where it comes from: is it a name we have introduced, or does it come from a library? If so, from which one? With using directives, we lose that information, meaning that the program becomes less readable and more difficult to maintain.

2.1.18 Goals

Dispositional. At this point, you should ...

1) understand the basic syntactical and semantical terms of C++, in particular *expression*, *operator*, *statement*, *lvalue*, *rvalue*, *literal*, *variable*, and *constant*;
2) understand syntax and semantics of the program `power8.cpp`.

Operational. In particular, you should be able to ...

(G1) describe the basic syntactical and semantical terms of C++ (as listed above) in your own words and give examples;

(G2) tell whether a given character sequence is an identifier;

(G3) tell whether a given character sequence is a simple expression, as defined below;

(G4) find out whether a given simple expression is an lvalue or an rvalue;

(G5) evaluate a given simple expression;

(G6) check whether a program of complexity similar to `power8.cpp` correctly uses the `const` keyword, and whether it follows the **Const Guideline** (meaning that it is const-correct).

(G7) read and write programs with functionality similar to `power8.cpp`.

A *simple expression* is an expression which only involves `int`-literals, identifiers, the binary multiplication operator $*$, the assignment operator, and parentheses.

2.1.19 Exercises

Exercise 1 *Which of the following character sequences are not C++ identifiers, and why not?* (G2)

(a)	`identifier`	*(b)*	`int`	*(c)*	`x_i`	*(d)*	`4x__`
(e)	`A99_`	*(f)*	`_tmp`	*(g)*	`T#`	*(h)*	`x12b`

Exercise 2 *Which of the following character sequences are not C++ expressions, and why not? Here,* `a` *and* `b` *are variables of type* `int`. (G3)

(a)	`1*(2*3)`	*(b)*	`a=(b=5)`	*(c)*	`1=a`	*(d)*	`(a=1)`
(e)	`(a=5)*(b=7)`	*(f)*	`(1`	*(g)*	`(a=b)*(b=5)`	*(h)*	`(a*3)=(b*5)`

Exercise 3 *For all of the expressions that you have identified in Exercise 2, decide whether these are lvalues or rvalues, and explain your decisions.* (G4)

Exercise 4 *Determine the values of the expressions that you have identified in Exercise 2 and explain how these values are obtained. Which of these values are unspecified and can therefore not be determined uniquely?* (G5)

Exercise 5 *Which of the following (rather stupid) programs are syntactically incorrect (w.r.t. the usage of* `const`*), and why? Among the correct ones, which programs do not adhere to the* **Const Guideline** *of Section 2.1.9, and why?* (G6).

a)
```
int main ()
{
  const int a = 5;
  int b = a;
  b = b*2;
  return 0;
}
```

b)
```
#include <iostream>
int main ()
{
  const int a = 5;
  std::cin >> a;
  std::cout << a + 5;
  return 0;
}
```

c)
```
#include <iostream>
int main ()
{
  const int a;
  int b;
  std::cin >> b;
  std::cout << a;
  return 0;
}
```

d)
```
int main ()
{
  const int a = 5;
  int b = 2*a;
  int c = 2*b;
  b = b*b;
  return 0;
}
```

e)
```
int main ()
{
  const int a = 5;
  const int b = (a = 6);
  return 0;
}
```

f)
```
int main ()
{
  const int a = 5;
  a = 5;
  return 0;
}
```

g)
```
#include<iostream>
int main ()
{
  int a = 5;
  a = a*a;
  int b = a;
  b = b*b;
  const int c = b;
  std::cout << c*c;
  return 0;
}
```

Exercise 6 *What is the smallest natural number that is divisible by all numbers between 2 and* (G7)

a) 10 ?

b) 20 ?

c) 30 ?

The result is also known as the *least common multiple* of the respective numbers.

Note: This exercise does not require you to write a program, but you may use a program to help you in the computations.

Exercise 7 *Write a program* `multhree.cpp` *that reads three integers* a, b, c *from standard input and outputs their product* abc. (G7)

Exercise 8 *Write a program* `power20.cpp` *that reads an integer* a *from standard input and outputs* a^{20} *using at most five multiplications.* (G7)

Exercise 9 *During an electronic transmission, the following C++ program got garbled. As you can see, the layout got messed up, but at the same time, some errors got introduced as well.* (G4)(G7)

```
#include <iostream>
int main[]{int a;int b;int c;std::cin >> a;
cin >> b;c = a * b;std::cout << c*c;return 0;}
```

a) Write the program down in a well-formatted way.

b) The program contains two syntax errors. Fix them! What does the fixed program do?

c) The (fixed) program contains a number of composite expressions. List them all, and decide for each composite expression whether it is an rvalue or an lvalue. Recall that a composite expression may consist of primary expressions, but also of other composite expressions.

d) Add sensible comments to the program; most notably, there should be a comment in the beginning that says what the program does.

e) Move the variable declarations to their logical places (immediately before first use), and make the program const-correct.

Exercise 10 *Write a program* `age_verification.cpp` *that defines the current year (e.g. 2009) as a constant, and then outputs the age groups that are not allowed to buy alcohol. (If you are a manager at Coop, you can now post this at every checkout). For the year 2009, the output should be* (G7)

```
No alcohol to people born in the years 1992 - 2009!
For people born in 1991, check the id!
```

2.1.20 Challenges

Exercise 11 *The obvious "slow" method for computing the eighth power of an integer* a *needs seven multiplications. Program 2.1 requires only three, and we believe that this should be faster. The goal of this challenge is to find out: how much faster? For example, if we compute the eighth power of a* $10,000$*-digit number using both methods, what will be the difference in runtime? Using the type* `int`*, though, we cannot correctly compute with* $10,000$*-digit numbers (as you will easily notice when you start* `power8.cpp` *with somewhat larger inputs, see Section 2.1.16). For this reason, you should use the type* `ifmp::integer` *for your computations.*

Write two programs, `power8_slow.cpp` *and* `power8_fast.cpp` *that compute the eighth power of an integer with* 7 *and* 3 *multiplications, respectively (over the exact type* `ifmp::integer`*). Since we want to measure runtimes, there should be no output (you don't want to read it, anyway). In order to be able to use the same large inputs for both programs, it is beneficial to have the programs read the input from a file. For example, if you have a file* `power8.dat` *with contents* `1234567`*, you can tell the program* `power8_exact.cpp` *to read its input from this file by starting it with the command*

```
./power8_exact < power8.dat
```

Now that you have both programs, create an input file `power8.dat`*, and fill it with larger and larger numbers (each time doubling the number of digits, for example). Then measure the times taken by each of the programs* `power8_slow.cpp` *and* `power8_fast.cpp` *on these inputs. You can simply do this using your watch (for sufficiently many digits both programs will be slow enough), or you can start the programs like this:*

```
time ./power8_fast < power8.dat
```

This command will run the program and afterwards output the number of seconds that it took (the first number, the one ending in `u`*, is the relevant one).*

What do you observe? Is `power8_fast.cpp` *more than twice as fast as* `power8_slow.cpp` *(this is what you might expect from the number of multiplications)? Or do you observe a speedup factor quite different from 2? And is this factor stable as the input numbers get larger?*

Whatever you observe, try to explain your observations!

Exercise 12 *Let $\ell(n)$ be the smallest number of multiplications that are needed in order to compute the n-th power a^n of an integer a. Since $\ell(n)$ may depend on what we consider as a "computation", we make $\ell(n)$ well-defined by restricting to the following kind of computations. Let a_0 denote the input number a. A computation consists of t steps, where t is some natural number, and step $i, 1 \leq i \leq t$ has the form*

$a_i = a_j * a_k$

with $j, k < i$. The computation is correct if $a_t = a^n$. For example, to compute a^8 in three steps (three multiplications) as in `power8.cpp`*, we can use the computation*

$a_1 = a_0 * a_0$
$a_2 = a_1 * a_1$
$a_3 = a_2 * a_2$

Now, $\ell(n)$ is defined as the smallest value of t such that there exists a correct t-step computation for a^n.

a) In the above model of computation, prove that for all $n \geq 1$,

$$\lambda(n) \leq \ell(n) \leq \lambda(n) + \nu(n) - 1,$$

where $\lambda(n)$ is one less than the number of significant bits of n in the binary representation of n (see Section 2.2.8), and $\nu(n)$ is the number of 1's in the binary representation of n. For example, the binary representation of 20 is 10100, and hence $\lambda(20) = 4$ and $\nu(20) = 2$, resulting in $\ell(n) \leq 5$.

b) Either prove that the upper bound in a) is always best possible, or find a value n^* such that $\ell(n^*) < \lambda(n^*) + \nu(n^*) - 1$.

2.2 Integers

Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk.

Leopold Kronecker, in a lecture to the Berliner Naturforscher-Versammlung (1886)

This section discusses the types `int` *and* `unsigned int` *for representing integers and natural numbers, respectively. You will learn how to evaluate arithmetic expressions over both types. You will also understand the limitations of these types, and—related to this—how their values can be represented in the computer's memory.*

Here is our next C++ program. It asks the user to input a temperature in degrees Celsius, and outputs it in degrees Fahrenheit. The conversion is defined by the following formula.

$$\text{Degrees Fahrenheit} = \frac{9 \cdot \text{Degrees Celsius}}{5} + 32.$$

```
// Program: fahrenheit.cpp
// Convert temperatures from Celsius to Fahrenheit.

#include <iostream>

int main()
{
  // Input
  std::cout << "Temperature in degrees Celsius =? ";
  int celsius;
  std::cin >> celsius;

  // Computation and output
  std::cout << celsius << " degrees Celsius are "
            << 9 * celsius / 5 + 32 << " degrees Fahrenheit.\n";
  return 0;
}
```

Program 2.5: *../progs/lecture/fahrenheit.cpp*

If you try out the program on the input of 15 degrees Celsius, you will get the following output.

```
15 degrees Celsius are 59 degrees Fahrenheit.
```

This output is produced when the expression statement in lines 14–15 of the program is executed. Here we focus on the evaluation of the arithmetic expression

```
9 * celsius / 5 + 32
```

in line 15. This expression contains the primary expressions 9, 5, 32, and `celsius`, where `celsius` is a variable of type `int`. This fundamental type is one of the *arithmetic types* in C++.

Literals of type int. 9, 5 and 32 are decimal literals of type `int`, with their values immediately apparent. Decimal literals of type `int` consist of a sequence of digits from 0 to 9, where the first digit must not be 0. The value of a decimal literal is the decimal number represented by the sequence of digits. There are no literals for negative integers. You can get value -9 by writing -9, but this is a composite expression built from the unary subtraction operator (Section 2.2.4) and the literal 9.

2.2.1 Associativity and precedence of operators

The evaluation of an expression is to a large extent governed by the *associativities* and *precedences* of the involved operators. In short, associativities and precedences determine the logical parentheses in an expression that is not, or only incompletely, parenthesized. We have already touched associativity in connection with the output operator in Section 2.1.14.

C++ allows incompletely parenthesized expressions in order to save parentheses at obvious places. This is like in mathematics, where we write $3 + 4 \cdot 5$ when we mean $3+(4\cdot 5)$. We also write $3+4+5$, even though it is not a priori clear whether this means $(3+4)+5$ or $3+(4+5)$. Here, the justification is that addition is *associative*, so it does not matter which variant we mean.

The price to pay for less parentheses is that we have to know the *logical* parentheses. But this is a moderate price, since the two rules that are used most frequently are quite intuitive and easy to remember. Also, there is always the option of explicitly adding parentheses in case you are not sure where C++ would put them. Let us start with the two essential rules for arithmetic expressions.

Arithmetic Evaluation Rule 1: Multiplicative operators have higher precedence than additive operators.

The expression `9 * celsius / 5 + 32` involves the multiplication operator `*`, the division operator `/`, and the addition operator `+`. All three are binary operators. In C++ as in mathematics, the multiplicative operators `*` and `/` have *higher* precedence than the additive operators `+` and `-`. We also say that multiplicative operators *bind* more strongly than additive ones. This means, our expression contains the logical parentheses `(9 * celsius / 5) + 32`: it is a composite expression built from the addition operator and its operands `9 * celsius / 5` and 32.

Arithmetic Evaluation Rule 2: Binary arithmetic operators are left associative.

In mathematics, it does not matter how the sub-expression `9 * celsius / 5` is parenthesized. But in C++, it is done from *left to right*, that is, the two leftmost sub-expressions are grouped together. This is a consequence of the fact that the binary arithmetic operators are defined to be *left* associative. The expression `9 * celsius / 5` is therefore logically parenthesized as `(9 * celsius) / 5`, and our original expression has to be read as

```
((9 * celsius) / 5) + 32
```

Identifying the operators in an expression. There is one issue we haven't discussed yet, namely that *different* C++ operators may have the *same* token. For example, `-` can be a binary operator as in `3 - 4`, but it can also be a unary operator as in `-5`. Which one is meant must be inferred from the context. Usually, this is clear, and in cases where it is not (but also in other cases), it is probably a good idea to add some extra parentheses to make the expression more readable (see also the Details section below).

Let us consider another concrete example, the expression `-3 - 4`. It is clear that the first `-` must be unary (there is no left hand side operand), while the second one is binary (there are operands on both sides). But is this expression logically parenthesized as `-(3 - 4)`, or as `(-3) - 4`? Since we get different values in both cases, we better make sure that we know the answer.

The correct logical parentheses are

$$((-3) - 4),$$

so the value of the expression `-3 - 4` is -7. This follows from the third most important rule for arithmetic expressions.

Arithmetic Evaluation Rule 3: Unary operators + and - have higher precedence than their binary counterparts.

By using (explicit) parentheses as in `9 * (celsius + 5) * 32`, precedences can be overruled. To get the logical parentheses for such a partially parenthesized expression, we apply the rules from above, considering the already parenthesized parts as operands. In the example, this leads to the logical parentheses `(9 * (celsius + 5)) * 32`.

The Details section discusses how to parenthesize a general expression involving arbitrary C++ operators, using their arities, precedences and associativities.

2.2.2 Expression trees

In every composite expression, the logical parentheses determine a unique "top-level" operator, namely the one that appears within the smallest number of parentheses. The

expression is then a composite expression, built from the top-level operator and its operands that are again expressions.

The structure of an expression can nicely be visualized in the form of an *expression tree*. In Figure 3, the expression tree for the expression `9 * celsius / 5 + 32` is shown.

Figure 3: *An expression tree for* `9 * celsius / 5 + 32` *and its logical parentheses* `((9 * celsius) / 5) + 32`*. Nodes are labeled from one to seven.*

How do we get this tree? The expression itself defines the *root* of the tree, and the operands of the top-level operator become the root's *children* in the tree. Each operand then serves as the root of another subtree. When we reach a primary expression, it defines a *leaf* in the tree, with no further children.

2.2.3 Evaluating expressions

From an expression tree we can easily read off the possible *evaluation sequences* for the arithmetic expression. Such a sequence contains all sub-expressions occurring in the tree, ordered by their time of evaluation. For this sequence to be valid, we have to make sure that we evaluate an expression only *after* the expressions corresponding to *all* its children have been evaluated. By looking at Figure 3, this becomes clear: before evaluating `9 * celsius`, we have to evaluate `9` and `celsius`, otherwise, we don't have enough information to perform the evaluation.

When we associate the evaluation sequence with the corresponding sequence of nodes in the tree, a valid node sequence *topologically sorts* the tree. This means that every node in the sequence occurs only after all its children have occurred. In Figure 3, for example, the node sequence $(1, 2, 5, 3, 6, 4, 7)$ induces a valid evaluation sequence. Assuming that the variable `celsius` has value 15, we obtain the following evaluation sequence. (In each step, the sub-expression to be evaluated next is marked by a surrounding box).

$$
\begin{aligned}
\boxed{9}\ \texttt{*}\ \texttt{celsius}\ \texttt{/}\ 5 + 32 &\longrightarrow^{1} 9\ \texttt{*}\ \boxed{\texttt{celsius}}\ \texttt{/}\ 5 + 32 \\
&\longrightarrow^{2} \boxed{9\ \texttt{*}\ 15}\ \texttt{/}\ 5 + 32 \\
&\longrightarrow^{5} 135\ \texttt{/}\ \boxed{5} + 32 \\
&\longrightarrow^{3} \boxed{135\ \texttt{/}\ 5} + 32 \\
&\longrightarrow^{6} 27 + \boxed{32} \\
&\longrightarrow^{4} \boxed{27\ \texttt{+}\ 32} \\
&\longrightarrow^{7} 59
\end{aligned}
$$

The sequence $(1,2,3,4,5,6,7)$ is another valid node sequence, inducing a different evaluation sequence; the resulting value of 59 is the same. There are much more evaluation sequences, of course, and it is unspecified by the C++ standard which one is to be used.

In our small example, all possible evaluation sequences will result in value 59, but it is also not hard to write down expressions whose values and effects depend on the evaluation sequence being chosen (see Exercise 2*(g)*, Exercise 15*(h)*, and the Details section below). A program that contains such an expression might exhibit unspecified behavior. But through good programming style, this issue is easy to avoid, since it typically only occurs when one tries to squeeze too much functionality into a single line of code.

2.2.4 Arithmetic operators on the type int

In the program `fahrenheit.cpp`, we have already encountered the multiplicative operators `*` and `/`, as well as the binary addition operator `+`. Its obvious counterpart is the binary subtraction operator `-`.

Table 1 lists arithmetic operators (and the derived *assignment operators*) that are available for the type `int`, with their arities, precedences and associativities. The actual numbers that appear in the precedence column are not relevant: it is the *order* among precedences that matters.

Let us discuss the functionalities of these operators in turn, where `*`, `+` and `-` are self-explanatory. But already the division operator requires a discussion.

The division operator. According to the rules of mathematics, we could replace the expression

```
9 * celsius / 5 + 32
```

by the expression

```
9 / 5 * celsius + 32
```

without affecting its value and the functionality of the program `fahrenheit.cpp`. But if we run the program with the latter version of the expression on the input of 15 degrees Celsius, we get the following output:

Description	Operator	Arity	Prec.	Assoc.
post-increment	++	1	17	left
post-decrement	−−	1	17	left
pre-increment	++	1	16	right
pre-decrement	−−	1	16	right
sign	+	1	16	right
sign	−	1	16	right
multiplication	*	2	14	left
division (integer)	/	2	14	left
modulus	%	2	14	left
addition	+	2	13	left
subtraction	−	2	13	left
assignment	=	2	3	right
mult assignment	*=	2	3	right
div assignment	/=	2	3	right
mod assignment	%=	2	3	right
add assignment	+=	2	3	right
sub assignment	−=	2	3	right

Table 1: *Arithmetic and assignment operators for the type* `int`*. Each increment or decrement operator expects an lvalue. The composite expression is an lvalue (pre-increment and pre-decrement), or an rvalue (post-increment and post-decrement). Each assignment operator expects an lvalue as left operand and an rvalue as right operand; the composite expression is an lvalue. All other operators involve rvalues only and have no effects.*

```
15 degrees Celsius are 47 degrees Fahrenheit.
```

This result is fairly different from our previous (and correct) result of 59 degrees Fahrenheit, so what is going on here? The answer is that the binary division operator / on the type `int` implements the *integer division*, in mathematics denoted by div. This does not correspond to the regular division where the quotient of two integers is in general a non-integral rational number.

The modulus operator. The remainder of the integer division can be obtained with the binary *modulus* operator %, in mathematics denoted by mod. The mathematical rule

$$a = (a \operatorname{div} b)b + a \bmod b$$

also holds in C++: for example, if a and b are variables of type `int`, the value of b being non-zero, the expression

```
(a / b) * b + a % b
```

has the same value as a. The modulus operator is considered as a multiplicative operator and has the same precedence (14) and associativity (left) as the other two multiplicative operators * and /.

If both a and b have non-negative values, then a % b has a non-negative value as well. This implies that the integer division rounds *down* in this case. If (at least) one of a or b has a negative value, it is implementation defined whether division rounds up or down. Note that by the identity a = (a / b) * b + a % b, the rounding mode for division also determines the functionality of the modulus operator. If b has value 0, the values of a / b and a % b are undefined.

Coming back to our example (and taking precedences and associativities into account), we get the following valid evaluation sequence for our alternative Celsius-to-Fahrenheit conversion.

```
9 / 5 * celsius + 32 ⟶ 1 * celsius + 32
                     ⟶ 1 * 15 + 32
                     ⟶ 15 + 32
                     ⟶ 47
```

Here we see the "error" made by the integer division: 9 / 5 has value 1.

Unary additive operators. We have already touched the unary − operator, and this operator does what one expects: the value of the composite expression −*expr* is the negative of the value of *expr*. There is a unary + operator, for completeness, although its "functionality" is non-existing: the value of the composite expression +*expr* is the same as the value of *expr*.

Increment and decrement operators. Each of the tokens ++ and -- is associated with two *distinct* unary operators that differ in precedence and associativity.

The pre-increment ++ and the pre-decrement -- are right associative, For a unary operator, this means that the argument appears to the right of the operator token. The effect of the composite expressions ++*expr* and --*expr* is to increase (decrease, respectively) the value of *expr* by 1. Then, the object referred to by *expr* is returned. For this to make sense, *expr* has to be an lvalue. We also say that pre-increment is ++ in *prefix notation*, and similarly for --.

The post-increment ++ and the post-decrement -- are left associative. As before, the effect of the composite expressions *expr*++ and *expr*-- is to increase (respectively decrease) the value of *expr* by 1, and *expr* has to be an lvalue for this to work. The return value, though, is an rvalue corresponding to the *old* value of *expr* *before* the increment or decrement took place. We also say that post-increment is ++ in *postfix notation*, and similarly for --.

The difference between the increment operators in pre- and postfix notation is illustrated in the following example program.

```
#include <iostream>
int main() {
  int a = 7;
  std::cout << ++a << "\n"; // outputs 8
  std::cout << a++ << "\n"; // outputs 8
  std::cout << a   << "\n"; // outputs 9
  return 0;
}
```

You may argue that the increment and decrement operators are superfluous, since their functionality can be realized by combining the assignment operator (Section 2.1.14) with an additive operator. Indeed, if a is a variable, the expression ++a is equivalent in value and effect to the expression a = a + 1. There is one subtlety, though: if *expr* is a general lvalue, ++*expr* is *not* necessarily equivalent to *expr* = *expr* + 1. The reason is that in the former expression, *expr* is evaluated once only, while in the latter, it is evaluated *twice*. If *expr* has an effect, this can make a difference.

On the other hand, this subtlety is not the reason why increment and decrement operators are so popular and widely used in C++. The truth is that incrementing or decrementing values by 1 are such frequent operations in typical C++ code that it pays off to have shortcuts for them.

Prefer pre-increment over post-increment. The statements ++i; and i++; are obviously equivalent, as their effect is the same and the value of the expression is not used. You can exchange them with each other arbitrarily without affecting the behavior of the surrounding program. Whenever you have this choice, you should opt for the pre-increment operator. Pre-increment is the simpler operation because the value of ++i can simply be read off the variable i. In contrast, the post-increment has to "remember" the original

value of `i`. As pre-increment is simpler, it also tends to be more efficient.

Remark: We write "pre-increment tends to be more efficient" because in many cases the compiler realizes when the value of an expression is not used. In such a case, the compiler may choose on its own to replace the post-increment in the source code by a "pre-increment" in machine language as an optimization. However, there is absolutely no benefit in choosing a post-increment where a pre-increment would do as well. In this case, you should take the burden from the compiler and optimize by yourself.

Also, post-increment and post-decrement are the only unary C++ operators that are left associative. This makes their usage appear somewhat counterintuitive.

Why C++ should rather be called ++C. The language C++ is a further development of the language C. And indeed, you can read the *effect* of the "expression" C++ as "take C one step further". But the name C++ is still a misnomer, because the *value* of the "expression" C++ is the plain old C, after all. A better name would be ++C: the "value" of this is really the new language created from taking C one step further. Bjarne Stroustrup, the designer of C++, writes that "Connoisseurs of C semantics find C++ inferior to ++C".

Assignment operators. The assignment operator = is available for all types, see Section 2.1.14. But there are specific operators that combine the arithmetic operators with an assignment. These are the binary operators `+=`, `-=`, `*=`, `/=` and `%=`. The expression *expr1* `+=` *expr2* has the effect of adding the value of *expr2* (an rvalue) to the value of *expr1* (an lvalue). The object referred to by *expr1* is returned. This is a generalization of the pre-increment: the expression `++`*expr* is equivalent to *expr* `+= 1`. As before, *expr1* `+=` *expr2* is *not* equivalent to *expr1* `=` *expr1* `+` *expr2* in general, since the latter expression evaluates *expr1* twice.

The operators `-=`, `*=`, `/=` and `%=` work in the same fashion, based on the subtraction, multiplication, division, and modulus operator, respectively.

All the assignment operators have precedence 4, i.e. they bind more weakly than the other arithmetic operators. This is quite intuitive: `a=b*c-d`, say, means `a=(b*c-d)`.

2.2.5 Value range

A variable of type `int` is associated with a *fixed* number of memory cells, and therefore also with a fixed number of bits, say b. We call this a b-*bit representation*.

Such a representation implies that an object of type `int` can assume only finitely many different values. Since every bit can independently have two states, the maximum number of representable values is 2^b, and the actual value range is defined as the set

$$\{-2^{b-1}, -2^{b-1}+1, \ldots, -1, 0, 1, \ldots, 2^{b-1}-1\} \subset \mathbb{Z}$$

of 2^b numbers. The C++ standard does not prescribe this, but any different choice of value range would be somewhat unreasonable, given other requirements imposed by the

standard. You can find out the smallest and largest int values on your platform, using the library limits. The corresponding code is given in Program 2.6.

```
// Program: limits.cpp
// Output the smallest and the largest value of type int.

#include <iostream>
#include <limits>

int main()
{
  std::cout << "Minimum int value is "
            << std::numeric_limits<int>::min() << ".\n"
            << "Maximum int value is "
            << std::numeric_limits<int>::max() << ".\n";
  return 0;
}
```

Program 2.6: *../progs/lecture/limits.cpp*

When you run the program limits.cpp on a 32-bit system, you may get the following output.

```
Minimum int value is -2147483648.
Maximum int value is 2147483647.
```

Indeed, as $2147483647 = 2^{31}-1$, you can deduce that the number of bits used to represent an int value on your platform is 32. At this point, you are not supposed to understand expressions like std::numeric_limits<int>::min() in detail, but we believe that you get their idea. We cannot resist to note in passing that $2147483647 = 2^{31}-1$ is the number from Mersenne's conjecture that Euler has proved to be prime in 1772, see Section 1.1.

It is clear that the arithmetic operators (except the unary + and the binary / and %) cannot work exactly like their mathematical counterparts, even when their arguments are restricted to representable int values. The reason is that the values of composite expressions constructed from these operators can under- or overflow the value range of the type int. The most obvious such example is the expression 2147483647+1. As we have just seen, its mathematically correct value of 2147483648 may not be representable over the type int on your platform, in which case you will inevitably get some other value.

Such under- and overflows are a severe problem in many practical applications, but it would be an even more severe problem not to know that they can occur.

2.2.6 The type unsigned int

An object of type int can have negative values, but often we only work with natural numbers. (For us, the set $\mathbb{N}$ of natural numbers starts with 0, $\mathbb{N} = \{0, 1, 2, \ldots\}$.) Using

a type that represents only non-negative values allows to extend the range of positive values without using more bits. C++ provides such a type, it is called `unsigned int`. On this type, we have all the arithmetic operators we also have for `int`, with the same arities, precedences and associativities. Given a b-bit representation, the value range of `unsigned int` is the set

$$\{0, 1, \ldots, 2^b - 1\} \subset \mathbb{N}$$

of 2^b natural numbers. Indeed, when you replace all occurrences of `int` by `unsigned int` in the program `limits.cpp`, it may produce the following output.

```
Minimum value of an unsigned int object is 0.
Maximum value of an unsigned int object is 4294967295.
```

Literals of type `unsigned int` look like literals of type `int`, followed by either the letter `u` or `U`. For example, `127u` and `0U` are literals of type `unsigned int`, with their values immediately apparent.

2.2.7 Mixed expressions and conversions

Expressions may involve sub-expressions of type `int` *and* of type `unsigned int`. For example `17+17u` is a legal arithmetic expression, but what are its type and value? In such *mixed expressions*, the operands are implicitly *converted* to the *more general* type. By the C++ standard, the more general type is `unsigned int`. Therefore, the expression `17+17u` is of type `unsigned int` and gets evaluated step by step as

`17+17u` $\longrightarrow$ `17u+17u` $\longrightarrow$ `34u`

This might be somewhat confusing, since in mathematics, it is just the other way around: $\mathbb{Z}$ (the set of integers) is more general than $\mathbb{N}$ (the set of natural numbers). We are not aware of any deeper justification for the way it is done in C++, but at least the conversion is well-defined:

Non-negative `int` values are "converted" to the same value of type `unsigned int`; negative `int` values are converted to the `unsigned int` value that results from (mathematically) adding 2^b. This rule establishes a bijection between the value ranges of `int` and `unsigned int`.

Implicit conversions in the other direction may also occur but are not always well-defined. Consider for example the declarations

```
int a = 3u;
int b = 4294967295u;
```

The value of `a` is 3, since this value is in the range of the type `int`. But if we assume the 32-bit system from above, the value of `b` is implementation defined according to the C++ standard, since the literal `4294967295` is outside the range of `int`.

2.2.8 Binary representation

Assuming b-bit representation, we already know that the type `int` covers the values

$$-2^{b-1}, \ldots, 2^{b-1} - 1,$$

while `unsigned int` covers

$$0, \ldots, 2^b - 1.$$

In this subsection, we want to take a closer look at how these values are represented in memory, using the b available bits. This will also shed more light on some of the material in the previous subsection.

The *binary expansion* of a natural number $n \in \mathbb{N}$ is the sum

$$n = \sum_{i=0}^{\infty} b_i 2^i,$$

where the b_i are uniquely determined coefficients from $\{0, 1\}$, with only finitely many of them being nonzero. For example,

$$13 = 1 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^3.$$

The sequence of the b_i in reverse order is called the binary representation of n. The binary representation of 13 is 1101, for example.

Conversion decimal $\rightarrow$ binary. The identity

$$n = \sum_{i=0}^{\infty} b_i 2^i = b_0 + \sum_{i=1}^{\infty} b_i 2^i = b_0 + \sum_{i=0}^{\infty} b_{i+1} 2^{i+1} = b_0 + 2 \underbrace{\sum_{i=0}^{\infty} b_{i+1} 2^i}_{=:n'}$$

provides a simple algorithm to compute the binary representation of a given decimal number $n \in \mathbb{N}$. The least significant coefficient b_0 of the binary expansion of n is $n \bmod 2$. The other coefficients b_i, $i \geq 1$, can subsequently be extracted by applying the same technique to $n' = (n - b_0)/2$.

For example, for $n = 14$ we get $b_0 = 14 \bmod 2 = 0$ and $n' = (14 - 0)/2 = 7$. We continue with $n = 7$ and get $b_1 = 7 \bmod 2 = 1$ and $n' = (7 - 1)/2 = 3$. For $n = 3$ we get $b_2 = 3 \bmod 2 = 1$ and $n' = (3 - 1)/2 = 1$ which leaves us with $n = b_3 = 1$. In summary, the binary representation of 14 is $b_3 b_2 b_1 b_0 = 1110$.

Conversion binary $\rightarrow$ decimal. To convert a given binary number $b_k \ldots b_0$ into decimal representation, we can once again use the identity from above.

$$\sum_{i=0}^{k} b_i 2^i = b_0 + 2 \sum_{i=0}^{k-1} b_{i+1} 2^i = \ldots = b_0 + 2(b_1 + 2(b_2 + 2(\cdots + 2b_k) \ldots))$$

For example, to convert the binary number $b_4b_3b_2b_1b_0 = 10100$ into decimal representation, we compute

$$(((b_4 \cdot 2 + b_3) \cdot 2 + b_2) \cdot 2 + b_1) \cdot 2 + b_0 \;=\; (((1 \cdot 2 + 0) \cdot 2 + 1) \cdot 2 + 0) \cdot 2 + 0 \;=\; 20.$$

Representing unsigned int values. Since every `unsigned int` value

$$n \in \{0, \ldots, 2^b - 1\}$$

has a binary representation of length exactly b (filling up with leading zeros), this binary representation is a canonical format for storing n using the b available bits. Like the value range itself, this storage format is not explicitly prescribed by the C++ standard, but hardly anything else makes sense in practice. As there are 2^b `unsigned int` values, and the same number of b-bit patterns, each pattern encodes one value. For $b = 3$, this looks as follows.

n	representation
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

Representing int values. A common way of representing `int` values using the same b bits goes as follows. If the value n is non-negative, we store the binary representation of n itself—a number from

$$\{0, \ldots, 2^{b-1} - 1\}.$$

That way we use all the b-bit patterns that start with 0.

If the value n is negative, we store the binary representation of $n + 2^b$, a number from

$$\{2^{b-1}, \ldots, 2^b - 1\}.$$

This yields the missing b-bit patterns, the ones that start with 1. For $b = 3$, the resulting representations are

n	representation
-4	100
-3	101
-2	110
-1	111
0	000
1	001
2	010
3	011

This is called the *two's complement* representation. In this representation, adding two `int` values n and n' is very easy: simply add the representations according to the usual rules of binary number addition, and ignore the overflow bit (if any). For example, to add -2 and -1 in case of $b = 3$, we compute

$$\begin{array}{rr} & 110 \\ + & 111 \\ \hline & 1101 \end{array}$$

Ignoring the leftmost overflow bit, this gives 101, the representation of the result -3 in two's complement. This works since the binary number behind the encoding of n is either n or $n + 2^b$. Thus, when we add the binary numbers for n and n', the result is congruent to $n + n'$ modulo 2^b and therefore agrees with $n + n'$ in the b rightmost bits.

Using the two's complement representation we can now better understand what happens when a negative `int` value n gets converted to type `unsigned int`. The standard specifies that for this, n has to be incremented by 2^b. But under the two's complement, the negative `int` value n and the resulting positive `unsigned int` value $n + 2^b$ have the same representation! This means that the conversion is purely conceptual, and no actual computation takes place.

The C++ standard does not prescribe the use of the two's complement, but the rule for conversion from `int` to `unsigned int` is clearly motivated by it.

2.2.9 Integral types

There is a number of other fundamental types to represent signed and unsigned integers, see the Details section. These types may differ from `int` and `unsigned int` with respect to their value range. All these types are called *integral types*, and for each of them, all the operators in Table 1 (Page 51) are available, with the same arities, precedences, associativities and functionalities (up to the obvious limits dictated by the respective value ranges).

2.2.10 Details

Literals. There are also non-decimal literals of type `int`. An *octal* literal starts with the digit `0`, followed by a sequence of digits from `0` to `7`. The value is the octal number represented by the sequence of digits following the leading `0`. For example, the literal `011` has value $9 = 1 \cdot 8^1 + 1 \cdot 8^0$.

Hexadecimal literals start with `0x`, followed by a sequence of digits from `0` to `9` and letters from `A` to `F` (representing the hexadecimal digits of values $10, \ldots, 15$). The value is the hexadecimal number represented by the sequence of digits and letters following the leading `0x`. For example, the literal `0x1F` has value $31 = 1 \cdot 16^1 + 15 \cdot 16^0$.

Logically parenthesizing a general expression. Given an expression that consists of a sequence of operators and operands, we want to deduce the logical parentheses. For each operator in the sequence, we know its arity, its precedence (a number between 1 and 18, see Table 1 on Page 51 for the arithmetic operators), and its associativity (left or right). In case of a unary operator, the associativity specifies on which side of the operator its operand is to be found.

Let us consider the following abstract example to emphasize that what we do here is completely general and not restricted to arithmetic expressions.

expression	x_1	op_1	x_2	op_2	x_3	op_3	op_4	x_4
arity		2		2		2	1	
precedence		4		13		13	16	
associativity		r		l		l	r	

Here is how the parentheses are obtained: for each operator, we identify its *leading* operand, defined as the left hand side operand for left associative operators, and as the right hand side operand otherwise. The leading operand for op_i includes everything to the relevant side between op_i and the next operator of *lower* precedence than op_i. In other words, everything in between these two operators is "grabbed" by the "stronger" operator.

In our example, the leading operand of op_3 is the subsequence $x_2 \;\; op_2 \;\; x_3$ to the left of op_3, since the next operator of lower precedence to the left of op_3 is op_1.

In the case of binary operators, we also find the *secondary* operand, the one to the other side of the leading operand. The secondary operand for op_i includes everything to the relevant side between op_i and the next operator of *the same or lower* precedence than op_i. The only difference to the leading operand rule is that the secondary operand already ends when an operator of the *same* precedence appears.

According to this definition, the secondary operand of op_3 is $op_4 \;\; x_4$ in our example.

Finally, we put a pair of parentheses around the subsequence corresponding to the leading operand, the operator itself, and the secondary operand (if any).

Here is the table for our example again, enhanced with the subsequences of all four operators that are put in parentheses according to the rules just described.

expression	x_1	op_1	x_2	op_2	x_3	op_3	op_4	x_4
arity		2		2		2	1	
precedence		4		13		13	16	
associativity		r		l		l	r	
op_1	(x_1	op_1	x_2	op_2	x_3	op_3	op_4	x_4)
op_2			(x_2	op_2	x_3)			
op_3			(x_2	op_2	x_3	op_3	op_4	x_4)
op_4							(op_4	x_4)

Now we simply put together all parentheses that we have obtained, taking their multiplicities into account. In our example we get the expression

$$(\ x_1 \ \ op_1 \ \ ((\ x_2 \ \ op_2 \ \ x_3 \ ) \ \ op_3 \ \ (op_4 \ \ x_4 \ ))).$$

By some magic, this worked out, and we have a fully parenthesized expression (the outer pair of parentheses can be dropped again, of course). But note that we cannot expect such nice behavior in general. Consider the following example.

expression	x_1	op_1	x_2	op_2	x_3
arity		2		2	
precedence		13		13	
associativity		r		l	
op_1	(x_1	op_1	x_2	op_2	x_3)
op_2	(x_1	op_1	x_2	op_2	x_3)

The resulting parenthesized expression is

$$((\ x_1 \ \ op_1 \ \ x_2 \ \ op_2 \ \ x_3 \ )),$$

which does not specify the evaluation order. What comes to our rescue is that C++ only allows expressions for which the magic works out! The previous bad case is impossible, for example, since all binary operators of the same precedence also have the same associativity.

Unsigned arithmetic. We have discussed how `int` values are converted to `unsigned int` values, and vice versa. The main issue (what to do with non-representable values) also occurs during evaluation of arithmetic expressions involving only *one* of the types. The C++ standard contains one rule for this. For all unsigned integral types, the arithmetic operators work modulo 2^b, given b-bit representation. This means that the value of every arithmetic operation with operands of type `unsigned int` is well-defined. It does not necessarily give the mathematically correct value, but the unique value in the `unsigned int` range that is congruent to it modulo 2^b. For example, if a is a variable of type `unsigned int` with non-zero value, then `-a` has value $2^b - a$.

No such rule exists for the signed integral types, meaning that over- and underflow are dealt with at the discretion of the compiler.

Sequences of + and −. We have argued above that it is usually clear which operators occur in an expression, even though some of them share their token. But since the characters + and − are heavily overused in operator tokens, special rules are needed to resolve the meanings of sequences of +'s, or of −'s.

For example, only from arities, precedences and associativities it is not clear how to interpret the expressions `a+++b` or `---a`. The first expression could mean `(a++)+b`, but it could as well mean `a+(++b)` or `a+(+(+b)`. Similarly, the second expression could either mean `-(--a)`, `--(-a)` or `-(-(-a)`.

The C++ standard resolves this dilemma by defining that a sequence consisting only of +'s, or only of −'s, has to be grouped into pairs from left to right, with possibly one remaining + or − at the end. Thus, `a+++b` means `(a++)+b`, and `---a` means `--(-a)`. Note that for example the expression `a++b` *would* make sense when parenthesized as `a+(+b)`, but according to the rule just established, it is not a well-formed expression, since a unary operator ++ cannot have operands on both sides. The expression `---a` with its logical parentheses `--(-a)` is invalid for another reason: the operand of the pre-increment must be an lvalue, but the expression `-a` is an rvalue.

Other integral types. C++ contains a number of fundamental *signed* and *unsigned* integral types. The signed ones are `signed char`, `short int`, `int` and `long int`. The standard specifies that each of them is represented by at least as many bits as the previous one in the list. The number of bits used to represent `int` values depends on the platform. The corresponding sequence of unsigned types is `unsigned char`, `unsigned short int`, `unsigned int` and `unsigned long int`.

These types give compilers the freedom of offering integers with larger or smaller value ranges than `int` and `unsigned int`. Smaller value ranges are useful when memory consumption is a concern, and larger ones are attractive when over- and underflow occurs. The significance of these types (which are already present in the C programming language) has faded in C++. The reason is that we can quite easily implement our own tailor-made integral types in C++, if we need them. In C this is much more cumbersome. Consequently, many C++ compilers simply make `short int` and `long int` an alias for `int`, and the same holds for the corresponding unsigned types.

Order of effects and sequence points Increment and decrement operators as well as assignment operators construct expressions with an effect. Such operators have to be used with care for two reasons.

The obvious reason is that (as we have already learned in the end of Section 2.1.1) the evaluation order for the sub-expressions of a given expression is not specified in general. Consequently, value and effect may depend on the evaluation order. Consider the expression

```
++i + i
```

where we suppose that `i` is a variable of type `int`. If `i` is initially 5, say, then the value of the composite expression may in practice be 11 or 12. The result depends on whether

or not the effect of the left operand `++i` of the addition is processed before the right operand `i` is evaluated. The value of the expression `++i + i` is therefore unspecified by the C++ standard.

To explain the second (and much less obvious, but fortunately also much less relevant) reason, let us consider the following innocent looking expression that involves a variable `i` of type `int`.

```
i = ++i + 1
```

This expression has two effects: the increment of `i` and the assignment to `i`. Because the assignment can only happen after the operands have been evaluated, it seems that the order of the two effects is clear: the increment comes before the assignment, and the overall value and effect are well-defined.

However, this is not true, for reasons that have to do with our underlying computer model, the von Neumann architecture. From the computer's point of view, the evaluation of the sub-expression `++i` consists of the following steps.

1. Copy the value of `i` from the main memory into one of the CPU registers;
2. Add 1 to this value in the register;
3. Write the register content back to main memory, at the address of `i`;

Clearly, the first two steps are necessary to obtain the value of the expression `++i` and, hence, have to be processed before the assignment. But the third step does not necessarily have to be completed before the assignment. In order to allow the compiler to optimize the transfer of data between CPU registers and main memory (which is very much platform dependent), this order has not been specified. In fact, it is not unreasonable to assume that the traffic between registers and main memory is organized such that several items are transferred at once or quickly after another, using so-called *bursts*.

Suppose as before that `i` initially has value 5. If the assignment is performed after the register content is written back to main memory, `i = ++i + 1` sets `i` to 7. But if the assignment happens *before*, the later transfer of the register value 6 overrides the previous value of 7, and `i` is set to 6 instead.

The C++standard defines a *sequence point* to be a point during the evaluation sequence of an expression at which is guaranteed that all effects of previously evaluated (sub)expressions have been carried out. It was probably the existence of highly optimized C compilers that let the C++ standard refrain from declaring the assignment as a sequence point. In other words, when the assignment to `i` takes place in the evaluation `i = ++i + 1`, it is not specified whether the effect of the previously evaluated increment operator has been carried out or not. In contrast, the semicolon that terminates an expression statement is always a sequence point.

Therefore, we only have an issue with expressions that have more than one effect. Hence, if you prefer not to worry about effect order, ensure that each expression that you write generates at most one effect. Expressions with more than one effect can make

sense, though, and they are ok, as long as some sequence points separate the effects and put them into a well-defined order. This is summarized in the following rule.

Single Modification Rule: Between two sequence points, the evaluation of an expression may modify the value of an object of fundamental type at most once.

An expression like `i = ++i + 1` that violates this rule is considered semantically illegal and leads to undefined behavior.

If you perceive this example as artificial, here is a "more natural" violation of the single modification rule: if `nextvalue` is a variable of type `int`, it might seem that

```
nextvalue = 5 * nextvalue + 3
```

could more compactly be written as

```
(nextvalue *= 5) += 3
```

This will compile: `(nextvalue *= 5)` is an lvalue, so we can assign to it. Still, the latter expression is invalid since it modifies `nextvalue` twice.

At this point, an attentive reader should wonder how an expression that involves several output operators complies with the Single Modification Rule. Indeed, an expression like

```
std::cout << a << "^8 = " << b * b << ".\n"
```

has several effects all of which modify the lvalue `std::cout`. This works since the type of `std::cout` (which we will not discuss here) is *not* fundamental and, hence, the Single Modification Rule does not apply in this case.

2.2.11 Goals

Dispositional. At this point, you should ...

1) know the three Arithmetic Evaluation Rules;
2) understand the concepts of operator precedence and associativity;
3) know the arithmetic operators for the types `int` and `unsigned int`;
4) be aware that computations involving the types `int` and `unsigned int` may deliver incorrect results, due to possible over- and underflows.

Operational. In particular, you should be able to ...

(G1) parenthesize and evaluate a given arithmetic expression involving operands of types `unsigned int` and `int`, the binary arithmetic operators `+`,`-`, `*`, `/`, `%`, and the unary `-` (the paragraph on parenthesizing a general expression in the Details section enables you to do this for all arithmetic operators);

(G2) derive basic statements about arithmetic expressions;

(G3) convert a given decimal number into binary representation and vice versa;

(G4) derive the *two's complement* representation of a given number in b-bit representation, for some $b \in \mathbb{N}$;

(G5) write programs whose output is determined by a fixed number of arithmetic expressions involving literals and input variables of types `int` and `unsigned int`;

(G6) determine the value range of integral types on a given machine (using a program).

2.2.12 Exercises

Exercise 13 *Parenthesize the following expressions and then evaluate them step by step. This means that types and values of all intermediate results that are computed during the evaluation should be provided.* (G1)

a) `-2-4*3` *b)* `10%6*8%3` *c)* `6-3+4*5`
d) `5u+5*3u` *e)* `31/4/2` *f)* `-1-1u+1-(-1)`

Exercise 14 *Which of the following character sequences are not legal expressions, and why? (Here,* `a`, `b`, *and* `c` *are variables of type* `int`*.) For the ones that are legal, give the logical parentheses. (In order to avoid (misleading?) hints, we have removed the spaces that we usually include for the sake of better readability.)* (G1)

a) `c=a+7+--b` *b)* `c=-a=b` *c)* `c=a=-b`
d) `a-a/b*b` *e)* `b*=++a+b` *f)* `a-a*+-b`
g) `7+a=b*2` *h)* `a+3*--b+a++` *i)* `b+++--a`

These exercises require you to read the paragraph on logically parenthesizing a general expression in the Details section. Exercise *i)* also requires you to read the paragraph on sequences of + and - in the Details section.

Exercise 15 *For all legal expressions from Exercise 14, provide a step-by-step evaluation, supposing that initially* `a` *has value* 5, `b` *has value* 2, *and the value of* `c` *is undefined. Which of the expressions result in unspecified or undefined behavior?* (G1)

Exercise 16 *Prove that for all* $a \geq 0$ *and* $b, c > 0$, *the following equation holds.*

$$(a \operatorname{div} b) \operatorname{div} c = a \operatorname{div}(bc).$$

Does this imply that the two expressions `a/b/c` *and* `a/(b*c)` *are equivalent for all such values of the variables* `a`, `b`, *and* `c` *(which are assumed to be of type* `unsigned int`*)?* (G2)

Exercise 17 *Compute by hand binary representations of the following decimal numbers.* (G3)

a) 15 *b)* 172 *c)* 329 *d)* 1022

Exercise 18 *Compute by hand decimal representations of the following binary numbers.* (G3)

a) 110111 *b)* 1000001 *c)* 11101001 *d)* 101010101

Exercise 19 *By September 2009, the largest known Mersenne Prime is* $2^{43,112,609} - 1$, *see Section 1.1. What is the number of decimal digits that this number has? Explain how you got your answer!* (G3)

Hint: You may need the basic rules of logarithms and a pocket calculator.

Exercise 20 *Assuming a* 4*-bit representation, compute the binary two's complement representations of the following decimal numbers.* (G4)

a) 6 *b)* -4 *c)* -8 *d)* 9 *e)* -3

Exercise 21 *Suppose that someone drives from* A *to* B *at an average speed of 50 km/h. On the way back from* B *to* A*, there is a traffic jam, and the average speed is only 30 km/h. What is the average speed over the whole roundtrip?*

When confronted with this question, many people would answer "40 km/h," but this is wrong. Write a program that lets the user enter two average speeds in km/h ($A \to B$ *and* $B \to A$*) and computes from this the average speed over the whole roundtrip (*$A \to B \to A$*). Both inputs should be positive integers, and the output should be rounded down to the next smaller integer.*

Exercise 22 *Write a program* `celsius.cpp` *that converts temperatures from degrees Fahrenheit into degrees Celsius.*

The initial output that prompts the user to enter the temperature in degrees Fahrenheit should also contain lower *and* upper *bounds for the allowed inputs. These bounds should be chosen such that no over- and underflows can occur in the subsequent computations, given that the user respects the bounds. You may for this exercise assume that the integer division rounds towards zero for all operands: for example,* `-5 / 2` *then rounds the exact result* -2.5 *to* -2.

The program should output the correct *result in degrees Celsius as a mixed rational number of the form* $x\frac{y}{9}$ *(meaning* $x + y/9$*), where* $x, y \in \mathbb{Z}$ *and* $|y| \leq 8$. *For example,* $13\frac{4}{9}$ *could be output simply as* `13 4/9`. *We also allow for example the output* `-1 -1/9` *(meaning* $-1 - 1/9 = -10/9$*).* (G5)

Exercise 23 *Write a program* `threebin.cpp` *that reads a (decimal) number* $a \geq 0$ *from standard input and outputs the* last three *bits of* a*'s binary representation. Fill up with leading zeros in case the binary representation has less than three bits.* (G5)

Exercise 24 *Write a program* `vat.cpp` *that computes from a net amount (in integral units of CHF) a total amount, including value-added tax (VAT). The VAT rate should be provided to the computation in form of a constant (in one tenth of a percent, to allow VAT rates like 7.6%). The net amount is the input. The output*

(VAT and total amount) should be rounded (down) to two digits (Rp.) after the decimal point. (G5)

An example run of the program may look like this (assuming a VAT rate of 7.6%).

```
Net amount =?  59
VAT          = 4,48
Total amount = 63,48
```

2.2.13 Challenges

Exercise 25 *Josephus was a Jewish military leader in the Jewish-Roman war of 66-73. After the Romans had invaded his garrison town, the few soldiers (among them Josephus) that had survived the killings by the Romans decided to commit suicide. But somehow, Josephus and one of his comrades managed to surrender to the Roman forces without being killed (Josephus later became a Roman citizen and well-known historian).*

This historical event is the background for the Josephus Problem *that offers a (mythical) explanation about how Josephus was able to avoid suicide. Here is the problem.*

41 people (numbered $0, 1, ..., 40$*) are standing in a circle, and every* k*-th person is killed until no one survives. For* $k = 3$*, the killing order is therefore* $2, 5, 8, ..., 38, 0, 4, ...$*. Where in the circle does Josephus have to position himself in order to be the last survivor (who then obviously doesn't need to kill himself)?*

a) *Write a program that solves the Josephus problem; the program should receive as input the number* k *and output the number* $p(k) \in \{0, ..., 40\}$ *of the last survivor.*

b) *Let us assume that Josephus is not able to chose his position in the circle, but that he can in return choose the parameter* $k \in \{1, \ldots, 41\}$*. Is it possible for him to survive then, no matter where he initially stands?*

Hint: *This exercise has a theoretical part in which you need to come up with a formula for the survivor number that you can evaluate using the subset of the C++ language that you know so far.*

Exercise 26 *A triple* (a, b, c) *of positive integers is called a* Pythagorean triple *if* $a^2 + b^2 = c^2$*. For example,* $(3, 4, 5)$ *is a Pythagorean triple. Write a program* `pythagoras.cpp` *that allows you to list all Pythagorean triples for which* $a + b + c = 1000$*. We're not demanding that the program lists them directly, but the program should "prove" that your list is correct. How many such Pythagorean triples are there? (This is a slight variation of Problem 9 from the* Project Euler*, see* `http://projecteuler.net/`*.)*

2.3 Booleans

The truth always lies somewhere else.

Folklore

This section discusses the type `bool` *used to represent truth values or Booleans, for short. You will see a number of operations on Booleans and why only few of these operations suffice to express all the others. You will learn how to evaluate expressions involving the type* `bool`*, using* short-circuit evaluation.

What is the simplest C++ type you can think of? If we think of types in terms of their value ranges, then you will probably come up with a type whose value range is empty or consists of one possible value only. Arguably, values of such types are very easy to represent, even without spending any memory resources. However, although such types are useful in certain circumstances, you can't do a lot of interesting computations with them. After all, there is no operation on them other than the identity.

So, let us rephrase the above question: What is the simplest non-trivial C++ type you can think of? After the above discussion we certainly have one candidate: a type with a value range that consists of exactly two elements. At first sight, such a type may again appear very limited. Nevertheless, we will see below that it allows for many interesting operations. Actually, such a type is sufficient as a basis for all kinds of computations you can imagine. (Recall, for example, that integral numbers can be represented in binary format, that is, using the two values 0 and 1 only.)

2.3.1 Boolean functions

The name "Boolean" stems from the British mathematician George Boole (1815–1864) who pioneered in establishing connections between logic and symbolic algebra. By the term *Boolean function* we denote a function $f : \mathcal{B}^n \to \mathcal{B}$, where $\mathcal{B} := \{0, 1\}$ and $n \in \mathbb{N}$. (Read 0 as *false* and 1 as *true*.)

Clearly the number of different Boolean functions is finite for every fixed n; Exercise 27 asks you to show what exactly their number is. To give you a first hint: For $n = 1$ there are only four Boolean functions, the two constant functions $c_0 : x \mapsto 0$ and $c_1 : x \mapsto 1$, the identity $\mathrm{id} : x \mapsto x$ and the negation $\mathrm{NOT} : x \mapsto \overline{x}$, where $\overline{0} := 1$ and $\overline{1} := 0$.

In the following we restrict our focus to unary and binary Boolean functions, that is, functions from $\mathcal{B}$ or $\mathcal{B}^2$ to $\mathcal{B}$. Such functions are most conveniently described as a small table that lists the function values for all possible arguments. An example for a binary Boolean function is $\mathrm{AND} : (x, y) \mapsto x \wedge y$ shown in Figure 4(a). It is named AND because $x \wedge y = 1$ if and only if $x = 1$ *and* $y = 1$. You may guess why the

function $f : (x, y) \mapsto x \vee y$ defined in Figure 4(b) is called OR. In fact, there are two possible interpretations of the word "or": You can read it as "at least one of", but just as well it can mean "either ...or", that is, "exactly one of". The function that corresponds to the latter interpretation is shown in Figure 4(c). It is usually referred to as $\mathrm{XOR} : (x, y) \mapsto x \oplus y$ or *exclusive or*. Figure 4(e) depicts the table for the unary function NOT.

x	y	$x \wedge y$
0	0	0
0	1	0
1	0	0
1	1	1

(a) AND.

x	y	$x \vee y$
0	0	0
0	1	1
1	0	1
1	1	1

(b) OR.

x	y	$x \oplus y$
0	0	0
0	1	1
1	0	1
1	1	0

(c) XOR.

x	y	$x \uparrow y$
0	0	1
0	1	1
1	0	1
1	1	0

(d) NAND.

x	$\overline{x}$
0	1
1	0

(e) NOT.

Figure 4: *Examples for Boolean functions.*

Completeness. Figure 4 shows just a few examples. However, in a certain sense, it shows you everything about binary Boolean functions. Some of these functions are so fundamental that *every* binary Boolean function can be generated from them. For example, XOR can be generated from AND, OR and NOT:

$$\mathrm{XOR}(x, y) = \mathrm{AND}(\mathrm{OR}(x, y), \mathrm{NOT}(\mathrm{AND}(x, y))).$$

Informally, "either or" means "or" but not "and". Formulas like this are easily checked by going through all (four) possible combinations of arguments.

Similarly, the function $\mathrm{NAND} : (x, y) \mapsto x \uparrow y$ described in Figure 4(d) can be generated from NOT and AND (hence the name):

$$\mathrm{NAND}(x, y) = \mathrm{NOT}(\mathrm{AND}(x, y)).$$

Let us define what we mean by "generate".

Definition 2 *Consider a set F of boolean functions. A binary boolean function f is called* **generated** *by F if f can be expressed by a formula that only contains the variables x and y, the constants 0 and 1, and the functions from F.*

For a set $\mathcal{F}$ of binary functions, a set F of binary functions is said to be **complete** *if and only if every function $f \in \mathcal{F}$ can be generated by F.*

We are now prepared for a completeness proof.

Theorem 1 *The set of functions $\{\mathrm{AND}, \mathrm{OR}, \mathrm{NOT}\}$ is complete for the set of binary Boolean functions.*

Proof. Every binary Boolean function f is completely described by its *characteristic vector* $(f(0,0), f(0,1), f(1,0), f(1,1))$. For example, AND has characteristic vector $(0,0,0,1)$, or 0001 for short. Let $f_{b_1b_2b_3b_4}$ denote the Boolean function with characteristic vector $b_1b_2b_3b_4$. For example, $\text{AND} = f_{0001}$.

In the first step of the proof, we show that all those functions can be generated whose characteristic vector contains a single 1. Indeed,

$$\begin{aligned}
f_{0001}(x,y) &= \text{AND}(x,y),\\
f_{0010}(x,y) &= \text{AND}(x,\text{NOT}(y)),\\
f_{0100}(x,y) &= \text{AND}(y,\text{NOT}(x)),\\
f_{1000}(x,y) &= \text{NOT}(\text{OR}(x,y)).
\end{aligned}$$

To check the formula for f_{0010}, for example, we can create a table for the function $\text{AND}(x,\text{NOT}(y))$ as in Figure 4 and convince ourselves that the resulting characteristic vector is 0010.

In the second step, we show that every function whose characteristic vector is nonzero can be generated. This is done by combining the already generated "single-1" functions through OR, which simply adds up their 1's. For example,

$$\begin{aligned}
f_{1100}(x,y) &= \text{OR}(f_{1000}(x,y), f_{0100}(x,y)),\\
f_{0111}(x,y) &= \text{OR}(\text{OR}(f_{0100}(x,y), f_{0010}(x,y)), f_{0001}(x,y)).
\end{aligned}$$

We abstain from working this argument out formally, since we believe that you get its idea. Finally, we generate f_{0000} as

$$f_{0000}(x,y) = 0.$$

□

Exercise 30 asks you to show that the sets $\{\text{AND}, \text{NOT}\}$, $\{\text{OR}, \text{NOT}\}$, and even the set that consists of the single function $\{\text{NAND}\}$ are complete for the set of binary Boolean functions.

2.3.2 The type bool

In C++, Booleans are represented by the fundamental type `bool`. Its value range consists of the two elements *true* and *false* that are associated with the literals `true` and `false`, respectively. For example,

```
bool b = true;
```

defines a variable `b` of type `bool` and initializes it to *true*.

Formally, the type `bool` is an integral type, defined to be less general than `int` (which in turn is less general than `unsigned int`, see Section 2.2.7). An expression of type `bool`, or of a type whose values can be converted to `bool`, is called a *predicate*.

Logical operators. The complete set of binary Boolean functions is available via the *logical operators* `&&` (AND), `||` (OR), and `!` (NOT). Compared to the notation used in Section 2.3.1 we simply identify 1 with *true* and 0 with *false*. Both `&&` and `||` are binary operators, while `!` is unary. All operands are rvalues of type `bool`, and all logical operators also return rvalues of type `bool`. Like in logics, `&&` binds more strongly than `||`, and `!` binds more strongly than `&&` (recall that an operator binds more strongly than another if its has higher precedence).

Relational operators. There is also a number of operators on arithmetic types whose result is of type `bool`. For each arithmetic type there exist the six *relational operators* `<`, `>`, `<=`, `>=`, `==`, and `!=`. These are binary operators whose two rvalue operands are of some arithmetic type and whose result is an rvalue of type `bool`. The operators `<=` and `>=` correspond to the mathematical relations $\leq$ and $\geq$, respectively. The operator `==` tests for equality and `!=` tests for inequality.

Since `bool` is an integral type, the relational operators may also have operands of type `bool`. The respective comparisons are done according to the convention *false*<*true*.

> **Watch out!** A frequent beginner's mistake is to use the assignment operator `=` where the equality operator `==` is meant.

As a general rule, arithmetic operators bind more strongly than relational ones, and these in turn bind more strongly than the logical operators.

> **Boolean Evaluation Rule:** Binary arithmetic operators have higher precedence than relational operators, and these have higher precedence than binary logical operators.

All binary relational and logical operators are (as the binary arithmetic operators) left-associative, see also Table 2. For example, the expression

```
7 + x < y && y != 3 * z
```

is logically parenthesized as

```
((7 + x) < y) && (y != (3 * z)).
```

Be careful with mathematical shortcut notation such as $a = b = c$. As a C++ expression,

```
a == b == c
```

is not equivalent to

```
a == b && b == c.
```

By left associativity of `==`, the expression `a == b == c` is logically parenthesized as `(a == b) == c`. If all of `a`, `b`, and `c` are variables of type `int` with value 0, the evaluation yields

$$(0 \texttt{ == } 0) \texttt{ == } 0 \longrightarrow \mathit{true} \texttt{ == } 0 \longrightarrow 1 \texttt{ == } 0 \longrightarrow \mathit{false}\,,$$

just the opposite of what you usually mean by $a = b = c$.

De Morgan's laws. The well-known formulae of how to express AND in terms of OR and vice versa with the help of NOT, are named after the British mathematician Augustus De Morgan (1806–1871). He was a pioneer in symbolic algebra and logics. Also the rigorous formulation of "mathematical induction" as we know and use it today goes back to him. The de-Morgan-formulae state that (in C++-language)

```
!(x && y) == (!x || !y)
```

and

```
!(x || y) == (!x && !y) .
```

These formulae can often be used to transform a *Boolean expression* (an expression of type `bool`) into a "simpler" equivalent form. For example,

```
!(x < y || x + 1 > z) && !(y <= 5 * z || !(y > 7 * z))
```

can equivalently be written as

```
x >= y && x + 1 <= z && y > 5 * z && y > 7 * z
```

which is clearly preferable in terms of readability.

For more details about precedences and associativities of the logical and relational operators, see Table 2. You may find this information helpful in order to solve Exercise 32.

Description	Operator	Arity	Prec.	Assoc.
logical not	!	1	16	right
less	<	2	11	left
greater	>	2	11	left
less or equal	<=	2	11	left
greater or equal	>=	2	11	left
equality	==	2	10	left
inequality	!=	2	10	left
logical and	&&	2	6	left
logical or	||	2	5	left

Table 2: *Precedences and associativities of logical and relational operators. All operands and return values are rvalues.*

Conversion and promotion. It is possible that the two operands of a relational operator have different type. This case is treated in the same way as for the arithmetic operators. The composite expression is evaluated on the more general type, to which the operand of the less general type is implicitly converted. In particular, `bool` operands are converted to the respective integral type of the other operand. Here, the value *false* is converted to 0, and *true* to 1. If the integral type is `int`, this conversion is defined to be a *promotion*. A promotion is a special conversion for which the C++ standard guarantees that no information gets lost.

The conversion goes into the other direction for logical operators. In mixed expressions, the integral operands of logical operators are converted to `bool` in such a way that 0 is converted to *false* and any other value is converted to *true*.

These conversions also take place in initializations and assignments, as in the following examples.

```
bool b = 5; // b is initialized to true
int i = b;  // i is initialized to 1
```

2.3.3 Short circuit evaluation

The evaluation of expressions involving logical and relational operators proceeds according to the general rules, as discussed in Sections 2.2.1 and 2.2.3. However, there is one important difference regarding the order in which the operands of an operator are evaluated. While in general this order is unspecified, the binary logical operators `&&` and `||` always guarantee that their left operand is evaluated first. Moreover, if the value of the composite expression is already determined by the value of the left operand then the right operand is *not evaluated* at all. This evaluation scheme is known as *short circuit evaluation*.

How can it happen that the final value is already determined by the left operand only? Suppose that in an `&&` operator the left operand evaluates to *false*; then no matter what the right operand gives, the result will always be *false*. Hence, there is no need to evaluate the right operand at all. The analogous situation occurs if in an `||` operator the left operand evaluates to *true*.

At first sight it looks as if short circuit evaluation is merely a matter of efficiency. But there is another benefit. It occurs when dealing with expressions that are defined for certain parameters only. Consider for example the division operation that is defined for a nonzero divisor only. Due to short circuit evaluation, we can write

```
x != 0 && z / x > y
```

and be sure that this expression is always valid. If the right operand was evaluated for x having value 0, then the result would be undefined.

2.3.4 Details

Naming. The XOR function is also frequently called *antivalence* and denoted by $\nleftrightarrow$. The NAND function is also known as *alternate denial* or *Sheffer stroke*. The latter name is after the American mathematician Henry M. Sheffer (1883–1964) who proved that all other logical operations can be expressed in terms of NAND.

Bitwise operators. We have seen in Section 2.2.8 that integers can be represented in binary format, that is, as a sequence of bits each of which is either 0 or 1. Boolean

functions can naturally be extended to integral types by applying them bitwise to the binary representations.

Definition 3 *Consider a nonnegative integer* b *and two integers* $x = \sum_{i=0}^{b} a_i 2^i$ *and* $y = \sum_{i=0}^{b} b_i 2^i$*, for which* $a_i, b_i \in \{0, 1\}$ *for all* $0 \leq i \leq b$*.*

For a unary Boolean function $f : \{0, 1\} \to \{0, 1\}$ *the* **bitwise operator** φ_f *corresponding to* f *is defined as* $\varphi_f(x) = \sum_{i=0}^{b} f(a_i) 2^i$*.*

For a binary Boolean function $g : \{0, 1\}^2 \to \{0, 1\}$ *the* **bitwise operator** φ_g *corresponding to* g *is defined as* $\varphi_g(x, y) = \sum_{i=0}^{b} g(a_i, b_i) 2^i$*.*

For illustration, suppose that we have an unsigned integral type with a 4-bit representation. That is, 0000 represents 0, 0001 represents 1, and so on, up to 1111 which represents 15.

Then you can check that $\varphi_{\text{OR}}(4, 13) = 13$, $\varphi_{\text{NAND}}(13, 9) = 6$, and $\varphi_{\text{NOT}}(2) = 13$.

Several bitwise operators are defined for the integral types in C++. There is a bitwise AND `&`, a bitwise OR `|`, and a bitwise XOR `^`, as well as a bitwise NOT `~` that is usually referred to as *complement*. As the arithmetic operators, the binary bitwise operators (except for `~`) have a corresponding assignment operator. The precedences and associativity of these operators are listed in Table 3.

Description	Operator	Arity	Prec.	Assoc.
bitwise complement	`~`	1	16	right
bitwise and	`&`	2	9	left
bitwise xor	`^`	2	8	left
bitwise or	`|`	2	7	left
and assignment	`&=`	2	4	right
xor assignment	`^=`	2	4	right
or assignment	`|=`	2	4	right

Table 3: *Precedence and associativity of bitwise operators.*

Note that the functionality of these operators is implementation defined, since the bitwise representations of integral type values are not specified by the C++ standard. We have only discussed the most frequent (and most likely) such representations in Section 2.2.8. You should therefore *only* use these operators when you *know* the representation. Even then, expressions involving the bitwise operators are implementation defined.

This is most obvious with the bitwise complement: even if we assume the standard binary representation of Section 2.2.8, the value of the expression `~0` depends on the number b of bits in the representation. This value therefore changes when you switch from a 32-bit machine to a 64-bit machine.

2.3.5 Goals

Dispositional. At this point, you should ...

1) know the basic terminology around Boolean functions and understand the concept of completeness;
2) know the type `bool`, its value range, and the conversions and operations involving `bool`;
3) understand the evaluation of expressions involving logical and relational operators, in particular the Boolean Evaluation Rule and the concept of short circuit evaluation.

Operational. In particular, you should be able to ...

(G1) prove or disprove basic statements about Boolean functions;

(G2) prove whether or not a given set of binary Boolean functions is complete;

(G3) evaluate a given expression involving arithmetic, logical, and relational operators;

(G4) read and understand a given simple program (see below), involving objects of arithmetic type (including `bool`) and arithmetic, logical, and relational operators.

The term *simple program* refers to a program that consists of a main function which in turn consists of a sequence of declaration and expression statements. Naturally, only the fundamental types and operations discussed in the preceding sections are used.

2.3.6 Exercises

Exercise 27 *For* $n \in \mathbb{N}$, *how many different Boolean functions* $f : \mathcal{B}^n \to \mathcal{B}$ *exist?* (G1)

Exercise 28 *Prove or disprove that for all* $x, y, z \in \mathcal{B}$ (G1)

a) $(x \oplus y) \oplus z = x \oplus (y \oplus z)$. (i.e., XOR is associative)

b) $(x \wedge y) \vee z = (x \vee z) \wedge (y \vee z)$. (i.e., (AND, OR) is distributive)

c) $(x \vee y) \wedge z = (x \wedge z) \vee (y \wedge z)$. (i.e., (OR, AND) is distributive)

d) $(x \uparrow y) \uparrow z = x \uparrow (y \uparrow z)$. (i.e., NAND is associative)

Exercise 29 *For* $x_1, \ldots, x_n$, $n \in \mathbb{N}$, *give a verbal description of* $x_1 \oplus x_2 \oplus \ldots \oplus x_n$ *in terms of the* x_i, $1 \leq i \leq n$. (G1)

Exercise 30 *Show that the following sets of functions are complete for the set of binary Boolean functions.* (G2)

a) $\{\text{AND}, \text{NOT}\}$

b) $\{\text{OR}, \text{NOT}\}$

c) $\{\text{NAND}\}$

d) $\{\text{NOR}\}$, where $\text{NOR} := \text{NOT} \circ \text{OR}$.

e) $\{\text{XOR}, \text{AND}\}$

You may use the fact that set $\{\text{AND}, \text{OR}, \text{NOT}\}$ is complete for the set of binary Boolean functions.

Exercise 31 *Suppose* `a`, `b`, *and* `c` *are all variables of type* `int`. *Find values for* `a`, `b`, *and* `c` *for which the expressions* `a < b < c` *and* `a < b && b < c` *yield different results.* (G3)

Exercise 32 *Parenthesize the following expressions according to operator precedences and associativities.* (G3)

a) `x != 3 < 2 || y && -3 <= 4 - 2 * 3`

b) `z > 1 && ! x != 2 - 2 == 1 && y`

c) `3 * z > z || 1 / x != 0 && 3 + 4 >= 7`

Exercise 33 *Evaluate the expressions given in Exercise 32 step-by-step, assuming that* `x`, `y`, *and* `z` *are all of type* `int` *with* `x==0`, `y==1`, *and* `z==2`. (G3)

Exercise 34 *What can you say about the output of the following program? Characterize it depending on the input and explain your reasoning.* (G4)

```
1   #include <iostream>
2   int main()
3   {
4     int a;
5     std::cin >> a;
6     std::cout << (a++ < 3) << ".\n";
7     const bool b = a * 3 > a + 4 && !(a >= 5);
8     std::cout << (!b || ++a > 4) << ".\n";
9     return 0;
10  }
```

Exercise 35 *Find the logical parentheses in lines 9 and 12 of the following program. What can you say about the output of the following program? Characterize it depending on the input and explain your reasoning.* (G4)

```
1  #include <iostream>
2
3  int main ()
4  {
5    unsigned int a;
6    std::cin >> a;
7
8    unsigned int b = a;
9    b /= 2 + b / 2;
10   std::cout << b << "\n";
11
12   const bool c = a < 1 || b != 0 && 2 * a / (a - 1) > 2;
13   std::cout << c << "\n";
14
15   return 0;
16 }
```

2.3.7 Challenges

Exercise 36 *The* Reverse Polish Notation *(RPN) is a format of writing expressions without any parentheses. RPN became popular in the late nineteensixties when the company Hewlett-Packard started to use it as input format for expressions on their desktop and handheld calculators.*

In RPN, we first write the operands, and then the operator (that's what the Reverse *stands for). For example, the expression*

$$\mathrm{AND}(\mathrm{OR}(0, \mathrm{NOT}(\mathrm{AND}(0, 1))), 1)$$

can be written like this in RPN:

$$0\ 0\ 1\ \mathrm{AND}\ \mathrm{NOT}\ \mathrm{OR}\ 1\ \mathrm{AND}.$$

The latter sequence of operands and operators defines a specific evaluation sequence of the expression, see Section 2.2.3. To evaluate an expression in RPN, we go through the sequence from left to right; whenever we find an operand, we don't do anything, but when we find an operator (of arity n*), we evaluate it for the* n *operands directly to the left of it and replace the involved* $n+1$ *sequence elements by the result of the evaluation. Then we go to the next sequence element. In case of our example above, this proceeds as follows (currently processed operator in bold):*

$$0\ \underbrace{0\ 1\ \textbf{AND}}_{0}\ \text{NOT OR 1 AND}$$

$$0\ \underbrace{0\ \textbf{NOT}}_{1}\ \text{OR 1 AND}$$

$$\underbrace{0\ 1\ \textbf{OR}}_{1}\ \text{1 AND}$$

$$\underbrace{1\ 1\ \textbf{AND}}_{1}$$

$$1$$

To see that this is indeed a way of evaluating the original expression

$$\text{AND}(\text{OR}(0, \text{NOT}(\text{AND}(0, 1))), 1),$$

you can for example make a bottom-up drawing of an expression tree (Section 2.2.2) that corresponds to the evaluation sequence in RPN. You will find that this tree is also valid for the original expression.

Here comes the actual exercise. Write programs `and.cpp`, `or.cpp`, *and* `not.cpp` *that receive as input a sequence* s *of boolean values in* $\{0, 1\}$ *("all operands to the left of the operator"). The output should be the sequence* s' *that we get by replacing the last* n *operands in* s *with the result of evaluating the respective operator for them. In case of* `and.cpp` *and* `or.cpp`, *we use* $n = 2$, *and for* `not.cpp` $n = 1$. *For example, on input* $(1, 1, 0)$, *program* `and` *should output the sequence* $(1, 0)$, *while* `not` *should yields* $(1, 1, 1)$.

In addition, write programs `zero.cpp` *and* `one.cpp` *that output the sequence* s' *obtained by appending a* 0 *or* 1 *to the input* s. *Finally, write a program* `eval.cpp` *(with no input) that outputs the empty sequence.*

The goal of all this is to evaluate boolean functions in RPN by simply calling the corresponding sequence of programs (preceded by a call to `eval`*), where the output of one program is used as input for the next one in the sequence. In Unix and Linux this can elegantly be done via a* pipe. *For example, to evaluate the example expression from above in RPN, we simply type the command*

```
./eval |./zero |./zero |./one |./and |./not |./or |./one |./and
```

This calls all listed programs in turn, where a separating pipe symbol `|` *has the effect that the output of the program to the left of it is used as the input for ("is piped into") the program to the right of it.*

Consequently, the whole aforementioned command should simply write `1` *to standard output, the result of the evaluation. Also test your programs with some other RPN sequences, in particular the "obvious" ones of the form*

```
./eval |./zero |./one |./or
```

(this one should output `1`*) to make sure that they work as expected.*

Hint: *It is not necessary that your programs accept sequences* `s` *of arbitrary length as input. A maximum length of 32, for example, is sufficient for all practical purposes; in this case, the sequence can be encoded by one value of type* `unsigned int`*.*

2.4 Control statements

We are what we repeatedly do. Excellence, then, is not an act but a habit.

Will Durant in a summary of Aristotle's ideas,
The Story of Philosophy: The Lives and Opinions
of the World's Greatest Philosophers (1926)

This section introduces four concepts to control the execution of a program: selection, iteration, blocks, *and* jumps. *These concepts enable us to deviate from the default linear control flow which executes statement by statement from top to bottom. You will learn how these concepts are implemented in C++, and how to apply them to create interesting programs.*

The programs that we have seen so far are all pretty simple. They consist of a sequence of statements that are executed one by one from the first to the last. Such a program is said to have a *linear control flow*. This type of control flow is quite restrictive, as each statement in the source code is executed at most once during the execution of the program. Suppose you want to implement an algorithm that performs 10,000 steps for some input. Then you would have to write a program with at least 10,000 lines of code. Obviously this is undesirable. Therefore, in order to implement non-trivial algorithms, more powerful mechanisms to control the flow of a program are needed.

2.4.1 Selection: if– and if-else statements

One particularly simple way to deviate from linear control flow is to select whether or not a particular statement is executed. In C++ this can be done via an `if` *statement*. The syntax is

```
if ( condition )
  statement
```

where *condition* is an expression or variable declaration of a type whose values can be converted to `bool`, and *statement* — as the name suggests — is a statement. In case you are missing a semicolon after *statement*: recall that this semicolon is part of the statement. The semantics is the following: *condition* is evaluated; if and only if its value is *true*, *statement* is executed afterwards. In other words, an `if` statement splits the control flow into two branches. The value of *condition* selects which of these branches is executed. For example, the following lines of code

```
int a;
std::cin >> a;
if (a % 2 == 0) std::cout << "even";
```

read a number from standard input into the variable a and write "even" to standard output if and only if a is even.

Optionally, an `if` statement can be complemented by an `else`-*branch*. The syntax is

```
if ( condition )
  statement1
else
  statement2
```

and the semantics is as follows: *condition* is evaluated; if its value is *true*, *statement1* is executed afterwards; otherwise, *statement2* is executed afterwards. For example, the following lines of code

```
int a;
std::cin >> a;
if (a % 2 == 0)
  std::cout << "even";
else
  std::cout << "odd";
```

read a number from standard input into the variable a. Then if a is even, "even" is written to standard output; otherwise, "odd" is written to standard output.

When formatting an `if` statement, it is common to insert a line break before *statement1*, before `else`, and before *statement2*. Moreover, *statement1* and *statement2* are indented and `else` is aligned with `if`, as shown in the example above. If the whole statement fits on a single line then it can also be typeset as a single line.

Collectively, `if`- and `if-else` statements are known as *selection statements*.

2.4.2 Iteration: for statements

A much more powerful way of manipulating the control flow is provided by *iteration statements*. Iteration allows to execute a statement many times, possibly with different parameters each time. Iteration statements are also called *loops*, as they "loop through" a statement (potentially) several times. Selection and iteration statements are collectively referred to as *control statements*.

Consider the problem of computing the sum $S_n = \sum_{i=1}^{n} i$ of the first n natural numbers, for a given $n \in \mathbb{N}$. Program 2.7 reads in a variable n from standard input, defines another variable s to contain the result, computes the result and finally outputs it. In order to understand why the program `sum_n.cpp` indeed behaves as claimed, we have to explain the different parts of a `for` *statement*.

```
// Program: sum_n.cpp
// Compute the sum of the first n natural numbers.

#include <iostream>

int main()
{
  // input
  std::cout << "Compute the sum 1+...+n for n =? ";
  unsigned int n;
  std::cin >> n;

  // computation of sum_{i=1}^n i
  unsigned int s = 0;
  for (unsigned int i = 1; i <= n; ++i) s += i;

  // output
  std::cout << "1+...+" << n << " = " << s << ".\n";
  return 0;
}
```

Program 2.7: *../progs/lecture/sum_n.cpp*

for statement. The `for` statement is a very compact form of an iteration statement, as it combines three statements or expressions into one. In most cases, the `for` statement serves as a "counting loop" as in Program 2.7. Its syntax is defined by

`for` (*init-statement condition*; *expression*)
  statement

where *init-statement* is an expression statement, a declaration statement, or the null statement, see Section 2.1.15. In all of these cases, *init-statement* ends with a semicolon such that there are always two semicolons in between the parentheses after a `for`. Usually *init-statement* defines and initializes a variable that is used to control and eventually end the iteration statement's execution. In `sum_n.cpp`, *init-statement* is a declaration statement that defines the variable `i`.

As in an `if` statement, *condition* is an expression or variable declaration whose type can be converted to `bool`. It defines how long the iteration goes on, namely as long as *condition* returns *true*. It is allowed that *condition* is empty in which case its value is interpreted as *true*. As the name suggests, *expression* is an arbitrary expression that may also be empty (in which case it has no effect). Finally, *statement* is an arbitrary statement. It is referred to as the *body* of the `for` statement.

Typically, *expression* has the effect of changing a value that appears in *condition*. Such an effect is said to "make progress towards termination". The goal is that *condition* is *false* after *expression* has been evaluated a finite number of times. In that sense, every evaluation of *expression* makes a step towards the end of the `for` statement. In `sum_n.cpp`, *expression* increments the variable `i` which is bounded from above in *condition*. In cases like this where the value of a single variable is accessed and changed by *condition* and *expression*, we call this variable the *control variable* of the `for` statement.

We are now ready to precisely define the semantics of a `for` statement. First, *init-statement* is executed once. Thereafter *condition* is evaluated. If it returns *true*, an *iteration* of the loop starts. If *condition* returns *false*, the `for` statement *terminates*, that is, its processing ends immediately.

Each single iteration of a `for` statement consists of first executing *statement* and then evaluating *expression*. After each iteration, *condition* is evaluated again. If it returns *true*, another iteration follows. If *condition* returns *false*, the `for` statement terminates. The execution order is therefore *init-statement*, *condition*, *statement*, *expression*, *condition*, *statement*, *expression*, ... until *condition* returns *false*.

Let's see this in action: Consider the `for` statement

```
for (unsigned int i = 1; i <= n; ++i) s += i;
```

in `sum_n.cpp` and suppose `n == 2`. First, the variable `i` is defined and initialized to 1. Then it is tested whether `i <= n`. As `1 <= 2` is *true*, the first iteration starts. The statement `s += i` is executed, setting `s` to 1, and thereafter `i` is incremented by one such that `i == 2`. One iteration is now complete. As a next step, the condition `i <= n` is evaluated again. As `2 <= 2` is *true*, another iteration follows. First `s += i` is executed, setting `s` to 3. Thereafter, `i` is incremented by one such that `i == 3`. The second iteration is now complete. The subsequent evaluation of `i <= n` entails `3 <= 2` which is *false*. Thus, no further iteration takes place and the processing of the `for` statement ends. The value of `s` is now 3, the sum of the first 2 natural numbers.

Infinite loops. It is easily possible to create loops that do not terminate. For example, recall that both *condition* and *expression* may be empty. Moreover, both *init-statement* and *statement* can be the null statement. In this case we get the `for` statement

```
for (;;);
```

As the empty *condition* has value *true*, executing this statement runs through iteration after iteration without actually doing anything. Therefore, `for (;;)` may be read as "forever". In general, a statement which does not terminate is called an *infinite loop*.

Clearly, infinite loops are extremely undesirable and programmers try hard to avoid them. Nevertheless, sometimes such loops occur even in real life software. If you regularly use a computer, you have probably experienced this kind of phenomenon: a program "hangs".

You may ask: Why doesn't the compiler simply detect infinite loops and warns me about them just as it complains about syntax errors? Indeed, this would be a great thing to have and it would solve many problems in software development. The problem is that infinite loops are not always as easy to spot as in the above example. Loops can be pretty complicated, and possibly they loop infinitely when executed in certain program states only.

In fact, the situation is hopeless: It can be shown that the problem of detecting infinite loops (commonly referred to as the *halting problem*) cannot be solved by a computer, as we have and understand it today (see the Details). Therefore, some care is needed when designing loops. We have to check "by hand" that the iteration statement terminates for all possible program states that can occur.

Gauss. You may know or have realized that our program `sum_n.cpp` is actually a bad example. It is bad in the sense that it does not convincingly demonstrate the power of control statements.

In his primary school days, the German mathematician Carl Friedrich Gauss (1777–1855) was told to sum up the numbers $1, 2, 3, \ldots, 100$. The teacher had planned to keep his students busy for a while, but Gauss came up with the correct result 5050 very quickly. He had imagined writing down the numbers in increasing order, and one line below once again in decreasing order. Clearly, the two numbers in each column sum up to 101; hence, the overall sum is $100 \cdot 101 = 10100$, half of which is the number that was asked for.

1	2	3	...	98	99	100
100	99	98	...	3	2	1
101	101	101	...	101	101	101

In this way, Gauss discovered the formula

$$\sum_{i=1}^{n} i = n(n+1)/2,$$

for every $n \in \mathbb{N}$. The `for` statement in `sum_n.cpp` can therefore be replaced by the much more elegant and efficient statement

```
s = n * (n + 1) / 2;
```

Note that in this statement, the integer division coincides with the real division, since for all n, the product $n(n+1)$ is even. We next get to a real application of selection and iteration statements.

Prime numbers. In the introductory Section 1.1, we have talked a lot about prime numbers. How would a program look like that tests whether or not a given number is prime? A number $n \in \mathbb{N}, n \geq 2$ is prime if and only if it is not divisible by any number $d \in \{2, \ldots, n-1\}$. The strategy for our program is therefore clear: Write a loop that runs

through all these numbers, and test each of them for being a divisor of n. If a divisor is found, we can stop and output a factorization of n into two numbers, proving that n is not prime. Otherwise, we output that n is prime. Program 2.8 implements this strategy in C++, using one `for` statement, and one `if` statement. Remarkably, the `for` statement has an empty body, since we have put the divisibility test into the *condition*. The important observation is that the *condition* `n % d != 0` definitely returns *false* for `d == n`, so that the loop is guaranteed to terminate; if (and only if) *condition* returns *false* earlier, we have found a divisor of n in the range $\{2, \ldots, n-1\}$.

```
1  // Program: prime.cpp
2  // Test if a given natural number is prime.
3
4  #include <iostream>
5
6  int main ()
7  {
8    // Input
9    unsigned int n;
10   std::cout << "Test if n>1 is prime for n =? ";
11   std::cin >> n;
12
13   // Computation: test possible divisors d
14   unsigned int d;
15   for (d = 2; n % d != 0; ++d);
16
17   // Output
18   if (d < n)
19     // d is a divisor of n in {2,...,n-1}
20     std::cout << n << " = " << d << " * " << n / d << ".\n";
21   else
22     // no proper divisor found
23     std::cout << n << " is prime.\n";
24
25   return 0;
26 }
```

Program 2.8: *../progs/lecture/prime.cpp*

We would like to point out that the above arguments are only valid for $n \geq 2$; Exercise 45 asks you to consider the cases $n = 0, 1$.

2.4.3 Blocks and scope

In C++ it is possible to group a sequence of one or more statements into one single statement that is then called a *compound statement*, or simply a *block*. This mechanism

does not manipulate the control flow directly. Blocks allow to structure a program by grouping statements that logically belong together. In particular, they are a tool to design powerful and at the same time readable control statements.

Syntactically, a block is simply a sequence of zero or more statements that are enclosed in curly braces.

{ *statement1 statement2 ... statementN* }

Each of the statements may in particular be a block, so it is possible to have nested blocks. The simplest block is the empty block {}.

You have already seen blocks. Each program contains a special block, the so-called *function body* of the `main` function. This block encloses the sequence of statements that is executed when the `main` function is called by the operating system.

Using blocks, one can create selection and iteration statements whose body contains a sequence of two or more statements. For example, suppose that for testing purposes we would like to write out all partial sums during the computation in `sum_n.cpp`:

```
for (unsigned int i = 1; i <= n; ++i) {
  s += i;
  std::cerr << i << "-th partial sum is " << s << "\n";
}
```

Here two statements are executed in each iteration of the loop. First, the next summand is added to `s`, then the current value of `s` is written to standard error output.

Blocks should in general be formatted as shown above. That is, a line break appears after the opening and before the closing brace, and all lines in between are indented one level. Only if the block consists of just one single statement and it all fits on one line, the block can be formatted as one single line.

The kind of test output we have created in the previous example is called *debugging output*. A *bug* is a commonly used term to denote a programming error, hence "debugging" is the process of finding and eliminating such errors. It is good practice to write debugging output to standard error output since it can then more easily be separated from the "real" program output that usually goes to standard output.

Visibility. Blocks do not only structure a program visually but they also provide a logical boundary around declarations (of variables, for example). Every declaration that appears inside a block is called *local* to that block. A local declaration extends only until the end of the block in which it appears. A name that is introduced by a local declaration is not "visible" outside of the block where it is declared. For example, in

```
1 int main()
2 {
3   {
4     const int i = 2;
5   }
```

```
6    std::cout << i; // error, undeclared identifier
7    return 0;
8  }
```

the variable i declared inside the block in line 3–5 is not visible in the output statement in line 6. Thus, if you confront the compiler with this code, it issues an error message.

Control statements and blocks. Control statements act like blocks themselves. Therefore every declaration appearing in a control statement is local to that control statement. In particular, this applies to a variable defined in the *init-statement* of a for statement. For example, in

```
1  int main()
2  {
3    for (unsigned int i = 0; i < 10; ++i) s += i;
4    std::cout << i; // error, undeclared identifier
5    return 0;
6  }
```

the expression i in line 4 does *not* refer to the variable i defined in line 3.

Declarative region. After having seen these first examples, we will now introduce the precise terminology that allows us to deduce which names can be used where in the program. Each declaration has an associated *declarative region*. This region is the part of the program in which the declaration appears. Such a region can be a block, a function definition, or a control statement. In all these cases the declaration is said to have *local scope*. A declaration can also have *namespace scope*, if it appears inside a namespace, see Section 2.1.3. Finally, a declaration that is outside of any particular other structure has *global scope*.

Scope. A name introduced by a declaration D is *valid* or *visible* in a part of its declaration's declarative region, called the *scope* of the declaration. Within the scope of D, the name introduced by D may be used and actually refers to the declaration D. In most cases, the scope of a declaration is equal to its *potential scope*.

The *potential scope* of a declaration starts at the point where the declaration appears. For the name to be declared this is called its *point of declaration*. The potential scope extends until the end of the declarative region.

To get the scope of a declaration, we start from its potential scope but we possibly have to remove some parts of it. This happens when the potential scope contains one or more declarations of the *same* name. As an example, consider Program 2.9.

```
1  #include <iostream>
2
3  int main()
```

```
 4 {
 5   const int i = 2;
 6   for (int i = 0; i < 5; ++i)
 7     std::cout << i;    // outputs 0, 1, 2, 3, 4
 8   std::cout << i;      // outputs 2
 9   return 0;
10 }
```

Program 2.9: *../progs/lecture/scope.cpp*

The i in line 7 refers to the declaration from line 6, whereas the i in line 8 refers to the declaration from line 5. Therefore, the program outputs first 0, 1, 2, 3, 4, and then 2. In some sense, the declaration in line 6 temporarily hides the previous declaration of i from line 5. This phenomenon is called *name hiding*. But when the declarative region of the second declaration ends in line 7, the second declaration "becomes invisible" (we say: "it runs out of scope") and the first declaration takes over again. In particular, since the name i in line 8 refers to the constant defined in line 5, we get the output 2 in line 8.

It is good practice to avoid name hiding since this unnecessarily obfuscates the program. On the other hand, name hiding allows us (like in Program 2.9) to use our favorite identifier i as the name of the control variable in a for statement, without having to check whether there is some other name i somewhere else in the program. This is an acceptable and even useful application of name hiding.

Now we can get to the formal definition of scope in the general case (possible presence of multiple declarations of the same name). The *scope* of a declaration D is obtained from its potential scope as follows: For each declaration E in the potential scope of D such that both D and E declare the same name, the potential scope of E is removed from the scope of D. Figure 5 gives a symbolic picture of the situation.

In Program 2.9, the declarative region of the declaration in line 5 is line 4–10 (a block), its potential scope is line 5–10, and its scope is line 5 plus line 8–10. For the declaration in line 6, declarative region (a control statement), potential scope and scope are line 6–7.

Breaking down the scopes into lines is in general not possible, of course, since line breaks may (or may not) appear almost anywhere. If we want to talk about scope on a line-by-line basis, we have to format the program accordingly.

Storage duration. Related to the scope of a variable is its *storage duration*. This term denotes the time in which the address of the variable is valid, that is, some memory location is assigned to it.

For a variable with local scope, the storage duration is usually the time in which the program's control is in the variable's potential scope. During program execution, this means that whenever the variable declaration is reached, some memory location is assigned and the address becomes valid. And whenever the execution gets to the end of the declarative region, the associated memory is freed and the variable's address becomes

Figure 5: *Potential scopes of declarations* D, E_1, E_2, E_3 *of the same name, drawn as rectangles with the corresponding declaration in the upper left corner (left); on the right, we see the resulting scopes of* D *(dark gray),* E_1, E_3 *(light gray) and* E_2 *(white).*

invalid. We therefore get a "fresh instance" of the variable everytime its declaration is executed.

This behavior is called *automatic storage duration*. For example, in

```
for (unsigned int i = 0; i < 10; ++i) {
  const int k = i;
  ...
}
```

the initialization takes place in each iteration, and it yields a different value each time. This is fine, since the "constness" individually refers to each of the ten instances of `k` that are being generated throughout the loop.

As a more concrete example, consider the following code fragment.

```
1 int i = 5;
2 for (int j = 0; j < 5; ++j) {
3   std::cout << ++i; // outputs 6, 7, 8, 9, 10
4   int k = 2;
5   std::cout << --k; // outputs 1, 1, 1, 1, 1
6 }
```

Since line 3 belongs to the scope of the declaration in line 1, the effect of line 3 is to increment the variable defined in line 1 in every iteration of the `for` statement. Line 5, on the other hand, belongs to the scope of the declaration in line 4; the effect of line 5 is therefore to decrement the "fresh" variable `k` in every iteration, and this always results in value 1.

In contrast, a variable that is defined in namespace scope or global scope has *static*

storage duration. This means that its address is determined at the beginning of the program's execution, and it does not change (hence "static") nor become invalid until the execution of the program ends. The variables named by `std::cin` and `std::cout`, for instance, have static storage duration. Variables with static storage duration are also referred to as *static variables*.

2.4.4 Iteration: while statements

So far, we have seen one iteration statement, the `for` statement. The `while` *statement* is a simplified `for` statement, where both *init-statement* and *expression* are omitted. Its syntax is

```
while ( condition )
  statement
```

where *condition* and *statement* are as in a `for` statement. As before, *statement* is referred to as the body of the `while` statement. Semantically, a `while` statement is equivalent to the corresponding `for` statement

```
for ( ; condition ; )
  statement
```

The execution order is therefore *condition*, *statement*, *condition*,... until *condition* returns *false*.

Since `while` statements are so easy to rewrite as `for` statements, why do we need them? The main reason is readability. As its name suggests, a `for` statement is typically perceived as a counting loop in which the increment (or decrement) of a single variable is responsible for the progress towards termination. In this case, the progress is most conveniently made in the `for` statement's *expression*. But the situation can be more complex: the progress may depend on the values of several variables, or on some condition that we check in the loop's body. In some of these cases, a `while` statement is preferable. The next section describes an example.

The Collatz problem. Given a natural number $n \in \mathbb{N}$, we consider the *Collatz sequence* $n_0, n_1, n_2, \ldots$ with $n_0 = n$ and

$$n_i = \begin{cases} n_{i-1}/2, & \text{if } n_{i-1} \text{ is even} \\ 3n_{i-1} + 1, & \text{if } n_{i-1} \text{ is odd} \end{cases} \qquad i \geq 1.$$

For example, if $n = 5$, we get the sequence $5, 16, 8, 4, 2, 1, 4, 2, 1, \ldots$. Since the sequence gets repetitive as soon as 1 appears, we may stop at this point. Program 2.10 reads in a number n and outputs the elements of the sequence $(n_i)_{i \geq 1}$ until the number 1 appears.

```
// Program: collatz.cpp
// Compute the Collatz sequence of a number n.

#include <iostream>

int main()
{
  // Input
  std::cout << "Compute the Collatz sequence for n =? ";
  unsigned int n;
  std::cin >> n;

  // Iteration
  while (n > 1) {
    if (n % 2 == 0)
      n = n / 2;
    else
      n = 3 * n + 1;
    std::cout << n << " ";
  }
  std::cout << "\n";
  return 0;
}
```

Program 2.10: *../progs/lecture/collatz.cpp*

The loop can of course be written as a `for` statement with empty *init-statement* and *expression*, but the resulting variant of the program is less readable since it tries to advertise the rather complicated iteration as a simple counting loop. As a rule of thumb, if there is a simple *expression* that captures the loop's progress, use a `for` statement. Otherwise, consider formulating your loop as a `while` statement.

Talking about progress: is it clear that the number 1 always appears? If not, the program `collatz.cpp` contains an infinite loop for certain values of `n`. If you play with the program, you will observe that 1 indeed appears for all numbers you try, although this may take a while. You will find, for example, that the Collatz sequence for $n = 27$ is

> 27, 82, 41, 124, 62, 31, 94, 47, 142, 71, 214, 107, 322, 161, 484, 242, 121, 364, 182, 91, 274, 137, 412, 206, 103, 310, 155, 466, 233, 700, 350, 175, 526, 263, 790, 395, 1186, 593, 1780, 890, 445, 1336, 668, 334, 167, 502, 251, 754, 377, 1132, 566, 283, 850, 425, 1276, 638, 319, 958, 479, 1438, 719, 2158, 1079, 3238, 1619, 4858, 2429, 7288, 3644, 1822, 911, 2734, 1367, 4102, 2051, 6154, 3077, 9232, 4616, 2308, 1154, 577, 1732, 866, 433, 1300, 650, 325, 976, 488, 244, 122, 61, 184, 92, 46, 23, 70, 35, 106, 53, 160, 80, 40, 20, 10, 5, 16, 8, 4, 2, 1.

It is generally believed that 1 eventually comes up for all values of n, but mathematicians have not yet been able to produce a proof of this conjecture. As innocent as it looks, this problem seems to be a very hard mathematical nut to crack (see also the Details section), but you are certainly invited to give it a try!

2.4.5 Iteration: do statements

`Do` *statements* are similar to `while` statements, except that the condition is evaluated *after* every iteration of the loop instead of *before* every iteration. Therefore, in contrast to for– and `while` statements, the body of a `do` statement is executed at least once. The syntax of a `do` statement is as follows.

```
do
  statement
while ( expression );
```

where *expression* is of a type whose values can be converted to `bool`.

The semantics is defined as follows. An iteration of the loop consists of first executing *statement* and then evaluating *expression*. If *expression* returns *true* then another iteration follows. Otherwise, the `do` statement terminates. The execution order is therefore *statement*, *expression*, *statement*, *expression*, ... until *expression* returns *false*.

Alternatively, the semantics could be defined in terms of the following equivalent `for` statement.

```
for ( bool firsttime = true; firsttime || expression; firsttime = false )
  statement
```

This behaves like our "simulation" of the `while` statement, except that in the first iteration, *expression* is not evaluated (due to short circuit evaluation, see Section 2.3.3), and *statement* is executed unconditionally.

Consider a simple calculator-type application in which the user enters a sequence of numbers, and after each number the program outputs the sum of the numbers entered so far. By entering `0`, the user indicates that the program should stop. This is most naturally written using a `do` statement, since the termination condition can only be checked *after* the next number has been entered.

```
int a;      // next input value
int s = 0; // sum of values so far
do {
  std::cout << "next number =? ";
  std::cin >> a;
  s += a;
  std::cout << "sum = " << s << "\n";
} while (a != 0);
```

In this case, it is *not* possible to declare a where we would usually do it, namely immediately before the input statement. The reason is that a would then be local to the body of the do statement and would not be visible in the do statement's *expression* a != 0.

2.4.6 Jump statements

At this point, we would like to extend our arsenal of control statements with a special type of statements that are referred to as *jump statements*. These statements are not necessary in the sense that they would allow you to do something which is not possible otherwise. Instead, just like while– and do statements (which are also unnecessary in that sense), jump statements provide additional flexibility in designing iteration statements. You should use this flexibility wherever it allows you to improve your code. However, be also warned that jump statements should be used with care since they tend to complicate the control flow. The complication of the control flow has to be balanced by a significant gain in other categories (such as code readability). Therefore, think carefully before introducing a jump statement!

When a jump statement is executed, the program flow unconditionally "jumps" to a certain point. There are two different jump statements that we want to discuss here.

The first jump statement is called a break *statement*; its syntax is rather simple.

```
break;
```

When a break statement is executed within an iteration statement, the smallest enclosing iteration statement terminates immediately. The execution continues at the statement after the iteration statement (if any). For example,

```
for (;;) break;
```

is not an infinite loop but rather a complicated way of writing a null statement. Here is a more useful appearance of break. In our calculator example from Page 92, it would be more elegant to suppress the irrelevant addition of 0 in the last iteration. This can be done with the following loop.

```
for (;;) {
  std::cout <<  "next number =? ";
  std::cin >> a;
  if (a == 0) break;
  s += a;
  std::cout << "sum = " << s << "\n";
}
```

Here, we see the typical usage of break, namely the termination of a loop "somewhere in the middle". Note that we could equivalently write

```
do {
  std::cout <<  "next number =? ";
```

```
  std::cin >> a;
  if (a == 0) break;
  s += a;
  std::cout << "sum = " << s << "\n";
} while (true);
```

In this case `for` is preferable, though, since it nicely reads as "forever". Of course, the same functionality is possible without `break`, but the resulting code requires an additional block and evaluates `a != 0` twice.

```
do {
  std::cout <<  "next number =? ";
  std::cin >> a;
  if (a != 0) {
    s += a;
    std::cout << "sum = " << s << "\n";
  }
} while (a != 0);
```

The second jump statement is called a `continue` *statement*; again the syntax is simple.

```
continue;
```

When a continue statement is executed, the remainder of the smallest enclosing iteration statement's body is skipped, and execution continues at the end of the body. The iteration statement itself is *not* terminated.

If the surrounding iteration statement is a `while` or `do` statement, the execution therefore continues by evaluating its *condition* (*expression*, respectively). If the surrounding iteration statement is a `for` statement, the execution continues by evaluating its *expression* and then its *condition*. Like the `break` statement, the `continue` statement can therefore be used to manipulate the control flow "in the middle" of a loop.

In our calculator example, the following variant of the loop ignores negative input. Again, it would be possible to do this without `continue`, at the expense of another nested block.

```
for (;;) {
  std::cout <<  "next number =? ";
  std::cin >> a;
  if (a < 0) continue;
  if (a == 0) break;
  s += a;
  std::cout << "sum = " << s << "\n";
}
```

2.4.7 Equivalence of iteration statements

In terms of pure functionality, the `while`– and `do` statements are redundant, as both of them can equivalently be expressed using a `for` statement. This may create the impression that `for` statements have more expressive power than `while`– and `do` statements. In this section we show that this is not the case: all three iteration statements are functionally equivalent. More precisely, we show how to use

- `do` statements to express `while` statements, and
- `while` statements to express `for` statements.

If we denote "A can be used to express B" by $A \Rightarrow B$, we therefore have

`do` statement $\Rightarrow$ `while` statement $\Rightarrow$ `for` statement $\Rightarrow$ `do` statement,

where we know the last implication from the previous section. Together, this clearly "proves" the claimed equivalence.

Note that we put the word *proves* in quotes, as our reasoning cannot be considered a formal proof. In order to really prove a statement like this, we first of all would have to be more formal in defining the semantics of statements. Semantics of programming languages is a subject of its own, and the formal treatment of semantics is way beyond what we can do here. In other words: The following is as much of a "proof" as you will get here, but it is sufficient to understand the relations between the three iteration statements.

do statement $\Rightarrow$ while statement. Consider the `while` statement

```
while ( condition )
  statement
```

Your first idea how to simulate this using a `do` statement might look like this:

```
if ( condition )
  do
    statement
  while ( condition );
```

Indeed, this induces the execution order *condition*, *statement*, *condition*,... until *condition* returns *false* and the statement terminates. But there is a simple technical problem: if *condition* is a variable declaration, we can't use it as the *expression* in the `do` statement. Here is a reformulation that works. (We are not saying that this should be done in practice. On the contrary, this should *never* be done in practice. This section is about *conceptual* equivalence, not about practical equivalence.)

```
do
  if ( condition )
```

```
    statement
  else
    break;
while ( true );
```

This induces exactly the `while` statement's execution order *condition*, *statement*, *condition*,... until *condition* returns *false* and the loop is terminated using `break`.

while statement ⇒ for statement. Simulating the `for` statement

```
for ( init-statement condition; expression )
  statement
```

by a `while` statement seems easy:

```
{
  init-statement
  while ( condition ) {
    statement
    expression;
  }
}
```

Indeed, this will work, *unless statement* contains a `continue`. In the `for` statement, execution would then proceed with the evaluation of *expression*, but in the simulating `while` statement, *expression* is skipped, and *condition* comes next. This reformulation is therefore wrong. Here is a version that works:

```
{
  init-statement
  while ( condition ) {
    bool b = false;
    while ( b = !b )
      statement
    if ( b ) break;
    expression;
  }
}
```

This looks somewhat more complicated, so let us explain what is going on.

We may suppose that the identifier `b` does not appear in the given `for` statement (otherwise we choose a different name). Note that the whole statement forms a separate block, as does a `for` statement. A potential declaration in *init-statement* as well as the scope of `b` is thus limited to this block.

Consider an execution of the outer `while` statement. First, *condition* is evaluated, and if it returns *false* the statement terminates. Otherwise, the variable `b` is set to *true* in the inner `while` statement's condition, meaning that *statement* is executed next. (Recall

that the assignment operator returns the new value of its left operand.) If *statement* does not contain a `break`, the inner loop evaluates its condition for the second time. In doing so, `b` is set to *false*, and the condition returns *false*. Therefore, the inner loop terminates. Since `b` is now *false*, *expression* is evaluated next, followed by *condition*. This induces the `for` statement's execution order *condition*, *statement*, *expression*, *condition*,... until *condition* returns *false* and the outer loop terminates.

In the case where *statement* contains a `break`, the inner loop terminates immediately, and `b` remains *true*. In this case, we also terminate the outer loop that represents our original `for` statement.

In retrospect, we should now check that jump statements cause no harm in our previous reformulation of the `while` statement in terms of the `do` statement. We leave this as an exercise.

2.4.8 Choosing the "right" iteration statements

We have seen that from a functional point of view, the `for` statement, the `while` statement and the `do` statement are equivalent. Moreover, the `break` and `continue` statements are redundant. Still, C++ offers all of these statements, and this gives you the freedom (but also the burden) of choosing the appropriate control statements for your particular program.

Writing programs is a dynamic process. Even though the program may do what you want at some point, the requirements change, and you will keep changing the program in the future. Even if there is currently no need to change the functionality of the program, you may want to replace a complicated iteration statement by an equivalent simpler formulation. The general theme here is *refactoring*: the process of rewriting a program to improve its readability or structure, while keeping its functionality unchanged.

Here is a simple guideline for writing "good" loops. Choose the loop that leads to the most *readable* and *concise* formulation. This means

- few statements,
- few lines of code,
- simple control flow, and
- simple expressions.

Almost never there is *the* one and only best formulation; however, there are always arguably bad choices which you should try to avoid. Usually, there are some tradeoffs, like fewer lines of code versus more complicated expressions, and there is also some amount of personal taste involved. You should experience and find out what suits you best.

Let us look at some examples to show what we mean. Suppose that you want to output the odd numbers between 0 and 100. Having just learned about the `continue` statement, you may write the following loop.

```
for (unsigned int i = 0; i < 100; ++i) {
  if (i % 2 == 0) continue;
  std::cout << i << "\n";
}
```

This is perfectly correct, but the following version is preferable since it has **fewer statements** and **fewer lines of code**.

```
for (unsigned int i = 0; i < 100; ++i)
  if (i % 2 != 0) std::cout << i << "\n";
```

This variant still contains nested control statements; but you can get rid of the `if` statement and obtain code with **simpler control flow**.

```
for (unsigned int i = 1; i < 100; i += 2)
  std::cout << i << "\n";
```

The same output can be produced with a `while` statement and equally simple control flow.

```
int i = -1;
while ((i += 2) < 100)
  std::cout << i << "\n";
```

But here, the condition is more complicated, since it combines assignment and comparison operators. Such expressions are comparatively difficult to understand due to the effect of the assignment operation. Also, the initialization of `i` to -1 is counter-intuitive, given that we deal with natural numbers.

You can solve the latter problem and at the same time get **simpler expressions** by writing

```
unsigned int i = 1;
while (i < 100) {
  std::cout << i << "\n";
  i += 2;
}
```

The price to pay is that you get less concise code; there are now five lines instead of the two lines that the `for` statement needs. It seems that for the simple problem of writing out odd numbers, a `for` statement with *expression* `i += 2` is the loop of choice.

2.4.9 Details

Nested if-else statements. Consider the statement

```
if(true) if (false); else std::cout << "Where do I belong?";
```

It is not a priori clear what its effect is: if the `else` branch belongs to the outer `if`, there will be no output (since the condition has value *true*), but if the `else` branch belongs to the inner `if`, we get the output `Where do I belong?`

The intuitive rule is that the `else` branch belongs to the `if` immediately preceding it, in our case to the inner `if`. Therefore, the output is `Where do I belong?`, and we should actually format the statement like this:

```
if(true)
   if (false)
     ; // null statement
   else
     std::cout << "Where do I belong?";
```

Whenever you are unsure about rules like this, you can make the structure clear through explicit blocks:

```
if(true) {
  if (false) {
    ; // null statement
  }
  else {
    std::cout << "Where do I belong?";
  }
}
```

The switch statement. Besides `if...else` there exists a second selection statement in C++: the `switch` *statement*. It is useful to select between many alternative statements, using the following syntax.

`switch` (*condition*)
  statement

The value of *condition* must be convertible to an integral type. This is in contrast to the other control statements where *condition* has to be convertible to `bool`.

statement is usually a block that contains several *labels* of the form

`case` *literal*:

where *literal* is a literal of integral type. For no two labels shall these literals have the same value. There can also be a label `default:`.

The semantics of a `switch` statement is the following. *condition* is evaluated and the result is compared to each of the literals which appear in a label in *statement*. If for any of them the values agree, the execution continues at the statement immediately following the label. If there is no agreement but a `default:` label, the execution continues at the statement immediately following the `default:` label. Otherwise, *statement* is ignored and the execution continues after the `switch` statement.

Note that `switch` only selects an entry point for the processing of *statement*, it does not exit when the execution reaches another label. If one wants to separate the different alternatives, one has to use `break` (and this is the only legal use of `break` outside of

an iteration statement). Consider for example the following piece of code, and let us suppose that x is a variable of type int.

```
switch (x) {
  case 0: std::cout << "0";
  case 1: std::cout << "1"; break;
  default: std::cout << "whatever";
}
```

For x==0 the output is 01; for x==1 the output is 1; otherwise we get the output whatever.

The switch statement is powerful in the sense that it allows the different alternatives to share code. However, this power also makes switch statements hard to read and error prone. A frequent problem is that one forgets to put a break where there should be one. Therefore, we mention switch here for completeness only. Whenever there are only a few alternatives to be distinguished, play it safe and use if...else rather than switch.

The Halting Problem, Decidability, and Computability. The halting problem is one of the fundamental problems in the theory of computation. Informally speaking, the problem is to decide (using an algorithm) whether a given program halts (terminates) when executed on a given input (program state). The term "program" may refer to a C++-program, but also to a program in any other common programming language.

To attack the problem formally, the British mathematician Alan Turing (1912-1954) defined in a seminal paper a "minimal" programming language; a program in this language is known as a *Turing machine*.

Turing proved that the halting problem is undecidable for Turing machines, but the same arguments can also be used to prove the same statement for C++ programs.

What does "undecidable" mean? We have seen a simple loop for which it was painfully evident that it is an infinite loop, haven't we? Yes, indeed one *can* decide the halting problem for many concrete programs. Undecidable means that (in a particular model of computation) there cannot be an algorithm that decides the halting problem for *all possible* programs.

Despite their simplicity, Turing machines are a widely accepted model of computation; in fact, just like machine language, Turing machines can do everything that C++ programs can do, except that they usually need a huge number of very primitive operations for that.

At the same time as Turing, the American mathematician Alonzo Church (1903–1995) developed a computational model called λ-calculus. As it turned out, his model is equivalent to Turing machines in terms of computational power. The Church-Turing thesis states that "every function that is naturally regarded as computable can be computed by a Turing machine". As there is no rigorous definition of what is "naturally regarded as computable", this statement is not a theorem but a hypothesis that cannot be proven mathematically. As of today, the hypothesis has not been disproved. In theoretical computer science the term *computable* used without further qualification is a synonym for "computable by a Turing machine" (equivalently, a C++ program).

Point of declaration. Our approach of defining potential scope and scope line by line is a simplification, even if the code is suitably formatted and we only have one declaration per line. The truth is that the point of declaration of `i` in

```
int i = 5;
```

is in the *middle* of the declaration, after the name `i` has appeared. The potential scope therefore does not include the full line, but only the part starting from =. This explains what happens in the following code fragment, but fortunately this is consistent with our line-by-line approach. In

```
1 int i = 5;
2 {
3   int i = i;
4 }
```

the name `i` after the = in line 3 refers to the declaration in line 3. Consequently, `i` is initialized with itself in this line, meaning that its value will be undefined, and not 5.

In other situations it may happen, though, that the appearance of a name in the declaration of the same name refers to a *previous* declaration of this name. For now, we can easily avoid such subtleties by the following rule: any declaration should contain the name to be declared once only.

The Collatz problem and the ?-operator. The Collatz sequence goes back to the German mathematician Lothar Collatz (1910–1990) who studied it in the 1930's. Several prizes have been offered to anyone who proves or disproves the conjecture that the number 1 appears in the Collatz sequence of every number $n \geq 1$. The famous Hungarian mathematician Paul Erdős offered $500, which is much by his standards (he used to offer much lower amounts for very difficult problems). Erdős said that "Mathematics is not yet ready for such problems". Indeed, the conjecture is still unsolved.

We have presented the computation of the Collatz sequence as an application of the `while` statement, pointing out that the conditional change of `n` is too complicated to put it into a `for` statement's *expression*. Well, that's not exactly true: the designers of C, the precursor to C++, had a weakness for very compact code and came up with the *conditional operator* that allows us to simulate `if` statements by expressions. The syntax of this *ternary* operator (arity 3) is

condition ? *expression1* : *expression2*

Here, *condition* is an expression of a type whose values can be converted to `bool`, and *expression1* and *expression2* are expressions. The semantics is as follows. First, *condition* is evaluated. If it returns *true*, *expression1* is evaluated, and its value is returned as the value of the composite expression. Otherwise (if *condition* returns *false*), *expression2* is evaluated, and its value is returned. The token ? is a sequence point (see Section 2.2.10), meaning that all effects of *condition* are processed before either *expression1* or *expression2* are evaluated.

Using the conditional operator, the loop of Program 2.10 could quite compactly be written as follows.

```
for ( ; n > 1; std::cout << (n % 2 == 0 ? n=n/2 : n=3*n+1) << " ");
```

We leave it up to you to decide whether you like this variant better.

Static variables. The discussion about storage duration above does not tell the whole story: it is also possible to define variables with local scope that have static storage duration.

This is done by prepending the keyword `static` to the variable declaration. For example, in

```
for (int i = 0; i < 5; ++i) {
  static int k = i;
  k += i;
  std::cout << k << "\n";
}
```

the address of `k` remains the same during all iterations, and `k` is initialized to `i` *once only*, in the first iteration. The above piece of code will therefore output the sequence of values $0, 1, 3, 6, 10$ (remember Gauss). Without the `static` keyword, the result would simply be the sequence of even numbers $0, 2, 4, 6, 8$.

Static variables have been quite useful in C, for example to count how often a specific piece of code is executed; in C++, they are less important.

For variables of fundamental type the initial value may be undefined, as in the definition `int x;`. However, the value is undefined only if `x` has automatic storage duration. In contrast, variables with static storage duration are always *zero-initialized*, that is, filled with a "zero" of the appropriate type.

Jump statements. There are two more jump statements in C++ that we haven't discussed in this section. One of them is the `return` statement that you already know (Section 2.1.15): it may occur only in a function, and its execution lets the program flow jump to the end of the corresponding function body. The other jump statement is the `goto` statement, but since this one is rarely needed (and somewhat difficult to use), we omit it.

2.4.10 Goals

Dispositional. At this point, you should ...

1) know the syntax and semantics of `if...else`–, `for`–, `while`–, and `do` statements;
2) understand the concepts block, selection, iteration, declarative region, scope, and storage duration;

3) understand the concept of an infinite loop and be aware of the difficulty of detecting such loops;
4) understand the conceptual equivalence of for–, while–, and do statements;
5) know the syntax and semantics of `continue`– and `break` statements;
6) know at least four criteria to judge the code quality of iteration statements.

Operational. In particular, you should be able to ...

(G1) check a given simple program (as defined below) for syntactical correctness and point out possible errors;
(G2) read and understand a given simple program and explain what happens during its execution;
(G3) find (potential) infinite loops in a given simple program;
(G4) find the matching declaration for a given identifier;
(G5) determine declarative region and scope of a given declaration;
(G6) reformulate a given `for`–, `while`–, or `do` statement equivalently using any of the other two statements;
(G7) compare the code quality of two given iteration statements and pick the one that is preferable (if any);
(G8) design simple programs for given tasks.

The term *simple program* refers to a program that consists of a main function in which up to four (possibly nested) iteration statements appear, plus some selection statements. Naturally, only the fundamental types and operations discussed in the preceding sections are used.

2.4.11 Exercises

Exercise 37 *Correct all syntax errors in the program below. What does the resulting program output for the following inputs?*
(a) -4 *(b)* 0 *(c)* 1 *(d)* 3 (G1)(G2)

```
1   #include <iostraem>
2   int main()
3   {
4     unsinged int x = +1;
5     { std::cin >> x; }
6     for (int y = 0u; y < x) {
7       std:cout << ++y;
8     return 0;
9   }
```

Exercise 38 *What is the problem with the code below? Fix it and explain what the resulting code computes.* (G2)(G3)

```
unsigned int s = 0;
do {
  int i = 1;
  if (i % 2 == 1) s *= i;
} while (++i < 10);
```

Exercise 39 *For each variable declaration in the following program give its declarative region and its scope in the form "line x–y". What is the output of the program?* (G2)(G5)

```
#include <iostream>
int main()
{
  int s = 0;
  {
    int i = 0;
    while (i < 4)
    {
      ++i;
      const int f = i + 1;
      s += f;
      const int s = 3;
      i += s;
    }
    const unsigned int t = 2;
    std::cout << s + t << "\n";
  }
  const int k = 1;
  return 0;
}
```

Exercise 40 *Consider the program given below for each of the listed input numbers. Determine the values of* `x`, `s`, *and* `i` *at begin of the first five iterations of the for-loop, before the condition is evaluated. What does the program output for these inputs? (a)* -1 *(b)* 1 *(c)* 2 *(d)* 3 (G2)(G3)

```
#include <iostream>
int main()
{
  int x;
  std::cin >> x;
  int s = 0;
  for (int i = 0; i < x; ++i) {
```

```
      s += i;
      x += s / 2;
    }
    std::cout << s << "\n";
    return 0;
  }
```

Exercise 41 *Find at least four problems in the code given below.* (G3)(G4)(G5)

```
#include <iostream>
int main()
{
  { unsigned int x; }
  std::cin << x;
  unsigned int y = x;
  for (unsigned int s = 0; y >= 0; --y)
    s += y;
  std::cout << "s=" << s << "\n";
  return 0;
}
```

Exercise 42 *For which input numbers is the output of the program given below well defined? List those input/output pairs and argue why your list is complete.* (G3)(G4)(G5)

```
#include <iostream>
int main()
{
  unsigned int x;
  std::cin >> x;
  int s = 0;
  for (unsigned int y = 1 + x; y > 0; y -= x)
    s += y;
  std::cout << "s=" << s << "\n";
  return 0;
}
```

Exercise 43 *Reformulate the code below equivalently in order to improve its readability. Describe the program's output as a function of its input* `n`. (G2)(G6)(G7)

```
unsigned int n;
std::cin >> n;
int x = 1;
if (n > 0) {
  int k = 0;
  bool e = true;
  do {
```

```
      if (++k == n) e = false;
      x *= 2;
    } while (e);
  }
  std::cout << x;
```

Exercise 44 *Reformulate the program below equivalently in order to improve its readability and efficiency. Describe the program's output as a function of its input* x. (G2)(G6)(G7)

```
#include <iostream>
int main()
{
  int x;
  std::cin >> x;
  int s = 0;
  int i = -10;
  do
    for (int j = 1;;)
      if (j++ < i) s += j - 1; else break;
  while (++i <= x);
  std::cout << s << "\n";
  return 0;
}
```

Exercise 45 *What is the behavior of Program 2.8 on Page 85 if the user inputs* 0 *or* 1*? Rewrite the program (if this is necessary at all) so that it correctly handles all possible inputs (we adopt the convention that* 0 *and* 1 *are not prime numbers).* (G2)(G3) (G8)

Exercise 46 *Write a program* fak-1.cpp *to compute the factorial* $n!$ *of a given input number* n. (G8)

Exercise 47 *Write a program* dec2bin.cpp *that inputs a natural number* n *and outputs the binary digits of* n *in reverse order. For example, for* n==2 *the output is* 01 *and for* n==11 *the output is* 1101 *(see also Exercise 50).* (G8)

Exercise 48 *Write a program* cross_sum.cpp *that inputs a natural number* n *and outputs the sum of the (decimal) digits of* n*. For example, for* n==10 *the output is* 1 *and for* n==112 *the output is* 4*.* (G8)

Exercise 49 *Write a program* perfect.cpp *to test whether a given natural number* n *is perfect. A number* $n \in \mathbb{N}$ *is called* perfect *if and only if it is equal to the sum of its proper divisors, that is,* $n = \sum_{k \in \mathbb{N},\, s.t.\, k<n \,\wedge\, k|n} k$*. For example,* $28 = 1+2+4+7+14$ *is perfect, while* $12 < 1+2+3+4+6$ *is not.*

Extend the program to find all perfect numbers between 1 *and* n. *How many perfect numbers exist in the range* $[1, 50000]$*?* (G8)

Exercise 50 *Write a program* `dec2bin2.cpp` *that inputs a natural number* n *and outputs the binary digits of* n *in the correct order. For example, for* n==2 *the output is* 10 *and for* n==11 *the output is* 1011 *(see also Exercise 47).* (G8)

Exercise 51 *Pete and Colin play a dice game against each other. Pete has three four-sided (pyramidal) dice, each with faces numbered 1, 2, 3, 4. Colin has two six-sided (cubical) dice, each with faces numbered 1, 2, 3, 4, 5, 6. Peter and Colin roll their dice and compare totals: the highest total wins. The result is a draw if the totals are equal.*

What is the probability that Pyramidal Pete beats Cubic Colin? What is the probability that Cubic Colin beats Pyramidal Pete? And what is the probability of a draw? As a consequence, is it a fair game, and if not, who would you rather be?

Write a program `dice.cpp` *that outputs the aforementioned probabilities as rational numbers of the form* p/q. *(This is a simplified version of Problem 205 from the* Project Euler, *see* `http://projecteuler.net/`.*)* (G8)

Exercise 52 *We know from Section 1.1 that it took Frank Nelson Cole around three years to find the factorization*

$$761838257287 \cdot 193707721$$

of the Mersenne number $2^{67} - 1$ *by hand calculations. Write a program* `cole.cpp` *that performs the same task (hopefully in less than three years).* (G8)

Hint: *You will need the type* `ifmp::integer`, *see Section 2.1.16.*

2.4.12 Challenges

Exercise 53 *The Ulam spiral is named after the Polish mathematician Stanislaw Ulam who discovered it accidentally in 1963 while being bored during a scientific meeting.*

The Ulam spiral is a method of "drawing" prime numbers. For this, one first lists all natural numbers in a spiral fashion, so that the integer grid $\mathbb{Z}^2$ *gets filled with the natural numbers (see Figure 6(a)). For the drawing, every grid point containing a prime number is replaced with a black pixel, while the composite numbers are replaced with white pixels (Figure 6(b)). When this is done for the first* n *numbers (*n *large), one observes a surprising phenomenon: the prime numbers tend to accumulate along diagonals (see Figure 6(c)).*

Write a program that outputs the Ulam spiral generated by the first n *numbers, where* n *is read from the input. You may restrict* n *such that the drawing fits into a window of* 500×500 *pixels, say.* (G8)

Hint: You may use the `libwindow` library to produce the drawing. The example program in its documentation should give you an idea how this can be done.

```
37—36—35—34—33—32—31
 |                  |
38  17—16—15—14—13  30
 |   |           |   |
39  18   5— 4— 3  12  29
 |   |   |       |   |   |
40  19   6   1— 2  11  28
 |   |   |       |   |
41  20   7— 8— 9—10  27
 |   |               |
42  21—22—23—24—25—26
 |
43—44—45—46—47—48—49...
```

(a) A spiral-shape listing of the natural numbers

(b) Replacing prime numbers with black and composite numbers with white pixels

(c) A larger spiral (around 200x 200 grid points)

Figure 6: *The Ulam Spiral*

Exercise 54 *The* n-queens problem *is to place* n *queens on an* $n \times n$ *chessboard such that no two queens threaten each other. Formally, this means that there is no horizontal, vertical, or diagonal with more than one queen in it. Write a program that outputs the number of different solutions to the* n*-queens problem for a given input* n*. Assuming a 32 bit system, the program should work up to* $n = 9$ *at least. Check through a web search whether the numbers that your program computes are correct.*

Exercise 55 *The largest Mersenne prime known as of September 2009 is*

$$2^{43,112,609} - 1,$$

see Section 1.1. In Exercise 19, we have asked you to find the number of decimal digits that this number has. In this challenge, we originally wanted to ask you to list all these digits, but in the interest of the TA that has to mark your solution, we decided to switch to the following variant: Write a program `famous_last_digits.cpp` *that computes and outputs the last 10 decimal digits of the above Mersenne prime!*

2.5 Floating point numbers

Furthermore, it has revealed the ratio of the chord and arc of ninety degrees, which is as seven to eight, and also the ratio of the diagonal and one side of a square which is as ten to seven, disclosing the fourth important fact, that the ratio of the diameter and circumference is as five-fourths to four.

Indiana House Bill No. 246, defining $\frac{2\sqrt{2}}{\pi} = \frac{7}{8}$, $\sqrt{2} = \frac{10}{7}$, *and* $\frac{1}{\pi} = 5/16$ *(1897)*

This section discusses the floating point number types `float` *and* `double` *for approximating real numbers. You will learn about floating point number systems in general, and about the IEEE standard 754 that describes two specific floating point number systems. We will point out the strengths and weaknesses of floating point numbers and give you three guidelines to avoid common pitfalls in computing with floating point numbers.*

When converting degrees Celsius into Fahrenheit with the program `fahrenheit.cpp` in Section 2.2, we make mistakes. For example, 28 degrees Celsius are 82.4 degrees Fahrenheit, but not 82 as output by `fahrenheit.cpp`. The reason for this mistake is that the integer division employed in the program simply "cuts off" the fractional part. What we need is a type that allows us to represent and compute with fractional numbers like 82.4.

For this, C++ provides two *floating point number* types `float` and `double`. Indeed, if we simply replace the declaration `int celsius` in `fahrenheit.cpp` by `float celsius`, the resulting program outputs `82.4` for an input value of 28. Floating point numbers also solve another problem that we had with the types `int` and `unsigned int`: `float` and `double` have a much larger value range and are therefore suitable for "serious" computations. In fact, computations with floating point numbers are very fast on modern platforms, due to specialized processors.

Fixed versus floating point. If you think about how to represent decimal numbers like 82.4 using a fixed number of decimal digits (10 digits, say), a natural solution is this: you partition the 10 available digits into 7 digits before the decimal point, say, and 3 digits after the decimal point. Then you can represent all decimal numbers of the form

$$\sum_{i=-3}^{6} \beta_i 10^i,$$

with $\beta_i \in \{0, \ldots, 9\}$ for all i. This is called a *fixed point representation.*

There are, however, two obvious disadvantages of a fixed point representation. On the one hand, the value range is very limited. We have already seen in Section 2.2.5 that the largest `int` value is so small that it hardly allows any interesting computations (as an example, try out Program 2.1 on some larger input). A fixed point representation is even worse in this respect, since it reserves some of our precious digits for the fractional part after the decimal point, even if these digits are not—or not fully—needed (as in 82.4).

The second disadvantage is closely related: even though the two numbers 82.4 and 0.0824 have the same number of significant digits (namely 3), the latter number is not representable with only 3 digits after the decimal point. Here, we are wasting the 7 digits before the decimal point, but we are lacking digits after the decimal point.

A *floating point representation* resolves both issues by representing a number simply as its sequence of decimal digits (an integer called the *significand*), *plus* the information "where the decimal point is". Technically, one possibility to realize this is to store an *exponent* such that the represented number is of the form

$$\mathit{significand} \cdot 10^{\mathit{exponent}}.$$

For example,

$$\begin{aligned} 82.4 &= 824 \cdot 10^{-1}, \\ 0.0824 &= 824 \cdot 10^{-4}. \end{aligned}$$

2.5.1 The types float and double

The types `float` and `double` are fundamental types provided by C++, and they store numbers in floating point representation.

While the fundamental types `int` and `unsigned int` are meant to approximate the "mathematical types" $\mathbb{Z}$ and $\mathbb{N}$, respectively, the goal of both `float` and `double` is to approximate the set $\mathbb{R}$ of real numbers. Since there are much more real numbers than integers, this goal seems even more ambitious (and less realistic) than trying to approximate $\mathbb{Z}$, say, with a finite value range. Nevertheless, the two types `float` and `double` are very useful in practical applications. The floating point representation allows values that are much larger than any value of type `int` and `unsigned int`. In fact, the value ranges of the floating point number types `float` and `double` are sufficient in most applications.

Values of these two types are referred to as *floating point numbers*, where `double` usually allows higher (namely, *double*) precision in approximating real numbers.

On the types `float` and `double` we have the same arithmetic, relational, and assignment operators as on integral types, with the same associativities and precedences. The only exception is that the modulus operators `%` and `%=` are available for integral types only. This makes sense, since division over `float` and `double` is meant to model the true division over $\mathbb{R}$ which has no remainder.

Like integral types, the floating point number types are *arithmetic types*, and this completes the list of fundamental arithmetic types in C++.

Literals of type float and double. Literals of types `float` and `double` are more complicated than literals of type `int` or `unsigned int`. For example, `1.23e-7` is a valid `double` literal, representing the value $1.23 \cdot 10^{-7}$. Literals of type `float` look the same as literals of type `double`, followed by the letter `f` or `F`.

In its most general form, a `double` literal consists of an *integer part*, followed by a *fractional part* (starting with the *decimal point* `.`), and an *exponential part* (starting with the letter `e` or `E`). The literal `1.23e-7` has all of these parts.

Both the integer part as well as the fractional part (after the decimal point) are sequences of digits from `0` to `9`, where *one* of them may be empty, like in `.1` (meaning 0.1) and in `1.` (meaning 1.0). The exponential part (after the letter `e` or `E`) is also a sequence of digits, preceded by an optional `+` or `-`. *Either* the fractional part *or* the exponential part may be omitted. Thus, `123e-9` and `1.23` are valid `double` literals, but `123` is not, in order to avoid confusion with `int` literals.

The value of the literal is obtained by scaling the fractional decimal value defined by the integer part and the fractional part by 10^e, where e is the (signed) decimal integer in the exponential part (defined as 0, if the exponential part is missing).

To show floating point numbers in action, let us write a program that "computes" a fully-fledged real number, namely the Euler constant

$$\sum_{i=0}^{\infty} \frac{1}{i!} = 2.71828\ldots$$

You may recall that this sum converges quickly, so we should already get a good approximation for the Euler constant when we sum up the first 10 terms, say. Program 2.11 does exactly this.

```
// Program: euler.cpp
// Approximate the Euler number e.

#include <iostream>

int main ()
{
  // values for term i, initialized for i = 0
  float t = 1.0f;    // 1/i!
  float e = 1.0f;    // i-th approximation of e

  std::cout << "Approximating the Euler number...\n";
  // steps 1,...,n
  for (unsigned int i = 1; i < 10; ++i) {
```

```
15      e += t /= i;     // compact form of t = t / i; e = e + t
16      std::cout << "Value after term " << i << ": " << e << "\n";
17    }
18
19    return 0;
20  }
```

Program 2.11: *../progs/lecture/euler.cpp*

When you run the program, its output may look like this.

```
Approximating the Euler constant...
Value after term 1: 2
Value after term 2: 2.5
Value after term 3: 2.66667
Value after term 4: 2.70833
Value after term 5: 2.71667
Value after term 6: 2.71806
Value after term 7: 2.71825
Value after term 8: 2.71828
Value after term 9: 2.71828
```

It seems that we do get a good approximation of the Euler constant in this way. What remains to be explained is how the mixed expression `e += t /= i` in line 15 is dealt with that contains operands of types `unsigned int` and `float`. Note that since the arithmetic assignment operators are right-associative (Table 1 on Page 51), this expression is implicitly parenthesized as `e += (t /= i)`. When evaluated in iteration `i`, it therefore first divides `t` by `i` (corresponding to the step from $1/(i-1)!$ to $1/i!$), and then it adds the resulting value $1/i!$ to the approximation `e`.

2.5.2 Mixed expressions, conversions, and promotions

The floating point number types are defined to be more general than any integral type. Thus, in mixed composite expressions, integral operands get converted to the respective floating point number type (see also Section 2.2.7 where we first saw this mechanism for mixed expressions over the types `int` and `unsigned int`). The resulting value is the representable value *nearest* to the original value, where ties are broken in an implementation-defined fashion. In particular, if the original integer value is in the value range of the relevant floating point number type, the value remains unchanged.

This in particular explains why the change of `int celsius` to `float celsius` in the program `fahrenheit.cpp` leads to the behavior we want: during evaluation of the expression `9 * celsius / 5 + 32`, all integral operands are eventually converted to `float`, so that the computation takes place exclusively over the type `float`.

In the program `euler.cpp`, we have the same kind of conversion: in the mixed expression `t /= i`, the `unsigned int` operand `i` gets converted to the type `float` of the other operand `t`.

The type `double` is defined to be more general than the type `float`. Thus, a composite expression involving operands of types `float` and `double` is of type `double`. When such an expression gets evaluated, every operand of type `float` is *promoted* to `double`. Recall from Section 2.3.2 that promotion is a term used to denote certain privileged conversions in which no information gets lost. In particular, the value range of `double` must contain the value range of `float`.

In summary, the hierarchy of arithmetic types from the least general to the most general type is

$$\texttt{bool} \prec \texttt{int} \prec \texttt{unsigned int} \prec \texttt{float} \prec \texttt{double}.$$

We already know that a conversion may also go from the more general to the less general type, see Section 2.2.7. This happens for example in the declaration statement

```
int i = -1.6f;
```

When a floating point number is converted to an integer, the fractional part is discarded. If the resulting value is in the value range of the target type, we get this value, otherwise the result of the conversion is undefined. In the previous example, this rule initializes `i` with -1 (and *not* with the nearest representable value -2).

When `double` values are converted to `float`, we again get the nearest representable value (with ties broken in an implementation-dependent way), *unless* the original value is larger or smaller than any `float` value. In this latter case, the conversion is undefined.

2.5.3 Explicit conversions

Conversions between integral and floating point number types are common in practice. For example, the conversion of a nonnegative `float` value x to the type `unsigned int` corresponds to the well-known *floor function* $\lfloor x \rfloor$ that rounds down to the next integer. Conversely, it can make sense to perform an integral computation over a floating point number type, if this latter type has a larger value range.

Explicit conversion allows to convert a value of any arithmetic type directly into any other arithmetic type, without the detour of defining an extra variable like in `int i = -1.6f;` To obtain the `int` value resulting from the `float` value -1.6, we can simply write the expression `int(-1.6f)`.

The general syntax of an explicit conversion, also called a *cast expression*, is

T (*expr*)

where T is a type, and *expr* is an expression. The cast expression is valid if and only if the corresponding conversion of *expr* to the type T (as in `T x = expr`) is defined.

For certain "complicated" type names T, it is necessary to parenthesize T, like in the cast expression `(unsigned int)(1.6f)`.

2.5.4 Value range

For integral types, the arithmetic operations may fail to compute correct results only due to over- or underflow. This is because the value range of each integral type is a *contiguous* subset of $\mathbb{Z}$, with no "holes" in between.

For floating point number types, this is not true: with finite (and even with countable) value range, it is impossible to represent a subset of $\mathbb{R}$ with more than one element but no holes. In contrast, over- or underflows are less of an issue: the representable values usually span a huge interval, much larger than for integral types. If you print the largest `double` value on your platform via the expression

```
std::numeric_limits<double>::max()
```

you might for example get the output `1.79769e+308`. Recall that this means $1.79769 \cdot 10^{308}$, a pretty large number.

Let us approach the issue of holes with a very simple program that asks the user to input two floating point numbers *and* their difference. The program then checks whether this is indeed the correct difference. Program 2.12 performs this task.

```
// Program: diff.cpp
// Check subtraction of two floating point numbers

#include <iostream>

int main()
{
  // Input
  float n1;
  std::cout << "First number     =? ";
  std::cin >> n1;

  float n2;
  std::cout << "Second number    =? ";
  std::cin >> n2;

  float d;
  std::cout << "Their difference =? ";
  std::cin >> d;

  // Computation and output
  std::cout << "Computed difference - input difference = "
            << n1 - n2 - d << ".\n";
  return 0;
}
```

Program 2.12: *../progs/lecture/diff.cpp*

Here is an example run showing that the authors are able to correctly subtract 1 from 1.5.

```
First number      =? 1.5
Second number     =? 1.0
Their difference =? 0.5
Computed difference - input difference = 0.
```

But the authors can apparently *not* correctly subtract 1 from 1.1:

```
First number      =? 1.1
Second number     =? 1.0
Their difference =? 0.1
Computed difference - input difference = 2.23517e-08.
```

What is going on here? After double checking our mental arithmetic, we must conclude that it's the *computer* and not us who cannot correctly subtract. To understand why, we have to take a somewhat closer look at floating point numbers in general.

2.5.5 Floating point number systems

A *finite floating point number system* is a finite subset of $\mathbb{R}$, defined by four numbers $2 \leq \beta \in \mathbb{N}$ (the *base*), $1 \leq p \in \mathbb{N}$ (the *precision*), $e_{\min} \in \mathbb{Z}$ (the *smallest exponent*) and $e_{\max} \in \mathbb{Z}$ (the *largest exponent*).

The set $\mathcal{F}(\beta, p, e_{\min}, e_{\max})$ of real numbers represented by this system consists of all floating point numbers of the form

$$s \cdot \sum_{i=0}^{p-1} d_i \beta^{-i} \cdot \beta^e,$$

where $s \in \{-1, 1\}$, $d_i \in \{0, \ldots, \beta - 1\}$ for all i, and $e \in \{e_{\min}, \ldots, e_{\max}\}$.

The number s is the *sign*, the sequence $d_0 d_1 \ldots d_{p-1}$ is called the *significand* (an older equivalent term is *mantissa*), and the number e is the *exponent*.

We sometimes write a floating point number in the form

$$\pm d_0.d_1 \ldots d_{p-1} \cdot \beta^e.$$

For example, using base $\beta = 10$, the number 0.1 can be written as $1.0 \cdot 10^{-1}$, and as $0.1 \cdot 10^0$, $0.01 \cdot 10^1$ and in many other ways.

The representation of a number becomes unique when we restrict ourselves to the set $\mathcal{F}^*(\beta, p, e_{\min}, e_{\max})$ of *normalized* numbers, i.e. the ones with $d_0 \neq 0$. The downside of this is that we lose some numbers (in particular the number 0, but let's not worry about this now). More precisely, normalization loses exactly the numbers of absolute value smaller than $\beta^{e_{\min}}$ (see also Exercise 62).

For a fixed exponent e, the smallest positive normalized number is

$$1.0 \ldots 0 \cdot \beta^e = \beta^e,$$

while the largest one is (by the formula $\sum_{i=0}^{n} x^i = (x^{n+1}-1)/(x-1)$ for $x \neq 1$)

$$(\beta-1).(\beta-1)\ldots(\beta-1)\cdot\beta^e = \sum_{i=0}^{p-1}(\beta-1)\beta^{-i}\cdot\beta^e = \left(1-\left(\frac{1}{\beta}\right)^p\right)\beta^{e+1} < \beta^{e+1}.$$

This means that the normalized numbers are "sorted by exponent".

Most floating point number systems used in practice are *binary*, meaning that they have base $\beta = 2$. In a binary system, the decimal numbers 1.1 and 0.1 are not representable, as we will see next; consequently, errors are made in converting them to floating point numbers, and this explains the strange behavior of Program 2.12.

Computing the floating point representation. In order to convert a given positive decimal number x into a normalized binary floating point number system $\mathcal{F}^*(2, p, e_{\min}, e_{\max})$, we first compute its *binary expansion*

$$x = \sum_{i=-\infty}^{\infty} b_i 2^i, \quad b_i \in \{0,1\} \text{ for all } i.$$

This is similar to the binary expansion of a natural number as discussed in Section 2.2.8. The only difference is that we have to allow all negative powers of 2, since x can be arbitrarily close to 0. The binary expansion of 1.25 for example is

$$1.25 = 1\cdot 2^{-2} + 1\cdot 2^0.$$

We then determine the smallest and largest values of i, $\underline{i}$ and $\bar{i}$, for which b_i is nonzero (note that $\underline{i}$ may be $-\infty$, but $\bar{i}$ is finite since x is finite). The number $\bar{i}-\underline{i}+1 \in \mathbb{N}\cup\{\infty\}$ is the number of *significant digits* of x.

With $d_i := b_{\bar{i}-i}$, we get $d_0 \neq 0$ and

$$x = \sum_{i=\underline{i}}^{\bar{i}} b_i 2^i = \sum_{i=0}^{\bar{i}-\underline{i}} b_{\bar{i}-i} 2^{\bar{i}-i} = \sum_{i=0}^{\bar{i}-\underline{i}} d_i 2^{-i}\cdot 2^{\bar{i}}.$$

This implies that $x \in \mathcal{F}^*(2, p, e_{\min}, e_{\max})$ if and only if $\bar{i}-\underline{i} < p$ and $e_{\min} \leq \bar{i} \leq e_{\max}$. Equivalently, if the binary expansion of x has at most p significant digits, and the exponent of the normalized representation is within the allowable range.

In computing the binary expansion of $x > 0$, let us assume for simplicity that $x < 2$. This is sufficient to explain the issue with the decimal numbers 1.1 and 0.1, and all other cases can be reduced to this case by separately dealing with the largest even integer smaller or equal to x: writing $x = y + 2k$ with $k \in \mathbb{N}$ and $y < 2$, we get the binary expansion of x by combining the expansions of y and $2k$.

For $x < 2$, we have

$$x = \sum_{i=-\infty}^{0} b_i 2^i = b_0 + \sum_{i=-\infty}^{-1} b_i 2^i = b_0 + \sum_{i=-\infty}^{0} b_{i-1} 2^{i-1} = b_0 + \frac{1}{2}\underbrace{\sum_{i=-\infty}^{0} b_{i-1} 2^i}_{=:x'}.$$

This identity provides a simple algorithm to compute the binary expansion of x. If $x \geq 1$, the most significant digit b_0 is 1, otherwise it is 0. The other digits b_i, $i \leq -1$, can subsequently be extracted by applying the same technique to $x' = 2(x - b_0)$.

Doing this for $x = 1.1$ yields the following sequence of digits.

$$
\begin{array}{rcrcrcl}
 & & & & 1.1 & \rightarrow & b_0 = 1 \\
2(1.1-1) & = & 2 \cdot 0.1 & = & 0.2 & \rightarrow & b_{-1} = 0 \\
2(0.2-0) & = & 2 \cdot 0.2 & = & 0.4 & \rightarrow & b_{-2} = 0 \\
2(0.4-0) & = & 2 \cdot 0.4 & = & 0.8 & \rightarrow & b_{-3} = 0 \\
2(0.8-0) & = & 2 \cdot 0.8 & = & 1.6 & \rightarrow & b_{-4} = 1 \\
2(1.6-1) & = & 2 \cdot 0.6 & = & 1.2 & \rightarrow & b_{-5} = 1 \\
2(1.2-1) & = & 2 \cdot 0.2 & = & 0.4 & \rightarrow & b_{-6} = 0 \\
 & & & & & \vdots &
\end{array}
$$

We now see that the binary expansion of the decimal number 1.1 is periodic: the corresponding binary number is $1.0\overline{0011}$, and it has infinitely many significant digits. Since all numbers in the floating point number systems $\mathcal{F}(2, p, e_{min}, e_{max})$ and $\mathcal{F}^*(2, p, e_{min}, e_{max})$ have at most p significant digits, it follows that $x = 1.1$ is not representable in a binary floating point number system, regardless of p, e_{min} and e_{max}. The same is true for $x = 0.1$.

The Excel 2007 bug. We have shown in the previous paragraph that it is impossible to convert some common decimal numbers (like 1.1 or 0.1) into binary floating-point numbers, without making small errors. This has the embarrassing consequence that the types `float` and `double` are unable to represent the values of some of their own literals.

Despite this problem, a huge number of decimal-to-binary conversions take place on computers worldwide, the minute you read this. For example, whenever you enter a number into a spreadsheet, you do it in decimal format. But chances are high that internally, the number is converted to and represented in binary floating-point format. The small errors themselves are usually not the problem; but the resulting "weird" floating-point numbers extremely close to some "nice" decimal value may expose other problems in the program.

A recent such issue that has received a lot of attention is known as the the *Excel 2007 bug*. Users have reported that the multiplication of 77.1 with 850 in Microsoft Excel does not yield 65,535 (the mathematically correct result) but 100,000.

Microsoft reacted to this by admitting the bug, but at the same time pointing out that the *computed value* is correct, and that the error only happens when this value is *displayed* in the sheet. But how can it happen that the nice integer value 65,535 is incorrectly displayed? Well, it doesn't happen: when you multiply 65,535 with 1, for example, the result is correctly displayed as 65,535.

The point is that the computed value is *not* 65,535, but some other number extremely close to it. The reason is that a small but unavoidable error is made in converting the

decimal value 77.1 into the floating-point number system internally used by Excel: like 1.1 and 0.1, the number 77.1 has no finite binary representation.

This error can of course not be "repaired" by the multiplication with 850, so Excel gets a value only very close to $65,535$. This would be acceptable, but exactly for *this* value (and 11 others, according to Microsoft), the display functionality has a bug. Naturally, if only 12 "weird" numbers out of all floating-point numbers are affected by this bug, it is easy not to detect the bug during regular tests.

While Microsoft earned quite some ridicule for the Excel 2007 bug (for which it quickly offered a fix), it should in all fairness be admitted that such bugs could still be hidden in software of other vendors as well.

Relative error. If we are not able to represent a real number x exactly as a binary floating point number in the system $\mathcal{F}^*(2, p, e_{\min}, e_{\max})$, it is natural to approximate it by the floating point number *nearest* to x. What is the error we make in this approximation?

Suppose that x is positive and has binary expansion

$$x = \sum_{i=-\infty}^{\bar{i}} b_i 2^i = b_{\bar{i}}.b_{\bar{i}-1} \ldots \cdot 2^{\bar{i}}, \quad \text{where } b_{\bar{i}} = 1.$$

There are two natural ways of approximating x with p or less significant digits. One way is to round down, resulting in the number

$$\underline{x} = b_{\bar{i}}.b_{\bar{i}-1} \ldots b_{\bar{i}-p+1} \cdot 2^{\bar{i}} = \sum_{i=\bar{i}-p+1}^{\bar{i}} b_i 2^i.$$

This truncates all the digits $b_i, i \leq \bar{i} - p$, and the error we make is

$$x - \underline{x} = \sum_{i=-\infty}^{\bar{i}-p} b_i 2^i \leq \sum_{i=-\infty}^{\bar{i}-p} 2^i = 2^{\bar{i}-p+1}.$$

Alternatively, we could round up to the number

$$\overline{x} = \underline{x} + 2^{\bar{i}-p+1}$$

where our previous error estimate shows that indeed, $x \leq \overline{x}$ holds. For this, we have to check that $\overline{x}$ has at most p significant digits. This is true if $b_{\bar{i}-p+1} = 0$, since then the addition of $2^{\bar{i}-p+1}$ adds exactly one digit to the at most $p-1$ significant digits of $\underline{x}$. And if $b_{\bar{i}-p+1} = 1$, the addition of $2^{\bar{i}-p+1}$ removes the least significant coefficient of $2^{\bar{i}-p+1}$ and *may* create one extra carry digit at the other end.

This means that x is between two numbers that are $2^{\bar{i}-p+1}$ apart, so the nearer of the two numbers is at most $2^{\bar{i}-p}$ away from x. On the other hand, x has size *at least* $2^{\bar{i}}$, meaning that

$$|x - \hat{x}|/x \leq 2^{-p},$$

where $\hat{x}$ is the floating point number nearest to x. The number 2^{-p}, referred to as the *machine epsilon*, is the *relative error* made in approximating x with its nearest floating point number $\hat{x}$.

The previous inequality also holds for negative x and their corresponding best approximation $\hat{x}$, so that we get the general *relative error* formula

$$\frac{|x - \hat{x}|}{|x|} \leq 2^{-p}, \quad x \neq 0.$$

This means that the distance of x to its nearest floating point number is in the worst case proportional to the size of x. This is because the floating point numbers are not equally spaced along the real line. Close to 0, their density is high, but the more we go away from 0, the sparser they become. As a simple example, consider the normalized floating point number system $F^*(2, 3, -2, 2)$. The smallest positive number is $1.00 \cdot 2^{-2} = 1/4$, and the largest one is $1.11 \cdot 2^2 = 7$ (recall that the digits are binary). The distribution of all positive numbers over the interval $[1/4, 7]$ is shown in the following picture.

From this picture, it is clear that the relative error formula cannot hold for very large x. But also if x is very close to zero, the relative error formula may fail. In fact, there is a substantial gap between 0 and the smallest positive normalized number. Numbers x in that gap are not necessarily approximable by normalized floating point numbers with relative error at most 2^{-p}.

Where is the mistake in our calculations, then? There is no mistake, but the calculations are only applicable if the floating point number $\hat{x}$ nearest to x *is* in fact a floating point number in the system we consider, i.e. if it has its exponent in the allowed range $\{e_{\min}, \ldots, e_{\max}\}$. This fails if $\hat{x}$ is too large or too small.

Arithmetic operations. Performing addition, subtraction, multiplication, and division with floating point numbers is easy in theory: as these are real numbers, we simply perform the arithmetic operations over the set $\mathbb{R}$ of real numbers; if the result is not representable in our floating point number system, we apply some rounding rule (such as choosing the nearest representable floating point number).

In practice, floating point number arithmetic is not more difficult than integer arithmetic. Let us illustrate this with an example. Suppose that $p = 4$, and that we have a binary system; we want to perform the addition

$$\begin{array}{rl} & 1.111 \cdot 2^{-2} \\ + & 1.011 \cdot 2^{-1} \ . \end{array}$$

The first step is to *align* the two numbers such that they have the same exponent. This means to "denormalize" one of the two numbers, e.g. the second one:

$$\begin{array}{rr} & 1.111 \cdot 2^{-2} \\ + & 10.110 \cdot 2^{-2} \ . \end{array}$$

Now we can simply add up the two significands, just like we add integers in binary representation. The result is

$$100.101 \cdot 2^{-2}.$$

Finally, we renormalize and obtain

$$1.00101 \cdot 2^{0}.$$

We now realize that this exact result is not representable with $p = 4$ significant digits, so we have to round. In this case, the nearest representable number is obtained by simply dropping the last two digits:

$$1.001 \cdot 2^{0}.$$

2.5.6 The IEEE standard 754

Value range. The C++ standard does not prescribe the value range of the types `float` and `double`. It only stipulates that the value range of `float` is contained in the value range of `double` such that a `float` value can be promoted to a `double` value.

In practice, most platforms support (variants of) the *IEEE standard 754* for representing and computing with floating point numbers. Under this standard, the value range of the type `float` is the set

$$F^*(2, 24, -126, 127)$$

of *single precision* normalized floating point numbers, plus some special numbers (conveniently, 0 is one of these special numbers). The value range of `double` is the set

$$F^*(2, 53, -1022, 1023)$$

of *double precision* normalized floating point numbers, again with some special numbers added, including 0.

These parameters may seem somewhat arbitrary at first, but they are motivated by a common memory layout in which 32 bits can be manipulated at once. Indeed, 32 bits of memory are used to represent a single precision number. The significand requires 23 bits; recall that in a normalized binary floating point number system, the first digit of the significand is always 1, hence it need not explicitly be stored. The exponent requires another 8 bits for representing its $254 = 2^8 - 2$ possible values, and another bit is needed for the sign.

For double precision numbers, the significand requires 52 bits, the exponent has 11 bits for its $2046 = 2^{11} - 2$ possible values, and one bit is needed for the sign. In total, this gives 64 bits.

Note that in both cases, two more exponent values could be accommodated without increasing the total number of bits. These extra values are in fact used for representing the special numbers mentioned above, including 0.

Requirements for the arithmetic operations. The C++ standard does not prescribe the accuracy of arithmetic operations over the types `float` and `double`, but the IEEE standard 754 does. The requirements are as strict as possible: the result of any addition, subtraction, multiplication, or division is the representable value *nearest* to the true value. If there are two nearest values (meaning that the true value is halfway in between them), the one that has least significant digit $d_{p-1} = 0$ is chosen. This is called *round-to-even*; other rounding modes can be enabled if necessary. The same rule applies to the conversion of decimal values like 1.1 to their binary floating point representation.

Moreover, comparisons of values have to be exact under all relational operators (Section 2.3.2).

2.5.7 Computing with floating point numbers

We have seen that for every floating point number system, there are numbers that it cannot represent, and these are not necessarily very exotic, as our example with the decimal number 1.1 shows. On the other hand, the IEEE standard 754 guarantees that we will get the nearest representable number, and the same holds for the result of every arithmetic operation, up to (rare) over- and underflows. Given this, one might be tempted to believe that the results of most computations involving floating point numbers are close to the mathematically correct results, with respect to relative error.

Indeed, this is true in many cases. For example, our initial program `euler.cpp` computes a pretty good approximation of the Euler constant. Nevertheless, some care has to be taken in general. The goal of this section is to point out common pitfalls, and to provide resulting guidelines for "safe" computations with floating point numbers.

We start with the first and most important guideline that may already be obvious to you at this point.

Floating Point Arithmetic Guideline 1: Never compare two floating point numbers for equality, if at least one of them results from inexact floating point computations.

Even very simple expressions involving floating point numbers may be mathematically equivalent, but still return different values, since intermediate results are rounded. Two such expressions are `x * x - y * y` and `(x + y) * (x - y)`. Therefore, testing the results of two floating point computations for equality using the relational operators `==` or `!=` makes little sense. Since equality is sensitive to the tiniest errors, we won't get equality in most cases, even if mathematically, we would.

Given the formulation of the above guideline, you may wonder how to tell whether a particular floating point computation is exact or not. Exactness usually depends on the representation and is, therefore, hard to claim in general. However, there are certain operations which are easily seen to be exact. For instance, multiplication and division by a power of the base (usually, 2) do not change the significand, but only the exponent. Thus, these operations are exact, unless they lead to an over- or underflow in the exponent.

Moreover, it is safe to assume that the largest exponent is (much) higher than the precision p, and in this case we can also exactly represent all integers of absolute value smaller than β^p. Consequently, integer additions, subtractions, and multiplications within this range are exact.

The next two guidelines are somewhat less obvious, and we motivate them by first showing the underlying problem. Throughout, we assume a binary floating point number system of precision p.

Adding numbers of different sizes. Suppose we want to add the two floating point numbers 2^p and 1. What will be the result? Mathematically, it is

$$2^p + 1 = \sum_{i=0}^{p} b_i 2^i,$$

with $(b_p, b_{p-1}, \ldots, b_0) = (1, 0, \ldots, 0, 1)$. Since this binary expansion has $p+1$ significant digits, $2^p + 1$ is not representable with precision p. Under the IEEE standard 754, the result of the addition is 2^p (chosen from the two nearest candidates 2^p and $2^p + 2$), so this addition has no effect.

The general phenomenon here is that adding floating point numbers of different sizes "kills" the less significant digits of the smaller number (in our example, *all* its digits). The larger the size difference, the more drastic is the effect.

To convince you that this is not an artificial phenomenon, let us consider the problem of computing Harmonic numbers (search the web for the *coupon collector's problem* to find an interesting occurrence of Harmonic numbers). For $n \in \mathbb{N}$, the n-th *Harmonic number* H_n is defined as the sum of the reciprocals of the first n natural numbers, that is,

$$H_n = \sum_{i=1}^{n} \frac{1}{i}.$$

It should now be an easy exercise for you to write a program that computes H_n for a given $n \in \mathbb{N}$. You only need a single loop running through the numbers 1 up to n, adding their reciprocals. Just as well you can make your loop run from n down to 1 and sum up the reciprocals. Why not, that should not make any difference, right? Let us try both variants and see what we get. The program `harmonic.cpp` shown below computes the two sums and outputs them.

```
1  // Program: harmonic.cpp
2  // Compute the n-th harmonic number in two ways.
3
4  #include <iostream>
5
6  int main()
7  {
8    // Input
9    std::cout << "Compute H_n for n =? ";
10   unsigned int n;
11   std::cin >> n;
12
13   // Forward sum
14   float fs = 0;
15   for (unsigned int i = 1; i <= n; ++i)
16     fs += 1.0f / i;
17
18   // Backward sum
19   float bs = 0;
20   for (unsigned int i = n; i >= 1; --i)
21     bs += 1.0f / i;
22
23   // Output
24   std::cout << "Forward sum  = " << fs << "\n"
25             << "Backward sum = " << bs << "\n";
26   return 0;
27 }
```

Program 2.13: *../progs/lecture/harmonic.cpp*

Think for a second and recall why it is important to *not* write `1 / i` in line 16 and line 21. Now let us have a look at an execution of the program. On the platform of the authors, the following happens.

```
Compute H_n for n =? 10000000
Forward sum  = 15.4037
Backward sum = 16.686
```

The results differ significantly. The difference becomes even more apparent when we try larger inputs.

```
Compute H_n for n =? 100000000
Forward sum  = 15.4037
Backward sum = 18.8079
```

Notice that the forward sum did not change, which cannot be correct. Using the

approximation

$$\frac{1}{2(n+1)} < H_n - \ln n - \gamma < \frac{1}{2n},$$

where $\gamma = 0.57721666\ldots$ is the Euler-Mascheroni constant, we get $H_n \approx 18.998$ for $n = 10^8$. That is, the backward sum provides a much better approximation of H_n.

Why does the forward sum behave so badly? The reason is simple: As the larger summands are added up first, the intermediate value of the sum to be computed grows (comparatively) fast. At some point, the size difference between the partial sum and the summand $\frac{1}{i}$ to be added is so large that the addition does not change the partial sum anymore, just like in $2^p + 1 \text{ }''=''\text{ } 2^p$. Thus, regardless of how many more summands are added to it, the sum stays the same.

In contrast, the backward sum starts to add up the small summands first. Therefore, the value of the partial sum grows (comparatively) slowly, allowing the small summands to contribute. The summands treated in the end of the summation have still a good chance to influence the significand of the partial sum, since they are (comparatively) large.

The phenomenon just observed leads us to our second guideline.

Floating Point Arithmetic Guideline 2: Avoid adding two numbers that considerably differ in size.

Cancellation. Consider the quadratic equation

$$ax^2 + bx + c = 0, \quad a \neq 0.$$

It is well known that its two roots are given by

$$r_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

In a program that computes these roots, we might therefore want to compute the value $d = b^2 - 4ac$ of the *discriminant*. If b^2 and $4ac$ are representable as floating point numbers with precision p, our previous error estimates guarantee that the result $\hat{d}$ of the final subtraction has small relative error: $|d - \hat{d}| \leq 2^{-p}|d|$. This means, even if d is close to zero, $\hat{d}$ will be away from d by *much less* than the distance of d to zero.

The problem arises if the numbers b^2 and/or $4ac$ are not representable as floating point numbers, in which case errors are made in computing them. Assume $b = 2^p, a = 2^{p-1} - 1, c = 2^{p-1} + 1$ (all these numbers are exactly representable). Then the exact value of d is 4. The value $b^2 = 2^{2p}$ is a representable floating point number, but $4ac = 2^{2p} - 4$ is not, since this number has $2p - 2$ significant digits (all of them equal to 1) in its binary expansion. The nearest floating point number is obtained by rounding up (adding 4), and

after the (error-free) subtraction, we get $\hat{d} = 0$. The relative error of this computation is therefore 1 instead of 2^{-p}.

The reason is that in subtracting two numbers that are almost equal, the more significant digits cancel each other. If, on the other hand, the remaining less significant digits already carry some errors from previous computations, the subtraction hugely amplifies these errors: the cancellation promotes the previously less significant digits to much more significant digits of the result.

Again, the example we gave here is artificial, but be assured that cancellation happens in practice. Even in the quadratic equation example, it might be that the equations that come up in an application have the property that their discriminant $b^2 - 4ac$ is much smaller than a, b and c themselves. In this case, cancellation *will* happen.

The discussion can be summarized in form of a third guideline.

Floating Point Arithmetic Guideline 3: Avoid subtracting two numbers of almost equal size, if these numbers are results of other floating point computations.

2.5.8 Details

Other floating point number systems. The IEEE standard 754 defines two more floating point number systems, *single-extended precision* ($p = 32$), and *double-extended precision* ($p = 64$), and some platforms offer implementations of these types. There is also the IEEE standard 854 that allows base $\beta = 10$, for obvious reasons: the decimal format is the one in which we think about numbers, and in which we usually represent numbers. In particular, a base-10 system has no holes in the value range at decimal fractional numbers like 1.1 and 0.1.

IEEE compliance. While on most platforms, the types `float` and `double` correspond to the single and double precision floating point numbers of the IEEE standard 754, this correspondence is usually not one-to-one. For example, if you are trying to reproduce the cancellation example we gave, you might write

```
const float b = 16777216.0f;  // 2^24
const float a =  8388607.0f;  // 2^23 - 1
const float c =  8388609.0f;  // 2^23 + 1

std::cout << b * b - 4.0f * a * c << "\n";
```

and expect to get the predicted wrong result 0. But it may easily happen that you get the correct result 4, even though your platform claims to follow the IEEE standard 754. The most likely reason is that the platform internally uses a register with more bits to perform the computation. While this seems like a good idea in general, it can be fatal for a program whose functionality critically relies on the IEEE standard 754.

You *will* most likely see the cancellation effect in the following seemingly equivalent variant of the above code.

```
float b = 16777216.0f;  // 2^24
float a =  8388607.0f;  // 2^23 - 1
float c =  8388609.0f;  // 2^23 + 1

float bb = b * b;
float ac4 = 4.0f * a * c;

std::cout << bb - ac4 << "\n";
```

Here, the results of the intermediate computations are written back to `float` variables, probably resulting in the expected rounding of $4ac$. Then the final subtraction reveals the cancellation effect. *Unless*, of course, the compiler decides to keep the variable `ac4` in a register with more precision. For this reason, you can typically provide a compiler option to make sure that floating point numbers are not kept in registers.

What is the morale of this? You usually cannot fully trust the *IEEE compliance* of a platform, and it is neither easy nor worthwhile to predict how floating point numbers exactly behave on a specific platform. It is more important for you to know and understand floating point number systems in general, along with their limitations. This knowledge will allow you to identify and work around problems that might come up on specific platforms.

The type long double. The C++ standard prescribes another fundamental floating point number type called `long double`. Its literals end with the letter `l` or `L`, and it is guaranteed that the value range of `double` is contained in the value range of `long double`. Despite this, the conversion from `double` to `long double` is not defined to be a promotion by the C++ standard.

While `float` and `double` usually correspond to single and double precision of the IEEE standard 754, there is no such default choice for `long double`. In practice, `long double` might simply be a synonym for `double`, but it might also be something else. On the platform used by the authors, for example, `long double` corresponds to the normalized floating point number system $F^*(2, 64, -16382, 16384)$— this is exactly the double-extended precision of the IEEE standard 754.

Numeric limits. If you want to know the parameters of the floating point number systems behind `float`, `double` and `long double` on your platform, you can employ the `numeric_limits` we have used before in the program `limits.cpp` in Section 2.2.5. Here are the relevant expressions together with their meanings, shown for the type `float`.

expression (of type `int`)	meaning
`std::numeric_limits<float>::radix`	β
`std::numeric_limits<float>::digits`	p
`std::numeric_limits<float>::min_exponent`	$e_{\min} + 1$
`std::numeric_limits<float>::max_exponent`	$e_{\max} + 1$

We remark that `std::numeric_limits<float>::min()` does *not* give the smallest `float` value (because of the sign bit, this smallest value is simply the negative of the largest value), but the smallest normalized *positive* value.

Special numbers. We have mentioned that the floating point systems prescribed by the IEEE standard 754 contain some special numbers; their encoding uses exponent values that do not occur in normalized numbers.

On the one hand, there are the *denormalized* numbers of the form

$$\pm d_0.d_1 \ldots d_{p-1} \cdot \beta^{e_{min}},$$

with $d_0 = 0$. A denormalized number has smaller absolute value than any normalized number. In particular, 0 is a denormalized number.

The other special numbers cannot really be called numbers. There are values representing $+\infty$ and $-\infty$, and they are returned by overflowing operations. Then there are several values called `NaN`s (for "not a number") that are returned by operations with undefined result, like taking the square root of a negative number. The idea behind these values is to provide more flexibility in dealing with exceptional situations. Instead of simply aborting the program when some operation fails, it makes sense to return an exceptional value. The caller of the operation can then decide how to deal with the situation.

2.5.9 Goals

Dispositional. At this point, you should ...

1) know the floating point number types `float` and `double`, and that they are more general than the integral types;
2) understand the concept of a floating point number system, and in particular its advantages over a fixed point number system;
3) know that the IEEE standard 754 describes specific floating point number systems used as models for `float` and `double` on many platforms;
4) know the three Floating Point Arithmetic Guidelines;
5) be aware that computations involving the types `float` and `double` may deliver inexact results, mostly due to holes in the value range.

Operational. In particular, you should be able to ...

(G1) evaluate expressions involving the arithmetic types `int`, `unsigned int`, `float` and `double`;

(G2) compute the binary representation of a given real number;

(G3) compute the floating point number nearest to a given real number, with respect to a finite floating point number system;

(G4) work with a given floating point number system;

(G5) recognize usage of floating point numbers that violates any of the three Floating Point Arithmetic Guidelines;

(G6) write programs that perform computations with floating point numbers.

2.5.10 Exercises

Exercise 56 *For every expression in the following list, determine its type, its value, and whether it is an rvalue or an lvalue. In each of the expressions, the variable* `x` *is of type* `int` *and has value* `1`. (G1)

a) `1 + true == 2`

b) `3 % 2 + 1 * 4`

c) `x = 10 / 2 / 5 / 2`

d) `x / 2.0`

e) `1 + x++`

f) `++x`

Exercise 57 *For every expression in the following list, determine its type and its value. We assume a floating point representation according to IEEE 754, that is,* `float` *corresponds to* $F^*(2, 24, -126, 127)$ *and* `double` *to* $F^*(2, 53, -1022, 1023)$. *We also assume that 32 bits are used to represent* `int` *values.* (G1)

a) `2e2-3e3f>-23.0`

b) `-7+7.5`

c) `1.0f/2+1/3+1/4`

d) `true||false&&false`

e) `1u-2u<0`

f) `1+2*3+4`

g) `int(8.5)-int(7.6)`

h) `100*1.1==110`

i) `10*11.0==110`

j) `4+12/4.0/2`

Exercise 58 *Evaluate the following expressions step-by-step, according to the conversion rules of mixed expressions. We assume a floating point representation according to IEEE 754, that is,* `float` *corresponds to* $F^*(2, 24, -126, 127)$ *and* `double` *corresponds to* $F^*(2, 53, -1022, 1023)$. *We also assume that* 32 *bits are used to represent* `int` *values.* (G1)

a) `6 / 4 * 2.0f - 3`

b) `2 + 15.0e7f - 3 / 2.0 * 1.0e8`

c) `392593 * 2735.0f - 8192 * 131072 + 1.0`

d) `16 * (0.2f + 262144 - 262144.0)`

Exercise 59 *Compute the binary expansions of the following decimal numbers.*
a) 0.25 *b)* 1.52 *c)* 1.3 *d)* 11.1 (G2)

Exercise 60 *For the numbers in Exercise 59, compute nearest floating point numbers in the systems* $\mathcal{F}^*(2, 5, -1, 2)$ *and* $\mathcal{F}(2, 5, -1, 2)$. (G3)

Exercise 61 *What are the largest and smallest* positive *normalized single and double precision floating point numbers, according to the IEEE standard 754?* (G4)

Exercise 62 *How many floating point numbers do the systems* $\mathcal{F}^*(\beta, p, e_{min}, e_{max})$ *and* $\mathcal{F}(\beta, p, e_{min}, e_{max})$ *contain?* (G4)

Exercise 63 *Compute the value of the variable* `d` *after the declaration statement*

```
float d = 0.1;
```

Assume the IEEE standard 754. (G3)

Exercise 64 *What is the problem with the following loop (assuming the IEEE standard 754)?* (G5)

```
for (float i = 0.1f; i != 1.0f; i += 0.1f)
  std::cout << i << "\n";
```

Exercise 65 *What is the problem with the following loop (assuming the IEEE standard 754)?* (G5)

```
for (float i = 0.0f; i < 100000000.0f; ++i)
   std::cout << i << "\n";
```

Exercise 66 *Write a program that outputs for a given decimal input number* x, $0 < x < 2$, *its normalized* `float` *value on your platform. The output should contain the (binary) digits of the significand, starting with* 1, *and the (decimal) exponent. You may assume that the floating point number system underlying the type* `float` *has base* $\beta = 2$. (G3)(G6)

Exercise 67 *Write a program that tests whether a given value of type* `double` *is actually an integer, and test the program with various inputs like* `0.5`, `1`, `1234567890`, `1234567890.2`. *Simply converting to a value of type* `int` *and checking whether this changes the value does not work in general, since the given value might be an integer outside the value range of* `int`. *You may assume that the floating point number system underlying the type* `double` *has base* $\beta = 2$. (G3)(G6)

Exercise 68 *The number* π *can be defined through various infinite sums. Here are two of them.*

$$\begin{aligned}\frac{\pi}{4} &= 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots \\ \frac{\pi}{2} &= 1 + \frac{1}{3} + \frac{1 \cdot 2}{3 \cdot 5} + \frac{1 \cdot 2 \cdot 3}{3 \cdot 5 \cdot 7} + \cdots\end{aligned}$$

Write a program for computing an approximation of π, *based on these formulas. Which formula is better for that purpose?* (G6)

Exercise 69 *The function* $\sin(x)$ *can be defined through the following power series.*

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots .$$

a) *Implement and test—based on the first* 20 *terms of this power series—a function*

```
// POST: returns an approximation of sin(x)
double sinus (double x);
```

b) *What happens if the argument value is very large,* $x = 10,000$, *say? Does your function still yield reasonable results, then? If not, do you see a way to fix this?* (G6)

Exercise 70 *There is a well-known iterative procedure (the* Babylonian method*) for computing the square root of a positive real number* s. *Starting from any value* $x_0 > 0$, *we compute a sequence* $x_0, x_1, x_2, \ldots$ *of values according to the formula*

$$x_n = \frac{1}{2}(x_{n-1} + \frac{s}{x_{n-1}}).$$

It can be shown that

$$\lim_{n\to\infty} x_n = \sqrt{s}.$$

Write a program `babylonian.cpp` *that reads in the number* s *and computes an approximation of* $\sqrt{s}$ *using the Babylonian method. To be concrete, the program should output the first number* x_i *such that* (G6)

$$|x_i^2 - s| < 0.001.$$

Exercise 71 *Write a program* `fpsys.cpp` *to visualize a normalized floating point number system* $\mathcal{F}^*(2, p, e_{min}, e_{max})$. *The program should read the parameters* p, e_{min}, *and* e_{max} *as inputs and for each positive number* x *from* $\mathcal{F}^*(2, p, e_{min}, e_{max})$ *draw a circle of radius* x *around the origin. Use the library* `libwindow` *that is available at the course homepage to create graphical output. Use the program to verify the numbers you computed in Exercise 62.* (G4)(G6)

Exercise 72 *We have seen that the decimal number* 0.1 *has no finite representation in a binary floating-point number system* ($\beta = 2$). *Mr. X. M. Plestudent claims that this is due to* $\beta < 10$. *He suggests to work with a hexadecimal system* ($\beta = 16$) *and argues that in such a system,* 0.1 does *have a finite representation. Is Mr. Plestudent right or not?*

Somewhat more formally, is there a natural number p *and numbers* $d_1, \ldots, d_p$ *with* $d_i \in \{0, 1, \ldots, 15\}$ *for all* i, *such that*

$$\frac{1}{10} = \sum_{i=1}^{p} d_i 16^{-i} = \text{“}0.d_1 \ldots d_p\text{”}$$

holds? (G4)

Exercise 73 *We have seen that there are decimal numbers without a finite binary representation (such as* 1.1 *and* 0.1*). Conversely, every (fractional) binary number* does *have a finite decimal representation, a fact that may be somewhat surprising at first sight. Prove this fact!*

More formally, given a number b *of the form*

$$b = \sum_{i=1}^{k} b_i 2^{-i}, \quad b_1, b_2, \ldots, b_k \in \{0, 1\},$$

prove that there is a natural number ℓ *such that* b *can be written as an* ℓ*-digit decimal number*

$$b = \sum_{i=1}^{\ell} d_i 10^{-i}, \quad d_1, d_2, \ldots, d_\ell \in \{0, 1, \ldots, 9\}.$$

(G4)

2.5.11 Challenges

Exercise 74 *We have seen that decimal numbers do not necessarily have finite binary representations (examples are the decimal numbers* 1.1 *and* 0.1*). Vice versa, binary numbers* do *have finite decimal representations (Exercise 73). And Exercise 72 asks whether every decimal number has a finite* hexadecimal *representation. The goal of this challenge is to understand the general picture.*

Let $\beta \geq 2$ *and* $\gamma \geq 2$ *be two natural numbers We say that* γ refines β *if for every* p, e_{min}, e_{max}*, there are* q, f_{min}, f_{max} *such that*

$$\mathcal{F}(\beta, p, e_{min}, e_{max}) \subseteq \mathcal{F}(\gamma, q, f_{min}, f_{max}).$$

In other words, every floating-point number system to the base β *is contained in some floating-point number system to the base* γ*.*

In this language, the result of Section 2.5.5 is that 2 *does not refine* 10*. Exercise 73 implies that* 10 *refines* 2*, and Exercise 72 asks whether* 16 *refines* 10*.*

Here is the challenge: characterize the pairs (β, γ) *for which* γ *refines* β*!*

Exercise 75 *The* Mandelbrot set *is a subset of the complex plane that became popular through its* fractal *shape and the beautiful drawings of it. Below you see the set's* main cardioid *and a detail of it at much higher zoom scale.*

The Mandelbrot set is defined as follows. For $c \in \mathbb{C}$*, we consider the sequence* $z_0(c), z_1(c), \ldots$ *of complex numbers given by* $z_0(c) = 0$ *and*

$$z_n(c) = z_{n-1}(c)^2 + c, \quad n > 0.$$

There are two cases: either $|z_n(c)| \leq 2$ *for all* n *(this obviously happens for example if* $c = 0$*), or* $|z_n(c)| > 2$ *for some* n *(this obviously happens for example if* $|c| > 2$*). The Mandelbrot set consists of all* c *for which we are in the first case. It follows that the Mandelbrot set contains* 0 *and is contained in a disk of radius* 2 *around* 0 *in the complex plane.*

Write a program that draws (an approximation of) the Mandelbrot set, restricted to a rectangular subset of the complex plane. It should be possible to zoom in, meaning that the rectangular subset becomes smaller, and more details become visible in

the drawing window. Obviously, you can't process all infinitely many complex numbers c *in the rectangle, and for given* c*, you cannot really check whether* $|z_n(c)| \leq 2$ *for* all n*, so it is necessary to discretize the rectangle into pixels, and to establish some upper bound* N *on the number of iterations. If* $|z_n(c)| \leq 2$ *for all* $n \leq N$*, you may simply assume that* c *is in the Mandelbrot set. Per se, there is no guarantee that the resulting drawing is even close to the Mandelbrot set (especially at finer level of detail), but for the sake of obtaining nice pictures, we can generously gloss over this issue.*

Hint: *You may use the* `libwindow` *library to produce the drawing. The example program in its documentation should give you an idea how this can be done.*

Exercise 76 *The following email was sent to a mailing list for users of the software library CGAL.*

```
Hi all,

This should be a very easy question.

When I check if the points (0.14, 0.22), (0.15, 0.21) and (0.19,0.17) are
collinear, using CGAL::orientation, it returns CGAL::LEFT_TURN, which is
false, because those points are in fact collinear.

However, if I do the same with the points (14, 22), (15, 21) and (19, 17) I
get the correct answer: CGAL::COLLINEAR.
```

a) *Find out what this email is about; in particular, what is CGAL, what is the orientation of a point triple, what is* `CGAL::orientation`*, what does "collinear" mean, and why is the writer of the email surprised about the observed behavior?*

b) *Draft an answer to this email that explains the observations of the CGAL user that wrote it.*

2.6 A first C++ function

Garbage in, garbage out.

Attributed to George Fuechsel, IBM, late 1950's

This section introduces C++ functions as a means to encapsulate and reuse functionality, and to subdivide a program into subtasks. You will learn how to add functions to your programs, and how to call them. We also explain how functions can efficiently be made available for many programs at the same time, through separate compilation and libraries.

In many numerical calculations, computing powers is a fundamental operation (see Section 2.5), and there are many other operations that occur frequently in applications. In C++, *functions* are used to encapsulate such frequently used operations, making it easy to invoke them many times, with different arguments, and from different programs, but *without* having to reprogram them every time.

Even more importantly, functions are used to structure a program. In practice, large programs consist of many small functions, each of which serves a clearly defined subtask. This makes it a lot easier to read, understand, and maintain the program.

We have already seen quite a number of functions, since the `main` function of every C++ program is a special function (Section 2.1.4).

Program 2.14 emphasizes the encapsulation aspect and shows how functions can be used. It first defines a function for computing the value b^e for a given real number b and given integer e (possibly negative). It then calls this function for several values of b and e. The computations are performed over the floating point number type `double`.

```
// Prog: callpow.cpp
// Define and call a function for computing powers.

#include <iostream>

// PRE:  e >= 0 || b != 0.0
// POST: return value is b^e
double pow (double b, int e)
{
  double result = 1.0;
  if (e < 0) {
    // b^e = (1/b)^(-e)
    b = 1.0/b;
    e = -e;
  }
  for (int i = 0; i < e; ++i) result *= b;
```

```
17    return result;
18  }
19
20  int main()
21  {
22    std::cout << pow( 2.0, -2) << "\n"; // outputs 0.25
23    std::cout << pow( 1.5,  2) << "\n"; // outputs 2.25
24    std::cout << pow( 5.0,  1) << "\n"; // outputs 5
25    std::cout << pow( 3.0,  4) << "\n"; // outputs 81
26    std::cout << pow(-2.0,  9) << "\n"; // outputs -512
27
28    return 0;
29  }
```

Program 2.14: *../progs/lecture/callpow.cpp*

Before we explain the concepts necessary to understand this program in detail, let us get an overview of what is going on in the function pow. For nonnegative exponents e, b^e is obtained from the initial value of 1 by e-fold multiplication with b. This is what the for-loop does. The case of negative e can be handled by the formula $b^e = (1/b)^{-e}$: after inverting b and negating e in the if-statement, we have an equivalent problem with a positive exponent. The latter only works if $b \neq 0$, and indeed, negative powers of 0 are mathematically undefined.

2.6.1 Pre- and postconditions

Even a very simple function should document its *precondition* and its *postcondition*, in the form of comments. The precondition specifies what has to hold when the function is called, and the postcondition describes value and effect of the function. This information allows us to understand the function without looking at the actual sourcecode; this in turn is a necessary for keeping track of larger programs. In case of the function pow, the precondition

```
// PRE:   e >= 0 || b != 0.0
```

tells us that e must be nonnegative, or (if e is negative) that $b \neq 0$ must hold. The postcondition

```
// POST: return value is b^e
```

tells us the function value, depending on the arguments. In this case, there is no effect.

The pre- and postconditions specify the function in a mathematical sense. At first sight, functions with values *and* effect do not fit into the framework of mathematical functions which only have values. But using the concept of *program states* (Section 1.2.3), a C++ function can be considered as a mathematical function that maps program states (immediately *before* the function call) to program states (immediately *after* the function call).

Under this point of view, the precondition specifies the *domain* of the function, the set of program states in which the function may be called. In case of `pow`, these are all program states in which the arguments `b` and `e` are in a suitable relation. The postcondition describes the function itself by specifying how the (relevant part of the) program state gets transformed. In case of `pow`, the return value b^e will (temporarily) be put at some memory location.

To summarize, the postcondition tells us what happens when the precondition is satisfied. On the other hand, the postcondition gives *no guarantee whatsoever* for the case where the precondition is not satisfied. From a mathematical point of view, this is fine: a function is simply not defined for arguments outside its domain.

Arithmetic pre- and postconditions. The careful reader of Section 2.5 might have realized that both pre- and postcondition of the function `pow` cannot be correct. If e is too large, for example, the computation might overflow, but such e are not excluded by the precondition. Even if there is no overflow, the value range of the type `double` may have a hole at b^e, meaning that this value cannot be returned by the function. The postcondition is therefore imprecise as well.

In the context of arithmetic operations over the fundamental C++ types, it is often tedious and even undesirable to write down precise pre- and postconditions; part of the problem is that fundamental types may behave differently on different platforms. Therefore, we often confine ourselves to pre- and postconditions that document our *mathematical* intention, but we have to keep in mind that in reality, the function might behave differently.

Assertions. So far, our preconditions are just comments like in

```
// PRE:   e >= 0 || b != 0.0
```

Therefore, if the function `pow` is called with arguments `b` and `e` that violate the precondition, this passes unnoticed. On the syntactical level, there is nothing we can do about it: the function call `pow (0.0, -1)`, for example, will compile. But we can make sure that this blunder is detected at runtime. A simple way to do this uses *assertions*. An assertion has the form

assert (*expr*)

where *expr* is a predicate, an expression of a type whose values can be converted to `bool`. No comma is allowed in *expr*, a consequence of the fact that `assert` is not a function but a *macro*. A macro is a piece of meta-code that the compiler replaces with actual C++ code prior to compilation.

With assertions, `pow` can be written as follows.

```
// PRE:   e >= 0 || b != 0.0
// POST: return value is b^e
```

```
double pow (double b, int e)
{
  assert (e >= 0 || b != 0.0);
  double result = 1.0;
  // the remainder is as before
  ...
}
```

The purpose of an assertion is to check whether a certain predicate holds at a certain point. The precise semantics of an assertion is as follows. *expr* is evaluated, and if it returns *false*, execution of the program terminates immediately with an error message telling us that the respective assertion was violated. If *expr* returns *true*, execution continues normally. In our case, this means that the evaluation of the expression `pow (0.0,-1)` leads to a *runtime error*. This might not be a very polite way of telling the user that the arguments were illegal but the point will surely come across.

You can argue that it is costly to test the assertion in every function call, just to catch a few "bad" calls. However, it is possible to tell the compiler to ignore the `assert` macro, meaning that an empty piece of C++ code replaces it. The usual way to go is therefore as follows: during code development, put assertions everywhere you want to be sure that something holds. When the code is put into use, tell the compiler to remove the assertions. The machine language code is then as efficient as if you would never have written the assertions in the first place.

To use the `assert` macro, we have to include the header `cassert`.

2.6.2 Function definitions

Lines 8–18 of Program 2.14 define a function called `pow`. The syntax of a function definition is as follows.

T fname (*T1 pname1*, *T2 pname2*, ..., *Tk pnamek*)
  block

This defines a function called *fname*, with *return type T*, and with *formal arguments pname1*,..., *pnamek* of types *T1*,..., *Tk*, respectively, and with a *function body block*.

Syntactically, *T* and *T1*,..., *Tk* are type names, *fname* as well as *pname1*,..., *pnamek* are identifiers (Section 2.1.10), and *block* is a block, a sequence of statements enclosed in curly braces (Section 2.4.3).

We can think of the formal arguments as placeholders for the actual arguments that are supplied (or "passed") during a function call.

Function definitions must not appear inside blocks, other functions, or control statements. They may appear inside namespaces, though, or at global scope, like in `callpow.cpp`. A program may contain an arbitrary number of function definitions, appearing one af-

ter another without any delimiters between them. In fact, the program `callpow.cpp` consists of *two* function definitions, since the `main` function is a function as well.

2.6.3 Function calls

In Program 2.14, `pow(2.0,-2)` is one of five function calls. Formally, a function call is an expression. The syntax of a function call that matches the general function definition from above is as follows.

fname (*expr1*, ..., *exprk*)

Here, *expr1*,...,*exprk* must be expressions of types whose values can be converted to the formal argument types *T1*,...,*Tk*. These expressions are the *call arguments*. For all types that we know so far, the call arguments as well as the function call itself are rvalues. The type of the function call is the function's return type *T*.

When a function call is evaluated, the call arguments are evaluated first (in an order that is unspecified by the C++ standard). The resulting values are then used to initialize the formal arguments. Finally, the function body is executed; in this execution, the formal arguments behave like they were variables defined in the beginning of *block*, initialized with the values of the call arguments.

In particular, if the formal arguments are of const-qualified type, they cannot be modified within the function body. For example, the following version of the function `pow` will not compile.

```
double pow (const double b, const int e)
{
  double result = 1.0;
  if (e < 0) {
    // b^e = (1/b)^(-e)
    b = 1.0/b;            // error: b is constant
    e = -e;               // error: e is constant
  }
  for (int i = 0; i < e; ++i) result *= b;
  return result;
}
```

Const-correctness extends to formal function arguments: whenever the formal argument is *meant* to be kept constant within the function body, this should be documented by giving it a const-qualified type. From what we have said above about how function calls are evaluated, it follows that the "outside behavior" of the function (as documented by the pre- and postconditions) does not depend on whether the argument type is `const int` or `int` (see also Section 2.6.5 below). In this case, the effect of `const` is purely internal and determines whether the formal argument in question acts as a variable or as a constant within the function body.

For a fundamental type T, there is no difference between return type T and return type `const` T.

The evaluation of a function call terminates as soon as a return statement is reached, see Section 2.1.15. This return statement must be of the form

`return` *expr*;

where *expr* is an expression of a type whose values can be converted to the return type T. The resulting value is the value of the function call. The effect of the function call is determined by the joint effects of the call argument evaluations, and of executing *block*.

The function body may contain several return statements, but if no return statement is reached during the execution of *block*, value and effect of the function call are undefined (unless the return type is `void`, see Section 2.6.4 below).

For example, during the execution of *block* in `pow(2.0,-2)`, `b` and `e` initially have values 2 and -2. These values are changed in the `if`-statement to 0.5 and 2, before the subsequent loop sets `result` to 0.5 in its first and to 0.25 in its second (and last) iteration. This value is returned and becomes the value of the function call expression `pow(2.0,-2)`.

2.6.4 The type void

In C++, there is a fundamental type called `void`, used as return type for functions that only have an effect, but no value. Such functions are also called *void functions*.

As an example, consider the following program (note that the function `print_pair` requires no precondition, since it works for every combination of `int` values).

```
1  #include <iostream>
2
3  // POST: "(i, j)" has been written to standard output
4  void print_pair (const int i, const int j)
5  {
6    std::cout << "(" << i << ", " << j << ")\n";
7  }
8
9  int main()
10 {
11   print_pair(3,4); // outputs  (3, 4)
12 }
```

The type `void` has empty value range, and there are no literals, variables, or formal function arguments of type `void`. There are expressions of type `void`, though, for example `print_pair(3,4)`.

A void function does not require a return statement, but it may contain return statements with *expr* of type `void`, or return statements of the form

```
return;
```

Evaluation of a void function call terminates when a return statement is reached, *or* when the execution of *block* is finished.

2.6.5 Functions and scope

The parenthesized part of a function definition contains the declarations of the formal arguments. For all of them, the declarative region is the function definition, so the formal arguments have local scope (Section 2.4.3). The potential scope of a formal argument declaration begins after the declaration and extends until the end of the function body. Therefore, the formal arguments are not visible outside the function definition. Within the body, the formal arguments behave like variables that are local to *block*.

In particular, changes made to the values of formal arguments (like in the function pow) are "lost" after the function call and have no effect on the values of the call arguments. This is not surprising, since the call arguments are rvalues, but to make the point clear, let us consider the following alternative main function in callpow.cpp.

```
1 int main() {
2   double b = 2.0;
3   int e = -2;
4   std::cout << pow(b,e); // outputs 0.25
5   std::cout << b;        // outputs 2
6   std::cout << e;        // outputs -2
7
8   return 0;
9 }
```

The variables b and e defined in lines 2–3 are unaffected by the function call, since the function body of pow is not in the scope of their declarations, for two reasons. First, the definition of pow appears *before* the declarations of b and e in lines 2–3, so the body of pow cannot even be in the *potential* scope of these declarations. Second, even if we would move the declarations of the variables b and e to the beginning of the program (before the definition of pow, so that they have global scope), their scope would exclude the body of pow, since that body is in the potential scopes of redeclarations of the names b and e (the formal arguments), see Section 2.4.3.

But the general scope rules of Section 2.4.3 *do* allow function bodies to use names of global or namespace scope; the program on page 139 for example uses std::cout as such a name. Here is a contrived program that demonstrates how a program may modify a *global variable* (a variable whose declaration has global scope). While such constructions may be useful in certain cases, they usually make the program less readable, since the effect of a function call may then become very non-local.

```
1 #include<iostream>
```

```
 2
 3  int i = 0; // global variable
 4
 5  void f()
 6  {
 7    ++i;     // in the scope of declaration in line 3
 8  }
 9
10  int main()
11  {
12    f();
13    std::cout << i << "\n"; // outputs 1
14
15    return 0;
16  }
```

Since the formal arguments of a function have local scope, they also have automatic storage duration. This means that we get a "fresh" set of formal arguments every time the function is called, with memory assigned to them only until the respective function call terminates.

Function declarations. A function itself also has a scope, and the function can only be called within its scope. The scope of a function is obtained by combining the scopes of all its *declarations*. The part of the function definition before *block* is a declaration, but there may be function declarations that have no subsequent *block*. This is in contrast to variables where every declaration is at the same time a definition. A function may be declared several times, but it can be defined once only.

The following program, for example, does not compile, since the call of `f` in `main` is not in the scope of `f`.

```
#include<iostream>

int main()
{
  std::cout << f(1); // f undeclared
  return 0;
}

int f (const int i) // scope of f begins here
{
  return i;
}
```

But we can put `f` into the scope of `main` by adding a declaration before `main`, and this yields a valid program.

```
#include<iostream>
```

```
int f (int i); // scope of f begins here

int main()
{
  std::cout << f(1); // ok, call is in scope of f
  return 0;
}

int f (const int i)
{
  return i;
}
```

In declarations, we omit const-qualifications of arguments of fundamental type. As explained in Section 2.6.3, such qualifications are useful only internally, but the declaration is about the outside behavior of the function. For the compiler, the two declarations

```
int f (int i);
```

and

```
int f (const int i);
```

are identical.

In the previous program, we could get rid of the extra declaration by simply defining `f` before `main`, but sometimes, separate function declarations are indeed necessary. Consider two functions `f` and `g` such that `g` is called in the function body of `f`, and `f` is called in the function body of `g`. We *have* to define one of the two functions first (`f`, say), but since we call `g` within the body of `f`, `g` must have a declaration *before* the definition of `f`.

2.6.6 Procedural programming

So far, we have been able to "live" without functions only because the programs that we have written are pretty simple. But even some of these simple ones would benefit from functions. Consider as an example the program `perfect.cpp` from Exercise 49. In this exercise, we have asked you to find the perfect numbers between 1 and n, for a given input number n. The solution so far uses one "big" *double loop* (loop within a loop) that in turn contains two `if` statements. Although in this case, the "big" loop is still small enough to be read without difficulties, it doesn't really reflect the logical structure of the solution. Once we get to triple or quadruple loops, the program may become very hard to follow.

But what *is* the logical structure of the solution? For every number i between 1 and n, we have to **test whether i is perfect**; and to do the latter, we have to **compute the sum of all proper divisors of** i and check whether it is equal to i. Thus, we have two clearly defined subtasks that the program has to solve for every number i, and it is best to encapsulate

these into functions. Program 2.15 shows how this is done. Note that the program is now almost self-explanatory: the postconditions can more or less directly be read off the function names.

```
// Program: perfect2.cpp
// Find all perfect numbers up to an input number n

#include <iostream>

// POST: return value is the sum of all divisors of i
//       that are smaller than i
unsigned int sum_of_proper_divisors (const unsigned int i)
{
  unsigned int sum = 0;
  for (unsigned int d = 1; d < i; ++d)
    if (i % d == 0) sum += d;
  return sum;
}

// POST: return value is true if and only if i is a
//       perfect number
bool is_perfect (const unsigned int i)
{
  return sum_of_proper_divisors (i) == i;
}

int main()
{
  // input
  std::cout << "Find perfect numbers up to n =? ";
  unsigned int n;
  std::cin >> n;

  // computation and output
  std::cout << "The following numbers are perfect.\n";
  for (unsigned int i = 1; i <= n ; ++i)
    if (is_perfect (i)) std::cout << i << " ";
  std::cout << "\n";

  return 0;
}
```

Program 2.15: *../progs/lecture/perfect2.cpp*

Admittedly, the program is longer than perfect.cpp, but it is more readable, and it has simpler control flow. In particular, the double loop has disappeared.

The larger a program gets, the more important is it to subdivide it into small subtasks, in order not to lose track of what is going on in the program on the whole; this is the *procedural programming* paradigm, and in C++, it is realized with functions.

The procedural programming paradigm is not so self-evident as it may seem today. The first programming language that became accessible to a general audience since the 1960's was BASIC (**B**eginner's **A**ll-purpose **S**ymbolic **I**nstruction **C**ode).

In BASIC, there were no functions; in order to execute a code fragment responsible for a subtask, you had to use the `GOTO` statement (with a line number)—or `GOSUB` in many dialects—to jump to that code, and then jump back using another `GOTO` (`RETURN`, respectively). The result was often referred to as *spaghetti code*, due to the control flow meandering like a boiled spaghetti on a plate. Moreover, programmers often didn't think in terms of clearly defined subtasks, simply because the language did not support it. This usually lowered the code quality even further.

Despite this, BASIC was an extremely successful programming language. It reached the peak of its popularity in the late 1970's and early 1980's when the proud owners of the first home computers (among them the authors) created programs of fairly high complexity in BASIC.

2.6.7 Modularization

There are functions that are tailor-made for a specific program, and it would not make sense to use them in another program. But there are also general purpose functions that are useful in many programs. It is clearly undesirable to copy the corresponding function definition into every program that calls the function; what we need is *modularization*, a subdivision of the program into independent parts.

The power function `pow` from Program 2.14 is certainly general purpose. In order to make it available to all our programs, we can simply put the function definition into a separate sourcecode file `pow.cpp`, say, in our working directory.

```
#include <cassert>

// PRE:  e >= 0 || b != 0.0
// POST: return value is b^e
double pow (double b, int e)
{
  assert (e >= 0 || b != 0.0);
  double result = 1.0;
  if (e < 0) {
    // b^e = (1/b)^(-e)
    b = 1.0/b;
    e = -e;
  }
  for (int i=0; i<e; ++i) result *= b;
```

```
15   return result;
16 }
```

Program 2.16: *../progs/lecture/pow.cpp*

Then we can include this file from our main program as follows.

```
 1 // Prog: callpow2.cpp
 2 // Call a function for computing powers.
 3
 4 #include <iostream>
 5 #include "pow.cpp"
 6
 7 int main()
 8 {
 9   std::cout << pow( 2.0, -2) << "\n"; // outputs 0.25
10   std::cout << pow( 1.5,  2) << "\n"; // outputs 2.25
11   std::cout << pow( 5.0,  1) << "\n"; // outputs 5
12   std::cout << pow( 3.0,  4) << "\n"; // outputs 81
13   std::cout << pow(-2.0,  9) << "\n"; // outputs -512
14
15   return 0;
16 }
```

Program 2.17: *../progs/lecture/callpow2.cpp*

An include directive of the form

`#include "`*filename*`"`

logically replaces the include directive by the contents of the specified file. Usually, *filename* is interpreted relative to the working directory.

Separate compilation and object code files. The code separation mechanism from the previous paragraph has one major drawback: the compiler does not "see" it. Before compilation, `pow.cpp` is logically copied back into the main file, so the compiler still has to translate the function definition into machine language *every time* it compiles a program that calls `pow`. This is a waste of time that can be avoided by *separate compilation*.

In our case, we would compile the file `pow.cpp` separately. We only have to tell the compiler that it should not generate an executable program (it can't, since there is no `main` function) but an *object code* file, called `pow.o`, say. This file contains the machine language instructions that correspond to the C++ statements in the function body of `pow`.

Header files. The separate compilation concept is more powerful than we have seen so far: surprisingly, even programs that call the function `pow` can be compiled separately,

without knowing about the source code file pow.cpp or the object code file pow.o. What the compiler needs to have, though, is a declaration of the function pow.

This function declaration is best put into a separate file as well. In our case, this file pow.h, say, is very short; it contains just the lines

```
// PRE:  e >= 0 || b != 0.0
// POST: return value is b^e
double pow (double b, int e);
```

Since this is the "header" of the function pow, the file pow.h is called a *header file*. In the calling Program 2.17, we simply replace the inclusion of pow.cpp by the inclusion of pow.h, resulting in the following program.

```
1  // Prog: callpow3.cpp
2  // Call a function for computing powers.
3
4  #include <iostream>
5  #include "pow.h"
6
7  int main()
8  {
9    std::cout << pow( 2.0, -2) << "\n"; // outputs 0.25
10   std::cout << pow( 1.5,  2) << "\n"; // outputs 2.25
11   std::cout << pow( 5.0,  1) << "\n"; // outputs 5
12   std::cout << pow( 3.0,  4) << "\n"; // outputs 81
13   std::cout << pow(-2.0,  9) << "\n"; // outputs -512
14
15   return 0;
16 }
```

Program 2.18: *../progs/lecture/callpow3.cpp*

From this program, the compiler can then generate an object code file callpow3.o. Instead of the machine language instructions for executing the body of pow, this object code contains a *placeholder* for the location under which these instructions are to be found in the executable program. It is important to understand that callpow3.o cannot be an executable program yet: it *does* contain machine language code for main, but *not* for another function that it needs, namely pow.

The linker. Only when an executable program is built from callpow3.o, the object code file pow.o comes into play. Given all object files that are involved, the *linker* builds the executable program by gluing together machine language code for function calls (in callpow3.o) with machine language code for the corresponding function bodies (in pow.o). Technically, this is done by putting all object files together into a single executable file, and by filling placeholders for function body locations with the actual locations in the executable.

Separate compilation is very useful. It allows to change the definition of a function without having to recompile a single program that calls it. As long as the function declaration remains unchanged, it is only the linker that has to work in the end; and the linker is usually very fast. It follows that separate compilation also makes sense for functions that are specific to one program only.

Separate compilation reflects the "customer" view of the calling program: as long as a function does what its pre- and postcondition promise in the header file, it is not important to know *how* it does this. On the other hand, if the function definition is hidden from the calling program, clean pre- and postconditions are of critical importance, since they may be the only information available about the function's behavior.

Availability of sourcecode. If you have carefully gone through what we have done so far, you realize that we could in principle delete the sourcecode file `pow.cpp` after having generated `pow.o`, since later, the function definition is not needed anymore. When you buy commercial software, you are often faced with the absence of sourcecode files, since the vendor does not want customers to modify the sourcecode instead of buying updates, or to discover how much money they have paid for lousy software. (To be fair, we want to remark that there are also more honest reasons for not giving away sourcecode.)

In academic software, availability of sourcecode goes without saying. In order to evaluate or reproduce the contribution of such software to the respective area of research, it is necessary to have sourcecode. Even in commercial contexts, *open source* software is advancing. The most prominent software that comes with all sourcecode files is the operating system *Linux*. Open source software can very efficiently be adapted and improved if many people contribute. But such a contribution is possible only when the sourcecode is available.

Libraries. The function `pow` will not be the only mathematical function that we want to use in our programs. To make the addition of new functions easy, we can put the definition of `pow` (and similar functions that we may add later) into a single sourcecode file `math.cpp`, say, and the corresponding declarations into a single header file `math.h`. The object code file `math.o` then contains machine language code for all our mathematical functions.

Although not strictly necessary, it is good practice to include `math.h` in the beginning of `math.cpp`. This ensures consistency between function declarations and function definitions and puts the code in `math.cpp` into the scope of all functions declared in `math.h`, see Section 2.6.5. In all function bodies in `math.cpp`, we can therefore call the other functions, without having to think about whether these functions have already been declared.

In general, several object code files may be needed to generate an executable program, and it would be cumbersome to tell the linker about all of them. Instead, object code files that logically belong together can be *archived* into a *library*. Only the name of this library must then be given to the linker in order to have all library functions available for

the executable program. In our case, we so far have only one object file `math.o` resulting from `math.cpp`, but we can still build a library file `libmath.a`, say, from it.

Figure 7 schematically shows how object code files, a library and finally an executable program are obtained from a number of sourcecode files.

Figure 7: *Building object code files, libraries and executable programs.*

Centralization and namespaces It is clear that we do not want to keep header files and libraries of general interest in our working directory, since we (and others) may have many working directories. Header files and libraries should be at some central place.

We can make our programs independent from the location of header files by writing

```
#include <filename>
```

but in this case, we have to tell the compiler (when we start it) where to search for files to be included. This is exactly the way that headers like `iostream` from the standard library are included; their locations are known to the compiler, so we don't have to provide any information here. Similarly, we can tell the linker where the libraries we need are to be found. Again, for the various libraries of the standard library, the compiler knows this. We want to remark that *filename* is not necessarily the name of a physical file; the mapping of *filename* to actual files is implementation defined.

Finally, it is good practice to put all functions of a library into a namespace, in order to avoid clashes with user-declared names, see Section 2.1.3. Let us use the namespace `ifmp` here.

Here are the header and implementation files `math.h` and `math.cpp` that result from these guidelines for our intended library of mathematical functions (that currently contains `pow` only).

```
// math.h
// A small library of mathematical functions.

namespace ifmp {
  // PRE:  e >= 0 || b != 0.0
  // POST: return value is b^e
  double pow (double b, int e);
}
```

Program 2.19: *../progs/lecture/math.h*

```
// math.cpp
// A small library of mathematical functions.

#include <cassert>
#include <IFMP/math.h>

namespace ifmp {

  double pow (double b, int e)
  {
    assert (e >= 0 || b != 0.0);
    // PRE:  e >= 0 || b != 0.0
    // POST: return value is b^e
    double result = 1.0;
    if (e < 0) {
      // b^e = (1/b)^(-e)
      b = 1.0/b;
      e = -e;
    }
    for (int i=0; i<e; ++i) result *= b;
    return result;
  }

}
```

Program 2.20: *../progs/lecture/math.cpp*

Finally, the program `callpow4.cpp` calls our library function `ifmp::pow`. It includes the header file `math.h` from a central directory `IFMP`.

```
// Prog: callpow4.cpp
// Call library function for computing powers.

#include <iostream>
#include <IFMP/math.h>

int main()
{
  std::cout << ifmp::pow( 2.0, -2) << "\n"; // outputs 0.25
  std::cout << ifmp::pow( 1.5,  2) << "\n"; // outputs 2.25
  std::cout << ifmp::pow( 5.0,  1) << "\n"; // outputs 5
  std::cout << ifmp::pow( 3.0,  4) << "\n"; // outputs 81
  std::cout << ifmp::pow(-2.0,  9) << "\n"; // outputs -512

  return 0;
}
```

Program 2.21: *../progs/lecture/callpow4.cpp*

2.6.8 Using library functions

You can imagine that we were not the first to put a function like `pow` into a library. Indeed, the standard library contains a function `std::pow` that is even more general than ours: it can compute b^e for *real* exponents e. Accordingly, the arguments of `std::pow` and its return value are of type `double`. In order to use this function, we have to include the header `cmath`. This header contains declarations for a variety of other numerical functions.

Using functions from the standard library can help us to get shorter, better, or more efficient code, without having to write a single new line by ourselves. For example, computing *square roots* can speed up our primality test in Program 2.8. You might have realized this much earlier, but when we are looking for some proper divisor of a natural number $n \geq 2$, it is sufficient to search in the range $\{2, \ldots, \lfloor\sqrt{n}\rfloor\}$. Indeed, if n can be written as a product $n = dd'$, then the smaller of d and d' must be bounded by $\sqrt{n}$; since the divisors are integral, we even get a bound of $\lfloor\sqrt{n}\rfloor$, $\sqrt{n}$ rounded down.

The primality test could therefore be written more efficiently; Program 2.22 does this, using the function `std::sqrt` from the library `cmath`, whose argument and return types are `double`.

```
// Program: prime2.cpp
// Test if a given natural number is prime.

```

```
#include <iostream>
#include <cmath>

int main ()
{
  // Input
  unsigned int n;
  std::cout << "Test if n>1 is prime for n =? ";
  std::cin >> n;

  // Computation: test possible divisors d up to sqrt(n)
  const unsigned int bound = (unsigned int)(std::sqrt(n));
  unsigned int d;
  for (d = 2; d <= bound && n % d != 0; ++d);

  // Output
  if (d <= bound)
    // d is a divisor of n in {2,...,[sqrt(n)]}
    std::cout << n << " = " << d << " * " << n / d << ".\n";
  else
    // no proper divisor found
    std::cout << n << " is prime.\n";

  return 0;
}
```

Program 2.22: *../progs/lecture/prime2.cpp*

The program is correct: if `d <= bound` still holds after the loop, we have left the loop because the *other* condition `n % d != 0` has failed. This means that we have found a divisor. If `d > bound` holds after the loop, we have tried all possible divisors smaller or equal to `bound` (whose value is $\lfloor\sqrt{n}\rfloor$, since the explicit conversion rounds down, see Section 2.5.3), so we certainly have not missed any divisor. But we have to be a little careful here: our arguments assume that `std::sqrt` works correctly for squares. For example, `std::sqrt(121)` must return 11 (a little more wouldn't hurt), but *not* 10.99998, say. In that latter case, `(unsigned int)(std::sqrt(121))` would have value 10, and by making this our bound, we miss the divisor 11 of 121, erroneously concluding that 121 is prime.

It is generally not safe to rely on some precise semantics of library functions, even if your platform implements floating point arithmetic according to the IEEE standard 754 (see Section 2.5.6). The square root function is special in the sense that the IEEE standard still guarantees the result of `std::sqrt` to be the floating point number closest to the real square root; consequently, our above implementation of the primality test is safe. But similar guarantees do *not* necessarily hold for other library functions.

2.6.9 Details

Function signatures. In function declarations, the formal argument names *pname1*,..., *pnamek* can be omitted.

This makes sense since these names are only needed in the function definition. The important information, namely domain and range of the function, are already specified by the argument types and the return type. All these types together form the *signature* of the function.

In `math.h`, we could therefore equivalently write the declaration

```
double pow (double, int);
```

The only problem is that we need the formal argument names to specify pre- and post-conditions, without going to lengthy formulations involving "the first argument" and "the second argument". Therefore, we usually write the formal argument names even in function declarations.

Mathematical functions. Many of the mathematical functions that are available on scientific pocket calculators are also available from the math library `cmath`. The following table lists some of them. All are available for the three floating point number types `float`, `double` and `long double`.

name	function
`std::abs`	$\lvert x \rvert$
`std::sin`	$\sin(x)$
`std::cos`	$\cos(x)$
`std::tan`	$\tan(x)$
`std::asin`	$\sin^{-1}(x)$
`std::acos`	$\cos^{-1}(x)$
`std::atan`	$\tan^{-1}(x)$
`std::exp`	e^x
`std::log`	$\ln x$
`std::log10`	$\log_{10} x$
`std::sqrt`	$\sqrt{x}$

2.6.10 Goals

Dispositional. At this point, you should ...

1) be able to explain the purpose of functions in C++;
2) understand the syntax and semantics of C++ function definitions and declarations;
3) know what the term "procedural programming" means;
4) know why it makes sense to compile function definitions separately, and to put functions into libraries.

Operational. In particular, you should be able to ...

(G1) give two reasons why it is desirable to subdivide programs into functions;

(G2) find pre- and postconditions for given functions, where the preconditions should be as *weak* as possible, and the postconditions should be as *strong* as possible;

(G3) find syntactical and semantical errors in function definitions, and in programs that contain function definitions;

(G4) evaluate given function call expressions;

(G5) subdivide a given task into small subtasks, and write a program for the given task that uses functions to realize the subtasks;

(G6) build a library on your platform, given that you are told the necessary technical details.

2.6.11 Exercises

Exercise 77 *Find pre- and postconditions for the following functions.* (G2)(G4)

a)
```
double f (const double i, const double j, const double k)
{
  if (i > j)
    if (i > k)
      return i;
    else
      return k;
  else
    if (j > k)
      return j;
    else
      return k;
}
```

b)
```
double g (const int i, const int j)
{
  double r = 0.0;
  for (int k = i; k <= j; ++k)
    r += 1.0 / k;
  return r;
}
```

Exercise 78 *What are the problems (if any) with the following functions? Fix them and find appropriate pre- and postconditions.* (G2)(G3)

a)
```
bool is_even (const int i)
{
  if (i % 2 == 0) return true;
}
```

b)
```
double inverse (const double x)
{
  double result;
  if (x != 0.0)
    result = 1.0 / x;
  return result;
}
```

Exercise 79 *What is the output of the following program, depending on the input number* `i`*? Describe the output in mathematical terms, ignoring possible over- and underflows.* (G4)

```
#include<iostream>

int f (const int i)
{
  return i * i;
}

int g (const int i)
{
  return i * f(i) * f(f(i));
}

void h (const int i)
{
  std::cout << g(i) << "\n";
}

int main()
{
  int i;
  std::cin >> i;
  h(i);

  return 0;
}
```

Exercise 80 *Find three problems in the following program.* (G3)(G4)

```
#include<iostream>

double f (const double x)
{
  return g(2.0 * x);
}

bool g (const double x)
{
  return x % 2.0 == 0;
}

void h ()
{
  std::cout << result;
}

int main()
{
  const double result = f(3.0);
  h();

  return 0;
}
```

Exercise 81 *Simplify the program from Exercise 71 by using the library function* `std::pow`. (G5)

Exercise 82 *Assume that on your platform, the library function* `std::sqrt` *is not very reliable. For* x *a value of type* `double` *(*$x \geq 0$*), we let* $s(x)$ *be the value returned by* `std::sqrt(expr)`*, if expr has value* x*, and we assume that we only know that for some positive value* $\varepsilon \leq 1/2$*, the relative error satisfies*

$$\frac{|s(x) - \sqrt{x}|}{\sqrt{x}} \leq \varepsilon, \quad \forall x.$$

How can you change Program 2.22 such that it correctly works under this relative error bound? You may assume that the floating point number system used on your platform is binary, and that all values of type `unsigned int` *are exactly representable in this system. (This is a theory exercise.)*

Exercise 83

a) Write a function

```
// POST: return value is true if and only if n is prime
bool is_prime (unsigned int n);
```

and use this function in a program to count the number of twin primes *in the range* $\{2, \ldots, 10000000\}$ *(two up to ten millions). A twin prime is a pair of numbers* $(i, i+2)$ *both of which are prime.*

b) *Is the approach of a) the best (most efficient) one to this problem? If you can think of a better approach, you are free to implement it instead of the one outlined in a).*

(G5)

Exercise 84 *The function* `pow` *in Program 2.14 needs* $|e|$ *multiplications to compute* b^e. *Change the function body such that less multiplications are performed. You may use the following fact. If* $e \geq 0$ *and* e *has binary representation*

$$e = \sum_{i=0}^{\infty} b_i 2^i,$$

then

$$b^e = \prod_{i=0}^{\infty} \left(b^{2^i}\right)^{b_i}.$$

(G5)

Exercise 85 *A* perpetual calendar *can be used to determine the weekday (Monday, ..., Sunday) of any given date. You may for example know that the Berlin wall came down on November 9, 1989, but what was the weekday? (It was a Thursday.) Or what is the weekday of the 1000th anniversary of the Swiss confederation, to be celebrated on August 1, 2291? (Quite adequately, it will be a Saturday.)*

a) *The task is to write a program that outputs the weekday (Monday, ..., Sunday) of a given input date.*

   Identify a set of subtasks to which you can reduce this task. Such a set is not unique, of course, but all individual subtasks should be small (so small that they could be realized with few lines of code). It is of course possible for a subtask in your set to reduce to other subtasks. (Without giving away anything, one subtask that you certainly need is to determine whether a given year is a leap year).

b) *Write a program* `perpetual_calendar.cpp` *that reads a date from the input and outputs the corresponding weekday. The range of dates that the program can process should start no later than January 1, 1900 (Monday). The program should check whether the input is a legal date, and if not, reject it. An example run of the program might look like this.*

```
day =? 13
month =? 11
year =? 2007
Tuesday
```

To structure your program, implement the subtasks from a) as functions, and put the program together from these functions.

(G5)

Exercise 86 *Build a library on your platform from the files* `math.h` *and* `math.cpp` *in Program 2.19 and Program 2.20. Use this library to generate an executable program from Program 2.21.* (G5)(G6)

Exercise 87

a) *Implement the following function and test it. You may assume that the type* `double` *complies with the IEEE standard 754, see Section 2.5.6. The function is only required to work correctly, if the nearest integer is in the value range of the type* `int`. (G5)

```
// POST: return value is the integer nearest to x
int round (double x);
```

b) *The postcondition of the function does not say what happens if there are two nearest integers. Specify the behavior of your implementation in the postcondition of your function.* (G2)

c) Add a declaration of your function to the file `math.h` (Program 2.19) and a definition to `math.cpp` (Program 2.20). Build a library from these two files, and rewrite your test function from a) to call the library version of the function `round`. (G6)

Exercise 88 *This is another (not too difficult) one from* Project Euler *(Problem 56,* `http://projecteuler.net/`*). Find natural numbers* $a, b < 100$ *for which* a^b *has the largest cross sum (sum of decimal digits). Let us say upfront that* 99^{99} *is not the answer.*

Write a program `power_cross_sums.cpp` *that computes the best* a *and* b *(within reasonable time).*

Can you also find the best a, b *up to* $1,000$*?*

2.6.12 Challenges

Exercise 89 *(This is a theory challenge.) The simplest computer model that is being studied in theoretical computer science is the* deterministic finite automaton *(DFA).*

Such an automaton is defined over a finite alphabet Σ *(for example* $\Sigma = \{0, 1\}$*). Then it has a finite set of* states Q*. The main ingredient is the* transition function

$$\delta : Q \times \Sigma \to Q.$$

We can visualize this function as follows: whenever $\delta(q, \sigma) = q'$*, we draw an arrow from state* q *to state* q'*, labeled with* σ*.*

Finally, there is a starting state $s \in Q$ *and a subset* $F \subseteq Q$ *of* accepting states*. Figure Figure 8 depicts a DFA with state set* $Q = \{0, 1, 2\}$*. The starting state is indicated by an arrow coming in from nowhere, and the accepting states are marked with double circles (in this case, there is only one).*

Figure 8: *A deterministic finite automaton (DFA)*

Why can we call such an automaton a computer model? Because it performs a computation, namely the following: given an input word $w \in \Sigma^*$ *(finite sequence of symbols from the alphabet* Σ*), the automaton either* accepts*, or* rejects *it. To do this, the word* w *is processed symbol by symbol, starting in* s*. Whenever the automaton is in some state* q *and the next symbol is* σ*, the automaton switches to state* $q' = \delta(q, \sigma)$*. When all symbols have been processed, the automaton is either in an accepting state* $q \in F$ *(in which case* w *is accepted), or in a non-accepting state* $q \in Q \setminus F$ *(in which case* w *is rejected).*

For example, when we feed the automaton of Figure 8 with the word $w = 0101$*, the sequence of states that are being visited is* $0, 0, 1, 2, 2$*. Consequently,* w *is rejected.*

The language L *of the automaton is the set of accepted words. This is a (generally infinite) subset of* Σ^**. Let's try to determine the language of the automaton in Figure 8.*

It turns out that this is not such a straightforward task, and you need the right idea. (To be honest, we had the idea first and then came up with an automaton that realizes it). We claim that the automaton accepts exactly all the words that

are divisible by 3 *if you interpret the word as a binary number (where the empty word is interpreted as* 0*). For example,* 0101 *is the binary number*

$$0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 5,$$

and indeed 5 *is not divisible by* 3 *(and hence rejected). In fact (and this is the key to the proof of our claim), the state after processing* w *is the one numbered with* $w \bmod 3$*. You can therefore say that the DFA of Figure 8 is a computer (with a built-in program) that can solve the decision problem of checking whether a given number is divisible by 3.*

We are slowly approaching the actual challenge. For every subset L *of* $\{0,1\}^*$ *from the following list, either find a DFA that has* L *as its language, or prove that such a DFA cannot exist (which would show that DFA are limited in their computational power).*

a) $L = \{w \in \{0,1\}^* \mid w$ *has an even number of zeros and an even number of ones*$\}$

b) $L = \{w \in \{0,1\}^* \mid w$ *is divisible by* 5 *when interpreted as a binary number*$\}$

c) $L = \{w \in \{0,1\}^* \mid w$ *has more zeros than ones*$\}$

d) $L = \{w \in \{0,1\}^* \mid w$ *does not contain three consecutive ones*$\}$

Exercise 90 *The two logicians Mr. Sum and Mr. Product are friends who one day have a phone conversation. Before, Mr. Sum only knows the sum* $s = a + b$ *of two unknown integer numbers* $1 < a < 100$ *and* $1 < b < 100$*. Mr. Product, on the other hand, only knows the product* $p = a \cdot b$ *of* a *and* b*. The conversation goes as follows:*

Mr. Product:	*"I don't know the numbers* a *and* b*."*
Mr. Sum:	*"I knew that you don't know them."*
Mr. Product:	*"Ah... but now I know them."*
Mr. Sum:	*"Then I know them too, now."*

What is the set of numbers $\{a, b\}$*? You have to assume, of course, that all statements made during this conversation are true. Write a program that computes* $\{a, b\}$*, or submit a convincing written solution!*

2.7 Arrays

As all real programmers know, the only useful data structure is the array.

Ed Post, Real Programmers don't use Pascal (1983)

Reading into an array without making a "silly error" is beyond the ability of complete novices - by the time you get that right, you are no longer a complete novice.

Bjarne Stroustrup, C++ Style and Technique FAQ

This section introduces arrays to store sequences of objects of the same type, with random access to individual members of the sequence. A static array is the most primitive but at the same time the most efficient type for storing, processing, and iterating over large amounts of data. You will also learn about vectors from the C++ standard library as generally better and more flexible arrays.

In Section 2.4 on control statements, we have learned about the concept of iteration. For example, we can now iterate over the sequence of numbers $1, 2, \ldots, n$ and perform some operations like adding up all the numbers, or identifying the prime numbers among them. Similarly, we can iterate over the odd numbers, the powers of two, etc.

In real applications, however, we often have to process (and in particular traverse) sequences of *data*. For example, if you want to identify the movie theaters in town that show your desired movie tonight, you have to traverse the sequence of movie theater repertoires. These repertoires must be stored somewhere, and there must be a way to inspect them in turn. In C++, we can deal with such tasks by using *arrays*.

2.7.1 Static arrays

An array of length n aggregates n objects of the *same* type T into a sequence. To access one of the aggregated objects (the *elements*), we use its *index* or *subscript* (position) in the sequence. All these length-n sequences form an array type whose value range corresponds to the mathematical type T^n. In the computer's main memory, an array occupies a contiguous part, with the elements stored side-by-side (see Figure 9).

Let us start by showing an array in action: *Eratosthenes' Sieve* is a fast method for computing all prime numbers smaller than a given number n, based on crossing out the numbers that are not prime. It works like this: you write down the sequence of numbers between 2 and $n-1$. Starting from 2, you always go to the next number not crossed out yet, report it as prime, and then cross out all its proper multiples.

Let's not dwell on the correctness of this method but go right to the implementation. If you think about it for a minute, the major question is this: how do we cross out numbers?

The following program uses a *static array* type variable crossed_out for the list, where every value crossed_out[i] is of type bool and represents the (changing) information whether the number i has already been crossed out or not. Array indices always start from 0, so in order to get to index $n-1$, we need an array of length n. The program runs Eratosthenes' Sieve for $n = 1,000$.

```
// Program: eratosthenes.cpp
// Calculate prime numbers in {2,...,n-1} using
// Eratosthenes' sieve.

#include <iostream>

int main()
{
  const unsigned int n = 1000;

  // definition and initialization: provides us with
  // Booleans crossed_out[0],..., crossed_out[n-1]
  bool crossed_out[n];
  for (unsigned int i = 0; i < n; ++i)
    crossed_out[i] = false;

  // computation and output
  std::cout << "Prime numbers in {2,...," << n-1 << "}:\n";
  for (unsigned int i = 2; i < n; ++i)
    if (!crossed_out[i]) {
      // i is prime
      std::cout << i << " ";
      // cross out all proper multiples of i
      for (unsigned int m = 2*i; m < n; m += i)
        crossed_out[m] = true;
    }
  std::cout << "\n";

  return 0;
}
```

Program 2.23: *../progs/lecture/eratosthenes.cpp*

Definition. A static array variable (or simply static array) a with $n > 0$ elements of *underlying type* T is defined through the following declaration.

T a[*expr*]

Here, *expr* must be a *constant expression* of integral type whose value is n. For example, literals like 1000, arithmetic expressions over literals (like 1+1), and constants (Section 2.1.9) are constant expressions; all of them have the property that their value is known at compile time. This allows the compiler to figure out how much memory the static array variable needs.

The type of *a* is "$T[n]$", but we put this in double quotes here (only to omit them later). The reason is that $T[n]$ is not the official name: we can't write `int[5] a`, for example, to declare a static array a of type `int[5]`.

The value range of $T[n]$ is T^n, the set of all sequences $(t_1, t_2, \ldots, t_n)$ with all t_i being of type T.

The fact that the static array length must be known at compile time clearly limits the usefulness of static array variables. For example, this limitation does not allow us to write a version of Eratosthenes' sieve in which the number n is read from the input. But we will shortly see how this restriction can be overcome—for the time being, let's simply live with it.

2.7.2 Initializing static arrays

The definition of a static array with underlying fundamental type does not initialize the values of the array elements. We can assign values to the elements afterwards (like we do it in Program 2.23), but we can also provide the values directly, as in the following declaration statement.

```
int a[5] = {4,3,5,2,1};
```

Since the number of array elements can be deduced from the length of the *initializer list*, we can also write

```
int a[] = {4,3,5,2,1};
```

The declaration `int a[]` without any initialization is invalid, though, since it does not fully determine the type of a. We say that a has *incomplete type* in this case.

2.7.3 Random access to elements

The most common and useful way of accessing and modifying the elements of an array is by *random access*. If *expr* is of integral type and has value i, the lvalue

a[*expr*]

is of the type underlying the static array a and refers to the i-th element (counting from 0) of a. The number i is called the *index* or *subscript* of the element. If n is the length of a, the index i must satisfy $0 \leq i < n$. The operator `[]` is called the *subscript operator*.

The somewhat strange declaration format of a static array is motivated by the subscript operator. Indeed, the declaration

T a[*expr*]

can be read as "a[*expr*] is of type T". In this sense, it is an implicit definition of a's type.

> **Watch out!** *You* as a programmer are responsible for making sure that a given array index i indeed satisfies $0 \leq i < n$, where n is the length of the array. Indices that are not in this range are called *out of bound*. Unless your compiler offers specific debugging facilities, the usage of out-of-bound indices in the subscript operator is *not* checked at runtime and leads to undefined behavior of the program.

We have already discussed the term random access in connection with the computer's main memory (Section 1.2.3); random access means that *every* array element can be accessed in the same uniform way, and with (almost) the same access time, no matter what its index is. Evaluating the expression `a[0]` is as fast as evaluating `a[10000]`. In contrast, the thick pile of pending invoices, bank transfers and various other papers on your desk does not support random access: the time to find an item is roughly proportional to its depth within the pile.

In fact, random access in an array directly reduces to random access in the computer's main memory, since an array always occupies a contiguous set of memory cells, see Figure 9.

Figure 9: *An array occupies a contiguous part of the main memory. Every element in turn occupies* s *memory cells, where* s *is the number of memory cells required to store a single value of the underlying type* T.

To access the element of index i in the array a, a simple computation with addresses therefore suffices. If p is the address (position) where the first element of a "starts", and s is the number of memory cells that a single value of the underlying type T occupies, then the element of index i starts at the memory cell whose address is $p + si$, see Figure 10.

2.7.4 Static arrays are primitive

Static array types are exceptional in C++. The following code fragment illustrates this:

```
int a[5] = {4,3,5,2,1}; // array of type int[5]
```

Figure 10: *The array element of index* i *starts at the address* p + si.

```
int b[5];
b = a;                    // error: we cannot assign to an array
```

We also cannot initialize a static array from another one. Why is this? Static arrays are a dead hand from the programming language C, and the design of arrays in C is quite primitive. C was designed to compete with machine language in efficiency, and this didn't leave room for luxury. We *can* of course copy one static array into another manually via a simple loop, and early C programmers were not yet spoiled enough to complain about this bit of extra work.

When C++ was developed much later, one design goal was to have C as a subset. As a consequence, static arrays are still around in C++, in their original form. Internally, a static array is represented by the address of its first memory cell, and by its memory size in bytes, see Section 1.2.3. In general, if *obj* is any object of any type, the expression

```
sizeof(obj)
```

returns the number of memory cells used by its argument, under the agreement that one memory cell is required to store an object of type `char`.

The argument of `sizeof` can also be the name of a type, in which case the number of memory cells occupied by one object of this type is returned (the number s in Figure 9). Thus, to find out how many elements a static array a of underlying type `int` has, we may use the somewhat clumsy expression

```
sizeof(a) / sizeof(int)
```

Luckily, the C++ standard library offers less primitive array types. In the following, we will see how *vectors* can be used and what some of their benefits are over plain arrays.

2.7.5 Vectors

Let us go back to Program 2.23. Its main drawback is that the number n is hardwired as 1,000 in this program, just because the length of a static array has to be known at compile time.

At least in this respect, static arrays are nothing special, though. All types that we have met earlier (`int`, `unsigned int`, and `bool`) have the property that a single object of the type occupies a fixed amount of memory known to the compiler (for example, 32 bits for an `int` object on many platforms). With arrays, an obvious need arises to circumvent this restriction.

Here is how this works for Eratosthenes' Sieve. Remember that we want the list of prime numbers between 2 and $n-1$. The following variant reads the number n from standard input and then defines a *vector* variable `crossed_out` of length n, with all its elements (of type `bool`, to be provided in angle brackets) set to *false*. This is done using the single declaration

```
std::vector<bool> crossed_out (n, false);
```

The remainder of the program is exactly as before. In order to be able to use vectors, we need to

```
#include <vector>
```

at the beginning of our program.

```
// Program: eratosthenes2.cpp
// Calculate prime numbers in {2,...,n-1} using
// Eratosthenes' sieve.

#include <iostream>
#include <vector>  // vectors are standard containers with
                   // array functionality
int main()
{
  // input
  std::cout << "Compute prime numbers in {2,...,n-1} for n =? ";
  unsigned int n;
  std::cin >> n;

  // definition and initialization: provides us with
  // Booleans crossed_out[0],..., crossed_out[n-1],
  // initialized to false
  std::vector<bool> crossed_out (n, false);

  // computation and output
  std::cout << "Prime numbers in {2,...," << n-1 << "}:\n";
  for (unsigned int i = 2; i < n; ++i)
    if (!crossed_out[i]) {
      // i is prime
      std::cout << i << " ";
      // cross out all proper multiples of i
      for (unsigned int m = 2*i; m < n; m += i)
```

```
28          crossed_out[m] = true;
29      }
30    std::cout << "\n";
31
32    return 0;
33  }
```

Program 2.24: *../progs/lecture/eratosthenes2.cpp*

A vector is an array with extra functionality that makes it more convenient to use. One piece of extra functionality is the initialization. While we needed an explicit loop in Program 2.23 in order to initialize the static array `crossed_out`, the vector `crossed_out` can directly be initialized by a *constructor* call.

We formally get to constructors only later; here we just need to know that a constructor is a function—in our case with the list of two arguments `(n, false)`—that takes care of initialization, according to postconditions that we can look up in the standard library reference. To call the constructor, we simply append the arguments to the name of the new variable to be initialized.

2.7.6 Strings

Sequences of characters enclosed in double quotes as in

```
std::cout << "Prime numbers in {2,...,"
```

are called *string literals*.

So far we have used string literals only within output expressions, but we can work with them in other contexts as well. Most notably, a string literal can be used to initialize a *string* which is an array of *characters* with some additional functionality. Characters are the building blocks of text as we know it. In C++, they are modeled by the fundamental type `char` that we briefly discuss next.

The type char. The fundamental type `char` represents characters. Characters include the *letters* a through z (along with their capital versions A through Z), the *digits* 0 through 9, as well as numerous other *special characters* like % or $. The line

```
char c = 'a';
```

defines a variable `c` of type `char` and value `'a'`, representing the letter a. The expression `'a'` is a literal of type `char`. The quotes around the actual character symbol are necessary in order to distinguish the literal `'a'` from the identifier `a`.

Formally, the type `char` is an integral type: it has the same operators as the types `int` or `unsigned int`, and the C++ standard even postulates a promotion from `char` to `int` or `unsigned int`. It is *not* specified, though, to which integer the character `'a'`, say, will be promoted. Under the widely used `ASCII` code (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange), it is the integer 97.

This setting may not seem very useful, and indeed it makes little sense to divide one character by another. On the other hand, we can for example print the alphabet through one simple loop (assuming ASCII encoding). Execution of the for-loop

```
for (char c = 'a'; c <= 'z'; ++c)
  std::cout << c;
```

writes the character sequence

```
abcdefghijklmnopqrstuvwxyz
```

to standard output. Given this, you may think that the line

```
std::cout << 'a' + 1;
```

prints 'b', but it doesn't. Since the operands of the composite expression 'a'+1 are of different types, the left operand of type char will automatically be promoted to the more general type int of the right operand. Therefore, the type of the expression 'a'+1 is int, and its value is 98 (assuming ASCII encoding); and that's what gets printed. If you want 'b' to be printed, you must use the explicit conversion char('a'+1).

The category of special characters also includes *control characters* that *do* something when printed. These are written with a leading backslash, and the most important control character for us is '\n', which causes a line break.

A char value occupies 8 bits (one byte) of memory; whether the value range correspond to the set of integers $\{-128, \ldots, 127\}$ (the signed case) or the set $\{0, \ldots, 255\}$ (the unsigned case) is implementation defined. Since all ASCII characters have integer values in $\{0, \ldots, 127\}$, they can be represented in both cases.

From characters to text. A text is simply a sequence of characters and can be modeled in C++ through a value of the type std::string For example, the declaration

```
std::string text = "bool"
```

defines a string of length 4 that represents the text *bool*. Before using strings from the standard library, we need to

```
#include <string>
```

Here is an application of strings. *String matching* is the problem of finding the first or all occurrences of a given pattern (usually short) in a given text (usually long).

The obvious solution is the following: assuming that the pattern has length m, we compare it characterwise with the elements $1, 2, ..., m$ of the text. If a mismatch is found for some element, we stop and next compare the pattern with the elements $2, 3, \ldots, m+1$ of the text, and so on. Sets of m consecutive elements $i, i+1, \ldots, i+m-1$ in the text are called a *window*.

This algorithm is fast as long as the pattern is short, but it may become inefficient for long patterns (see Exercise 95). There is a more sophisticated algorithm (the *Knuth-Morris-Pratt algorithm*) that is always fast.

The following Program 2.25 implements the obvious algorithm. It can be used to search for a pattern in a very long text, since it does not store the whole text but only keeps the current window to be compared with the pattern. Such an algorithm is known as a *streaming algorithm*, since the input is considered as a stream just "flowing through" the program.

The program maintains two strings, one for the pattern, and one for the current window. The input text is read character by character, and after each character, the window is advanced and compared with the pattern.

```
// Program: string_matching.cpp
// find the first occurrence of a fixed pattern within the
// input text, and output the text read so far

#include<iostream>
#include<string>     // strings are sequences of characters
                     // with array functionality
#include<ios>        // for std::noskipws

// POST: returns window, after removing the first character and
//       appending next as a new last character
std::string advance (std::string window, const char next)
{
  int last_index = window.length()-1;
  for (int i=0; i<last_index; ++i)
     window[i] = window[i+1];
  window[last_index] = next;
  return window;
}

int main ()
{
  // search pattern
  const std::string pattern = "bool";

  // "empty" text window of pattern length
  std::string window (pattern.length(), '\0');

  // read characterwise from standard input until
  // search pattern was found
  std::cin >> std::noskipws; // don't skip whitespaces!
  char c;                    // next character to be read
  while (std::cin >> c) {
    std::cout << c;
    window = advance (window, c);
    if (window == pattern) break;
```

```
  }
  std::cout << "\n";

  return 0;
}
```

Program 2.25: *../progs/lecture/string_matching.cpp*

When we apply the program to the text of the file `eratosthenes.cpp`, the program outputs Program 2.23 up to the first occurrence of the string `"bool"`:

```
// Program: eratosthenes.cpp
// Calculate prime numbers in {2,...,n-1} using
// Eratosthenes' sieve.

#include <iostream>

int main()
{
  const unsigned int n = 1000;

  // definition and initialization: provides us with
  // Booleans crossed_out[0],..., crossed_out[n-1]
  bool
```

The initialization of the window happens in a way that should be familiar from Program 2.24: the declaration

```
std::string window (pattern.length(), '\0')
```

uses a constructor call to define a new variable `window` of type `std::string` and initialize it to a sequence of *null characters* whose length equals the length of the search pattern.

A new feature of Program 2.25 is the use of *member function* calls of the form

```
pattern.length()
```

As with constructor calls before, we will get to the precise details only later; at this point, we need to understand that the above expression is a call of a function `length` with no *explicit* arguments (we have an empty argument list) but one *implicit* argument, namely the expression `pattern` of type `std::string`. The `.` syntax designates that the expression before the dot is an implicit argument of the function after the dot. In our case, the value of the expression `pattern.length()` is (unsurprisingly) the length of the pattern.

The function `advance` returns the new window resulting from moving the old window one place to the right, i.e. throwing away the first character, and appending the newly read character. Here we see that a string can be used just as an array: we have random access to the single characters of a string.

Redirecting standard input. A few comments need to be made with respect to the handling of standard input here. The program reads the text character by character from `std::cin`, until this stream becomes "empty". To test this, we use the fact that stream values can implicitly be converted to `bool`, with the result being *true* as long as there was no attempt at reading past the end of the stream. Since the value of `std::cin >> c` is the stream *after* removal of one character, the conversion to `bool` exactly tells us whether there still was a character in the stream, or not.

Most conveniently, the program is run by redirecting standard input to a file containing the text. In this case, the stream `std::cin` will become empty exactly at the end of the file. The line

```
std::cin >> std::noskipws; // don't skip whitespaces!
```

is necessary to tell the stream that *whitespaces* (blanks, newlines, etc.) should not be ignored (by default, they are). This allows us to search for strings that contain whitespaces, and it allows us to output the text (up to the first occurrence of the search string) in its original layout.

2.7.7 Multidimensional arrays

In C++, we can have arrays of arrays. For example, the declaration

```
int a[2][3]
```

declares `a` to be a static array of length 2 whose elements are static arrays of length 3 with underlying type `int`. We also say that `a` is a *multidimensional array* (in this case of dimensions 2 and 3). The type of `a` is "`int[2][3]`", and the underlying type is `int[3]`. In general, the declaration

T a[*expr1*]...[*exprk*]

defines a static array `a` of length n_1 (value of *expr1*) whose elements are static arrays of length n_2 (value of *expr2*) whose elements are... you get the picture. The values $n_1, \ldots, n_k$ are called the *dimensions* of the array, and the expressions *expr1*,..., *exprk* must be constant expressions of integral type and positive value.

Random access in multidimensional static arrays works as expected: `a[i]` is the element of index `i`, and this element is a static array itself. Consequently, `a[i][j]` is the element of index `j` in the array `a[i]`, and so on.

Although we usually think of multidimensional arrays as tables or matrices, the memory layout is "flat" like for one-dimensional arrays. For example, the twodimensional array declared through `int a[2][3]` occupies a contiguous part of the memory, with space for $6 = 2 \times 3$ objects of type `int`, see Figure 11.

Multidimensional static arrays can be initialized in a way similar to onedimensional static arrays; the value for the *first* (and *only* the first) dimension may be omitted:

```
int a[][3] = { {2,4,6}, {1,3,5} };
```

Figure 11: *Memory layout of a twodimensional static array*

This defines a static array of type `int[2][3]` where `{2,4,6}` is used to initialize the element `a[0]`, and `{1,3,5}` is used for `a[1]`.

Vectors of vectors. The required dimensions of a multidimensional array may not be known at compile time in which case we will work with vectors of vectors (of vectors...).

Let us discuss the twodimensional case only to avoid lengthy formulae.

To get a two-dimensional vector variable a with dimensions n and m, we create a vector with n elements, where each element is in turn a vector with m elements. Using the constructor call syntax introduced in Program 2.24, this looks somewhat formidable, but conceptually, it is simple:

```
std::vector<std::vector<int> > a (n, std::vector<int>(m));
```

This means that a is a vector whose elements are of type `std::vector<int>` (the type in the outer angle brackets). In the constructor call, we are saying that a should have n elements, each of which will be initialized with a vector of m `ints`'s (the type in the inner angle brackets). These integers themselves don't get initialized, but we could also have initialized all of them with 0, say, by writing

```
std::vector<std::vector<int> > a (n, std::vector<int>(m, 0));
```

The type of a is `std::vector<std::vector<int> >`; hence, each element `a[i]` is of type `std::vector<int>`. Consequently, `a[i][j]` is an lvalue of type `int`, just like in a "regular" twodimensional static array.

Computing shortest paths. Let us conclude this section with an interesting application of (multidimensional) arrays. Imagine a rectangular factory floor, subdivided into square cells. Some of the cells are blocked with obstacles (these could for example be machines or cupboards, but let us abstractly call them "walls"). A robot is initially located at some cell S (the source), and the goal is to move the robot to some other cell T (the target). At any time, the robot can make one step from its current cell to any of the four adjacent cells, but for obvious reasons it may only use cells that are empty.

Given this setup, we want to find a shortest possible robot path from S to T (or find out that no such path exists). Here, the length of a robot path is the number of steps taken by the robot during its motion from S to T (the initial cell S does not count; in

particular, it takes 0 steps to reach S from S). Figure 12 (left) shows an example with 8×12 cells.

Figure 12: *Left: What is a shortest robot path from* S *to* T*? Right: This one!*

In this example, a little thinking reveals that there are essentially two different possibilities for the robot to reach T: it can pass below the component of walls adjacent to S, or above. It turns out that passing above is faster, and a resulting shortest path (of length 21) is depicted in Figure 12 (right). Note that in general there is not a unique shortest path. In our example, the final right turn of the path could also have been made one or two cells further down.

We want to write a program that finds a shortest robot path, given the dimensions n (number of rows) and m (number of columns) of the factory floor, the coordinates of source and target, and the walls. How can this be done? Before reading further, we encourage you to think about this problem for a while. Please note that the *brute-force approach* of trying all possible paths and selecting the shortest one is not an option, since the number of such paths is simply too large already for moderate floor dimensions. (Besides, how do you even generate all these paths?)

Here is an approach based on *dynamic programming*. This general technique is applicable to problems whose solutions can quickly be obtained from the solutions to smaller subproblems of the same structure. The art in dynamic programming is to find the "right" subproblems, and this may require a more or less far-reaching generalization of the original problem.

Once we have identified suitable subproblems, we solve all of them in turn, from the smaller to the larger ones, and memorize the solutions. That way, we have all the information that we need in order to quickly compute the solution to a given subproblem from the solutions of the (already solved) smaller subproblems.

In our case, we generalize the problem as follows: for *all* empty cells C on the floor, compute the *length* of a shortest path from S to C (where the value is ∞ if no such path exists). We claim that this also solves our original problem of computing a shortest path from S to T: Assume that the length of a shortest path from S to T is $\ell < \infty$ (otherwise we know right away that there is no path at all). We also say that T is *reachable* from

S in ℓ steps.

Now if $T \neq S$, there must be a cell adjacent to T that is reachable from S in $\ell - 1$ steps, and adjacent to this a cell reachable in $\ell - 2$ steps etc. Following such a chain of cells until we get to S gives us a path of length ℓ which is shortest possible.

Let us rephrase the generalized problem: we want to label every empty cell C with a nonnegative integer (possibly ∞) that indicates the length of a shortest path from S to C. Here are the subproblems to which we plan to reduce this: for a given integer $i \geq 0$, label all the cells that are reachable from S in at most i steps. For $i = nm - 1$ (actually, for some smaller value), this labels all cells that are reachable from S at all, since a shortest path will never enter a cell twice.

Here is the reduction from larger to smaller subproblems: assume that we have already solved the subproblem for $i - 1$, i.e. we have labeled all cells that are reachable from S within $i - 1$ or less steps. In order to solve the subproblem for i, we still need to label the cells that are reachable in i steps (but not less). But this is simple, since these cells are exactly the unlabeled ones adjacent to cells with label $i - 1$.

Figure 13 illustrates how the frontier of labeled cells grows in this process, for $i = 0, 1, 2, 3$.

Continuing in this fashion, we finally arrive at the situation depicted in Figure 14: all empty cells have been labeled (and are in fact reachable from S in this example). To find a shortest path from S to T, we start from T (which has label 21) and follow any path of decreasing labels $(20, 19, \ldots)$ until we finally reach S.

The shortest path program. Let's get to the C++ implementation of the above method. We represent the floor by a twodimensional array `floor` with dimensions $n+2$ and $m+2$ and entries of type `int`. (Formally, `floor` is a vector of vectors, but we still call this a twodimensional array). These dimensions leave space for extra walls surrounding the floor. Such extra walls allow us to get rid of special cases: floor cells having less than four adjacent cells. In general, an artificial data item that guards the actual data against special cases is called a *sentinel*.

The heart of the program (which appears as Program 2.26 below) is a loop that computes the solution to subproblem i from the solution to subproblem $i - 1$, for $i = 1, 2, \ldots$. The solution to subproblem 0 is readily available: we set the `floor` entry corresponding to S to 0, and the entries corresponding to the empty cells to -1 (this is meant to indicate that the cell has not been labeled yet). Walls are always labeled with the integer -2.

In iteration i of the loop, we simply go through all the yet unlabeled cells and label exactly the ones with i that have an adjacent cell with label $i - 1$. The loop terminates as soon as no progress is made anymore, meaning that no new cell could be labeled in the current iteration. Here is the code.

```
// main loop: find and label cells reachable in i=1,2,... steps
for (int i=1;; ++i) {
  bool progress = false;
```

Figure 13: *The solution to subproblem* i *labels all cells* C *reachable from* S *within at most* i *steps with the length of the shortest path from* S *to* C.

```
  for (int r=1; r<n+1; ++r)
     for (int c=1; c<m+1; ++c) {
      if (floor[r][c] != -1) continue; // wall, or labeled before
      // is any neighbor reachable in i-1 steps?
      if (floor[r-1][c] == i-1 || floor[r+1][c] == i-1 ||
          floor[r][c-1] == i-1 || floor[r][c+1] == i-1 ) {
        floor[r][c] = i; // label cell with i
        progress = true;
      }
    }
  if (!progress) break;
}
```

The other parts of the main function are more or less straightforward. Initially, we read the dimensions from standard input and do the dynamic allocation.

```
// read floor dimensions
```

Figure 14: *The solution to subproblem* $\mathtt{i} = 23$ *solves the generalized problem and the original problem (a shortest path is obtained by starting from* T *and following a path of decreasing labels).*

```
int n; std::cin >> n; // number of rows
int m; std::cin >> m; // number of columns

// define a twodimensional array of dimensions
// (n+2) x (m+2) to hold the floor plus extra walls around
std::vector<std::vector<int> > floor (n+2, std::vector<int>(m+2));
```

Next, we read the floor plan from standard input. We assume that it is given rowwise as a sequence of $\mathtt{n}\mathtt{m}$ characters, where 'S' and 'T' stand for source and target, 'X' represents a wall, and '-' an empty cell. The input file for our initial example from Figure 12 would then look as in Figure 15

```
8 12
------X-----
-XXX--X-----
--SX--------
---X---XXX--
---X---X----
---X---X----
---X---X-T--
-------X----
```

Figure 15: *Input for Program 2.26 corresponding to the example of Figure 12*

If other characters are found in the input (or if the input prematurely becomes empty), we generate empty cells. While reading the floor plan, we put the appropriate integers

into the entries of `floor`, and we remember the target position for later.

```
// target coordinates, set upon reading 'T'
int tr = 0;
int tc = 0;

// assign initial floor values from input:
// source:     'S'  ->     0 (source reached in 0 steps)
// target:     'T'  ->    -1 (number of steps still unknown)
// wall:       'X'  ->    -2
// empty cell: '-'  ->    -1 (number of steps still unknown)
for (int r=1; r<n+1; ++r)
  for (int c=1; c<m+1; ++c) {
    char entry = '-';
    std::cin >> entry;
    if      (entry == 'S') floor[r][c] = 0;
    else if (entry == 'T') floor[tr = r][tc = c] = -1;
    else if (entry == 'X') floor[r][c] = -2;
    else if (entry == '-') floor[r][c] = -1;
  }
```

Now we add the surrounding walls as sentinels.

```
// add surrounding walls
for (int r=0; r<n+2; ++r)
  floor[r][0] = floor[r][m+1] = -2;
for (int c=0; c<m+2; ++c)
  floor[0][c] = floor[n+1][c] = -2;
```

Next comes the main loop that we have already discussed above. It labels all reachable cells, so that we obtain a labeling as in Figure 14. From this labeling, we must now extract the shortest path from S to T. As explained above, this can be done by following a chain of adjacent cells with decreasing labels. For every cell on this path (except S), we put the integer -3 into the corresponding `floor` entry; this allows us to draw the path in the subsequent output. If no path was found (or if there is no target), the body of the `while` statement in the following code fragment is (correctly) not executed at all.

```
// mark shortest path from source to target (if there is one)
int r = tr; int c = tc; // start from target
while (floor[r][c] > 0) {
  const int d = floor[r][c] - 1; // distance one less
  floor[r][c] = -3; // mark cell as being on shortest path
  // go to some neighbor with distance d
  if      (floor[r-1][c] == d) --r;
  else if (floor[r+1][c] == d) ++r;
  else if (floor[r][c-1] == d) --c;
  else                         ++c; // (floor[r][c+1] == d)
}
```

Finally, the output: we map the integer entries of `floor` back to characters, where -3 becomes `'o'`, our path symbol. Inserting `'\n'` at the right places, we obtain a copy of the input floor, with the shortest path appearing in addition.

```
// print floor with shortest path
for (int r=1; r<n+1; ++r) {
  for (int c=1; c<m+1; ++c)
    if      (floor[r][c] == 0)   std::cout << 'S';
    else if (r == tr && c == tc) std::cout << 'T';
    else if (floor[r][c] == -3)  std::cout << 'o';
    else if (floor[r][c] == -2)  std::cout << 'X';
    else                         std::cout << '-';
  std::cout << "\n";
}
```

In case of our initial example, the output looks like in Figure 16. Program 2.26 shows the complete source code.

```
oooooX-----
oXXX-oX-----
ooSX-oooooo-
---X---XXXo-
---X---X-oo-
---X---X-o--
---X---X-T--
-------X----
```

Figure 16: *Output of Program 2.26 on the input of Figure 15*

```
1  #include<iostream>
2  #include<vector>
3
4  int main()
5  {
6    // read floor dimensions
7    int n; std::cin >> n; // number of rows
8    int m; std::cin >> m; // number of columns
9
10   // define a twodimensional array of dimensions
11   // (n+2) x (m+2) to hold the floor plus extra walls around
12   std::vector<std::vector<int> > floor (n+2, std::vector<int>(m+2));
13
14   // target coordinates, set upon reading 'T'
15   int tr = 0;
```

```
int tc = 0;

// assign initial floor values from input:
// source:     'S'  ->     0 (source reached in 0 steps)
// target:     'T'  ->    -1 (number of steps still unknown)
// wall:       'X'  ->    -2
// empty cell: '-'  ->    -1 (number of steps still unknown)
for (int r=1; r<n+1; ++r)
  for (int c=1; c<m+1; ++c) {
    char entry = '-';
    std::cin >> entry;
    if      (entry == 'S') floor[r][c] = 0;
    else if (entry == 'T') floor[tr = r][tc = c] = -1;
    else if (entry == 'X') floor[r][c] = -2;
    else if (entry == '-') floor[r][c] = -1;
  }

// add surrounding walls
for (int r=0; r<n+2; ++r)
  floor[r][0] = floor[r][m+1] = -2;
for (int c=0; c<m+2; ++c)
  floor[0][c] = floor[n+1][c] = -2;

// main loop: find and label cells reachable in i=1,2,... steps
for (int i=1;; ++i) {
  bool progress = false;
  for (int r=1; r<n+1; ++r)
    for (int c=1; c<m+1; ++c) {
      if (floor[r][c] != -1) continue; // wall, or labeled before
      // is any neighbor reachable in i-1 steps?
      if (floor[r-1][c] == i-1 || floor[r+1][c] == i-1 ||
          floor[r][c-1] == i-1 || floor[r][c+1] == i-1 ) {
        floor[r][c] = i; // label cell with i
        progress = true;
      }
    }
  if (!progress) break;
}

// mark shortest path from source to target (if there is one)
int r = tr; int c = tc; // start from target
while (floor[r][c] > 0) {
  const int d = floor[r][c] - 1; // distance one less
  floor[r][c] = -3; // mark cell as being on shortest path
  // go to some neighbor with distance d
```

```
    if      (floor[r-1][c] == d) --r;
    else if (floor[r+1][c] == d) ++r;
    else if (floor[r][c-1] == d) --c;
    else                         ++c; // (floor[r][c+1] == d)
  }

  // print floor with shortest path
  for (int r=1; r<n+1; ++r) {
    for (int c=1; c<m+1; ++c)
      if      (floor[r][c] == 0)   std::cout << 'S';
      else if (r == tr && c == tc) std::cout << 'T';
      else if (floor[r][c] == -3)  std::cout << 'o';
      else if (floor[r][c] == -2)  std::cout << 'X';
      else                         std::cout << '-';
    std::cout << "\n";
  }

  return 0;
}
```

Program 2.26: *../progs/lecture/shortest_path.cpp*

2.7.8 Details

Command line arguments. In Program 2.25 for string matching, it is not very convenient that the search string is *fixed*. We then have to recompile the program every time we want to search for another string.

A more flexible alternative is to pass the search string as a *command line argument* that we provide upon calling the program.

The main function can access such command line arguments if we provide suitable parameters. Here is how the first lines of Program 2.25 have to be changed in order to make this work.

```
// Program: string_matching2.cpp
// find the first occurrence of a string (provided as command
// line argument) within the input text, and output text so far

#include<iostream>
#include<ios> // for std::noskipws

int main (int argc, char* argv[])
{
  if (argc < 2) {
    // no command line arguments (except program name)
    std::cout << "Usage: string_matching2 <string>\n";
```

```
    return 1;
  }

  // search pattern
  const std::string pattern = argv[1];
```

The values of argc and and argv[] (which is an array of strings) are initialized by the operating system when it calls the main function. We will explain function parameters in detail later, here we will be satisfied with an example. Suppose that we call the program like this (assuming a Unix-type system):

```
./string_matching2 bool
```

Then argc (which counts the number of command line arguments) gets initialized with value 2. This count includes the program name itself ("string_matching2" in this case), and any additional strings provided on the command line (just the single string "bool" in this case). The 2 arrays argv[0] and argv[1] get initialized with the strings "string_matching2" and "bool" as described in Section 2.7.6 above. Consequently, after its definition, the variable pattern in the above piece of code holds the string "bool". This gets us back to the situation in Program 2.25, and the remainders of both programs are identical.

2.7.9 Goals

Dispositional. At this point, you should ...

1) know what an array is, and what random access means in static arrays and vectors
2) know that characters and strings (arrays of characters with added functionality) can be used to perform basic text processing tasks;
3) know that (multidimensional) arrays of variable length can be obtained through vectors of vectors (of vectors...)

Operational. In particular, you should be able to ...

(G1) write programs that define static array variables or vector variables;

(G2) write programs that perform simple data processing tasks by using random access in (multidimensional) arrays as the major tool;

(G3) write programs that perform simple text processing tasks with characters and strings

2.7.10 Exercises

Exercise 91 *Let us call a natural number* k-composite *if and only if it is divisible by exactly* k *different prime numbers. For example, prime powers are 1-composite, and*

$6 = 2 \cdot 3$ *as well as* $20 = 2 \cdot 2 \cdot 5$ *are 2-composite. Write a program* `k_composite.cpp` *that reads numbers* $n \geq 0$ *and* $k \geq 0$ *from the input and then outputs all* k-*composite numbers in* $\{2, \ldots, n-1\}$. *How many 7-composite numbers are there for* $n = 1,000,000$*?* (G1)(G2)

Exercise 92 *Write a program* `inverse_matrix.cpp` *that inverts a* 3×3 *matrix* A *with real entries. The program should read the nine matrix entries from the input, and then output the inverse matrix* A^{-1} *(or the information that the matrix* A *is not invertible). In addition, the program should output the matrix* AA^{-1} *in order to let the user check whether the computation of the inverse was accurate (in the fully accurate case, the latter product is the identity matrix).*

Hint: *For the computation of the inverse, you can employ* Cramer's rule. *Applied to the computation of the inverse, it yields that* A^{-1}_{ij} *(the entry of* A^{-1} *in row* i *and column* j*) is given by*

$$A^{-1}_{ij} = \frac{(-1)^{i+j} \det(A^{ji})}{\det(A)},$$

where $\det(M)$ *is the determinant of a square matrix* M*, and* A^{ji} *is the* 2×2 *matrix obtained from* A *by deleting row* j *and column* i.

To compute the determinant of a 3×3 *matrix, you might want to use the well-known* Sarrus' rule. (G1)(G2)

Exercise 93 *Write a program* `read_array` *that reads a sequence of* n *integers from standard input into an array (realized by a vector). The number* n *is the first input, and then the program expects you to input another* n *values. After reading the* n *values, the program should output them in the same order. (If you can do this, you have proven that you are no longer a complete novice, according to Stroustrup.) For example, on input* `5 4 3 6 1 2` *the program should output* `4 3 6 1 2`. (G1)(G2)

Exercise 94 *Enhance the program* `read_array.cpp` *from Exercise 93 so that the resulting program* `sort_array.cpp` *sorts the array elements into ascending order before outputting them. Your sorting algorithm does not have to be particularly efficient, the main thing here is that it works correctly. Test your program on some larger inputs (preferably read from a file, after redirecting standard input). For example, on input* `5 4 3 6 1 2` *the program should output* `1 2 3 4 6`. (G1)(G2)

Exercise 95 *Consider the string matching algorithm of Program 2.25. Prove that for all* $m > 1, n \geq m$*, there exists a search string* s *of length* m *and a text* t *of length* n *on which the algorithm in Program 2.25 performs* $m(n - m + 1)$ *comparisons between single characters.* (G3)

Exercise 96 *Write a program* `frequencies.cpp` *that reads a text from standard input (like in Program 2.25) and outputs the frequencies of the letters in the text, where*

we do not distinguish between lower and upper case letters. For this exercise, you may assume that the type `char` *implements ASCII encoding. This means that all characters have integer values in* $\{0, 1, \ldots, 127\}$*. Moreover, in ASCII, the values of the 26 upper case literals* `'A'` *up to* `'Z'` *are consecutive numbers in* $\{65, \ldots, 90\}$*; for the lower case literals* `'a'` *up to* `'z'`*, the value range is* $\{97, \ldots, 122\}$*.* (G3)

Running this on the lyrics of *Yesterday* (The Beatles) for example should yield the following output.

```
Frequencies:
a:      45 of 520
b:      5 of 520
c:      5 of 520
d:      28 of 520
e:      65 of 520
f:      4 of 520
g:      13 of 520
h:      27 of 520
i:      27 of 520
j:      0 of 520
k:      3 of 520
l:      20 of 520
m:      10 of 520
n:      30 of 520
o:      43 of 520
p:      4 of 520
q:      0 of 520
r:      19 of 520
s:      36 of 520
t:      31 of 520
u:      9 of 520
v:      6 of 520
w:      19 of 520
x:      0 of 520
y:      34 of 520
z:      0 of 520
Other:  37 of 520
```

2.7.11 Challenges

Exercise 97 *For larger floors, Program 2.26 can become quite inefficient, since every step* i *examines* all *cells of the floor in order to find the (possibly very few) ones that have to be labeled with* i *in that step. A better solution would be to examine only the neighbors of the cells that are already labeled with* $i - 1$*, since only these are candidates for getting label* i*.*

Write a program `shortest_path_fast.cpp` *that realizes this idea, and measure the performance gain on some larger floors of your choice.*

Exercise 98 *In 1772, Leonhard Euler discovered the quadratic polynomial*

$$n^2 + n + 41$$

with the following remarkable property: if you evaluate it for $n = 0, 1, \ldots, 39$*, you always get a prime number, and moreover, all these prime numbers are different. Here is the list of all the* 40 *prime numbers generated by Euler's polynomial:*

41, 43, 47, 53, 61, 71, 83, 97, 113, 131, 151, 173, 197, 223, 251, 281, 313, 347, 383, 421, 461, 503, 547, 593, 641, 691, 743, 797, 853, 911, 971, 1033, 1097, 1163, 1231, 1301, 1373, 1447, 1523, 1601.

Here we are concerned with the question whether there are still better quadratic polynomials in the sense that they generate even more prime numbers. We say that a quadratic polynomial $an^2 + bn + c$ *has* Euler quality p *if the* p *numbers*

$$|an^2 + bn + c|, \quad n = 0, ..., p - 1$$

are different prime numbers. By taking absolute values, we therefore also allow "negative primes". As an example, let us look at the polynomial $n^2 - 10n + 2$*. For* $n = 0$*, we get* 2 *(prime), for* $n = 1$ *we obtain* -7 *(negative prime), and* $n = 2$ *gives* -14 *(no (negative) prime). The Euler quality of* $n^2 - 10n + 2$ *is therefore* 2 *but not higher.*

Here is the challenge: write a program that systematically searches for a quadratic polynomial with high Euler quality. The goal is to find a polynomial with Euler quality larger than 40*, in order to "beat"* $n^2 + n + 41$*. What is the highest Euler quality that you can find?*

Exercise 99 *The* `XBM` *file format is a format for storing monochrome (black & white) images. The format is somewhat outdated, but many browsers (Internet Explorer is a notable exception) can still display images in* `XBM` *format.*

An `XBM` *image file for an image named* `test` *might look like this (taken from Wikipedia's* `XBM` *page).*

```
#define test_width 16
#define test_height 7
static unsigned char test_bits[] = {
  0x13, 0x00, 0x15, 0x00, 0x93, 0xcd, 0x55,
  0xa5, 0x93, 0xc5, 0x00, 0x80, 0x00, 0x60};
```

As you can guess from this, `XBM` *files are designed to be integrated into C and C++ source code which makes it easy to process them (there is no need to read in the data; simply include the file from the C++ program that needs to process the image). In our example,* `test_width` *and* `test_height` *denote the width and height of the image in pixels. Formally, these names are* macros*, but in the program they can be used like constant expressions.* `test_bits` *is a static array of characters that encodes the colors of the* 16×7 *pixels in the image. Every* hexadecimal literal *of the form* $0xd_1d_2$ *encodes eight pixels, where the order is row by row. In our case,* `0x13` *and* `0x00` *encode the 16 pixels of the first row, while* `0x15` *and* `0x00` *are for the second row, etc.*

Here is how a two-digit hexadecimal literal encodes the colors of eight consecutive pixels within a row. If the width is not a multiple of 8, the superfluous color values from the last hexadecimal literal of each row are being ignored.

Every hexadecimal digit d_i *is from the set* $\{0, ..., 9, a, \ldots, f\}$ *where* a *up to* f *stand for* $10, \ldots, 15$*. The actual number encoded by a hexadecimal literal is* $16d_1 + d_2 \in \{0, \ldots, 255\}$*. If the type* `char` *has value range* $\{-128, \ldots, 127\}$*, the silent assumption is that a literal value* α *larger than 127 converts to* $\alpha - 256$*, which has the same representation under two's complement. For example,* `0x13` *has value* $1 \cdot 16 + 3 = 19$*.*

Now, any number in $\{0, \ldots, 255\}$ *has a binary representation with* 8 *bits.* 19*, for example, has binary representation* 00010011*. The pixel colors are obtained by reading this backwards, and interpreting* 1 *as black and* 0 *as white. Thus, the first eight pixels in row* 1 *of the* `test` *image are black, black, white, white, black, white, white, white. The complete* `test` *image looks like this:*

Write a program `xbm.cpp` *that* `#include`*s an* `XBM` *file of your choice (you may search the web to find suitable* `XBM` *files, or you may use an image manipulation program such as* `gimp` *to convert your favorite image into* `XBM` *format), and that outputs an* `XBM` *file for the same image,* rotated by 90 degrees. *The program may write the resulting file to standard output. In case of the* `test` *image, the resulting* `XBM` *file and the resulting rotated image are as follows.*

```
#define rotated_width 7
#define rotated_height 16
static unsigned char rotated_bits[] =
   0x3c, 0x54, 0x48, 0x00,
   0x04, 0x1c, 0x00, 0x1c,
   0x14, 0x08, 0x00, 0x1f,
   0x00, 0x0a, 0x15, 0x1f};
```

Note that we now have 16 instead of 14 hexadecimal literals. This is due to the fact that each of the 16 rows needs one literal for its 7 pixels, where the leading bits of the binary representations are being ignored.

You may extend your program to perform other kinds of image processing tasks of your choice. Examples include color inversion (replace black with white, and vice versa), computing a mirror image, scaling the image (so that it occupies less or more pixels), etc.

2.8 Pointers, Algorithms, Iterators and Containers

These boots are made for walking, and that's just what they'll do.
One of these days these boots are gonna walk all over you.

Nancy Sinatra (1966)

You will learn how to use static arrays in functions. This requires the concept of pointers to represent addresses of array elements. You will see how pointers are used by standard algorithms for processing arrays. You will learn about vector iterators as "pointers" for vectors that allow you to also process vectors in a uniform way, using functions and standard algorithms. Finally, you will understand that static arrays and vectors are special cases of containers, and that each container comes with its own iterator type for processing the container in a uniform way that is independent from the specifics of the container itself.

2.8.1 Arrays and functions

So far we have carefully avoided to write functions that do things with arrays. The reason is that it's not so easy to understand how and why this works. As a concrete example, let us consider the following program that defines and uses a function to set all elements of an array to a fixed value.

```
// Program: fill.cpp
// define and use a function to fill an array
#include<iostream>

// PRE:  a[0],...,a[n-1] are elements of an array
// POST: a[i] is set to value, for 0 <= i < n
void fill_n (int a[], const int n, const int value) {
  for (int i=0; i<n; ++i)
    a[i] = value;
}

int main()
{
  int a[5];
  fill_n (a, 5, 1);
  for (int i=0; i<5; ++i)
    std::cout << a[i] << " ";  // 1 1 1 1 1
  std::cout << "\n";
```

```

  return 0;
}
```

Program 2.27: *../progs/lecture/fill.cpp*

If you have digested what we said about arrays and functions so far, you should now at least scratch your head. If you don't, then try to answer these two questions for yourself first:

Question 1: Since static arrays cannot be initialized from other arrays—as we have explicitly seen on Page 163—how can a static array be passed as an argument to a function? The usual first step in a function call evaluation (the call arguments are evaluated, and their values are used to initialize the formal arguments) won't work with static arrays. On top of that, the formal argument `int a[]` seems to have incomplete type, according to Section 2.7.2, since it does not specify the length of the array to be passed as a call argument.

Question 2: Even if C++ contains some trickery to allow static arrays as function arguments, how is it possible that the function call changes the value of the array by assigning new values to its elements? On Page 140, we have explicitly made the point that function call arguments are rvalues, so any changes made to the formal arguments within the function body should be "lost", right?

To answer Question 1 right away: the function `fill_n` only *appears* to have a static array argument. The compiler, however, internally *adjusts* this to the completely equivalent declaration

```
// PRE:  a[0],...,a[n-1] are elements of an array
// POST: a[i] is set to value, for 0 <= i < n
void fill_n (int* a, const int n, const int value);
```

where `int*` is a *pointer type* that we will discuss next. The same adjustment would happen for the formal argument `int a[5]`, say, meaning that the array length is ignored. You could in fact (legally, but quite confusingly) have a formal argument `int a[5]`, and then use a static array of length 10 as call argument.

The moral is that in reality, no formal arguments of static array type exist, and in order to avoid confusion, it is better not to pretend otherwise.

2.8.2 Pointer types and functionality

For every type T the corresponding *pointer type* is

$T*$

We call T the underlying type of T*. An expression of type T* is called a *pointer* (to T).

The value of a pointer to T is the address of (the first memory cell of) an object of type T. We call this the object *pointed to* by the pointer.

We can visualize a pointer p as an arrow pointing to a cell in the computer's main memory—the cell where the object pointed to starts, see Figure 17.

Figure 17: *A pointer to T represents the address of an object of type T in the computer's main memory.*

Initialization, the assignment operator =, and the comparison operators == and != are defined for every pointer type T*. The latter simply test whether the addresses in question are the same or not.

Initialization and assignment copy the value (as usual), which in this case means to copy an address; thus, if j points to some object, the assignment i = j has the effect that i now also points to this object. The object itself is not copied. We remark that pointer initialization and assignment require the types of both operands to be exactly the same—implicit conversions don't work. If you think about it, this is clear. Imagine that p is of type int*, pointing to some int value, and now you write

```
int* i = p;      // ok
double* j = i;   // error: cannot convert
```

Since double objects usually require more memory cells than int objects, j would be a pointer to a double object that includes memory cells originally not belonging to i. This can hardly be called a "conversion". In fact, since we only copy an address, there cannot be any physical conversion of the stored value, even if the memory requirements of the two types happen to be the same.

2.8.3 Array-to-pointer conversion

Any static array of type $T[n]$ can implicitly be converted to type T*. The resulting value is the address of the first element of the array. For example, we can write

```
int a[5];
int* begin = a; // begin points to a[0]
```

The array-to-pointer conversion is trivial on the machine side. For this, we recall from our earlier discussion in Section 2.7.4 that in C and therefore also in C++, the address of the first element is part of a static array's internal representation. It is important

to understand, though, that the *length* of the array gets lost during array-to-pointer conversion: the resulting pointer is just a pointer, and nothing else.

Array-to-pointer conversion automatically takes place when a static array appears in an expression. The only exception are expressions of the form `sizeof(a)`, where a is a static array. Bjarne Stroustrup, the designer of C++, illustrates this by saying that *the name of an array converts to a pointer to its first element at the slightest provocation*. In still other words, there are no operations on static arrays, except the length computations through `sizeof`: everything that we conceptually do with a static array is in reality done with a pointer to its first element; this in particular applies to the random access operation, see the paragraph called "Pointer subscripting, or the truth about random access" in Section 2.8.5 below.

Coming back to Program 2.27 above, this also explains why the two declarations

```
void fill_n (int a[], const int n, const int value)
```

and

```
void fill_n (int* a, const int n, const int value)
```

are equivalent. C++ allows us to use formal arguments of static array type, but in actually passing the array, it anyway converts to a pointer to its first element. Hence we can as well use a formal argument of pointer type `int*`.

2.8.4 Objects and addresses

The address operator. We can obtain a pointer to a given object by applying the unary *address operator* to an lvalue that refers to the object. If the lvalue is of type T, then the result is an rvalue of type T*. The syntax of an address operator call is

&lvalue

In the following code fragment we use the address operator to initialize a variable `iptr` of type `int*` with the address of an object of type `int` named `i`.

```
int i = 5;
int* iptr = &i; // iptr initialized with the address of i
```

The dereference operator. From a pointer, we can get back to the object pointed to through *dereferencing* or *indirection*. The unary *dereference operator* $*$ applied to an rvalue of pointer type yields an lvalue referring to the object pointed to. If the rvalue is of type T*, then the result is of type T. The syntax of a dereference operator call is

**rvalue*

Following up on our previous code fragment, we can therefore write

```
int i = 5;
int* iptr = &i; // iptr initialized with the address of i
int j = *iptr;  // j == 5
```

The naming scheme of pointer types is motivated by the dereference operator. The declaration

$T*\ p$

can also be read (and in fact legally be written; we don't do this, though) as

$T\ *p$

The second version implicitly defines the type of p by saying that $*p$ is of type T. This is the same kind of implicit definition that we already know from array declarations.

Figure 18 illustrates address and dereference operator.

Figure 18: *The address operator (left) and its inverse, the dereference operator (right)*

The null pointer. For every pointer type there is a value distinguishable from any other pointer value. This value is called the *null pointer value*. The integer value 0 can be converted to every pointer type, and the value after conversion is the null pointer value. In the declaration `int* iptr = 0`, for example, the variable `iptr` gets initialized with the null pointer value. We also say that `iptr` is a *null pointer*. The null pointer value must not be dereferenced, since it does not correspond to any existing address.

Using the null pointer value is the safe way of indicating that there is no object (yet) to point to. The alternative of leaving the pointer uninitialized is bad: there is no way of testing whether a pointer that is not a null pointer holds the address of a legitimate object, or whether it holds some "random" address resulting from leaving the pointer uninitialized.

In the latter case, dereferencing the pointer usually crashes the program. Consider this code:

```
int* iptr;      // uninitialized pointer
int j = *iptr; // trouble!
```

After its declaration, the pointer `iptr` has undefined value, which in practice means that it may correspond to an arbitrary address in memory; dereferencing it means to access the memory content at this address. In general, this address will not belong to the part of memory to which the program has access; the operating system will then deny access to it and terminate the program, typically with a very brief error message such as *segmentation fault* or *access violation*.

2.8.5 Pointer arithmetic

It's time to get back to Question 2 on Page 186, asking why and how it is possible for the function `fill_n` in Program 2.27 to change the value of its call argument array. The situation has become even more obscure, now that we know that the formal argument a is in reality not even a static array but a pointer to the first element of a static array; in this case, what is `a[i]` supposed to mean?

In order to understand this, we have to understand *pointer arithmetic*, the art of computing with addresses. We deliberately call this an "art", since pointer arithmetic comes with a lot of pitfalls, but without a safety net. On the other hand, the authors feel that there is also a certain beauty in the minimalism of pointer arithmetic. It's like driving an oldtimer: it's loud, it's difficult to steer, seats are uncomfortable, and there's no heating. But the heck with it! The oldtimer looks so much better than a modern car. Nevertheless, after driving the oldtimer for a while, it will probably turn out that beauty is not enough, and that safety and usability are more important factors in the long run. Still, if you want to understand the essentials of how a car works, it's better to start with the oldtimer.

Adding integers to pointers. The binary addition operators +, − are defined for left operands of any pointer type *T** and right operands of any integral type. Recall that if an array is provided as the left operand, it will implicitly be converted to a pointer using array-to-pointer conversion.

For the behavior of + to be defined, there must be an array of some length n, such that the left operand *ptr* is a pointer to the element of some index $k, 0 \leq k \leq n$, in the array. The case $k = n$ is allowed and corresponds to the situation where *ptr* is a pointer one past the last element of the array (we call this a *past-the-end pointer*; note that such a pointer must not be dereferenced).

If the second operand *expr* has some value i such that $0 \leq k + i \leq n$, then

ptr + *expr*

is a pointer to the $(k + i)$-th element of the same array. Informally, we get a pointer that has been moved "i elements to the right" (which actually means to the left if i is

negative). Therefore, if p is the value of `ptr` (an address), then the value of `ptr + expr` is the address $p + si$, assuming that a value of the underlying type occupies s memory cells. The pleasing fact is that we don't have to care about s; the operation `ptr + expr` (which knows s from the type of `ptr`) does this for us and offers a type-independent way of moving a pointer i elements to the right.

As before, if $k + i = n$, we get a past-the-end pointer. Values of i such that $k + i$ is not between 0 and n lead to undefined behavior.

Let us repeat the point that we have made before in connection with random access in Section 2.7.3: by default, there are absolutely no checks that the above requirements indeed hold, and it is entirely your responsibility to make sure that this is the case. Failure to do so will result in program crashes, strange behavior of the program, or (probably the worst scenario) seemingly normal behavior, but with the potential of turning into strange behavior at any time, or on any other machine.

Therefore, let us summarize the requirements once more:

- *ptr* must point to the element of index k in some array of length n, where $0 \leq k \leq n$, and
- *expr* must have some value i such that $0 \leq k + i \leq n$.

Pointer values are not integers. The discussion here might suggest that pointer values (addresses) are just integers, and that you can compute with them accordingly. But this is not true. On a high level, we *do* consider an address as an integer (see Section 1.2.3), but adding 1 to it using the + operator from the previous paragraph doesn't give us the next larger integer, it gives us the address of the next element in the array. What the internal integer representation of this address is depends on the memory requirements of the type underlying the array, plus a number of other machine-dependent factors. We should therefore never think of an address as an absolute integer whose value tells us anything useful. In particular, pointer values and integers cannot be converted into each other, since such a conversion would be meaningless:

```
int* ptr = 5; // error: invalid conversion from int to int*
int a = ptr;  // error: invalid conversion from int* to int
```

Pointer subscripting, or the truth about random access. In reality, the subscript operator [] as introduced in Section 2.7.3 does not operate on arrays, but on pointers: Invoking this operator on a static array constructs an expression and therefore triggers an array-to-pointer conversion.

Given a pointer *ptr* and an expression *expr* of integral type, the expression

ptr[*expr*]

is equivalent (also in its requirements on *ptr* and *expr*) to

$*(ptr + expr)$

If *expr* has value i, the latter expression yields the array element i places to the right of the one pointed to by *ptr*. In particular, if *ptr* results from an array-to-pointer conversion, this agrees with the semantics of random access for arrays as introduced in Section 2.7.3. This finally explains why in the function `fill_n` of Program 2.27, the loop

```
for (int i=0; i<n; ++i)
  a[i] = value;
```

sets all array elements to `value`. An equivalent version of the loop would be

```
for (int i=0; i<n; ++i)
  *(a+i) = value;
```

Since `a` is a pointer to the first element of the original array (the one of index 0), `a+i` is a pointer to the element of index i. Dereferencing this pointer through `*(a+i)` yields an lvalue for the element of index i itself. Assigning `value` to it thus changes the value of the array element with index i.

Subtracting integers from pointers. Binary subtraction is similar. If *expr* has value i such that $0 \leq k - i \leq n$, then

$$\boxed{ptr - expr}$$

yields a pointer to the array element of index $k - i$.

The assignment versions += and −= of the two operators can be used with left operands of pointer type as well, with the usual meaning. Similarly, the unary increment and decrement operators ++ and −− are available for pointers. Since precedences and associativities are tied to the operator symbols, they are as in Table 1 on page 51.

Pointer comparison. We have already discussed the relational operators == and != that simply test whether the two pointers in question point to the same object. But we can also compare two pointers using the operators <, <=, >, and >=. Again, precedences and associativities of all relational operators are as in Table 2 on page 72.

For the result to be specified, there must be an array of some length n, such that the left operand *ptr1* is a pointer to the element of some index $k_1, 0 \leq k_1 \leq n$ in the array, and the second operand *ptr2* is a pointer to the element of some index $k_2, 0 \leq k_2 \leq n$ in the same array. Again, $k_1 = n$ and $k_2 = n$ are allowed and correspond to the past-the-end case.

Given this, the result of the pointer comparison is determined by the integer comparison of k_1 and k_2. In other words (and quite intuitively), the pointer to the element that comes first in the array is the smaller one.

Comparing two pointers that do not meet the above requirements leads to unspecified results in the four operators <, <=, >, and >=.

Pointer subtraction. There is one more arithmetic operation on pointers. Assume that *ptr1* is a pointer to the element of some index $k_1, 0 \leq k_1 \leq n$ in some array of length n, and the second operand *ptr2* is a pointer to the element of some index $k_2, 0 \leq k_2 \leq n$ in the same array (past-the-end pointers allowed). Then the result of the *pointer subtraction*

ptr1 − ptr2

is the integer $k_1 - k_2$. Thus, pointer subtraction tells us "how far apart" the two array elements are. The behavior of pointer subtraction is undefined if *ptr1* and *ptr2* are not pointers to elements in (or past-the-end pointers of) the same array.

A typical use of pointer subtraction is to determine the number of elements in an array that is given by a pointer to its first element and a past-the-end pointer. Table 4 summarizes the new pointer-specific binary operators.

Description	Operator	Arity	Prec.	Assoc.
subscript	[]	2	17	left
dereference	*	1	16	right
address	&	1	16	right

Table 4: *Precedences and associativities of pointer operators. The subscript operator expects rvalues as operands and returns an lvalue. The dereference operator expects an rvalue and returns an lvalue, while the address operator expects an lvalue and returns an rvalue.*

2.8.6 Traversing arrays

Let's look at the `fill_n` function from Program 2.27 again:

```
void fill_n (int a[], const int n, const int value) {
  for (int i=0; i<n; ++i)
    a[i] = value;
```

Its main functionality is to traverse the elements of the array a, and to do something with each element. Given what we have learned about pointers and pointer arithmetic above, this kind of traversal might now strike us as somewhat unnatural. Here's why.

Random access. We have seen that the random access functionality of arrays is internally based on address arithmetic. If p denotes the address of the first element of a, and s denotes the number of memory cells occupied by a value of type `int`, then during the above loop, the following addresses are computed: $p, p + s, p + 2s, \ldots, p + (n - 1)s$.

This requires one multiplication and one addition for every address except the first. But if you think about it, the multiplication only comes in because we compute each

address from scratch, independently from the previous ones. In fact, the same set of addresses could more efficiently and more naturally be computed by starting with p and repeatedly adding s ("going to the next element"; this is what we call *iteration*).

Using random access, we merely *simulate* iteration over an array, but we are missing the operation of "going to the next element". The following analogy illustrates the point: you *can* of course read a book by starting with page 1, then closing the book, opening it again on pages $2-3$, closing it, opening it on pages $4-5$, etc. But unless you're somewhat eccentric, you probably prefer to just turn the pages in between.

Iteration. Well, we *do* know how to go to the next element after the one pointed to by `ptr`: it's the one pointed to by `ptr+1`; in fact, we can change `ptr` to point to the next element simply through the expression `++ptr`. Internally, this just adds s to the address represented by `ptr` and hence corresponds to "turning the page".

Let us rewrite our `fill_n` function accordingly by making the loop advance a pointer `ptr` instead of an index `i`. It seems that we still need the index to check for termination (`i<n`), but we can also get rid of this. The standard way to "get an array into a function" is to pass *two* pointers, one to the first element, and a past-the-end pointer (Page 190). This also uniquely describes the array. We may actually process a contiguous *subarray* of the original array. Such a subarray is described by a *range* `[begin, end)`, where the halfopen interval notation `[begin, end)` means that the range is given by the values `begin`, `begin+1`, ...,`end-1`. A *valid range* contains the addresses of a set of consecutive array elements, where `[begin, begin)` is an empty range.

Here is our updated program to fill an array. The fill function is now simply called `fill`, since there is no explicit length argument anymore.

```
// Program: fill2.cpp
// define and use a function to fill a range
#include<iostream>

// PRE:  [begin, end) is a valid range
// POST: *p is set to value, for p in [begin, end)
void fill (int* begin, int* end, const int value) {
  for (int* ptr = begin; ptr != end; ++ptr)
    *ptr = value;
}

int main()
{
  int a[5];
  fill (a, a+5, 1);
  for (int i=0; i<5; ++i)
    std::cout << a[i] << " ";  // 1 1 1 1 1
  std::cout << "\n";
```

```

  return 0;
}
```

Program 2.28: *../progs/lecture/fill2.cpp*

To fill an array a with 5 elements, we use the statement

```
fill (a, a+5, 1);
```

We know that a automatically converts to a pointer to the first element of the array, which is used to initialize the begin pointer. a+5 on the other hand, points to *after* the last element of a (which has index 4). Hence, a+5 is a past-the-end pointer that is used to initialize the end pointer. Let's take a closer look at the main loop

```
for (int* ptr = begin; ptr != end; ++ptr)
  *ptr = value;
```

Starting with a pointer to the first element described by the range (ptr = begin), the element pointed to is set to the desired value (*ptr = value). Then we increment ptr so that it points to the next element (++ptr). We repeat this as long as ptr is different from the past-the-end pointer end. Thus, the loop sets all elements described by the range [begin, end) to the desired value.

2.8.7 Const pointers

There is a substantial difference between the function pow on the one hand (see Program 2.14 on Page 134), and the function fill from Program 2.28 on the other hand. A call to the function pow has no effect, since the computations only modify formal argument values; these values are "local" to the function call and are lost upon termination. With pow, it's the *value* of a function call that we are interested in.

Calls to the function fill, on the other hand, have effects: they modify the values of array elements, and these values are *not* local to the function call. When we write

```
int a[5];
fill (a, a+5, 1);
```

the effect of the expression fill (a, a+5, 1) is that all elements of a receive value 1. This is possible since there are formal arguments of pointer type. When the function call fill (a, a+5, 1) is evaluated, the formal argument begin is initialized with the *address* of a's first element. In the function body, the value at this address is modified through the lvalue *ptr, and the same happens for the other four array elements in turn.

Formal arguments of pointer type are therefore a means of constructing functions with value-modifying effects. Such functions are called *mutating*.

Of course, there are also functions on arrays that are non-mutating, and in this case, const-correctness requires us to document this in the types of the formal arguments. As an example, consider the following function for finding the minimum integer described by a nonempty range.

```
// PRE:  [begin, end) is a valid and nonempty range
// POST: the smallest of the values described by [begin, end)
//       is returned
int min (const int* begin, const int* end)
{
  assert (begin != end);
  int m = *begin++; // current candidate for the minimum
  for (const int *ptr = begin; ptr != end; ++ptr)
    if (*ptr < m) m = *ptr;
  return m;
}
```

As the function min does not intend to modify the values of the objects described by the range [begin, end), the respective pointers have *const-qualified* underlying type const int and are called *const pointers*.

For all formal arguments of *non*-pointer type that we know so far, the const keyword makes no difference outside the function body, since the values of the call arguments cannot be changed by the function anyway. For formal arguments of pointer type, the difference is essential and lets us distinguish mutating from non-mutating functions.

Const-qualified types should be used as underlying types of arrays (and pointers to array elements) *if and only if* the program does not intend to modify the values of the array elements.

What exactly is constant? With a pointer p of type const T*, we cannot achieve "absolute constness" of the object pointed to by p. We only achieve that the object can neither directly nor indirectly be modified through the pointer p. But we *could* modify it through another pointer. Consider the code fragment

```
int a[5];
const int* begin1 = a;
int*       begin2 = a;
*begin1 = 1;                    // error
*begin2 = 1;                    // ok
```

Here, begin1 and begin2 are two pointers to the same object, namely the first element of the array a. With respect to the pointer begin1, this element is constant, but with respect to begin2, it is not.

2.8.8 Algorithms

For many routine tasks, the C++ standard library already contains solutions. One such task is the one we have prominently discussed before: filling an array with a specified value. There is no need to write code for such routine tasks ourselves. For filling an array, we could instead use an *algorithm* from the standard library. The benefit is that we eliminate possible sources of error (even a trivial loop has the potential of being wrong),

and that we simplify the control flow (see also Section 2.4.8). Here is the corresponding program.

```
// Program: fill3.cpp
// use a standard algorithm to fill a range
#include<iostream>
#include<algorithm> // needed for std::fill

int main()
{
  int a[5];
  std::fill (a, a+5, 1);
  for (int i=0; i<5; ++i)
    std::cout << a[i] << " ";  // 1 1 1 1 1
  std::cout << "\n";

  return 0;
}
```

Program 2.29: *../progs/lecture/fill3.cpp*

In this program, we have replaced the call of our handwritten `fill` function with a call of `std::fill`, and we see that both functions are called in exactly the same way. If you haven't believed it so far, you now see that processing static arrays through ranges is indeed the standard way of processing static arrays in C++.

It may sound a bit pompous to use the term "algorithm" for the obvious recipe of filling an array with a specified value, but that's how it is with the standard library. In all fairness it should be said, though, that there are also more elaborate algorithms around, for example for sorting arrays.

One important feature of algorithms such as `std::fill` is that they do not only work for static arrays with underlying type `int`, but for *all* underlying types. Obviously, the recipe of filling the array is the same for all underlying types, and this reflects in the usage. For example, the following program is exactly the same as Program 2.29, exept that it fills a static array of underlying type `std::string`.

```
// Program: fill4.cpp
// use a standard algorithm to fill a range
#include<iostream>
#include<string>
#include<algorithm> // needed for std::fill

int main()
{
  std::string a[5];
  std::fill (a, a+5, "bla");
```

```
  for (int i=0; i<5; ++i)
    std::cout << a[i] << " ";  // bla bla bla bla bla
  std::cout << "\n";

  return 0;
}
```

Program 2.30: *../progs/lecture/fill4.cpp*

Again, you might scratch your head about this, if you have carefully followed Section 2.6. What are actually the formal argument types of `std::fill`? How can it be possible to call the same function with an integer on the one hand, and a string on the other hand?

Function templates. The full picture is beyond the scope of this book, but in a nutshell, `std::fill` is not a conventional function but a function *template*. This means that the definition of `std::fill` also has a type argument, where the compiler simply deduces the appropriate type from the call arguments. We can even show a poor man's version of this by turning our handwritten version of `fill` into a function template that works for static arrays of all underlying types. Here is the program.

```
// Program: fill5.cpp
// define and use a function template to fill a range
#include<iostream>
#include<string>

// PRE:  [begin, end) is a valid range
// POST: *p is set to value, for p in [begin, end)
template <typename T>
void fill (T* begin, T* end, const T value) {
  for (T* ptr = begin; ptr != end; ++ptr)
    *ptr = value;
}

int main()
{
  int a[5];
  fill (a, a+5, 1);
  for (int i=0; i<5; ++i)
    std::cout << a[i] << " ";  // 1 1 1 1 1
  std::cout << "\n";

  std::string b[5];
  fill (b, b+5, "bla");
  for (int i=0; i<5; ++i)
```

```
    std::cout << b[i] << " ";  // bla bla bla bla bla
  std::cout << "\n";

  return 0;
}
```

Program 2.31: *../progs/lecture/fill5.cpp*

The angle brackets are familiar from `std::vector<int>`, and this is not a coincidence. In fact, `std::vector` is a *class template*, i.e. it also has a type argument to specify the type of the elements it stores.

An algorithm such as `std::fill` that works in the same way for different types is called *generic*, but the authors are afraid that it's not possible here to be less generic on the issue of generic algorithms.

2.8.9 Vector iterators

Now suppose that we want to fill a *vector* with a specified value, and not a static array. Our poor man's version of a generic fill algorithm depicted in Program 2.31 is not generic enough for this purpose (and that's exactly why it's only a poor man's version).

The problem is that we cannot iterate over a vector with pointers. Pointers are tied to static arrays. On the other hand, a vector was designed to be a luxury edition of the primitive static array, so it would be a shame if `std::fill` wouldn't work for vectors. And indeed, it does work. The following program shows how, where we afterwards explain in detail what is going on.

```
// Program: fill6.cpp
// use a standard algorithm to fill a vector
#include<iostream>
#include<vector>
#include<algorithm> // needed for std::fill

int main()
{
  std::vector<int> v(5);
  std::fill (v.begin(), v.end(), 1);
  for (int i=0; i<5; ++i)
    std::cout << v[i] << " ";  // 1 1 1 1 1
  std::cout << "\n";

  return 0;
}
```

Program 2.32: *../progs/lecture/fill6.cpp*

Every vector type comes with two *vector iterator* types. These play exactly the role of the pointer types T^* and const T^* for static arrays with underlying type T. A vector iterator is not a pointer, but it *behaves* like a pointer in all the relevant aspects. That means, it points to a vector element and can be dereferenced to yield the vector element it points to. It can also be used in the same kind of (pointer) arithmetic expressions as discussed in Section 2.8.5.

While array-to-pointer conversion is automatic (Section 2.8.3), vector-to-iterator conversion needs to be done explicitly. If v is a vector, then `v.begin()` yields a vector iterator pointing to the first element of the vector v, while `v.end()` yields a past-the-end iterator for v. We have already encountered such member function calls on Page 169 in connection with strings.

The following program demonstrates the use of vector iterators.

```
// Program: vector_iterators.cpp
// demonstrates vector iterators
#include<iostream>
#include<vector>

typedef std::vector<int>::const_iterator Cvit;
typedef std::vector<int>::iterator       Vit;

int main()
{
  std::vector<int> v(5, 0);   // 0 0 0 0 0

  // output all elements of a, using iteration
  for (Cvit it = v.begin(); it != v.end(); ++it)
    std::cout << *it << " ";
  std::cout << "\n";

  // manually set all elements to 1
  for (Vit it = v.begin(); it != v.end(); ++it)
    *it = 1;

  // output all elements again, using random access
  for (int i=0; i<5; ++i)
    std::cout << v.begin()[i] << " ";
  std::cout << "\n";

  return 0;
}
```

Program 2.33: *../progs/lecture/vector_iterators.cpp*

The names of the vector iterator types seem somewhat long (but compared to other

type names in real-life C++ they are *very* short). We use

```
std::vector<int>::iterator
```

where we would use `int*` for a static arry with underlying type `int`, and we use

```
std::vector<int>::const_iterator
```

where we would use the const pointer type `const int*` for the static array. Here, `::` is the scope operator, indicating that the types `iterator` and `const_iterator` are not global types (such as `int` and `double`), but are defined inside the vector class template. Still, in order not to clutter up the code, we are using a *typedef declaration* Such a declaration introduces a new (short) name for an existing type into its scope. It does *not* introduce a new type. In fact, the new name can be used synonymously with the old name in all contexts.

We see in Program 2.33 that vector iterators can indeed be used like pointers. If `it` is a vector iterator, `*it` is the vector element it points to, and `++it` advances the iterator to point to the next vector element. The last loop of Program 2.33 shows that vector iterators support random access just like pointers, where also here, `v.begin()[i]` is a shorthand for `*(v.begin()+i)`.

2.8.10 Containers and iterators

Let's take a step back, forget about arrays for a moment, and go for a bigger picture. We have already indicated in the introduction to this section that the process of traversing a sequence of data is ubiquitous. Typically, the data are stored in some *container*, and we need to perform a certain operation for all elements in the container. In general, a container is an object that can store other objects (its elements), and that offers some ways of accessing these elements. The only "hard" requirement here is that a container must offer the possibility of traversing all its elements. In this informal sense, static arrays and vectors are indeed containers, since traversal is possible through pointers and vector iterators, respectively. We can even traverse the elements in two ways: by random access, or by iteration.

But arrays are by far not the only useful containers, even though Ed Post humorously claims otherwise in his 1983 essay *Real Programmers don't use Pascal.* In fact, the C++ standard library contains many different containers, one of which we will discuss next.

Sets. Mathematically, a set is an unordered collection of elements, where every element occurs only once. For example, the set $\{1, 2\}$ is the same as the set $\{2, 1\}$, or the set $\{1, 2, 1\}$. In this respect, sets are fundamentally different from sequences (that are modeled by arrays). In C++, the type `std::set<T>` can be used to store a set with elements of type T.

Here is an application of sets: we have a text and want to know all the characters that appear at least once in the text. To do this, we turn the text into a set (so that

duplicates are removed) and then iterate over the set, outputting all its elements. Here is a corresponding program that shows how easy this is using `std::set<char>` and the corresponding *set iterators*. On the side, we also show how a (generic) algorithm (in this case `std::find`) can be used to check whether a certain element is present in a set; as is customary, we provide the set to the algorithm in form of two set iterators.

```
1  // Prog: set.cpp
2  // demonstrates usage of the std::set container
3  #include<iostream>
4  #include<set>
5  #include<algorithm>
6
7  typedef std::set<char>::const_iterator Sit;
8
9  int main()
10 {
11   std::string text =
12     "What are the distinct characters in this string?";
13   std::set<char> s (text.begin(), text.end());
14
15   // check whether text contains a question mark
16   if (std::find (s.begin(), s.end(), '?') != s.end())
17     std::cout << "Good question!\n";
18
19   // output all distinct characters
20   for (Sit it = s.begin(); it != s.end(); ++it)
21     std::cout << *it;
22   std::cout << "\n";
23
24   return 0;
25 }
```

Program 2.34: *../progs/lecture/set.cpp*

The output of this program is

```
Good question!
 ?Wacdeghinrst
```

We see that surprisingly few distinct characters appear in our text. We also see that iteration over the set `s` yields the characters in lexicographical order, but for the purposes of a set, any other order would have been fine as well. In this sense, all we can assume is that `[s.begin(), s.end)]` describes the elements of the set in *some* order.

What happens if we are trying to traverse the set using random access, with the following loop?

```
// output all distinct characters
```

```
for (int i=0; i<s.size(); ++i)
  std::cout << s.begin()[i];
std::cout << "\n";
```

The compiler will issue an error message, telling us that no subscript operator [] is available for set iterators such as `s.begin()`. This makes sense: since sets are unordered, there is no notion of "the i-th element of a set." This means that set iterators are weaker than pointers or vector iterators. They support the operation of "going to the next element" (in no specified order), using the ++ operator, but they don't support random access.

Iterator concepts. In C++, there are various iterator types. All of them support dereferencing via * and the operation of "going to the next element" via ++; that's precisely what defines iterators. On top of that, some iterators can do more. For example, "going to the previous element" via -- is (maybe surprisingly) something that not every iterator type supports. But again, this makes sense: in the context of streaming algorithms—a topic that we briefly touched in Section 2.7.6 on Page 168—an element may be read, processed, and discarded immediately afterwards, so there is no way of "going to the previous element". An iterator that allows random access is a Mercedes among the iterators, but you don't always need a Mercedes to get from A to B. For example, the `std::find` algorithm just requires the minimum iterator functionality (* and ++) in order to work.

In fact, *every* container algorithm of the C++ standard library works in this way: it expects the underlying container to offer iterator types conforming to some well-defined *iterator concept*. Such a concept is a list of requirements that the iterator types need to satisfy in order for the algorithm to work. For example, some algorithms require random access iterators, some don't. The specifics of the container itself are completely irrelevant for the algorithms. For example, `std::find` can be used to search for elements in arrays or in sets in the same way, although these two types of containers are very different. This is a very strong point of the C++ standard library; in general, the approach of parameterizing algorithms with requirements on the data instead of actual data types is called *generic programming* and deserves a full lecture on its own.

Pointers, and why we need them after all. The pointer concept is usually hard to understand for many students, and its details are perceived as very technical. Section 2.8.5 on pointer arithmetic seems to support this perception. On top of that, students often fail to see why pointers are useful in connection with processing arrays. Isn't it much simpler and more readable to work with indices in connection with arrays, rather than pointers? Indeed, given the choice between the following two loops, the authors would prefer the first index-based one, due to its better readability. The second one seems to be slightly more efficient due to less address computations (see Section 2.8.6), but any modern compiler would optimize the index-based loop to be as efficient as the pointer-based one.

```
for (int i=0; i<n; ++i)
  a[i] = 0;
```

```
for (int* ptr=a; ptr<a+n; ++ptr)
  *ptr = 0;
```

But here is the point: in (good) practice, we would not use any of these two loops, but we would employ `std::fill` to do the job for us:

```
std::fill (a, a+n, 0);
```

And here we have no choice: if we want to use standard library algorithms, we have to pass even a simple static array as a range of pointers. We do not need to do all the pointer arithmetic ourselves, but we do need to know that

- a static array variable `a` is at the same time a pointer to the first element of the array, and that
- `a+i` is a pointer to the element of index i (or a past-the-end pointer if $i = n$, the length of the array).

If we are using standard library algorithms with other containers, we do not use pointers but the corresponding container iterators. But we still think of them as pointers, and we need to understand basic pointer functionality in order to use the algorithms. Here is another example. For a given iterator range `[begin, end)`, the expression

```
std::min_element (begin, end)
```

returns the smallest element described by the range. However, in order to be able to cope with an empty range, it does not return the element itself but an *iterator* pointing to it (which has the value of `end` if the range is empty). Hence, to get the actual smallest element of a nonempty range, we still have to dereference the result and use the expression

```
*std::min_element (begin, end)
```

How would you understand or even do this if you had never heard about pointers? The morale is: in manually processing arrays, use indices for better readability, but in processing them with standard algorithms, have a basic understanding of the concept of pointers.

2.8.11 Details

Constant pointers. We have seen that

```
const int* ptr = a;
```

defines a const pointer to the first element of the array `a`, meaning that `a` cannot be modified through this pointer. We can still advance this pointer in a loop to read the elements of `a`, but not change their values. This means that the dereferenced value `*ptr` is constant (of type `const int`), but not the pointer `ptr` itself (we can advance it). If we wanted a *constant pointer* instead, we would write

```
int* const ptr = a;
```

Luckily, this is rarely needed.

Default arguments. Some functions have the property that there are "natural" values for one or more of their formal arguments. For example, when filling an array of underlying type `int`, the value 0 is such a natural value. In such a case, it is possible to specify this value as a *default argument*; this allows the caller of the function to omit the corresponding call argument and let the compiler insert the default value instead. In case of the function `fill` from Program 2.28, this would look as follows.

```
// PRE:  [first, last) is a valid range
// POST: *p is set to value, for p in [first, last)
void fill (int* first, int* last, const int value = 0) {
  // iteration by pointer
  for (int* p = first; p != last; ++p)
    *p = value;
}
```

This function can now be called with either two or three arguments, as follows.

```
int a[5];
fill (a, a+5);     // means: fill (a, a+5, 0)
fill (a, a+5, 1);
```

In general, there can be default values for any number of formal arguments, but these arguments must be at consecutive positions $i, i+1, \ldots, k$ among the k arguments, for some i. The function can then be called with any number of call arguments between $i-1$ and k, and the compiler automatically inserts the default values for the missing call arguments.

A function may have a separate declaration that specifies default arguments, like in the following declaration of `fill`.

```
// PRE:  [first, last) is a valid range
// POST: *p is set to value, for p in [first, last)
void fill (int* first, int* last, int value = 0);
```

In this case, the actual definition must not repeat the default arguments (the actual rules are a bit more complicated, but this is the upshot).

2.8.12 Goals

Dispositional. At this point, you should ...

1) understand the functions `fill_n` and `fill` from Program 2.27 and Program 2.28 ;
2) know that formal arguments of pointer type can be used to write array-processing functions, and mutating functions;
3) understand the pointer concept, and how to compute with addresses;
4) be aware that (and understand why) arrays and pointers must be used with care;
5) understand vector iterators and their role as "pointers" for vectors;

6) know what a container is, and that it can be processed through its iterators by algorithms.

Operational. In particular, you should be able to ...

(G1) read, understand, and argue about simple programs involving arrays and pointers; relevant aspects are in particular pointer arthmetic and const-correctness

(G2) write programs that read a sequence of data into a (multidimensional) array;

(G3) within programs, traverse an array by using indices, pointers, or vector iterators

(G4) understand and write functions that perform simple array processing tasks;

(G5) write (mutating) functions for given tasks, and write programs for given tasks that use these functions, observing const-correctness;

(G6) write simple programs for given tasks that use standard algorithms on arrays.

2.8.13 Exercises

Exercise 100 *Write a program* `swap.cpp` *that defines and calls a function for interchanging the values of two* `int` *objects. The program should have the following structure.*

```
#include<iostream>

// your function definition goes here

int main() {
  // input
  std::cout << "i =? ";
  int i; std::cin >> i;

  std::cout << "j =? ";
  int j; std::cin >> j;

  // your function call goes here

  // output
  std::cout << "Values after swapping: i = " << i
            << ", j = " << j << ".\n";

  return 0;
}
```

Here is an example run of the completed program:

```
i =? 5
j =? 8
Values after swapping: i = 8, j = 5.
```

(G5)

Exercise 101 *Write a program* unique.cpp *that implements and tests the following function.*

```
// PRE:  [begin, end) is a valid range and describes a sequence
//       of elements that are sorted in nondecreasing order
// POST: the return value is true if and only if no element
//       occurs twice in the sequence
typedef std::vector<int>::const_iterator Cit;
bool unique (Cit first, Cit last);
```

(G4)(G5)

Exercise 102 *Modify the program* sort_array.cpp *from Exercise 94 in such way that the resulting program* sort_array2.cpp *defines and calls a function*

```
// PRE: [begin, end) is a valid range
// POST: the elements *p, p in [begin, end) are
//       in ascending order
typedef std::vector<int>::iterator It;
void sort (It begin, It end);
```

to perform the sorting of the array into ascending order. It may be tempting (but not allowed for obvious reasons) to use std::sort *or similar standard library functions in the body of the function* sort *that is to be defined. It* is *allowed, though, to compare the efficiency of your* sort *function with that of* std::sort *(which has the same pre- and postconditions and can be used after* include<algorithm>*).*

For this exercise, it is desirable (but not strictly necessary) to use pointer increment (++p*) as the only operation on pointers (apart from initialization and assignment, of course). If you succeed in doing so, your sorting function has the potential of working for containers that do not offer random access (see also Section 2.8.10).* (G4)(G5)

Exercise 103 *Provide the postcondition for the following function. The postcondition must completely describe the behavior of the function for all valid inputs.*(G1)(G4)(G5)

a)
```
// PRE: [b, e) and [o, o+(e-b)) are disjoint valid ranges
void f (int* b, int* e, int* o)
{
  while (b != e) *(o++) = *(--e);
}
```

b) Which of the three following function calls are valid according to the precondition?

```
int a[5] = {1,2,3,4,5};

f(a, a+5, a+5);
f(a, a+2, a+3);
f(a, a+3, a+2);
```

c) Is the function f implemented in a const-correct fashion? If not, where are const's to be added?

Exercise 104 *What does the following function do if* `e - b` *has value* 5*? To answer this, write down the values of* `b[0]`, `b[1]`,..., `b[4]` *after a call to* `f(b, b+5)`.

```
// PRE: [b, e) is a valid range
void f(unsigned int* b, unsigned int* e)
{
  int n = e - b;
  for (int i = 0; i < n; ++i) {
    b[i] = 1;
    for (int j = i - 1; j > 0; --j)
    b[j] += b[j - 1];
  }
}
```

Can you describe the behavior (and thus provide a postcondition) for general value of `e-b`*?* (G1)(G4)

Exercise 105

a) *What does the following program output, and why?*

```
#include<iostream>

int main()
{
  int a[] = {5, 6, 2, 3, 1, 4, 0};
  int* p = a;
  do {
    std::cout << *p << " ";
    p = a + *p;
  } while (p != a);

  return 0;
}
```

b) *More generally, suppose that in the previous program,* `a` *is initialized with some sequence of* n different *numbers in* $\{0, \ldots, n-1\}$ *(we see this for* $n = 7$ *in the previous program). Prove that the program terminates in this case.*

(G1)

Exercise 106 *Assume that in some program,* a *is an array of underlying type* int *and length* n.

a) *Given a variable* i *of type* int *with value* $0 \leq i \leq n$, *how can you obtain a pointer* p *to the element of index* i *in* a*? (Note: if* $i = n$, *this is asking for a past-the-end pointer.)*

b) *Given a pointer* p *to some element in* a, *how can you obtain the index* i *of this element? (Note: if* p *is a past-the-end pointer, the index is defined as* n.*)*

Write code fragments that compute p *from* i *in a) and* i *from* p *in b).* (G1)

Exercise 107 *Find and fix at least 4 problems in the following program. The fixed program should indeed correctly do what it claims to do. Is the fixed program const-correct? If not, make it const-correct! (This is a theory exercise, but you may of course use the computer to help you.)* (G1)

```
#include<iostream>

int main()
{
  int a[7] = {0, 6, 5, 3, 2, 4, 1}; // static array
  int b[7];
  const int* c = b;

  // copy a into b using pointers
  for (int* p = a; p <= a+7; ++p)
    *c++ = *p;

  // cross-check with random access
  for (int i = 0; i <= 7; ++i)
    if (a[i] != c[i])
      std::cout << "Oops, copy error...\n";

  return 0;
}
```

Exercise 108 *Enhance the program* read_array.cpp *from Exercise 93 so that the resulting program* cycles.cpp *interprets the input sequence of* n *integers as a* permutation π *of* $\{0, \ldots, n-1\}$, *and that it outputs the* cycle decomposition *of* π.

Some explanations are in order: a permutation π *is a bijective mapping from the set* $\{0, \ldots, n-1\}$ *to itself; therefore, the input sequence can be interpreted as the sequence of values* $\pi(0), \ldots, \pi(n-1)$ *of a permutation* π *if and only if it contains every number from* $\{0, \ldots, n-1\}$ *exactly once.*

The program `cycles.cpp` *should first check whether the input sequence satisfies this condition, and if not, terminate with a corresponding message. If the input indeed encodes a permutation* π*, the program should output the* cycle decomposition *of* π*. A cycle in* π *is any sequence of the form* $(\ n_1\ n_2\ \cdots\ n_k\ )$ *such that*

- $n_2 = \pi(n_1), \quad n_3 = \pi(n_2), \quad \ldots, \quad n_k = \pi(n_{k-1})$*, and* $n_1 = \pi(n_k)$*, and*
- n_1 *is the smallest element among* $n_1, \ldots, n_k$*.*

Any cycle uniquely determines the π*-values of all its elements; on the other hand, every element appears in some cycle (which might be of the trivial form* (n_1)*, meaning that* $\pi(n_1) = n_1$*). This implies that the permutation decomposes into a unique set of cycles. For example, the permutation* π *given by*

$$\pi(0) = 4, \quad \pi(1) = 2, \quad \pi(2) = 3, \quad \pi(3) = 1, \quad \pi(4) = 0$$

decomposes into the two cycles $(\ 0\ 4\ )$ *and* $(\ 1\ 2\ 3\ )$*.* (G1)(G2)(G3)

Exercise 109 *a) Enhance the program* `read_array.cpp` *from Exercise 93 in such a way that it outputs the* index *of the first occurrence of a given element in the array. You may use the algorithm* `std::find` *to do this.*

b) Modify the program so that it outputs the index of the last *occurrence of an element.* (G2)(G3)(G6)

Exercise 110 *A Sudoku puzzle is posed on a grid of* 9×9 *cells, subdivided into 9 square boxes of* 3×3 *cells each. Some grid cells are already filled by numbers between* 1 *and 9; the goal is to fill the remaining cells by numbers between* 1 *and 9 in such a way that within each row, column, and box of the completed grid, every number occurs exactly once. Here is an example of a Sudoku puzzle:*

			1			7	4	
	5			9			3	2
		6	7			9		
4			8					
	2						1	
					9			5
		4			7	3		
7	3			2			6	
	6	5			4			

In solving the puzzle, one may try to deduce from the already filled numbers that exactly one *number is a candidate for a suitable empty cell. Then this number is filled into the cell, and the deduction process is repeated. There are two situations*

where such a deduction for the cell in row r */ column* c *and number* n *is particularly easy and follows the Sherlock Holmes approach (*"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?"*).*

1. *All numbers distinct from* n *already appear somewhere in the same row, column, or 3x3 box. This necessarily means that the cell has to be filled with* n*, since we have eliminated all other* numbers *as impossible.*

2. *All other cells in the same row, or in the same column, or in the same 3x3 box are already known not to contain* n*. Again, the cell has to be filled by* n *then, since we have eliminated all other* cells *for the number* n *within the row, column, or box.*

Write a program sudoku.cpp *that takes as input a Sudoku puzzle in form of a sequence of* 81 *numbers between* 0 *and* 9 *(the grid numbers given row by row, where* 0 *indicates an empty cell). The numbers might be separated by whitespaces, so that the Sudoku puzzle from above could conveniently be encoded like this in a file:*

```
0 0 0   1 0 0   7 4 0
0 5 0   0 9 0   0 3 2
0 0 6   7 0 0   9 0 0

4 0 0   8 0 0   0 0 0
0 2 0   0 0 0   0 1 0
0 0 0   0 0 9   0 0 5

0 0 4   0 0 7   3 0 0
7 3 0   0 2 0   0 6 0
0 6 5   0 0 4   0 0 0
```

The program should now try to solve the puzzle by using only the two Sherlock-Holmes-type deductions from above. The output should be a (partially) completed grid that is either the solution to the puzzle, or the unique (why?) partial solution in which no Sherlock-Holmes-type deductions apply anymore (again, empty cells should be indicated by the digit 0*).*

In the above example, the output of a correct program will be the solution:

```
3 9 2   1 8 5   7 4 6
8 5 7   4 9 6   1 3 2
1 4 6   7 3 2   9 5 8

4 7 9   8 5 1   6 2 3
5 2 8   6 7 3   4 1 9
6 1 3   2 4 9   8 7 5

2 8 4   5 6 7   3 9 1
```

```
7 3 1   9 2 8   5 6 4
9 6 5   3 1 4   2 8 7
```

For reading the input from a file, it can be convenient to redirect the standard input to the file containing the puzzle data. For checking whether any Sherlock-Holmes-type deductions apply, it can be useful to maintain (and update) for every triple (r, c, n) *the information whether* n *is still a possible candidate for the cell in row* r */ column* c*.*

You will discover that many Sudoku puzzles that typically appear in newspapers can be solved by your program and are therefore easy, even if they are labeled as medium *or* hard*.*

Hint: *It is advisable not to optimize for efficiency here, since this will only lead to more complicated and error-prone code. Given the very small problem size, such optimizations won't have a noticeable effect anyway.*

2.9 Recursion

Mir san mir.

Bavarian dictum, meaning "we are we".

Beware of bugs in the above code; I have only proved it correct, not tried it.

D./ E./ Knuth, in a letter to van Emde Boas (1977)

This section introduces recursive functions, functions that directly or indirectly call themselves. You will see that recursive functions are very natural in many situations, and that they lead to compact and readable code close to mathematical function definitions. We will also explain how recursive function calls are processed, and how recursion can (in principle) be replaced with iteration. In the end, you will see two applications (sorting, and drawing fractals) that demonstrate the power of recursion.

2.9.1 A warm-up

Many mathematical functions are naturally defined *recursively*, meaning that the function to be defined appears in its own definition. For example, for any $n \in \mathbb{N}$, the number $n!$ can recursively be defined as follows.

$$n! := \begin{cases} 1, & \text{if } n \leq 1 \\ n \cdot (n-1)!, & \text{if } n > 1. \end{cases}$$

In C++ we can also have recursive functions: a function may call itself. This is nothing exotic, since after all, a function call is just an expression that can in principle appear anywhere in the function's scope, and that scope includes the function body. Here is a recursive function for computing $n!$; in fact, this definition exactly matches the mathematical definition from above.

```
// POST: return value is n!
unsigned int fac (const unsigned int n)
{
  if (n <= 1) return 1;
  return n * fac(n-1); // n > 1
}
```

Here, the expression `fac(n-1)` is a *recursive call* of `fac`.

Infinite recursion. With recursive functions, we have the same issue as with loops (Section 2.4.2): it is easy to write down function calls whose evaluation does not terminate. Here is the shortest way of creating an infinite recursion: define the function

```
void f()
{
  f();
}
```

with no arguments and evaluate the expression `f()`. The reason for non-termination is clear: the evaluation of `f()` consists of an evaluation of `f()` which consists of an evaluation of `f()` which... you get the picture.

As for loops, the function definition has to make sure that progress towards termination is made in every function call. For the function `fac` above, this is the case: each time `fac` is called recursively, the value of the call argument becomes smaller, and when the value reaches 1, no more recursive calls are performed: we say that the recursion "bottoms out".

2.9.2 The call stack

Let's try to understand what exactly happens during the evaluation of `fac(3)`, say. The formal argument `n` is initialized with 3, and since this is greater than 1, the statement `return n * fac(n-1);` is executed next. This first evaluates the expression `n * fac(n-1)` and in particular the right operand `fac(n-1)`. Since `n-1` has value 2, the formal argument `n` is therefore initialized with 2.

But wait: what is "the" formal argument? Automatic storage duration implies that each function call has its "own" fresh instance of the formal argument, and the lifetime of this instance is the respective function call. In evaluating `f(n-1)`, we therefore get a new instance of the formal argument `n`, on top of the previous instance from the call `f(3)` (that has not yet terminated). But which instance of `n` do we use in the evaluation of `f(n-1)`? Quite naturally, it will be the new one, the one that "belongs" to the call `f(n-1)`. This rule is in line with the general scope rules from Section 2.4.3: the relevant declaration is always the most recent one that is still visible.

The technical realization of this is very simple. Everytime a function is called, the call argument is evaluated, and the resulting value is put on the *call stack* which is simply a region in the computer's memory. If the function has several arguments, several values are put on the call stack; to keep the description simple, we concentrate on the case of one argument.

Like a stack of papers on your desk, the call stack has the property that the object that came last is "on top". Upon termination of a function call, the top object is taken off the stack again. Whenever a function call accesses or changes its formal argument, it does so by accessing or changing the corresponding object on top of the stack.

This has all the properties we want: every function call works with its own instance of the formal argument; when it calls another function (or the function itself recursively),

this instance becomes temporarily hidden, until the nested call has terminated. At that point, the instance reappears on top of the stack and allows the original function call to work with it again.

Table 5 shows how this looks like for `f(3)`, assuming that the right operand of the multiplication operator is always evaluated first. Putting an object on the stack "pushes" it, and taking the top object off "pops" it.

call stack (bottom ⟷ top)				evaluation sequence	action
···				`fac(3)`	push 3
···	n: 3			`n * fac(n-1)`	
···	n: 3			`n * fac(2)`	push 2
···	n: 3	n: 2		`n * (n * fac(n-1))`	
···	n: 3	n: 2		`n * (n * fac(1))`	push 1
···	n: 3	n: 2	n: 1	`n * (n * 1)`	pop
···	n: 3	n: 2		`n * (2 * 1)`	
···	n: 3	n: 2		`n * 2`	pop
···	n: 3			`3 * 2`	
···	n: 3			`6`	pop
···					

Table 5: *The call stack, and how it evolves during an evaluation of* `fac(3)`*; the respective value of* `n` *to use is always the one on top*

Because of the call stack, infinite recursions do not only consume time but also memory. Unlike infinite loops, they usually lead to a program abortion as soon as the memory reserved for the call stack is full.

2.9.3 Basic practice

Let us consider two more simple recursive functions that are somewhat more interesting than `fac`. They show that recursive functions are particularly amenable to correctness proofs of their postconditions, and this makes them attractive. On the other hand, we also see that it is easy to write innocent-looking recursive functions that are very inefficient to evaluate.

Greatest common divisor. Consider the problem of finding the greatest common divisor $\gcd(a, b)$ of two natural numbers a, b. This is defined as the largest natural number that divides both a and b without remainder. In particular, $\gcd(n, 0) = \gcd(0, n) = n$ for $n > 0$; let us also define $\gcd(0, 0) := 0$.

The *Euclidean algorithm* finds $\gcd(a, b)$, based on the following

Lemma 1 *If* $b > 0$, *then*

$$\gcd(a, b) = \gcd(b, a \bmod b).$$

Proof. Let k be a divisor of b. From

$$a = (a \operatorname{div} b)b + a \bmod b$$

it follows that

$$\frac{a}{k} = (a \operatorname{div} b)\frac{b}{k} + \frac{a \bmod b}{k}.$$

Since $a \operatorname{div} b$ and b/k are integers, we get

$$\frac{a \bmod b}{k} \in \mathbb{N} \quad \Leftrightarrow \quad \frac{a}{k} \in \mathbb{N}.$$

In words, if k is a divisor of b, then k divides a if and only if k divides $a \bmod b$. This means, the divisors of a *and* b are exactly the divisors of b *and* $a \bmod b$. This proves that $\gcd(a, b)$ and $\gcd(b, a \bmod b)$ are equal. □

Here is the corresponding C++ function for computing the greatest common divisor of two `unsigned int` values, according to the Euclidean algorithm.

```
// POST: return value is the greatest common divisor of a and b
unsigned int gcd (const unsigned int a, const unsigned int b)
{
  if (b == 0) return a;
  return gcd(b, a % b);  // b != 0
}
```

The Euclidean algorithm is very fast. We can easily call it for any `unsigned int` values on our platform, without noticing any delay in the evaluation.

Correctness and termination. For recursive functions, it is often very easy to prove that the postcondition is correct, by using the underlying mathematical definition directly (such as $n!$ for `fac`), or by using some facts that follow from the mathematical definition (such as Lemma 1 for gcd).

The correctness proof must involve a termination proof, so let's start with this: every call to gcd terminates, since the value b of the second argument is bounded from below by 0 and gets smaller in every recursive call (we have $a \bmod b < b$).

Given this, the correctness of the postcondition follows from Lemma 1 by induction on b. For $b = 0$, this is clear. For $b > 0$, we inductively assume that the postcondition is correct for all calls to gcd where the second argument has value $b' < b$. Since $b' = a \bmod b$ satisfies $b' < b$, we may assume that the call `gcd(b, a % b)` correctly returns $\gcd(b, a \bmod b)$. But by the lemma, $\gcd(b, a \bmod b) = \gcd(a, b)$, so the statement

```
return gcd(b, a % b);
```

correctly returns $\gcd(a, b)$.

Fibonacci numbers. The sequence $0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots$ of *Fibonacci numbers* is one of the most famous sequences in mathematics. Formally, the sequence is defined as follows.

$$\begin{aligned} F_0 &:= 0, \\ F_1 &:= 1, \\ F_n &:= F_{n-1} + F_{n-2}, \quad n > 1. \end{aligned}$$

This means, every element of the sequence is the sum of the two previous ones. From this definition, we can immediately write down a recursive C++ function for computing Fibonacci numbers, getting termination and correctness for free.

```
// POST: return value is the n-th Fibonacci number F_n
unsigned int fib (const unsigned int n)
{
  if (n == 0) return 0;
  if (n == 1) return 1;
  return fib(n-1) + fib(n-2); // n > 1
}
```

If you write a program to compute the Fibonacci number F_n using this function, you will notice that somewhere between $n = 30$ and $n = 50$, the program becomes very slow. You even notice how much slower it becomes when you increase n by just 1.

The reason is that the mathematical definition of F_n does not lead to an efficient algorithm, since all values $F_i, i < n-1$, are repeatedly computed, some of them extremely often. You can for example check that the call to `fib(50)` computes F_{48} already twice (once directly in `fib(n-2)`, and once indirectly from `fib(n-1)`. F_{47} is computed three times, F_{46} five times, and F_{45} eight times (do you see a pattern?).

2.9.4 Recursion versus iteration

From a strictly functional point of view, recursion is superfluous, since it can be simulated through iteration (and a call stack explicitly maintained by the program; we could simulate the call stack with an array). We don't have the means to prove this here, but we want to show it for the recursive functions that we have seen in the previous section.

The function gcd is very easy to write iteratively, since it is *tail-end recursive*. This means that there is only one recursive call, and that one appears at the very end of the function body. Tail-end recursion can be replaced by a simple loop that iteratively updates the formal arguments until the termination condition is satisfied. In the case of gcd, this update corresponds to the transformation $(a, b) \to (b, a \bmod b)$.

```
// POST: return value is the greatest common divisor of a and b
unsigned int gcd2 (unsigned int a, unsigned int b)
{
  while (b != 0) {
    const unsigned int a_prev = a;
```

```
    a = b;
    b = a_prev % b;
  }
  return a;
}
```

You see that we get longer and less readable code, and that we need an extra variable to remember the previous value of a before the update step; in the spirit of Section 2.4.8, we should therefore use the original recursive formulation.

Our function `fib` for computing Fibonacci numbers is not tail-end recursive, but it is still easy to write it iteratively. Remember that F_n is the sum of F_{n-1} and F_{n-2}. We can therefore write a loop whose iteration i computes F_i from the previously computed values F_{i-2} and F_{i-1} that we maintain in the variables a and b.

```
// POST: return value is the n-th Fibonacci number F_n
unsigned int fib2 (const unsigned int n)
{
  if (n == 0) return 0;
  if (n <= 2) return 1;
  unsigned int a = 1;  // F_1
  unsigned int b = 1;  // F_2
  for (unsigned int i = 3; i <= n; ++i) {
    const unsigned int a_prev = a;  // F_{i-2}
    a = b;                          // F_{i-1}
    b += a_prev;                    // F_{i-1} += F_{i-2} -> F_i
  }
  return b;
}
```

Again, this non-recursive version `fib2` is substantially longer and more difficult to understand than `fib`, but this time there is a benefit: `fib2` is much faster, since it computes every number $F_i, i \leq n$ *exactly once*. While we would grow old in waiting for the call `fib(50)` to terminate, `fib2(50)` gives us the answer in no time. (Unfortunately, this answer may be incorrect, since F_{50} could exceed the value range of the type `unsigned int`.)

In this case we would prefer `fib2` over `fib`, simply since `fib` is too inefficient for practical use. The more complicated function definition of `fib2` is a moderate price to pay for the speedup that we get.

2.9.5 Primitive recursion

Roughly speaking, a mathematical function is *primitive recursive* if it can be written as a C++ function f in such a way that f neither directly nor indirectly calls itself with call arguments depending on f. For example,

```
unsigned int f (const unsigned int n)
```

```
{
  if (n == 0) return 1;
  return f(f(n-1) - 1);
}
```

is not allowed, since f recursively calls itself with an argument depending of f. This does *not* mean that the underlying mathematical function is not primitive recursive, it just means that we have chosen the wrong C++ implementation. Indeed, the above f implements the mathematical function satisfying $f(n) = 1$ for all n, and this function is obviously primitive recursive.

In the early 20-th century, it was believed that the functions whose values can in principle be computed by a machine are exactly the primitive recursive ones. Indeed, the function values one computes in practice (including $\gcd(a, b)$ and F_n) come from primitive recursive functions.

It later turned out that there are computable functions that are not primitive recursive. A simple and well-known example is the binary *Ackermann function* $A(m, n)$, defined by

$$A(m,n) = \begin{cases} n+1, & \text{if } m = 0 \\ A(m-1,1), & \text{if } m > 0, n = 0 \\ A(m-1, A(m, n-1)), & \text{if } m > 0, n > 0. \end{cases}$$

The fact that this function is not primitive recursive requires a proof (that we don't give here). As already noted above, it is necessary but not sufficient that this definition recursively uses A with an argument that depends on A.

It may not be immediately clear that the corresponding C++ function

```
// POST: return value is the Ackermann function value A(m,n)
unsigned int A (const unsigned int m, const unsigned int n) {
  if (m == 0) return n+1;
  if (n == 0) return A(m-1,1);
  return A(m-1, A(m, n-1));
}
```

always terminates, but Exercise 112 asks you to show this. Table 6 lists some Ackermann function values. For $m \leq 3$, $A(m, n)$ looks quite moderate, but starting from $m = 4$, the values get extremely large. You can still compute $A(4, 1)$, although this takes surprisingly long already. You *might* be able to compute $A(4, 2)$; after all, $2^{65536}-3$ has "only" around $20,000$ decimal digits. But the call to A(4,3) will not terminate within any observable period.

It can in fact be shown that $A(n, n)$ grows faster than any primitive recursive function in n (and this is a proof that A cannot be primitive recursive). Recursion is a powerful but also dangerous tool, since it is easy to encode (too) complicated computations with very few lines of code.

		n					
		0	1	2	3	$\cdots$	n
	0	1	2	3	4	$\cdots$	$n+1$
	1	2	3	4	5	$\cdots$	$n+2$
m	2	3	5	7	9	$\cdots$	$2n+3$
	3	5	13	29	61	$\cdots$	$2^{n+3}-3$
	4	13	65533	$2^{65536}-3$	$2^{2^{65536}}-3$	$\cdots$	$\underbrace{2^{2^{\cdot^{\cdot^{\cdot^{2}}}}}}_{n+3}-3$

Table 6: *Some values of Ackermann's function*

2.9.6 Sorting

Sorting a sequence of values (numbers, texts, etc.) into ascending order is a very basic and important operation. For example, a specific value can be found much faster in a sorted than in an unsorted sequence (see Exercise 120). You know this from daily life, and that's why you sort your CDs, and why the entries in a telephone directory are sorted by name.

We have asked you in Exercise 94 to write a program that sorts a given sequence of integers; Exercise 102 was about making this into a function that sorts all numbers described by a given iterator range. In both exercises, you were not supposed to do any efficiency considerations.

Here we want to catch up on this and investigate the *complexity* of the sorting problem. Roughly speaking, the complexity of a problem is defined as the complexity (runtime) of the fastest algorithm that solves the problem. In computing Fibonacci numbers in Section 2.9.3 and Section 2.9.4, we have already seen that the runtimes of different algorithms for the same problem may vary a lot. The same is true for sorting algorithms, as we will discover shortly.

Let us start by analyzing one of the "obvious" sorting algorithms that you may have come up with in Exercise 94. The simplest one that the authors can think of is *minimum-sort*. Given the sequence of values (let's assume they are integers), *minimum-sort* first finds the smallest element of the sequence; then it interchanges this element with the first element. The sequence now starts with the smallest element, as desired, but the remainder of the sequence still needs to be sorted. But this is done in the same way: the smallest element among the remaining ones is found and interchanged with the *second* element of the sequence, and so on.

We are assuming that the sequence to sort is described by a vector iterator range, so we are sorting a vector (of integers). As before, in order to get rid of the lengthy vector iterator type name, we are starting with an appropriate typedef (see Page 201). Now *minimum-sort* can be realized as follows.

```
typedef std::vector<int>::iterator Vit;
```

```
// PRE: [begin, end) is a valid range
// POST: the elements *p, p in [begin, end) are in ascending order
void minimum_sort (Vit begin, Vit end)
{
  for (Vit p = begin; p != end; ++p) {
    // find minimum in nonempty range described by [p, end)
    Vit p_min = p; // iterator pointing to current minimum
    Vit q = p;     // iterator pointing to current element
    while (++q != end)
      if (*q < *p_min) p_min = q;
    // interchange *p with *p_min
    std::iter_swap (p, p_min);
  }
}
```

The standard library function `std::iter_swap` (also from `<algorithms>`) interchanges the values of the objects pointed to by its two arguments. There is also a function `std::min_element` (with similar functionality as our home-made `min` on page 195) that we could use to get rid of the inner loop; however, since we want to analyze the function `minimum_sort` in detail, we refrain from calling any nontrivial standard library function here.

What can we say about the runtime of `minimum_sort` for a given range? That it depends on the platform, this is for sure. On a modern PC, the algorithm will run much faster than on a vintage computer from the twentieth century. There is no such thing as "the" runtime. But if we look at what the algorithm does, we can find a measure of runtime that is platform-independent.

A dominating operation in the sense that it occurs very frequently during a call to `minimum_sort` is the comparison `*q < *p_min`. We can even exactly count the number of such comparisons, depending on the number of elements n that are to be sorted. In the first execution of the `while` statement, the first element is compared with all $n-1$ succeeding elements. In the second execution, the second element in compared with all the $n-2$ succeeding elements, and so on. In the second-to-last execution of the `while` statement, finally, we have one comparison, and that's it. We therefore have the following

Observation 1 *The function* `minimum_sort` *sorts a sequence of* n *elements with*

$$1+2+\ldots n-1 = \frac{n(n-1)}{2}$$

comparisons between sequence elements.

Why do we specifically count these comparisons? Because any other operation is either performed much less frequently (for example, the declaration statement `int* q = p` is executed only n times), or with approximately the same frequency. This concerns the assignment `p_min = q` which may happen up to $n(n-1)/2$ times, and the expression `++q != last`; this one is evaluated even more frequently, namely $n(n-1)/2+n$ times.

The total number of operations is therefore at most $c_1 n(n-1)/2 + c_2 n$ for some constants c_1, c_2. For large n, the linear term $c_2 n$ is negligible compared to the quadratic term $c_1 n(n-1)/2$; we can therefore conclude that the total number of operations needed to sort n numbers is proportional to the number of comparisons between sequence elements.

This implies the following: if you measure the runtime of the whole sorting algorithm, the resulting time T_{total} will be proportional to the time T_{comp} that is being spent with comparisons between sequence elements. (However, due to the effects of caching and other add-ons to the von-Neumann architecture, this is not necessarily true on every platform.) Since T_{comp} is in turn proportional to the number of comparisons itself, this number is a good indicator for the efficiency of the algorithm.

If you think about sorting more complicated values (like names in a telephone directory), a comparison between two elements might even become the single most time-consuming operation. In such a scenario, T_{comp} may eat up almost everything of T_{total}, making the comparison count an even more appropriate measure of efficiency.

To check that all this is not only pure theory, let us make some experiments and measure the time that it takes to execute Program 2.35 below, for various values of n (read from a file in order not to measure the time it takes us to enter n). The program first brings the sequence $0, 1, \ldots, n-1$ into random order, using the standard library function `std::random_shuffle`. Then it calls the function `minimum_sort` and finally checks whether we now indeed have the ascending sequence $0, 1, \ldots, n-1$. Yes, this program does other things apart from the actual sorting, but all additional operations are "cheap" in the sense that their number is proportional to n at most; according to our above line of arguments, they should therefore not matter.

```
// Prog: minimum_sort.cpp
// implements and tests minimum-sort on random input

#include<iostream>
#include<algorithm>
#include<vector>

typedef std::vector<int>::iterator Vit;

// PRE: [begin, end) is a valid range
// POST: the elements *p, p in [begin, end) are in ascending order
void minimum_sort (Vit begin, Vit end)
{
  for (Vit p = begin; p != end; ++p) {
    // find minimum in nonempty range described by [p, end)
    Vit p_min = p; // iterator pointing to current minimum
    Vit q = p;     // iterator pointing to current element
    while (++q != end)
      if (*q < *p_min) p_min = q;
    // interchange *p with *p_min
```

n	100,000	200,000	400,000	800,000	1,600,000
Gcomp	5	20	80	320	1280
Time (min)	0:15	1:05	4:26	15:39	64:22
sec/Gcomp	3.0	3.25	3.325	2.93	3.01

Table 7: *Number of comparisons and runtime of* minimum-sort

```
    std::iter_swap (p, p_min);
  }
}

int main()
{
  // input of number of values to be sorted
  int n;
  std::cin >> n;

  std::vector<int> v(n);

  std::cout << "Sorting " << n << " integers...\n";

  // create random sequence
  for (int i=0; i<n; ++i) v[i] = i;
  std::random_shuffle (v.begin(), v.end());

  // sort into ascending order
  minimum_sort (v.begin(), v.end());

  // is it really sorted ?
  for (int i=0; i<n-1;++i)
    if (v[i] != i) std::cout << "Sorting error!\n";

  return 0;
}
```

Program 2.35: *../progs/lecture/minimum_sort.cpp*

Table 7 summarizes the results. For every value of n, **Gcomp** is the number of Gigacomparisons (10^9 comparisons), according to Observation 1. In other words, **Gcomp**$= 10^{-9}n(n-1)/2$. **Time** is the absolute runtime of the program in minutes and seconds, on a modern PC. **sec/Gcomp** is **Time** (in seconds) divided by **Gcomp** and tells us how many seconds the program needs to perform one Gigacomparison.

The table shows that the number of seconds per Gigacomparison is around 3 for all considered values of n. As predicted above, the runtime in practice is therefore indeed proportional to the number of comparisons between sequence elements. This number

quadruples from one column to the next, and so does the runtime.

We also see that sorting numbers using *minimum-sort* appears to be pretty inefficient. 1,600,000 is not large by today's standards, but to sort that many numbers takes more than one hour! Given that **sec/Gcomp** appears to be constant, we can even estimate the time that it would take to sort 10,000,000 numbers, say. For this, we derive from Observation 1 the required number of Gigacomparisons (50,000) and multiply it with 3. The resulting 150,000 seconds are almost two days.

Essentially the same figures result from running other well-known simple sorting algorithms like *bubble-sort* or *insert-sort*. Can we do better? Yes, we can, and recursion helps us to do it!

Merge-sort. The paradigm behind the *merge-sort* algorithm is this: if a problem is too large to be solved directly, subdivide it into smaller subproblems that are easier to solve, and then put the overall solution together from the solutions of the subproblems. This paradigm is known as *divide and conquer*.

Here is how this works for sorting. Let us imagine that the numbers to be sorted come as a deck of cards, with the numbers written on them. The first step is to partition the deck into two smaller decks of half the size each. These two decks are then sorted independently from each other, with the same method; but the main ingredient of this method comes only now: we have to merge the two sorted decks into one sorted deck. But this is not hard: we put the two decks in front of us (both now have their smallest card on top); as long as there are still cards in one or both of the decks, the smaller of the two top cards (or the single remaining top card) is taken off and put upside down on a new deck that in the end represents the result of the overall sorting process. Figure 19 visualizes the merge step.

Figure 19: *Merging two sorted decks of cards into one sorted deck*

Here is how *merge-sort* can be realized in C++, assuming that we have a function merge that performs the above operation of merging two sorted sequences into a single sorted sequence. Again, we are sorting a vector.

```
typedef std::vector<int>::iterator Vit;

// PRE: [begin, end) is a valid range
```

```
// POST: the elements *p, p in [begin, end) are in ascending order
void merge_sort (Vit begin, Vit end)
{
  const int n = end - begin;
  if (n <= 1) return;                // nothing to do
  const Vit middle = begin + n/2;
  merge_sort (begin, middle);        // sort first half
  merge_sort (middle, end);          // sort second half
  merge (begin, middle, end);        // merge both halfs
}
```

If there is more than one element to sort, the function splits the range `[first, last)` into two ranges `[first, middle)` and `[middle, last)` of lengths $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$. Just as a reminder, for any real number x, $\lceil x \rceil$ is the smallest integer greater or equal to x ("x rounded up"), and $\lfloor x \rfloor$ is the largest integer smaller or equal to x ("x rounded down"). If n is even, both values $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ are equal to $n/2$, and otherwise, the first value is smaller by one.

As its next step, the algorithm recursively sorts the elements described by both ranges. In the end, it calls the function `merge` on the two ranges. In commenting the latter function, we stick to the deck analogy that we have used above. If you have understood the deck merging process, you will perceive the definition of `merge` as being straightforward, despite the iterator handling.

```
// PRE: [begin, middle), [middle, end) are valid ranges; in
//      both of them, the elements are in ascending order
void merge (Vit begin, Vit middle, Vit end)
{
  const int n = end - begin;    // total number of cards
  std::vector<int> deck(n);     // new deck to be built

  Vit left = begin;    // top card of left deck
  Vit right = middle; // top card of right deck
  for (Vit d = deck.begin(); d != deck.end(); ++d)
    // put next card onto new deck
    if      (left == middle) *d = *right++; // left deck is empty
    else if (right == end)   *d = *left++;  // right deck is empty
    else if (*left < *right) *d = *left++;  // smaller top card left
    else                     *d = *right++; // smaller top card right

  // copy new deck back into [begin, end)
  Vit d = deck.begin();
  while (begin != middle) *begin++ = *d++;
  while (middle != end) *middle++ = *d++;
}
```

Analyzing merge-sort. As for *minimum-sort*, we will count the number of comparisons between sequence elements that occur when a sequence of n numbers is being sorted. Again, we can argue that the total number of operations is proportional to this number of comparisons. For *merge-sort*, this fact is not so immediate, though, and we don't expect you to understand it now. But for the benefit of (not only) the skeptic reader, we will check this fact experimentally below, as we did for *minimum-sort*.

All the comparisons take place during the calls to the function `merge` at the various levels of recursion, so let us first count the number of comparisons between sequence elements that one call to `merge` performs in order to create a sorted deck of n cards from two sorted decks.

It is apparent from the function body (and also from our informal description of the merging process above) that *at most one* comparison is needed for every card that is put on the new deck. Indeed, we may have to compare the two top cards of the left and the right deck in order to find out which card to take off next. But if one of the two decks becomes empty (this situation definitely occurs before the last card is put on the new deck), we don't do any further comparisons. This means that *at most* $n-1$ comparisons between sequence elements are performed in merging two sorted decks into one sorted deck with n cards.

Knowing this, we can now prove our main result.

Theorem 2 *The function* `merge_sort` *sorts a sequence of* $n \geq 1$ *elements with at most*

$$(n-1)\lceil \log_2 n \rceil$$

comparisons between sequence elements.

Proof. We define $T(n)$ to be the *maximum* possible number of comparisons between sequence elements that can occur during a call to `merge_sort` with an argument range of length n. For example, $T(0) = T(1) = 0$, since for ranges of lengths 0 and 1, no comparisons are made. We also get $T(2) = 1$, since for a range of length 2, *merge-sort* performs one comparison (in merging two sorted decks of one card each into one sorted deck of two cards). In a similar way, we can convince ourselves that $T(3) = 2$. There *are* sequences of length 3 for which one comparison suffices (the first card may be taken off the left deck which consists only of one card), but the maximum number that defines $T(3)$ is 2.

For general $n \geq 2$, we have the following *recurrence relation*:

$$T(n) \leq T(\lfloor \frac{n}{2} \rfloor) + T(\lceil \frac{n}{2} \rceil) + n - 1. \tag{2.1}$$

To see this, let us consider a sequence of n elements that actually requires the maximum number of $T(n)$ comparisons. This number of comparisons is the sum of the respective numbers in sorting the left and the right half, plus the number of comparisons during the merge step. The former two numbers are (by construction of `merge_sort` and definition of T) at most $T(\lfloor n/2 \rfloor)$ and $T(\lceil n/2 \rceil)$, while the latter number is at most $n-1$

n	100,000	200,000	400,000	800,000	1,600,000
Mcomp	1.7	3.6	7.6	16	33.6
Time (msec)	46	93	190	390	834
sec/Gcomp	27	25.8	25	24.4	25.1

Table 8: *Number of comparisons and runtime of* merge-sort

by our previous considerations regarding merge. It follows that $T(n)$, the actual number of comparisons, is bounded by the sum of all three numbers.

Now we can prove the actual statement of the theorem. Since the *merge-sort* algorithm is recursive, it is natural that the proof is inductive. For $n = 1$, we have $T(1) = 0 = (1-1)\lceil \log_2 2 \rceil$, so the statement holds for $n = 1$.

For $n \geq 2$, let us assume that the statement of the theorem holds for *all* values in $\{1, \ldots, n-1\}$ (this is the inductive hypothesis). From this hypothesis, we need to derive the validity of the statement for the number n itself (note that $\lfloor n/2 \rfloor, \lceil n/2 \rceil \geq 1$). This goes as follows.

$$
\begin{aligned}
T(n) &\leq T(\lfloor \frac{n}{2} \rfloor) + T(\lceil \frac{n}{2} \rceil) + n - 1 \quad \text{(Equation 2.1))} \\
&\leq (\lfloor \frac{n}{2} \rfloor - 1)\lceil \log_2 \lfloor \frac{n}{2} \rfloor \rceil + (\lceil \frac{n}{2} \rceil - 1)\lceil \log_2 \lceil \frac{n}{2} \rceil \rceil + n - 1 \quad \text{(inductive hypothesis)} \\
&\leq (\lfloor \frac{n}{2} \rfloor - 1)(\lceil \log_2 n \rceil - 1) + (\lceil \frac{n}{2} \rceil - 1)(\lceil \log_2 n \rceil - 1) + n - 1 \quad \text{(Exercise 121)} \\
&= (n-2)(\lceil \log_2 n \rceil - 1) + n - 1 \quad (\, n = \lfloor \tfrac{n}{2} \rfloor + \lceil \tfrac{n}{2} \rceil) \\
&\leq (n-1)(\lceil \log_2 n \rceil - 1) + n - 1 \\
&= (n-1)\lceil \log_2 n \rceil.
\end{aligned}
$$

□

As for *minimum-sort*, let us conclude with some experiments to check whether the number of comparisons between sequence elements is indeed a good indicator for the runtime in practice. The results of running *merge-sort* (Program 2.36 below) are given in Table 8 and look very different from the ones in Table 7.

```
// Prog: merge_sort.cpp
// implements and tests merge-sort on random input
#include<iostream>
#include<algorithm>
#include<vector>

typedef std::vector<int>::iterator Vit;

// PRE: [begin, middle), [middle, end) are valid ranges; in
//      both of them, the elements are in ascending order
void merge (Vit begin, Vit middle, Vit end)
```

```
{
  const int n = end - begin;   // total number of cards
  std::vector<int> deck(n);    // new deck to be built

  Vit left = begin;   // top card of left deck
  Vit right = middle; // top card of right deck
  for (Vit d = deck.begin(); d != deck.end(); ++d)
    // put next card onto new deck
    if      (left == middle) *d = *right++; // left deck is empty
    else if (right == end)   *d = *left++;  // right deck is empty
    else if (*left < *right) *d = *left++;  // smaller top card left
    else                     *d = *right++; // smaller top card right

  // copy new deck back into [begin, end)
  Vit d = deck.begin();
  while (begin != middle) *begin++ = *d++;
  while (middle != end) *middle++ = *d++;
}

// PRE: [begin, end) is a valid range
// POST: the elements *p, p in [begin, end) are in ascending order
void merge_sort (Vit begin, Vit end)
{
  const int n = end - begin;
  if (n <= 1) return;              // nothing to do
  const Vit middle = begin + n/2;
  merge_sort (begin, middle);      // sort first half
  merge_sort (middle, end);        // sort second half
  merge (begin, middle, end);      // merge both halfs
}

int main()
{
  // input of number of values to be sorted
  int n;
  std::cin >> n;

  std::vector<int> v(n);

  std::cout << "Sorting " << n << " integers...\n";

  // create sequence:
  for (int i=0; i<n; ++i) v[i] = i;
  std::random_shuffle (v.begin(), v.end());

```

```
  // sort into ascending order
  merge_sort (v.begin(), v.end());

  // is it really sorted ?
  for (int i=0; i<n-1;++i)
    if (v[i] != i) std::cout << "Sorting error!\n";

  return 0;
}
```

Program 2.36: *../progs/lecture/merge_sort.cpp*

Since merge_sort incurs much less comparisons than minimum_sort, our unit here is just **Mcomp**, the number of Megacomparisons (10^6 comparisons), according to Theorem 2. In other words, **Mcomp**$= 10^{-6}(n-1)\lceil \log_2 n \rceil$. **Time** is the absolute runtime of the program, this time in milliseconds and not minutes. But as in Table 7, **sec/Gcomp** tells us how many seconds the program needs to perform one Gigacomparison.

We first observe that this latter number is around 25 for all n, also confirming in this case that the runtime is proportional to the number of comparisons. On the other hand, the time needed by merge_sort per Gcomp is much higher than in minimum_sort (25 seconds for *merge-sort* vs. 3 seconds for *minimum-sort*. It may be surprising that the difference is this large, but the fact that it *is* larger can be explained. *Merge-sort* is a more complicated algorithm than *minimum-sort*, with its recursive structure, the extra memory needed for the new deck, etc. The price to pay is that less comparisons can be done per second, since a lot of time is needed for other operations. But this is a moderate price, since we can more than pay for it by the gain in total runtime.

This gets us to the most positive news of Table 8: *Merge-sort* is actually a very practical sorting algorithm. While it takes *minimum-sort* more than one hour to process $1,600,000$ numbers, *merge-sort* does the same in less than a second. Our experimental results show that this is mainly due to the fact that *merge-sort* needs much less comparisons (at most $(n-1)\lceil \log_2 n \rceil$) than *minimum-sort* with its $n(n-1)/2$ comparisons. Still, *merge-sort* is not the best sorting algorithm in practice, see Exercise 119.

2.9.7 Lindenmayer systems

In this final section we want to present another application in which recursion is predominant and difficult to avoid (an iterative version would indeed require an explicit stack). As a bonus, this applications lets us draw beautiful pictures.

Let us first fix an *alphabet* Σ which is simply a finite set of symbols, for example $\Sigma = \{F, +, -\}$. Let Σ^* denote the set of all *words* that we can form from symbols in Σ. For example, $F + F + \in \Sigma^*$.

Next, we fix a function $P : \Sigma \to \Sigma^*$. P maps every symbol to a word, and these are the *productions*. We might for example have the productions

Figure 20: *The turtle before and after processing the command sequence* $\mathsf{F}+\mathsf{F}+$

σ	$\mapsto$	$P(\sigma)$
F	$\mapsto$	$\mathsf{F}+\mathsf{F}+$
$+$	$\mapsto$	$+$
$-$	$\mapsto$	$-$

Finally, we fix an *initial word* $s \in \Sigma^*$, for example $s = \mathsf{F}$.

The triple $\mathcal{L} = (\Sigma, P, s)$ is called a *Lindenmayer system*. Such a system generates an infinite sequence of words $s = w_0, w_1, \ldots$ as follows. To get the next word w_i from the previous word w_{i-1}, we simply substitute all symbols in w_{i-1} by their productions.

In our example, this yields

$$\begin{aligned}
w_0 &= \mathsf{F}, \\
w_1 &= \mathsf{F}+\mathsf{F}+ \\
w_2 &= \mathsf{F}+\mathsf{F}++\mathsf{F}+\mathsf{F}++ \\
w_3 &= \mathsf{F}+\mathsf{F}++\mathsf{F}+\mathsf{F}+++\mathsf{F}+\mathsf{F}++\mathsf{F}+\mathsf{F}+++ \\
&\vdots
\end{aligned}$$

The next step is to "draw" these words, and this gives the pictures we were talking about.

Turtle graphics. Imagine a turtle sitting at some point p on a large piece of paper, with its head pointing in some direction, see Figure 20 (left). The turtle can understand the commands F, $+$, and $-$. F means "move one step forward", $+$ means "turn counter-clockwise by an angle of 90 degrees", and $-$ means "turn clockwise by an angle of 90 degrees". The turtle can process any sequence of such commands, by executing them one after another. We are interested in the resulting path taken by the turtle on the piece of paper. The path generated by the command sequence $\mathsf{F}+\mathsf{F}+$, for example, is shown in Figure 20 (right), along with the position and orientation of the turtle after processing the command sequence.

The turtle can therefore graphically interpret any word generated by a Lindenmayer system over the alphabet $\{\mathsf{F}, +, -\}$.

Recursively drawing Lindenmayer systems. For $\sigma \in \Sigma$, let w_i^σ denote the word resulting from σ by the i-fold substitution of all symbols according to their productions. In our running example, we have for example $w_2 = w_2^F = F+F++F+F++$ and $w_i^+ = +$ for all i.

The point is now that can we express w_i^σ in terms of the w_{i-1}'s of other symbols, and this is where recursion comes into play. Suppose that $P(\sigma) = \sigma_1 \cdots \sigma_k$. Then we can obtain w_i^σ as follows. We first substitute σ by $\sigma_1 \cdots \sigma_k$ (1-fold substitution), and in the resulting word $\sigma_1 \cdots \sigma_k$ we apply $(i-1)$-fold substitution to all the symbols. This shows that

$$w_i^\sigma = w_{i-1}^{\sigma_1} \cdots w_{i-1}^{\sigma_k}.$$

This formula also implies that the drawing of w_i^σ is obtained by simply concatenating the drawings for $w_{i-1}^{\sigma_1}, \ldots, w_{i-1}^{\sigma_k}$. To get the actual word w_i, we simply concatenate the drawings of all w_i^σ, for σ running through the symbols of the initial word s.

Program 2.37 shows how this works for our running example with productions $F \mapsto F+F+, + \mapsto +, - \mapsto -$ and initial word F. Since $P^i(+) = +, P^i(-) = -$ for all i, we do not need to substitute $+$ and $-$ and get

$$w_i = w_i^F = w_{i-1}^F + w_{i-1}^F + . \tag{2.2}$$

The program assumes the existence of a library `turtle` with predefined turtle command functions `forward`, `left` (counterclockwise rotation with some angle) and `right` (clockwise rotations with some angle) in namespace `ifmp`.

In the documentation of the program, we have omitted the "trivial" productions $+ \mapsto +, - \mapsto -$, and in specifying a Lindenmayer system, we can do so as well: we will usually only list productions for symbols that are not mapped to themselves.

```
// Prog: lindenmayer.cpp
// Draw turtle graphics for the Lindenmayer system with
// production F -> F+F+ and initial word F.

#include <iostream>
#include <IFMP/turtle>

// POST: the word w_i^F is drawn
void f (const unsigned int i) {
  if (i == 0)
    ifmp::forward();   // F
  else {
    f(i-1);            // w_{i-1}^F
    ifmp::left(90);     // +
    f(i-1);            // w_{i-1}^F
    ifmp::left(90);     // +
  }
```

```
}

int main () {
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // draw w_n = w_n(F)
  f(n);

  return 0;
}
```

Program 2.37: *../progs/lecture/lindenmayer.cpp*

For input $n = 14$, the program will produce the following drawing.

As n gets larger, the picture does not seem to change much; it rotates, and some more details develop, but apart from that the impression is the same. Assume you could draw the picture for $n = \infty$. Then equation (2.2) would give

$$w_\infty = w_\infty + w_\infty + .$$

This is a *self-similarity*: the drawing of w_∞ consists of two rotated drawings of itself. We have a *fractal*!

Additional features. We can extend the definition of a Lindenmayer system to include a rotation angle α that may be different from 90 degrees. This is shown in Program 2.38 that draws a snowflake for input $n = 5$.

```
// Prog: snowflake.cpp
// Draw turtle graphics for the Lindenmayer system with
// production F -> F-F++F-F, initial word F++F++F and
// rotation angle 60 degrees.
#include <iostream>
#include <IFMP/turtle>

// POST: the word w_i^F is drawn
void f (const unsigned int i) {
  if (i == 0)
    ifmp::forward();  // F
  else {
    f(i-1);           // w_{i-1}^F
    ifmp::right(60);  // -
    f(i-1);           // w_{i-1}^F
    ifmp::left(120);  // ++
    f(i-1);           // w_{i-1}^F
    ifmp::right(60);  // -
    f(i-1);           // w_{i-1}^F
  }
}

int main () {
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // draw w_n = w_n^F++w_n^F++w_n^F
  f(n);               // w_n^F
```

```
  ifmp::left(120);     // ++
  f(n);               // w_n^F
  ifmp::left(120);     // ++
  f(n);               // w_n^F

  return 0;
}
```

Program 2.38: *../progs/lecture/snowflake.cpp*

To get more flexibility, we can also extend the alphabet Σ of symbols. For example, we may add symbols without any graphical interpretation; these are still useful, though, since they may be used in productions. For example, the Lindenmayer system with $\Sigma = \{F, +, -, X, Y\}$, initial word X and productions

$$\begin{aligned} X &\mapsto X + YF + \\ Y &\mapsto -FX - Y \end{aligned}$$

yields the *dragon curve* (w_{14}, angle of 90 degrees).

The corresponding code is shown in Program 2.39.

```
// Prog: dragon.cpp
// Draw turtle graphics for the Lindenmayer system with
// productions X -> X+YF+, Y -> -FX-Y, initial word X
// and rotation angle 90 degrees
#include <iostream>
#include <IFMP/turtle>

// necessary: x and y call each other
void y (const unsigned int i);

// POST: w_i^X is drawn
```

```
void x (const unsigned int i) {
  if (i > 0) {
    x(i-1);            // w_{i-1}^X
    ifmp::left(90);     // +
    y(i-1);            // w_{i-1}^Y
    ifmp::forward();    // F
    ifmp::left(90);     // +
  }
}

// POST: w_i^Y is drawn
void y (const unsigned int i) {
  if (i > 0) {
    ifmp::right(90);    // -
    ifmp::forward();    // F
    x(i-1);            // w_{i-1}^X
    ifmp::right(90);    // -
    y(i-1);            // w_{i-1}^Y
  }
}

int main () {
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // draw w_n = w_n^X
  x(n);

  return 0;
}
```

Program 2.39: *../progs/lecture/dragon.cpp*

Finally, one can add symbols *with* graphical interpretation. Commonly used symbols are f (jump one step forward, this doesn't leave a trace), [(remember current position) and] (jump back to last remembered position). Our turtle library understands these three symbols as well. It is also typical to add new symbols with the same interpretation as F, say.

2.9.8 Details

Lindenmayer systems. Lindenmayer systems are named after the Hungarian biologist Aristide Lindenmayer (1925–1985) who proposed them in 1968 to model the growth of plants. Lindenmayer systems (with generalizations to 3-dimensional space) have found many ap-

plications in computer graphics.

2.9.9 Goals

Dispositional. At this point, you should ...

1) understand the concept of recursion, and why it makes sense to define a function through itself;
2) understand the semantics of recursive function calls and be aware that they do not always terminate;
3) appreciate the power of recursion in sorting and drawing Lindenmayer systems.

Operational. In particular, you should be able to ...

(G1) find pre- and postconditions for given recursive functions;

(G2) prove or disprove termination and correctness of recursive function calls;

(G3) translate recursive mathematical function definitions into C++ function definitions;

(G4) rewrite a given recursive function in iterative form;

(G5) recognize inefficient recursive functions and improve their performance;

(G6) count the number of operations of a given type in a recursive function call, using induction as the main tool;

(G7) write recursive functions for given tasks.

(G8) be able to compare algorithms with respect to their practical performance.

2.9.10 Exercises

Exercise 111 *Find pre- and postconditions for the following recursive functions.* (G1)

a)
```
bool f (const int n)
{
  if (n == 0) return false;
  return !f(n-1);
}
```

b)
```
void g (const unsigned int n)
{
  if (n == 0) {
    std::cout << "*";
    return;
  }
```

```
    g(n-1);
    g(n-1);
  }
```

c)
```
unsigned int h (const unsigned int n, const unsigned int b) {
  if (n == 1) return 0;
  return 1 + h (n / b, b);
}
```

Exercise 112 *Prove or disprove for any of the following recursive functions that it terminates for all possible arguments. In this theory exercise, overflow should not be taken into account, i.e. you should pretend that the value range of* `unsigned int` *is equal to* $\mathbb{N}$. (G2)

a)
```
unsigned int f (const unsigned int n)
{
  if (n == 0) return 1;
  return f(f(n-1));
}
```

b)
```
// POST: return value is the Ackermann function value A(m,n)
unsigned int A (const unsigned int m, const unsigned int n) {
  if (m == 0) return n+1;
  if (n == 0) return A(m-1,1);
  return A(m-1, A(m, n-1));
}
```

c)
```
unsigned int f (const unsigned int n, const unsigned int m)
{
  if (n == 0) return 0;
  return 1 + f ((n + m) / 2, 2 * m);
}
```

Exercise 113 *Consider the following recursive function defined on all nonnegative integers, also known as* McCarthy's 91 Function.

$$M(n) :== \begin{cases} n - 10, & \textit{if } n > 100 \\ M(M(n+11)), & \textit{if } n \leq 100. \end{cases}$$

a) *Provide a C++ function* `mccarthy` *that implements McCarthy's 91 Function.*

b) *What are the values of the following four function calls?*

   (i) `mccarthy(101)`

(iii) `mccarthy(100)`

(iii) `mccarthy(99)`

(iv) `mccarthy(91)`

c) *Explain why the function is called McCarthy's 91 Function! More precisely, what is the value of* $M(n)$ *for any given number* n*?*

(G3)(G7)

Exercise 114

a) *Write and test a C++ function that computes binomial coefficients* $\binom{n}{k}, n, k \in \mathbb{N}$*. These may be defined in various equivalent ways. For example,*

$$\binom{n}{k} := \frac{n!}{k!(n-k)!},$$

or

$$\binom{n}{k} := \begin{cases} 0, & \text{if } n < k \\ 1, & \text{if } n = k \text{ or } k = 0 \\ \binom{n-1}{k} + \binom{n-1}{k-1}, & \text{if } n > k, k > 0 \end{cases},$$

or

$$\binom{n}{k} := \begin{cases} 0, & \text{if } n < k \\ 1, & \text{if } n \geq k, k = 0 \\ \frac{n}{k}\binom{n-1}{k-1} & \text{if } n \geq k, k > 0 \end{cases}$$

b) *Which of the three variants is best suited for the implementation, and why? Argue theoretically, but underpin your arguments by comparing at least two different implementations of the function.*

(G3)(G5) (G7)

Exercise 115 *In how many ways can you own CHF 1 with bank notes and coins? Despite its somewhat philosophical appearance, the question is a mathematical one. Given some amount of money, in how many ways can you partition it using the available denominations (bank notes and coins)? Today's denominations in CHF are 1000, 200, 100, 50, 20, 10 (banknotes), 5, 2, 1, 0.5, 0.2, 0.1, 0.05(coins). The amount of CHF 0.2, for example, can be owned in four ways (to get integers, let's switch to centimes):* $(20), (10, 10), (10, 5, 5), (5, 5, 5, 5)$*. The amount of CHF 0.04 can be owned in no way with bank notes and coins, while there is exactly one way to own CHF 0 (you cannot have* 4 *centimes in your wallet, but you* can *have no money at all in your wallet).*

Solve the problem for a given input amount, by writing a program `partition.cpp` *that defines the following function (all values to be understood as centimes).*

```
// PRE: [begin, end) is a valid nonempty range that describes
//      a sequence of denominations d_1 > d_2 > ... > d_n > 0
// POST: return value is the number of ways to partition amount
//       using denominations from d_1, ..., d_n
unsigned int partitions (unsigned int amount,
                         const unsigned int* begin,
                         const unsigned int* end);
```

Use your program to determine in how many ways you can own CHF 1, and CHF 10. Can your program compute the number of ways for CHF 50? For CHF 100? (G7)(G5)

Exercise 116 *Suppose you want to crack somebody's secret code, consisting of* d *digits between* 1 *and* 9*. You have somehow found out that exactly* k *of these digits are* 1*'s.*

a) *Write a program that generates all possible codes. The program should contain a function that solves the problem for given arguments* d *and* k*.*

b) *Adapt the program so that it also outputs the number of possible codes.*

For example, if $d = 2$ *and* $k = 1$*, the output may look like this:*

```
12 13 14 15 16 17 18 19 21 31 41 51 61 71 81 91
There were 16 possible codes.
```

(G7)

Exercise 117 *Rewrite the following recursive function in iterative form and test with a program whether your iterative version is correct. What can you say about the absolute runtimes of both variants for values of* n *up to* 100*, say?* (G4)(G5)

```
unsigned int f (const unsigned int n)
{
  if (n <= 2) return 1;
  return f(n-1) + 2 * f(n-3);
}
```

Exercise 118 *Rewrite the following recursive function in iterative form and test with a program whether your iterative version is correct. What can you say about the runtimes of both variants for values of* n *up to* 100*, say?* (G4)(G5)

```
unsigned int f (const unsigned int n)
{
  if (n == 0) return 1;
  return f(n-1) + 2 * f(n/2);
}
```

Exercise 119 *Modify Program 2.36 (*`merge_sort.cpp`*) such that it calls the sorting function* `std::sort` *of the standard library instead of the function* `merge_sort`*. For this, you need the header* `algorithm` *that is already being included. Store the resulting program as* `std_sort.cpp` *and compare the total runtimes of both programs an larger inputs on your platform (start with* $n = 1,600,000$ *numbers and then keep doubling the input size).*

What do you observe? How many sec/Gcomp does `merge_sort` *take? Can you reproduce the "25" of Table 8, or do you get a different constant (explained by your computer having a speed different from the one of the authors)? How much faster (in absolute runtime) is* `std::sort` *on your platform? And is the speedup independent of* n*?* (G8)

Exercise 120 *The following function finds an element with a given value* x *in a sorted sequence (if there is such an element), using* binary search.

```
typedef std::vector<int>::const_iterator Cvit;

// PRE: [begin, end) is a valid range, and the elements *p,
//      p in [begin, end) are in ascending order
// POST: return value is an iterator p in [begin, end) such
//       that *p = x, or the pointer end, if so such pointer
//       exists
Cvit bin_search (const Cvit begin, const Cvit end, const int x)
{
  const int n = end - begin;
  if (n == 0) return end;          // empty range
  if (n == 1) {
    if (*begin == x)
      return begin;
    else
      return end;
  }
  // n >= 2
  const Cvit middle = begin + n/2;
  if (*middle > x) {
    // x can't be in [middle, end)
    const Cvit p = bin_search (begin, middle, x);
    if (p == middle)
      return end; // x not found
    else
      return p;
  } else
    // *middle <= x; we may skip [begin, middle)
    return bin_search (middle, end, x);
}
```

What is the maximum number $T(n)$ *of comparisons between sequence elements and* x *that this function performs if the number of sequence elements is* n*? Try to find an upper bound on* $T(n)$ *that is as good as possible. (You may use the statement of Exercise 121.)* (G6)

Exercise 121 *For any natural number* $n \geq 2$*, prove the following two (in)equalities.* (G6)

$$\lceil \log_2 \lfloor \frac{n}{2} \rfloor \rceil \leq \lceil \log_2 \lceil \frac{n}{2} \rceil \rceil = \lceil \log_2 n \rceil - 1.$$

Exercise 122 *Write programs that produce turtle graphics drawings for the following Lindenmayer systems* (Σ, P, s)*.* (G7)

a) $\Sigma = \{F, +, -\}$, $s = F + F + F + F$ and P given by

$$F \mapsto FF + F + F + F + F + F - F.$$

b) $\Sigma = \{X, Y, +, -\}$, $s = Y$, and P given by

$$\begin{aligned} X &\mapsto Y + X + Y \\ Y &\mapsto X - Y - X. \end{aligned}$$

For the drawing, use rotation angle $\alpha = 60$ degrees and interpret *both* X and Y as "move one step forward".

c) Like b), but with the productions

$$\begin{aligned} X &\mapsto X + Y + +Y - X - -XX - Y + \\ Y &\mapsto -X + YY + +Y + X - -X - Y. \end{aligned}$$

Exercise 123 *The* Towers of Hanoi *puzzle (that can actually be bought from shops) is the following. There are three wooden pegs labeled* $1, 2, 3$*, where the first peg holds a stack of* n *disks, stacked in decreasing order of size, see Figure 21.*

The goal is to transfer the stack of disks to peg 3, by moving one disk at a time from one peg to another. The rule is that at no time, a larger disk may be on top of a smaller one. For example, we could start by moving the topmost disk to peg 2 *(move* $(1, 2)$*), then move the next disk from peg* 1 *to peg* 3 *(move* $(1, 3)$*), then move the smaller disk from peg* 2 *onto the larger disk on peg* 3 *(move* $(2, 3)$*), etc.*

Write a program `hanoi.cpp` *that outputs a sequence of moves that does the required transfer, for given input* n*. For example, if* $n = 2$*, the above initial sequence* $(1, 2)(1, 3)(2, 3)$ *is already complete and solves the puzzle. Check the correctness of your program by hand at least for* $n = 3$*, by manually reproducing the sequence of moves on a piece of paper (or an actual Tower of Hanoi, if you have one).* (G7)

Figure 21: *The Tower of Hanoi*

2.9.11 Challenges

Exercise 124 *This challenge lets you take a look behind the scenes of GIMPS (the* **G***reat* **I***nternet* **M***ersenne* **P***rime* **S***earch,* `www.mersenne.org`*). This project is concerned with the problem of finding large Mersenne primes, see Section 1.1. The essential tool used by GIMPS is the* Lucas-Lehmer *test, a simple procedure to check whether a number of the form* $2^n - 1$ *is prime or not.*

Find out what the Lucas-Lehmer test is and write a program that employs it to find all Mersenne primes of the form $2^n - 1$ *with* $n \leq 1,000$ *(you may go higher if you can). The resulting list will allow you to rediscover the five errors that Mersenne made in his original conjecture (Section 1.1), and it will tell you what happens next after* $n = 257$*, the largest exponent covered by Mersenne's conjecture.*

Exercise 125 *On the occasion of major sports events, the Italian company Panini sells stickers to be collected in an album. For the EURO 2008 soccer championship, the collection comprised of 555 different stickers, available in packages of five stickers each.*

When buying a package, you cannot see which stickers it contains. The company only guarantees that each package contains five different stickers. Let us assume that each possible selection of five different stickers is equally likely to be contained in any given package. How many packages do you need to buy on average in order to have all the stickers?

For the case of EURO 2008 with 555 stickers, a newspaper claimed (based on consulting a math professor) that this number is 763. How did the professor arrive at that number, and is it correct?

Write a program that computes the average number of packages that you need to buy for a collection of size n*. (As a simple check, you should get one package on average if* $n = 5$*). What do you get for* $n = 555$*?*

Note: *In order to solve this challenge in a mathematically sound way, you need some basic knowledge of probability theory. But for our purposes, it is also ok to just handwave why your program is correct.*

Exercise 126 *Lindenmayer systems can also be used to draw (quite realistic) plants, with the growth process simulated by the various iterations. For this, however, there must be a possibility of creating branches. Let us therefore enhance our default set* $\{\mathsf{F}, +, -\}$ *of symbols with fixed meaning and now use* $\Sigma = \{\mathsf{F}, +, -, [,], \mathsf{f}\}$. *The symbol* [*is defined to have the following effect: the current state of the turtle (position and direction) is put on top of the* state stack *which is initially empty. The symbol*] *sets the state of the turtle back to the one found on top of the state stack, and removes the top state from the stack. This mechanism can be used to remember a certain state and return to it later.*

For example, if the rotation angle is 45 *degrees, the word* $\mathsf{FF}[+\mathsf{F}][-\mathsf{F}]$ *produces the drawing of Figure 22.*

Figure 22: *The turtle after processing the command sequence* $\mathsf{FF}[+\mathsf{F}][-\mathsf{F}]$

This does not look like a very sophisticated plant yet, but if you for example try the production

$$\mathsf{F} \quad \mapsto \quad \mathsf{FF} + [+\mathsf{F} - \mathsf{F} - \mathsf{F}] - [-\mathsf{F} + \mathsf{F} + \mathsf{F}]$$

with initial word F, *rotation angle* 22 *degrees, and four iterations, you will see what we mean.*

It remains to explain what the symbol f *means. It has the same effect on the state of the turtle as* F, *except that it does not draw a line. You can imagine that* f *makes the turtle "jump".*

Here are the functions of the library `turtle` *that correspond to this additional functionality.* `jump` *realizes* f, *while* `save` *and* `restore` *are for* [*and*]. *In order to draw Figure 22, we can therefore use the following statements.*

```
ifm::forward(2);
ifm::save();
ifm::left(45);
ifm::forward();
ifm::restore();
ifm::save();
ifm::right(45);
```

```
ifm::forward();
ifm::restore();
```

Here you see that we can provide an integer to `forward` *telling it how many steps we want to move forward (the default that we always used before is* 1*).*

Now here comes the challenge: write a turtle graphics program `amazing.cpp` *that will knock our socks off! In other words, we are asking for the most beautiful picture that you can produce using the recursive drawing scheme on top of the turtle graphics commands introduced so far (there are still more commands that are more or less common, but our turtle library stops at* $\Sigma = \{F, +, -, [,], f\}$*).*

If you think that you can submit a crappy program and still earn full points, you're right. But we count on your sportsmanship to give your best!

Chapter 3

Classes

3.1 Structs

> *A POD-struct is an aggregate class that has no nonstatic data members of type pointer to member, non-POD struct, non-POD union (or array of such types) or reference, and has no user-defined copy-assignment operator and no user-defined destructor.*
>
> *Section 9, paragraph 4, of the ISO/IEC Standard 14882 (C++ Standard)*

In this section, we show how structs *are used to group data and to obtain new types with application-specific functionality. You will also see how* operator overloading *can help in making new types easy and intuitive to use.*

Suppose we want to use *rational numbers* in a program, i.e., numbers of the form n/d, where both the numerator n and the denominator d are integers. C++ does not have a fundamental type for rational numbers, so we have to implement it ourselves.

We could of course represent a rational number simply by two values of type `int`, but this would not be in line with our perception of the rational numbers as a distinct mathematical concept. The two numbers n and d "belong together", and this is also reflected in mathematical notation: the symbol $\mathbb{Q}$ for the set of rational numbers indicates that we are dealing with a mathematical type, defined by its value range *and* its functionality (see Section 2.1.6). Ideally, we would like to get a C++ type that can be used like existing arithmetic types; the following piece of code (for adding two rational numbers) shows how this could look like.

```
// input
std::cout << "Rational number r:\n";
rational r;
std::cin >> r;

std::cout << "Rational number s:\n";
rational s;
std::cin >> s;

// computation and output
std::cout << "Sum is " << r + s << ".\n";
```

And we would like a sample run of this to look as follows.

```
Rational number r:
1/2
Rational number s:
```

```
1/3
Sum is 5/6.
```

In C++ this can be done. C++ offers the possibility of defining new types based on existing types. In this section, we introduce as a first step the concept of *structs*. A struct is used to aggregate several values of different types into one value of a new type. With this, we can easily model the mathematical type $\mathbb{Q}$ as a new type in C++. Here is a working program that makes a first step toward the desired piece of code above.

```
// Program: userational.cpp
// Add two rational numbers.
#include <iostream>

// the new type rational
struct rational {
  int n;
  int d; // INV: d != 0
};

// POST: return value is the sum of a and b
rational add (const rational a, const rational b)
{
  rational result;
  result.n = a.n * b.d + a.d * b.n;
  result.d = a.d * b.d;
  return result;
}

int main ()
{
  // input
  std::cout << "Rational number r:\n";
  rational r;
  std::cout << " numerator =?   "; std::cin >> r.n;
  std::cout << " denominator =? "; std::cin >> r.d;

  std::cout << "Rational number s:\n";
  rational s;
  std::cout << " numerator =?   "; std::cin >> s.n;
  std::cout << " denominator =? "; std::cin >> s.d;

  // computation
  const rational t = add (r, s);

  // output
  std::cout << "Sum is " << t.n << "/" << t.d << ".\n";
```

```

  return 0;
}
```

Program 3.1: *../progs/lecture/userational.cpp*

In C++, a struct defines a new type whose value range is the *Cartesian product* of a fixed number of types. (Here and in the following, we identify a type with its value range to avoid clumsy formulations.) In our case, we define a new type named `rational` whose value range is the Cartesian product `int`$\times$`int`, where we interpret a value (n, d) as the quotient n/d.

Since there is no type for the denominator with the appropriate value range `int`$\setminus\{0\}$, we specify the requirement $d \neq 0$ by an informal *invariant*, a condition that has to hold for all legal combinations of values. Such an invariant is indicated by a comment starting with

```
// INV:
```

Like pre- and postconditions of functions (see Section 2.6.1), invariants are an informal way of documenting the program; they are not standardized, and our way of writing them is one possible convention.

The type `rational` is referred to as a *struct*, and it can be used like any other type; for example, it may appear as argument type and return type in functions like add.

A struct defines a type, not variables. Let's get rid of one possible confusion right from the beginning. The definition

```
struct rational {
  int n;
  int d; // INV: d != 0
};
```

does *not* define variables `n` and `d` of type `int`, although the two middle lines look like variable declarations as we know them. Rather, all four lines together define a *type* of the name `rational`, but at that point, neither a variable of that new type, nor variables of type `int` have been defined. The two middle lines

```
  int n;
  int d; // INV: d != 0
```

specify that every actual *object* of the new type (i.e. a concrete rational number) "has" (is represented by) two objects of type `int` that can be accessed through the names `n` and `d`; see the member access below. This specification is important if we want to implement operations on our new type like in the function add.

Here is an analogy for the situation. If the university administration wants to specify how a student is represented in their files, they might come up with three pieces of data that are necessary: a name, an identification number, and a program of study. This

defines the "type" of a student and allows functionality (registration, change of program of study, etc.) to be realized, long before any students actually show up.

3.1.1 Struct definitions.

In general, a struct definition looks as follows.

```
struct T {
  T1 name1;
  T2 name2;
  ...
  TN nameN;
};
```

Here, T is the name of the newly introduced struct (this name must be an identifier, Section 2.1.10), and $T1,\ldots,TN$ are names of existing types. These are called the *underlying types* of T. The identifiers *name1*, *name2*,..., *nameN* are the *data members* of the new type T.

The value range of T is $T1 \times T2 \times \ldots \times TN$. This means, a value of type T is an N-tuple $(t_1, t_2, \ldots, t_N)$ where $t_i \in Ti$.

"Existing types" might be fundamental types, but also user-defined types. For example, consider the vector space $\mathbb{Q}^3$ over the field $\mathbb{Q}$. Given the type `rational` as above, we could model $\mathbb{Q}^3$ as follows.

```
struct rational_vector_3 {
  rational x;
  rational y;
  rational z;
};
```

Although it follows from the definition, let us make it explicit: the types $T1,\ldots,TN$ need not be the same. Here is an example: If $0, 1, \ldots, U$ is the value range of the type `unsigned int`, we can get a variant of the type `int` with value range

$$\{-U, -U+1, \ldots, -1, 0, 1, \ldots, U-1, U\}$$

as follows.

```
struct extended_int {
  // represents u if n==false and -u otherwise
  unsigned int u;  // absolute value
  bool         n;  // sign bit
};
```

The value range of this type is $\{0, 1, \ldots, U\} \times \{\mathit{true}, \mathit{false}\}$, but like in the rational case, we interpret values differently: a value (u, n) "means" u if $n = \mathit{false}$ and $-u$ if $n = \mathit{true}$.

Even if two struct definitions have the same *member specification* (the part of the definition enclosed in curly braces), they define *different* types, and it is not possible to replace one for the other. Consider this trivial but instructive example with two apparently equal structs defined over an empty set of existing types.

```
struct S {
};

struct T {
};

void foo (const S s) {}

int main() {
  S s;
  T t;
  foo (s); // ok
  foo (t); // error: type mismatch
  return 0;
}
```

It is also possible to use array members in structs. For example, the field $\mathbb{Q}^3$ that we have discussed above could alternatively be modeled like this.

```
struct rational_vector_3 {
  rational v[3];
};
```

3.1.2 Structs and scope

The scope of a struct is the part of the program in which it can be used (in a variable declaration, or as a formal function argument type, for example). Structs behave similar to functions here: the scope of a struct is the union of the scopes of all its *declarations*, where a struct declaration has the form

struct T

The struct definition is a declaration as well, and usually one actually needs the definition in order to use a struct. This is easy to explain: in order to translate a variable declaration of struct type, or a function with formal arguments of struct type into machine language, the compiler needs to know the amount of memory required by an object of the struct. But this information is only obtainable from the definition of the struct; as long as the compiler has only seen a declaration of T, the struct T is said to have *incomplete type*.

3.1.3 Member access

A struct is more than the Cartesian product of its underlying types—it offers some basic functionality on its own that we explain next. The most important (and also most visible) functionality of a struct is the access to the data members, and here is where the identifiers *name1*,..., *nameN* come in. If *expr* is an expression of type T with value $(t_1, \ldots, t_N)$, then t_K—the K-th component of its value—can be accessed as

expr.nameK

Here, '.' is the *member access operator* (see Table 9 in the Appendix for its specifics). The composite expression *expr.nameK* is an lvalue if *expr* itself is an lvalue, and we say that the data member *nameK* is *accessed for expr*.

Lines 25 and 26 of Program 3.1 assign values to the rational numbers `r` through the member access operator, while line 37 employs the member access operator to output the value of the rational number `t`. The additional output of '/' indicates that we interpret the 2-tuple (n, d) as the quotient n/d.

3.1.4 Initialization and assignment

We can initialize objects of struct type and assign values to them, just like we do it for fundamental types.

In line 34 of Program 3.1 for example, the variable `t` of type `rational` is initialized with the value of the expression `add (r, s)`. In a struct, initialization is quite naturally done member-wise, i.e. for each data member separately. Under the hood, the declaration statement

```
const rational t = add (r, s);
```

therefore has the effect of initializing `t.n` (with the first component of the value of `add (r, s)`) and `t.d` (with the second component). Interestingly, this also works with array members. Structs therefore provide a way of forging array initialization and assignment by wrapping the array into a struct. Here is an example to show what we mean.

```
#include<iostream>

struct point {
  double coord[2];
};

int main()
{
  point p;
  p.coord[0] = 1;
  p.coord[1] = 2;
```

```
  point q = p;
  std::cout << q.coord[0] << " "     // 1
            << q.coord[1] << "\n";   // 2

  return 0;
}
```

In the same way (memberwise initialization), the formal arguments a and b of the function add are initialized from the values of r and s; the value of add (r, s) itself also results from an initialization of a (temporary) object when the return statement of the function add is executed.

Instead of the above declaration statement that initializes t we could also have written

```
rational t;
t = add (r, s);
```

Here, t is *default-initialized* first, and this default-initializes the data members. In our case, they are of type int; for fundamental types, default-initialization does nothing, so the values of the data members are undefined after default-initialization (see also Section 2.1.8). In the next line, the value of add (r, s) is assigned to t, and this assignment again happens member-wise.

What about other operations? For every fundamental type T, two expressions of type T can be tested for equality, using the operators == and !=. It would therefore seem natural to have these two operators also for structs, implemented in such a way that they compare member-wise.

Formally, this would be correct: if $t = (t_1, \ldots, t_N)$ and $t' = (t'_1, \ldots t'_N)$, then we have $t = t'$ if and only if $t_K = t'_K$ for $K = 1, \ldots, N$.

But our type rational already shows that this won't work: under member-wise equality, we would erroneously conclude that $2/3 \neq 4/6$. The problem is that the *syntactical value range* int×int of the type rational does not coincide with the *semantical value range* in which we identify pairs (n, d) that define the same rational number n/d.

The same happens with our type extended_int from above: since both pairs $(0, \mathit{false})$ and $(0, \mathit{true})$ are interpreted as 0, member-wise equality would give us "$0 \neq 0$" in this case.

Only the implementor of a struct knows the semantical value range, and for this reason, C++ neither provides equality operators for structs, nor any other operations beyond the member access, initialization, and assignment discussed above. Operations that respect the semantical value range can be provided by the implementor, though, see next section.

You might argue that even member-wise initialization and assignment could be inconsistent with the semantics of the type. Later, we will indeed encounter such a situation,

and we will show how it can be dealt with elegantly.

3.1.5 User-defined operators

New types require new operations, but when it comes to the naming of such operations, one less nice aspect of Program 3.1 shows in line 34. By defining the function add, we were able to perform the operation $t := r + s$ through the statement

```
const rational t = add (r, s);
```

Ideally, however, we would like to add rational numbers like we add integers or floating-point numbers, by simply writing (in our case)

```
const rational t = r + s;
```

The benefit of this might not be immediately obvious, in particular since the naming of the function add seems to be quite reasonable; but consider the expression

```
const rational t = subtract (multiply (p, q), multiply (r, s));
```

and its "natural" counterpart

```
const rational t = p * q - r * s;
```

to get an idea what we mean.

The natural notation can indeed be achieved: a key feature of the C++ language is that most of its operators (see Table 9 in the Appendix for the full list) can be *overloaded* to work for other types as well. This means that we can use the same operator token to implement various operators: we "overload" the token.

In principle, this is nothing new: we already know that the binary operator + is available for several types, for example `int` and `double`. What is new is that we can add even more overloads on our own, and simply let the compiler figure out from the call argument types which one is needed in a certain context.

In overloading an operator, we cannot change the operator's arity, precedence or associativity, but we can create versions of it with arbitrary formal argument and return types.

Operator overloading is simply a special case of *function overloading*. For example, having the structs `rational` and `extended_int` available, we could declare the following two functions in the same program, without creating a name clash: for any call to the function square in the program, the compiler can find out from the call argument type which of the two functions we mean.

```
// POST: returns a * a
rational square (rational a);

// POST: returns a * a
extended_int square (extended_int a);
```

Function overloading in general is useful, but not nearly as useful as operator overloading. To define an overloaded operator, we have to use the *functional operator notation*. In this notation, the name of the operator is obtained by appending its token to the prefix operator. In case of the binary addition operator for the type `rational`, this looks as follows and replaces the function add.

```
// POST: return value is the sum of a and b
rational operator+ (const rational a, const rational b)
{
  rational result;
  result.n = a.n * b.d + a.d * b.n;
  result.d = a.d * b.d;
  return result;
}
```

In Program 3.1, we can now replace line 34 by

```
const rational t = r + s; // equivalent to operator+ (r, s);
```

Here, the comment refers to the fact that an operator can also be *called* in functional notation; in contrast, it appears in *infix notation* in `r + s`. The call in functional notation can be useful for didactic purposes, since it emphasizes the fact that an operator is simply a special function; in an application, however, the point is to avoid functional notation and use the infix notation.

The other three basic arithmetic operations are similar, and here we only give their declarations.

```
// POST: return value is the difference of a and b
rational operator- (rational a, rational b);

// POST: return value is the product of a and b
rational operator* (rational a, rational b);

// POST: return value is the quotient of a and b
// PRE:  b != 0
rational operator/ (rational a, rational b);
```

We can also overload the unary - operator; in functional operator notation, it has the same name as the binary version, but it has only one instead of two arguments. In the following implementation, we use the (modified) "local copy" of the call argument a as the return value.

```
// POST: return value is -a
rational operator- (rational a)
{
  a.n = -a.n;
  return a;
}
```

In order to compare rational numbers, we need the relational operators as well. Here is the equality operator as an example.

```
// POST: return value is true if and only if a == b
bool operator== (const rational a, const rational b)
{
  return a.n * b.d == a.d * b.n;
}
```

3.1.6 Details

Overloading resolution. If there are several functions or operators of the same name in a program, the compiler has to figure out which one is meant in a certain function call. This process is called *overloading resolution* and only depends on the types of the call arguments. Overloading resolution is therefore done at compile time. There are two cases that we need to consider: we can either have an *unqualified* function call (like `add (r, s)` in Program 3.1), or a *qualified* function call (like `std::sqrt(2.0)`). To process an unqualified function call of the form

fname (*expression1*, ..., *expressionN*)

the compiler has to find a matching function declaration. Candidates are all functions f of name *fname* such that the function call is in the scope of some declaration of f. In addition, the number of formal arguments must match the number of call arguments, and each call argument must be of a type whose values can be converted to the corresponding formal argument types.

In a qualified function call of the form

X::*fname* (*expression1*, ..., *expressionN*)

where X is a namespace, only this namespace is searched for candidates.

Argument-dependent name lookup (Koenig lookup). There is one special rule that sometimes makes the list of candidates larger. If some call argument type of an unqualified function call is defined in a namespace X (for example the namespace `std`), then the compiler also searches for candidates in X. This is useful mainly for operators and allows them to be called unqualified in infix notation. The point of using operators in infix notation would be spoiled if we had to mention a namespace somewhere in the operator call.

Resolution: Finding the best match. For each candidate function and each call argument, it is checked how well the call argument type matches the corresponding formal argument type. There are four quality levels, going from better to worse, given in the following list.

(1) EXACT MATCH. The types of the call argument and the formal argument are the same.

(2) PROMOTION MATCH. There is a promotion from the call argument type to the formal argument type. We have seen some examples for promotions, like from `bool` to `int` and from `float` to `double`.

(3) STANDARD CONVERSION MATCH. There is a standard conversion from the call argument type to the formal argument type. We have seen that all fundamental arithmetic types can be converted into each other by standard conversions.

(4) USER-DEFINED CONVERSION MATCH. There is a user-defined conversion from the call argument type to the formal argument type. We will get to user-defined conversions only later in this book.

A function f is called *better* than g with respect to an argument, if the match that f induces on that argument is at least as good as the match induced by g. If the match is really better, f is called *strictly better* for the argument.

A function f is called a *best match* if it is better than any other candidate g in all arguments, and strictly better than g in at least one argument.

Under this definition, there is at most one best match, but it may happen that there is no best match, in which case the function call is *ambiguous*, and the compiler issues an error message.

Here is an example. Consider the two overloaded function declarations

```
void foo(double d);
void foo(unsigned int u);
```

In the code fragment

```
float f = 1.0f;
foo(f);
```

the first overload is chosen, since `float` can be promoted to `double`, but only standard-converted to `unsigned int`. In

```
int i = 1;
foo(i);
```

the call is ambiguous, since `int` can be standard-converted to both `double` and `unsigned int`.

3.1.7 Goals

Dispositional. At this point, you should ...

1) know how structs can be used to aggregate several different types into one new type;
2) understand the difference between the syntactical and semantical value range of a struct;
3) know that C++ functions and operators can be overloaded.

Operational. In particular, you should be able to ...

(G1) define structs whose semantical value ranges correspond to that of given mathematical sets;

(G2) provide definitions of functions and overloaded operators on structs, according to given functionality;

(G3) write programs that define and use structs according to given functionality.

3.1.8 Exercises

Exercise 127 *A struct may have data members of struct type that may have data members of struct type, and so on. It can be quite cumbersome to set all the data members of the data members of the data members manually whenever an object of the struct type needs to be initialized. A preferable solution is to write a function that does this. As an example, consider the type*

```
struct rational_vector_3 {
  rational x;
  rational y;
  rational z;
};
```

where `rational` *is as before defined as*

```
struct rational {
  int n;
  int d;  // INV: d != 0
};
```

Write a function

```
rational_vector_3 create_rational_vector_3
   (int n1, int d1, int n2, int d2, int n3, int d3)
```

that returns the rational vector $(n_1/d_1, n_2/d_2, n_3/d_3)$. *The function should also make sure that* d_1, d_2, d_3 *are nonzero.* (G3)

Exercise 128 *Define a type* `Tribool` *for three-valued logic; in three-valued logic, we have the truth values true, false, and unknown.*

For the type `Tribool`*, implement the logical operators*

```
// POST: returns x AND y
Tribool operator&& (Tribool x, Tribool y);

// POST: returns x OR y
Tribool operator|| (Tribool x, Tribool y);
```

where AND ($\wedge$) and OR ($\vee$) are defined according to the following two tables. (G1)(G2)(G3)

$\wedge$	*false*	*unknown*	*true*
false	*false*	*false*	*false*
unknown	*false*	*unknown*	*unknown*
true	*false*	*unknown*	*true*

$\vee$	*false*	*unknown*	*true*
false	*false*	*unknown*	*true*
unknown	*unknown*	*unknown*	*true*
true	*true*	*true*	*true*

Test your type by writing a program that outputs these truth tables in some format of your choice.

Exercise 129 *Define a type* `Z_7` *for computing with integers modulo 7. Mathematically, this corresponds to the finite ring $\mathbb{Z}_7 = \mathbb{Z}/7\mathbb{Z}$ of residue classes modulo 7.*

For the type `Z_7`*, implement addition and subtraction operators*

```
// POST: return value is the sum of a and b
Z_7 operator+ (Z_7 a, Z_7 b);

// POST: return value is the difference of a and b
Z_7 operator- (Z_7 a, Z_7 b);
```

according to the following table (this table also defines subtraction: $x - y$ is the unique number $z \in \{0, \ldots, 6\}$ such that $x = y + z$). (G1)(G2)

+	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

Exercise 130 *Provide definitions for the following binary arithmetic operators on the type* `rational`. (G2)(G3)

```
// POST: return value is the difference of a and b
rational operator- (rational a, rational b);

// POST: return value is the product of a and b
rational operator* (rational a, rational b);

// POST: return value is the quotient of a and b
// PRE:  b != 0
rational operator/ (rational a, rational b);
```

Exercise 131 *(a) Provide definitions for the following binary relational operators on the type* `rational`*. (b)Provide a main function to test the correctness of your implementation. In doing this, try to reuse operators that are already defined.* (G2)(G3)

```
// POST: return value is true if and only if a != b
bool operator!= (rational a, rational b);

// POST: return value is true if and only if a < b
bool operator< (rational a, rational b);

// POST: return value is true if and only if a <= b
bool operator<= (rational a, rational b);

// POST: return value is true if and only if a > b
bool operator> (rational a, rational b);

// POST: return value is true if and only if a >= b
bool operator>= (rational a, rational b);
```

Exercise 132 *Provide definitions for the following binary arithmetic operators on the type* `extended_int` *(Page 249), and test them in a program (for that it could be helpful to provide an output facility for the type* `extended_int`*, and a function that assigns to an* `extended_int` *value a value of type* `int`*). As in the previous exercise, try to reuse code.* (G2)(G3)

```
// POST: return value is the sum of a and b
extended_int operator+ (extended_int a, extended_int b);

// POST: return value is the difference of a and b
extended_int operator- (extended_int a, extended_int b);

// POST: return value is the product of a and b
extended_int operator* (extended_int a, extended_int b);

// POST: return value is -a
extended_int operator- (extended_int a);
```

Exercise 133 *Consider the following set of three functions.*

```
void foo(double, double)       { ... }  // function A
void foo(unsigned int, int)    { ... }  // function B
void foo(float, unsigned int)  { ... }  // function C
```

For each of the following function calls, decide to which of the functions (A, B, C) *it resolves to, or decide that the call is ambiguous. Explain your decisions! This*

exercise requires you to read the paragraph on overloading resolution in the Details section.

a) `foo(1, 1)`

b) `foo(1u, 1.0f)`

c) `foo(1.0, 1)`

d) `foo(1, 1u)`

e) `foo(1, 1.0f)`

f) `foo(1.0f, 1.0)`

3.1.9 Challenges

Exercise 134 *This challenge has a computer graphics flavor. Write a program that allows you to visualize and manipulate a 3-dimensional object. For the sake of concreteness, think of a* wireframe model *of a cube given by 12 edges in threedimensional space.*

The program should be able to draw the object in perspective view and at least provide the user with a possibility of rotating the object around the three axes. The drawing window might for example look like in Figure 23.

Instead of a cube, you may want to take another platonic solid, you may read the wireframe model from a file, you may add the possibility of scaling the object, translating it, etc. Use the library `libwindow` *that is available at the course homepage to create the graphical output.*

If you don't know (or have forgotten) how to rotate and project threedimensional points, here is a crash course.

Rotating a point $(x, y) \in \mathbb{R}^2$ *around the origin by an angle of* α *(radians) results in the point* (x', y') *with coordinates*

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

In order to rotate a point $(x, y, z) \in \mathbb{R}^3$ *around the* z*-axis, you simply keep* z *unchanged and rotate* (x, y) *as above. By symmetry, you can figure out how this works for the other axes.*

General perspective projection is not so easy, but if you want to project a point onto the $z = 0$*-plane (imagine that this plane is the computer screen that you want to draw on), this is not hard. Imagine that* $v = (v_x, v_y, v_z)$ *is the viewpoint (position of your eye).* $v_z > 0$ *for example means that you are sitting in front of the*

Figure 23: *Sample drawing window for Exercise 134*

screen. When you project the point $p = (x, y, z)$ *onto the screen, the image point has coordinates*

$$(x - t(v_x - x), y - t(v_y - y)),$$

where

$$t = \frac{z}{v_z - z}.$$

The projection thus only works if $v_z \neq z$.

3.2 Reference Types

Whereas Europeans generally pronounce my name the right way ('Ni-klows Wirt'), Americans invariably mangle it into 'Nick-les Worth'. This is to say that Europeans call me by name, but Americans call me by value.

Attributed to Niklaus Wirth

This section explains reference types *that enable functions to accept and return lvalues and in particular change the values of their call arguments. Reference types naturally come up in implementing functionality for structs.*

3.2.1 The trouble with `operator+=`

Let us try to implement the addition assignment operator += for the struct `rational` from Program 3.1. Here is an attempt:

```
rational operator+= (rational a, const rational b) {
  a.n = a.n * b.d + a.d * b.n;
  a.d *= b.d;
  return a;
}
```

With this, we can write

```
rational r;
r.n = 1; r.d = 2;                          // 1/2

rational s;
s.n = 1; s.d = 3;                          // 1/3

r += s;
std::cout << r.n << "/" << r.d << "\n";
```

You may already see that the output of this will not be the desired 5/6. Recall from Section 2.6.3 what happens when `r += s` (equivalently, `operator+= (r, s)`) is evaluated: `r` and `s` are evaluated, and the resulting values are used to initialize the formal arguments `a` and `b` of the function `operator+=`. The values of `r` and `s` are not changed by the function call.

Hence, with the above implementation of `operator+=`, the *value* of the expression `r += s` is indeed 5/6, but the desired *effect*, the change of `r`'s value, does not happen. That's why we get 1/2 as output in the above piece of code.

In order to implement operator+= properly, we must enable functions to change the values of their call arguments. The careful reader may remember that pointer types (Section 2.8.2) can indirectly be used to achieve this, but with a somewhat clumsy syntax and potentially unsafe behavior due to pointers "going nowhere", see Exercise 100.

Reference types are very similar to pointer types in their philosophy, but in most contexts, they are more natural to use and at the same time more safe, since they render the use of explicit (and possibly invalid) pointers unnecessary.

3.2.2 Reference Types: Definition

If T is any type, then

T&

is the corresponding *reference type* (read *T&* as "T reference" or "reference to T"). In value range and functionality, *T&* is identical to T. The difference is only in the initialization and assignment semantics. An expression of reference type is called a *reference*.

A variable of reference type *T&* can be initialized only from an *lvalue* of type T. The initialization makes it an *alias* of the lvalue: another name for the object behind the lvalue. We also say that the reference *refers* to that object. The following example shows this.

```
int i = 5;
int& j = i;                 // j becomes an alias of i

j = 6;                      // changes the values of i
std::cout << i << "\n";     // outputs 6
```

Declarations such as

```
int& j;                     // error: no object to refer to
int& k = 5;                 // error: 5 is not an lvalue
```

are invalid.

A reference behaves like (and is implemented) by a constant pointer to the object that it refers to, a pointer that is dereferenced whenever the reference is evaluated. But the pointer and its value—the address of the object—are hidden. Here is the explicit-pointer variant of the above code fragment that changes the value of i to 6:

```
int i = 5;
int* const j = &i;          // j points to i

*j = 6;                     // changes the values of i
std::cout << i << "\n";     // outputs 6
```

Apart from the syntax, the main difference between the two declarations

```
int& j = lvalue
```

and

```
int* const j = ptr
```

is that in the former case, the constant pointer hidden behind the reference j is *guaranteed* to represent a valid address, namely the one of *lvalue*. In the latter case, the explicit constant pointer j may or may not be valid, depending on whether the pointer *ptr* is valid. From a safety point of view, working with references is therefore preferable to working with pointers.

The fact that a reference is implemented by a constant pointer implies that a reference cannot be changed to refer to another object after initialization. If we later assign something to the reference, we in fact assign to the *object* referred to by it. In writing j = 6 in the above piece of code, we therefore change the value of i to 6, since j is an alias of i.

Every reference is an lvalue itself. We can therefore use a reference *r* to initialize another reference *r'*, but then we don't get a reference to *r*, but another reference to the object referred to by *r*:

```
int i = 5;
int& j = i;  // j becomes an alias of i
int& k = j;  // k becomes another alias of i
```

3.2.3 Call by value and call by reference

When a function has a formal argument of reference type, the corresponding call argument must be an lvalue; when the function call is evaluated, the initialization of the formal argument makes it an alias of the call argument. In this way, we can implement functions that change the values of their call arguments. Here is an example.

```
void increment (int& i)
{
  ++i;
}

int main ()
{
  int j = 5;
  increment (j);
  std::cout << j << "\n"; // outputs 6

  return 0;
}
```

If a formal argument of a function has reference type, we have *call-by-reference* semantics with respect to that argument. Equivalently, we say that we *pass* the argument by reference. Another frequently used term for call by reference is *call by name*.

If the formal argument is not of reference type, we have *call-by-value* semantics: we pass the argument by value. Under call by reference, the address of (or a reference to) the call argument is used to initialize the formal argument; under call-by-value semantics, it is the value of the call argument that is used for initialization.

The basic rule is to pass an argument by reference only if the function in question actually needs to change the call argument value. If that is not the case, call by value is more flexible, since it allows a larger class of call arguments (lvalues *and* rvalues instead of lvalues only).

3.2.4 Return by value and return by reference

The return type of a function can be a reference type as well, in which case we have *return-by-reference* semantics (otherwise, we *return by value*). If the function returns a reference, the function call expression is an lvalue itself, and we can use it wherever lvalues are expected.

This means that the function itself chooses (by using reference types or not) whether its call arguments and return value are lvalues or rvalues. Section 2.1.14 and Section 2.2.4 document these choices for some of the operators on fundamental types, but only now we understand the mechanism that makes such choices possible.

As a concrete example, let us consider the following version of the function `increment` that exactly models the behavior of the pre-increment operator ++: it increments its lvalue argument and returns it as an lvalue.

```
int& increment (int& i)
{
  return ++i;
}
```

In general, we must make sure that an expression of reference type that we return refers to a *non-temporary* object. To understand what a temporary object is, let us consider the following function.

```
int& foo (const int i)
{
  return i;
}
```

This is asking for trouble, since the formal argument `i` runs out of scope when the function call terminates. This means that the associated memory is freed and the address expires (see Section 2.4.3). If we now write for example

```
int i = 3;
int& j = foo(i);          // j refers to expired object
std::cout << j << "\n"; // undefined behavior
```

the reference `j` refers to an expired object, and the resulting behavior of the program is undefined. References are safer than pointers, but not absolutely safe. In working with references, we need to adhere to the following guideline.

Reference Guideline: Whenever you create an alias for an object, ensure that the object does not expire before the alias.

The compiler usually notices violations of the Reference Guideline and issues a warning.

3.2.5 More user-defined operators

Rational numbers: addition assignment. Let's get back to the addition assignment operator for our new struct `rational`. In order to fix our failed attempt from the beginning of this section, we need to add two characters only.

As in the previous function `increment`, the formal argument `a` must be passed as a reference, and to be compliant with the usual semantics of `+=`, we also return the result as a reference:

```
// POST: b has been added to a; return value is the new value of a
rational& operator+= (rational& a, const rational b)
{
  a.n = a.n * b.d + a.d * b.n;
  a.d *= b.d;
  return a;
}
```

The other arithmetic assignment operators are similar, and we don't list them here explicitly. Together with the arithmetic and relational operators discussed in Section 3.1.5, we now have a useful set of operations on rational numbers.

Rational numbers: input and output. Let us look at Program 3.1 once more, with the function name `add` replaced by `operator+` and the function call `add (r, s)` replaced by `r + s`. Still, we can spot potential improvements: instead of writing

```
std::cout << "Sum is " << t.n << "/" << t.d << "\n";
```

in line 37, we'd rather write

```
std::cout << "Sum is " << t << "\n";
```

just like we are doing it for fundamental types.

From what we have done above, you can guess that all we have to do is to overload the output operator `<<`. In discussing the output operator in Section 2.1.14 we have argued that the output stream passed to and returned by the output operator must be an lvalue, since the output operator modifies the stream. Having reference types at our disposal, this can easily be done: we simply pass and return the output stream (whose type is `std::ostream`) as a reference:

```
// POST: a has been written to o
std::ostream& operator<< (std::ostream& o, const rational r)
{
  return o << r.n << "/" << r.d;
}
```

There is no reason to stop here: for the input, we would in the same fashion like to replace the two input statements `std::cin >> r.n;` and `std::cin >> r.d;` by the single statement

```
std::cin >> r;
```

(and the same for the input of `s`). Again, we need to pass and return the input stream (of type `std::istream`) as a reference. In addition, we must pass the rational number that we want to read as a reference, since the input operator has to modify its value.

The operator first reads the numerator from the stream, followed by a separating character, and finally the denominator. Thus, we can read a rational number in one go by entering for example 1/2.

```
// POST: r has been read from i
// PRE:  i starts with a rational number of the form "n/d"
std::istream& operator>> (std::istream& i, rational& r)
{
  char c; // separating character, e.g. '/'
  return i >> r.n >> c >> r.d;
}
```

In contrast to `operator<<`, things can go wrong, e.g., if the user enters the character sequence "`A/B`" when prompted for a rational number. Also, we probably don't want to accept `3.4` as a rational number as our input operator does. There are mechanisms to deal with such issues, but we won't discuss them here.

Let us conclude this section with a beautified version of Program 3.1 (see Program 3.2 below). What makes this version even nicer is the fact that the new type is used exactly as fundamental type such as `int`. All the operations necessary for this are outsourced into Program 3.3.

In the spirit of Section 2.6.7 on modularization, we should actually split the program into three files: a file `rational.h` that contains the definition of the struct `rational`, along with declarations of the overloaded operators; a file `rational.cpp` that contains the definitions of these operators; and finally, a file `userational2.cpp` that contains the main program. At the same time, we should put our new type `rational` and the operations on it into namespace `ifmp` in order to avoid possible name clashes. (You can still write the expression `r + s` in Program 3.2, without mentioning the namespace in which the `operator+` in question is defined. The Details of Section 3.1 explains this in the paragraph on argument-dependent name lookup.) Exercise 143 asks you to do all this and integrate the new rational number type into the math library that you have built in Exercise 86, so that Program 3.2 below can be compiled using this library.

```
// Program: userational2.cpp
// Add two rational numbers.
#include <iostream>
#include "rational.cpp"

int main ()
{
  // input
  std::cout << "Rational number r:\n";
  rational r;
  std::cin >> r;

  std::cout << "Rational number s:\n";
  rational s;
  std::cin >> s;

  // computation and output
  std::cout << "Sum is " << r + s << ".\n";

  return 0;
}
```

Program 3.2: *../progs/lecture/userational2.cpp*

```
// Program: rational.cpp
// Define a type rational and operations on it

// the new type rational
struct rational {
  int n;
  int d; // INV: d != 0
};

// POST: b has been added to a; return value is the
// new value of a
rational& operator+= (rational& a, const rational b) {
  a.n = a.n * b.d + a.d * b.n;
  a.d *= b.d;
  return a;
}

// POST: return value is the sum of a and b
rational operator+ (const rational a, const rational b)
{
```

```
  // reduce to operator +=
  rational result = a;
  return result += b;
}

// POST: return value is true if and only if a == b
bool operator== (const rational a, const rational b)
{
  return a.n * b.d == a.d * b.n;
}

// POST: a has been written to o
std::ostream& operator<< (std::ostream& o, const rational a)
{
  return o << a.n << "/" << a.d;
}

// POST: a has been read from i
// PRE:  i starts with a rational number of the form "n/d"
std::istream& operator>> (std::istream& i, rational& a)
{
  char c; // separating character, e.g. '/'
  return i >> a.n >> c >> a.d;
}
```

Program 3.3: *../progs/lecture/rational.cpp*

Here is an example run of Program 3.2, showing that we have now achieved what we set out to do in the beginning of Section 3.1.

```
Rational number r:
1/2
Rational number s:
1/3
Sum is 5/6.
```

3.2.6 Const references

Let us come back to the addition operator for rational numbers defined in Program 3.3:

```
// POST: return value is the sum of a and b
rational operator+ (const rational a, const rational b)
{
  rational result = a;
  return result += b;
}
```

The efficiency fanatic in you might already have anticipated that we can slightly speed up this operator by using *const-qualified* reference types:

```
// POST: return value is the sum of a and b
rational operator+ (const rational& a, const rational& b)
{
  rational result = a;
  return result += b;
}
```

Indeed, this version is potentially faster than the previous one, since the initialization of a formal argument is done by copying just *one* address, rather than *two* `int` values as in the member-wise copy that takes place under the call-by-value semantics. Even if the saving is small in this example, you can imagine that member-wise copy can be pretty expensive in structs that are more elaborate than `rational`; in contrast, call by reference is fast for *all* types, even the most complicated ones.

As usual, the `const` keyword makes sure that the formal arguments cannot be changed within the function body, but in addition, the `const` promise extends to the *object* referred to by the call argument, since the formal argument will become an alias for it. Under call by reference, const-correctness is thus much more important than under call by value. Here is an instructive example. Consider the unary `operator-` from Section 3.1.5:

```
// POST: return value is -a
rational operator- (rational a)
{
  a.n = -a.n;
  return a;
}
```

Blindly speeding this up to

```
rational operator- (rational& a)
{
  a.n = -a.n;
  return a;
}
```

has a drastic (and undesired) consequence: the expression `-a` will still have the same value as before, but it will have the additional effect of negating the value of the call argument. We have "accidentally" created a completely different operator. The fact that we have started out with a non-constant formal argument in the first place is taking revenge when we move to call by reference semantics.

Const-correctness requires that whenever an argument is passed by reference to a non-mutating function, it must have const-qualified reference type. In case of the unary minus, this would have revealed our programming error:

```
rational operator- (const rational& a)
```

```
{
  a.n = -a.n; // error: a was promised to be constant
  return a;
}
```

In compiling this variant of the operator, the compiler will issue an error message, pointing out the mistake. We can then fix it by either going back to call-by-value semantics, or by introducing a result variable like in operator+ above:

```
rational operator- (const rational& a)
{
  rational result = a;
  result.n = -result.n;
  return result;
}
```

Let us discuss the const mechanism in context of references in more detail. The situation is very similar to that of const pointers, see Section 2.8.7.

A *const reference* is a reference for which the *object* referred to is constant. A const reference can for example be obtained as follows.

```
int i = 5;
const int& j = i; // j becomes a non-modifiable alias of i
```

Now, writing j = 6 leads to an error message, since i and its alias j have been promised to be constant.

A const reference behave like (and is implemented by) a constant *const* pointer to the object that it refers to.

Const references of const-qualified type const *T&* can as usual be initialized with lvalues of type *T*, but also with rvalues of type *T*, or of types whose values are convertible to *T*. We can for example write

```
const int& j = 5;
```

Under the hood, the compiler generates a temporary object that holds the value 5, and it initializes j with the address of the temporary object. This is useful in operator+ that can still be called with rvalue arguments under formal arguments of const-qualified reference type const rational&. A formal argument of type const *T&* is therefore the all-in-one device suitable for every purpose: *if* the call argument is an lvalue, the initialization is very efficient (only its address needs to be copied), and otherwise, we essentially fall back to call-by-value semantics.

Despite this, there are still situations where *T* is preferable over const *T&* as argument type. If *T* is a fundamental type or a struct with small memory requirements, it does not pay off to move to const *T&*, since the saving in handling lvalue arguments is so small (or even nonexistent) that it won't compensate for the (slightly) more costly access to the formal function argument in the function body. Indeed, call by reference adds one indirection: to look up the value of a formal function argument under call-by-reference semantics, we *first* have to look up its address and then look up the actual value at that

address. Under call-by-value semantics, the address of the value is "hardwired" (and refers to some object on the call stack, see Section 2.9.2).

Also, it is often convenient to use the formal argument as a local variable and modify its value (see operator- above); for that, its type must not be a const-type.

What exactly is constant? We already had a paragraph with the same title in Section 2.8.7, and what is going on here is also the same: an expression of type const *T&* is the alias of an object, but that object is *not* necessarily constant itself. The const-qualification in this case is merely a promise that the object's value will not be modified *through* the alias in question. Here is an example that illustrates this point.

```
int n = 5;
const int& i = n; // i becomes a non-modifiable alias of n
int& j = n;       // j becomes a modifiable alias of n
i = 6;            // error: n is modified through const-reference
j = 6;            // ok: n receives value 6
```

Here, we do not have a constant *object*, but a constant *alias* (namely i).

3.2.7 Const-types as return types.

Const-types may also appear as return types of functions, just like any other types. In that case, the const promises that the function call expression itself is constant.

It is *not* generally valid to replace return type *T* by const *T&*; while this safely works for the formal argument types, it can for the return type result in syntactically correct but semantically wrong code.

As an example, let's replace rational by const rational& as the return type of operator+, with the hope of realizing another speedup:

```
const rational& operator+ (const rational& a, const rational& b) {
  rational result = a;
  return result += b;
}
```

In executing the return statement, the return value (in this case a const-reference) to be passed to the caller of the function is initialized with the expression result. Now recall that the initialization of a (const-)reference from an lvalue simply makes it an alias of the lvalue. But the lvalue in question (namely result) is a local variable whose memory is freed and whose address becomes invalid when the function call terminates (see Section 2.4.3 and Section 3.2.1). The consequence is that the returned reference will be the alias of an expired object, and using this reference results in undefined behavior of the program. The issue is again the one of temporary objects, see the discussion and the Reference Guideline on 265.

Errors like this are very hard to find (and we cannot reliably count on compiler warnings here), since the program *may* work as intended, for example if the memory

that was associated to the expired object is not immediately reused. But on another platform, the program may behave differently or even crash.

3.2.8 Goals

Dispositional. At this point, you should ...

1) understand the alias concept behind reference types and the Reference Guideline;
2) understand the difference between *call by value* and *call by reference* semantics for function arguments;
3) understand const references.

Operational. In particular, you should be able to ...

(G1) state exact pre-and postconditions for functions involving formal argument types or return types of reference and/or const-type;

(G2) write functions that modify (some of) their call arguments;

(G3) find syntactical and semantical errors in programs that are due to improper handling of reference types;

(G4) find syntactical and semantical errors in programs that are due to improper handling of const-types;

(G5) find the declarations in a given program whose types should be const-according to the Const Guideline.

(G6) understand the fact that using const and non-const reference argument will result in function overloading.

3.2.9 Exercises

Exercise 135 *Consider the following family of functions:*

```
T foo (S i)
{
  return ++i;
}
```

with T being one of the types `int`, `int&` *and* `const int&`, *and S being one of the types* `int`, `const int`, `int&` *and* `const int&`. *This defines 12 different functions, and all of those combinations are syntactically correct.*

a) Find the combinations of T and S for which the resulting function definition is semantically valid, meaning, for example, that the constness of variables and references is respected. Semantical correctness also means that the compiler will accept the code, because we have already established syntactical correctness. Explain your answer.

b) *Among the combinations found in a), find the combinations of T and S for which the resulting function definition is also valid during runtime, meaning that function calls always have well-defined value and effect; explain your answer.*

c) *For all combinations found in b), give precise postconditions for the corresponding function* `foo`.

(G1)(G3)(G4)

Exercise 136 *Write a function that swaps the values of two* `int`*-variables.* (G2)
For example,

```
int a = 5;
int b = 6;
// here comes your function call
std::cout << a << "\n"; // outputs 6
std::cout << b << "\n"; // outputs 5
```

Exercise 137 *We want to have a function that* normalizes *a rational number, i.e. transforms it into the unique representation in which numerator and denominator are relatively prime, and the denominator is positive. For example,*

$$\frac{21}{-14}$$

is normalized to

$$\frac{-3}{2}.$$

There are two natural versions of this function:

```
// POST: r is normalized
void normalize (rational& r);

// POST: return value is the normalization of r
rational normalize (const rational& r);
```

Implement one of them, and argue why you have chosen it over the other one.
Hint: *you may want to use the function* gcd *from Section 2.9, modified for arguments of type* `int` *(how does this modification look like?).* (G2)(G6)

Exercise 138 *Provide a definition of the following function.*

```
// POST: return value indicates whether the linear equation
//       a * x + b = 0 has a real solution x ; if true is
//       returned, the value s satisfies a * s + b = 0
bool solve (double a, double b, double& s);
```

Test your function in a program for at least the pairs (a, b) *from the set*

$$\{(2,1),(0,2),(0,0),(3,-4)\}.$$

(G2)

Exercise 139 *Find all mistakes (if any) in the following programs, and explain why these are mistakes. All programs share the following two function definitions and only differ in their* `main` *functions.* (G1)(G3)

```
int foo (int& i) {
  return i += 2;
}

const int& bar (int &i) {
  return i += 2;
}
```

a)
```
int main()
{
  const int i = 5;
  int& j = foo (i);
}
```

b)
```
int main()
{
  int i = 5;
  const int& j = foo (i);
}
```

c)
```
int main()
{
  int i = 5;
  const int& j = bar (foo (i));
}
```

d)
```
int main()
{
  int i = 5;
  const int& j = foo( bar (i));
}
```

e)
```
int main()
{
  int i = 5;
  const int j = bar (++i);
}
```

Exercise 140 *The C++ standard library also contains a type for computing with* complex numbers. *A complex number where both the real and the imaginary part are* doubles *has type* `std::complex<double>` *(you need to* `#include <complex>` *in order to get this type). In order to get a a complex number with real part* r *and imaginary part* i, *you can use the expression*

```
std::complex<double>(r,i); // r and i are of type double
```

Otherwise, complex numbers work as expected. All the standard operators (arithmetic, relational) and mathematical functions (`std::sqrt`, `std::abs`, `std::pow`*...) are available. The operators also work in mixed expressions where one operand is of type* `std::complex<double>` *and the other one of type* `double`. *Of course, you can also input and output complex numbers.*

Here is the actual exercise. Implement the following function for solving quadratic equations over the complex numbers:

```
// POST: return value is the number of distinct complex solutions
//       of the quadratic equation ax^2 + bx + c = 0. If there
//       are infinitely many solutions (a=b=c=0), the return
//       value is -1. Otherwise, the return value is a number n
//       from {0,1,2}, and the solutions are written to s1,..,sn
int solve_quadratic_equation (std::complex<double> a,
                              std::complex<double> b,
                              std::complex<double> c,
                              std::complex<double>& s1,
                              std::complex<double>& s2);
```

Test your function in a program for at least the triples (a, b, c) *from the set*

$$\{(0,0,0),(0,0,2),(0,2,2),(2,2,2),(1,2,1),(i,1,1)\}. \tag{G2}$$

3.2.10 Challenges

Exercise 141 *Implement the following function for solving cubic equations over the complex numbers:*

```
// POST: return value is the number of distinct (complex) solutions
//       of the cubic equation ax^3 + bx^2 + cx + d = 0. If there
//       are infinitely many solutions (a=b=c=d=0), the return
//       value is -1. Otherwise, the return value is a number n
//       from {0,1,2,3}, and the solutions are written to s1,..,sn
int solve_cubic_equation (std::complex<double> a,
                          std::complex<double> b,
                          std::complex<double> c,
                          std::complex<double> d,
```

```
                                std::complex<double>& s1,
                                std::complex<double>& s2,
                                std::complex<double>& s3);
```

You find a brief description of the type `std::complex<double>` *in the text of Exercise 140.*

Write a program that tests your function. For example, you may substitute the solutions returned by the above function into $ax^3+bx^2+cx+d=0$ *and check whether the expression indeed evaluates to (approximately) zero.*

Hint: *You find the necessary theory under the keyword* Cardano's formula.

3.3 Class types

Let me tell you this in closing
I know we might seem imposing
But trust me if we ever show in your section
Believe me its for your own protection

Will Smith, Men in Black (1997)

This section introduces the concept of classes as an extension of the struct concept from Section 3.1. You will learn about data encapsulation as a distinguishing feature of classes. This feature makes type implementations more safe and flexible. You will first learn classes feature by feature for rational numbers, and then see two complete classes in connection with random number generation.

3.3.1 Encapsulation

In the previous two sections, we have defined a new struct type `rational` whose value range models the mathematical type $\mathbb{Q}$ (the set of rational numbers), and we have shown how it can be equipped with some useful functionality (arithmetic and relational operators, input and output).

To motivate the transition from structs to classes in this section (and in particular the aspect of encapsulation), let us start off with a thought experiment. Suppose you have put the struct `rational` and all the functionality that we have developed into a nice library. In Exercise 143 you have actually done this, for the very basic version of the type `rational` from Program 3.3. Now you have sold the library to a customer; let's call it RAT (*Rational Thinking Inc.*). RAT is initially happy with the functionality that the library provides, and starts working with it. But then some unpleasant issues come up.

Issue 1: Initialization is cumbersome. Some code developed at RAT needs to initialize a new variable `r` with the rational number $1/2$; for this, the programmer in charge must write

```
rational r; // default-initialization of r
r.n = 1;    // assignment to data member
r.d = 2;    // assignment to data member
```

The declaration `rational r` default-initializes `r`, but the actual value of `r` must be provided through *two* assignments later. RAT tell you that they would prefer to initialize `r` from the numerator and denominator in one go, and you realize that they have a point

here. Indeed, if the programmer at RAT forgets one of the assignments, `r` has undefined value (and you get to handle the bug reports). If the struct is larger (consider the example of `rational_vector_3` on page 249), the problem is amplified.

Issue 2: Invariants cannot be guaranteed. Every legal value of the type `rational` must have a nonzero denominator. You have stipulated this as an invariant in Program 3.3, but there is no way of enforcing this invariant. It is possible for anyone to write

```
rational r;
r.n = 1;
r.d = 0;
```

and thus violate the *integrity* of the type, the correctness of the internal representation.

You might argue that it would be quite stupid to write `r.d = 0`, and even the programmer at RAT can't be that stupid. But in RAT's application, the values of rational numbers arise from complicated computations somewhere else in the program; these computations may result in a zero denominator simply by mistake, and in allowing value 0 to be assigned to `r.d`, the mistake further propagates instead of being withdrawn from circulation (again, you get to handle the bug reports).

You think about how both issues could be addressed in the next release of the rational numbers library, and you come up with the following solution: As another piece of functionality on the type `rational`, you define a function that creates a value of type `rational` from two values of type `int`. (We have proposed such a solution already in Exercise 127 in connection with nested structs).

```
// PRE:  d != 0
// POST: return value is n/d
rational create_rational (const int n, const int d) {
  assert (d != 0);
  rational result;
  result.n = n;
  result.d = d;
  return result;
}
```

You then advise RAT to use this function whenever they want to initialize or assign to a rational number. For example,

```
rational r = create_rational (1, 2);
```

would initialize `r` with 1/2 in one go, and at the same time make sure that the denominator is nonzero.

Such a creation function certainly makes sense for structs in general, but the two issues above don't really go away. The reason is that this safe creation can be circumvented by not using it. In fact, your advice might not have reached the programmer at RAT, and even if it did, the programmer might be too lazy to follow it. It is therefore still possible

to write `rational r;` and forget about data member assignment, and it is still possible to assign 0 to `r.d`. Behind this lies in fact a much larger problem, as you discover next.

Issue 3: The internal representation cannot be changed. After having used the rational numbers library for some time, RAT approaches you with a request for a version with a larger value range, since they have observed that intermediate values sometimes overflow.

You recall the type `extended_int` from Page 249 and realize that one thing you could easily do is to change the type of numerator and denominator from `int` to `unsigned int` and store the sign of the rational number separately as a data member of type `bool`. for example like this:

```
struct rational {
  unsigned int n;    // absolute value of numerator
  unsigned int d;    // absolute value of denominator
  bool is_negative; // sign of the rational number
};
```

It is also not too hard to rewrite the library files `rational.h` and `rational.cpp` to reflect this change in representation.

But shortly after you have shipped the new version of your library to RAT (you have even included the safe creation function `create_rational` from above in the hope to resolve issues 1 and 2 above), you receive an angry phone call from the management of RAT: the programmer reports that although the application code still compiles with the new version of the library, *nothing* works anymore!

After taking a quick look at the application code, you suddenly realize what the problem is: the code is cluttered up with expressions of the form *expr*`.n` and *expr*`.d`, as in

```
rational r;
r.n = 1;
r.d = 2;
```

Already this particular piece of code does not work anymore: a rational number is now represented by *three* data members, but the (old) application code obviously does not initialize the (new) member of type `bool`. Now you regret not to have provided the `create_rational` function in the first place; indeed, the statement

```
rational r = create_rational (1, 2);
```

would still work, assuming that you have correctly adapted the definition of the function `create_rational` to deal with the new representation. But the problem is much more far-reaching and manifests itself in each and every occurrence of *expr*`.n` or *expr*`.d` in the application code, since the data members have changed their meaning (they might even have changed their names): in letting RAT access numerator and denominator through data members that are specific to a certain representation, you are now committed to that representation, and you can't change it without asking RAT to change its application code as well (which they will refuse, of course).

When the RAT management realizes that the new rational numbers with extended value range are useless for them, they terminate the contract with you. Disappointed as you are, you still realize that what you need to avoid such troubles in the future is *encapsulation*: a mechanism that *hides* the actual representation of the type `rational` from the customer, and at the same time offers the customer *representation-independent* ways of working with rational numbers.

In C++, encapsulation is available through the use of classes, and we start by explaining how to hide the representation of a type from the customer. In the following, the term "customer" is used in a broader sense for all programs that use the class.

3.3.2 Public and private

Here is a preliminary `class` version of `struct rational` that takes care of data hiding.

```
class rational {
private:
   int n;
   int d; // INV: d!= 0
};
```

In the same way as a `struct`, a `class` aggregates several different types into a new type, but the `class` keyword indicates that *access restriction* may occur, realized through the keywords `public` and `private`.

A data member is *public* if and only if its declaration appears somewhere after a `public:` specifier, and with no `private:` specifier in between. It is *private* otherwise. In particular, if the class definition (see Section 3.3.9 below for the precise meaning of this term) contains no `public:` specifier, all data members are private by default. In contrast, a struct is a class where all data members are public by default.

If a data member is private, it cannot be accessed by customers through the member access operator. If a data member is public, there are no such restrictions. Under our above definition of `class rational`, the following will therefore not compile:

```
rational r;
r.n = 1;      // error: n is private
r.d = 2;      // error: d is private
int i = r.n; // error: n is private
```

In particular, the assignment `r.d = 0` becomes impossible (which is good), but at a (too) high price: now your customer cannot do anything with a rational number, and even *you* cannot implement `operator+=`, say, as you used to do it in Program 3.3. What we are still lacking is some way of accessing the encapsulated representation. This functionality is provided by a second category of class members, namely *member functions*.

3.3.3 Member functions

Let us now add the missing functionality to `class rational` through member functions. It would seem natural to start with safe creation, but since there are specific member functions reserved for this purpose, let us first show two "general" member functions that grant safe access to the numerator and denominator of a rational number (we'll discuss below what `*this` and `const` mean here; and if you wonder why we can use `n` and `d` before they are declared: this is a special feature of class scope, explained in Section 3.3.9).

```
class rational {
public:
   // POST: return value is the numerator of *this
   int numerator () const
   {
     return n;
   }
   // POST: return value is the denominator of *this
   int denominator () const
   {
     return d;
   }
private:
   int n;
   int d; // INV: d!= 0
};
```

If `r` is a variable of type `rational`, for example, the customer can then write

```
int n = r.numerator();     // get numerator of r
int d = r.denominator();   // get denominator of r
```

using the member access operator as for data members. The customer can call these two functions, since they are declared `public`. Access specifiers have the same meaning for member functions as for data members: a private member function cannot be called by the customer. This kind of access to the representation is flexible, since the corresponding member functions can easily be adapted to a new representation; it is also safe, since it is not possible to change the values of the data members through the functions `numerator` and `denominator`. As a general rule of thumb, all data members of a class should be private (otherwise, you encourage the customer to access the data members, with the ugly consequences mentioned in Issue 3 above).

The implicit call argument and *this. In order to call a member function, we need an expression of the class type for which we *access* the function, and this expression (appearing before the `.`) is an *implicit call argument* whose value may or may not be modified by the function call.

Within each member function, the lvalue `*this` refers to this implicit call argument and explains the appearance of `*this` in the postconditions of the two member functions above. It does not explain why an asterisk appears in `*this`, but we will get to this later.

Const member functions. A `const` keyword after the formal argument list of a member function refers to the implicit argument `*this` and therefore promises that the member function call does not change the value (represented by the values of the data members) of `*this`. We call such a member function a *const member function*.

Member function call. The general syntax of a member function call is

expr.fname (*expr1*, ..., *exprN*)

Here, *expr* is an expression of a class type for which a member function called *fname* is declared, *expr1*,..., *exprN* are the call arguments, and `.` is the member access operator. In most cases, *expr* is an lvalue of the class type, typically a variable.

Access to members within member functions. Within the body of a member function `f` of a class, any member (data member of member function) of the same class can be accessed without a prefix *expr*`.`; in this case, we implicitly access it for `*this`. In our example, the expression `n` in the return statement of the member function `numerator` refers to the data member `n` of `*this`. The call `r.numerator()` therefore does what we expect: it returns the numerator of the rational number `r`.

Within member functions, we can also access members for other expressions of the same class type through the member access operator (like a customer would do it). All accesses to class members within member functions of the same class are *unrestricted*, regardless of whether the member in question is public or private. The `public:` and `private:` specifiers are only relevant for the customer, but not for member functions of the class itself.

Member functions are sometimes also referred to as *methods* of the class.

Member functions and modularization. In the spirit of Section 2.6.7, it would be useful to source out the member function definitions, in order to allow separate compilation. This works like for ordinary functions, except that in a member function definition outside of the class definition, the function name must be qualified with the class name. In the header file `rational.h` we would then write only the declarations (as usual within namespace `ifmp`):

```
class rational {
public:
   // POST: return value is the numerator of *this
   int numerator () const;
```

```
   // POST: return value is the denominator of *this
   int denominator () const;
private:
   int n;
   int d; // INV: d!= 0
};
```

The matching definitions would then appear in the source code file `rational.cpp` (again within namespace `ifmp`, and after including `rational.h`) as follows.

```
int rational::numerator () const
{
  return n;
}
int rational::denominator () const
{
  return d;
}
```

3.3.4 Constructors

A constructor is a special member function that provides safe initialization of class values. The name of a constructor coincides with the name of the class, and—this distinguishes constructors from other functions—it does not have a return type, and consequently no return value. A class usually has several constructors, and the compiler figures out which one is meant in a given context (using the rules of overloading resolution, see the Details of Section 3.1).

The syntax of a constructor definition for a class T is as follows.

T (*T1 pname1*, *T2 pname2*, ..., *TN pnameN*)
  : *name1* (*expr1*), ..., *nameM* (*exprM*)
 block

Here, *pname1*,..., *pnameN* are the formal arguments of the constructor. In the *initializer*

: *name1* (*expr1*), ..., *nameM* (*exprM*)

name1,..., *nameM* are data members, and *expr1*,...,*exprM* are expressions of types whose values can be converted to the respective data member types. These values are used to initialize the data members, before *block* is executed, and in the order in which the members are declared in the class. In other words, the order in the initializer is ignored, but it is good practice to use the declaration order here as well. If a data member is not listed in the initializer, it is default-initialized. In the constructor body *block*, we can still set or change the values of some of the data members.

For the type `rational`, here is a constructor that initializes a rational number from two integers.

```
// PRE: d != 0
// POST: *this is initialized with numerator / denominator
rational (const int numerator, const int denominator)
  : n (numerator), d (denominator)
{
  assert (d != 0);
}
```

To use this constructor in a variable declaration, we would for example write

```
rational r (1,2); // initializes r with value 1/2
```

In general, the declaration

T x (*expr1*, ..., *exprN*)

defines a variable *x* of type *T* and at the same time initializes it by calling the appropriate constructor with call arguments *expr1*,..., *exprN*.

The constructor can also be called explicitly as in

```
rational r = rational (1, 2);
```

This initializes `r` not directly from two integers, but from an expression of type `rational` that is constructed by the explicit constructor call `rational(1,2)` (which is of type `rational`).

3.3.5 Default constructor

In Section 3.1.4, we have introduced the term *default-initialization* for the kind of initialization that takes place in declarations like

```
rational r;
```

For fundamental types, default-initialization leaves the value in question undefined, but for class types, the *default constructor* is automatically called to initialize the value. If present, the default constructor is the unique constructor with an empty formal argument list.

By providing a default constructor, we can thus make sure that class type values are *always* properly initialized. In case of the class `rational` (or any arithmetic type), default-initialization with value 0 seems to be the canonical choice, and here is the corresponding default constructor.

```
// POST: *this is initialized with 0
rational ()
  : n (0), d (1)
{}
```

In fact, we *must* provide a default constructor if we want the compiler to accept the declaration `rational r`. This makes class types safer than fundamental types, since it is not possible to circumvent a constructor call in declaring a variable.

The careful reader will notice that there must be an exception to this rule: Program 3.1 in Section 3.1 contains the declaration statement `rational r;` although in that program, the type `rational` is a struct without any constructors. This is in fact the only exception: for a class without any constructors, the default constructor is implicitly provided by the compiler, and it simply default-initializes the data members; if a data member is of class type, this in turn calls the default constructor of the corresponding class. This exception has been made so that structs (which C++ has inherited from its precursor C) fit into the class concept of C++.

3.3.6 User-defined conversions

Constructors with one argument play a special role: they are *user-defined conversions*. For the class `rational`, the constructor

```
// POST: *this is initialized with value i
rational (const int i)
  : n (i), d (1)
{}
```

is a user-defined conversion from `int` to `rational`. Under this constructor, `int` becomes a "type whose values can be converted to `rational`". This for example means that we can provide a call argument of type `int` whenever a formal function argument of type `rational` is expected; in the implicit conversion that takes place, the converting constructor is called. With user-defined conversions, we go beyond the set of *standard conversions* that are built-in (like the one from `int` to `double`), but in contrast to the (sometimes incomplete) standard conversion rules stipulated by the C++ standard, we make the rules ourselves.

There are meaningful user-defined conversions that can't be realized by constructors. For example, if we want a conversion from `rational` to `double`, we can't add a corresponding constructor to the type `double`, since `double` is not a class type. Even conversions to some class type T might not be possible in this way: if T is not "our" type (but comes from a library, say), we cannot simply add a constructor to T. In such situations, we simply tell *our* type how its values should be converted to the target type. The conversion from `rational` to `double`, for example, could be done through a member function named `operator double` like this.

```
// POST: return value is double-approximation of *this
operator double () const
{
  return double(n)/d;
}
```

In general, the member function operator S has implicit return type S and induces a user-defined conversion to the type S that is automatically invoked whenever this is necessary.

3.3.7 Member operators

All functionality of rational numbers that we have previously provided through "global" functions (operator+, operator+=,...) must now be reconsidered, since directly accessing the data members is no longer possible. For example, operator+= as in Program 3.3 needs to change the value of a rational number, but there there is no specific member function that allows us to do this.

The way to go is to realize operator+ as a public member function (a *member operator*), having only one formal argument (for b), and *this taking the role of a. This looks as follows.

```
// POST: b has been added to *this; return value is
//       the new value of *this
rational& operator+= (const rational& b)
{
  n = n * b.d + d * b.n;
  d *= b.d;
  return *this;
}
```

Within this member function, there is no problem in accessing the data members directly, since the access restrictions do not apply to member functions. This version of operator+= is as efficient as the one previously used for struct rational.

Prefer nonmember operators over member operators. You might argue that even operator+ should become a member function of class rational, and indeed, this would probably allow a slightly more efficient implementation (without calling operator+=). There is one important reason to keep this operator global, though, and this has to do with user-defined conversions.

Having the conversion from int to rational that we get through the constructor

```
// POST: *this is initialized with value i
rational (int i);
```

we can for example write expressions like r + 2 or 2 + r, where r is of type rational. In compiling this, the compiler automatically inserts a converting constructor call. Now, having operator+ as a member would remove the second possibility of writing 2 + r. Why? Let's first see what happens when r + 2 is compiled. If operator+ is a member function, then r + 2 "means"

```
r.operator+ (2)
```

In compiling this, the compiler inserts the conversion from the call argument type `int` to the formal argument type `rational` of `operator+`, and everything works as expected. `2 + r`, however, would mean

```
2.operator+ (r)
```

which makes no sense whatsoever. If we write a binary operator as a member function, then the first call argument *must* be of the respective class type. Implicit conversions do not work here: they only adapt call arguments to formal argument types of concrete functions, but they cannot be expected to "find" the class whose `operator+` has to be applied.

3.3.8 Nested types

There is a third category of class members, and these are *nested types*. To motivate these, let us come back to Issue 3 above, the one concerning the internal representation of rational numbers. If you think about consequently hiding the representation of a rational number from the customer, then you probably also want to hide the numerator and denominator type. As indicated in the example, these types might internally change, but in the member functions `numerator` and `denominator`, you still promise to return `int`-values.

A better solution would be to promise only a type with certain properties, by saying for example that the functions `numerator` and `denominator` return an integral type (Section 2.2.9). Then you can internally change from one integral type to a different one without annoying the customer. Technically, this can be done as follows.

```
class rational {
public:
   // nested type for numerator and denominator
   typedef int rat_int;
   ...
   // realize all functionality in terms of rat_int
   // instead of int, e.g.
   rational (rat_int numerator, rat_int denominator); // constructor
   rat_int numerator() const;                         // numerator
   ...
private:
   rat_int n;
   rat_int d; // INV: d!= 0
};
```

In customer code, this can be used for example like this.

```
typedef rational::rat_int rat_int;
rational r (1,2);
rat_int numerator = r.numerator();        // 1
rat_int denominator = r.denominator();    // 2
```

We already see one of the properties that the nested type `rational::rat_int` must have in order for this to work. For example, values of type `int` must be convertible to it. If you have set up everything cleanly, you can now for example replace the line

```
typedef int rat_int;
```

by the lines

```
typedef ifm::integer rat_int;
```

and thus immediately get exact rational numbers without any overflow issues; see also Exercise 142.

In the above code, we are using two typedef declarations (Page 201) in order to hide the actual type from the customer. Within the class `rational`, the typedef declaration introduces a nested type `rat_int`, a new name for the type `int`. In the customer code, the class's nested type (that can be accessed using the scope operator, if the nested type declaration is public) receives a new (shorter) name.

In real-life C++ code, there are nested types of nested types of nested types,..., and typenames tend to get very long due to this. The typedef mechanism allows us to keep our code readable.

3.3.9 Class definitions

We now have seen the major ingredients of a class. Formally, a class definition has the form

```
class T {
  class−element ... class−element
};
```

where T is an identifier. The sequence of *class-element*'s may be empty. Each *class-element* is an *access specifier* (`public:` or `private:`), or a *member declaration*. A member declaration is a declaration statement that typically declares a member function, a data member, or a nested type. Collectively, these are called *members* of the class, and their names must be identifiers. A class definition introduces a new type, and this type is called a *class type*, as opposed to a fundamental type.

A member function definition is a declaration as well, but if the class definition does not contain the definition of a member function, this function must have a matching definition somewhere else (see Section 3.3.3). All member function definitions together form the *class implementation*.

Class Scope. Any member declaration of a class is said to have *class scope*. Its declarative region is the class definition. Class scope differs from local scope (Section 2.4.3) in one aspect. The potential scope of a member declaration is not only the part of the class definition "below" the declaration, but it spans the *whole* class definition, *and* the

formal argument lists and bodies of all member function definitions. In short, a class member can be used "everywhere" in the class.

If two class definitions form disjoint declarative regions, there is no problem in using the same name for members of both classes.

3.3.10 Random numbers

We now have all the means to put together a complete and useful implementation of the type `rational` as a class in C++; but since we have already seen most of the necessary code in Section 3.1 and in this section, we leave this as Exercise 142 and continue here with a fresh class that has a little more entertainment in store.

Playing games on the computer would be pretty boring without some unpredictability: a chess program should not always come up with the same old moves in reaction to your same old moves, and in an action game, the enemies should not always pop up at the same time and location. In order to achieve unpredictability, the program typically uses a *random number generator*. This term is misleading, though, since the numbers are in reality generated according to some fixed rule, in such a way that they *appear* to be random. But for many purposes (including games), this is completely sufficient, and we call such numbers *pseudorandom*.

Linear congruential generators. A simple and widely-used technique of getting a sequence of pseudorandom numbers is the *linear congruential method*. Given a *multiplier* $a \in \mathbb{N}$, an *offset* $c \in \mathbb{N}$, a *modulus* $m \in \mathbb{N}$ and a *seed* $x_0 \in \mathbb{N}$, let us consider the sequence $x_1, x_2, \ldots$ of natural numbers defined by the rule

$$x_i = (ax_{i-1} + c) \bmod m, \quad i > 0.$$

A small example is the pseudorandom number generator `knuth8`, defined by the following parameters.

$$a = 137, \quad c = 187, \quad m = 2^8 = 256, \quad x_0 = 0.$$

The sequence $x_1, x_2, \ldots$ of numbers that we get from this is

187, 206, 249, 252, 151, 138, 149, 120, 243, 198, 177, 116, 207, 130, 77, 240, 43, 190, 105, 236, 7, 122, 5, 104, 99, 182, 33, 100, 63, 114, 189, 224, 155, 174, 217, 220, 119, 106, 117, 88, 211, 166, 145, 84, 175, 98, 45, 208, 11, 158, 73, 204, 231, 90, 229, 72, 67, 150, 1, 68, 31, 82, 157, 192, 123, 142, 185, 188, 87, 74, 85, 56, 179, 134, 113, 52, 143, 66, 13, 176, 235, 126, 41, 172, 199, 58, 197, 40, 35, 118, 225, 36, 255, 50, 125, 160, 91, 110, 153, 156, 55, 42, 53, 24, 147, 102, 81, 20, 111, 34, 237, 144, 203, 94, 9, 140, 167, 26, 165, 8, 3, 86, 193, 4, 223, 18, 93, 128, 59, 78, 121, 124, 23, 10, 21, 248, 115, 70, 49, 244, 79, 2, 205, 112, 171, 62, 233, 108, 135, 250, 133, 232, 227, 54, 161, 228, 191, 242, 61, 96, 27, 46, 89, 92, 247, 234, 245, 216, 83, 38, 17, 212, 47, 226, 173, 80, 139,

30, 201, 76, 103, 218, 101, 200, 195, 22, 129, 196, 159, 210, 29, 64, 251, 14, 57, 60, 215, 202, 213, 184, 51, 6, 241, 180, 15, 194, 141, 48, 107, 254, 169, 44, 71, 186, 69, 168, 163, 246, 97, 164, 127, 178, 253, 32, 219, 238, 25, 28, 183, 170, 181, 152, 19, 230, 209, 148, 239, 162, 109, 16, 75, 222, 137, 12, 39, 154, 37, 136, 131, 214, 65, 132, 95, 146, 221, 0, 187, ...

From here on, the sequence repeats itself (in general, the period can never be longer than m). But until this point, it appears to be pretty random (although a closer look reveals that it is not random at all; do you discover a striking sign of nonrandomness?).

In order to make the magnitude of the random numbers independent from the modulus, it is common practice to *normalize* the numbers so that they are real numbers in the interval $[0, 1)$.

Program 3.4 below contains the definition of a class `random` in namespace `ifmp` for generating normalized pseudorandom numbers according to the linear congruential method. There is a constructor that allows the customer to provide the parameters a, c, m, x_0, and a member function `operator()` to get the respective next element in the sequence of the x_i.

```
// Prog: random.h
// define a class for pseudorandom numbers.

namespace ifmp {
  // class random: definition
  class random {
  public:
    // POST: *this is initialized with the linear congruential
    //       random number generator
    //           x_i = ( a * x_{i-1} + c) mod m
    //       with seed x0.
    random(const unsigned int a, const unsigned int c,
           const unsigned int m, const unsigned int x0);

    // POST: return value is the next pseudorandom number
    //       in the sequence of the x_i, divided by m
    double operator()();

  private:
    const unsigned int a_; // multiplier
    const unsigned int c_; // offset
    const unsigned int m_; // modulus
    unsigned int xi_;      // current sequence element
  };
} // end namespace ifmp
```

Program 3.4: *../progs/lecture/random.h*

The function `operator()` has no arguments in our case (that's why its declaration is `operator()()`, which admittedly looks a bit funny), and it overloads the *function call operator*, see Table 9 in the Appendix. In general, if *expr* is an expression of some class type which has the member function

`operator()`(*T1 name1*,..., *TN nameN*)

then *expr* can be used like a function: the expression

expr (*expr1*,..., *exprN*)

is equivalent to a call of the member function `operator()` with arguments *expr1*,...,*exprN* for the expression *expr*. We will see such calls in Program 3.7 and Program 3.8 below.

Here is the implementation of the class `random` in which we see how `operator()` updates the value of the data member `x_i` to be the respective next element in the sequence of the x_i. Since the class implementation appears outside of the class definition (actually, in a separate file), we need to qualify all member functions with the name of the class to which they pertain. Hence, the class implementation needs to contain a definition of the function `random::operator()`.

```
// Prog: random.cpp
// implement a class for pseudorandom numbers.

#include <IFMP/random.h>

namespace ifmp {
  // class random: implementation
  random::random(const unsigned int a, const unsigned int c,
                 const unsigned int m, const unsigned int x0)
    : a_(a), c_(c), m_(m), xi_(x0)
  {}

  double random::operator()()
  {
    // update xi acording to formula,...
    xi_ = (a_ * xi_ + c_) % m_;
    // ...normalize it to [0,1), and return it
    return double(xi_) / m_;
  }
} // end namespace ifmp
```

Program 3.5: *../progs/lecture/random.cpp*

Many commonly used random number generators are obtained in exactly this way. For example, the well-known generator `drand48` returns pseudorandom numbers in $[0, 1)$

according to the parameters

$$a = 25214903917, \quad c = 11, \quad m = 2^{48},$$

and a seed chosen by the customer. Assuming that the value range of `unsigned int` is $\{0, \ldots, 2^{32} - 1\}$, we can't realize this generator using our class `random`. Doing all the computations over the type `double` and simulating the modulo operator in a suitable way is the way to go here. It is clear that we need a large modulus to obtain a useful generator, since m is an upper bound for the number of different numbers that we can possibly get from the generator. This means that `knuth8` from above is rather a toy generator.

The game of choosing numbers. Here is a game that you could play with your friend while waiting for a delayed train. Each of you independently writes down an integer between 1 and 6. Then the numbers are compared. If they are equal, the game is a draw. If the numbers differ by one, the player with the smaller number gets CHF 2 from the one with the larger number. If the two numbers differ by two or more, the player with the larger number gets CHF 1 from the one with the smaller number. You can repeat this until the train arrives (or until one of you runs out of cash, and hopefully it's your friend).

If you think about how to play this game, it's not obvious what to do. One thing *is* obvious, though: you should not write down the same number in every round, since then your friend quickly learns to exploit this by writing down a number that beats your number (by design of the game, this is always possible).

You should therefore add some unpredictability to your choices. You could, for example, secretly roll a dice in every round and write down the number that it shows. But Exercise 148 reveals that your friend can exploit this as well.

You must somehow finetune your random choices, but how? In order to experiment with different distributions, you decide to define and implement a class `loaded_dice` that rolls the dice in such a way that the probability for number i to come up is equal to a prespecified value p_i (a fair dice has $p_i = 1/6$ for all $i \in \{1, \ldots, 6\}$). Then you could let different loaded dices play against each other, and in this way discover suitable probabilities to use against your friend (who is by the way not studying computer science).

Program 3.6 shows a suitable class definition (that in turn relies on the class `random` from above, with the normalization to the interval $[0, 1)$). We will get to the class implementation (and the meaning of the data members) in Program 3.7 below.

```
link ../libraries/librandom/include/IFMP/loaded_dice.h
```

Program 3.6: *../progs/lecture/loaded_dice.h*

To initialize the loaded dice, we have to provide the probabilities $p_1, \ldots, p_5$ ($p_6 = 1 - \sum_{i=1}^{5} p_i$), and the random number generator that is being used to actually roll the dice. Again, we overload `operator()` to realize the functionality of rolling the dice once.

How do we implement this functionality? We partition the interval $[0, 1)$ into 6 right-open intervals, where interval i has length p_i:

Then we draw a number x at random from $[0, 1)$, using our generator. If the number that we get were truly random, then it would end up in interval i with probability exactly p_i. Under the assumption that our pseudorandom numbers behave like random numbers in a suitable way, we therefore declare i as the outcome of rolling the dice if and only if x ends up in interval i. This is the case if and only if

$$p_1 + \ldots + p_{i-1} \leq x < p_1 + \ldots + p_i.$$

This explains the data members `p_upto_1`,..., `p_upto_5` (we don't need `p_upto_0` ($= 0$) and `p_upto_6` ($= 1$)). The constructor in Program 3.7 simply sets these members from the data provided, and the implementation of `operator()` uses them in exactly the way that was envisioned by the previous equation.

```
// Prog: loaded_dice.cpp
// implement a class for rolling a loaded dice.

#include <IFMP/loaded_dice.h>

namespace ifmp {
  // class loaded_dice: implementation
  loaded_dice::loaded_dice
  (double p1, double p2, double p3, double p4, double p5,
   ifmpp::random& generator)
    : p_upto_1 (p1),
      p_upto_2 (p_upto_1 + p2),
      p_upto_3 (p_upto_2 + p3),
      p_upto_4 (p_upto_3 + p4),
      p_upto_5 (p_upto_4 + p5),
      g (generator)
  {}

  unsigned int loaded_dice::operator()()
  {
    const double x = g();
    if (x <= p_upto_1) return 1;
    if (x <= p_upto_2) return 2;
    if (x <= p_upto_3) return 3;
```

```
    if (x <= p_upto_4) return 4;
    if (x <= p_upto_5) return 5;
    return 6;
  }
} // end namespace ifmp
```

Program 3.7: *../progs/lecture/loaded_dice.cpp*

Now you can compare two different loaded dices to find out which one is better in the game of choosing numbers. Program 3.8 does this, assuming that you are using a loaded dice that prefers larger numbers, and your friend uses a loaded dice that stays more in the middle. It turns out that in this setting, you win in the long run, but not by much (CHF 0.12 on average per round). Exercise 149 challenges you to find the best loaded dice that you could possibly use in this game.

```
// Prog: choosing_numbers.cpp
// let your loaded dice play against your friend's dice
// in the game of choosing numbers.

#include <iostream>
#include <IFMP/loaded_dice.h>

// POST: return value is the payoff to you (possibly negative),
//       given the numbers of you and your friend
int your_payoff (const unsigned int you, const unsigned int your_friend)
{
  if (you == your_friend) return 0;        // draw
  if (you < your_friend) {
    if (you + 1 == your_friend) return 2; // you win 2
    return -1;                            // you lose 1
  } // now we have your_friend < you
  if (your_friend + 1 == you) return -2;  // you lose 2
  return 1;                               // you win 1
}

int main() {
  // the random number generator; let us use the generator
  // ANSIC instead of the toy generator knuth8; m = 2^31;
  ifmp::random ansic (1103515245u, 12345u, 2147483648u, 12345u);

  // your strategy may be to prefer larger numbers and use
  // the distribution (1/21, 2/21, 3/21, 4/21, 5/21, 6/21)
  const double p = 1.0/21.0;
  ifmp::loaded_dice you (p, 2*p, 3*p, 4*p, 5*p, ansic);

  // your friend's strategy may be to stay more in the middle
```

```
  // and use the distribution (1/12, 2/12, 3/12, 3/12, 2/12, 1/12)
  const double q = 1.0/12.0;
  ifmp::loaded_dice your_friend (q, 2*q, 3*q, 3*q, 2*q, ansic);

  // now simulate 1 million rounds (the train may be very late...)
  int your_total_payoff = 0;
  for (unsigned int round = 0; round < 1000000; round++) {
    your_total_payoff += your_payoff (you(), your_friend());
  }

  // output the result:
  std::cout << "Your total payoff is "
            << your_total_payoff << "\n";

  return 0;
}
```

Program 3.8: *../progs/lecture/choosing_numbers.cpp*

3.3.11 Details

Friend functions. Sometimes, we want to grant nonmember functions access to the internal representation of a class. Typical functions for which this makes sense are the in- and output operators operator<< and operator>>. Indeed, writing out or reading into the internal representation often requires some knowledge of this representation that goes beyond what other functions need.

We cannot reasonably write operator<< and operator>> as members (why not?), but we can make these functions *friends* of the class. As a friend, a function has unrestricted access to the private class members. It is clear that the *class* must declare a function to be its friend, and not the other way around, since it's the class that has to protect its privacy, and not the function. Formally, a *friend declaration* is a member declaration of the form

friend *function—declaration*;

This declaration makes the respective function a friend of the class and grants access to all data members, whether they are public or private. For the class rational, we could rewrite the private section as follows to declare in- and output operators to be friends of the class.

```
class rational {
private:
  friend std::ostream& operator<< (std::ostream& o, const rational& r);
  friend std::istream& operator>> (std::istream& i, rational& r);
```

```
  int n;
  int d; // INV: d!= 0
};
```

In the definition of these operators, we can then access the numerator and denominator through `.n` and `.d` as we used to do it in Section 3.2.5. If possible, friend declarations should be avoided, since they compromise encapsulation; but sometimes, they are useful in order to save unnecessary member functions.

3.3.12 Goals

Dispositional. At this point, you should ...

1) be able to explain the purpose of a class in C++;
2) understand the new syntactical and semantical terms associated with C++ classes, in particular *access specifiers*, *member functions*, and *constructors*;
3) understand the classes `ifmp::random` and `ifmp::loaded_dice` in detail.

Operational. In particular, you should be able to ...

(G1) find syntactical and semantical errors in a given class definition and implementation;

(G2) describe value range and functionality of a type given by a class definition;

(G3) add functionality to a given class through member functions;

(G4) write simple classes on your own;

(G5) work with and argue about pseudorandom numbers.

3.3.13 Exercises

Exercise 142 *Provide a full implementation of rational numbers as a class type, and test it. The type should offer all arithmetic operators (including in- and decrement, and the arithmetic assignments), relational operators, as well as in- and output and user-defined conversions (from* `int` *and to* `double`*). As an invariant, it should hold that the internal representation is normalized (see also Exercise 137). For all the functionality you provide, decide whether it should be realized by member functions, or by nonmember functions. The class should also have a nested numerator and denominator type to achieve more flexibility, and there should be a conversion function from values of this type. Test this class with internal integral type* `ifmp::integer`*!* (G3)(G4)

Exercise 143 *Split program Program 3.3 into a header file* `rational.h` *and a source file* `rational.cpp`*. Build object code file* `rational.o` *from* `rational.cpp` *and integrate it into the library* `libmath.a` *that you have created in Exercise 86. Modify the main program Program 3.2 to use this library.*

Exercise 144 *Rewrite the struct* `Tribool` *that you have developed in Exercise 128 into a class, by*

a) *making the data members private,*

b) *adding corresponding access functions,*

c) *adding an access function* `is_bool() const` *that returns true if and only if the value is not unknown, and*

d) *adding user-defined conversions from and to the type* `bool`*.*

(G4)

Exercise 145

a) *Find all errors and violations of const-correctness in the following program. Fix them and describe the functionality of the type* `Clock`*, by providing pre-and postconditions for the member functions.* (G1)(G2)

```
#include <iostream>

class Clock {
  Clock(unsigned int h, unsigned int m, unsigned int s);
  void tick();
  void time(unsigned int h, unsigned int m,
            unsigned int s);
private:
  unsigned int h_;
  unsigned int m_;
  unsigned int s_;
};

Clock::Clock(unsigned int& h,
             unsigned int& m,
             unsigned int& s)
    : h_(h), m_(m), s_(s)
{}

void Clock::tick()
{
  h_ += (m_ += (s_ += 1) / 60) / 60;
  h_ %= 24; m_ %= 60; s_ %= 60;
}

void Clock::time(unsigned int& h,
                 unsigned int& m,
                 unsigned int& s)
{
```

```
  h = h_;
  m = m_;
  s = s_;
}

int main() {
  Clock c1 (23, 59, 58);
  tick();

  unsigned int h;
  unsigned int m;
  unsigned int s;
  time(h, m, s);

  std::cout << h << ":" << m << ":" << s << "\n";

  return 0;
}
```

b) Implement an output operator for the class `Clock`. (G3)

Exercise 146 *Write a program* `random_triangle.cpp` *to simulate the following random process graphically. Consider a fixed triangle* t *and choose an arbitrary vertex of* t *as a starting point. In each step, choose as a next point the midpoint between the current point and a (uniformly) randomly selected vertex of* t.

The simulation at each step draws the current point into a Window. Use the window object `ifmp::wio` *defined in* `<IFM/window>` *for graphical output, and choose the triangle with vertices* $(0,0)$, $(512,0)$, *and* $(256,512)$. *Use the random number generator* `ansic` *from Program 3.8. At begin, the program should read in a seed for the random number generator and the number of simulation steps to perform. For testing purposes, let the simulation run for about 100,000 steps.* (G5)

Exercise 147 *Consider the generator* `ansic` *used in Program 3.8. Since the modulus is* $\mathrm{m} = 2^{31}$, *the internal computations of the generator will certainly overflow if 32 bits are used to represent* `unsigned int` *values. Despite this, the sequence of pseudorandom numbers computed by the generator is correct and coincides with its mathematical definition. Explain this!* (G5)

Exercise 148 *Find a loaded dice that beats the fair dice in the game of choosing numbers. (This is a theory exercise.)* (G5)

3.3.14 Challenges

Exercise 149 *What is the best loaded dice for playing the game of choosing numbers? Give its distribution! You could try to approximate the distribution experimentally, or somehow compute it. (***Hint***: in order to find a suitable theoretical model, search for the term "zero-sum games", or directly go to the corresponding chapter in* `http://www.inf.ethz.ch/personal/gaertner/cv/lecturenotes/ra.pdf`. *Once*

you have formulated the problem as a zero-sum game, you can solve it using for example the web-interface `http://banach.lse.ac.uk/form.html` (G5)

Exercise 150 *This is about* visual cryptography. *You may know the riddle of the rich Bedouin on his deathbed who wants to split his 17 camels among his three sons. Here we consider a modern version of this riddle where a* secret image *should be split among two people. You can for example imagine that the rich Bedouin wants to pass a treasure map to his two sons (these days, even Bedouins have less children than they used to have). Unfortunately, the two sons are at odds with each other, so that the Bedouin can't put the map into his testament. He is afraid that the son who finds the testament first keeps the treasure for himself.*

But the clever Bedouin finds a way of "splitting" the treasure map in such a way that the two sons must *cooperate in order to decipher the map and find the treasure. He calls his two sons and gives each of them one image, printed on a transparency. By themselves, both images look (and really are) absolutely random, and the sons are puzzled. But then the father tells the two that in order to see the treasure map, they simply have to overlay the two transparencies!*

Now how does this work? We assume for simplicity that the image to be split is black and white. The splitting happens pixel by pixel, according to the result (head or tail) of a fair coin flip. In each of the two transparencies, the original pixel is replaced by a "superpixel" of four pixels, according to the scheme in Figure 24.

Figure 24: *Visual cryptography: one pixel is split into two random "superpixels" of four pixels each, depending on the outcome of a fair coin flip*

In overlaying the two transparencies, we get a black superpixel for every black pixel, and a gray superpixel for every white pixel. Thus, the overlay reveals the secret image, except that it now has double size, and white appears as gray.

Each transparency by itself does not contain any information about the secret image. Independently of the color of the original pixel, every superpixel is with the same probability in one of the two different gray states.

Now the challenge: write a program `visual_crypto.cpp` *that splits an image in* `XBM` *format into two images (again in* `XBM` *format) according to the scheme just outlined. The* `XBM` *format is described in the text of Exercise 99 on page 183. You may use command line arguments (see the Details of Section 2.7 to specify the transparency to be generated.*

Test your program by printing the resulting images on transparencies, and then overlaying them. If you cannot print on transparencies (or if you feel that you cannot afford it anymore, after a few dozen failed attempts to fix your buggy program), write a program `xbm_merge.cpp` *that simulates the overlay of two transparencies in* `XBM` *format.*

3.4 Dynamic Types

It took us so long
and we worked so hard
We came so far just to compete
But don't forget all
the love and laughter
now the world is at our feet

Pet Shop Boys, Winner (2012)

This section explains how dynamic types can be realized in C++. Objects of fundamental types and of class types as developed so far occupy a fixed amount of memory known at compile time. In contrast, objects of dynamic types may vary in their memory consumption during a run of the program. We introduce the C++ concepts necessary to implement a dynamic type as a class. You will learn about explicit memory management through the `new` *and* `delete` *operators, and about the copy constructor, the assignment operator, and the destructor of a class type. As a running example, we discuss lists which are containers for sequences of elements that—unlike arrays—allow insertion and removal of elements "in the middle".*

Not every sequence of elements is properly modeled by an array. A sequence of customers standing in line at the supermarket checkout is a good example. Unlike an array, the sequence changes in length as customers at the front get served, and new customers line up at the end. Also, customers in the middle of the line may realize that they have to catch a train before they can expect to get served, so they either drop out of the line (appreciated by everyone), or ask to move ahead of somebody closer to the front (not appreciated by everyone).

Similar lines exist and need to be maintained in computing environments—just think of printer queues where documents line up to get printed. For storing the elements in such a line, an array is not a good choice. For example, how could we insert a new element 8 (an impatient customer) between 2 and 9 into this array

1	5	6	2	9	7	7	0

of 8 elements? We would first have to create space for a new element at the end, move

the elements $9, 7, 7, 0$ one index to the right, and then put 8 at the free index:

	1	5	6	2	9	7	7	0	
→	1	5	6	2	9	7	7	0	
→	1	5	6	2		9	7	7	0
→	1	5	6	2	8	9	7	7	0

With static arrays, we simply cannot create space for a new element at the end. With vectors, we can, although we have not shown it yet: vectors *are* dynamic. If v is a vector of length n, the expression

```
v.resize (n+1)
```

has the effect of making the vector one element longer. But after that, we still need to move potentially a lot of elements one index to the right before we can insert the new element. This is a lot of work and does not correspond to what happens at the supermarket checkout. People may complain about somebody entering the line in the middle, but hardly anyone physically moves when this happens. The "insertion" of an impatient customer into a checkout line is a very local operation that affects only the immediate neighborhood.

Removing an element from an array is equally cumbersome, unless it is the last one in which case we can get away with pretending that the array just got one element shorter. Suppose we want to remove element 2 from the array

1	5	6	2	8	9	7	7	0

of 9 elements. Taking out 2 creates a 'hole' that we need to fill by moving the elements $8, 9, 7, 7, 0$ one index to the left. Then we need to get rid of the last element of the array:

	1	5	6	2	8	9	7	7	0
→	1	5	6		8	9	7	7	0
→	1	5	6	8	9	7	7	0	
→	1	5	6	8	9	7	7	0	

With static arrays, the final step cannot physically be done, but we could simply ignore the last element from now on. Vectors, on the other hand, can explicitly be shrunk. If v has n elements, the expression

```
v.resize (n-1)
```

removes the last element from v. But before doing that, we needed to move a lot of elements to the left.

You might argue that it is not necessary to actually move elements. Instead, we could simply replace the removed element by a special value not occuring otherwise (-1, say):

1	5	6	-1	8	9	7	7	0

But this is not a real solution, since it messes up random access in the array. For example, 8 physically has index 4 still, but logically, it should have index 3 (skipping the -1 that stands for a removed element). So an "array with holes" can no longer be used as a real array.

3.4.1 Lists

Graphically, a data structure that naturally and efficiently supports the "checkout line" operations discussed above looks like in Figure 25.

Figure 25: *A list: a sequence of elements in which each element is pointing to the next one.*

Here, the elements of the sequence are not ordered by index (the irregular arrangement is meant to illustrate this); instead every element explicitly "knows" its successor element in the sequence (indicated by an arrow). The last element has a $\multimap$ arrow to indicate that it has no successor. Such a data structure is called a *list*, or *singly-linked list* in order to distinguish it from a *doubly-linked list* in which every element also knows its predecessor.

Insertion of an element into and removal of an element from a list is easy. The other elements "stay where they are", only one successor arrow needs to be changed. Following up on our above example, inserting 8 between 2 and 9 is done as follows: a new element 8 is created, with successor 9, and the successor of 2 is changed to 8. Figure 26 illustrates this change.

Removing 2 from this list of nine elements is equally simply. We delete the element 2 and change the successor of 6 from 2 to 8; see Figure 27.

Lists versus arrays. Let us summarize the main points: insertion and removal of elements are cheap local operations in a list, but expensive global operations in an array. On the

Figure 26: *Inserting a new element (8) between two elements (2 and 9) of a list.*

Figure 27: *Removing an element (2) from a list.*

other hand, a list does not allow random access anymore, since the elements are not consecutive in memory: if we want the i-th element in a list, we cannot "directly go there" through a simple address calculation, but we need to follow $i-1$ successor arrows until we arrive at the element.

We therefore cannot say that a list is better or worse than an array; both represent a sequence of values of the same type, and which representation is more suitable depends on what we want to do with the sequence.

3.4.2 Dynamic memory allocation

We now get to the C++ aspects of lists. In the above graphical description, we have talked about successor arrows, and if you go back to Figure 17 on Page 187, it is quite intuitive that such arrows can be realized with pointers. We have also talked about "creating" a new element (when inserting it into a list) and "deleting" an element (when removing it from a list). In order to do this in C++, we need the concept of *dynamic memory allocation*. Through such an allocation, we create an object with *dynamic* storage duration.

Most objects that we have seen so far were tied to variables, in which case memory gets assigned to them (and is freed again) at predetermined points during program execution (automatic and static storage duration, Section 2.4.3). Objects of dynamic storage duration are not tied to variables, and they may "start to live" (get memory assigned to them) and "die" (get their memory freed) at *any* point during program execution. The programmer can determine these points via `new` and `delete` expressions.

This is exactly what we need when we insert an element into a list, or remove an element from a list.

The program has some (typically quite large) region of the computer's main memory available to store dynamically allocated objects. This region is called the *heap*. It is initially unused, but when an object is dynamically allocated, it is being stored on the heap, so that the memory actually used by the program grows.

The new expression. For every type T, a `new` expression can come in one of the following two variants.

```
new T
new T(...)
```

In both cases, the expression returns an rvalue of pointer type T*. Its value is the address of an object of type T that has dynamically been allocated on the heap. The object itself is anonymous, but we usually store the resulting address under a variable name.

Variant 1 initializes the new object using T's default constructor (which leaves it uninitialized if T is a fundamental type), while variant 2 initializes the new object by calling the appropriate constructor with the arguments provided. For example, the following declarations initialize the variables `i` and `r`, of types `int*` and `rational*`, with the addresses of two new objects of types `int` and `rational`, respectively. In the `rational` case, this calls the constructor with two integers, see Page 285 in Section 3.3.4.

```
int*      iptr = new int;            // *iptr is undefined
rational* rptr = new rational(2,3);  // *rptr has value 2/3
```

We could also properly initialize `*i` using the "constructor syntax"

```
int*      iptr = new int(6);         // *iptr has value 6

std::cout << *iptr << "\n";          // 6
std::cout << *rptr << "\n";          // 2/3
```

The situation is visualized in Figure 28.

The delete expression. Dynamically allocated memory that is no longer needed should be freed. In C++, the programmer decides at which point this is the case. There are programming languages (Java, for example) that automatically detect and free unused memory on the heap. This automatic process is called *garbage collection*. It is generally more user-friendly than the manual deletion process in C++, but requires a more sophisticated implementation.

Dynamic storage duration implies that dynamically allocated objects live until the program terminates, unless they are explicitly freed. Dynamically allocated memory is more flexible than static memory, but in return it also involves some administrative effort. The `delete` expressions take care of freeing memory.

Figure 28: *Dynamically allocating an integer and a rational number on the heap. The new expression returns the address of the newly allocated object; this address can be stored in a pointer variable through which the object can be accessed from now on.*

`delete` *ptr*

Here, *ptr* may be a null pointer, in which case the `delete` expression has no effect. Otherwise, *ptr* must be a pointer to an object that has previously been dynamically allocated with a `new` expression. The effect is to make the corresponding memory available again for subsequent dynamic allocations on the heap.

For example, at a point in the program where the objects dynamically allocated through

```
int*      iptr = new int;             // *iptr is undefined
rational* rptr = new rational(2,3);   // *rptr has value 2/3
```

are no longer needed, we would write

```
delete rptr;
delete iptr;
```

The order of deletion does not matter here, but many programmers consider it logical to `delete` pointers in the inverse order of dynamic allocation: If you need to undo two steps, you first undo the second step.

There is one possible pitfall here: After the above deletion, the two objects pointed to are "gone" from the heap, but `iptr` and `rptr` still hold their expired addresses. Subsequently dereferencing the pointers by accident leads to undefined behavior. Thus, we better write

```
delete rptr; rptr = 0;
delete iptr; iptr = 0;
```

if we want to make sure that using any of these two pointers later on leads to a controlled runtime error. The situation after such a safe deletion is depicted in Figure 29.

Figure 29: *Safely deleting dynamically allocated objects: The two* `delete` *expressions involving* `iptr` *and* `rptr` *free the memory of the objects pointed to on the heap. Afterwards, the two pointers are set to* `0` *(visualized with* $\multimap$*) in order to prevent accidental use of the expired object addresses.*

Memory leaks. Although all memory allocated by a program is automatically freed when the program terminates normally, it is very bad practice to rely on this fact for freeing dynamically allocated memory. If a program does not explicitly free all dynamically allocated memory it is said to have a *memory leak*. Such leaks are often a sign of bad coding. They usually have no immediate consequences, but without freeing unused memory, a program running for a long time (think of operating system routines) may at some point simply exhaust the available heap memory.

Therefore, we have the following guideline.

Dynamic Memory Guideline: new and delete expressions always come in matching pairs.

3.4.3 List nodes

We now have all ingredients to get started with lists in C++. We represent the list elements as *nodes*, where a node stores

- the actual element which we call the *key* here (as in all the examples above, we use integer keys), and

- a pointer to the next node in the list. If the node is the last one in the list, this will be a null pointer.

Below is the definition of the class node. It has two data members, one for the key, and one for the pointer to the next node. Following good practice, we declare these data members private but offer public member functions get_key and get_node to access them in a read-only fashion.

When you go back to Figure 26 and Figure 27, you will remember that we frequently needed the operation of changing the successor of a node. For this, we provide a public member function set_next.

Finally, we need a constructor that is invoked when a new list node is created. As arguments, it has the new key and the successor pointer. The successor pointer has a default value of 0 (standing for no successor); see the Details of Section 2.8 on Page 205 for the explanation of default arguments.

```
1  // Prog: node.h
2  // definition of the class node
3
4  class node {
5   public:
6    // Constructor
7    node(int key, node* next = 0);
8
9    int get_key() const;
10   node* get_next() const;
11   void set_next (node* next);
12
13  private:
14   int key_;
15   node* next_;
16 };
```

Program 3.9: *../progs/lecture/node.h*

The implementation of this class (that we put into a separate cpp-file) is completely straightforward, so we just list it here without further comments. Note that—as already seen on Page 292 in the context of random numbers— we need to qualify (::) the member function names with the class name in such a separate file. Unlike for random numbers, we do not plan to make a library out of our lists, so we refrain from using namespace ifmp here and just concentrate on the plain code.

```
1  // Prog: node.cpp
2  // implementation of the class node
3
4  #include <iostream>
```

```
 5 #include <cassert>
 6 #include "node.h"
 7
 8 // Constructor
 9 node::node (int key, node* next)
10   : key_(key), next_(next)
11 {}
12
13 int node::get_key() const
14 {
15   return key_;
16 }
17
18 node* node::get_next() const
19 {
20   return next_;
21 }
22
23 void node::set_next(node* next)
24 {
25   next_ = next;
26 }
```

Program 3.10: *../progs/lecture/node.cpp*

3.4.4 Data members and default constructor

Having list nodes at our disposal, we will now represent a list as a sequence of list nodes, where each node holds a `key_` (the actual element) and a pointer `next_` to the successor node in the list. We thus mimick the graphcial representation of a list as in Figure 25; the sequence of elements is implicitly represented by the `next_` pointers of all list nodes, and the only explicit data member that we need in our class `list` is a pointer `head_` to the first list node. In Figure 25, this is the leftmost arrow coming out of the blue. In this way, we can also represent an empty list by a `head_` pointer of value 0.

In fact, default-initialization as it happens in a declaration statement of the form

```
list l;
```

should naturally create such an empty list. Here is a partial definition of the class `list` that works accordingly, along with an implementation of the default constructor.

```
// from list.h
class list{
 public:
  // default constructor:
  // POST: *this is an empty list
```

```
  list();
 ...
 private:
  node* head_;
...
};

// from list.cpp
list::list()
  : head_(0)
{}
```

3.4.5 Inserting an element at the front

If we want to manipulate a list, the easiest operation is the insertion of a new element at the beginning of the list. For this, we define and implement a member function push_front:

```
// POST: key was added before first element of *this
void push_front (int key);
```

Graphically, this looks as in Figure 30. To perform the insertion, we first create a

Figure 30: *Inserting a new element (8) at the beginning of the list* $9 \rightarrow 7 \rightarrow 7 \rightarrow 0$.

new list node with the provided key; since the successor of this new element is the former first element (pointed to by head_), the successor pointer gets initialized with head_. Afterwards, head_ is changed to point to the newly created element. Here is the implementation of push_front. Although the function body has just one line of code, we encourage the reader to go through this line in detail and match it with the above verbal description.

```
void list::push_front (int key)
{
  head_ = new node (key, head_);
}
```

3.4.6 Outputting a list

We have already seen in connection with our type `rational` in Section 3.2.5 that it is possible to overload the input and output operators for user-defined types. We will now provide `list` output. In customer code, the usage should be as follows.

```
list l1;
l1.push_front (1);
l1.push_front (3);
l1.push_front (2);
std::cout << l1 << "\n";  // 2 3 1
```

The implementation of operator<< is straightforward but nicely illustrates how lists can be traversed. We maintain a const pointer p to go through all list elements in turn. Initially, p is the lists's head pointer. Then, while p is not zero, we output the key of the element pointed to and advance p to point to the next element.

The -> operator. The operation of advancing a pointer in the way just described is very frequent in working with dynamic types; in our case, it can be done through the assignment

```
p = (*p).get_next()
```

where the brackets are necessary since the member access operator . has higher precedence than the dereferencing operator *. This looks somewhat ugly, though, and C++ provides an equivalent alternative, involving a new type of member access operator:

```
p = p->get_next()
```

The general syntax of this form of member access is

ptr—>*dname*
ptr—>*fname* (*expression1*, ..., *expressionN*)

where *ptr* is an expression of a pointer type whose underlying type has a data member *dname*, or a member function *fname*. The composite expressions are completely equivalent to

(∗*ptr*).*dname*
(∗*ptr*).*fname* (*expression1*, ..., *expressionN*)

Here is the implementation of our list's output operator, using the new form of member access just introduced. Note that as before, operator<< is implemented as a global (non-member) function.

```
// POST: *this is written to std::cout
std::ostream& operator<< (std::ostream& o, const list& l) {
  const node* p = l.get_head();
```

```
  while (p != 0) {
    o << p->get_key() << " ";
    p = p->get_next();
  }
  return o;
}
```

Figure 31 visualizes the operation of advancing a pointer.

Figure 31: *Setting p to p->get_next() advances the pointer p (left arrow) to point to the next element (right arrow).*

3.4.7 Appending an element at the end

While push_front was very easy to implement, the more natural operation may be to append a new element at the end of the list:

```
// POST: key was added after last element of *this
void push_back (int key);
```

This requires some work, though: since we do not explicitly store a pointer to the last element, we have to traverse the list in order to find the last element, using an advancing pointer p as before (Exercise 151 asks you to change the list type such that it also maintains a pointer to the last element).

If the list is not empty, the last element is the one for which p->get_next() == 0. Having found this element, we create a new last element with the provided key (a null pointer to the successor element is the default); then the formerly last element's successor pointer is changed to point to the new element.

If the list is empty, we proceed as for insertion at the front.

```
void list::push_back (int key)
{
  if (head_ == 0)
    head_ = new node (key);
  else {
    node* p = head_;
    while (p->get_next() != 0)
      p = p->get_next();
    p->set_next (new node (key));
  }
```

3.4.8 The copy constructor

The problem with member-wise initialization. Suppose that we enhance our previous customer code segment as follows.

```
1  list l1;
2  l1.push_front (1);
3  l1.push_front (3);
4  l1.push_front (2);
5  std::cout << l1 << "\n";  // 2 3 1
6
7  list l2 = l1;
8  std::cout << l2 << "\n";  // 2 3 1
9
10 l1.push_back(5);
11 l1.push_back(6);
12 std::cout << l1 << "\n"; // 2 3 1 5 6
13 std::cout << l2 << "\n"; // ?
```

What will be the output in Line 13? Let's analyze the code: In Line 7, we initialize list `l2` with `l1`, and at that point, `l1` represents the sequence $2, 3, 1$. Consequently, that's also what we see when we output `l2` in Line 8. Subsequently, we manipulate `l1` by appending two more elements, but we leave `l2` alone. So we expect to see

```
std::cout << l2 << "\n"; // 2 3 1
```

But instead, we will see this:

```
std::cout << l2 << "\n"; // 2 3 1 5 6
```

Somehow, the changes done to `l1` have also affected `l2`! What is going on?

Recall from Section 3.1.4 how initialization works for struct types: in

```
list l2 = l1
```

each data member of `l2` is initialized separately from the corresponding data member of `l1`. For our class `list`, there is only one data member, i.e. the `head_` pointer of `l2` is initialized with the `head_` pointer of `l1`. This means that afterwards, both heads point to the same object in memory, the first element of `l1`. Hence `l2` is not a copy of `l1` as intended but merely an alias—another name for the same list in memory. Figure 32 shows how this looks like. Now, since `l2` is just an alias of `l1`, everything that happens with `l1` will also happen with `l2`; this explains why outputting `l2` in Line 13 of the above code fragment actually outputs `l1`.

User-defined initialization. This situation is of course unacceptable. A customer of the type `list` does not know or even care how we implement lists; the cutomer simply expects that the declaration

```
list l2 = l1
```

Figure 32: *In the initialization* `list l2 = l1`*, the head pointer of* `l2` *is initialized with the head pointer of* `l1`*, so that afterwards,* `l2` *is not a copy of* `l1` *but an alias.*

produces a proper copy of `l1` in the sense that future changes to `l1` will not affect `l2`. In order to ensure this, C++ allows us to change the default behavior of initialization for a class type that we are writing. If memberwise initialization is not what we want, we can provide a *copy constructor* to override default initialization.

For a class T, the copy constructor is the unique constructor with signature

T (`const` T&)

In our case, it has the following declaration.

```
list (const list& l)
```

When pre- and postconditions are omitted, we mean that the copy constructor creates a new list as a proper copy of the argument list `l`. The copy constructor is automatically called whenever an object of type T is initialized from an object of (or convertible to) type T. If no copy constructor is declared, an *implicit copy constructor* is provided automatically, and this one performs the default member-wise initialization of data members.

A copy constructor for lists. Here is a copy constructor for the `list` class that properly copies its argument list. It first creates an empty list and then calls a private member function

```
// PRE: *this is empty
// POST: list starting at from was copied to *this
void copy (const node* from);
```

in order to do the actual work.

```
list::list (const list& l)
  : head_ (0)
{
  copy (l.head_);
}
```

The implementation of copy below appears to be somewhat lengthy, but it is again straightforward. We need to advance *two* pointers now, one pointer p for the list to be copied, and another pointer q for the list `*this` to be built up. In the beginning of each loop iteration, p points to the next element to be copied, while q points to the currently last element of the list being built up. Then, a copy of the element pointed to by p is appended as a new last element to `*this`, using the `set_next()` method for q.

```
void list::copy (const node* from)
{
  assert (head_ == 0);
  if (from != 0) {
    // copy first element to head
    head_ = new node (from->get_key());
    const node* p = from->get_next();
    node* q = head_;
    // copy remaining elements
    while (p != 0) {
      q->set_next (new node (p->get_key()));
      p = p->get_next();
      q = q->get_next();
    }
  }
}
```

If you see (or think that there should be) a simpler way of doing this, you might want to look at Exercise 152.

The copy constructor argument type. It is important that the formal argument of a copy constructor for the class T is of type `const` T`&` and not just T. The reason is not efficiency (our first motivation of const references in Section 3.2.6); the reason is that an argument of type T simply wouldn't work. The copy constructor is about *defining* what happens when values of type T are initialized, but in order to even pass a call argument of type T to the copy constructor, the compiler would already need to know how to do the initialization that we are about to define. In contrast, call by reference with argument type `const` T`&` implicitly passes only a pointer to the call argument.

3.4.9 The assignment operator

We already know that assignment and initialization are two different things. Initialization is applied to newly created objects, while an assignment is changing the value of an existing object (Section 3.1.4). Let us further extend our customer code above by adding the following lines.

```
1 list l3;
2 l3 = l1;
```

```
3 std::cout << l3 << "\n"; // 2 3 1 5 6
4
5 l1.push_back(7);
6 std::cout << l1 << "\n"; // 2 3 1 5 6 7
7 std::cout << l3 << "\n"; // ?
```

Here, l3 is default-initialized (to an empty list) in Line 1. Afterwards, the value of l1 is assigned to l3 which shows in the subsequent output. Again, we want to understand the output in the last Line 7 after manipulating l1 but leaving alone l3. You will not be surprised to see

```
std::cout << l3 << "\n"; // 2 3 1 5 6 7
```

even though l3 was not explicitly changed. The problem is as with initialization: also assignment is member-wise by default, meaning that l3 becomes an alias of l1 in Line 2, but not a proper copy.

User-defined assignment. To get user-defined assignment semantics for a class type T that we are implementing, we simply provide an overload of the assignment operator= with signature

T& operator= (const T&)

as a member function of the class T. If no such operator is present, an implicit assignment operator is generated, with the default member-wise asssignment semantics.

An assignment operator for lists. In case of our class list, the assignment operator becomes

```
list& operator= (const list& l);
```

where as before, absence of pre- and postconditions mean that we are aiming at proper copies in assignments. The implementation again uses the private member function copy, but more needs to be done here. In general, when we write

```
l2 = l1;
```

the target list l2 may be nonempty, in which case we first need to delete all its nodes that have dynamically been allocated at some earlier time with new. Not doing this creates memory leaks and might be fatal in the long run, see Section 3.4.2. Only after this deletion, we can copy the new content into l2. The implementation of operator= for our lists therefore uses a public member function clear that takes care of deletion. It is public since the customer may also use it to empty an existing list.

```
// POST: *this was emptied
void clear();
```

With this, here is our actual assignment operator.

```
list& list::operator= (const list& l)
{
  if (head_ != l.head_) { // avoid self-assignment
    clear();
    copy (l.head_);
  }
  return *this;
}
```

The check `head_l != l.head_` is necessary to catch *self-assignments* of the form

```
l1 = l1;
```

that should have no effect. Without the check, the content of `l1` would just vanish into thin air. It is customary that the assignment operator returns the left operand as an lvalue (Section 2.1.14), and this is what we are also doing here with return type `list&` and return expression `return *this`: since the infix notation

```
l1 = l2
```

already discussed in Section 3.1.5 is just a shorthand for the operator notation

```
l1.operator= (l2)
```

the left operand is in fact the implicit call argument `*this`. Finally, we provide the implementation of `clear`. With an advancing pointer p, we go through the list, "misusing" head_ to always point to the previous element that we then delete.

```
void list::clear ()
{
  node* p = head_;
  while (p != 0) {
    p = p->get_next();
    delete head_;
    head_ = p;
  }
  assert (head_ == 0);
}
```

3.4.10 The destructor

Consider the following example program where a list "runs out of scope" at the end of a block (which is artificial here, just to make the point).

```
1 // Prog: list_example.cpp
2 #include<iostream>
3 #include "list.h"
4
5 int main() {
```

```
  {
    list l;
    l.push_front (1);
    l.push_back (3);
    l.push_front (2);
    std::cout << l << "\n";  // 2 1 3
  } // l runs out of scope

  return 0;
}
```

Program 3.11: *../progs/lecture/list_example.cpp*

According to Section 2.4.3, the list `l` has automatic storage duration, so its memory will automatically be freed at the end of the inner block. You may have thought that a list has dynamic storage duration, but this is not true. Its *nodes* have dynamic storage duration, since they were allocated using `new` expressions. But `l` itself wasn't allocated with `new`, and since there is also no `static` keyword, automatic storage duration applies to the list `l`.

When an object of struct type runs out of scope, the default behavior is that the memory previously assigned to the inidividual data members is freed again. In our case, the memory cells storing the `head_` pointer of `l` are marked as free. However, the list's *nodes* are not automatically deleted. Each node has dynamic storage duration which means that its memory is freed only upon an explicit `delete`.

Hence, what we get is a memory leak: The list's head pointer is gone, and the list elements become useless junk on the heap.

User-defined deletion. Again, the solution is to change the default behavior at the end of automatic storage duration. If we need to do more than just member-wise deletion, we can define a *destructor* to customize the deletion process. A class T can have only one destructor that has signature

$\sim T()$

If no destructor is provided, it is generated implicity and performs member-wise deletion. The destructor is a member function with no arguments that is automatically called when the automatic storage duration of an object of type T ends, or when we explicitly `delete` the object.

A destructor for lists. The destructor is declared as follows and needs no pre- and post-conditions if it just `delete`'s the existing objects previously allocated with `new`.

```
~list();
```

The implementation is utterly simple: we just call the member function `clear` which explicitly `delete`'s all the list nodes.

```
list::~list()
{
  clear();
}
```

Going back to Program 3.11 above, this destructor is now automatically invoked for the list `l` at the point where `l` runs out of scope. Only after the destructor call is finished (with the deletion of all list nodes), the memory for the `head_` data member itself is being freed.

3.4.11 Definition of a dynamic type

We have introduced the class `list` as an instance of a dynamic type. In general, a *dynamic type* is a class type that performs dynamic memory allocation and has at least

- a default constructor,
- a copy constructor,
- an assignment operator, and
- a destructor

to properly deal with the dynamic memory that it manages.

A dynamic type completely hides the dynamic aspect from the customer. In using the type, the customer does not need to know that dynamic memory allocation is going on behind the scenes. The type can be used like any fundamental or user-defined static type (a type that does not perform dynamic memory allocation): For example, objects of dynamic types can be passed to functions, and under call by value semantics, the customer can be sure that the formal argument is initialized with the call argument by means of a proper copy (copy constructor). Also, the customer can rely on objects of dynamic types not leaking memory when their automatic storage duration ends (destructor).

Examples of dynamic types that we have seen are the vector types `std::vector<`*T*`>`.

3.4.12 More list functionality

One of our main motivations for lists was the possibility of efficiently removing an element "in the middle" of a list, see Figure 27. Indeed, if we have a pointer to the element to be removed *and* a pointer to the previous element, this is easy:

```
// PRE: p points to an element, and previous points to the
//      element before (= 0 if p points to the first element)
// POST: the element pointed to by p was removed from *this
void remove (node* previous, const node* p);
```

There are two cases: if previous is a null pointer, we need to change head_ to p->get_next(). Otherwise, previous_ needs to be changed to p->get_next(). In both cases, p gets deleted afterwards.

```
void list::remove (node* previous, const node* p)
{
  if (previous != 0)
    previous->set_next(p->get_next());
  else
    head_ = p->get_next();
  delete p;
}
```

Note that in a doubly-linked list, a single pointer to the element to be deleted would suffice (we leave the details to the reader, see Exercise 153). Now suppose that we want to remove the first occurrence of a specific key from the list:

```
// POST: the first occurrence of key was removed from *this
//       if *this does not contain key, nothing happened
void remove (int key);
```

This can be reduced to the removal by pointers: we advance two pointers previous and p until p->get_key() yields the specified key. Then we call remove with the current pointer values. Here is the straightforward implementation.

```
void list::remove (int key)
{
  node* previous = 0;
  node* p = head_;
  while (p != 0) {
    if (p->get_key() == key) {
      remove (previous, p);
      return;
    }
    previous = p;
    p = p->get_next();
  }
}
```

Many other list operations are desirable:

- searching for an element in the list;
- inserting a new element "in the middle";
- reversing the order of elements in the list;
- removing duplicate keys from a list;
- sorting the list;

- splitting a list in the middle, etc.;

Some of these operations are left to the reader as exercises.

3.4.13 Standard lists

If we need lists in practice, we should probably not use our own `list` class as implemented above. We have gone through this implementation mainly in order to explain the *concepts* behind dynamic types on a simple example. But our class has some obvious shortcomings: it is limited to key type `int`, and it is missing a lot of functionality still. A better alternative is to use `std::list<`T`>` from the standard library in order to represent lists of elements of type T. We will not elaborate on this type here but encourage the reader to look up its usage if needed.

3.4.14 Details

Dynamic arrays. It is also possible to get plain arrays of variable length through a `new` expression of this format:

`new` T[*expr*]

If *expr* has integer value $n \geq 0$, the effect is to dynamically allocate an array of length n with underlying type T on the heap. The return value is the address of the first element. As usual, the n array elements remain uninitialized if T is a fundamental type.

For example, in Program 2.24, we could instead of

```
std::vector<bool> crossed_out (n, false);
```

have used

```
bool* crossed_out = new bool[n];
for (int i=0; i<n; ++i) crossed_out[i] = false;
```

The expression

`delete`[] *ptr*

removes the array from the heap again. Here, *ptr* must be a pointer to the first element of an array that has previously been dynamically allocated with a `new` expression. The whole memory occupied by the array is freed on the heap. Under the aforementioned change to Program 2.24, we would therefore need to provide a matching

```
delete[] crossed_out;
```

at the end in order to avoid memory leaks.

If the plain `delete` is applied to a non-null pointer that does not point to a dynamically allocated single object, the behavior is undefined. The same is true if one tries to `delete[]` a single object. As always with pointers, the C++ language does not offer any means of detecting such errors.

3.4.15 Goals

Dispositional. At this point, you should ...

1) know the concept of a list as an abstract data structure;
2) know that pointers along with dynamic memory allocation can be used to maintain objects of dynamic type, such as lists;
3) know the concepts of user-defined initialization, user-defined asssignment and user-defined deletion, and understand why they are necessary for dynamic types;
4) know the definition of a dynamic type in general, and the implications for customer code using such a type.

Operational. In particular, you should be able to ...

(G1) enhance an existing implementation of a list type with additional functionality;

(G2) argue about the efficiency of list-manipulating algorithms;

3.4.16 Exercises

Exercise 151 *Implement a variant of the class* `list` *that in addition to* `head_` *(pointer to the first element) stores a pointer* `last_` *to the last element. In particular, change the function* `push_back` *to take advantage of this additional information.* (G1)

Exercise 152 *The following variant of the member function* `copy` *discussed in Section 3.4.8 is much shorter and more readable. Can you give a reason why we might still prefer the original variant?*

```
void list::copy (const node* from)
{
  assert (head_ == 0);
  const node* p = from;
  while (p != 0) {
    push_back (p->get_key());
    p = p->get_next();
  }
}
```

(G2)

Exercise 153 *Implement a variant of the class* `list` *in which every element has not only a pointer to the successor element, but also a pointer to the predecessor element. This is known as a* doubly-linked list. *Simplify in particular the* `remove` *member functions of the class.* (G1)

Exercise 154 *Implement some or all of extra list functionality mentioned in Section 3.4.12 for doubly-linked lists!* (G1)

Exercise 155 *Implement a comparison operator for the class* `list`*, in form of a global function* `operator==`*:*

```
// POST: returns true if and only if l1 and l2
//       store the same sequence of keys
bool operator== (const list& l1, const list& l2)
```

(G1)

Appendix A

C++ Operators

Description	Operator	Arity	Prec.	Assoc.
scope	`::`	2	18	right
subscript	`[]`	2	17	left
function call	`()`	2	17	left
construction	`type()`	1	17	right
member access	`.`	2	17	left
member access	`->`	2	17	left
post-increment	`++`	1	17	left
post-decrement	`--`	1	17	left
dynamic cast	`dynamic_cast<>`	1	17	right
static cast	`static_cast<>`	1	17	right
reinterpret cast	`reinterpret_cast<>`	1	17	right
const cast	`const_cast<>`	1	17	right
type identification	`typeid`	1	17	right
pre-increment	`++`	1	16	right
pre-decrement	`--`	1	16	right
dereference	`*`	1	16	right
address	`&`	1	16	right
bitwise complement	`~`	1	16	right
logical not	`!`	1	16	right
sign	`+`	1	16	right
sign	`-`	1	16	right
sizeof	`sizeof`	1	16	right
new	`new`	1	16	right
delete	`delete`	1	16	right
cast	`(type)`	1	16	right
member pointer	`->*`	2	15	left
member pointer	`.*`	2	15	left
	...			

	...			
multiplication	*	2	14	left
division (integer)	/	2	14	left
modulus	%	2	14	left
addition	+	2	13	left
subtraction	-	2	13	left
output/left shift	<<	2	12	left
input/right shift	>>	2	12	left
less	<	2	11	left
greater	>	2	11	left
less equal	<=	2	11	left
greater equal	>=	2	11	left
equality	==	2	10	left
inequality	!=	2	10	left
bitwise and	&	2	9	left
bitwise xor	^	2	8	left
bitwise or	|	2	7	left
logical and	&&	2	6	left
logical or	||	2	5	left
selection	?	3	4	right
assignment	=	2	3	right
mult assignment	*=	2	3	right
div assignment	/=	2	3	right
mod assignment	%=	2	3	right
add assignment	+=	2	3	right
sub assignment	-=	2	3	right
rshift assignment	>>=	2	3	right
lshift assignment	<<=	2	3	right
and assignment	&=	2	3	right
xor assignment	^=	2	3	right
or assignment	|=	2	3	right
exception	`throw`	1	2	right
sequencing	,	2	1	left

Table 9: *Precedences and associativities of C++ operators.*

Appendix B

Solutions

B.1 A first C++ program

Solution to Exercise 1. (d) and (f) are not identifiers, since they do not start with a letter. (g) is not an identifier, since it contains the character #. (b) is not allowed as a variable name, but it is a valid identifier.

Solution to Exercise 2. (c) is not an expression, since the first operand of the assignment operator must be an lvalue, but `1` is a literal, hence an rvalue. (f) is not an expression, since there is no closing parenthesis for the opening one. (h) is invalid, since `(a*3)` is an rvalue, but the left operand of the assignment operator must be an lvalue.

Solution to Exercise 3. (a), (e), and (g) are rvalues by definition of the binary multiplication operator. (b) and (d) are lvalues by definition of the assignment operator.

Solution to Exercise 4. (a) has value 6, obtained by multiplying the value of the primary expression `1` with the value of the composite expression `(2*3)`. The latter value is 6, for the same reason. (b) has value 5, obtained by assigning value 5 to `b` first (right assignment), and then to `a` (left assignment). (d) has value 1, by definition of the assignment operator. (e) has value 35, since the operands `(a=5)` and `(b=7)` have values 5 and 7, respectively.

In case of (g), the value is unspecified. If the right operand is evaluated first, we get value 25, but if the left operand comes first, `b` may have some value other than 5, and the left operand evaluates to this other value. The final result will not be 25, then.

Solution to Exercise 5. b) is incorrect, since `std::cin >> a` tries to change the value of the constant `a`. c) is incorrect, since there is an uninitialized constant `a`. e) is incorrect, since the assignment `a = 6` tries to change the value of the constant `a`. Same thing with f): although the value will not be changed, it is incorrect. The exact wording was that it is impossible to store anything under the address of a constant, even if it is the same

value as before. The correct programs are thus a), d), g). Among them, only d) violates the **Const Guideline**, since c's value never changes, but yet c is not defined as a constant.

Solution to Exercise 6. The least common multiple of $2, \ldots, n$ is the product of all *maximal* prime powers less or equal to n. For $n = 10$, for example, we get $2^3 \cdot 3^2 \cdot 5 \cdot 7 = 2520$. For $n = 20$, we get

$$2^4 \cdot 3^2 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 = 232792560,$$

and for $n = 30$

$$2^4 \cdot 3^3 \cdot 5^2 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 \cdot 29 = 2329089562800.$$

The following program computes these numbers. Interestingly, declaring `lcm` of type `int` leads to an incorrect value for $n = 30$ on a 32-bit system, due to overflow. But using `ifmp::integer`, our arbitrary-precision integers also mentioned in Challenge 11, yields the correct results.

```
// Prog: lcm.cpp (least common multiple)
// computes the least common multiple of 10, 20, and 30

#include<iostream>
#include<IFMP/integer.h>

int main()
{
  // the least common multiple of the number 2,...,n is the product
  // of all maximal prime powers that are less or equal to n.
  // Example: the least common multiple of 2,...,10 is 2^3 * 3^2 * 5 * 7

  std::cout << "Least common multiple of 2 up to...\n";
  // 10: 2^3 * 3^2 * 5 * 7
  ifmp::integer lcm = 2*2*2 * 3*3 * 5 * 7;
  std::cout << 10 << ": " << lcm << "\n";

  // 20: 2^4 * 3^2 * 5 * 7 * 11 * 13 * 17 * 19
  lcm = lcm * 2 * 11 * 13 * 17 * 19;
  std::cout << 20 << ": " << lcm << "\n";

  // 30: 2^4 * 3^3 * 5^2 * 7 * 11 * 13 * 17 * 19 * 23 * 29
  lcm = lcm * 3 * 5 *23 * 29;
  std::cout << 30 << ": " << lcm << "\n";

  return 0;

}
```

Solution to Exercise 7.

```
// Program: multhree.cpp
// Compute the product of three numbers.

#include <iostream>

int main()
```

```
{
  // input of a, b and c
  std::cout << "Compute a * b * c for a =? ";
  int a;
  std::cin >> a;

  std::cout << "... and b =? ";
  int b;
  std::cin >> b;

  std::cout << "... and c =? ";
  int c;
  std::cin >> c;

  // output a * b * c,
  std::cout << a << " * " << b  << " * " << c << " = "
            << a * b * c << ".\n";
  return 0;
}
```

Solution to Exercise 8.

```
// Program: power20.cpp
// Raise a number to the power twenty.

#include <iostream>

int main()
{
  // input
  std::cout << "Compute a^20 for a =? ";
  int a;
  std::cin >> a;

  // computation
  const int b = a * a; // b = a^2
  const int c = b * b; // c = a^4
  const int d = c * c; // d = a^8
  const int e = d * d; // e = a^16

  // output e * c, i.e. a^20
  std::cout << a << "^20 = " << e * c << ".\n";
  return 0;
}
```

Solution to Exercise 9. Here is the well-formatted program, complete with informative output and sensible comments. We have also fixed the two errors (`main()` instead of `main[]`, and `std::cin >> b` instead of `cin >> b`). This solves parts a), b), d), and e).

```
// SquareProduct.cpp
// Reads in two numbers a and b and outputs (a*b)^2,
// the square of their product
#include <iostream>

int main() {

    // input
    std::cout << "a =? ";
    int a;
```

```
    std::cin >> a;

    std::cout << "b =? ";
    int b;
    std::cin >> b;

    // computation and output
    const int c = a * b;
    std::cout << "(a*b)^2 = ";
    std::cout << c * c << ".\n";

    return 0;
}
```

For part c), here is the list of composite expressions from the original (fixed) program, along with their status.

- `std::cin >> a` (lvalue)
- `std::cin >> b` (lvalue)
- `a * b` (rvalue)
- `c = a * b` (lvalue)
- `c * c` (rvalue)
- `std::cout << c * c` (lvalue)

Solution to Exercise 10.

```
// Prog: age_verification.cpp
// outputs the age group that is not allowed to buy alcohol

#include<iostream>

int main()
{
  const int year = 2009;
  std::cout << "No alcohol to people born in the years "
            << year-17 << " - " << year << "!\n";
  std::cout << "For people born in " << year-18
            << ", check the id!\n";

  return 0;
}
```

Solution to Exercise 11. Here are the two programs:

```
// Program: power8_slow.cpp
// Raise a number to the eighth power,
// using integers of arbitrary size
// and with seven multiplications

#include <iostream>
#include <IFMP/integer.h>
```

```
 8
 9  int main()
10  {
11    // input (no prompt, as we intend to read from file)
12    ifmp::integer a;
13    std::cin >> a;
14
15    // computation
16    ifmp::integer b = a * a; // b = a^2
17    b = b * a;               // b = a^3
18    b = b * a;               // b = a^4
19    b = b * a;               // b = a^5
20    b = b * a;               // b = a^6
21    b = b * a;               // b = a^7
22    b = b * a;               // b = a^8
23
24    // no output, as we are interested in computation time
25    return 0;
26  }
```

```
 1  // Program: power8_fast.cpp
 2  // Raise a number to the eighth power,
 3  // using integers of arbitrary size
 4  // and with three multiplications
 5
 6  #include <iostream>
 7  #include <IFMP/integer.h>
 8
 9  int main()
10  {
11    // input (no prompt, as we intend to read from file)
12    ifmp::integer a;
13    std::cin >> a;
14
15    // computation
16    ifmp::integer b = a * a; // b = a^2
17    b = b * b;               // b = a^4
18    b = b * b;               // b = a^8
19
20    // no output, as we are interested in computation time
21    return 0;
22  }
```

Running them on inputs up to 100,000 decimal digits, you will see that `power8_fast.cpp` is indeed faster. But while the number of multiplications performed by `power8_fast.cpp` is only 42% of the corresponding number for `power8_slow.cpp` (3 vs. 7), this does not directly translate to the runtimes. Instead, you will observe that `power8_fast.cpp` needs around 75% of the time required by `power8_slow.cpp` (and this remains stable as the inputs get larger). Thus the speedup is much less than you might have expected from just counting multiplications. Why is this so?

In order to really answer this, you would have to know how the type `ifmp::integer` is implemented; but for our discussion, it is enough to know that multiplication of two `ifmp::integers` works according to the school method. Do you remember how this is done? When you multiply two numbers on a piece of paper, you multiply the first number with each individual digit of the second number and write down all the results (properly aligned). Then you just add them up (see Page 11 for an example of the school

method).

If the two numbers have m and n digits, respectively, you have n intermediate results, with m or $m+1$ digits each, meaning that in total, you will write down roughly mn digits. This is also what the computer does, and the time to do it will be roughly proportional to mn. Let us cheat a little and assume that the time is really mn (it could in reality be around $10mn$, or any other factor times mn, but the speedup of `power8_fast.cpp` compared to `power8_slow.cpp` does not depend on this).

With this knowledge at hand, we can already understand the 75% from above. Let's look at `power8_slow.cpp` first, and let us assume that the input number `a` has m digits. Then the first multiplication (`a * a`) needs time m^2 and results in the number `b` with (roughly) $2m$ digits. The next multiplication `b * a` features numbers with $2m$ and m digits and therefore takes time $2m^2$. The result has (roughly) $3m$ digits. Then we multiply numbers with $3m$ and m digits, in time $3m^2$, and the result has (roughly) $4m$ digits. Continuing in this way, we see that the time to perform all multiplications is (roughly) $(1+2+\cdots+7)m^2 = 28m^2$.

Now comes `power8_fast.cpp`. The first multiplication again takes time m^2, but since the second one features two numbers with (roughly) $2m$ digits, the time for the second multiplication is $4m^2$ and results in a number with (roughly) $4m$ digits. In the third multiplication, we therefore multiply two $4m$-digit numbers, in time $16m^2$. In total, this takes time $(1+4+16)m^2 = 21m^2$. And since 21 is 75% of 28, this calculation is a pretty good explanation of the experimental observations.

Solution to Exercise 12. For the lower bound in a), we argue by induction that $a_i \leq a^{2^i}$ for all i (we can't do more than double the number in each step). In order to get $a_t = n$, we therefore must have

$$a^{2^t} \geq a_t = a^n,$$

or $2^t \geq n$. It follows that

$$t \geq \lg n \geq \lfloor \lg n \rfloor = \lambda(n).$$

For the upper bound, we have to come up with a computation for a^n that needs at most $\lambda(n) + \nu(n) - 1$ steps. This is simple (and called the binary method). By doubling a $\lambda(n)$ times, we can compute in $\lambda(n)$ steps all powers of the form a^{2^i} that are less or equal to a^n. For example, in $\lambda(20) = 4$ steps, we can get the values a, a^2, a^4, a^8, a^{16}. Since n is the sum of exactly $\nu(n)$ of these 2^i, a^n is the *product* of exactly $\nu(n)$ of these a^{2^i} (this is a simple consequence of the formula $a^{n+m} = a^n \cdot a^m$). It follows that we can obtain a^n by simply multiplying these $\nu(n)$ values together, and since we already have them, this can be done in $\nu(n) - 1$ further multiplications.

For b), we give an example where the upper bound is not tight. Consider $n = 15$ (1111 in binary). We have $\lambda(15) = 3$ and $\nu(15) = 4$, so the binary method would need 6 multiplications. But we can do it with five multiplications, as follows:

```
a1 = a0 * a0   // a^2
a2 = a1 * a0   // a^3
a3 = a2 * a2   // a^6
a4 = a3 * a3   // a^12
a5 = a4 * a2   // a^15
```

In general, no exact formula for $\ell(n)$ is known.

B.2 Integers

Solution to Exercise 13. Types of intermediate results are either `int` or `unsigned int` (indicated with `u`).

a) `-2-4*3` is parenthesized as `(-2)-(4*3)` since unary minus binds more strongly than binary minus, and since multiplicative operators bind more strongly than additive ones. Evaluation: `(-2)-(4*3)` $\rightarrow$ `(-2)-12` $\rightarrow$ `-14`.

b) `10%6*8%3` is parenthesized as `((10%6)*8)%3` since binary (multiplicative) arithmetic operators are left-associative. Evaluation: `((10%6)*8)%3` $\rightarrow$ `(4*8)%3` $\rightarrow$ `(32)%3` $\rightarrow$ `2`.

c) `6-3+4*5` is parenthesized as `(6-3)+(4*5)` since multiplicative operators bind more strongly than additive ones, and since (additive) arithmetic operators are left-associative. Evaluation: `(6-3)+(4*5)` $\rightarrow$ `3+(4*5)` $\rightarrow$ `3+20` $\rightarrow$ `23`.

d) `5u+5*3u` is parenthesized as `5u+(5*3u)` since multiplicative operators bind more strongly than additive ones. Evaluation: `5u+(5*3u)` $\rightarrow$ `5u+15u` $\rightarrow$ `20u`. Here, the `int` value 5 is implicitly converted to `unsigned int`.

e) `31/4/2` is parenthesized as `(31/4)/2` since (multiplicative) arithmetic operators are left-associative. Evaluation: `(31/4)/2` $\rightarrow$ `7/2` $\rightarrow$ `3`.

f) `-1-1u+1-(-1)` is parenthesized as `(((-1)-1u)+1)-(-1)` since (additive) arithmetic operators are left-associative. Evaluation: `(((-1)-1u)+1)-(-1)` $\rightarrow$ `(((`2^b-1`)u-1u)+1)-(-1)` $\rightarrow$ `((`2^b-2`)u+1)-(-1)` $\rightarrow$ `(`2^b-1`)u-(-1)` $\rightarrow$ `(`2^b-1`)u-(`2^b-1`)u` $\rightarrow$ `0u`, due to the conversion rules.

Solution to Exercise 14. *b)* and *g)* are illegal because `-a` and `7+a`, respectively, are rvalues and the left operand of an assignment must be an lvalue. For the others, the parentheses are as follows.

a)	`c=((a+7)+(--b))`			*c)*	`c=(a=(-b))`
d)	`a-((a/b)*b)`	*e)*	`b*=((++a)+b)`	*f)*	`a-(a*(+(-b)))`
		h)	`(a+(3*(--b)))+(a++)`	*i)*	`(b++)+(--a)`

Solution to Exercise 15. *h)* induces unspecified behavior because it is not specified whether the increment of `a` happens before or after the first `a` is evaluated. The remaining expressions can be evaluated as follows.

a) `c=((a+7)+(--b))` $\rightarrow$ `c=((5+7)+1)` $\rightarrow$ `c=(12+1)` $\rightarrow$ `c=13` $\rightarrow$ 13.

c) `c=(a=(-b))` $\rightarrow$ `c=(a=(-2))` $\rightarrow$ `c=(a=`-2`)` $\rightarrow$ `c=`-2 $\rightarrow$ -2.

d) `a-((a/b)*b)` $\rightarrow$ `5-((5/2)*2)` $\rightarrow$ `5-(2*2)` $\rightarrow$ `5-(4)` $\rightarrow$ 1.

e) `b*=((++a)+b)` $\rightarrow$ `b*=(6+2)` $\rightarrow$ `b*=8` $\rightarrow$ 16.

f) `a-(a*(+(-b)))` $\rightarrow$ `5-(5*(+(-2)))` $\rightarrow$ `5-(5*(+−2))` $\rightarrow$ `5-(5*−2)` $\rightarrow$ `5-−10` $\rightarrow$ 15.

i) `(b++)+(--a)` $\rightarrow$ `2+4` $\rightarrow$ 6.

Solution to Exercise 16. For any positive d, we can uniquely write n in the form

$$n = xd + y,$$

where $x, y \in \mathbb{N}$ and $y < d$. In fact, these define div and mod via $x = n \operatorname{div} d$ and $y = n \bmod d$.

For $n = a, d = bc$, we get

$$a = pbc + q,$$

where $p, q \in \mathbb{N}$ and $q < bc$. In particular, $p = a \operatorname{div}(bc)$. It remains to prove that

$$(a \operatorname{div} b) \operatorname{div} c = p.$$

To this end, we use $n = q, d = b$ to get

$$q = rb + s,$$

where $r, s \in \mathbb{N}, s < b$ and $r < c$ (since $rb \leq q < cb$). Then we have

$$a = (pc + r)b + s, s < b,$$

and this means that

$$a \operatorname{div} b = pc + r, r < c.$$

This in turn means that $p = (a \operatorname{div} b) \operatorname{div} c$.

The implication holds as long as the mathematical value of the expression `b*c` is representable as an `unsigned int` value. This holds since the other operations involving `/` are always error-free. If there is an overflow in `b*c`, the expressions `a/b/c` and `a/(b*c)` might yield different values.

Solution to Exercise 17. *a)* 1111 *b)* 10101100 *c)* 101001001 *d)* 1111111110

Solution to Exercise 18. *a)* 55 *b)* 65 *c)* 233 *d)* 341

Solution to Exercise 19. We first need the fact that a natural number $n \geq 1$ has $\lfloor \log_{10} n \rfloor + 1$ decimal digits ($\lfloor x \rfloor$ is x rounded down to the next smaller integer). This is not hard to see: if n has decimal expansion

$$n = \sum_{i=0}^{\infty} b_i 10^i,$$

the largest i with nonzero b_i occurs for the unique value i_0 such that $10^{i_0} \leq n < 10^{i_0+1}$. By taking logarithms, this is equivalent to $i_0 \leq \log_{10} n < i_0 + 1$, meaning that $i_0 = \lfloor \log_{10} n \rfloor$. Since we now have

$$n = \sum_{i=0}^{i_0} b_i 10^i,$$

with $b_{i_0} \neq 0$, n has exactly $\lfloor \log_{10} n \rfloor + 1$ decimal digits.

Now for the actual question: first, it is not hard to see that the two numbers $2^{43,112,609} - 1$ and $2^{43,112,609}$ have the same number of decimal digits, since a power of two always has a last decimal digit $2, 4, 8$, or 6.

The number of digits is therefore

$$\begin{aligned} \lfloor \log_{10} 2^{43,112,609} \rfloor + 1 &= \lfloor 43,112,609 \cdot \log_{10} 2 \rfloor + 1 \\ &= \lfloor 43,112,609 \cdot 0.3010299957 \rfloor + 1 \\ &= \lfloor 12,978,188.5 \rfloor + 1 \\ &= 12,978,189. \end{aligned}$$

We have to be a bit careful here: if we compute $\log_{10} 2$ with insufficient precision, we may easily get a wrong answer. In fact, how can we be sure that the above answer is correct? Assuming that 0.3010299957 is correct up to possibly a difference of one in the last digit, we have an additive error of at most 10^{-10} in this approximation of $\log_{10} 2$. Even if we multiply this approximation with the large number $43,112,609$, the additive error is therefore at most around 0.0043 which is not enough to change the value of $\lfloor 12,978,188.5 \rfloor$.

These observations imply that we need to have $\log_{10} 2$ rounded to eight digits in order to guarantee the correct value. Most pocket calculators will be able to compute $\log_{10} 2$ up to that precision.

Solution to Exercise 20.

a) 0110 *b)* 1100 *c)* 1000 *d)* not representable *e)* 1101

Solution to Exercise 21.

```
// Prog: average_speed.cpp
// computes the average speed over a roundtrip A->B->A, given
// the average speeds for A->B and B->A
//
// Here is the "theory": let's say that the distance A->B is
// d km, and let x and y be the two average speeds in km/h.
// Then the trip A->B takes d/x hours, while B->A needs d/y
// hours. This is a total of d/x+d/y hours for a trip of 2d km.
// Thus, the average speed is
//      2d/(d/x+d/y) = 2(1/x+1/y) = 2/((x+y)/xy) = 2xy/(x+y).
// We simply compute this over the type unsigned int, with the
// desired effect that the mathematically correct result gets
// rounded down.

```

```
#include<iostream>

int main()
{
  // inputs
  std::cout << "Average speed in km/h from A to B =? ";
  unsigned int x;
  std::cin >> x;
  std::cout << "Average speed in km/h from B to A =? ";
  unsigned int y;
  std::cin >> y;

  // computation and output
  std::cout << "Average speed in km/h from A to B and back is "
            << 2*x*y/(x+y) << "\n";

  return 0;
}
```

Solution to Exercise 22. We start by solving the Celsius-to-Fahrenheit equation for Degrees Celsius and get

$$\text{Degrees Celsius} = \frac{5(\text{Degrees Fahrenheit} - 32)}{9}.$$

The critical operation is the multiplication by 5. If $\underline{m}, \overline{m}$ are the smallest and largest representable `int` values, we must guarantee that

$$\underline{m} \leq 5(\text{Degrees Fahrenheit} - 32) \leq \overline{m}$$

in order to avoid over- and underflows. Solving this for Degrees Fahrenheit yields the following bounds.

$$\frac{\underline{m}}{5} + 32 \leq \text{Degrees Fahrenheit} \leq \frac{\overline{m}}{5} + 32.$$

Since integer division rounds down for positive numerators, a valid upper bound is obtained from the expression

```
std::numeric_limits<int>::max() / 5 + 32
```

For the lower bound, we have assumed for this exercise that the integer division rounds towards zero (up) for negative numerators. Therefore, the value

```
std::numeric_limits<int>::min() / 5 + 32
```

is a valid lower bound. The resulting code is given below (the output of the mixed number simply employs the div and mod operators).

```
// Program: celsius.cpp
// Convert temperatures from Fahrenheit to Celsius.

#include<iostream>
#include<limits>

```

```
int main ()
{
  // Input
  std::cout << "Temperature in degrees Fahrenheit =? \n"
            << " (from the interval ["
            << std::numeric_limits<int>::min() / 5 + 32
            << ", "
            << std::numeric_limits<int>::max() / 5 + 32
            << "])\n";
  int fahrenheit;
  std::cin >> fahrenheit;

  // Computation
  const int ncelsius = 5 * (fahrenheit - 32); // numerator

  // Output as mixed number
  std::cout << fahrenheit << " degrees Fahrenheit are "
            << ncelsius / 9 << " " << ncelsius % 9
            << "/9 degrees Celsius.\n";

  return 0;
}
```

Solution to Exercise 23. According to Section 2.2.8, the least significant binary digit b_0 of n is $n \bmod 2$, and the remaining digits are obtained by applying the same technique to $(n - b_0)/2 = n \operatorname{div} 2$. It follows that the second to last digit is $(n \operatorname{div} 2) \bmod 2$, and that the third to last one is $((n \operatorname{div} 2) \operatorname{div} 2) \bmod 2$, or (as one can prove; nice Exercise for next time) $(n \operatorname{div} 4) \bmod 2$.

```
// Program: threebin.cpp
// Output the last three binary digits of a number n.

#include<iostream>

int main ()
{
  // input
  std::cout << "Last three binary digits of n =? ";
  unsigned int n;
  std::cin >> n;

  // computation and output
  std::cout << "The digits are "
            << n / 2 / 2 % 2 << n / 2 % 2 << n % 2 << ".\n";

  return 0;
}
```

Solution to Exercise 24.

```
// Prog: vat.cpp
// given a net amount, the program computes the total amount,
// including VAT

#include<iostream>

int main()
{
```

```
    const int vat_rate = 76;        // tenth of a percent

    // input
    std::cout << "Net amount =?  "; // CHF
    unsigned int net;
    std::cin >> net;

    // computation (in CHF/1000)
    const unsigned int vat = vat_rate * net;
    const unsigned int tot = 1000 * net + vat;

    // convert to CHF/Rp.
    // digits before the decimal point
    const unsigned int vat_chf = vat / 1000;
    const unsigned int tot_chf = tot / 1000;
    // first digit after the decimal point
    const unsigned int vat_rp1 = vat % 1000 / 100;
    const unsigned int tot_rp1 = tot % 1000 / 100;
    // second digit after the decimal point
    const unsigned int vat_rp2 = vat % 100 / 10;
    const unsigned int tot_rp2 = tot % 100 / 10;


    // output
    std::cout << "VAT          = "
              << vat_chf << "," << vat_rp1 << vat_rp2 << "\n";
    std::cout << "Total amount = "
              << tot_chf << "," << tot_rp1 << tot_rp2 << "\n";

    return 0;
}
```

Solution to Exercise 25. For part a), we first have to find a theoretical way of computing the survivor $p(k)$. One way of doing it is to generalize the problem first. Let $p_n(k)$ be the last survivor in a circle of n people when every k-th person is killed. Then $p(k) = p_{41}(k)$. Here is the crucial insight that leads to a recursive formula for $p_n(k)$:

$$p_n(k) = (p_{n-1}(k) + k) \bmod n, \quad n \geq 2,$$

where $p_1(k) = 0$ (in a circle of 1, person 0 is the only person and therefore the last survivor). To see that this formula holds, we consider the n-circle after the first person (numbered $k-1$) has been killed. We now have an $(n-1)$-circle, and if we renumber the positions in such a way that the position formerly being k is now position 0, $p_{n-1}(k)$ tells us the last survivor in this new numbering. To get the survivor $p_n(k)$ in the old numbering, we simply have to add k, *but modulo* n. This takes a little thought but is easily deduced from the following table.

old numbering	0	1	$\cdots$	$k-2$	$k-1$	k	$\cdots$	$n-1$
new numbering	$n-k$	$n-k+1$	$\cdots$	$n-2$	$-$	0	$\cdots$	$n-k-1$

Here is the program that determines $p(k)$ by iteratively computing the sequence $p_1(k), p_2(k), ..., p_{41}(k)$, according to the formula above.

```
// Program: josephus.cpp
// determines the survivor number in the Josephus Problem

#include<iostream>

int main ()
{
  // input
  std::cout << "Every k-th person is killed for k =? ";
  unsigned int k;
  std::cin >> k;

  // computation of p_41(k)
  unsigned int n = 1; // circle size, runs through 1,...,41
  unsigned int p = 0; // survivor number, runs through p_1(k),...,p_{41}(k)

  // each of the following 40 identical statements increments n and updates
  // p according to the formula p_n(k) = (p_{n-1}(k) + k) mod n
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;
  p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n; p = (p + k) % ++n;

  // now p = p_{41}(k) = p(k); output
  std::cout << "The survivor is the person number " << p << ".\n";

  return 0;
}
```

Part b) is solved by playing with the program a little, and the answer is no. There are 41 possible initial positions for Josephus, and 41 ways of choosing k. Both for $k = 3$ and for $k = 7$, the survivor number is 30, so in the sequence $p(1), \ldots, p(41)$, at least one of the 41 number $\{0, \ldots, 40\}$ appears twice. But this implies that there must also be at least one other number in $\{0, \ldots, 40\}$ that does not appear at all. The initial position with this number is therefore fatal for all $k \in \{1, \ldots, 41\}$.

Interestingly, if we allow k to be *any* natural number, then it is possible to survive from any initial position.

Solution to Exercise 26.

```
// Prog: pythagoras.cpp
// helper program for computing all positive integers a,b,c
// such that
//     a^2 + b^2 = c^2 and a + b + c = 1000
#include<iostream>

// Here is our 'Ansatz':
// Substituting c = 1000 - (a+b) into a^2 + b^2 = c^2 implies
//
//                0 = 1000000 - 2000(a+b) + 2ab,
//    <=>    ab/1000 = a+b - 500 (*)
//
// This means that ab must be divisible by 1000 = 2^3*5^3. Therefore,
```

```
// one of a and b must be divisible by 5^2 = 25. Because the situation
// is completely symmetric in a and b, we will assume that it's a. In
// other words, we must only check the 39 values a=25,50,75,...,975.
//
// Given a, (*) can be used to compute b = 1000 - 500000/(1000-a).
// Since b must be integer, 500000/(1000-a) must be integer as well.
//
// The following outputs the four values
//   r = 500000 mod (1000-a),
//   a,
//   b = 1000 - 500000/(1000-a),
//   c = 1000 - a - b
// for all 39 values of a that need to be considered. (a,b,c) is
// a solution if and only if r = 0 and b,c > 0.

int main()
{
  int a = 25;
  int r;
  int b;
  int c;
  // now come 39 copies of the same two lines
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
```

```
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;
  r = 500000 % (1000-a); b = 1000 - 500000 / (1000 - a); c = 1000 - a - b;
  std::cout << r << ", " << a << ", " << b << ", " << c << "\n"; a +=25;

  // The unique solution is {a, b} = {375, 200}, c = 425
  return 0;
}
```

B.3 Booleans

Solution to Exercise 27. A Boolean function has 2^n possible n-tuples of arguments, and the function is specified by assigning the value 0 or 1 to each of them. It follows that there are 2^{2^n} possibilities, and this is also the number of Boolean functions.

Solution to Exercise 28. This can be done by going through the truth tables.

a) XOR is associative.

x	y	z	$(x \oplus y)$	$\oplus$	z			x	$\oplus$	$(y \oplus z)$		
0	0	0	0	$\oplus$	0	$=$	0	0	$\oplus$	0	$=$	0
0	0	1	0	$\oplus$	1	$=$	1	0	$\oplus$	1	$=$	1
0	1	0	1	$\oplus$	0	$=$	1	0	$\oplus$	1	$=$	1
0	1	1	1	$\oplus$	1	$=$	0	0	$\oplus$	0	$=$	0
1	0	0	1	$\oplus$	0	$=$	1	1	$\oplus$	0	$=$	1
1	0	1	1	$\oplus$	1	$=$	0	1	$\oplus$	1	$=$	0
1	1	0	0	$\oplus$	0	$=$	0	1	$\oplus$	1	$=$	0
1	1	1	0	$\oplus$	1	$=$	1	1	$\oplus$	0	$=$	1

b) (AND, OR) is distributive.

x	y	z	$(x \wedge y)$	$\vee$	z			$(x \vee z)$	$\wedge$	$(y \vee z)$		
0	0	0	0	$\vee$	0	$=$	0	0	$\wedge$	0	$=$	0
0	0	1	0	$\vee$	1	$=$	1	1	$\wedge$	1	$=$	1
0	1	0	0	$\vee$	0	$=$	0	0	$\wedge$	1	$=$	0
0	1	1	0	$\vee$	1	$=$	1	1	$\wedge$	1	$=$	1
1	0	0	0	$\vee$	0	$=$	0	1	$\wedge$	0	$=$	0
1	0	1	0	$\vee$	1	$=$	1	1	$\wedge$	1	$=$	1
1	1	0	1	$\vee$	0	$=$	1	1	$\wedge$	1	$=$	1
1	1	1	1	$\vee$	1	$=$	1	1	$\wedge$	1	$=$	1

c) (OR, AND) is distributive.

x	y	z	$(x \vee y)$	$\wedge$	z			$(x \wedge z)$	$\vee$	$(y \wedge z)$		
0	0	0	0	$\wedge$	0	$=$	0	0	$\vee$	0	$=$	0
0	0	1	0	$\wedge$	1	$=$	0	0	$\vee$	0	$=$	0
0	1	0	1	$\wedge$	0	$=$	0	0	$\vee$	0	$=$	0
0	1	1	1	$\wedge$	1	$=$	1	0	$\vee$	1	$=$	1
1	0	0	1	$\wedge$	0	$=$	0	0	$\vee$	0	$=$	0
1	0	1	1	$\wedge$	1	$=$	1	1	$\vee$	0	$=$	1
1	1	0	1	$\wedge$	0	$=$	0	0	$\vee$	0	$=$	0
1	1	1	1	$\wedge$	1	$=$	1	1	$\vee$	1	$=$	1

d) NAND is not associative, since for example $(0 \uparrow 0) \uparrow 1 = 1 \uparrow 1 = 0$, but $0 \uparrow (0 \uparrow 1) = 0 \uparrow 1 = 1$.

Solution to Exercise 29. $x_1 \oplus x_2 \oplus \cdots \oplus x_n$ is true if and only if an *odd* number of the x_i are 1.

Solution to Exercise 30.

a) We only need to show that $\{\text{AND}, \text{NOT}\}$ generates OR, and then use the completeness of $\{\text{AND}, \text{OR}, \text{NOT}\}$. This is simply DeMorgan's Law: we have

$$\text{OR}(x, y) = \text{NOT}(\text{AND}(\text{NOT}(x), \text{NOT}(y))).$$

b) This is similar; we only need to express AND using DeMorgan's Law:

$$\text{AND}(x, y) = \text{NOT}(\text{OR}(\text{NOT}(x), \text{NOT}(y))).$$

c) If we can generate NOT from NAND, we are done, since NOT and NAND can obviously generate AND, and then we use a). Indeed,

$$\text{NOT}(x) = \text{NAND}(x, x).$$

d) As in c), we are done once we can generate NOT from NOR, since we can then get OR and use b). Indeed,

$$\text{NOT}(x) = \text{NOR}(x, x).$$

e) Once we are able to generate NOT, we are done, since we can use a). But this is easy:

$$\text{NOT}(x) = \text{XOR}(x, 1).$$

Solution to Exercise 31. Values that work are -1 for `a`, 0 for `b` and 1 for `c`. Since `a < b < c` is parenthesized as `(a < b) < c` and `a< b` evaluates to *true*, the latter value is implicitly converted to the integer value 1. The comparison `1 < 1` then yields value *false*. The expression `a < b && b < c`, on the other hand, yields value *true*.

Solution to Exercise 32.

a) `(x != (3 < 2)) || (y && ((-3) <= (4 - (2 * 3))))`

b) `((z > 1) && (((! x) != (2 - 2)) == 1)) && y`

c) `((3 * z) > z) || (((1 / x) != 0) && ((3 + 4) >= 7))`

Solution to Exercise 33.

```
a) (x != (3 < 2)) || (y && ((-3) <= (4 - (2 * 3)))) →
 (0 != (3 < 2)) || (y && ((-3) <= (4 - (2 * 3)))) →
 (0 != false) || (y && ((-3) <= (4 - (2 * 3)))) →
 (0 != 0) || (y && ((-3) <= (4 - (2 * 3)))) →
 false || (y && ((-3) <= (4 - (2 * 3)))) →
 false || (true && ((-3) <= (4 - (2 * 3)))) →
 false || (true && (−3 <= (4 - (2 * 3)))) →
 false || (true && (−3 <= (4 - 6))) →
 false || (true && (−3 <= −2)) →
 false || (true && true) →
 false || true →
 true

b) ((z > 1) && (((! x) != (2 - 2)) == 1)) && y →
 ((2 > 1) && (((! x) != (2 - 2)) == 1)) && y →
 (true && (((! x) != (2 - 2)) == 1)) && y →
 (true && (((! 0) != (2 - 2)) == 1)) && y →
 (true && ((1 != (2 - 2)) == 1)) && y →
 (true && ((1 != 0) == 1)) && y →
 (true && (true == 1)) && y →
 (true && (1 == 1)) && y →
 (true && true) && y →
 (true && true) && 1 →
 (true && true) && true →
 true && true →
 true

c) ((3 * z) > z) || (((1 / x) != 0) && ((3 + 4) >= 7)) →
 ((3 * 2) > 2) || (((1 / x) != 0) && ((3 + 4) >= 7)) →
 (6 > 2) || (((1 / x) != 0) && ((3 + 4) >= 7)) →
 true || (((1 / x) != 0) && ((3 + 4) >= 7)) →
 true
```

Solution to Exercise 34. The output from line 6 is 1 (*true*) if and only if the input number n that is read in line 5 is less than 3. The expression `!b` is by de Morgan equivalent to

```
a * 3 <= a + 4 || a >= 5
```

which in turn is equivalent to `a <= 2 || a >= 5`. Observe that the variable `a` is incremented by one in line 6; therefore the expression `!b` in line 8 is *true*, unless $n \in \{2, 3\}$. Furthermore, for $n = 3$ the expression `++a > 4` in line 8 evaluates to 5>4, that is, *true*. Thus, the output from line 8 is 1 (*true*) if and only if $n \neq 2$.

Solution to Exercise 35. Here are the logical parentheses:

```
b /= (2 + (b / 2))
```

```
(a < 1) || ((b != 0) && (((2 * a) / (a - 1)) > 2))
```

Note that $2 + \mathtt{b} \operatorname{div} 2 > \mathtt{b}/2$, so b will be replaced by 1 at most. In fact, it is almost always replaced by 1, except for a <= 2. The first output is therefore 0 if a <= 2 and 1, otherwise.

The second expression is definitely *true* for a == 0. For a >= 1, the value of the expression is determined by

```
(b != 0) && (((2 * a) / (a - 1)) > 2)
```

For a <= 2, we have b == 0, meaning that the value is *false*, and short circuit evaluation ensures that we don't divide by a non-positive number in this case.

For a > 2, we have b == 1, and the value of the expression is determined by

```
((2 * a) / (a - 1)) > 2
```

This yields *true* exactly for a <= 3. To summarize, the second output is 1 if and only if a has value 0 or 3.

Solution to Exercise 36. We can model the sequence s for example by a variable of type unsigned int. The binary representation of its value then yields the sequence of bits. To append a bit in the end, we simply multiply by two and add the bit. To remove bits in the end, we (repeatedly) divide by two. The last bit can at any stage be obtained through the modulus operator. To realize the replacement of bits in s, we use these mechanisms, along with the implicit conversions from bool to unsigned int and vice versa.

The empty sequence can be represented by value 0, but if you think about it, this value also represents *any* sequence of zeros. This means that the following programs can't issue an error message if the sequence of operands is too short for the current operator: logically, the sequence is filled up with zeros until "$-\infty$".

We could alternatively say that we only use unsigned int values whose binary representation starts with a 1 (and then value 1 would correspond to the empty sequence). However, at this stage we have no way of testing in a C++ program whether a number equals 1, so we can't issue an error message anyway.

Here are the six required programs.

```
// Program: and.cpp
// Replace last two bits in a sequence of bits by their AND

#include<iostream>

int main()
{
  // input sequence: the binary representation of s
  unsigned int s;
  std::cin >> s;

  const bool x = s / 2 % 2; // second-to-last bit
  const bool y = s % 2; // last bit
```

```
  std::cout << 2 * (s / 4)  + (x && y); //  x, y -> AND(x,y)
  return 0;
}
```

```
// Program: or.cpp
// Replace last two bits in a sequence of bits by their OR

#include<iostream>

int main()
{
  // input sequence: the binary representation of s
  unsigned int s;
  std::cin >> s;

  const bool x = s / 2 % 2; // second-to-last bit
  const bool y = s % 2; // last bit
  std::cout << 2 * (s / 4)  + (x || y); //  x, y -> OR(x,y)
  return 0;
}
```

```
// Program: not.cpp
// Flip last bit in a sequence of bits

#include<iostream>

int main()
{
  // input sequence: the binary representation of s
  unsigned int s;
  std::cin >> s;

  const bool x = s % 2; // last bit
  std::cout << 2 * (s / 2) + !x; // x -> NOT(x)
  return 0;
}
```

```
// Program: zero.cpp
// Append 0 to a sequence of bits

#include<iostream>

int main()
{
  // input sequence: the binary representation of s
  unsigned int s;
  std::cin >> s;

  std::cout << 2 * s; // append 0
  return 0;
}
```

```
// Program: one.cpp
// Append 1 to a sequence of bits

#include<iostream>

int main()
```

```
{
  // input sequence: the binary representation of s
  unsigned int s;
  std::cin >> s;

  std::cout << 2 * s + 1; // append 1
  return 0;
}
```

```
// Program: eval.cpp
// Output an empty sequence of bits

#include<iostream>

int main()
{
  // behaves like an infinite sequence of 0's
  std::cout << 0;
  return 0;
}
```

B.4 Control statements

Solution to Exercise 37. This program contains four errors.

- line 1: `iostraem` ⟶ `iostream`
- line 4: `unsinged` ⟶ `unsigned`
- line 6: semicolon missing after *condition* of `for`-statement
- line 6: superfluous opening curly brace `{`

Things that are ok are in particular the following.

- The `+1` in line 4 is a valid expression of type `int` (and value 1), constructed from the unary addition operator and the literal operand `1`. In the assignment, this value is converted to the type `unsigned int`.
- `{ std::cin >> x; }` The variable `x` here is the one defined in line 4, even though we have put the input statement into an extra scope.
- The *init-statement* of the `for`-loop is ok, since we may assign an `unsigned int` value to an `int` variable.

The fixed program looks like this.

```
1  #include<iostream>
2
3  int main()
4  {
5    unsigned int x = +1;
6    { std::cin >> x; }
7    for (int y=0u; y < x;)
8      std::cout << ++y;
9    return 0;
10 }
```

It outputs the sequence of positive natural numbers smaller or equal to the input number (without any blanks, this will be difficult to read, though).

Solution to Exercise 38. There are two problems in the code. First, the declarative region of the declaration of `i` in line 3 is the body of the loop only. Therefore, the expression `++i < 10` in line 5 is undefined. Moreover, even if the declaration would extend up to line 5, the loop would be infinite because `i` is initialized with 1 in every iteration. Thus, the condition `++i < 10` will never be fulfilled. This problem can be fixed by moving the declaration of `i` out of the loop.

Second, as the variable `s` is initialized with `0`, none of the multiplications in the loop changes the value of `s`. This problem can be fixed by initializing `s` with `1`. The corrected program then looks as shown below.

```
unsigned int s = 1;
int i = 1;
do {
  if (i % 2 == 1) s *= i;
} while (++i < 10);
```

This computes the product of the odd numbers $1, 3, \ldots, 9$.

Solution to Exercise 39.

declaration in line	declarative region	scope
4	3–20	4–11, 15–20
6	5–17	6–17
10	8–14	10–14
12	8–14	12–14
15	5–17	15–17
18	3–20	18–20

The output of the program is 4. There is only one iteration of the loop, and that one sets the s relevant in line 16 to 2.

Solution to Exercise 40.

- For input -1: no iteration, output is 0.
- For input 1:

begin of iteration	s	x	i
1	0	1	0
2	0	1	1

  No second iteration, output is 0.

- For input 2:

begin of iteration	s	x	i
1	0	2	0
2	0	2	1
3	1	2	2

  No third iteration, output is 1.

- For input 3:

begin of iteration	s	x	i
1	0	3	0
2	0	3	1
3	1	3	2
4	3	4	3
5	6	7	4

We see that `x` grows faster than `i`, so the condition `i < x` will mathematically never be satisfied. Due to overflow, it might eventually be satisfied in practice, but the output is undefined.

Solution to Exercise 41. Here are the problems.

- line 5: `x` is undeclared, because the `x` in line 4 is local to the block in line 4.
- line 5: it should be `std::cin >> x` and not `std::cin << x`.
- line 7: this is an infinite loop: since `y` is of type `unsigned int`, its value will always be nonnegative. The condition `y >= 0` is therefore always satisfied, regardless of the way `y` is changed in the loop.
- line 9: `s` appears outside of its scope (lines 7-8), so it is undefined here.

Solution to Exercise 42. Let x and y be the values of `x` and `y`. For $x = 0$, we get an infinite loop, since `y` is initialized with 1 but never changes. For $x \geq 1$, `y` is initialized with $x + 1$ and decreased by x in the subsequent iterations. Unless $x = 1$, y therefore becomes "negative" in the second iteration, leading to a not well-defined positive value of `y`. For $x = 1$, the loop terminates after two iterations and outputs value $2 + 1 = 3$.

To summarize, the behavior is well-defined exactly for the following input/output pair.

input	output
1	3

Solution to Exercise 43. Let's analyze what the code does. In the do-statement, `k` is incremented (starting from 0), until it reaches the value n of `n`. After this, `e` is set to *false*, and the loop terminates. Therefore, `k` increases n times, meaning that `x` (initially of value 1) is multiplied by 2^n.

The output is therefore 2^n, for input n. Here is a more readable version of this code.

```
1   unsigned int n;
2   std::cin >> n;
3
4   unsigned int x = 1;
5   for (int k = 0; k < n; ++k)
```

```
    x *= 2;

  std::cout << x;
```

Solution to Exercise 44. Let's first see what the inner `for`-loop does. For given value i of `i`, `j` starts from 1 and gets incremented by 1 during the `if`-statement's condition. The inner loop terminates with `break` if the value *before* the increment reaches i. This means, `s += j - 1` is executed for `j` running from 2 through i. The inner loop therefore increments `s` by

$$\sum_{k=1}^{i-1} k = \frac{i(i-1)}{2}, \quad i \geq 2.$$

The other do-loop increments i (starting from -10) by 1, as long as the incremented i is smaller or equal to x. The inner loop is therefore executed for the values $i = -10, -9, \ldots, x$, and we get

$$s = \sum_{i=2}^{x} \frac{i(i-1)}{2}.$$

A more readable version is therefore

```
#include <iostream>
int main()
{
  int x;
  std::cin >> x;
  int s = 0;
  for (int i = 2; i <= x; ++i)
    s += i * (i-1) / 2;
  std::cout << s << "\n";
  return 0;
}
```

Solution to Exercise 45. For $n = 0$, the condition `n % d != 0` is never satisfied (0 is divisible by any positive integer). It follows that the `for` statement performs no iteration, meaning that `d` has value 2 after the `for` statement. The output is therefore (incorrectly)

```
0 is prime.
```

For $n = 1$, the program enters an infinite loop: Since 1 is *not* divisible by any number bigger than 1, the condition `n % d != 0` will never be satisfied, and the `for` statement will not terminate.

The most transparent solution is to handle these two special cases separately, as follows:

```
// Program: prime_all.cpp
// Test if a given natural number is prime.

#include <iostream>

int main ()
{
  // Input
  unsigned int n;
  std::cout << "Test if n>=0 is prime for n =? ";
  std::cin >> n;

  // deal with 0,1
  if (n < 2) {
    std::cout << n << " is not prime.\n";
    return 0;
  }

  // Computation: test possible divisors d
  unsigned int d;
  for (d = 2; n % d != 0; ++d);

  // Output
  if (d < n)
    // d is a divisor of n in {2,...,n-1}
    std::cout << n << " = " << d << " * " << n / d << ".\n";
  else
    // no proper divisor found
    std::cout << n << " is prime.\n";

  return 0;
}
```

Solution to Exercise 46.

```
// Prog: fak-1.cpp
// compute the n!

#include <iostream>

int main ()
{
  std::cout << "Factorial of n =? ";
  int n;
  std::cin >> n;

  // computation
  unsigned int fak = 1;
  for (unsigned int i = 2; i <= n; ++i) fak *= i;

  // output
  std::cout << "Factorial of " << n << " is " << fak << ".\n";

  return 0;
}
```

Solution to Exercise 47.

```
// Program: dec2bin.cpp
```

```
// Output reverse binary representation of a number n

#include <iostream>

int main()
{
  // input
  std::cout << "Number n =? ";
  unsigned int n;
  std::cin >> n;

  // computation and output
  std::cout << "Reverse binary representation is: ";
  do { // a while loop cannot correctly handle n == 0
    std::cout << n % 2;
    n /= 2;
  } while (n > 0);
  std::cout << "\n";

  return 0;
}
```

Solution to Exercise 48.

```
// Prog: cross_sum.cpp
// compute the cross sum of a natural number

#include <iostream>

int main ()
{
  std::cout << "Cross sum of n =? ";
  unsigned int n;
  std::cin >> n;

  // computation
  unsigned int cross = 0;
  for (unsigned int m = n; m > 0; m /= 10) cross += m % 10;

  // output
  std::cout << "Cross sum of " << n << " is "
            << cross << ".\n";

  return 0;
}
```

Solution to Exercise 49.

```
// Program: perfect.cpp
// Find all perfect numbers up to an input number n

#include <iostream>

int main()
{
  // input
  std::cout << "Find perfect numbers up to n =? ";
  unsigned int n;
  std::cin >> n;

```

```
  // computation and output
  std::cout << "The following numbers are perfect.\n";
  for (unsigned int i = 1; i <= n ; ++i) {
    // check whether i is perfect
    unsigned int sum = 0;
    for (unsigned int d = 1; d < i; ++d)
      if (i % d == 0) sum += d;
    if (sum == i)
      std::cout << i << " ";
  }
  std::cout << "\n";

  return 0;
}
```

The program output looks as follows.

```
Find perfect numbers from 1 to n =? 50000
The following numbers are perfect.
6 28 496 8128
```

Solution to Exercise 50. We start as in Exercise 47, but instead of outputting the digits (which would then appear in reverse order), we use them to build up a number m whose binary representation is the *reverse* of the binary representation of n. Then we proceed again as in Exercise 47, but this time starting with m. We have to be careful not to lose leading zeros in the binary representation of m, since we want to output them as well. Therefore we also need to count the number of binary digits in n.

```
// Program: dec2bin2.cpp
// Output binary representation of a number n

#include <iostream>

int main()
{
  // input
  std::cout << "Number n =? ";
  unsigned int n;
  std::cin >> n;

  // compute / count digits and build up number m with
  // reverse representation
  unsigned int digits = 0;
  unsigned int m = 0;
  do { // a while loop cannot correctly handle n == 0
    ++digits;
    m = m * 2 + n % 2;
    n /= 2;
  } while (n > 0);

  // now process m as in dec2bin.cpp
  std::cout << "Binary representation is: ";
  for (int d = 0; d < digits; ++d) {
    std::cout << m % 2;
    m /= 2;
  }
  std::cout << "\n";

  return 0;
```

```
}
```

Some people suggested that the problem can be solved with just one loop, as follows:

```
unsigned int s = 0;
for ( unsigned int k=1; n>0; n/=2 ){
  s += k * (n % 2);
  k *= 10;
}
std::cout << s;
```

This performs the following "hack": it builds up a *decimal* number s whose digits are just 0's and 1's. These digits are in fact exactly the digits of n's binary representation: the digits of n are computed in reverse order and are successively *prepended* to s. This means that in s, they appear in the right order.

The hack is that we can save the output loop as in dec2bin2.cpp above, by misusing standard integer output to do it for us. This is a very nice solution, but it has one drawback: the number s is substantially larger than n: where n needs k binary digits, s needs k *decimal* digits. This implies that (on a 32-bit system) only numbers n up to roughly $2^{10} = 1024$ can be processed, since beyond that, we get an overflow in s.

Solution to Exercise 51.

```
// Prog: dice.cpp
// Pete rolls three four-sided dice, and Colin rolls 2 six-sided dice.
// What is the probability that Pete / Colin / nobody has the higher
// total number of points?
#include<iostream>
#include<cassert>

int main()
{
  const unsigned int pete_throws =  4*4*4; // number of possibilities for Pete
  const unsigned int colin_throws = 6*6;  // number of possibilities for Colin
  const unsigned int pairs_of_throws = pete_throws * colin_throws;

  unsigned int pete_wins = 0; // pairs of throws won by Pete
  unsigned int colin_wins = 0; // pairs of throws won by Colin
  unsigned int draws = 0; // pairs of throws that are a draw

  // now do the following: for each possible pair of throws
  // (all of them have the same probability of occurring),
  // compute the scores of both players and determine the
  // winner (or that the throw is a draw)
  for (unsigned int pete=0; pete < pete_throws; ++pete) {
    // compute score of Pete, interpreting pete as a throw, i.e.
    // a base-4 number with three digits
    // (digit value i encodes i+1 points, i=0,1,2,3)
    unsigned int pete_score = 0;
    unsigned int p = pete;
    for (int i=0; i<3; ++i) {
      pete_score += p % 4 + 1;
      p /= 4;
    }
    for (unsigned int colin=0; colin < colin_throws; ++colin) {
      // compute score of Colin, interpreting colin as a throw, i.e.
```

```
      // base-6 number with two digits
      // (digit value i encodes i+1 points, i=0,1,2,3,4,5)
      unsigned int colin_score = 0;
      unsigned int c = colin;
      for (int i=0; i<2; ++i) {
        colin_score += c % 6 + 1;
        c /= 6;
      }
      // who wins in this pair of throws?
      if (pete_score >  colin_score) ++pete_wins;
      if (pete_score <  colin_score) ++colin_wins;
      if (pete_score == colin_score) ++draws;
    }
  }
  // compute the probabilities
  std::cout << "Pete wins with probability   "
            << pete_wins << "/" << pairs_of_throws << ".\n";
  std::cout << "Colin wins with probability  "
            << colin_wins << "/" << pairs_of_throws << ".\n";
  std::cout << "Draw occurs with probability "
            << draws << "/" << pairs_of_throws << ".\n";

  return 0;
}
```

The output is

```
Pete wins with probability   1152/2304.
Colin wins with probability  870/2304.
Draw occurs with probability 282/2304.
```

After cancellation, these numbers are $1/2$, $145/384$, and $47/384$. The game is therefore not fair, since Pete is much better off.

Solution to Exercise 52.

```
// Program: cole.cpp
// computes a nontrivial factor of 2^67-1

#include <iostream>
#include <IFMP/integer.h>

int main()
{
  // build up 2^67-1
  ifmp::integer p = 1;
  for (int i=0; i<67; ++i) p *= 2;
  --p;

  std::cout << "Factoring 2^67-1 = " << p << "...\n";

  // now try all potential divisors (odd numbers)
  for (ifmp::integer d = 3; d < p; d += 2)
    if (p % d == 0) {
      std::cout << p << " = " << d << " * " << p / d << ".\n";
      break;
    }

  return 0;
}
```

Solution to Exercise 53.

```
// Prog: ulam_spiral.cpp
// draws the Ulam spiral

#include<iostream>
#include<IFMP/window>

int main()
{
  // read in spiral length
  std::cout << "Draw Ulam spiral of length n =? (3 <= n <= 250000)\n";
  unsigned int n;
  std::cin >> n;

  // open display window of appropriate size
  ifmp::Wstream w (500, 500, "The Ulam spiral"); // fits all spirals

  int x = 250; int y = 250; // start in the middle

  // handle 2, 3 (they are prime)
  w << ifmp::Point (x, y);
  w << ifmp::Point (x, ++y);

  // drawing direction, initially left
  int dx = -1;
  int dy =  0;

  // how many pixels into this direction?
  int sidelength = 2; // for 4, 5

  // main drawing loop (completes the spiral arm containing n)
  for (unsigned int k = 4; k <= n; ) {
    for (int i = 0; i < sidelength; ++i) {
      x += dx; y += dy;
      // is k prime ?
      unsigned int d;
      for (d = 2; k % d != 0; ++d);
      if (d == k)
        w << ifmp::Point (x, y); // draw pixel
      ++k;
    }
    // change direction
    const int h = dx; dx = -dy; dy = h;
    // sidelength grows before going right/left
    if (dy == 0) ++sidelength;
  }

  // wait for right mouse click
  w.wait_for_mouse_click(3);

  return 0;
}
```

Solution to Exercise 54. Here is the idea: we build up the placements columnwise, where we represent a partial placement in the first $column$ columns by a $column$-digit number $placement$ in base n; digit i encodes the position of the queen in column i (from 0 (bottommost) to $n-1$ (topmost)).

In fact, we represent $placement$ by a value of type `unsigned int`; the digits in base n can be accessed through repeated modulus and integral division operations with sec-

ond operand n. The algorithm iteratively enumerates all threat-free partial placements. Whenever the partial placement is full, we increase a $\mathsf{solution}$ counter.

In the beginning of each iteration, we are given a partial placement $\mathsf{placement}$ up to column column, along with the information whether it is threat-free (in the beginning ($\mathsf{column} = 0$), the partial placement is empty ($\mathsf{placement} = 0$) and obviously threat-free). Moreover, we will maintain the invariant that $\mathsf{placement}$ restricted to columns $\{1, \ldots, c\}, c < \mathsf{column}$ is threat-free.

If the whole $\mathsf{placement}$ is threat-free, we extend it to the next column, where we start with the queen in the bottommost position. The invariant still holds in this case. The extension is done through a multiplication of $\mathsf{placement}$ by n. In the case where we have a threat in $\mathsf{placement}$, we switch to the next possible partial placement within our current set of columns. This can simply be done by adding one to $\mathsf{placement}$; the least significant base-n digits that have become 0 now correspond to columns where all possible queen positions 0 through $n - 1$ have been tried already. Therefore we remove these columns from our partial placement and continue with a shorter partial placement up to the last column where not everything has been tried yet. Since this new $\mathsf{placement}$ agrees with the old one except in its last column, we know that our invariant also holds in this case.

After updating $\mathsf{placement}$ that way, we now have a new partial placement for which we have to check whether it's threat-free. By the invariant we know that the partial placement excluding the last column is threat-free, so we only have to check for threats with the queen in the last column of our partial placement. This gives us the information whether the current $\mathsf{placement}$ is threat-free, and we start all over again with the next iteration. Here is the program.

```
1  // Program: n_queens.cpp
2  // compute number of solutions to the n-queens puzzle
3  // for given input number n (for 32-bit systems, this
4  // works up to n = 9 only); using the nonstandard g++
5  // type "unsigned long long" instead of "unsigned int"
6  // everywhere, the program works for larger values (but
7  // becomes very slow around n=14)
8
9  #include<iostream>
10
11 int main ()
12 {
13   // input
14   std::cout << "Solve n-queens puzzle for n =? ";
15   unsigned int n;
16   std::cin >> n;
17
18   unsigned int solutions = 0; // number of solutions
19   unsigned int placement = 0; // partial placement (base n)...
20   unsigned int column = 0;    // ...up to this column
21   bool threat_free = true;    // is the placement threat-free?
22
23   for(;;) {
24     if (threat_free && column < n) {
25       // extend placement to next column
26       placement *= n;
27       ++column;
```

```
    } else {
      if (threat_free) ++solutions; // n == column
      // try next placement up to this column
      placement += 1;
      while (placement % n == 0) {
        // column exhausted, remove from placement
        placement /= n;
        --column;
      }
      if (column == 0) break; // everything tried, stop
    }
    // check whether placement is threat-free
    threat_free = false;                        // we don't know yet
    const unsigned int q_column = placement % n; // queen in column
    unsigned int c = column-1;                  // column to the left
    unsigned int p_c = placement / n;           // placement up to c
    while (c > 0) {
      unsigned int q_c = p_c % n;          // queen in c
      // is q_column threatened by q_c?
      if (q_c == q_column) break;          // horizontal threat
      if (q_c - q_column == column - c) break; // diagonal threat 1
      if (q_column - q_c == column - c) break; // diagonal threat 2
      --c; p_c /= n; // go one column left
    }
    if (c == 0) threat_free = true;
  }

  std::cout << solutions << " solutions\n";

  return 0;
}
```

Solution to Exercise 55.

```
// Prog: famous_last_digits.cpp
// outputs the last 10 decimal digits of Mersenne prime 2^{43,112,609}-1

#include <iostream>
#include <IFMP/integer.h>

int main()
{
  // we compute 2^43112609-1 modulo 10^10 = 10,000,000,000;
  // this gives us the desired result. Since for all positive
  // integers a, b, m
  //    (a+b) mod m = (a mod m + b mod m) mod m , and
  //    (a*b) mod m = (a mod m * b mod m) mod m,
  // it follows that we can also compute all intermediate
  // results modulo 10^10, without making any mistake.
  // We need long integers, though, since we don't have a
  // sufficient number of digits otherwise

  ifmp::integer m = 100000; // 10^5;
  m = m * m;                // 10^10;

  // to compute 2^{43,112,609} quickly, we use "repeated squaring":
  // This works as follows: to compute n^p, we write p as 2q + r,
  // where r is 0 or 1. Then n^p = (n^2)^q * n^r.
  unsigned int p = 43112609;
  ifmp::integer n = 2;
  ifmp::integer result = 1;
  while (p > 0) {
    unsigned int r = p % 2;
```

```
    if (r > 0)
      result = result * n % m;  // *= n^r
    p = p / 2;                  // q
    n = n * n % m;              // n^2
  }

  // now subtract 1 and output
  std::cout << "Last 10 decimal digits: " << result - 1 << "\n";

  return 0;
}
```

The output is 6697152511.

B.5 Floating point numbers

Solution to Exercise 56.

	type	value	lavlue/rvalue
a)	`bool`	`true`	rvalue
b)	`int`	5	rvalue
c)	`int`	0	lvalue
d)	`double`	0.5	rvalue
e)	`int`	2	rvalue
f)	`int`	2	lvalue

Solution to Exercise 57.

	type	value
(a)	`bool`	`false`
(b)	`double`	0.5
(c)	`float`	0.5
(d)	`bool`	`true`
(e)	`bool`	`false`
(f)	`int`	11
(g)	`int`	1
(h)	`bool`	`false`
(i)	`bool`	`true`
(j)	`double`	5.5

Solution to Exercise 58.

a) `6 / 4 * 2.0f - 3` $\longrightarrow$
`1 * 2.0f - 3` $\longrightarrow$
`2.0f - 3` $\longrightarrow$
`-1.0f`

b) `2 + 15.0e7f - 3 / 2.0 * 1.0e8` $\longrightarrow^{1.5\cdot 10^8 > 2^{27}}$
`15.0e7f - 3 / 2.0 * 1.0e8` $\longrightarrow$
`15.0e7f - 1.5 * 1.0e8` $\longrightarrow$
`15.0e7f - 1.5e8` $\longrightarrow$
`0.0`

c) `392593 * 2735.0f - 8192 * 131072 + 1.0` $\longrightarrow$
binary: $1\underbrace{0\ldots0}_{\text{23 times}}\underbrace{0011111}_{\text{get lost}}$`f - 8192 * 131072 + 1.0` $\longrightarrow$
`1073741824.0f - 8192 * 131072 + 1.0` $\longrightarrow$
`1073741824.0f - 1073741824 + 1.0` $\longrightarrow$

```
0.0f + 1.0 ⟶
1.0
```

d) `16 * (0.2f + 262144 - 262144.0)` $\longrightarrow$

$$16 * (\text{binary:}1\underbrace{0\ldots0}_{\text{18 times}}.00110\,\underbrace{\overline{0110}}_{\text{get lost}}\texttt{f} - 262144.0) \longrightarrow$$

$$16 * \text{binary:}0.0011 \longrightarrow$$

$$3.0$$

Solution to Exercise 59.

a) This is easy and doesn't take any calculations: $0.25 = 1/4 = 1 \cdot 2^{-2}$. As a binary number, this is 0.01.

b) We employ the rules from Section 2.5.5.

$$\begin{array}{rcrcrcl}
 & & & & 1.52 & \to & b_0 = 1 \\
2(1.52-1) & = & 2\cdot 0.52 & = & 1.04 & \to & b_{-1} = 1 \\
2(1.04-1) & = & 2\cdot 0.04 & = & 0.08 & \to & b_{-2} = 0 \\
2(0.08-0) & = & 2\cdot 0.08 & = & 0.16 & \to & b_{-3} = 0 \\
2(0.16-0) & = & 2\cdot 0.16 & = & 0.32 & \to & b_{-4} = 0 \\
2(0.32-0) & = & 2\cdot 0.32 & = & 0.64 & \to & b_{-5} = 0 \\
2(0.64-0) & = & 2\cdot 0.64 & = & 1.28 & \to & b_{-6} = 1 \\
2(1.28-1) & = & 2\cdot 0.28 & = & 0.56 & \to & b_{-7} = 0 \\
2(0.56-0) & = & 2\cdot 0.56 & = & 1.12 & \to & b_{-8} = 1 \\
2(1.12-1) & = & 2\cdot 0.12 & = & 0.24 & \to & b_{-9} = 0 \\
2(0.24-0) & = & 2\cdot 0.24 & = & 0.48 & \to & b_{-10} = 0 \\
2(0.48-0) & = & 2\cdot 0.48 & = & 0.96 & \to & b_{-11} = 0 \\
2(0.96-0) & = & 2\cdot 0.96 & = & 1.92 & \to & b_{-12} = 1 \\
2(1.92-1) & = & 2\cdot 0.92 & = & 1.84 & \to & b_{-13} = 1 \\
2(1.84-1) & = & 2\cdot 0.84 & = & 1.68 & \to & b_{-14} = 1 \\
2(1.68-1) & = & 2\cdot 0.68 & = & 1.36 & \to & b_{-15} = 1 \\
2(1.36-1) & = & 2\cdot 0.36 & = & 0.72 & \to & b_{-16} = 0 \\
2(0.72-0) & = & 2\cdot 0.72 & = & 1.44 & \to & b_{-17} = 1 \\
2(1.44-1) & = & 2\cdot 0.44 & = & 0.88 & \to & b_{-18} = 0 \\
2(0.88-0) & = & 2\cdot 0.88 & = & 1.76 & \to & b_{-19} = 1 \\
2(1.76-1) & = & 2\cdot 0.76 & = & 1.52 & \to & b_{-20} = 1 \\
 & & & & & \vdots &
\end{array}$$

Phew, finally the sequence becomes periodic, and we get the binary expansion $1.\overline{10000101000111101011}$.

c) We employ the rules from Section 2.5.5.

$$
\begin{array}{rcrcrcl}
 & & & & 1.3 & \to & b_0 = 1 \\
2(1.3-1) & = & 2\cdot 0.3 & = & 0.6 & \to & b_{-1} = 0 \\
2(0.6-0) & = & 2\cdot 0.6 & = & 1.2 & \to & b_{-2} = 1 \\
2(1.2-1) & = & 2\cdot 0.2 & = & 0.4 & \to & b_{-3} = 0 \\
2(0.4-0) & = & 2\cdot 0.4 & = & 0.8 & \to & b_{-4} = 0 \\
2(0.8-0) & = & 2\cdot 0.8 & = & 1.6 & \to & b_{-5} = 1 \\
2(1.6-1) & = & 2\cdot 0.6 & = & 1.2 & \to & b_{-6} = 1 \\
 & & & & & \vdots &
\end{array}
$$

We see that the expansion is periodic and yields the binary number $1.0\overline{1001}$.

d) We write $11.1 = 10 + 1.1$ and add the binary expansion 1010.0 of 10 to the binary expansion $1.0\overline{0011}$ of 1.1 derived in Section 2.5.5. The resulting expansion is $1011.0\overline{0011}$.

Solution to Exercise 60.

a) 0.25 has normalized binary floating point representation $1.0 \cdot 2^{-2}$ and is therefore smaller than any number in $\mathcal{F}^*(2,5,-1,2)$. The nearest number is therefore the smallest number in this system, namely 0.5 with normalized binary representation $1.0 \cdot 2^{-1}$. In $\mathcal{F}(2,5,-1,2)$, we can represent 0.25 exactly as $0.1 \cdot 2^{-1}$.

b) 1.52 has normalized binary floating point representation $1.\overline{10000101000111101011} \cdot 2^0$. To get the nearest number in $\mathcal{F}^*(2,5,-1,2)$, we have to round to 5 significant digits. The result is $1.1000 \cdot 2^0 = 1.5$, obtained by rounding down, since $1.1001 \cdot 2^0 = 1.5625$, obtained by rounding up, is farther away. The nearest number in $\mathcal{F}(2,5,-1,2)$ is the same, since this system has only extra numbers *smaller* than any normalized number. Such numbers cannot be nearest to numbers *larger* than some normalized number.

c) 1.3 has normalized binary floating point representation $1.0\overline{1001}$. To get the nearest number in $\mathcal{F}^*(2,5,-1,2)$, we have to round to 5 significant digits. The result is $1.0101 \cdot 2^0 = 1.3125$, obtained from rounding up, since $1.0100 \cdot 2^0 = 1.25$, obtained from rounding down, is farther away. The nearest number in $\mathcal{F}(2,5,-1,2)$ is the same.

d) 11.1 is larger than any number in the system $\mathcal{F}^*(2,5,-1,2)$. Recall that the largest number is $1.1111 \cdot 2^2 = 4+2+1+1/2+1/4 = 7.75$, and this is the nearest number to 11.1, also in $\mathcal{F}(2,5,-1,2)$.

Solution to Exercise 61. The smallest normalized number is always $2^{e_{\min}}$. In case of single precision, this is 2^{-126}, for double precision, it is 2^{-1022}. Recall that the largest normalized number is

$$\left(1-\left(\frac{1}{\beta}\right)^{p}\right)\beta^{e_{\max}+1}.$$

For single precision, this yields

$$\left(1-\left(\frac{1}{2}\right)^{24}\right)2^{128}=2^{128}-2^{104}.$$

For double precision, we get

$$\left(1-\left(\frac{1}{2}\right)^{53}\right)2^{1024}=2^{1024}-2^{971}.$$

Solution to Exercise 62. For each exponent, $\mathcal{F}^*(\beta,p,e_{\min},e_{\max})$ has $\beta-1$ possibilities for the first digit, and β possibilities for the remaining $p-1$ digits. The size of $\mathcal{F}^*(\beta,p,e_{\min},e_{\max})$ is therefore

$$2(e_{\max}-e_{\min}+1)(\beta-1)\beta^{p-1},$$

if we take the two possible signs into account.

$\mathcal{F}(\beta,p,e_{\min},e_{\max})$ has extra nonnegative numbers of the form

$$0.d_1\ldots d_{p-1}2^{e_{\min}},$$

and there are β^{p-1} of them. Adding the non-positive ones and subtracting 1 for counting 0 twice, we get

$$2\beta^{p-1}-1$$

extra numbers.

Solution to Exercise 63. The binary expansion of 0.1 is $0.0\overline{0011}$, obtained from the representation of 1.1 by subtracting 1. This value has to be rounded to the nearest value with 24 significant digits. Let us write out the expansion so that we get the first 26 significant digits of $0.0\overline{0011}$:

$$0.00011001100110011001100110011.$$

It follows that we have to round up to 1 at digit 24 to get the nearest `float` value

$$1.10011001100110011001101\cdot 2^{-4}.$$

To see how this value differs from 0.1, let's convert it back into decimal representation. Interestingly, this is always possible without any error, since 0.1 (binary) is 0.5 (decimal), 0.01 (binary) is 0.25 (decimal), and so on. The decimal value that we obtain is

$$0.100000001490116119384765625.$$

Solution to Exercise 64. We compare floating point numbers for equality in `i != 1.0`, although one of them (namely the value of `i`) is the result of inexact computations, assuming a base-2 floating point number system. The inexactness comes from the rounding of 0.1 to a floating point number, and from the subsequent addition of numbers. In practice, this leads to an infinite loop, since `i != 1.0` will always be satisfied.

Solution to Exercise 65. The problem is that this is an infinite loop. Here is why: The literal `100000000.0f` has as its value the nearest `float` value to 10^8. What is this value? It *is* 10^8: the binary representation of 10^8 is 101111101011110000100000000 (as you can for example see by running program `dec2bin2.cpp`, see Exercise 50). This has 19 significant digits and therefore easily fits into a `float` value. But the total number of binary digits of 10^8 is 27. Now, the loop tries to run through all numbers up to 10^8. This works as long as the numbers have 24 or less binary digits in total, since they are exactly representable as `float` values in this case.

But at some point, the loop reaches the number 2^{24} which is the first one with a 25-digit binary representation, namely 1000000000000000000000000. This value is still representable over the type `float`, since the number of significant digits is 1. The next number is $2^{24} + 1$ with binary representation 1000000000000000000000001. This is still a number with 25 binary digits, but now all of them are significant. A `float` value can accommodate only 24 significant digits, and the value is therefore being rounded to the next representable number. In this case, there are two such numbers, 2^{24} and $2^{24} + 2$, and by the round-to-even rule, 2^{24} is chosen.

Thus, the loop does not make progress beyond 2^{24} which means that it never terminates. The following program checks for the first repetition in the value of `i`.

```
// demonstrates the effect of floating-point arithmetic on
// a simple loop

#include<iostream>

int main()
{
  // we suspect that i does not grow beyond 2^24,
  // so we check whether this is really the case
  float old_i = -1.0f;
  for (float i = 0.0f; i < 100000000.0f; ++i)
    if (i == old_i) {
      // infinite loop!
      std::cout << "loop gets stuck with i = " << int(i) << ".\n";
      break;
    } else
      old_i = i;
  return 0;
}
```

On the authors' platform, the first repeating value is indeed 2^{24}:

```
loop gets stuck with i = 16777216.
```

Solution to Exercise 66.

```
// Prog: dec2float.cpp
// compute the float representation of a number
// in the open interval (0,2)

#include <iostream>

int main ()
{
  // input
  std::cout << "Decimal number x (0 < x < 2) =? ";
  float x;
  std::cin >> x;

  // x = w * 2^e
  float w = x;
  int e = 0;

  // as long as w < 1, decrement e and double w
  for ( ; w < 1.0f; w *= 2.0f) --e;

  // Now we have 1 <= w < 2, apply rule from lecture
  std::cout << "Significand: ";
  for ( ; w != 0.0; w = 2.0f * (w - int(w)))
    std::cout << int(w);

  std::cout << "\nExponent: " << e << "\n";

  return 0;
}
```

Solution to Exercise 67.

```
// Prog: double_integer.cpp
// tests whether a given double value is integer
//
#include<iostream>

int main()
{
  // input
  std::cout << "Decimal number =? ";
  double d;
  std::cin >> d;
  const double abs_d = d > 0 ? d: -d; // |d|

  // |d| can be written in the form m * 2^e, where m is a
  // natural number whose last binary digit is 1. Then |d| is
  // integer if and only if e>=0. Having m, we can therefore
  // conclude that d is integer if and only if |d| >= m

  // step 1: normalize such that number is in [1,2)
  double dd = abs_d;
  while (dd >=2) dd/=2;  // ensure dd < 2;
  while (dd < 1) dd*=2;  // ensure dd >=1;
```

```

  // step 2: compute binary expansion m like in the lecture notes
  double m = 0;
  while (dd > 0) {
    // move last binary digit of dd into m
    m *= 2;
    if (dd >= 1) {
      m += 1;
      dd = 2*(dd-1);
    } else
      dd = 2*dd;
  }

  // step 3: compare with abs_d
  std::cout << d;
  if (abs_d >= m)
    std::cout << " is integer.\n";
  else
    std::cout << " is not integer.\n";

  return 0;
}
```

Solution to Exercise 68. Here is the program based on the first formula.

```
// Prog: pi1.cpp
// approximate pi according to first n terms of the formula
// pi = 4 - 4/3 + 4/5 - 4/7 ...

#include <iostream>

int main ()
{
  // input
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // computation (forward sum)
  double pif = 0.0;
  for (int i = 1; i < 2*n; i += 2)
    if (i % 4 == 1)
      pif += 4.0 / i;
    else
      pif -= 4.0 / i;

  // computation (backward sum)
  double pib = 0.0;
  for (int i = 2*n-1; i > 0; i -= 2)
    if (i % 4 == 1)
      pib += 4.0 / i;
    else
      pib -= 4.0 / i;

  // output
   std::cout << "Pi is approximately "
            << pif << " (forward sum), or "
            << pib << " (backward sum); the difference is "
            << pif - pib << "\n";

  return 0;
}
```

When you run it for $n = 10,000$, for example, it gives on our platform the approximation 3.14139 (still off in the fourth digit after the decimal point). For $n = 100,000$, we get 3.14157 (still off in the fifth digit after the decimal point). For $n = 1,000,000$, finally, the result is correct to five digits after the decimal point: 3.14159.

Here is the approximation based on the second formula.

```
// Prog: pi2.cpp
// approximate pi according to the first n terms of the formula
// pi = 2 + 2*1 / 3 + 2*1*2 / 3*5 + 2*1*2*3 / 3*5*7

#include <iostream>

int main ()
{
  // input
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // auxiliary variables
  // initialized for first term of forward sum (i=0)
  double numer = 2.0; // numerator i-th term
  double denom = 1.0; // denominator i-th term

  // forward sum
  // pif: value after term i (i=0 initially, then i=1,2,...,n-1)
  double pif = 2.0;
  for (int i = 1; i < n; ++i)
    pif += (numer *= i) / (denom *= (2*i + 1)); // update to term i
  // now numer and denom are the ones for i=n-1

  // backward sum
  // pib: value after term i (i=n-1 initially, then i=n-2,...,1,0)
  double pib = numer / denom;
  for (int i = n-1; i >= 1; --i) {
    pib += (numer /= i) / (denom /= (2*i + 1)); // update to term i-1
  }

  // output
  std::cout << "Pi is approximately "
            << pif << " (forward sum), or "
            << pib << " (backward sum); the difference is "
            << pif - pib << "\n";

  return 0;
}
```

This already gives the result 3.14159 for $n = 17$ on our platform, so this version is obviously preferable.

Solution to Exercise 69. Here is the program, including the function(s) and a test; for large x, the idea is to use periodicity of $\sin(x)$, meaning that $\sin(x) = \sin(x - 2\pi)$ for all x. This means, we can reduce the argument modulo 2π to make it small, and only then call the function `sinus`.

```
// Prog: sinus.cpp
// implements and tests functions for computing sin(x)
#include<iostream>
```

```

// POST: returns an approximation of sin(x)
double sinus (double x) {
   double result = x;
   double temp = x;
   for (int i = 1; i < 20; ++i) {
      temp *= x / (2*i+1) * x / (2*i);
      if (i % 2 == 0) {
        result += temp;
      } else {
        result -= temp;
      }
   }

   return result;
}

// POST: returns an approximation of sin(x)
double sinus2  (double x) {
  // before we start, reduce x modulo 2pi. The following is not a clean
  // solution since it overflows for large x, but it shows the idea.
  double pi = 3.1415926535;
  int y = int(x / (2*pi));
  x -= y*2*pi;
  return sinus(x);
}

// test function
void test_sinus (double x) {
  std::cout << "sin (" << x << ") = "
            << sinus(x) << " ; " << sinus2(x) << "\n";
}

// now test it
int main() {
  double pi = 3.1415926535;
  test_sinus (0);                    // should be 0
  test_sinus (pi/6);                 // 30 degrees; should be 1/2
  test_sinus (pi/4);                 // 45 degrees; should be 1/2 * sqrt(2)
  test_sinus (pi/3);                 // 60 degrees; should be 1/2 * sqrt(3)
  test_sinus (pi/2);                 // should be 1
  test_sinus (pi/2 + 10000*2*pi);    // should be 1, due to periodicity

}
```

Solution to Exercise 70.

```
// Program: babylonian.cpp
// Approximation of the square root of a positive real number

#include <iostream>

int main(){

    // Read input
    double s; // input number
    const double eps = 0.001; // the epsilon, i.e. max square error

    std::cout << "Which number do you want to take the square root of?";
    std::cout << "\n";
    std::cin >> s;

    // Compute square root
```

```
    double x = s / 2.0; // initialize solution
    unsigned int n = 0; // counter for number of iterations

    while (x * x - s >= eps || s - x * x >= eps) {
        ++n;
        x = (x + s / x) / 2.0;
    }

    std::cout << "The square root of " << s << " is: " << x << std::endl;
    std::cout << "The number of iterations done: " << n << std::endl;

    return 0;
}
```

Solution to Exercise 71.

```
// Program: fpsys.cpp
// Provide a graphical representation of floating point numbers

#include <iostream>
#include <IFMP/window>

int main()
{
  // Input parameters of floating point system
  std::cout << "Draw F*(2,p,e_min,e_max).\np =? ";
  unsigned int p;
  std::cin >> p;
  std::cout << "e_min =? ";
  int emin;
  std::cin >> emin;
  std::cout << "e_max =? ";
  int emax;
  std::cin >> emax;

  // We compute significands using integral arithmetic, that is,
  // scaled by 2^(p-1).

  // compute the smallest normalized significand 2^(p-1)
  unsigned int smin = 1;
  for (unsigned int i = 1; i < p; ++i) smin *= 2;
  // compute the largest normalized significand (2^p)-1
  const unsigned int smax = 2 * smin - 1;
  // compute 2^emin
  double pemin = 1;
  for (int i = 0; i < emin; ++i) pemin *= 2;
  for (int i = 0; i > emin; --i) pemin /= 2;
  // compute 2^emax
  double pemax = 1;
  for (int i = 0; i < emax; ++i) pemax *= 2;
  for (int i = 0; i > emax; --i) pemax /= 2;

  // For each positive number x of the system draw a circle
  // with radius x around the window center

  // parameters to scale output
  const int cx = (ifmp::wio.xmax() - ifmp::wio.xmin()) / 2;
  const int cy = (ifmp::wio.ymax() - ifmp::wio.ymin()) / 2;
  const double scale = cx / (pemax * smax);

  // zero
  ifmp::wio << ifmp::Point(cx, cy);
  // loop over all normalized significands
```

```
  for (unsigned int i = smin; i <= smax; ++i)
    // loop over all exponents
    for (double m = pemin; m <= pemax; m *= 2)
      ifmp::wio << ifmp::Circle(cx, cy, int(m * i * scale));

  ifmp::wio.wait_for_mouse_click();
  return 0;
}
```

Solution to Exercise 72. Mr. Plestudent is wrong. When we use the algorithm for constructing the hexadecimal representation, we see a periodic pattern, just as for $\beta = 2$.

$$\begin{array}{rcrcrcl} & & & & 0.1 & \rightarrow & b_0 = 0 \\ 16(0.1-0) & = & 16\cdot 0.1 & = & 1.6 & \rightarrow & b_{-1} = 1 \\ 16(1.6-1) & = & 16\cdot 0.6 & = & 9.6 & \rightarrow & b_{-2} = 9 \\ 16(9.6-9) & = & 16\cdot 0.6 & = & 9.6 & \rightarrow & b_{-3} = 9 \\ & & & & & \vdots & \end{array}$$

Somewhat less formally, we can also argue as follows: the hexadecimal representation is obtained from the binary one by repeatedly grouping 4 consecutive binary digits in $\{0, 1\}$ into one hexadecimal digit in $\{0, 1, \ldots, 9, A, B, \ldots, F\}$. Then, if the binary expansion is infinite, so will be the hexadecimal one.

Again a bit more formal for the careful reader: Let's assume we have a representation

$$\frac{1}{10} = \sum_{i=1}^{p} d_i 16^{-i}.$$

We can write each d_i as a four-digit binary number:

$$d_i = d_{i_3}2^3 + d_{i_2}2^2 + d_{i_1}2^1 + d_{i_0}2^0, \quad d_{i_3}, d_{i_2}, d_{i_1}, d_{i_0} \in \{0, 1\}.$$

We then get

$$\frac{1}{10} = \sum_{i=1}^{p} d_i 16^{-i} = \sum_{i=1}^{p} d_i 2^{-4i} = \sum_{i=1}^{p} (\sum_{j=0}^{3} d_{i_j} 2^j) 2^{-4i} = \sum_{i=1}^{p} \sum_{j=0}^{3} d_{i_j} 2^{j-4i},$$

and this is a finite binary floating-point number, in contradiction to the known fact there is no such number.

Solution to Exercise 73. Because of

$$2^{-i} = 5^i \cdot 10^{-i},$$

we have

$$b = \sum_{i=1}^{k} b_i 2^{-i} = \sum_{i=1}^{k} (b_i 5^i) 10^{-i},$$

where all values $b_i 5^i$ are natural numbers. Thus, b is a finite sum of numbers of the form 10^{-i} that obviously have (very simple) finite decimal representations. But then the sum also has finite representation (just add the numbers up, using elementary school arithmetic).

Solution to Exercise 74. Here is the characterization. γ refines β if and only if all prime factors of β are also prime factors of γ. For example, 10 refines 2, since 2 is also a prime factor of 10. But 2 does not refine 10, since 5 is a prime factor of 10 but not of 2. Maybe surprisingly, a smaller number may refine a larger number, for example, 2 refines 16.

To prove the characterization, we need to show two directions. Let us first assume that all prime factors of β are also prime factors of γ. Now consider any number

$$r = s \cdot \sum_{i=0}^{p-1} d_i \beta^{-i} \cdot \beta^e \in \mathcal{F}(\beta, p, e_{min}, e_{max}),$$

i.e. $s \in \{-1, 1\}$, $d_i \in \{0, \ldots, \beta - 1\}$ for all i, and $e \in \{e_{min}, \ldots, e_{max}\}$. Since all prime factors of β are also prime factors of γ, there is some integer power γ^M of γ that is a multiple of β (choose M such that every prime factor of γ^M has at least the multiplicity that it has in β). So we have an integer m with

$$\beta \cdot m = \gamma^M,$$

which implies that

$$\beta^{-i} \cdot \beta^e = \beta^{e-i} = \begin{cases} \beta^{e-i}\gamma^0, & \text{if } e - i \geq 0, \\ m^{i-e}\gamma^{M(e-i)}, & \text{if } e - i < 0. \end{cases}$$

This means that r is a finite sum of integer powers of γ. Adding up all these powers as we do in elementary school shows that $r \in \mathcal{F}(\gamma, q, f_{min}, f_{max})$ for suitable q, f_{min}, f_{max}. Since $\mathcal{F}(\beta, p, e_{min}, e_{max})$ contains only finitely many numbers, we can choose q, f_{min}, f_{max} that work for all of them, and we are done.

For the other direction, let us now assume that there is a prime factor t of β that is not a prime factor of γ. We will show that there is a number in $\mathcal{F}(\beta, 2, 0, 0)$ that is not contained in any system of the form $\mathcal{F}(\gamma, q, f_{min}, f_{max})$. This number is simply β^{-1}. Suppose for a contradiction that we could write β^{-1} in the form

$$\beta^{-1} = \sum_{i=0}^{q-1} d_i \gamma^{-i} \cdot \gamma^f.$$

Multiplying the equation with a sufficiently large power γ^M of γ makes the right-hand side (and thus also the left-hand side) integer. Hence,

$$\frac{\gamma^M}{\beta}$$

is integer, so γ^M is a multiple of β. This, however, is a contradiction to γ not containing prime factor t.

Solution to Exercise 75.

```
// Prog: mandelbrot.cpp
// draws (a part of) the Mandelbrot set and allows the user to
// zoom in by clicking with the mouse on the region to be enlarged
//
// The Mandelbrot set is defined as the set of all complex numbers
// c such that the complex iteration formula z := z^2 + c (starting
// with z=0) always yields values z of absolute value at most two.
// In the computations below, we perform a large but fixed number
// of steps of this iteration for a given c; if all computed values
// are at most two in absolute value, we consider c as part of the
// Mandelbrot set (and depict its corresponding pixel in black),
// otherwise we draw a white pixel.

#include<IFMP/window>

int main()
{
  // the currently considered subset of the complex plane, initially
  // [-2, 1] x [-1, 1] (covers the so-called main cardioid of the
  // Mandelbrot set)
  double r_min = -2; double r_max = 1;
  double i_min = -1; double i_max= 1;

  // window scaling factor; change this for larger/smaller display
  // window
  double window_scale = 500;

  // zoom factor from one iteration to the next
  double zoom_factor = 10;

  // the display window dimensions in pixels (window should be
  // congruent to the current complex plane subset)
  int x_size = int (window_scale * (r_max - r_min));
  int y_size = int (window_scale * (i_max - i_min));

  // open the display window
  ifmp::Wstream w (x_size, y_size,
                   "The Mandelbrot set (click to zoom in)");

  // maximum number of iterations (the higher, the more accurate;
  // the lower, the faster)
  unsigned int max_iter = 500;

  // main drawing loop; one iteration for every zoom scale
  for(;;) {
    // go through all pixels
    for (int x=0; x<x_size; ++x)
      for (int y=0; y<y_size; ++y) {

        // compute corresponding point in complex plane
        const double r = r_min + x / window_scale;
        const double i = i_min + y / window_scale;

        // do the Mandelbrot iteration for that point
        // interpreted as complex number c = (r,i)
        double r_z = 0; // z (real part)
        double i_z = 0; // z (imaginary part)
        unsigned int iter = 0;
```

```
59          while (iter < max_iter && r_z * r_z + i_z * i_z <= 4) {
60            // |z| <= 2; replace z by z^2 + c
61            double h = r_z * r_z - i_z * i_z + r; // new z_r
62            i_z = 2 * r_z * i_z + i;              // new z_i
63            r_z = h;
64            ++iter;
65          }
66          // coloring: max_iter -> black, other -> white
67          if (iter == max_iter)
68            w.set_color (w.number_of_colors()-2); // black
69          else
70            w.set_color (w.number_of_colors()-1); // white
71          w << ifmp::Point (x, y);
72        }
73
74      // zoom in; new center is mouse click position
75      int x_c; int y_c;
76      w.get_mouse_click (x_c, y_c);
77      const double r_c = r_min + x_c / window_scale;
78      const double i_c = i_min + y_c / window_scale;
79      const double r_span = r_max - r_min;
80      const double i_span = i_max - i_min;
81      r_min = r_c - 0.5 * r_span / zoom_factor;
82      r_max = r_c + 0.5 * r_span / zoom_factor;
83      i_min = i_c - 0.5 * i_span / zoom_factor;
84      i_max = i_c + 0.5 * i_span / zoom_factor;
85      window_scale *= zoom_factor;
86      w.clear();
87    }
88
89    return 0;
90  }
```

Solution to Exercise 76. CGAL is the *Computational Geometry Algorithms Library*, an open source C++ library of data structures and algorithms for solving geometric problems. The CGAL homepage is `www.cgal.org`.

`CGAL::orientation` is a function that determines for three given points $p, q, r \in \mathbb{R}^2$ whether r lies to the left, on, or to the right of the oriented line through p and q. The resulting values (`CGAL::LEFTTURN`, `CGAL::COLLINEAR`, or `CGAL::RIGHTTURN`) define the orientation of the point triple $\{p, q, r\}$. `CGAL::LEFTTURN` means that p, q, r appear in counterclockwise order around the triangle spanned by p, q, r, while `CGAL::RIGHTTURN` signals clockwise order. `CGAL::COLLINEAR` means that all three points are on a common line, so the triangle is "flat".

The writer of the email is surprised since the orientation of a point triple should not change when all point coordinates are multiplied with a fixed scalar (in this case 100). But in reality, it does change, at least according to the function `CGAL::orientation`.

The reason is that the integer coordinates of the points $(14, 22), (15, 21), (19, 17)$ can be converted to `float` or `double` (we don't exactly know which of the two the writer of the email is using) without any error. In contrast, some of the coordinates of $(0.14, 0.22), (0.15, 0.21), (0.19, 0.17)$ don't have finite binary representations, so in converting them to `float` or `double`, errors are inevitable. Since the points are mathematically collinear (on the same line), it is clear that the tiniest error is enough to destroy this property. That's why `CGAL::orientation` delivers a result different from

CGAL::COLLINEAR.

Here is what you could answer the writer of the email.

```
Hi,

assuming that you use type float or double to represent the point
coordinates, the inconsistency that you reported is due to the
conversion of point coordinates from the decimal input format to the
internally used binary format. Decimal integers like 14, 22 etc. can be
represented exactly in binary format, and CGAL::orientation returns
the correct answer CGAL::COLLINEAR for the three points with integer
coordinates. But decimal fractions like 0.14, 0.22 etc. do not
necessarily have finite representations in binary format. This is like
trying to write the number 1/3 as a decimal fraction. The best you can
do is 0.33333... but wherever you stop, you make a small error.

Now, CGAL::orientation sees the points (0.14, 0.22), (0.15, 0.21) and
(0.19,0.17) only after the conversion to binary format, and this
conversion introduces some (tiny) errors. But since the points are
mathematically collinear, even the tiniest errors may have the effect
of destroying collinearity. This is exactly what you observed.

The problem is inevitable in working with floating-point numbers, since
you cannot circumvent the decimal-to-binary conversion. All you can do
is to only use point coordinates (integers, for example) for which the
conversion is exact.
```

B.6 A first C++ function

Solution to Exercise 77.

a)
```
// POST: return value is the maximum of i,j and k
```

b)
```
// PRE:  0 not contained in {i,...,j}
// POST: return value is the sum 1/i + 1/(i+1) + ... + 1/j
```

Solution to Exercise 78.

a) If `i` is odd, the execution does not reach a return statement, and the function call expression is invalid. It seems that we want a function that returns `true` if and only if `i` is even. This can be done as follows.

```
bool is_even (int i)
{
  // POST: return value is true if and only if i is even
  return (i % 2 == 0);
}
```

b) If `x` has value 0, `result` will never be set to a defined value. The corresponding function call expression has undefined value. To fix this, we can either make `x != 0.0` a precondition (but then we don't have to check it like the function does it), or we can invent some return value for the case where `x` has value 0. In any case, we don't need the variable `result`. Here are the two variants.

```
double inverse (double x)
{
  // PRE:  x != 0
  // POST: return value is 1/x
  return 1.0 / x;
}

double inverse (double x)
{
  // POST: return value is 1/x for x !=0, and 0 otherwise
  if (x != 0.0)
    return 1.0 / x;
  else
    return 0.0;
}
```

Solution to Exercise 79. The program outputs the 7-th power of the input value i. The computation takes place in the function g that multiplies i with f(i) (i^2) and f(f(i)) (i^4).

Solution to Exercise 80. Here are the three problems.

a) The call of g(2.0 * x) in the function body of f is not in the scope of the function g, since that function is only declared later through its definition. Consequently, g cannot be used in f.

b) In the function g, the modulus operator is used with double operands, but for floating point number type operands, there is no modulus operator.

c) The function body of h is not in the scope of the variable result, since that variable is only declared later. Consequently, result cannot be used in h.

Solution to Exercise 81.

```
// Program: fpsys2.cpp
// Provide a graphical representation of floating point numbers

#include <iostream>
#include <cmath>
#include <IFMP/window>


int main()
{
  // Input parameters of floating point system
  std::cout << "Draw F(2,p,e_min,e_max).\np =? ";
  unsigned int p;
  std::cin >> p;
  std::cout << "e_min =? ";
  int emin;
  std::cin >> emin;
  std::cout << "e_max =? ";
  int emax;
  std::cin >> emax;

  // compute the smallest normalized significand 2^(p-1)
  const unsigned int smin = (unsigned int)(std::pow (2.0, double(p-1)));
  // compute the largest normalized significand (2^p)-1
  const unsigned int smax = 2 * smin - 1;
  // compute 2^emin
  const double pemin = std::pow (2.0, double(emin));
  // compute 2^emax
  const double pemax = std::pow (2.0, double(emax));

  // For each positive number x of the system draw a circle
  // with radius x around the window center

  // parameters to scale output
  const int cx = (ifmp::wio.xmax() - ifmp::wio.xmin()) / 2;
  const int cy = (ifmp::wio.ymax() - ifmp::wio.ymin()) / 2;
  const double scale = cx / (pemax * smax);

  // zero
```

```
  ifmp::wio << ifmp::Point(cx, cy);
  // loop over all normalized significands
  for (unsigned int i = smin; i <= smax; ++i)
    // loop over all exponents
    for (double m = pemin; m <= pemax; m *= 2)
      ifmp::wio << ifmp::Circle(cx, cy, int(m * i * scale));

  ifmp::wio.wait_for_mouse_click();
  return 0;
}
```

Solution to Exercise 82. The program still works if $s(x) \geq \sqrt{x}$, so we only have a potential problem if $s(x) < \sqrt{x}$. Because $|s(x) - \sqrt{x}|$ equals $\sqrt{x} - s(x)$ in this case, the relative error bound gives us $\sqrt{x} - s(x) \leq \varepsilon\sqrt{x}$, and this implies

$$s(x) \geq (1 - \varepsilon)\sqrt{x} \geq \frac{1}{2}\sqrt{x}.$$

It follows that

$$2s(x) \geq \sqrt{x},$$

meaning that we can safely use `2*std::sqrt(n)` instead of `std::sqrt(n)` in Program 2.22. Note that $2s(x)$ is indeed representable as a floating point number again, since we have assumed the system to be binary.

Solution to Exercise 83. The desired number of twin primes is 58980 as the following program shows that implements the approach of a). It turns out that we have to use the fast prime number test from Program 2.22 in order not to wait too long (we still have to wait pretty long).

```
// Program: twinprimes.cpp
// Count twin primes in 2,...,10000000

#include <iostream>
#include <cmath>

// POST: return value is true if and only if n is prime
bool is_prime (const unsigned int n)
{
  if (n < 2) return false; // 0 and 1 are not prime

  // Computation: test possible divisors d up to sqrt(n)
  const unsigned int bound = (unsigned int)(std::sqrt(n));
  unsigned int d;
  for (d = 2; d <= bound && n % d != 0; ++d);

  // Output
  return d > bound;
}

int main ()
{
  // keep primality info for odd i and i+2
  bool curr = false; // i = 1
```

```
  bool next = true;  // i = 3
  int twins = 0;     // number of twins
  for (int i = 3; i < 9999999; i += 2) {
    curr = next;          // i
    next = is_prime(i+2); // i+2
    if (curr && next) ++twins;
  }
  std::cout << "Number of twin primes: " << twins << "\n";

  return 0;
}
```

A much faster approach is based on *Eratosthenes's sieve*. We simply compute all prime numbers in the given range, and then use this information to select the twin primes:

```
// Program: twinprimes2.cpp
// Count twin primes in 2,...,10000000

#include <iostream>
#include <algorithm>

int main()
{
  // definition and initialization: provides us with
  // Booleans crossed_out[0],..., crossed_out[9999999]
  bool crossed_out[10000000];
  std::fill (crossed_out, crossed_out + 10000000, false);

  // computation of all prime numbers in the range
  for (unsigned int i = 2; i < 10000000; ++i)
    if (!crossed_out[i])
      // cross out all proper multiples of i
      for (unsigned int m = 2*i; m < 10000000; m += i)
        crossed_out[m] = true;

  // now count twin primes: (i-2, i) = (3,5) is first pair
  unsigned int twins = 0;
  for (unsigned int i = 5; i < 10000000; i+=2)
    if (!crossed_out[i-2] && !crossed_out[i])
      ++twins;

  // output
  std::cout << "Number of twin primes: " << twins << "\n";

  return 0;
}
```

This shows that the subdivision of the task into subtasks according to a) is not appropriate in this case.

Solution to Exercise 84.

```
double pow (double b, int e)
{
  // PRE:  e >= 0 || b != 0.0
  // POST: return value is b^e
  double result = 1.0;
  if (e < 0) {
```

```
    // b^e = (1/b)^(-e)
    b = 1.0/b;
    e = -e;
  }
  // maintain bpow = b^(2^i), initialized for i=0
  for (double bpow = b; e != 0; e /= 2) {
    if (e % 2 == 1) result *= bpow;
    bpow *= bpow; // square bpow to get b^(2^(i+1))
  }
  return result;
}
```

Solution to Exercise 85. Given the input date, the main task is to count the number of days that have passed since January 1, 1900. Once we have that information, we can take the number modulo 7, and this determines the weekday.

Here is a list of the major subtasks (your list might be different).

1. find out whether the input date is legal;
2. count the number of days in all *years* preceding the input date;
3. count the number of days in all *months* preceding the input date (this has 2 as a subtask);
4. count the number of days preceding the input date (this has 3 as a subtask);
5. given this count, output the weekday.

Smaller subtasks needed by the above are the following.

1. find out whether a given year is a leap year;
2. find out how many days a given month has in a given year;
3. compute the weekday as a number in $\{0, \ldots, 6\}$;
4. transform the number of a weekday into its name.

Here is the program resulting from this subdivision into subtasks.

```
// Prog: perpetual_calendar
// compute the weekday for any given date >= 01.01.1900 (Monday)

#include<iostream>
#include<cassert>

// PRE:  year >= 1900
// POST: return value is true iff year is a leap year
// ---------------------------------------------------
bool is_leap_year (const unsigned int year)
```

```
{
  assert (year >= 1900);
  return year % 4 == 0 && (year % 100 != 0 || year % 400 == 0);
}

// PRE:  year >= 1900, 1 <= month <= 12
// POST: return value is the number of days in month.year
// -------------------------------------------------------
unsigned int days_in_month (const unsigned int month,
                            const unsigned int year)
{
  assert (year >= 1900);
  assert (1 <= month && month <= 12);
  unsigned int days[12] =
    {31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31};
  return days[month-1] + (month==2 && is_leap_year (year));
}

// POST: return value is true iff day.month.year is
//       an existing date  >= 01.01.1900
// ----------------------------------------------
bool is_date (const unsigned int day, const unsigned int month,
              const unsigned int year)
{
  return
    year >= 1900 && 1 <= month && month <= 12 &&
    1 <= day && day <= days_in_month (month, year);
}


// PRE:  year >= 1900
// POST: return value is the number of days in years [1900, ..., year)
// -------------------------------------------------------------------
unsigned int days_up_to_year (const unsigned int year)
{
  assert (year >= 1900);
  unsigned int d = 0;
  // the following could be done more efficiently, but why?
  for (unsigned int y = 1900; y < year; ++y)
      d += 365 + is_leap_year (y);
  return d;
}

// PRE:  year >= 1900, 1 <= month <= 12
// POST: return value is the number of days in
//       months [01.1900, ..., month.year)
// -----------------------------------------
unsigned int days_up_to_month (const unsigned int month,
                               const unsigned int year)
{
  assert (year >= 1900);
  assert (1 <= month && month <= 12);
  unsigned int days = days_up_to_year (year); // before year
  for (unsigned int m = 1; m < month; ++m)    // before month
    days += days_in_month (m, year);
  return days;
}

// PRE:  is_date (day, month, year)
// POST: return value is the number of days in
//       [01.01.1900, ..., day.month.year)
// ---------------------------------------
unsigned int days_up_to (const unsigned int day,
                         const unsigned int month,
                         const unsigned int year)
```

```
{
  assert (is_date (day, month, year));
  return days_up_to_month (month, year) + day - 1;
}

// PRE:  is_date (day, month, year)
// POST: return value is weekday (0 = Monday,...,6 = Sunday)
// ---------------------------------------------------------
unsigned int weekday (const unsigned int day,
                      const unsigned int month,
                      const unsigned int year)
{
  assert (is_date (day, month, year));
  return days_up_to (day, month, year) % 7;
}

// PRE:  weekday < 7
// POST: writes the name of the weekda to standard output
//       (0 = Monday,...,6 = Sunday)
// --------------------------------
void print_weekday (const unsigned int weekday) {
  assert (weekday < 7);
  if      (weekday == 0) std::cout << "Monday";
  else if (weekday == 1) std::cout << "Tuesday";
  else if (weekday == 2) std::cout << "Wednesday";
  else if (weekday == 3) std::cout << "Thursday";
  else if (weekday == 4) std::cout << "Friday";
  else if (weekday == 5) std::cout << "Saturday";
  else                   std::cout << "Sunday";
  std::cout << "\n";
}

int main()
{
  // input date
  std::cout << "Compute weekday of date (day/month/year) for\n";
  std::cout << "day =? ";
  unsigned int day; std::cin >> day;
  std::cout << "month =? ";
  unsigned int month; std::cin >> month;
  std::cout << "year =? ";
  unsigned int year; std::cin >> year;

  // check date
  if (!is_date (day, month, year)) {
    std::cout << "Illegal date.\n";
    return 1;
  }

  // output weekday
  print_weekday (weekday (day, month, year));

  return 0;
}
```

Solution to Exercise 86. Here we show how it is done under Unix-type platforms and the g++ compiler. Let's assume that your home directory is `/home/myhome/`. Under this directory, you now create a subdirectory named `libifm`, with two subdirectories `include` and `lib`. In the subdirectory `include`, you create another subdirectory `IFMP`.

Now you copy `math.h` into the `include/IFM` subdirectory and `math.cpp` into the `lib` subdirectory.

(i) Build the object code. For this, go to the `lib` subdirectory and tell the compiler to generate the object code file `math.o`. For this, you type

```
g++ -I/home/myhome/libifm/include -c math.cpp
```

The directory after the `-I` is the one where the compiler will look for your include files. The path `IFMP/math.h` in `math.cpp` is relative to that include directory. You can provide several include directories through several `-I`'s.

(ii) Build the library from the o-files in the `lib` subdirectory. For this, you type

```
ar r libmath.a *.o
```

This tells the archive program `ar` to put all object files found in the current directory into a single library called `libmath.a`.

(iii) Build the executable from whatever directory you want, by typing

```
g++ -I/home/myhome/libifm/include -L/home/myhome/libifm/lib
    callpow4.cpp -lmath -o callpow4
```

This tells g++ to compile (and link) the program `callpow4.cpp`, using the library `libmath.a` found in the directory that is specified after `-L`.

Solution to Exercise 87.

By the C++ standard, converting a floating point number to `int` cuts off the fractional part. Then we compute the error between the original number and its truncation. Under the IEEE standard 754, this difference has one significant digit less than and is therefore exactly representable as a `double` value, *unless* the input value is negative and has largest possible exponent. But in this case, the truncation is also out of range, and we don't care. Now we can exactly test whether the error is at most 0.5 (in which case we return the truncation), or more (in which case we return the next integer, going away from zero).

```
#include <iostream>

// POST: return value is the integer nearest to x; if there are
//       two nearest integers, the one closer to 0 is chosen
int round (const double x)
{
  const int trunc = int(x); // rounds towards 0 by standard
  const double error = x - trunc; // note: result is exact!
  if (error > 0.5)
    // x was positive, and trunc + 1 is nearer
    return trunc + 1;
  if (error < -0.5)
```

```
      // x was negative, and trunc - 1 is nearer
      return trunc - 1;
    // |error| <= 0.5, trunc is closest integer
    return trunc;
  }

  int main ()
  {
    for (double d = -2; d <= 2; d += 0.25)
      std::cout << "closest integer to " << d << " is "
                << round (d) << "\n";

    return 0;
  }
```

As far as the integration into the library is concerned, please follow the procedure in the solution to Exercise 86 above.

Solution to Exercise 88. We apply a slightly tuned brute-force solution (but maybe you found something better). We in fact compute all possible powers, and for each one compute the cross sum. We use the type `ifmp::integer` in order to have sufficient precision. The only slight improvement is that we compute the cross sum not by the straightforward method (going through the number digit by digit), but by breaking the number up into larger chunks through using a divisor larger than 10. On the platform of the authors, the divisor in the program below gives the fastest program. Its output is

```
Best a = 99
Best b = 95
```

The program below is too slow for $a, b < 1000$, so we don't know the answer for this case.

```
// Prog: power_cross_sums.cpp
// find the values a and b smaller than 100 such that
// a^b has maximum cross sum

#include <iostream>
#include <IFMP/integer.h>

// POST: returns sum of decimal digits of n
ifmp::integer cross_sum (ifmp::integer n)
{
  // the straightforward method
  ifmp::integer cross = 0;
  for (; n > 0; n /= 10) {
    cross += n % 10;
  }
  return cross;
}

// POST: returns sum of decimal digits of n
ifmp::integer cross_sum_big (ifmp::integer n)
{
  // a faster method; use a larger divisor
  // d to reduce n to n /d, and apply the
  // straightforward method to n % d
  const ifmp::integer d = 10000000;
  ifmp::integer cross = 0;
```

```
  while (n > 0) {
    ifmp::integer k = n / d;
    cross += cross_sum (n - k * d); // n % d
    n = k;
  }
  return cross;
}

int main()
{
  const int range = 100;

  ifmp::integer max_cross_sum = 0;
  unsigned int best_a = 0;
  unsigned int best_b = 0;
  for (int a=1; a<range; ++a) {
    ifmp::integer power = 1; // a^0
    for (int b=1; b<range; ++b) {
      // update to a^b
      power *= a;
      // count sum of digits
      ifmp::integer s = cross_sum_big (power);
      if (s >= max_cross_sum) {
        max_cross_sum = s;
        best_a = a;
        best_b = b;
      }
    }
  }

  // output
  std::cout << "Best a = " << best_a << "\n";
  std::cout << "Best b = " << best_b << "\n";
}
```

Solution to Exercise 89.

a) The idea is simple: we need to keep track of the parity (even or odd) of the number of zeros and ones processed so far. There are four possible parity combinations, and we have one state for each of them (ee, oe, eo, oo, where the first letter is for the parity of the zeros, the second for the parity of ones). When we are in state eo, for example, and we process the symbol 1, we move to state ee, because the parity of the number of ones changes. We accept if and only if we are in state ee which is also the starting state.

b) While we are processing w, we keep track of the value $w' \bmod 5$, where w' is the prefix of w processed so far. Initially, w' is empty and $w' \bmod = 0$. When we process the next symbol, it may be 0 and we have $w' := 2w'$, or it is 1 in which case we get $w' := 2w' + 1$. In all cases, we can say how $w' \bmod 5$ changes; here is

the table.

$w' \bmod 5$	$2w' \bmod 5$	$(2w'+1) \bmod 5$
0	0	1
1	2	3
2	4	0
3	1	2
4	3	4

We therefore need states $0, 1, 2, 3, 4$ corresponding to $w' \bmod 5$, connected by transitions as given in the above table, and we accept if and only if we are in state 0 in the end.

c) This language cannot be the language of any DFA. The intuitive reason is that an automaton can only "count up to some finite number", because it has only finitely many states. But in order to be able to decide whether a word contains more zeros than ones, it would have to count zeros and ones in arbitrarily long words.

But we van also prove formally that there cannot be a DFA for this language L. Assume for a contradiction that there is a DFA whose language is L, and assume that it has n states. The DFA surely accepts the word

$$\underbrace{0....0}_{n+1}\underbrace{1....1}_{n},$$

since it has more zeros than ones. Let s_i be the state the DFA is in after processing the i-th one, $i = 1, \ldots, n$. s_0 is the state before processing the first one. In total, these are $n+1$ states, and since there are only n distinct states, at least one state must occur twice in the sequence $s_0, s_1, \ldots, s_n$. So let $0 \leq j < k \leq n$ be indices such that $s_j = s_k$. This means that if the automaton is in state s_j and then processes $k-j$ ones, it gets back to state $s_j = s_k$: a loop. But then it would also get back to state j after $2(k-j)$ ones. Consequently, the DFA also accepts the word

$$\underbrace{0....0}_{n+1}\underbrace{1....1}_{n+k-j}$$

resulting from w by inserting $k-j$ additional ones after the k-th one. But this is a contradiction, since the latter word does not contain more zeros than ones.

This proof is actually an incarnation of the *pumping lemma*, a tool to prove that a language cannot be the language of a DFA. The class of languages L for which a DFA with language L exists is well-understood; it is called the class of *regular* languages.

d) Here we use the finite counting ability of DFA: we keep track of the longest run of ones that ends in the current symbol. Runs longer than two don't need to be counted, since we already know then that we will reject the word.

Solution to Exercise 90.

Fuzzy Products. Let's go step by step. Mr. Product says "I don't know the numbers a and b." This means that there must be more than one set $\{a, b\}$ such that $p = a \cdot b$. Let us call such a number p a *fuzzy product*. For example, $p = 12$ is a fuzzy product, because $12 = 2 \cdot 6 = 3 \cdot 4$. The number $p = 6$ is not a fuzzy product, since the only possible set is $\{2, 3\}$ (recall that $\{1, 6\}$ is not an option due to $1 < a, b$.

So, after the first statement of Mr. Product, we and Mr. Sum know that p must be fuzzy product.

Fuzzy Sums. Mr. Sum then says: "I knew that you don't know them." We can conclude that s must be a number, such that, for *every* way of writing it as a sum of two numbers a and b, the product $a \cdot b$ is a fuzzy product. Otherwise Mr. Sum would not have been able to be sure about his statement. Let us call such a number s a *fuzzy sum*. For example, $s = 11$ is a fuzzy sum, because

$$\begin{array}{rclcrclcl}
11 & = & 2+9 & \text{and} & 2\cdot 9 & = & 18 & = & 3\cdot 6 \\
11 & = & 3+8 & \text{and} & 3\cdot 8 & = & 24 & = & 4\cdot 6 \\
11 & = & 4+7 & \text{and} & 4\cdot 7 & = & 28 & = & 2\cdot 14 \\
11 & = & 5+6 & \text{and} & 5\cdot 6 & = & 30 & = & 3\cdot 10
\end{array}$$

So, after the first statement of Mr. Sum, we and Mr. Product know that s must be fuzzy sum.

Rare Products. In his mind, Mr. Product quickly computes all fuzzy sums between 4 and 198 (these are the candidates), and finds that there are only 10 of them: They are $11, 17, 23, 27, 29, 35, 37, 41, 47, 53$, and $s = a + b$ must be one of them. Now, Mr. Product, knowing p, is trying to solve— for each of the 10 candidates s—the equation system

$$\begin{array}{rcl}
s & = & a + b \\
p & = & ab
\end{array}$$

in order to find the two unknowns a and b. Since his next statement is "Ah... but now I know them.", we can conclude that the system has a valid solution for *exactly one* of the 10 fuzzy sums s. Therefore, p must have another special property: exactly one of the sets $\{a, b\}$ with $a \cdot b = p$ has a fuzzy sum $a + b$. We call such a p a *rare product*.

So, after the second statement of Mr. Product, we and Mr. Sum know that p must be a rare product.

Rare Sums. In his mind, Mr. Sum, knowing s, quickly computes all sets $\{a, b\}$ such that $s = a + b$ and $a \cdot b$ is a rare product. After this, he says "Then I know them too, now." We conclude that *exactly one* set $\{a, b\}$ such that $s = a + b$ has a rare product $a \cdot b$. We call such an s a *rare sum*. Mr. Sum now knows $\{a, b\}$, but we still have some work to do.

After the second statement of Mr. Sum, we know that s is both a fuzzy and a rare sum. And in order for the puzzle to be solvable for us, we know that exactly one s in the range $3 < s < 199$ must have this property.

Putting it Together. Now we write a program to search for the supposedly unique number $3 < s < 199$ that is both a fuzzy and a rare sum. And from the rare sum property of s, we can then uniquely deduce $\{a, b\}$. It turns out that

$$\begin{aligned} s &= 17, \\ p &= 52, \\ \{a, b\} &= \{4, 13\}. \end{aligned}$$

is indeed the unique solution.

```
// Prog: mr_prod_sum.cpp
// solves the puzzle of Mr. Sum and Mr. Product
#include <iostream>
#include <cassert>

// PRE:  1 < a < 100, s > 3
// POST: returns true iff 1 < a <= b := s-a < 100
bool legal_sum (int s, int a, int& b) {
  assert (s > 3);
  assert ((1 < a) && (a < 100));
  b = s - a;
  return ((a <= b) && (b < 100));
}

// PRE:  1 < a < 100, p > 3
// POST: returns true iff a divides p, and 1 < a <= b := p/a < 100
bool legal_product (int p, int a, int& b) {
  assert (p > 3);
  assert ((1 < a) && (a < 100));
  b = p / a;
  return ((p % a == 0) && (a <= b) && (b < 100));
}

// PRE:  p > 3
// POST: returns true iff p can be written in more
//       than one way as the product of two numbers
//       between 2 and 99
bool fuzzy_product(int p) {
  assert (p > 3);
  int count = 0;
  int b;
  for (int a = 2; a < 100; ++a) {
    if (legal_product (p, a, b))
      ++count;
  }
  return count > 1;
}

// PRE:  s > 3
// POST: Returns true iff for all a,b such that
//       s = a+b, the product a*b is a fuzzy product.
bool fuzzy_sum(int s) {
  assert (s > 3);
  int b;
```

```
  for (int a = 2; a < 100; ++a) {
    if (legal_sum (s, a, b) && !fuzzy_product( a * b )) {
      return false;
    }
  }
  return true;
}

// PRE:  p > 3
// POST: returns true iff exactly one of the sets {a,b}
//       with p = a*b has a fuzzy sum a+b.
bool rare_product(int p) {
  assert (p > 3);
  int count = 0;
  int b;
  for (int a = 2; a < 100; ++a) {
    if (legal_product (p, a, b) && fuzzy_sum (a + b))
      ++count;
  }
  return count == 1;
}


// PRE:  s > 3
// POST: returns true iff exactly one of the pairs (a,b)
//       with a+b=s has a rare product a*b.
//       If the return value is true, then the corresponding
//       pair {a,b} can be retrieved from the parameters.
bool rare_sum(int s, int& a, int& b) {
  assert (s > 3);
  int count = 0;
  int bb;
  for (int aa = 2; aa < 100; ++aa) {
    if (legal_sum (s, aa, bb) && rare_product( aa * bb )) {
      a = aa;
      b = bb;
      ++count;
    }
  }
  return count == 1;
}

int main() {
  std::cout << "Compute a and b in the range [2..99]...\n";
  int a;
  int b;

  for (int s = 4; s < 199; ++s) {
    //std::cout << "Checking sum: " << i << std::endl;
    if (fuzzy_sum(s) && rare_sum(s, a, b)) {
      std::cout << "Solution:\n";
      std::cout << "s= " << s << "\n";
      std::cout << "p= " << a*b << "\n";
      std::cout << "a= " << a << "\n";
      std::cout << "b= " << b << "\n";
    }
  }

  return 0;
}
```

B.7 Arrays and pointers

Solution to Exercise 91. We recycle the program eratosthenes2.cpp (Program 2.24); instead of maintaining the information whether a number has been crossed out, we maintain the information about the number of different prime divisors.

```
// Program: k_composite.cpp
// Calculate k-composite numbers in {2,...,n-1} using
// a variant of Eratosthenes' sieve.

#include <iostream>
#include <vector>

int main()
{
  // input of k
  std::cout << "Compute k-composite numbers for k =? ";
  unsigned int k;
  std::cin >> k;

  // input of n
  std::cout << "Compute " << k
            << "-composite numbers in {2,...,n-1} for n =? ";
  unsigned int n;
  std::cin >> n;

  // definition and initialization: provides us with
  // unsigned integers composition[0],..., composition[n-1]
  std::vector<unsigned int> composition(n, 0);

  // computation and output
  std::cout << k << "-composite numbers in {2,..,"
            << n-1 << "}:\n";
  for (unsigned int i = 2; i < n; ++i) {
    if (composition[i] == 0) {
      // i is prime: add 1 to composition number of all
      // multiples (including i)
      for (unsigned int m = i; m < n; m += i)
        ++composition[m];
    }
    // now the composition number of i is up-to-date
    if (composition[i] == k)
      std:: cout << i << " ";
  }
  std::cout << "\n";

  return 0;
}
```

The following are the 8 different 7-composite numbers smaller than $1,000,000$: 510510, 570570, 690690, 746130, 870870, 881790, 903210, 930930.

Solution to Exercise 92. One trick here is that computations with indices modulo 3 save us the signs $(-1)^{i+j}$.

```
// Prog: inverse_matrix.cpp
// read in a 3x3 matrix A, compute ins inverse A^{-1},
// and output it along with A x A^{-1} as a crosscheck

```

```
#include<iostream>

int main()
{
  // read in A (as a sequence of 9 numbers)
  double a[3][3];
  for (int i=0; i<3; ++i)
    for (int j=0; j<3; ++j)
      std::cin >> a[i][j];

  // compute determinant of A via Sarrus' rule
  double det = 0;
  for (int i=0; i<3; ++i)
    det += a[0][i] * a[1][(i+1)%3] * a[2][(i+2)%3]
      - a[2][i] * a[1][(i+1)%3] * a[0][(i+2)%3];

  // compute (and output) entries of A^{-1} through Cramer's rule
  std::cout << "A^-1 = \n";
  double a_inv[3][3];
  for (int i=0; i<3; ++i) {
    for (int j=0; j<3; ++j)
      std::cout
        << (
            a_inv[i][j] = ( (a[(j+1)%3][(i+1)%3]*a[(j+2)%3][(i+2)%3]) -
                            (a[(j+1)%3][(i+2)%3]*a[(j+2)%3][(i+1)%3]) ) / det)
        << " ";
    std::cout << "\n";
  }

  // crosscheck
  std::cout << "A * A^-1 = \n";
  for (int i=0; i<3; ++i) {
    for (int j=0; j<3; ++j)
      // output (A x A^{-1})_{ij}
      std::cout <<
        a[i][0]*a_inv[0][j]+
        a[i][1]*a_inv[1][j]+
        a[i][2]*a_inv[2][j] << " ";
    std::cout << "\n";
  }

  return 0;

}
```

Solution to Exercise 93.

```
// Program: read_array.cpp
// read a sequence of n numbers into an array
#include <iostream>
#include <vector>

int main()
{
  // input of n
  int n;
  std::cin >> n;

  // make a vector of length n
  std::vector<int> a(n);

  // read into the array
  for (int i=0; i<n; ++i) std::cin >> a[i];
```

```

  // output what we have
  for (int i=0; i<n; ++i) std::cout << a[i] << " ";
  std::cout << "\n";

  return 0;

}
```

Solution to Exercise 94.

```
// Program: sort_array.cpp
// read a sequence of n numbers into an array,
// sort them, and output the sorted sequence
#include <iostream>
#include <vector>

int main()
{
  // input of n
  unsigned int n;
  std::cin >> n;

  // get vector of length n
  std::vector<int> a(n);

  // read into the array
  for (int i=0; i<n; ++i) std::cin >> a[i];

  // sort array: in round i=0,...,n-2 we find
  // the smallest element in a[i],...,a[n-1] and
  // interchange it with a[i]
  for (int i=0; i<n-1; ++i) {
    // find minimum in a[i],...,a[n-1]
    int i_min = n-1; // index of minimum
    for (int j=i; j<n-1; ++j)
      if (a[j] < a[i_min]) i_min = j;
    // interchange a[i] with a[i_min]
    const int h = a[i]; a[i] = a[i_min]; a[i_min] = h;
  }

  // output sorted sequence
  for (int i=0; i<n; ++i) std::cout << a[i] << " ";
  std::cout << "\n";

  return 0;

}
```

Solution to Exercise 95. Let $s = a \dots ab$ (i.e. $m-1$ a's followed by one b); let $t = a \dots a$ (i.e. n a's). Then the algorithm must always go through all m characters of any window in order to find the mismatch with b at the last position. Since it in total processes $n - m + 1$ windows $\{1, \dots, m\}$ up to $\{n - m + 1, \dots, n\}$, the number of comparisons is $m(n - m + 1)$.

Solution to Exercise 96. The trick is to use characters (which have integral values) directly as array indices.

```
// Program: frequencies.cpp
// output frequencies of the letters in an input text

#include<iostream>

int main ()
{
  // array for number of occurences of every ASCII character
  int frequency[128];
  for (int i=0; i<128; ++i) frequency[i] = 0;

  // now scan the text
  char c;                    // next character
  unsigned int total = 0;    // text length
  while (std::cin >> c) {
    ++total;
    if ('a' <= c && c <= 'z' || 'A' <= c && c <= 'Z')
      ++frequency[c];        // handle letters only, assuming ASCII code
  }

  // output
  unsigned int letters = 0; // number of letters
  std::cout << "Frequencies: \n";
  for (char c = 'a'; c <= 'z'; ++c) {
    const int f = frequency[c] + frequency[c-32]; // lower + upper case c
    letters += f;
    std::cout << c << ":     " << f << " of " << total << "\n";
  }
  std::cout << "Other: " << total-letters << " of " << total << "\n";

  return 0;
}
```

Solution to Exercise 97. Here is the faster program.

```
#include<iostream>
#include<cassert>
#include<vector>

int main()
{
  // read floor dimensions
  int n; std::cin >> n; // number of rows
  int m; std::cin >> m; // number of columns

 // define a twodimensional vector of dimensions
  // (n+2) x (m+2) to hold the floor plus extra walls around
  std::vector<std::vector<int> > floor (n+2, std::vector<int>(m+2));

  // we need another two vectors for storing row and column
  // indices of already labeled cells;
  std::vector<int> labeled_r(n*m);
  std::vector<int> labeled_c(n*m);

  // in order to search for new cells to be labeled, we
  // always start from the first labeled cell whose neighbors
  // have not been looked at yet;
  int next_l = 0; // index of this cell

  // whenever we label a cell, we append it to the list of
  // labeled cells; that way, the cells are ordered by label
  // in the list
```

```
28    int last_l = 0; // one plus index of last cell in this list
29
30    // target coordinates, set upon reading 'T'
31    int tr = 0;
32    int tc = 0;
33
34    // assign initial floor values from input:
35    // source:     'S'  ->     0 (source reached in 0 steps)
36    // target:     'T'  ->    -1 (number of steps still unknown)
37    // wall:       'X'  ->    -2
38    // empty cell: '-'  ->    -1 (number of steps still unknown)
39    for (int r=1; r<n+1; ++r)
40      for (int c=1; c<m+1; ++c) {
41        char entry = '-';
42        std::cin >> entry;
43        if      (entry == 'S') {
44          floor[r][c] = 0;
45          labeled_r[last_l] = r;
46          labeled_c[last_l] = c;
47          ++last_l;
48        }
49        else if (entry == 'T') floor[tr = r][tc = c] = -1;
50        else if (entry == 'X') floor[r][c] = -2;
51        else if (entry == '-') floor[r][c] = -1;
52      }
53
54    // add surrounding walls
55    for (int r=0; r<n+2; ++r)
56      floor[r][0] = floor[r][m+1] = -2;
57    for (int c=0; c<m+2; ++c)
58      floor[0][c] = floor[n+1][c] = -2;
59
60    // main loop: process next labeled cell until done
61    while (next_l != last_l) {
62      const int r = labeled_r[next_l];
63      const int c = labeled_c[next_l];
64      const int i = floor[r][c];
65      assert (i >= 0);
66      // label the unlabeled neighbors by i+1
67      for (int rr = r-1; rr <= r+1; ++rr)
68        for (int cc = c-1; cc <= c+1; ++cc)
69          if ( (rr == r || cc == c) && floor[rr][cc] == -1) {
70            // we have a neighbor, and it's not labeled yet
71            floor[rr][cc] = i+1;
72            labeled_r[last_l] = rr;
73            labeled_c[last_l] = cc;
74            ++last_l;
75          }
76      ++next_l;
77    }
78
79    // mark shortest path from source to target (if there is one)
80    int r = tr; int c = tc; // start from target
81    while (floor[r][c] > 0) {
82      const int d = floor[r][c] - 1; // distance one less
83      floor[r][c] = -3; // mark cell as being on shortest path
84      // go to some neighbor with distance d
85      if      (floor[r-1][c] == d) --r;
86      else if (floor[r+1][c] == d) ++r;
87      else if (floor[r][c-1] == d) --c;
88      else                         ++c; // (floor[r][c+1] == d)
89    }
90
91    // print floor with shortest path
92    for (int r=1; r<n+1; ++r) {
```

```
    for (int c=1; c<m+1; ++c)
      if      (floor[r][c] == 0)   std::cout << 'S';
      else if (r == tr && c == tc) std::cout << 'T';
      else if (floor[r][c] == -3)  std::cout << 'o';
      else if (floor[r][c] == -2)  std::cout << 'X';
      else                         std::cout << '-';
    std::cout << "\n";
  }

  return 0;
}
```

Solution to Exercise 98. The idea is to try all possible combinations of a, b, c within prespecified ranges. In order to make this fast, some tricks are needed, though. First of all, we use Eratosthenes' Sieve in order to precompute the information whether a given number that may arise as $|an^2 + bn + c|$ is prime. Then we use another array to mark the primes that we have seen in a run of primes; the trick here is to delete the markers again in an efficient way. The following program discovers a quadratic polynomial of Euler quality 45, namely

$$36n^2 - 810n + 2753.$$

```
// Program: euler_prime.cpp
// Finds a,b,c within specified bounds such that
// the formula |an^2 + bn + c| produces the largest
// number of distinct consecutive primes, starting
// with n = 0

#include <iostream>
#include <cassert>

int main ()
{
  // by multiplying with -1, if necessary, we may assume c >= 0
  const int arange = 100;      // a in {-arange+1,...,arange-1}
  const int brange = 1000;    // b in {-brange+1,...,brange-1}
  const int crange = 10000;   // c in {2,...,crange-1}

  // first, compute all primes in the set {an^2 + bn + c} where a, b, c
  // run through their ranges and n is smaller than 100 (we're not
  // searching for longer runs of primes here)
  const int max_elem = arange * 10000 + brange * 100 + crange;

  // run Eratosthenes' sieve
  bool prime[max_elem];
  for (int i = 2; i < max_elem; ++i)
    prime[i] = true;
  for (int i = 2; i < max_elem; ++i)
    if (prime[i]) {
      // cross out all proper multiples of i
      for (int m = 2*i; m < max_elem; m += i)
        prime[m] = false;
    }

  // now search for the best a, b, c
  int best_run = 0;
  int best_a = arange;
  int best_b = brange;
```

```
  int best_c = crange;

  // array to keep track of primes we have already seen in a run
  bool seen_before[max_elem];
  for (int i = 2; i < max_elem; ++i)
    seen_before[i] = false;

  // loop over all candidate triples (a,b,c)
  for (int c = 2; c < crange; ++c) {
    if (!prime[c]) continue; // not even a prime for n = 0
    for (int a = -arange+1; a < arange; ++a)
      for (int b = -brange+1; b < brange; ++b) {
        // evaluate elem = an^2 + bn + c for n=0,1,...
        int n = 0;
        int elem = c;
        for (; n < 100; ++n) {
          const int abs_elem = elem < 0 ?  -elem: elem;
          if (abs_elem < 2 || !prime[abs_elem]) break; // not a prime
          if (seen_before[abs_elem]) break;            // repeated prime
          seen_before[abs_elem] = true;                // new prime
          // update element
          elem += a * (2*n + 1) + b;
        }
        // now we have seen a run of n primes (for 0,...,n-1)
        if (n > best_run) {
          best_run = n;
          best_a = a;
          best_b = b;
          best_c = c;
        }
        // remove the "seen_before" markers for the next run
        elem = c;
        for (int k=0; k<n; ++k) {
          const int abs_elem = elem < 0 ?  -elem: elem;
          seen_before[abs_elem] = false;
          elem += a * (2*k + 1) + b;
        }
      }
  }

  std::cout << "Best a        = " << best_a << ".\n";
  std::cout << "Best b        = " << best_b << ".\n";
  std::cout << "Best c        = " << best_c << ".\n";
  std::cout << "Euler quality = " << best_run << ".\n";

  return 0;
}
```

Solution to Exercise 99.

```
// Prog: xbm.cpp
// includes an xbm file and outputs the xbm-file that corresponds
// to the image rotated by 90 degrees

#include<iostream>
#include<cassert>
#include "original.xbm"
#define width original_width
#define height original_height
#define bits original_bits

int main()
{
```

```
14    // array for mapping from {0,...,15} to {'0',...,'f'}
15    char hex[] = {'0', '1', '2', '3', '4', '5', '6', '7',
16                  '8', '9', 'a', 'b', 'c', 'd', 'e', 'f'};
17
18    // output header for rotated image
19    std::cout << "#define rotated_width " << height << "\n";
20    std::cout << "#define rotated_height " << width << "\n";
21    std::cout << "static unsigned char rotated_bits[] = {\n";
22
23    // go through the pixels columnwise (from right to left),
24    // and within each column, proceed from top to bottom
25    const unsigned int columns = (width+7)/8; // bytes in row (original)
26    const unsigned int rows = (height+7)/8;   // bytes in column (rotated)
27    bool comma = false;                       // not before first byte
28    for (int c=columns-1; c>=0; --c) {
29      // go through columns from right to left
30      for (int d=128; d>0; d/=2) {
31        // for d = 2^j, we are in column c+j
32        for (int r=0; r<rows; ++r) {
33          // go through rows from top to bottom and build
34          // up one byte from any 8 pixels (least significant
35          // digit comes from pixel in first of these rows)
36          int byte = 0;
37          for (int i=7; i>=0; --i) {
38            // we are in row 8r+i
39            int pixel;
40            if (8*r+i >= height)
41              // non-existing row, fill up with zeros
42              pixel = 0;
43            else
44              // get pixel in row 8r+i and column c+j
45              pixel = (unsigned char)(bits[(8*r+i)*columns+c])/d%2;
46            byte = 2*byte+pixel;
47          }
48          if (comma) std::cout << ", "; comma = true;
49          std::cout << "0x" << hex[byte/16] << hex[byte%16];
50        }
51      }
52      std::cout <<"\n";
53    }
54    std::cout << "}\n;";
55
56    return 0;
57
58  }
```

B.8 Pointers, Algorithms, Iterators and Containers

Solution to Exercise 100.

```
#include<iostream>

// PRE:  i_ptr and j_ptr point to existing objects
// POST: the values of these two objects are swapped
void swap (int* i_ptr, int* j_ptr)
{
  const int h = *i_ptr;
  *i_ptr = *j_ptr;
  *j_ptr = h;
}

int main() {
  // input
  std::cout << "i =? ";
  int i; std::cin >> i;

  std::cout << "j =? ";
  int j; std::cin >> j;

  // function call
  swap(&i, &j);

  // output
  std::cout << "Values after swapping: i = " << i
            << ", j = " << j << ".\n";

  return 0;
}
```

Solution to Exercise 101.

```
// Prog: unique.cpp
// implemements and tests a function that checks whether every element
// in a sorted sequence is unique

#include<iostream>
#include<cassert>
#include<vector>

typedef std::vector<int>::const_iterator Cit;

// PRE:  [first, last) is a valid range and describes a sequence
//       of elements that are sorted in nondecreasing order
// POST: the return value is true if and only if no element
//       occurs twice in the sequence
bool unique(Cit first, Cit last)
{
  // handle the empty sequence
  if (first == last) return true;

  // simultaneously advance two iterators: next and first
  // first is pointing to the first unprocessed element, and
  // next to its successor, we have to check if *first == *next
  for (Cit next = first; ++next != last; ++first) {
    assert(*first <= *next);
    if (*first == *next)
```

```
      return false;
  }
  return true;
}

int main()
{
  int aTmp[5] = {1,2,3,4,4};
  int bTmp[5] = {5,6,7,8,9};

  std::vector<int> a(aTmp, aTmp + 5);
  std::vector<int> b(bTmp, bTmp + 5);
  std::cout << unique (a.begin(), a.end() ) << "\n"
            << unique (b.begin(), b.end() ) << "\n";

  return 0;
}
```

Solution to Exercise 102. We simply take the sorting loop out of `sort_array.cpp` and put it into a function. At the same time, we move from iteration by index to iteration by pointers. A less elegant but also valid solution is to keep the original code with iteration by index, after computing `n` as `last-first`.

```
// Program: sort_array2.cpp
// read a sequence of n numbers into an array,
// sort them using a function, and output the
// sorted sequence
#include <iostream>
#include <vector>

typedef std::vector<int>::iterator It;

// PRE: [first, last) is a valid range
// POST: the elements *p, p in [first, last) are in ascending order
void sort (It first, It last)
{
  // sort array: in round p=first,...,last-1 we find
  // the smallest element in the range described by
  // [p, last) and interchange it with *p
  for (It p = first; p != last; ++p) {
    // find minimum in nonempty range described by [p, last)
    It p_min = p; // iterator pointing to current minimum
    It q = p;     // iterator pointing to current element
    while (++q != last)
      if (*q < *p_min) p_min = q;
    // interchange *p with *p_min
    const int h = *p; *p = *p_min; *p_min = h;
  }
}

int main()
{
  // input of n
  int n;
  std::cin >> n;

  // construct vector
  std::vector<int> a(n);

  // read into the vector
  for (int i=0; i<n; ++i) std::cin >> a[i];
```

```
39
40    // sort
41    sort (a.begin(), a.end());
42
43    // output sorted sequence
44    for (int i=0; i<n; ++i) std::cout << a[i] << " ";
45    std::cout << "\n";
46
47    return 0;
48  }
```

Solution to Exercise 103. The postcondition is

```
// POST: The range [b,e) is copied in reverse order into the
// range [o, o+(e-b))
```

In b), the first call is invalid, since `[a+5,a+10)` is not a valid range. The second call is ok, but the third one is again invalid, since the ranges `[a,a+3)` and `[a+2,a+5)` overlap (in one element). For c), we observe that the elements described by the source range `[b,e)` are not modified. Thus, the pointers `b` and `e` should both have type `const int*`.

Solution to Exercise 104. The 5 values of `b[0]` up to `b[4]` after the function call are $(1, 4, 6, 4, 1)$. Incidentally, this is the fifth row of *Pascal's triangle*

$$
\begin{array}{ccccccccc}
 & & & & 1 & & & & \\
 & & & 1 & & 1 & & & \\
 & & 1 & & 2 & & 1 & & \\
 & 1 & & 3 & & 3 & & 1 & \\
1 & & 4 & & 6 & & 4 & & 1 \\
1 & 5 & & 10 & & 10 & & 5 & 1 \\
 & & & & \vdots & & & &
\end{array}
$$

where every entry is the sum of the two entries directly above it. And it is not hard to prove that this generalizes. If `e - b` has value n, the n values `b[0]` up to `b[n-1]` form the n-th row of Pascal's triangle. We prove this by induction on n, where the case $n = 1$ is easy (`b[0]==1`). Now let $n > 1$ and assume that the statement holds for $n - 1$. In going from $n - 1$ to n, there is one more outer loop iteration (with $i = n$) that first sets `b[n-1]` to 1 (as desired). The values `b[n-2]`, `b[n-3]`, `b[1]` that by hypothesis form the $(n - 1)$-st row of Pascal's triangle (without the first element) are updated as follows: to every such element, we add the one directly to the left of it. By definition of Pascal's triangle, this yields the correct elements in the n-th row.

We would like to remark that the k-th entry in the n-th row of Pascal's triangle is the binomial coefficient $\binom{n-1}{k-1}$; see also Exercise 114.

Solution to Exercise 105.

a) The program outputs `5 4 1 6 0`. The first array element is always the first one to be output by `std::cout << *p`, because `p` is initialized with `a`, and due to array-to-pointer conversion, `p` then points to the first element of `a`. The assignment

`p = a + *p` changes `p` into a pointer to the element of index `*p` in the array (see the paragraph on adding an integer to a pointer). Since `*p` is initially 5, we get the element of index 5 next, which is 4. The next element is the one of index 4 (1), followed by the ones of index 1 (6) and index 6 (0). At this point, `p` points again to the first element of the array, so the condition `p!=a` fails, and the loop terminates.

b) The general structure is this: Let i_t be the index of the array element pointed to by `p` after `t` iterations. We have $i_0 = 0$, and $i_t = a[i_{t-1}]$ for $t > 0$. The program terminates as soon as $i_t = 0$ again for some $t > 0$. We must prove that this always happens.

Assume for contradiction that 0 does not appear for a second time, so that we have an infinite loop. Since there are only n possible index values, some index distinct from 0 must appear twice in the infinite sequence $i_0, i_1, \ldots$. Let k be some value such that $i_k \neq 0$ already appears among $i_0, \ldots, i_{k-1}$, and let $0 < \ell < k$ be such that $i_\ell = i_k$. By definition of the sequence, it follows that

$$a[i_{k-1}] = i_k = i_\ell = a[i_{\ell-1}],$$

where $k-1 \neq \ell-1$. But this gives the desired contradiction, since the array `a` was initialized with a sequence of pairwise distinct numbers.

Solution to Exercise 106. In a), we use

```
int* p = a+i; // address of element of index i
```

Note that

```
int* p = &a[i]; // address of element of index i
```

also works, *unless* $i = n$, since n is an out-of-bound index. A past-the-end pointer can therefore only be obtained from the first variant.

In b), we use pointer subtraction:

```
int i = p-a;  // distance between *p and a[0] in the array
```

Solution to Exercise 107.

Problem 1:

`c` is a const pointer, but later, array elements are changed through `c`. This won't compile. Fix: remove the `const`. Now the program at least compiles...

Problem 2:

In the first loop, the range that `p` points to is too large by one address. Fix: `p <= a+7` should be `p < a+7`.

Problem 3

A similar problem also appears in the second loop. Inside the loop `a` and `c` are accessed by one too many values for index `i`. Fix: `i<=7` should be `i<7`.

Problem 4

The cross-check in the second loop goes wrong, because c doesn't point to the beginning of array b any more. Fix: compare with b[i] insteac of c[i].

If all these fixed are applied, there are still two logical consts missing for const-correctness. The loop pointer p should have underlying type const int to reflect the fact that the loop does not change the array a that it loops over. Also, the underlying type of the array a should be const int since the array elements are accessed in a read-only fashion.

Here is the resulting correct program.

```
#include<iostream>

int main()
{
  const int a[7] = {0, 6, 5, 3, 2, 4, 1}; // static array
  int b[7];
  int* c = b;

  // copy a into b using pointers
  for (const int* p = a; p < a+7; ++p)
    *c++ = *p;

  // cross-check with random access
  for (int i = 0; i < 7; ++i)
    if (a[i] != b[i])
      std::cout << "Oops, copy error...\n";

  return 0;
}
```

Solution to Exercise 108.

```
// Program: cycles.cpp
// read a sequence of n numbers into an array; if the sequence
// encodes a permutation of {0,...,n-1}, output its cycle
// decomposition

#include <iostream>
#include <vector>

int main()
{
  // input of n
  unsigned int n;
  std::cin >> n;

  // construct vector for the numbers, and a
  // second vector of booleans to keep track of which
  // numbers are present
  std::vector<unsigned int> a(n);
  std::vector<bool> present(n, false);

  // read number into the vector and remember that it was read
  for (std::vector<unsigned int>::iterator p = a.begin();p != a.end();++p) {
    std::cin >> *p;
    if (*p >= 0 && *p < n)
      present[*p] = true;
  }
```

```

  // check whether we have read all numbers in {0,...,n-1}
  bool ok = true;
  for (std::vector<bool>::iterator r = present.begin();r != present.end();++r)
    if (!*r) {
      std::cout << "input sequence does not encode a permutation.\n";
      ok = false;
      break;
    }

  if (ok) {
    // do the cycle decomposition. Here we reuse the vector present
    // and remove from it all numbers that we have already put into
    // some cycle
    std::cout << "cycle decomposition is ";
    unsigned int next = 0; // next number not yet put into a cycle
    while (next < n) {
      // output cycle starting with next; we must come back to next
      // below: assuming we would come back to some other element
      // on the cycle, that element would have two preimages under
      // pi, a contradiction
      const int first = next;
      std::cout << "( ";
      do {
                std::cout << next << " ";
                present[next] = false;
                next = a[next]; // next -> pi(next)
      } while (next != first);
      std::cout << ") ";

      // find start element of next cycle
      while (!present[next]) ++next;
    }
    std::cout << "\n";
  }

  return 0;

}
```

Solution to Exercise 109.

The solution code for problem (a) is in the following block.

```
// Program: first_occurrence.cpp
// read a sequence of n numbers into an array
// and output the index of the first occurrence of a given element
#include <iostream>
#include <vector>
#include <algorithm>

int main()
{
  // input of n
  unsigned int n;
  int target = 0;
  std::cin >> n;

  // make a vector of length n
  std::vector<int> a(n);

  // read into the array
  for (unsigned int i=0; i<n; ++i) std::cin >> a[i];
```

```

  // read into the value you want
  std::cin >> target;

  // find and output the index of the first occurrence
  std::vector<int>::iterator iter =
          find(a.begin(), a.end(), target);

  std::cout << (iter - a.begin() ) << std::endl;

  return 0;

}
```

The solution code for problem (b) is in the following block.

```
// Program: last_occurrence.cpp
// read a sequence of n numbers into an array
// and output the index of the last occurrence of a given element
#include <iostream>
#include <vector>
#include <algorithm>

int main()
{
  // input of n
  unsigned int n = 0;
  unsigned int id = 0;
  int target = 0;
  std::cin >> n;

  // make a vector of length n
  std::vector<int> a(n);

  // read into the array
  for (unsigned int i=0; i<n; ++i) std::cin >> a[i];

  // read into the value you want
  std::cin >> target;

  // find and output the index of the last occurrence
  for (std::vector<int>::const_iterator iter = a.begin(); iter != a.end(); iter++) {
        if (*iter == target)
                id = (iter - a.begin() );
  }

  std::cout << id << std::endl;

  return 0;

}
```

Solution to Exercise 110. The code is subdivided into several functions. The most important function `deducible` tries to make one Sherlock-Holmes-type deduction for given r, c and n. It is repeatedly called by the function `fill_cell` for all values of r, c, and n; if a deduction is found, `fill_cell` accordingly updates the board, a 3-dimensional array that maintains the information which numbers are candidates for which cells.

The function `deduce` tries the Sherlock-Holmes-type deductions in turn, where it distinguishes between deduction from row, column, or box in case 2.

```
// Prog: sudoku.cpp
// solves sudokus according to simple deduction heuristic (may fail)
#include<cassert>
#include<iostream>

// Here is the algorithm: given a partial filling of the 81 cells,
// we try to find a row r, a column c, and a number n such that n is
// the unique candidate to be filled into the empty cell (r,c) in row
// r and column c. There are two situations in which this is easy:
//  1. all numbers distinct from n already appear in the row, column,
//     or 3x3 box containing the cell (r,c)
//  2. we already know that n cannot appear in the other cells of the
//     row, column, or box containing (r,c)
// To check this, we maintain for every triple (r,c,n) the information
// whether n is still a candidate for cell (r,c). Whenever a cell is filled,
// we remove the filled number from the candidate list of all other cells
// in the same row, column, or box.

// the grid: a 3x3x3 array of boolean values, one for each triple (r,c,n)
//   grid[r][c][0] == true iff cell (r,c) not filled yet
//   grid[r][c][n] == true for n>0 iff n is still a candidate for cell (r,c)
bool grid[9][9][10];

// PRE:  cell (r,c) is empty, and n>0 is a candidate for (r,c)
// POST: puts n into (r,c); removes n as candidate from all other cells
//       of the row, the column, and the box of (r,c); removes all
//       numbers distinct from n as candidates from (r,c)
void update_grid (const unsigned int r, const unsigned int c,
                  const unsigned int n)
{
  // assert precondition
  assert (grid[r][c][0] && n > 0 && grid[r][c][n]);
  // go through row of (r,c)
  for (unsigned int j=0; j<9; ++j)
    grid[r][j][n] = false;
  // go through column of (r,c)
  for (unsigned int i=0; i<9; ++i)
    grid[i][c][n] = false;
  // go through box of (r,c)
  const unsigned int lr = r-r%3; // (lr, lc) is the lower left cell of
  const unsigned int lc = c-c%3; // this box
  for (unsigned int i=lr; i<lr+3; ++i)
    for (unsigned int j=lc; j<lc+3; ++j)
      grid[i][j][n] = false;
  // go through numbers
  for (unsigned int m=0; m<10; ++m)
    grid[r][c][m] = false;
  // fill cell
  grid[r][c][n] = true;
}

// POST: returns n > 0 if n is the unique candidate for cell (r,c) and
//       0 otherwise
int unique_number (const unsigned int r, const unsigned int c)
{
  int n = 0;
  bool found = false;
  for (unsigned int m = 1; m<10; ++m)
    if (grid[r][c][m]) {
      if (found) return 0; // we've already seen a candidate
      n = m;               // this is the first candidate
      found = true;
    }
  return n;
```

```
}

// PRE:  cell (r,c) is empty, and n>0 is a candidate for (r,c)
// POST: returns true if and only if n is not a candidate for any
//       cell distinct from (r,c) within the same row
bool deducible_from_row (const unsigned int r, const unsigned int c,
                         const unsigned int n)
{
  // assert preconditions
  assert (grid[r][c][0] && n > 0 && grid[r][c][n]);
  // go through the columns
  for (int j=0; j<9; ++j)
    if (j != c && grid[r][j][n]) return false; // candidate somewhere else
  return true;
}

// PRE:  cell (r,c) is empty, and n>0 is a candidate for (r,c)
// POST: returns true if and only if n is not a candidate for any
//       cell distinct from (r,c) within the same column
bool deducible_from_column (const unsigned int r, const unsigned int c,
                            const unsigned int n)
{
  // assert preconditions
  assert (grid[r][c][0] && n > 0 && grid[r][c][n]);
  // go through the rows
  for (int i=0; i<9; ++i)
    if (i != r && grid[i][c][n]) return false; // candidate somewhere else
  return true;
}

// PRE:  cell (r,c) is empty, and n>0 is a candidate for (r,c)
// POST: returns true if and only if n is not a candidate for any
//       cell distinct from (r,c) within the same box
bool deducible_from_box (const unsigned int r, const unsigned int c,
                         const unsigned int n)
{
  // assert preconditions
  assert (grid[r][c][0] && n > 0 && grid[r][c][n]);
  const unsigned int lr = r-r%3; // (lr, lc) is the lower left cell of
  const unsigned int lc = c-c%3; // this box
  // go through the box
  for (int i=lr; i<lr+3; ++i)
    for (int j=lc; j<lc+3; ++j)
      if ( (i != r || j != c) && grid[i][j][n])
        return false; // candidate somewhere else
  return true;
}

// POST: returns true iff (r,c) is empty, n>0 is a candidate for (r,c), and
//         - n is the unique candidate for the cell (r,c), or
//         - n is deducible as the unique candidate from the row,
//           column, or box of (r,c)
bool deducible (const unsigned int r, const unsigned int c,
                const unsigned int n)
{
  if (grid[r][c][0] && n > 0 && grid[r][c][n])
    return
      (n == unique_number (r,c))      ||  // unique candidate for (r,c) ?
      deducible_from_row (r, c, n)    ||
      deducible_from_column (r, c, n) ||
      deducible_from_box (r, c, n);
  return false;
}

// POST: returns true if and only if there is an empty cell (r,c) and a
```

```
//        number n>0 such that n can be deduced to be the number to be
//        put into (r,c); if the return value is true, the grid is updated
//        by filling an empty cell with an accordingly deduced number
bool fill_cell()
{
  for (unsigned int r=0; r<9; ++r)
    for (unsigned int c=0; c<9; ++c)
      for (unsigned int n=1; n<10; ++n)
        if (deducible (r, c, n)) {
          update_grid (r, c, n);
          return true;
        }
  return false;
}

int main()
{
  // set up empty grid
  for (int r=0; r<9; ++r)
    for (int c=0; c<9; ++c) {
      for (int n=0; n<10; ++n)
        grid[r][c][n] = true;
    }

  // input: 9 x 9 numbers in {0,..,9} rowwise (0 means no number)
  unsigned int filled_cells = 0; // total number of filled cells
  for (int r=0; r<9; ++r)
    for (int c=0; c<9; ++c) {
      int n;
      std::cin >> n;
      if (n > 0) {
        update_grid (r, c, n);
        ++filled_cells;
      }
    }

  // main loop
  while (fill_cell()) ++filled_cells;

  // sudoku is solved if all cells could be filled
  if (filled_cells == 81)
    std::cout << "Sudoku is solved:\n";
  else
    std::cout << "Could only fill " << filled_cells << " cells:\n";
  // output solution in 3 x 3 blocks (0 means no number deduced);
  for (unsigned int lr=0; lr<9; lr+=3) {
    for (unsigned int r=lr; r<lr+3; ++r) {
      for (unsigned int lc=0; lc<9; lc+=3) {
        for (unsigned int c=lc; c<lc+3; ++c)
          std::cout << unique_number (r, c) << " "; // blank after number
        std::cout << " "; // extra blank after every three columns
      }
      std::cout << "\n"; // extra line break after every third row
    }
    std::cout << "\n"; // line break after row
  }
}
```

B.9 Recursion

Solution to Exercise 111.

a)
```
// PRE: n >= 0
// POST: return value is false if n is even and true if n is odd
```

b)
```
// PRE: n >= 0
// POST: 2^n stars have been written to standard output
```

c) Let's analyze this function: if $n = b^e, e \geq 0$, then it is not hard to see that the function outputs $e = \log_b n$. An obvious precondition is $b \neq 0$. If n is not a power of b, things may go wrong, since the first argument may become 0 at some point, and then we enter an infinite recursion. This happens for example if $1 < n < b$. Also, if $b = 1$, we have an infinite recursion if $n > 1$. Let us therefore assume that $b \geq 2$.

Since n can always be written in the form

$$n = \sum_{i=0}^{k} \beta_i b^i, \quad 0 \leq \beta_i < b \ \forall i,$$

we can easily compute what happens, assuming that $k \geq 1$ and $\beta_k > 0$ (this just means $n \geq b$). The first integer division by b "cuts off" β_0 and yields

$$n' = \sum_{i=1}^{k} \beta_i b^{i-1}.$$

The next division cuts off β_1 and yields

$$n'' = \sum_{i=2}^{k} \beta_i b^{i-2}.$$

The pattern is clear: we eventually get the number β_k. If $\beta_k = 1$, the function call terminates, otherwise, the next division yields 0 as the first call parameter, and we have an infinite recursion. This yields the following pre- and postconditions.

```
// PRE: b >= 2, there exists e with b^e <= n < 2*b^e
// POST: return value is e = log_b(n), rounded down
```

Solution to Exercise 112.

a) This does not always terminate. Consider the call `f(1)`. This recursively calls `f(f(0))` which is `f(1)` again, and so on.

b) Let us move the call parameter `A(m, n-1)` out of the recursive call to make things more clear. This does not change anything since that parameter has to be evaluated before the recursive function call anyway.

```
unsigned int A (unsigned int m, unsigned int n) {
  // POST: return value is the Ackermann function value A(m,n)
  if (m == 0) return n+1;
  if (n == 0) return A(m-1,1);
  unsigned int param =  A(m, n-1);
  return A(m-1, param);
}
```

Now we see that the pair (m, n) gets *lexicographically* smaller in every recursive call. Under lexicographic order, $(m', n') < (m, n)$ if $m' < m$, or if $m' = m$ and $n' < n$. Therefore, starting from (m, n), the first parameter must go down by one after a finite number of recursive calls, meaning that it must also reach 0 after a finite number of recursive calls. At that point, the recursion bottoms out.

c) This one is somewhat tricky, since when you run it, it always seems to terminate. But this is only due to overflow in the arithmetic. Mathematically, this function does not terminate: if $m \geq n$, then no parameter decreases in the recursive call, and we again have

$$n' := (n + m) \operatorname{div} 2 \leq m < 2m =: m'.$$

Solution to Exercise 113. Here is the program for a) and b).

```
// Prog: mccarthy.cpp
// defines and calls McCarthy's 91 Function
#include <iostream>

// POST: return value is M(n), where M is McCarthy's 91 Function
unsigned int mccarthy(const unsigned int n) {
  if (n > 100)
    return n - 10;
  else
    return mccarthy(mccarthy(n + 11));
}

int main()
{
  // input
  std::cout << "Compute McCarthy's 91 Function M(n) for n =? ";
  unsigned int n;
  std::cin >> n;

  // computation and output
  std::cout << "M(" << n << ") = " << mccarthy(n) << "\n";

  return 0;
}
```

For c), you play with the program a little and start to guess that

$$M(n) = \begin{cases} n-10, & \text{if } n > 100 \\ 91, & \text{if} n \leq 100 \end{cases},$$

and this obviously explains the name *McCarthy's 91 Function*. Here is an inductive proof of this fact. Actually, we only need to handle the finitely many cases $n = 0, \ldots, 100$ since for $n > 100$, the result $n - 10$ follows from the definition. So we do backwards induction. Assume that we need to establish the validity of the formula for a given $n \leq 100$. We assume that the formula is already correct for all larger values of n. The definition gives us

$$M(n) = M(M(n+11)).$$

If $n + 11 > 100$, we thus get $M(n) = M(n + 11 - 10) = M(n + 1)$. If $n = 100$, this is $M(101) = 91$, and if $n < 100$, then $n+1 \leq 100$, so by induction we also get $M(n+1) = 91$. If $n + 11 \leq 100$, we inductively get $M(n + 11) = 91$, hence

$$\begin{aligned} M(n) &= M(91) := M(M(102)) \\ &= M(92) := M(M(103)) \\ &= M(93) := \cdots \\ &= M(99) := M(M(110)) \\ &= M(100) := M(M(111)) \\ &= M(101) = 91. \end{aligned}$$

Solution to Exercise 114. Let us start with part b). The first variant has the problem that if we *first* compute $n!$ and *only then* divide by $k!$ and $(n-k)!$, we can't compute many binomial coefficients, since $n!$ does not fit into an `unsigned int` variable already for small n (if we have 32-bit arithmetic, then 12! is the highest we can do).

The second variant is bad since it is very slow. We have the same phenomenon as with the Fibonacci numbers: the computation time is at least proportional to the computed number itself, and binomial coefficients are quite large.

The third solution seems to be best in practice, but even here, we have to be somewhat careful in the implementation. We should *first* multiply $\binom{n-1}{k-1}$ with n, and *only then* divide by k. First dividing by k is tempting in order to keep the numbers small but that doesn't work, since $\binom{n-1}{k-1}$ might not be divisible by k. This third method produces intermediate numbers that *are* larger than absolutely necessary, but only by at most n.

```
unsigned int binomial(unsigned int n, unsigned int k)
{
  if (n < k) return 0;
  if (k == 0) return 1;
  return n * binomial(n-1, k-1) / k;
}
```

Solution to Exercise 115.

```
// Prog: partition.cpp
// compute in how many ways a fixed amount of money can be
// partioned using the available denominations (banknotes
// and coins). This program is for the currency CHF, where
// the denominations are (in centimes)
//    100000, 20000, 10000, 5000, 2000, 1000 (banknotes)
//    500, 200, 100, 50, 20, 10, 5           (coins)
//
// Example: CHF 0,20 can be partitioned in four ways
//     (20), (10, 10), (10, 5, 5), and (5, 5, 5, 5)
#include<iostream>

// PRE: [begin, end) is a valid nonempty range that describes
//      a sequence of denominations d_1 > d_2 > ... > d_n > 0
// POST: return value is the number of ways to partition amount
//       using denominations from d_1, ..., d_n
unsigned int partitions (const unsigned int amount,
                         const unsigned int* begin,
                         const unsigned int* end)
{
  if (amount == 0) return 1;
  unsigned int ways = 0;
  // ways = ways_1 + ... + ways_n, where ways_i is the number
  // of ways to partition amount using d_i as the largest
  // denomination
  for (const unsigned int* d = begin; d != end; ++d)
    // ways_i = number of partitions of the form (d_i, X), with
    // (X) being a partition of amount-d_i using d_i,...,d_n
    if (amount >= *d) ways += partitions (amount-*d, d, end);

  return ways;
}

int main()
{
  // the 13 denominations of CHF
  unsigned int chf[] =
    {100000, 20000, 10000, 5000, 2000, 1000, 500, 200, 100, 50, 20, 10, 5};

  // input
  std::cout << "In how many ways can I own x CHF-centimes for x =? ";
  unsigned int x;
  std::cin >> x;

  // comutation and output
  std::cout << partitions (x, chf, chf+13) << "\n";

  return 0;
}
```

The number of ways in which you can own CHF 1 is 50, and CHF 10 can be owned in 104561 ways. The above program becomes very slow for larger values, since during the recursive calls, many values are computed over and over again. For CHF 50, we already have to wait "forever". We can speed things up by using dynamic programming. We don't even have to change the structure of our function, but we provide it with an additional twodimensional array to store the values that have already been computed. Whenever we need a value, we first check whether it has already been computed, and only if this is not the case, we recursively call the function.

```
// Prog: partition2.cpp
// compute in how many ways a fixed amount of money can be
// partioned using the available denominations (banknotes
// and coins). This program is for the currency CHF, where
// the denominations are (in centimes)
//    100000, 20000, 10000, 5000, 2000, 1000 (banknotes)
//    500, 200, 100, 50, 20, 10, 5           (coins)
//
// Example: CHF 0,20 can be partitioned in four ways
//     (20), (10, 10), (10, 5, 5), and (5, 5, 5, 5)
#include<iostream>
#include<algorithm>
#include<vector>

// PRE: [first, last) is a valid nonempty range that describes
//      a sequence of denominations d_1 > d_2 > ... > d_n > 0
//      memory is a pointer to a twodimensional array with
//      number of rows >= amount, and number of columns >=
//      last-first, with the following property:
//      for 0 < a < amount, and for d in [first, last),
//      memory[a-1][last-d-1] either has value -1,
//      or it is equal to dyn_prog_partitions (a, d, last, memory)
// POST: return value is the number of ways to partition amount
//       using denominations from d_1, ..., d_n

unsigned int dyn_prog_partitions (const unsigned int amount,
                  const unsigned int* first,
                  const unsigned int* last,
                  std::vector<std::vector<int> >* memory)
{
  if (amount == 0) return 1;
  unsigned int ways = 0;
  // ways = ways_1 + ... + ways_n, where ways_i is the number
  // of ways to partition amount using d_i as the largest
  // denomination
  for (const unsigned int* d = first; d != last; ++d)
    // ways_i = number of partitions of the form (d_i, X), with
    // (X) being a partition of amount-d_i using d_i,...,d_n
    if (amount >= *d) {
      // is ways_i already stored in memory?
      int stored_value = -1;
      if (amount > *d)
        stored_value = (*memory)[amount - *d - 1][last-d-1];
      if (stored_value != -1)
        ways += stored_value;
      else
        ways += dyn_prog_partitions (amount - *d, d, last, memory);
    }
  // store the new value
  (*memory)[amount-1][last-first-1] = ways;
  return ways;
}

// PRE: [first, last) is a valid nonempty range that describes
//      a sequence of denominations d_1 > d_2 > ... > d_n > 0
// POST: return value is the number of ways to partition amount
//       using denominations from d_1, ..., d_n
unsigned int partitions (unsigned int amount,
                         const unsigned int* first,
                         const unsigned int* last)
{
  // allocate memory for dynamic programming approach
  std::vector<std::vector<int> > memory(amount, std::vector<int> (last - first, -1) );

```

```
  // call the version with memory
  const unsigned int result =
    dyn_prog_partitions (amount, first, last, &memory);

  return result;
}

int main()
{
  // the 13 denominations of CHF
  unsigned int chf[] =
    {100000, 20000, 10000, 5000, 2000, 1000, 500, 200, 100, 50, 20, 10, 5};

  // input
  std::cout << "In how many ways can I own x CHF-centimes for x =? ";
  unsigned int x;
  std::cin >> x;

  // comutation and output
  std::cout << partitions (x, chf, chf+13) << "\n";

  return 0;
}
```

With this, we can very quickly compute the number of ways for CHF 50; it is 513269191. Beyond that, the type `unsigned int` is at some point no longer sufficient to represent the number of ways, since these are simply too many. But we may switch to `ifmp::integer` and then even get the answer for CHF 100 quickly (it is 47580509178).

```
// Prog: partition3.cpp
// compute in how many ways a fixed amount of money can be
// partioned using the available denominations (banknotes
// and coins). This program is for the currency CHF, where
// the denominations are (in centimes)
//    100000, 20000, 10000, 5000, 2000, 1000 (banknotes)
//    500, 200, 100, 50, 20, 10, 5           (coins)
//
// Example: CHF 0,20 can be partitioned in four ways
//     (20), (10, 10), (10, 5, 5), and (5, 5, 5, 5)
#include<iostream>
#include<algorithm>
#include<vector>
#include<IFMP/integer.h>


// PRE: [first, last) is a valid nonempty range that describes
//      a sequence of denominations d_1 > d_2 > ... > d_n > 0
//      memory is a pointer to a twodimensional array with
//      number of rows >= amount, and number of columns >=
//      last-first, with the following property:
//      for 0 < a < amount, and for d in [first, last),
//      memory[a-1][last-d-1] either has value -1,
//      or it is equal to dyn_prog_partitions (a, d, last, memory)
// POST: return value is the number of ways to partition amount
//       using denominations from d_1, ..., d_n

typedef std::vector<unsigned int>::const_iterator cIter;

ifmp::integer dyn_prog_partitions (const unsigned int amount,
                                   const unsigned int* first,
                                   const unsigned int* last,
                                   std::vector<std::vector<ifmp::integer> >* memory)
```

```
{
  if (amount == 0) return 1;
  ifmp::integer ways = 0;
  // ways = ways_1 + ... + ways_n, where ways_i is the number
  // of ways to partition amount using d_i as the largest
  // denomination
  for (const unsigned int* d = first; d != last; ++d)
    // ways_i = number of partitions of the form (d_i, X), with
    // (X) being a partition of amount-d_i using d_i,...,d_n
    if (amount >= *d) {
      // is ways_i already stored in memory?
      ifmp::integer stored_value = -1;
      if (amount > *d)
        stored_value = (*memory)[amount - *d - 1][last-d-1];
      if (stored_value != -1)
        ways += stored_value;
      else
        ways += dyn_prog_partitions (amount - *d, d, last, memory);
    }
  // store the new value
  (*memory)[amount-1][last-first-1] = ways;
  return ways;
}

// PRE: [first, last) is a valid nonempty range that describes
//      a sequence of denominations d_1 > d_2 > ... > d_n > 0
// POST: return value is the number of ways to partition amount
//       using denominations from d_1, ..., d_n
ifmp::integer partitions (unsigned int amount,
                          const unsigned int* first,
                          const unsigned int* last)
{
  // allocate memory for dynamic programming approach
          std::vector<std::vector<ifmp::integer> > memory(amount, std::vector<ifmp::integer> (last

  // call the version with memory
  const ifmp::integer result =
    dyn_prog_partitions (amount, first, last, &memory);

  return result;
}

int main()
{
  // the 13 denominations of CHF
  unsigned int chf[] =
    {100000, 20000, 10000, 5000, 2000, 1000, 500, 200, 100, 50, 20, 10, 5};

  // input
  std::cout << "In how many ways can I own x CHF-centimes for x =? ";
  unsigned int x;
  std::cin >> x;

  // comutation and output
  std::cout << partitions (x, chf, chf+13) << "\n";

  return 0;
}
```

Solution to Exercise 116.

```
#include<iostream>

```

```
// PRE:  d >= k
// POST: all codes with digits between 1 and 9 are output that
//       result from the partial code p by an extension with d
//       digits, k of which are 1; the return value is the
//       number of such codes
unsigned int crack (const unsigned int p, const unsigned int d,
                    const unsigned int k)
{
  if (d == 0) {
    // k == 0 as well by PRE, and we have a full code
    std::cout << p << " ";
    return 1;
  }

  // there are two possibilities to continue:
  //   next digit is 1, or not
  unsigned int n = 0; // total number of codes

  // poss 1: only if we still have 1's left
  if (k > 0)
    n += crack (10 * p + 1, d - 1, k - 1); // next digit is 1

  // poss 2: only if not all remaining digits have to be 1's
  if (d > k)
    for (unsigned int i=2; i<10; ++i)
      n += crack (10 * p + i, d - 1, k); // next digit is i != 1

  return n;
}

// PRE:  d >= k

// POST: all d-digit codes with digits between 1 and 9 are
//       output that have exactly k digits equal to one;
//       the return value is the number of such codes
unsigned int crack (const unsigned int d, const unsigned int k)
{
  return crack (0, d, k);
}

int main() {
  const int n = crack (2, 1);
  std::cout << "\nThere were " << n << " possible codes.\n";
  return 0;
}
```

Solution to Exercise 117. The function is rewritten in a way similar to what we did with the Fibonacci numbers. The recursive variant becomes very slow at some point, while the iterative version is always fast.

```
// Prog: rec2it2.cpp
// rewrites a recursive function in iterative form a la Fibonacci
#include<iostream>

unsigned int f (const unsigned int n)
{
  if (n <= 2) return 1;
  return f(n-1) + 2 * f(n-3);
}

unsigned int f_it (const unsigned int n)
{
```

```
    if (n <= 2) return 1;
    unsigned int a = 1;  // f(0)
    unsigned int b = 1;  // f(1)
    unsigned int c = 1;  // f(2)
    for (unsigned int i = 3; i < n; ++i) {
      const unsigned int a_prev = a;  // f(i-3)
      a = b;                          // f(i-2)
      b = c;                          // f(i-1)
      c = b + 2 * a_prev;             // f(i)
    }
    return c + 2 * a;                 // f(n-1) + 2 * f(n-3)
}

int main()
{
  std::cout << "Comparing f and f_it...\n";
  for (int n = 0; n < 100; ++n)
    std::cout << f(n) << " = " << f_it(n) << "\n";

  return 0;
}
```

Solution to Exercise 118. Here, we cannot directly use the Fibonacci trick, since it is not sufficient to remember only a constant number of previous function values. But we can simply remember *all* values by using an vector. Here, both versions are reasonably fast, since the second recursive call is to a problem of size $n/2$ only.

```
// Prog: rec2it2.cpp
// rewrites a recursive function in iterative form by using an vector
#include<iostream>
#include<vector>

unsigned int f (const unsigned int n)
{
  if (n == 0) return 1;
  return f(n-1) + 2 * f(n/2);
}

unsigned int f_it (const unsigned int n)
{
  if (n == 0) return 1;
  std::vector<unsigned int> f_values(n + 1, 0);
  f_values[0] = 1;
  for (unsigned int i=1; i<=n; ++i)
    f_values[i] = f_values[i-1] + 2 * f_values[i/2];
  return f_values[n];
}

int main()
{
  std::cout << "Comparing f and f_it...\n";
  for (int n = 0; n < 100; ++n)
    std::cout << f(n) << " = " << f_it(n) << "\n";

  return 0;
}
```

Solution to Exercise 119. Here is the sorting program.

n	1,600,000	3,200,000	6,400,000	128,000,000
merge_sort	0.83	1.75	3.65	7.4
std::sort	0.32	0.76	1.7	3.65
speedup	2.6	2.3	2.14	2.0

Table 10: *Runtime in seconds of* merge-sort *vs./* std::sort

```
// Prog: std_sort.cpp
// tests std::sort on random input

#include<iostream>
#include<algorithm>
#include<vector>

int main()
{
  // input of number of values to be sorted
  unsigned int n;
  std::cin >> n;

  std::vector<int> a(n, 0);

  std::cout << "Sorting " << n << " integers...\n";

  // create sequence:
  for (int i=0; i<n; ++i) a[i] = i;
  std::random_shuffle (a.begin(), a.end() );

  // sort into ascending order
  std::sort (a.begin(), a.end() );

  // is it really sorted ?
  for (int i=0; i<n-1;++i)
    if (a[i] != i) std::cout << "Sorting error!\n";

  return 0;
}
```

On the platform of the authors, the following happens Table 10). Here, *Speedup* is the factor by which `std::sort` is faster than `merge_sort` in absolute runtime. We see that the speedup is larger for smaller inputs but then approaches a factor of 2.

Solution to Exercise 120. As for *merge-sort*, it can be shown that $T(0) = 0, T(1) = 1$ and

$$T(n) \leq 1 + \max\left(T(\lfloor \frac{n}{2} \rfloor), T(\lceil \frac{n}{2} \rceil)\right), \quad n \geq 2.$$

Given this, a bound of $T(n) \leq 1 + \lceil \log_2 n \rceil$ can be shown by induction.

In addtion, we give a alternative iterative approach to do binary search.

```
// Prog: binary_search2.cpp
// implements and tests a iterative binary search function

#include <iostream>
#include <cassert>
#include <vector>
```

```

typedef std::vector<int>::const_iterator Cvit;

// PRE:  [begin, end) is a valid range, and the elements *p,
//       p in [begin, end) are in ascending order
// POST: return value is an iterator p in [begin, end) such
//       that *p = x, or the iterator end, if no such iterator
//       exists
Cvit bin_search (const Cvit begin, const Cvit end, const int x)
{
  Cvit first = begin;
  Cvit last = end;
  // in the loop we maintain the following two properties:
  // [first, last) is a valid range and if an iterator p exists with
  // *p == x, then p is in [first, last)
  while(last - first > 1) {
    const int n = last - first;    //number of elements in the range
    const Cvit middle = first + n / 2;
    if (*middle > x)   // x can't be in [middle, last)
      last = middle;
    else  //*middle <= x, continue the search in [middle, last)
      first = middle;
  }

  if (first == end || *first != x) //empty range or x not found
    return end;
  else
    return first;
}

int main()
{
  // input length of test array, n >= 0
  int n;
  std::cin >> n;

  std::vector<int> v(n);

  // 0, 2, 4, 6,...
  for (int i=0; i<n; ++i) v[i] = 2*i;

  // test binary search
  for (int i=0; i<2*n; i+=2)
    assert (bin_search (v.begin(), v.end(), i) == v.begin()+i/2);
  for (int i=-1; i<2*n+2; i+=2)
    assert (bin_search (v.begin(), v.end(), i) == v.end());

  // special case: empty range!
  assert( bin_search (v.begin(), v.begin(), 0) == v.begin());

  std::cout << "The search function seems to work as expected!\n";

  return 0;
}
```

Solution to Exercise 121. The inequality simply follows from the fact that $\log_2 x$ is a monotone increasing function. To see the equation, let us first consider the case where n is a power of two, $n = 2^k$, say, with $k \geq 1$. Then all involved numbers are integers, the rounding operation does nothing, and the statement follows from

$$\log_2 \frac{n}{2} = \log_2 n - 1.$$

Otherwise (if n is not a power of two), let us choose the unique number k such that

$$2^k < n < 2^{k+1}.$$

By taking logarithms, it follows that

$$k < \log_2 n < k + 1,$$

so

$$\lceil \log_2 n \rceil = k + 1. \tag{B.1}$$

We also have

$$2^{k-1} < \frac{n}{2} < 2^k,$$

and this yields

$$2^{k-1} < \lceil \frac{n}{2} \rceil \leq 2^k.$$

Taking logarithms, we get

$$k - 1 < \log_2 \lceil \frac{n}{2} \rceil \leq k,$$

and this means that

$$\lceil \log_2 \lceil \frac{n}{2} \rceil \rceil = k. \tag{B.2}$$

The desired equality now follows from (B.1) and (B.2).

Solution to Exercise 122.

a)

```
// Prog: lindenmayer_a.cpp
// Draw turtle graphics for the Lindenmayer system with
// production F -> FF+F+F+F+F+F-F, initial word F+F+F+F
// and rotation angle 90 degrees

#include <iostream>
#include <IFMP/turtle>

void f (const unsigned int i) {
  // POST: the word w_i^F is drawn
  if (i == 0)
    ifmp::forward();  // F
  else {
    f(i-1);           // w_{i-1}^F
    f(i-1);           // w_{i-1}^F
    ifmp::left(90);   // +
```

```
    f(i-1);           // w_{i-1}^F
    ifmp::left(90);    // +
    f(i-1);           // w_{i-1}^F
    ifmp::left(90);    // +
    f(i-1);           // w_{i-1}^F
    ifmp::left(90);    // +
    f(i-1);           // w_{i-1}^F
    ifmp::left(90);    // +
    f(i-1);           // w_{i-1}^F
    ifmp::right(90);   // -
    f(i-1);           // w_{i-1}^F
  }
}

int main () {
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // draw w_n = w_n(F+F+F+F)
  f(n); ifmp::left(90); f(n); ifmp::left(90);
  f(n); ifmp::left(90); f(n);
  return 0;
}
```

```
// Prog: lindenmayer_b.cpp
// Draw turtle graphics for the Lindenmayer system with
// productions X -> Y+X+Y, Y -> X-Y-X, initial word Y
// and rotation angle 60 degrees

#include <iostream>
#include <IFMP/turtle>

void y (const unsigned int i);
// necessary: x and y call each other

void x (const unsigned int i) {
  // POST: w_i^X is drawn
  if (i == 0)
    ifmp::forward();
  else {
    y(i-1);           // w_{i-1}^Y
    ifmp::left(60);    // +
    x(i-1);           // w_{i-1}^X
    ifmp::left(60);    // +
    y(i-1);           // w_{i-1}^Y
  }
}

void y (const unsigned int i) {
  // POST: w_i^Y is drawn
  if (i == 0)
    ifmp::forward();
  else {
    x(i-1);           // w_{i-1}^X
    ifmp::right(60);   // -
    y(i-1);           // w_{i-1}^Y
    ifmp::right(60);   // -
    x(i-1);           // w_{i-1}^X
  }
}

int main () {
```

```
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // draw w_n = w_n^Y
  y(n);

  return 0;
}
```

b)

```
// Prog: lindenmayer_c.cpp
// Draw turtle graphics for the Lindenmayer system with
// productions X ->  X+Y++Y-X--XX-Y+, Y -> -X+YY++Y+X--X-Y,
// initial word Y  and rotation angle 60 degrees

#include <iostream>
#include <IFMP/turtle>

void y (const unsigned int i);
// necessary: x and y call each other

void x (const unsigned int i) {
  // POST: w_i^X is drawn
  if (i == 0)
    ifmp::forward();
  else {
    x(i-1);            // w_{i-1}^X
    ifmp::left(60);     // +
    y(i-1);            // w_{i-1}^Y
    ifmp::left(60);     // +
    ifmp::left(60);     // +
    y(i-1);            // w_{i-1}^Y
    ifmp::right(60);    // -
    x(i-1);            // w_{i-1}^X
    ifmp::right(60);    // -
    ifmp::right(60);    // -
    x(i-1);            // w_{i-1}^X
    x(i-1);            // w_{i-1}^X
    ifmp::right(60);    // -
    y(i-1);            // w_{i-1}^Y
    ifmp::left(60);     // +
  }
}

void y (const unsigned int i) {
  // POST: w_i^Y is drawn
  if (i == 0)
    ifmp::forward();
  else {
    ifmp::right(60);    // -
    x(i-1);            // w_{i-1}^X
    ifmp::left(60);     // +
    y(i-1);            // w_{i-1}^Y
    y(i-1);            // w_{i-1}^Y
    ifmp::left(60);     // +
    ifmp::left(60);     // +
    y(i-1);            // w_{i-1}^Y
    ifmp::left(60);     // +
    x(i-1);            // w_{i-1}^X
    ifmp::right(60);    // -
    ifmp::right(60);    // -
    x(i-1);            // w_{i-1}^X
```

```
    ifmp::right(60);    // -
    y(i-1);            // w_{i-1}^Y
  }
}

int main () {
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // draw w_n = w_n^Y
  y(n);

  return 0;
}
```

Solution to Exercise 123. To move n disks from peg s to peg t, we can first move the topmost $n-1$ disks to the helper peg, then move the bottommost disk directly to peg t, and then move the $n-1$ disks again, this time from the helper peg to peg t. In the following program, we do this through a recursive function. As a small hack, we determine the number of the third helper peg as $6-s-t$.

```
// Prog: hanoi.cpp
// solves the Tower of Hanoi puzzle

#include<iostream>

// PRE:  s and t are different and both in {1,2,3}
// POST: the sequence of moves necessary to transfer a stack of n
//       disks from peg s to peg t is written to standard output
void hanoi (const unsigned int n, const int s, const int t)
{
  if (n > 0) {
    // move topmost n-1 disks from s to helper peg 6-s-t
    hanoi (n-1, s, 6-s-t);
    // move bottommost disk from s to t
    std::cout << "(" << s << "," << t << ")";
    // move the n-1 disks from the helper peg to t
    hanoi (n-1, 6-s-t, t);
  }
}

int main()
{
  // input
  std::cout << "Move a stack of n disks for n =? ";
  unsigned int n;
  std::cin >> n;

  // output
  hanoi (n, 1, 3);
  std::cout << "\n";

  return 0;
}
```

Solution to Exercise 124. The Lucas-Lehmer test is based on the recursive sequence

$s_1, s_2, \ldots$ defined as follows.

$$\begin{aligned} s_1 &= 4, \\ s_i &= s_{i-1}^2 - 2, \quad i > 1. \end{aligned}$$

The actual test is provided by the following

> **Theorem.** Let $n \geq 3$ be prime. The Mersenne number $2^n - 1$ is prime if and only if s_{n-1} is divisible by $2^n - 1$.

Thus, we simply need to compute s_{n-1} and check this. The good thing is (see also Exercise 55) that in order to compute the remainder of s_{n-1} modulo $2^n - 1$, it suffices to do *all* calculations modulo $2^n - 1$. This is quite a saving, since the number s_{n-1} is huge compared to $2^n - 1$ itself, for larger n. Here is the program.

```
// Program: lucas_lehmer.cpp
// finds Mersenne primes using the Lucas-Lehmer test

#include <iostream>
#include <cmath>
#include <cassert>
#include <IFMP/integer.h>

// PRE:  n > 2
// POST: return value is true if and only if n is prime
bool is_prime (const unsigned int n)
{
  assert (n > 2);

  // Computation: test possible divisors d up to sqrt(n)
  const unsigned int bound = (unsigned int)(std::sqrt(n));
  unsigned int d;
  for (d = 2; d <= bound && n % d != 0; ++d);

  // Output
  return d > bound;
}

// PRE:  p > 2 is prime, M_p = 2^p-1
// POST: returns true if and only if M_p = 2^p-1 is prime
//       according to the Lucas-Lehmer test
bool lucas_lehmer (const unsigned int p, const ifmp::integer M_p)
{
  assert (p > 2 && is_prime (p));

  // compute (p-1)-st term of sequence s, modulo M_p
  ifmp::integer s = 4; // s_1
  for (unsigned int i = 2; i < p; ++i)
    s = (s * s - 2) % M_p;
  return (s == 0);
}

int main()
{
  // try all prime exponents
  unsigned int p = 3;
  ifmp::integer m_p = 8; // 2^p = M_p + 1
  std::cout << "List of Mersenne primes 2^p-1 (p > 2):\n";
  for(;;) {
```

```
    if (is_prime (p) && lucas_lehmer (p, m_p-1))
      std::cout << "2^" << p << "-1" << std::endl;
    ++p; m_p *= 2;
  }

  return 0;
}
```

Running it for something like eight days produced the following output (it gets really slow only beyond exponent $1,000$).

```
List of Mersenne primes 2^p-1 (p > 2):
2^3-1
2^5-1
2^7-1
2^13-1
2^17-1
2^19-1
2^31-1
2^61-1
2^89-1
2^107-1
2^127-1
2^521-1
2^607-1
2^1279-1
2^2203-1
2^2281-1
2^3217-1
2^4253-1
2^4423-1
2^9689-1
2^9941-1
2^11213-1
```

Solution to Exercise 125. Here comes our handwaving: what is the average number E_i of packages that you need to buy, given that you are still missing i stickers? Obviously, $E_0 = 0$, and E_n is the value that we are interested in.

If you are missing $i > 0$ stickers, you definitely have to buy another package, and this might decrease the number of missing stickers. In fact, the new number is between i (if you already have all five stickers that were in the package) and $i - 5$ (if all stickers were new). What is the probability that you go down to $i - k$ missing stickers, $k = 0, ..., 5$? For this event to happen, exactly k out of the 5 stickers have to come from the i that you are missing, and $5 - k$ have to come from the $n - i$ that you already have. The total number of 5-tuples of stickers is

$$\binom{n}{5} = \frac{n!}{5!(n-5)!},$$

and

$$\binom{i}{k}\binom{n-i}{5-k}$$

of them lead to the aforementioned event. (Recall that the binomial coefficient $\binom{m}{\ell}$ counts the number of ways in which ℓ objects can be selected from a total of m objects.)

This gives us the recursive formula

$$E_i = 1 + \sum_{k=0}^{5} \frac{\binom{i}{k}\binom{n-i}{5-k}}{\binom{n}{5}} E_{i-k}, \quad i > 0.$$

Note that for $i < 5$, this formula features negative subscripts $i - k$, but all these have multiplicative factor 0.

This formula is an incarnation of the partition theorem for conditional expectation, but the handwaving is this: given that you are missing $i > 0$ stickers, you need another package, and then the average number of packages for $i - k$ missing stickers, where the probability that a given k occurs is the above fraction.

The recursive formula is still not useful since E_i depends on E_i (for $k = 0$), but we can get rid of this by multiplying with $\binom{n}{5}$:

$$\left(\binom{n}{5} - \binom{n-i}{5}\right) E_i = \binom{n}{5} + \sum_{k=1}^{5} \binom{i}{k}\binom{n-i}{5-k} E_{i-k}.$$

Thus we get

$$E_i = \frac{\binom{n}{5}}{\binom{n}{5} - \binom{n-i}{5}} + \sum_{k=1}^{5} \frac{\binom{i}{k}\binom{n-i}{5-k}}{\binom{n}{5} - \binom{n-i}{5}} E_{i-k}, \quad i > 0.$$

The program now simply computes (and stores) the E_i's according to this formula, in the order E_0 (which we know), $E_1, E_2, \ldots, E_n$. In this way, we have always already computed the values that we need.

We use the third function from Exercise ?? to compute binomial coefficients, but since these may become pretty large, we work over the type `double` right away. $\binom{555}{5}$, the largest binomial coefficient that may come up, is 430960344486, a number that would require more than 32 bits but less than 53, so it will even exactly fit into a `double`).

```
// Prog: panini.cpp
// computes the expected number of 5-sticker packages that need
// to be purchased in order to have all n stickers in the collection

#include<iostream>

// POST: computes binomial coefficient "n choose k"
double binomial(const unsigned int n, const unsigned int k)
{
```

```
  if (n < k) return 0.0;
  if (k == 0) return 1.0;
  return n * binomial(n-1, k-1) / k;
}

int main()
{
  // input
  std::cout << "Collection size =? ";
  unsigned int n;
  std::cin >> n;

  // allocate the array...
  double* const E = new double[n+1]; // E_0,...,E_n

  // ...and fill it
  E[0] = 0;
  for (unsigned int i = 1; i <= n; ++i) {
    const double d = binomial(n, 5) - binomial(n-i, 5);
    E[i] =  binomial(n, 5) / d;
    for (unsigned int k = 1; k < 6 && k <= i; ++k)
      E[i] += binomial(i, k) * binomial(n-i, 5-k) / d * E[i-k];
  }

  // output
  std::cout << "Average number of packages is " << E[n] << ".\n";

  // clean up
  delete[] E;

  return 0;
}
```

According to this program, the expected number of packages for $n = 555$ is 763.213, so up to rounding to integer, the newspaper was right.

Solution to Exercise 126. Naturally, it's hard to present a solution here. Instead, let us just list the program corresponding to the production

$$F \mapsto FF + [+F - F - F] - [-F + F + F]$$

given as an example in the challenge.

```
// Prog: bush.cpp
// Draw turtle graphics for the Lindenmayer system with
// production F -> FF+[+F-F-F]-[-F+F+F], initial word F
// and rotation angle 22 degrees

#include <iostream>
#include <IFMP/turtle>

// POST: the word w_i^F is drawn
void f (const unsigned int i) {
  if (i == 0)
    ifmp::forward();  // F
  else {
    f(i-1);           // F
    f(i-1);           // F
    ifmp::left(22);    // +
    ifmp::save();      // [
    ifmp::left(22);    // +
```

```
    f(i-1);           // F
    ifmp::right(22);  // -
    f(i-1);           // F
    ifmp::right(22);  // -
    f(i-1);           // F
    ifmp::restore();  // ]
    ifmp::right(22);  // -
    ifmp::save();     // [
    ifmp::right(22);  // -
    f(i-1);           // F
    ifmp::left(22);   // +
    f(i-1);           // F
    ifmp::left(22);   // +
    f(i-1);           // F
    ifmp::restore();  // ]
  }
}

int main () {
  std::cout << "Number of iterations =? ";
  unsigned int n;
  std::cin >> n;

  // draw w_n = w_n(F), vertically
  ifmp::left(90);
  f(n);

  return 0;
}
```

B.10 Structs

Solution to Exercise 127.

```
// Prog: create_rational_vector_3.cpp
// writes a function that can be used to initialize a
// rational_vector_3 from 6 int's
#include <cassert>

struct rational {
  int n;
  int d;  // INV: d != 0
};

struct rational_vector_3 {
  rational x;
  rational y;
  rational z;
};

// PRE: d != 0
// POST: the rational number n/d is returned
rational create_rational (const int n, const int d)
{
  assert (d != 0);
  rational result;
  result.n = n;
  result.d = d;
  return result;
}

// PRE: d1 != 0, d2 != 0, d3 != 0
// POST: the rational_vector_3 (n1/d1, n2/d2, n3/d3) is returned
rational_vector_3 create_rational_vector_3
(const int n1, const int d1, const int n2,
 const int d2, const int n3, const int d3)
{
  rational_vector_3 result;
  result.x = create_rational (n1, d1);
  result.y = create_rational (n2, d2);
  result.z = create_rational (n3, d3);
  return result;
}

int main()
{
  const rational_vector_3 v = create_rational_vector_3 (1,2,3,4,5,6);
  return 0;
}
```

Solution to Exercise 128.

There are several possible representations for three-valued logic. A convenient one is based on the following observation: if we interpret *false* as 0, *unknown* as 1 and *true* as 2, then AND corresponds to the minimum of the two numbers, while OR corresponds to the maximum. The following program defines the type `Tribool` and the two operators, and it uses them to reproduce the truth tables.

```
// Prog: tribool.cpp
```

```
2  // implements three-valued logic
3
4  #include<iostream>
5  #include<cassert>
6
7  struct Tribool {
8    // INV : value in {0, 1, 2}
9    // 0 = false, 1 = unknown, 2 = true
10   unsigned int value;
11 };
12
13 // PRE: val in {0, 1, 2}
14 // POST: return value is a Tribool with the corresponding value
15 Tribool tribool (const unsigned int val)
16 {
17   assert (val <= 2);
18   Tribool result;
19   result.value = val;
20   return result;
21 }
22
23 // POST: returns x AND y
24 Tribool operator&& (const Tribool x, const Tribool y)
25 {
26   Tribool result;
27   if (x.value < y.value)
28     result.value = x.value;
29   else
30     result.value = y.value;
31   return result;
32 }
33
34 // POST: returns x OR y
35 Tribool operator|| (const Tribool x, const Tribool y)
36 {
37   Tribool result;
38   if (x.value > y.value)
39     result.value = x.value;
40   else
41     result.value = y.value;
42   return result;
43 }
44
45 // POST: Tribool value is written to std::cout
46 std::ostream& operator<< (std::ostream& o, const Tribool x)
47 {
48   if      (x.value == 0)  return std::cout << "false   ";
49   else if (x.value == 1)  return std::cout << "unknown ";
50   else                    return std::cout << "true    ";
51 }
52
53 int main()
54 {
55   // print 3 x 3 truth table for AND
56   for (int x_val = 0; x_val < 3; ++x_val) {
57     const Tribool x = tribool (x_val);
58     for (int y_val = 0; y_val < 3; ++y_val) {
59       const Tribool y = tribool (y_val);
60       std::cout << (x && y);
61     }
62     std::cout << "\n";
63   }
64   std::cout << "\n";
65
66   // print 3 x 3 truth table for OR
```

```
  for (int x_val = 0; x_val < 3; ++x_val) {
    const Tribool x = tribool (x_val);
    for (int y_val = 0; y_val < 3; ++y_val) {
      const Tribool y = tribool (y_val);
      std::cout << (x || y);
    }
    std::cout << "\n";
  }
  std::cout << "\n";


  return 0;
}
```

Solution to Exercise 129.

The most natural representation is by an unsigned int value in the range $\{0, \ldots, 6\}$. Addition then simply adds the values and takes the result modulo 7. Subtraction could in principle be realized by subtracting the values and taking the result modulo 7, but the problem is that the intermediate subtraction result could be negative (and therefore not representable as unsigned int value). To overcome this problem, we simply add 7 to the first value before we subtract the second one; this is guaranteed to yield a positive value, and modulo 7, it makes no difference. The following program defines the type Z_7 and the two operators, and it uses them to reproduce the addition and subtraction table.

```
// Prog: modulo7.cpp
// defines operations in the ring Z_7
#include<iostream>

struct Z_7 {
  // INV : value in {0, 1, 2, 3, 4, 5, 6}
  unsigned int value;
};

// POST: return value is the sum of a and b
Z_7 operator+ (const Z_7 a, const Z_7 b)
{
  Z_7 result;
  result.value = (a.value + b.value) % 7;
  return result;
}

// POST: return value is the difference of a and b
Z_7 operator- (const Z_7 a, const Z_7 b)
{
  Z_7 result;
  result.value = (7 + a.value - b.value) % 7;
  return result;
}

int main ()
{
  Z_7 a;
  Z_7 b;

  // print table for addition
  for (unsigned int a_val = 0; a_val < 7; ++a_val) {
    for (unsigned int b_val = 0; b_val < 7; ++b_val) {
```

```
      a.value = a_val;
      b.value = b_val;
      std::cout << (a + b).value << " ";
    }
    std::cout << "\n";
  }
  std::cout << "\n";

  // print table for subtraction
  for (unsigned int a_val = 0; a_val < 7; ++a_val) {
    for (unsigned int b_val = 0; b_val < 7; ++b_val) {
      a.value = a_val;
      b.value = b_val;
      std::cout << (a - b).value << " ";
    }
    std::cout << "\n";
  }
  std::cout << "\n";

  return 0;
}
```

Solution to Exercise 130. The following program is at the same time the solution to the next Exercise 131.

```
// Program: rational.cpp
// Define and use operations on rational numbers.

#include <iostream>

// the new type Rational
struct Rational {
  int n;
  int d; // INV: d != 0
};

// POST: return value is the sum of a and b
Rational operator+ (const Rational a, const Rational b)
{
  Rational result;
  result.n = a.n * b.d + a.d * b.n;
  result.d = a.d * b.d;
  return result;
}

// POST: return value is the difference of a and b
Rational operator- (const Rational a, const Rational b)
{
  Rational result;
  result.n = a.n * b.d - a.d * b.n;
  result.d = a.d * b.d;
  return result;
}

// POST: return value is the product of a and b
Rational operator* (const Rational a, const Rational b)
{
  Rational result;
  result.n = a.n * b.n;
  result.d = a.d * b.d;
  return result;
}
```

```

// POST: return value is the quotient of a and b
// PRE:  b != 0
Rational operator/ (const Rational a, const Rational b)
{
  Rational result;
  result.n = a.n * b.d;
  result.d = a.d * b.n;
  return result;
}

// POST: return value is true if and only if a == b
bool operator== (const Rational a, const Rational b)
{
  return a.n * b.d == a.d * b.n;
}

// POST: return value is true if and only if a != b
bool operator!= (const Rational a, const Rational b)
{
  return !(a == b);
}

// POST: return value is true if and only if a < b
bool operator< (const Rational a, const Rational b)
{
  // here we have to watch out for signs
  if (a.d > 0 && b.d > 0 || a.d < 0 && b.d < 0)
    // no sign reversal in multplying by a.d and b.d
    return a.n * b.d < a.d * b.n;
  else
    // sign reversal
    return a.n * b.d > a.d * b.n;
}

// POST: return value is true if and only if a <= b
bool operator<= (const Rational a, const Rational b)
{
  return a < b || a == b;
}

// POST: return value is true if and only if a > b
bool operator> (const Rational a, const Rational b)
{
  return b < a;
}

// POST: return value is true if and only if a >= b
bool operator>= (const Rational a, const Rational b)
{
  return a > b || a == b;
}


// POST: a has been written to o
std::ostream& operator<< (std::ostream& o, const Rational a)
{
  return o << a.n << "/" << a.d;
}

// POST: a has been read from i
// PRE:  i starts with a rational number of the form "n/d"
std::istream& operator>> (std::istream& i, Rational& a)
{
  char c; // separating character, e.g. '/'
```

```
  return i >> a.n >> c >> a.d;
}

int main ()
{
  // input
  std::cout << "Rational number r:\n";
  Rational r;
  std::cin >> r;

  std::cout << "Rational number s:\n";
  Rational s;
  std::cin >> s;

  // test the operations
  std::cout << "Sum is        " << r + s    << ".\n";
  std::cout << "Difference is " << r - s    << ".\n";
  std::cout << "Product is    " << r * s    << ".\n";
  std::cout << "Quotient is   " << r / s    << ".\n";
  std::cout << "r == s?       " << (r == s) << ".\n";
  std::cout << "r != s?       " << (r != s) << ".\n";
  std::cout << "r < s?        " << (r < s)  << ".\n";
  std::cout << "r <= s?       " << (r <= s) << ".\n";
  std::cout << "r > s?        " << (r > s)  << ".\n";
  std::cout << "r >= s?       " << (r >= s) << ".\n";

  return 0;
}
```

Solution to Exercise 131. See solution to Exercise 130 for the answer of part (a) and (b).

Solution to Exercise 132. It turns out that only `operator+` requires real work; the following program evaluates it according to the case distinction in the small table that is given.

```
#include<iostream>

struct extended_int {
  unsigned int u;  // the absolute value
  bool         n;  // the sign (true means negative)
};

// POST: return value is the sum of a and b
//              ||  a > b     ||  a <= b
// ====================================
// (a,+)+(b,+) ||          (a+b,+)
// (a,+)+(b,-) || (a-b,+)     ||  (b-a,-)
// (a,-)+(b,+) || (a-b,-)     ||  (b-a,+)
// (a,-)+(b,-) ||          (a+b,-)
extended_int operator+ (const extended_int a, const extended_int b)
{
  extended_int result;
  if (a.n == b.n) {
    result.u = a.u + b.u;
    result.n = a.n;
  }
  else
    if (a.u > b.u) {
      result.u = a.u - b.u;
      result.n = a.n;
    }
    else {
```

```
      result.u = b.u - a.u;
      result.n = b.n;
    }
  return result;
}

// POST: return value is -a
extended_int operator- (extended_int a)
{
  a.n = !a.n;
  return a;
}

// POST: return value is the difference of a and b
extended_int operator- (const extended_int a, const extended_int b)
{
  return a + (-b);
}

// POST: return value is the product of a and b
extended_int operator* (const extended_int a, const extended_int b)
{
  extended_int result;
  result.u = a.u * b.u;
  result.n = (a.n || b.n) && !(a.n && b.n); // XOR
  return result;
}

// POST: a has been written to o
std::ostream& operator<< (std::ostream& o, const extended_int a)
{
  if (a.n) o << "-";
  return o << a.u;
}

// POST: a has been set to i
void set (extended_int& a, const int i)
{
  if (i < 0) {
    a.u = -i; a.n = true;
  } else {
    a.u = i; a.n = false;
  }
}

// now test it
int main() {
  extended_int x;
  extended_int y;
  for (int i = -1; i < 2; ++i)
    for (int j = -1; j < 2; ++j) {
      // set x to i
      set (x, i);
      set (y, j);
      std::cout << "x = " << x << ", y = " << y << "\n";
      std::cout << " x + y = " << x + y << "\n";
      std::cout << " x - y = " << x - y << "\n";
      std::cout << " x * y = " << x * y << "\n";
    }
  return 0;
}
```

Solution to Exercise 133. Let us use the following shortcuts:

E exact match
P promotion
S standard conversion

Here is the table of match qualities for the two parameters.

	A	B	C
a)	(S,S)	**(S,E)**	(S,S)
b)	(S,P)	(E,S)	(S,S)
c)	(E,S)	(S,E)	(S,S)
d)	(S,S)	(S,S)	**(S,E)**
e)	**(S,P)**	(S,S)	(S,S)
f)	(P,E)	(S,S)	(E,S)

A best match is indicated in bold. This implies that a) resolves to B; b),c) are ambiguous; d) resolves to C; e) resolves to A; f) is ambiguous.

Solution to Exercise 134.

```
// Prog: geometry.cpp
// defines basic 3d geometry operations on points (rotations,
// projections), and uses them to rotate a cube

#include <IFMP/window>
#include <iostream>
#include <cmath>

// 2d points
// ---------
struct point2 {
  double x;
  double y;
};

// 3d points
// ---------
struct point3 {
  double x;
  double y;
  double z;
};

// rotations
// ---------
// POST: return value is p rotated around the x-axis by alpha (radians)
point3 rotate_x (const point3 p, const double alpha)
{
  point3 r;
  const double cosa = std::cos(alpha);
  const double sina = std::sin(alpha);
  r.x = p.x;
  r.y = cosa * p.y - sina * p.z;
  r.z = sina * p.y + cosa * p.z;
  return r;
}

// POST: return value is p rotated around the y-axis by alpha (radians)
point3 rotate_y (const point3 p, const double alpha)
{
  point3 r;
```

```
  const double cosa = std::cos(alpha);
  const double sina = std::sin(alpha);
  r.x = sina * p.z + cosa * p.x;
  r.y = p.y;
  r.z = cosa * p.z - sina * p.x;
  return r;
}

// POST: return value is p rotated around the z-axis by alpha (radians)
point3 rotate_z (const point3 p, const double alpha)
{
  point3 r;
  const double cosa = std::cos(alpha);
  const double sina = std::sin(alpha);
  r.x = cosa * p.x - sina * p.y;
  r.y = sina * p.x + cosa * p.y;
  r.z = p.z;
  return r;
}

// projection
// ----------
// PRE: v.z != p.z
// POST: return value is projection of p onto the plane z = 0,
//       with respect to view point v
point2 project (const point3 p, const point3 v)
{
  const double t = p.z / (v.z - p.z);
  point2 r;
  r.x = p.x - t * (v.x - p.x);
  r.y = p.y - t * (v.y - p.y);
  return r;
}

// line drawing
// ------------
// POST: draws a line from the projection of p to the projection
//       of q with respect to viewpoint v, and with offset o
void draw_line (ifmp::Wstream w, const point3 p, const point3 q,
                const point3 v, const point2 o)
{
  const point2 p2 = project (p, v);
  const point2 q2 = project (q, v);
  w << ifmp::Line
    (int(p2.x+o.x), int(p2.y+o.y), int(q2.x+o.x), int(q2.y+o.y));
}

int main()
{
  // construct a cube
  point3 cube[8];
  const int size = 200;
  int i = 0;
  for (double x=-size/2; x <= size/2; x += size)
    for (double y=-size/2; y <= size/2; y += size)
      for (double z=-size/2; z <= size/2; z += size) {
        cube[i].x = x;
        cube[i].y = y;
        cube[i].z = z;
        ++i;
      }

  // open window
  ifmp::Wstream w ("Rotate cube: press x, y, z or q to quit");

```

```
  // drawing loop
  const double alpha = 0.1; // rotation angle
  point3 v;           // viewpoint
  v.x = 200;
  v.y = 200;
  v.z = 1000;
  point2 o;           // offset
  o.x = 250;
  o.y = 250;
  for (;;) {
    // draw all the 12 dube edges
    w.clear();
    draw_line (w, cube[0], cube[1], v, o);
    draw_line (w, cube[0], cube[2], v, o);
    draw_line (w, cube[0], cube[4], v, o);
    draw_line (w, cube[1], cube[3], v, o);
    draw_line (w, cube[1], cube[5], v, o);
    draw_line (w, cube[2], cube[3], v, o);
    draw_line (w, cube[2], cube[6], v, o);
    draw_line (w, cube[3], cube[7], v, o);
    draw_line (w, cube[4], cube[5], v, o);
    draw_line (w, cube[4], cube[6], v, o);
    draw_line (w, cube[5], cube[7], v, o);
    draw_line (w, cube[6], cube[7], v, o);
    w << ifmp::flush;

    // wait for keystroke
    const int key = w.get_key();
    // we assume ASCII encoding
    switch (key) {
    case 'x':
      for (int i=0; i<8; ++i)
        cube[i] = rotate_x (cube[i], alpha);
      break;
    case 'y':
      for (int i=0; i<8; ++i)
        cube[i] = rotate_y (cube[i], alpha);
      break;
    case 'z':
      for (int i=0; i<8; ++i)
        cube[i] = rotate_z (cube[i], alpha);
      break;
    case 'q':
      return 0;
      break;
    default:
      break; // do nothing
    }
  }

  return 0;

}
```

B.11 Reference Types

Solution to Exercise 135.

a) Since i is changed in the function body, S may only be one of int and int&. T can be any of the three types int, int& and const int&, since values of all three types can be initialized from the lvalue ++i of type int.

b) If S is int, then T may only be int, since otherwise, the function returns a reference to a temporary object, namely the local copy of the formal parameter i. If S is int&, T can as before be any of the three types.

c) Here are the postconditions.

```
// POST: return value is i+1
int foo (int i);

// POST: i has been incremented by 1;
//       return value is the new value of i
int foo (int& i);

// POST: i has been incremented by 1 and
//       is returned as an lvalue
int& foo (int& i);

// POST: i has been incremented by 1 and
//       is returned as a non-modifiable lvalue
const int& foo (int& i);
```

Solution to Exercise 136.

```
#include<iostream>

// POST: the values of i and j are swapped
void swap (int& i, int& j)
{
  const int h = i;
  i = j;
  j = h;
}

int main() {
  // input
  std::cout << "i =? ";
  int i; std::cin >> i;

  std::cout << "j =? ";
  int j; std::cin >> j;

  // function call
  swap(i, j);
```

```

  // output
  std::cout << "Values after swapping: i = " << i
            << ", j = " << j << ".\n";

  return 0;
}
```

Solution to Exercise 137. We implement the second version, the one that returns the normalization of r. This one has the advantage that it works for rvalues.

The modification of the function gcd is as easy as it can be: we only need to replace the type unsigned int by int in the parameter and return types. Why does this still work? Let us go back to the proof of Lemma 1. Going through it, we realize that we never used nonnegativity of either a or b, so the statement extends to all pairs of integers with $b \neq 0$. It remains to prove termination. For this, we show that $|a \bmod b| < |b|$, so we indeed make progress towards termination.

Recall that

$$a \bmod b = a - (a \operatorname{div} b)b,$$

and that this equation also holds in C++. Furthermore, division may round up or down (we don't know), but in either case, the rounding makes a mistake of strictly less than 1. This means that

$$\frac{a}{b} - (a \operatorname{div} b)$$

has absolute value smaller than 1, and this implies (by multiplying with b) that

$$a - (a \operatorname{div} b)b = a \bmod b$$

has absolute value smaller than $|b|$.

```
#include <iostream>

struct rational {
  int n;
  int d; // INV: d != 0
};

// POST: a has been written to o
std::ostream& operator<< (std::ostream& o, const rational& a)
{
  return o << a.n << "/" << a.d;
}

// POST: a has been read from i
// PRE:  i starts with a rational number of the form "n/d"
std::istream& operator>> (std::istream& i, rational& a)
{
  char c; // separating character, e.g. '/'
  return i >> a.n >> c >> a.d;
}

```

```
// POST: return value is the greatest common divisor of a and b
int gcd (const int a, const int b)
{
  if (b == 0) return a;
  return gcd(b, a % b);  // b != 0
}

// POST: return value is the normalization of r
rational normalize (const rational& r)
{
  const int g = gcd (r.n, r.d);
  rational result;
  result.n = r.n / g;
  result.d = r.d / g;
  if (result.d < 0) {
    result.n = -result.n;
    result.d = -result.d;
  }
  return result;
}

int main ()
{
  std::cout << "rational number r =? ";
  rational r;
  std::cin >> r;
  std::cout << "Normalization is " << normalize(r) << ".\n";

  return 0;
}
```

Solution to Exercise 138.

```
#include <iostream>

// POST: return value indicates whether the linear equation
//       a * x + b = 0 has a real solution x ; if true is
//       returned, the value s satisfies a * s + b = 0
bool solve (const double a, const double b, double& s)
{
  // we have a solution if a is nonzero (s = -b/a),
  // or if both a and b are zero (take s = 0 in this case)
  if (a != 0.0) {
    s = -b / a;
    return true;
  }
  // now we have a == 0.0
  if (b == 0.0) {
    s = 0.0;
    return true;
  }
  return false;
}

int main()
{
  std::cout << "solve a * x + b = 0 for\n";
  std::cout << "a =? ";
  double a;
  std::cin >> a;
  std::cout << "b =? ";
  double b;
  std::cin >> b;
```

```

  double s;
  if (solve (a, b, s))
    std::cout << "Solution is " << s << ".\n";
  else
    std::cout << "Sorry, there is no solution.\n";

  return 0;
}
```

Solution to Exercise 139.

a) Problem: Initialization of non-const reference from const object. (The variable `i` is const-qualified and can, therefore, not be passed as a non-const reference argument to the function `foo`.)

   Problem: Initialization of non-const reference of type "`int&`" from an rvalue of type "`int`". We can fix this in two ways. Either make `j` `const int&` or `const int`.

b) ok. (The variable `j` is a const reference and may, therefore, be initialized from a temporary.)

c) Problem: Initialization of reference from temporary. (The return value of the function `foo` is of type `int` and, therefore, the corresponding object has temporary lifetime. Via the function `bar` that temporary is used to initialize the variable `j`.)

d) Problem: Initialization of non-const reference from const reference. (The function `bar` returns a const reference that cannot be passed as a non-const reference to the function `foo`.)

e) ok. (Remark: Does not violate the Single Modification Rule because there is a sequence point after all function arguments have been evaluated.)

Solution to Exercise 140.

```
// Prog: solve_quadratic_equation.cpp
// computes both (possibly complex) solutions to a quadratic equation
#include<iostream>
#include<complex>

// POST: return value is the number of distinct complex solutions
//       of the quadratic equation ax^2 + bx + c = 0. If there
//       are infinitely many solutions (a=b=c=0), the return
//       value is -1. Otherwise, the return value is a number n
//       from {0,1,2}, and the solutions are written to s1,..,sn
int solve_quadratic_equation (const std::complex<double> a,
                              const std::complex<double> b,
                              const std::complex<double> c,
                              std::complex<double>& s1,
                              std::complex<double>& s2)
{
  if (a == 0.0)
    // linear case: bx + c = 0
```

```
    if (b == 0.0)
      // trivial case: c = 0
      if (c == 0.0)
        return -1; // => infinitely many solutions
      else
        return 0;  // => no solution
    else {
      // bx + c = 0, b != 0   => one solution
      s1 = -c/b;
      return 1;
    }
  else {
    const std::complex<double> d = std::sqrt(b*b-4.0*a*c);
    if (d == 0.0) {
      // ax^2 + bx + c = 0, a != 0, d==0 => one solution
      s1 = -b / (2.0*a);
      return 1;
    } else {
      // ax^2 + bx + c = 0, a != 0, d!=0 => two solutions
      s1 = (-b + d) / (2.0*a);
      s2 = (-b - d) / (2.0*a);
      return 2;
    }
  }
}

// POST: solves the quadratic equation ax^2 + bx + c = 0 and outputs
//       the number of solutions as well as the solutions themselves
void test (const std::complex<double> a,
           const std::complex<double> b,
           const std::complex<double> c)
{
  std::complex<double> s1;
  std::complex<double> s2;

  const int n = solve_quadratic_equation (a, b, c, s1, s2);

  // output
  std::cout << "Number of solutions: " << n << "\n";
  std::cout << "Solutions:\n";
  if (n > 0) std::cout << " " << s1 << "\n";
  if (n > 1) std::cout << " " << s2 << "\n";
}

int main()
{
  test (0, 0, 0);
  // 0 = 0:              infinitely many solutions
  test (0, 0, 2);
  // 2 = 0:              no solution
  test (0, 2, 2);
  // 2x + 2 = 0:         one solution  (x=-1)
  test (2, 2, 2);
  // 2x^2 + 2x + 2 = 0:  two solutions (x=-1/2 +- sqrt(3)/2 i)
  test (1, 2, 1);
  //  x^2 + 2x + 1 = 0:  one solution  (x=-1)
  test (std::complex<double>(0,1), 1, 1);
  //  ix^2 + x + 1 = 0:  two solutions

  return 0;
}
```

We asked you to test your function for the triples (a, b, c) from the set

$$\{(0,0,0), (0,0,2), (0,2,2), (2,2,2), (1,2,1), (i,1,1)\}.$$

The output to these inputs should be:

- Input: $(0,0,0)$, number of solutions: infinitely many (`-1`)
- Input: $(0,0,2)$, number of solutions: `0`
- Input: $(0,2,2)$, number of solutions: `1`,
  solutions: `s1==(-1,0)`
- Input: $(2,2,2)$, number of solutions: `2`,
  solutions: `s1==(-0.5,0.866025)`, `s2==(-0.5,-0.866025)`
- Input: $(1,2,1)$, number of solutions: `1`,
  solutions: `s1==(-1,0)`
- Input: $(i,1,1)$, number of solutions: `2`,
  solutions: `s1==(-0.624811,-0.300243)`, `s2==(0.624811,1.30024)`

Solution to Exercise 141. We want to find all complex solutions to the equation

$$Ax^3 + Bx^2 + Cx + D = 0, \tag{B.3}$$

where $A, B, C, D \in \mathbb{C}$, and $A \neq 0$. The following method due to Scipione del Ferro and Tartaglia was published by Gerolamo Cardano in 1545. The method transforms the equation into an equivalent one that we can solve directly.

We first divide equation (B.3) by the leading coefficient A to arrive at an equivalent equation of the following form

$$x^3 + bx^2 + cx + d = 0, \tag{B.4}$$

where $b = B/A$, $c = C/A$, and $d = D/A$ are complex numbers.

Let us substitute $x = y - \frac{b}{3}$ into equation (B.4). Then the quadratic term disappears, and we get the equivalent equation

$$y^3 + 3qy - 2r = 0, \tag{B.5}$$

where

$$q = \frac{3c - b^2}{9}, \tag{B.6}$$

$$r = \frac{9bc - 27d - 2b^3}{54}. \tag{B.7}$$

You are, of course, invited to check that for yourself. The left-hand side of (B.5) is called a *depressed* cubic. If $q = 0$, we are done already, since the solutions to $y^3 - 2r = 0$ are just the three complex third roots of of $2r$.

Let us assume for the remainder that $q \neq 0$. Then we further massage equation (B.5) by making the substitution $y = z - \frac{q}{z}$ and multiplying with z^3 on both sides of the equation. Then we arrive at

$$z^6 - 2rz^3 - q^3 = 0, \tag{B.8}$$

which can be viewed as a quadratic equation of the unknown variable z^3,

$$(z^3)^2 - 2r(z^3) - q^3 = 0. \tag{B.9}$$

We know how to solve (B.9) for z^3 (and then also for z), but before doing this, let's take a step back and see whether this really solves our problem: we know that x solves (B.3) if and only if $y = x + \frac{b}{3}$ solves (B.5). Moreover, let y and $z \neq 0$ be such that $y = z - \frac{q}{z}$. Then y solves (B.5) if and only if z solves (B.9). This means, every solution $z \neq 0$ to (B.9) gives us a solution $y = z - \frac{q}{z}$ to (B.5) and thus a solution $x = y - \frac{b}{3}$ to (B.3). Vice versa, every solution x to (B.3) gives us a solution $y = x + \frac{b}{3}$ to (B.5) and thus two nonzero solutions $z = \frac{y}{2} \pm \sqrt[2]{\frac{y^2}{4} + q}$ to (B.9), using $q \neq 0$.

Note: In writing $\sqrt[k]{c}$ for a complex number c, we choose one of the k roots arbitrarily, but at the same time we need to make sure that the choice does not matter. Indeed, the set of *two* numbers $\pm\sqrt[2]{\frac{y^2}{4} + q}$ does not depend on the particular choice of the square root.

To summarize, our original problem is solved by finding the (nonzero) solutions to (B.9), so let's turn to this latter problem.

We use the solution formula for quadratic equations in order to obtain

$$z^3 = r \pm \sqrt[2]{r^2 + q^3}, \tag{B.10}$$

$$z = \triangleright\sqrt[3]{r \pm \sqrt[2]{r^2 + q^3}}. \tag{B.11}$$

Here, $\triangleright c$, for some number c, stands for one of the three numbers $c = c\xi_0, c\xi_1, c\xi_2$ where the ξ_i's are the third roots of unity, i.e. the three complex solutions to the equation $u^3 = 1$:

$$\xi_0 = 1, \quad \xi_1 = -\frac{1}{2} + \frac{\sqrt{3}}{2}i, \quad \xi_2 = -\frac{1}{2} - \frac{\sqrt{3}}{2}i.$$

Note that

$$\xi_i \xi_j = \xi_{i+j}, \tag{B.12}$$

where indices are taken modulo 3.

Equation (B.11) specifies a set of 6 values. Fixing our particular choices of roots, we can number them as

$$p_i = \xi_i \sqrt[3]{r + \sqrt[2]{r^2 + q^3}}, \quad i = 0, 1, 2, \tag{B.13}$$

$$n_i = \xi_i \sqrt[3]{r - \sqrt[2]{r^2 + q^3}}, \quad i = 0, 1, 2. \tag{B.14}$$

We want to argue that we only need to consider the three solutions $z = p_i$ in order to get *all* the solutions $y = z - \frac{q}{z}$ to (B.5). For this, we first observe that the set $\{p_i n_i \mid i = 0, 1, 2\}$ is the set of third roots of $-q^3$. Indeed, by

$$p_i^3 n_i^3 = (r + \sqrt[2]{r^2 + q^3})(r - \sqrt[2]{r^2 + q^3}) = -q^3, \quad i = 0, 1, 2 \tag{B.15}$$

all three numbers have third power $-q^3$, and using (B.12) on top of (B.13) and (B.14), we can show that they are distinct:

$$p_i n_i = \xi_i^2 \; p_0 n_0 = \xi_{2i} \; p_0 n_0, \quad i = 0, 1, 2 \; (\Rightarrow 2i = 0, 2, 1).$$

Consequently, there is an index $t \in \{0, 1, 2\}$ for which $p_t n_t = -q$, one of the third roots of $-q^3$. Having this, (B.13) and (B.14) together with (B.12) imply

$$p_t n_t = p_{t+1} n_{t+2} = p_{t+2} n_{t+1} = -q, \tag{B.16}$$

with indices taken modulo 3 again.

This in turn yields

$$\left(p_t - \frac{q}{p_t}, p_{t+1} - \frac{q}{p_{t+1}}, p_{t+2} - \frac{q}{p_{t+2}}\right) = \left(n_t - \frac{q}{n_t}, n_{t+2} - \frac{q}{n_{t+2}}, n_{t+1} - \frac{q}{n_{t+1}}\right), \tag{B.17}$$

meaning that in $y = z - \frac{q}{z}$, it does not matter whether we use $z = p_1, p_2, p_3$ or $z = n_1, n_2, n_3$.

That means we are done. We have found the three solutions

$$x_i = p_i - \frac{q}{p_i} - \frac{b}{3}, \quad i = 0, 1, 2$$

to equation (B.3), where p_i runs through the third roots of

$$r + \sqrt[2]{r^2 + q^3}$$

for $i = 0, 1, 2$.

The method described above is exactly the method implemented in the C++ program below.

```
// Prog: solve_cubic_equation.cpp
// computes the three complex solutions to a cubic equation
#include<iostream>
#include<complex>


// POST: return value is the number of distinct complex solutions
//       of the quadratic equation ax^2 + bx + c = 0. If there
//       are infinitely many solutions (a=b=c=0), the return
//       value is -1. Otherwise, the return value is a number n
//       from {0,1,2}, and the solutions are written to s1,..,sn
int solve_quadratic_equation (const std::complex<double> a,
                              const std::complex<double> b,
                              const std::complex<double> c,
                              std::complex<double>& s1,
                              std::complex<double>& s2)
{
  if (a == 0.0)
    // linear case: bx + c = 0
    if (b == 0.0)
      // trivial case: c = 0
      if (c == 0.0)
        return -1; // => infinitely many solutions
      else
        return 0;  // => no solution
    else {
      // bx + c = 0, b != 0   => one solution
      s1 = -c/b;
      return 1;
    }
  else {
    // ax^2 + bx + c = 0, a != 0 => two solutions
    const std::complex<double> d = std::sqrt(b*b-4.0*a*c);
    s1 = (-b + d) / (2.0*a);
    s2 = (-b - d) / (2.0*a);
    return 2;
  }
}

// POST: return value is the number of distinct complex solutions
//       of the cubic equation ax^3 + bx^2 + cx + d = 0. If there
//       are infinitely many solutions (a=b=c=d=0), the return
//       value is -1. Otherwise, the return value is a number n
//       from {0,1,2,3}, and the solutions are written to s1,..,sn
int solve_cubic_equation (const std::complex<double> a,
            const std::complex<double> b,
            const std::complex<double> c,
            const std::complex<double> d,
            std::complex<double>& s1,
            std::complex<double>& s2,
            std::complex<double>& s3)
{
  if (a == 0.0)
    // if a == 0 we can use the preceding function to solve
    // quadratic equations.
        return solve_quadratic_equation (b,c,d,s1,s2);
  else {
    // ax^3 + bx^2 + cx + d = 0, a != 0 => three solutions.
    // Some of the solutions may be equal due to multiple roots.

    // First we bring the equation  into the following form
    // x^3 + bn x^2 + cn x + dn = 0
    const std::complex<double> bn = b/a;
    const std::complex<double> cn = c/a;
```

```
    const std::complex<double> dn = d/a;

    // compute q and r
    const std::complex<double> q = (3.0*cn - bn*bn) / 9.0;
    const std::complex<double> r = (9.0*bn*cn - 27.0*dn - 2.0*bn*bn*bn) / 54.0;

    // define non-real root of unity
    const std::complex<double> rou(-0.5, std::sqrt(3.0)/2.0);
        // the variable temp is used to convert solutions y to x
    const std::complex<double> temp = bn/3.0;

    // check for the special case q == 0
    if (q == 0.0) {
      // compute        y, as in y^3 = 2r
      // substitution: x = y - b/3
      std::complex<double> y = std::pow(2.0*r, 1.0/3.0);

      // compute solutions for x
      s1 = y - temp;
      y *= rou; // rotate complex number y by -120 degrees
      s2 = y - temp;
      y *= rou; // rotate y by another -120 degrees
      s3 = y - temp;
    }
    else {
      // compute z, as in z^3 = r + sqrt(r^2 + q^3),
      // substitution 2: x = z - q/z - b/3
      std::complex<double> z = std::pow(r + std::sqrt(r*r + q*q*q), 1.0/3.0);

      // compute solutions for x
      s1 = z - q/z - temp;
      z *= rou; // rotate complex number z by -120 degrees
      s2 = z - q/z - temp;
      z *= rou; // rotate z by another -120 degrees
      s3 = z - q/z - temp;
    }

    return 3;
  }
}


// POST: Returns the magnitude of the complex number
// as^3 + bs^2 + cs + d. If s is an exact solution
// to the equation ax^3 + bx^2 + cx + d = 0 then this
// magnitude should be 0.
double check_solution(const std::complex<double> a,
            const std::complex<double> b,
            const std::complex<double> c,
            const std::complex<double> d,
            const std::complex<double> s) {

  // compute the ax^3 + bx^2 + cx + d and return its magnitude
  return std::abs(a*std::pow(s,3.0) + b*s*s + c*s + d);
}


int main()
{
  // input
  std::cout << "Solve cubic equation ax^3 + bx^2 + cx + d = 0 for\n";
  std::cout << "a =? ";
  std::complex<double> a;
  std::cin >> a;
  std::cout << "b =? ";
```

```
130     std::complex<double> b;
131     std::cin >> b;
132     std::cout << "c =? ";
133     std::complex<double> c;
134     std::cin >> c;
135     std::cout << "d =? ";
136     std::complex<double> d;
137     std::cin >> d;
138
139     // computation
140     std::complex<double> s1;
141     std::complex<double> s2;
142     std::complex<double> s3;
143
144     const int n = solve_cubic_equation (a, b, c, d, s1, s2, s3);
145
146     // output
147     std::cout << "Number of solutions: " << n << "\n";
148     std::cout << "Solutions:\n";
149     if (n > 0) {
150       std::cout << s1 << ", ";
151       std::cout << "Error: " << check_solution(a,b,c,d,s1) << "\n";
152     }
153     if (n > 1){
154       std::cout << s2 << ", ";
155       std::cout << "Error: " << check_solution(a,b,c,d,s2) << "\n";
156     }
157     if (n > 2){
158       std::cout << s3 << ", ";
159       std::cout << "Error: " << check_solution(a,b,c,d,s3) << "\n";
160     }
161
162     return 0;
163
164   }
```

B.12 Classes

Solution to Exercise 142. Here are the header and implementation files.

```
// Prog: rational.h
// Define a class for rational numbers
#include <iostream>
#include <cassert>
#include <IFMP/integer.h>

namespace math {

  // rationals
  class rational {
  public:
    // nested types
    typedef ifmp::integer rat_int;

    // public members
    rational();
    // POST: *this is initialized with value 0

    rational(const int n, const int d);
    // PRE: d != 0
    // POST: *this is initialized with value n / d

    rational(const int n);
    // POST: *this is initialized with value n

    rat_int n() const;
    // POST: return value is the numerator n

    rat_int d() const;
    // POST: return value is the denominator d

    rational& operator+= (const rational& s);
    // POST: *this has been incremented by s, and
    //       the return value is the new value

    rational& operator-= (const rational& s);
    // POST: *this has been decremented by s, and
    //       the return value is the new value

    rational& operator*= (const rational& s);
    // POST: *this has been multiplied by s, and
    //       the return value is the new value

    rational& operator/= (const rational& s);
    // POST: *this has been divided by s, and
    //       the return value is the new value

  private:
    rat_int n_;  // numerator
    rat_int d_;  // denominator; invariant: d_ != 0

    // POST: gcd(_n,_d) is 1 or -1
    void normalize();
  };

  // non-member operators
  rational operator+(const rational& r, const rational& s);
  rational operator-(const rational& r, const rational& s);
  rational operator*(const rational& r, const rational& s);
```

```
  rational operator/(const rational& r, const rational& s);
  bool operator==(const rational& r, const rational& s);
  bool operator!=(const rational& r, const rational& s);
  bool operator<(const rational& r, const rational& s);
  bool operator>(const rational& r, const rational& s);
  bool operator<=(const rational& r, const rational& s);
  bool operator>=(const rational& r, const rational& s);
  std::ostream& operator<<(std::ostream& o, const rational& r);
  std::istream& operator>>(std::ostream& i, rational& r);

}
```

```
// Prog: rational.cpp
// Define a class for rational numbers
#include <iostream>
#include <cassert>
#include <IFMP/rational.h>

namespace math {
  // global gcd function
  rational::rat_int gcd(const rational::rat_int& a,
                        const rational::rat_int& b)
  {
    // PRE: (a,b) != (0,0)
    // POST: return value is the greatest common divisor
    //       of a and b, where gcd(n,0):=gcd(0,n):=n.
    if (b == 0) return a;
    return gcd(b, a % b);
  }

  rational::rational()
    : n_(0), d_(1)
  {}

  rational::rational(const int n, const int d)
    : n_(n), d_(d)
  {
    assert (d != 0);
  }

  rational::rational(int n)
    : n_(n), d_(1)
  {}

  rational::rat_int rational::n() const { return n_; }
  rational::rat_int rational::d() const { return d_; }

  rational& rational::operator+= (const rational& s)
  {
    n_ = n_ * s.d_ + d_ * s.n_;   // increment *this by s
    d_ *= s.d_;
    normalize();
    return *this;                 // and return it
  }

  rational& rational::operator-= (const rational& s)
  {
    n_ = n_ * s.d_ - d_ * s.n_;   // decrement *this by s
    d_ *= s.d_;
    normalize();
    return *this;                 // and return it
  }

```

```
rational& rational::operator*= (const rational& s)
{
  n_ *= s.n_;    // multiply *this by s
  d_ *= s.d_;
  normalize();
  return *this;  // and return it
}

rational& rational::operator/= (const rational& s)
{
  n_ *= s.d_;    // multiply *this by 1/s
  d_ *= s.n_;
  normalize();
  return *this;  // and return it
}

// normalization
void rational::normalize()
{
  const rat_int g = gcd(n_,d_);
  n_ /= g;
  d_ /= g;
}

// non-member arithemtic operators
rational operator+(const rational& r, const rational& s)
{
  rational t = r;
  return t += s;
}

rational operator-(const rational& r, const rational& s)
{
  rational t = r;
  return t -= s;
}

rational operator*(const rational& r, const rational& s)
{
  rational t = r;
  return t *= s;
}

rational operator/(const rational& r, const rational& s)
{
  rational t = r;
  return t /= s;
}

// non-member relational operators
bool operator==(const rational& r, const rational& s)
{
  return r.n() * s.d() ==  s.n() * r.d();
}

bool operator!=(const rational& r, const rational& s)
{
  return !(r == s);
}

bool operator<(const rational& r, const rational& s)
{
  const rational::rat_int a = r.n() * s.d();
  const rational::rat_int b = s.n() * r.d();
  // watch out for the signs of denominators!
```

```
    if (r.d() > 0)
      if (s.d() > 0) return a < b; else return a > b;
    else
      if (s.d() > 0) return a > b; else return a < b;
  }

  bool operator>(const rational& r, const rational& s)
  {
    return s < r;
  }

  bool operator<=(const rational& r, const rational& s)
  {
    return r == s || r < s;
  }

  bool operator>=(const rational& r, const rational& s)
  {
    return r == s || s < r;
  }

  std::ostream& operator<<(std::ostream& o, const rational& r)
  {
    return o << r.n() << "/" << r.d();
  }

  std::istream& operator>> (std::istream& i, rational& a)
  {
    char c; // separating character, e.g. '/'
    int n;
    int d;
    i >> n >> c >> d;
    a = rational(n,d);
    return i;
  }

}
```

And here is a test program: it computes harmonic numbers backwards and forwards, showing that there is no difference in the two sums if rational numbers are used for the computations.

```
// Program: harmonic.cpp
// Compute the n-th harmonic number in two ways.

#include <iostream>
#include <IFMP/rational.h>

int main()
{
  // Input
  std::cout << "Compute H_n for n =? ";
  unsigned int n;
  std::cin >> n;

  // Forward sum
  math::rational fs (0, 1);
  for (unsigned int i = 1; i <= n; ++i)
    fs += math::rational (1, i);

  // Backward sum
  math::rational bs (0, 1);
```

```
  for (unsigned int i = n; i >= 1; --i)
    bs += math::rational (1, i);

  // Output
  std::cout << "Forward sum  = " << fs << "\n"
            << "Backward sum = " << bs << "\n";
  return 0;
}
```

Solution to Exercise 145.

```
#include <iostream>
#include <cassert>

// Class for representing time in hours : minutes : seconds
class Clock {
public: // error fix: public specifier was missing
  // PRE: h < 24, m < 60, s < 60
  // POST: *this is initialized with the time h:m:s
  Clock(const unsigned int h, const unsigned int m, const unsigned int s);
  // make it const-correct!
  // error fix: no reference types here!


  // POST: time stored by *this has been advanced by 1 second
  void tick();

  // POST: h, m, s are filled from the time stored by *this
  void time(unsigned int& h, unsigned int& m,
            unsigned int& s) const; // error fix: we need reference types!

private:
  unsigned int h_;
  unsigned int m_;
  unsigned int s_;
};

Clock::Clock(const unsigned int h,
             const unsigned int m,
             const unsigned int s)
    : h_(h), m_(m), s_(s)
{
  assert (h < 24);
  assert (m < 60);
  assert (s < 60); // error fix: enforce preconditions
}

void Clock::tick()
{
  h_ += (m_ += (s_ += 1) / 60) / 60;
  h_ %= 24; m_ %= 60; s_ %= 60;
}

void Clock::time(unsigned int& h,
                 unsigned int& m,
                 unsigned int& s) const
{
  h = h_;
  m = m_;
  s = s_;
}

std::ostream& operator<< (std::ostream& o, const Clock c)
```

```
{
  unsigned int h;
  unsigned int m;
  unsigned int s;
  c.time(h, m, s);
  o << h << ":";
  if (s < 10) o << "0";
  o << m << ":";
  if (s < 10) o << "0";
  o << s;
  return o;
}

int main() {
  Clock c1 (23, 9, 8);
  c1.tick(); // error fix: tick has to be called for an object

  unsigned int h;
  unsigned int m;
  unsigned int s;
  c1.time(h, m, s); // error fix: time has to be called for an object

  std::cout << c1 << "\n";

  return 0;
}
```

Solution to Exercise 146.

```
// Program: random_triangle.cpp
//
// A graphical random process:
//
//   * Start at an arbitrary vertex of a triangle t
//
//   * At each step, draw the current point p and choose
//     as a next point the midpoint of p and a (uniformly)
//     randomly selected vertex of t.

#include <iostream>
#include <IFMP/random.h>
#include <IFMP/window>

int main()
{
  std::cout << "Seed for random number generator =? ";
  unsigned int seed;
  std::cin >> seed;
  // use the ANSIC generator
  ifmp::random rnd(1103515245u, 12345u, 2147483648u, seed);

  std::cout << "Number of steps =? ";
  unsigned int n;
  std::cin >> n;

  // the vertices of the triangle
  const ifmp::Point v0(0, 0);
  const ifmp::Point v1(512, 0);
  const ifmp::Point v2(256, 512);

  ifmp::Point cur = v0; // current point
  for (unsigned int i = 0; i < n; ++i) {
    // output current point
```

```
      ifmp::wio << cur;
      // choose (uniformly) a random vertex of the triangle
      ifmp::Point rndvert;
      int r = int(rnd() * 3); // random integer in [0,2]
      if (r == 0) rndvert = v0;
      else if (r == 1) rndvert = v1;
      else rndvert = v2;
      // jump halfway to the chosen vertex
      cur = ifmp::Point((cur.x() + rndvert.x()) / 2,
                        (cur.y() + rndvert.y()) / 2);
    }
    ifmp::wio.wait_for_mouse_click();
    return 0;
  }
```

Solution to Exercise 147. If `unsigned int` computations overflow, they will still be correct modulo 2^{32} on a 32-bit system (or any other system with at least this many bits); this is guaranteed by the C++ standard according to which `unsigned int` behaves like the ring $Z_{2^{32}}$ under additions and multiplications. This means, if u is the mathematically correct value, the computed value is of the form $u + k2^{32}$, where k is the unique integer such that $u + k2^{32} \in \{0, \ldots, 2^{32} - 1\}$.

But now it is easy to see that taking u modulo 2^{31} gives the same result as taking $u + k2^{32}$ modulo 2^{31}, simply because $2^{32} \bmod 2^{31} = 0$.

Solution to Exercise 148. The opponent plays with a fair dice, i.e. he/she plays every number with an equal probability of $\frac{1}{6}$. The goal is to beat the fair dice. Let us for example consider the strategy of always choosing 1. Then our *average* gain (playing against the fair dice) can be computed from a table that lists what we get, depending on the number chosen by the opponent.

opponent's number	1	2	3	4	5	6
our gain (CHF) in choosing 1	0	2	-1	-1	-1	-1

Since the opponent plays each number with probability 1/6, our average gain in choosing number 1 is

$$\frac{1}{6}(0 + 2 - 1 - 1 - 1 - 1) = \frac{-2}{6} = -\frac{1}{3}.$$

Not good.

But we can do the same for all other numbers and then end up with the following table.

our number	1	2	3	4	5	6
our average gain (CHF)	$-\frac{1}{3}$	$-\frac{1}{2}$	$-\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Always choosing number 5 is therefore the best against a fair dice. This way, after n rounds of the game, we will have earned CHF $\frac{n}{2}$ on average.

In reality, our opponent would not be so stupid to stick with the fair dice if we always choose 5; he could switch to always choosing 4, making us lose in every round. But that's ok: the goal was to beat the (stupid) fair dice, and that's what we achieve by choosing 5.

Solution to Exercise 149. Here is a strategy that cannot lose in the long run, no matter what the opponent does. Choose the numbers according to the probability distribution

$$(p_1, p_2, p_3, p_4, p_5, p_6) = \left(0, \frac{1}{16}, \frac{5}{16}, \frac{1}{4}, \frac{5}{16}, \frac{1}{16}\right). \tag{B.18}$$

Now let's see what happens when the *opponent* plays some fixed number i. If we can show that our average gain is nonnegative, regardless of the opponent's number, then we have shown that we cannot lose in the long run. If the opponent plays 1, for example, we arrive at the following table.

our number	1	2	3	4	5	6
our gain (CHF) when opponent chooses 1	0	-2	1	1	1	1

Since we play as in (B.18), our average gain against the opponent choosing 1 is

$$0 \cdot 1 - \frac{1}{16} \cdot 2 + \frac{5}{16} \cdot 1 + \frac{1}{4} \cdot 1 + \frac{5}{16} \cdot 1 + \frac{1}{16} \cdot 1 = \frac{13}{16}.$$

Good.

We can do the same thing for all other numbers played by the opponent and then end up with the following (surprising) table.

opponent's number	1	2	3	4	5	6
our average gain (CHF)	$\frac{13}{16}$	0	0	0	0	0

Perfect! No matter which number the opponent chooses, we always earn at least CHF 0 on average. This means that we have found an unbeatable loaded dice.

Solution to Exercise 150. Let us start with the test program `xbm_merge.cpp`. It includes `transparency1.xbm` and `transparency2.xbm` and writes an `XBM` file (with header `overlay`) to standard output. **Note:** You have to recompile the program everytime you want to test with a different `XBM` input file!

```
#include<iostream>
#include<cassert>
#include "transparency1.xbm"
#include "transparency2.xbm"
#define width transparency1_width
#define height transparency1_height

int main()
{

  // array for mapping from {0,...,15} to {'0',...,'f'} for output
```

```
  char hex[] = {'0', '1', '2', '3', '4', '5', '6', '7',
                '8', '9', 'a', 'b', 'c', 'd', 'e', 'f'};

  // output header for merged image
  std::cout << "#define overlay_width "    << width << "\n";
  std::cout << "#define overlay_height "   << height << "\n";
  std::cout << "static unsigned char bits[] = {\n";

  bool comma = false;
  for (int i=0; i<sizeof(transparency1_bits); ++i) {
    unsigned char outbyte = 0;
    unsigned char transparency1 = transparency1_bits[i];
    unsigned char transparency2 = transparency2_bits[i];
    unsigned char fac=1;
    for (int j=0; j<8; ++j) {
      if ((transparency1 % 2) || (transparency2 % 2))
        outbyte += fac;
      transparency1/=2;
      transparency2/=2;
      fac *= 2;
    }
    if (comma) std::cout << ","; comma = true;
    std::cout << "0x" << hex[outbyte / 16] << hex[outbyte % 16];
  }
  std::cout << "};\n";
  return 0;
}
```

Now the actual encrypting program `visual_crypto.cpp`. It includes `original.xbm` and writes an XBM file with header `transparency1` to standard output, if called with `visual_crypto 1`. The call `visual_crypto 2` entails header `transparency2`. As with `xbm_merge.cpp`. you have to recompile the program everytime you want to test with a different XBM input file.

```
// Prog: visual_crypto.cpp
// includes an xbm image file and outputs two "random" xbm-files whose
// overlay corresponds to the original image; every pixel is replaced
// by four pixels in each image

#include<iostream>
#include<cassert>
#include<string>
#include<IFMP/random.h>
#include "original.xbm"
#define inwidth original_width
#define inheight original_height
#define inbits original_bits

// POST: returns ", ", unless called for the first time where we get ""
std::string comma() {
  static bool first_time = true;
  if (!first_time)
    return ", ";
  else {
    first_time = false;
    return "";
  }
}

// POST: return value is the hex-code of i for XBM output.
//       For example: 30 = 1*16 + 14 -> 0x1e
```

```
28  std::string hex (const unsigned char i) {
29    static std::string h[] = {"0", "1", "2", "3", "4", "5", "6", "7",
30                              "8", "9", "a", "b", "c", "d", "e", "f"};
31    return "0x" + h[i / 16] + h[i % 16];
32  }
33
34  // POST: computes the two-pixel replacement (01 = 1 or 10 = 2)
35  // of pixel p, for row 0 in image i, under (random) bit b, according
36  // to the following table. The replacement for row 1 is the reverse
37  //   bit              0 | 1            0 | 1
38  // ---------------------------------------------------
39  //   image 0:  0 -> 10 | 01;   1 -> 10 | 01   <- row 0
40  //                  01 | 10;        01 | 10   <- row 1
41  //
42  //   image 1:  0 -> 10 | 01;   1 -> 01 | 10   <- row 0
43  //                  01 | 10;        10 | 01   <- row 1
44  // ---------------------------------------------------
45  // => overlay: 0 -> 10 | 01;   1 -> 11 | 11   <- row 0
46  //                  01 | 10;        11 | 11   <- row 1
47  //                  (gray)          (black)
48  unsigned char split_pixel (const bool p, const bool i, const bool b)
49  {
50    if (p && i) return b ? 2 : 1; else return b ? 1 : 2;
51  }
52
53  // POST: computes the four replacement bytes for byte, for rows 0 and 1
54  //       in image i, using some fixed random number generator.
55  //       The replacement happens pixel by pixel, according to split_pixel.
56  //       The replacement bytes are written to outbyte01, outbyte02 for
57  //       row 0 and outbyte11, outbyte12 for row 1.
58  void split_byte (unsigned char byte, bool i,
59                   unsigned char& outbyte01, // for the first four pixels, row 0
60                   unsigned char& outbyte02, // for the last four pixels, row 0
61                   unsigned char& outbyte11, // for the first four pixels, row 1
62                   unsigned char& outbyte12) // for the last four pixels, row 1
63  {
64    // the random number generator; it is important that the sequence of
65    // random numbers is the same for every instance, since image 1 and 2
66    // need to be correlated
67    static ifmp::random ansic (1103515245u, 12345u, 2147483648u, 12345u);
68
69    // go through the pixels and build the outbytes
70    unsigned char outbyte[4];
71    for (int k=0; k<2; ++k) {
72      outbyte[k] = outbyte[k+2] = 0; // 0/2: first bytes, 1/3: second bytes
73      unsigned char power4 = 1;      // for prepending 01 or 10 to outbyte
74      for (int j=0; j<4; ++j) {
75        const bool p = byte % 2;           // last bit = next pixel
76        // prepend replacement bits to outbytes
77        const unsigned char split =  split_pixel (p, i, ansic()<0.5);
78        outbyte[k]   += power4 * split;     // row 0
79        outbyte[k+2] += power4 * (3-split); // row 1: 1 <-> 2
80        // shift power for next round
81        power4 *= 4;
82        // chop off processed bit
83        byte /= 2;
84      }
85    }
86    outbyte01 = outbyte[0];
87    outbyte02 = outbyte[1];
88    outbyte11 = outbyte[2];
89    outbyte12 = outbyte[3];
90  }
91
92  int main(int argc, char* argv[])
```

```
{
  if (argc < 2) {
    // no command line arguments (except program name)
    std::cout << "Usage: visual_crypto n (where n is 1 or 2)\n";
    return 1;
  }

  // read number of image to be generated (second command line argument)
  char* const s = argv[1];
  bool i;    // image number
  if      (*s == '1') i = 0;
  else if (*s == '2') i = 1;
  else {
    std::cerr << "Illegal argument (should be 0 or 1)\n";
    return 1;
  }

  // output header for encrypted image, containing number of image
  std::cout << "#define transparency"
            << *s << "_width "  << 2*inwidth << "\n";
  std::cout << "#define transparency"
            << *s << "_height " << 2*inheight << "\n";
  std::cout << "static unsigned char transparency"
            << *s << "_bits[] = {\n";

  const int bytes_per_inrow =  (inwidth+7)/8;
  const int bytes_per_outrow =
    2 * bytes_per_inrow - (inwidth % 8 > 0 && inwidth % 8 <= 4);
  // go through the input image row by row;
  for (int row = 0; row < inheight; ++row) {
    // compute the two replacement rows
    unsigned char outbytes0[2*bytes_per_inrow]; // replacement bytes, row 0
    unsigned char outbytes1[2*bytes_per_inrow]; // replacement bytes, row 1
    for (int b = 0; b < bytes_per_inrow; ++b)
      // split b-th input byte
      split_byte (inbits [row * bytes_per_inrow + b], i,
                  outbytes0[2*b], outbytes0[2*b+1],
                  outbytes1[2*b], outbytes1[2*b+1]);
    // write out the replacement rows
    for (int b = 0; b < bytes_per_outrow; ++b)
      std::cout << comma() << hex(outbytes0[b]); // row 0
    for (int b = 0; b < bytes_per_outrow; ++b)
      std::cout << comma() << hex(outbytes1[b]); // row 1
  }
  std::cout << "};\n";

  return 0;

}
```

B.13 Dynamic Types

Solution to Exercise ??. Here is a solution. We initially dynamically allocate an array of length $n = 1$ (pointed to by a pointer a), and whenever the next sequence element wouldn't fit anymore, we replace the array by a new one of length $2n$. For this, we first dynamically allocate a helper array, copy the contents of the current array into the helper array, delete the current array and then let a point to the newly allocated helper array.

```
// Program: read_array.cpp
// read a sequence of numbers into an array
#include <iostream>

int main()
{
  int n = 1; // current array size
  int k = 0; // number of elements read so far

  // dynamically allocate array
  int* a = new int[n]; // this time, a is NOT a constant

  // read into the array
  while (std::cin >> a[k]) {
    if (++k == n) {
      // next element wouldn't fit; replace the array a by
      // a new one of twice the size
      int* b = new int[n*=2];  // get pointer to new array
      for (int i=0; i<k; ++i)  // copy old array to new one
        b[i] = a[i];
      delete[] a;              // delete old array
      a = b;                   // let a point to new array
    }
  }

  // output the first k elements
  for (int i=0; i<k; ++i) std::cout << a[i] << " ";
  std::cout << "\n";

  // delete array
  delete[] a;

  return 0;

}
```

This is space and time efficient. The constant of proportionality in (i) is 3: whenever we grow the array (and these are the points in time where the ratio between memory cells in use and k is largest), we allocate a new array of length $2n = 2k$, in addition to the one of length k that we already have. This means that we have $3k$ memory cells in use at that time.

The constant of proportionality in (ii) is 3 as well. To see this, let us consider the situation after an execution of the `while` loop, where k is the number of elements read so far. There have been k assignments to array elements in the loop's condition, and some additional assignments during the copying of old to new array. Such assignments took place when the number of elements currently read was a power of two less or equal to k,

and the number of these additional assignments was exactly the power of two in question. Since the sum of all powers of two less or equal to k is at most $k + k/2 + k/4 + \cdots \leq 2k$, the total number of assignments is bounded by $k + 2k = 3k$.

Index

*this, 283
C++
    origin of name, 54
\n (line break), 167
32-bit system, 17

access
    data member, 251
    member function, 282
access restrictions
    private, 281
    public, 281
access specifier
    private, 281
    public, 281
access violation, 190
Ackermann function, 219
addition assignment operator, 54
addition operator, 50
address
    of a memory cell, 17
    of a variable, 27
address operator, 188
adjustment
    of array argument, 186
algorithm
    cache-oblivious, 19
    generic, 199
    standard, 196
    streaming, 168
alternate denial, 73
ANSIC
    random number generator, 294
antivalence, 73
application program, 10
argument-dependent lookup, 255
Arithmetic Evaluation Rule 1, 47
Arithmetic Evaluation Rule 2, 47
Arithmetic Evaluation Rule 3, 48
arithmetic expression, 30
arithmetic operators, 50
arithmetic type, 47
arity, 32
array
    as function argument, 186
    dimension, 170
    dynamic, 322
    element, 160
    incomplete type, 162
    index, 162
    initialization, 162
    initializer list, 162
    multidimensional, 170
    out-of-bound index, 163
    random access, 162
    size of, 164
    static, 161
    subscript, 162
    subscript operator, 162
    underlying type, 161
array-to-pointer conversion, 187
ASCII code, 166
asignment operator, 317
assertion, 136
assignment
    member-wise, 252
    of a struct value, 251

of reference, 264
pointer, 187
self, 318
user-defined, 317
assignment operator, 33
implicit, 317
associative operation, 47
associativity
left, 48
right, 48
associativity of operator, 47
automatic storage duration, 89
automaton
deterministic finite, 158

Babylonian method, 130
base
of a floating point number system, 115
BASIC
programming language, 144
behavior
implementation defined, 24
undefined, 24
unspecified, 24
binary expansion
of natural number, 57
of real number, 116
binary operator, 32
binary representation
of int value, 58
of unsigned int value , 58
binary search, 240
binary-to-decimal conversion, 58
binomial coefficient, 238
bit, 17
bitwise operators, 74
block, 85
BODMAS, 47
body
of do statement, 92
of for statement, 82
of function, 137
of while statement, 90
bool, 70
Boole, George, 68
Boolean, 68
Boolean Evaluation Rule, 71
Boolean expression, 72
Boolean function, 68
completeness, 69
break statement, 93
brute-force approach, 172
bubble-sort, 224
bug, 86
built-in type, 27
burst, 63
byte, 17

C++ standard, 24
cache, 19
cache-oblivious algorithm, 19
call
constructor, 166
member function, 169
call arguments, 138
call by name, 264
call by reference, 264
call by value, 265
call stack, 214
cancellation
in floating point computations, 124
Cardano's formula, 277
cast (functional notation), 113
cast expression, 113
central processing unit (CPU), 17
char
literal, 166
promotion to (unsigned) int, 166
type, 166
character, 166
control, 167
line break, 167
null, 169
special, 166
choosing numbers
game, 293

Church, Alonzo (1903–1995), 100
Church-Turing thesis, 100
class
    access specifier, 289
    constructor, 284
    copy constructor, 315
    definition, 289
    implementation, 289
    member declaration, 289
    member function, 282
    member function call, 283
    member operator, 287
    method, 283
    nested type, 288
    private member, 281
    public member, 281
    template, 199
class scope, 289
class type, 289
Collatz problem, 90
Collatz, Lothar (1910–1990), 101
command line argument, 179
command shell, 18
comment, 25
compilation, 16
    separate, 145
compiler, 16
complex number, 276
complexity
    of a problem, 220
    of an algorithm, 220
composite expression, 30
compound statement, 85
computable function, 100
computer
    main memory, 17
    memory cell, 17
    processor, 17
    von Neumann, 17
concept
    iterator, 203
condition
    of a while statement, 90
    of an if statement, 80
conditional operator, 101
const
    member function, 283
    pointer, 196
    reference, 271
Const Guideline, 29
const pointer, 196
const reference, 271
const-correctness, 29
const-pointer type
    in function argument, 196
const-qualified type, 28
    as formal argument type of function , 138
    as return type of function , 139
    in function argument, 196
constant, 28
constant expression, 162
constant pointer, 204
constructor, 284
    copy, 315
    default, 285
    explicit call, 285
    initializer, 284
constructor call, 166
container, 201
    traversal, 201
continue statement, 94
control character, 167
control flow, 80
    iteration, 81
    jump, 93
    linear, 80
    selection, 80
control statement, 81
control variable, 83
conversion
    array to pointer, 187
    explicit, 113
    floating point, 112
    implicit, 56
    integral, 56

promotion, 72
standard, 286
user-defined, 286
copy constructor, 315
implicit, 315
correctness proof, 216
coupon collector's problem, 122
CPU, 17
Cramer's rule, 181
cryptography
visual, 300
cryptosystem, 13

data encapsulation, 281
data member
access for, 251
of struct, 249
De Morgan's laws, 72
De Morgan, Augustus, 72
debugging output, 86
decimal-to-binary conversion, 57
declaration, 35
local, 86
of a class member, 289
of a function, 141
of a variable, 27
of friend, 296
struct, 250
declaration statement, 35
declarative region, 87
default argument
of a function, 205
default constructor, 285
default initialization
struct, 252
default-initialization
by default constructor, 285
definition
of a class, 289
of a function, 137
of a variable, 28
struct, 249
delete expression, 306
denormalized number, 127
dereference operator, 188
dereferencing, 188
destuctor, 319
deterministic finite automaton, 158
DFA, 158
dimension
(multidimensional) array, 170
directive
include, 25
using, 39
discriminant
of a quadratic equation, 124
divide and conquer, 224
division assignment operator, 54
do statement, 92
body, 92
domain
of a function, 136
double, 110
doubly-linked list, 304
drand48
random number generator, 293
dynamic
array, 322
dynamic memory allocation, 305
Dynamic Memory Guideline, 308
dynamic programming, 172
dynamic storage duration, 305
dynamic type, 320

editor, 15
effect
of a function, 26
of a statement, 35
of an expression, 30
effect (semantical term), 27
element
of array, 160
encapsulation
of data, 281
equality
pointer, 187

Eratosthenes' Sieve, 160
Erdős, Paul (1913–1996), 101
Euclidean algorithm, 215
Euler, Leonhard (1707 - 1783), 182
evaluation
    of an expression, 30
    order of operands, 32
    short circuit, 73
evaluation sequence, 49
Excel 2007 bug, 117
executable, 16
execution, 35
explicit conversion, 113
exponent
    of a floating point number, 115
expression, 30
    arithmetic, 30
    Boolean, 72
    cast, 113
    composite, 30
    constant, 162
    delete, 306
    effect of, 30
    evaluation of, 30
    evaluation sequence, 49
    function call, 138
    literal, 27
    lvalue, 31
    mixed, 56
    new, 306
    of type void, 139
    order of effects, 63
    primary, 30
    rvalue, 31
    type of, 30
    value of, 30
    variable, 27
expression statement, 35
expression tree, 49

factoring numbers, 12
fair dice, 293
false, 70
Fibonacci numbers, 217
file, 18
finite floating point number system, 115
fixed point number, 109
float, 110
Floating Point Arithmetic Guideline 1, 121
Floating Point Arithmetic Guideline 2, 124
floating point computations
    cancellation, 124
    different sizes, 124
    equality test, 121
    relative error, 119
floating point conversions, 112
floating point number
    denormalized, 127
    exponent, 115
    infinity, 127
    mantissa, 115
    NaN, 127
    normalized, 115
    sign, 115
    significand, 115
floating point number system, 115
    base, 115
    largest exponent, 115
    precision, 115
    smallest exponent, 115
floating point type, 109
floor function, 113
for statement, 82
    body, 82
    init-statement, 82
    iteration, 83
    termination, 83
formal argument, 137
fractal, 132, 232
friend declaration, 296
function
    Ackermann, 219
    body, 137
    call, 138
    call arguments, 138
    call by reference, 264

call by value, 265
const-pointer argument type, 196
const-qualified argument type, 138, 196
const-qualified return type, 139
declaration, 141
default argument, 205
definition, 137
domain, 136
effect, 135
formal argument, 137
formal argument of reference type, 264
main, 26
mutating, 195
overloading, 253
postcondition, 135
precondition, 135
recursive, 213
recursive call, 213
return by reference, 265
return by value, 265
return type, 137
return value of reference type, 265
scope, 141
scope of formal arguments, 140
signature, 152
template, 198
value, 135
void, 139
function call
qualified, 255
unqualified, 255
function call operator, 292
functional operator notation, 254
functional programming language, 26
functionality
of a struct, 251
of a type, 27
of an operator, 31
fundamental type, 27

game
choosing numbers, 293
garbage collection, 306
Gauss, Carl-Friedrich, 84
generic algorithm, 199
generic programming, 203
global scope, 87
global variable, 140
goto statement, 102
greatest common divisor, 215
guideline
Dynamic Memory, 308

halting problem, 84, 100
harmonic number, 122
header
file, 146
of the standard library, 26
iostream, 26
heap, 306
hexadecimal literal, 183
hiding
of name, 88

identifier, 29
IEEE compliance, 126
IEEE standard 754, 120
arithmetic operations, 121
double extended precision, 125
single extended precision, 125
value range, 120
IEEE standard 854, 125
if statement, 80
condition, 80
if-else statement, 81
ifm::integer, 37
implementation defined behavior, 24
implicit assignment operator, 317
implicit conversion, 56
stream to bool, 170
implicit copy constructor, 315
include directive, 25
variant with angle brackets, 148
variant with quotes, 145
incomplete array type, 162
incomplete type, 250

indentation, 25
index
    of array element, 162
indirection, 188
infinite loop, 83
infinite recursion, 214
infix operator notation, 254
initialization
    by constructor, 285
    by zero, 102
    member-wise, 251
    of a struct value, 251
    of a variable, 36
    of array, 162
    of reference, 263
    pointer, 187
    user-defined, 314
initializer
    of constructor, 284
initializer list
    array, 162
input operator, 33
input stream, 33
input/output efficiency, 20
insert-sort, 224
int, 47
integer division, 52
integer division operator, 52
integral conversions, 56
integral type, 59
integrity
    of representation, 279
invariant
    of a struct, 248
iostream, 26
ISO/IEC standard 14882, 24
iteration, 81, 194
iteration statements, 81
    equivalence of, 95
iterator, 203
    concept, 203
    set, 202
    vector, 200

Josephus problem, 67
jump statements, 93

Knuth, Donald E., 13
Knuth-Morris-Pratt algorithm, 167
knuth8
    random number generator, 290
Koenig lookup, 255

layout of program, 25
least common multiple, 43
left associativity, 48
left-associative, 34
library, 147
    standard, 26
Lindenmayer system, 230
    alphabet, 229
    fractal, 232
    graphical interpretation, 231
    initial word, 230
    productions, 229
Lindenmayer, Aristide (1925–1985), 235
line break character, 167
linear congruential generator, 290
linear congruential method, 290
linear control flow, 80
linker, 146
Linux, 18
list, 304
    doubly-linked, 304
    singly-linked, 304
    std, 322
literal, 27
    bool, 70
    char, 166
    double, 111
    float, 111
    hexadecimal, 60, 183
    int, 47
    long double, 126
    octal, 60
    string, 166
    unsigned int, 56

loaded dice, 293
local declaration, 86
local scope, 87
logical parentheses, 60
logical operators, 71
logical parentheses
    leading operand, 60
    secondary operand, 60
long double, 126
long int, 62
lookup
    argument-dependent, 255
loop, 81
    infinite, 83
    progress towards termination, 83
lvalue, 31
lvalue-to-rvalue conversion, 32

Mac OS, 18
machine epsilon, 119
machine language, 15, 18
macro, 136
main function, 26
main memory, 17
Mandelbrot set, 132
mantissa
    of a floating point number, 115
mathematical induction, 216
McCarthy's 91 Function, 237
member access
    in a struct, 251
    through pointer, 312
member access operator, 251
member function, 282
    access for, 282
    and const, 283
    call, 283
    implicit call argument, 282
member function call, 169
member operator
    of class, 287
member specification
    of a struct, 250
member-wise assignment, 252
member-wise initialization, 251
memory cell, 17
    address, 17
memory leak, 308
merge-sort, 224
    complexity, 226
Mersenne primes, 11
method
    of class, 283
minimum-sort, 220
    complexity, 221
mixed expression, 56
modularization, 144
modulus assignment operator, 54
modulus operator, 52
multidimensional array, 170
    dimension, 170
multiplication assignment operator, 54
multiplication operator, 33, 50
mutating function, 195

name
    clash, 26
    hiding, 88
    of a class, 289
    of a function, 137
    of a type, 27
    of a variable, 27
    of formal argument, 137
    qualified, 26
    unqualified, 26
namespace, 26
namespace scope, 87
nested type, 288
new expression, 306
normalized floating point number, 115
null character, 169
null pointer, 189
null pointer value, 189
null statement, 35
numeric limits
    of floating point types, 126

of integral types, 55

object, 30
unnamed, 30
object code, 145
open source software, 147
operand, 31
evaluation order, 32
operating system (OS), 18
Linux, 18
Mac OS, 18
Unix, 18
Windows, 18
operator
addition, 50
addition assignment, 54
address, 188
arithmetic, 50
arithmetic assignment, 54
arity, 32
assignment, 33, 317
associativity, 47
binary, 32
binding, 47
bitwise, 74
conditional, 101
conversion, 286
dereference, 188
division assignment, 54
function call, 292
functional notation, 254
functionality, 31
infix notation, 254
input, 33
integer division, 52
left-associative, 34
logical, 71
member access through pointer, 312
modulus, 52
modulus assignment, 54
multiplication, 33, 50
multiplication assignment, 54
operand, 31
output, 34
overloading, 253
post-decrement, 53
post-increment, 53
pre-decrement, 53
pre-increment, 53
precedence, 47
relational, 71
return value, 31
scope, 201
sizeof, 164
subscript, 162, 191
subtraction, 50
subtraction assignment, 54
ternary, 101
unary, 32
unary minus, 52
unary plus, 52
operator token, 32
overloaded, 48
order of effects, 63
OS, 18
out-of-bound array index, 163
output operator, 34
output stream, 34
overflow
of value range, 55
overloading
argument-dependent lookup, 255
best match, 255
of functions, 253
of operators, 253
overloading resolution, 255

Panini, 242
parallel computer, 20
Pascal's triangle, 401
past-the-end pointer, 190
PEMDAS, 47
permutation, 209
perpetual calendar, 156
pipe, 78
platform, 19

point of declaration, 87, 101
pointer, 187
    adding an integer, 190
    arithmetic, 190
    assignment, 187
    comparison, 192
    const, 196
    constant, 204
    equality, 187
    initialization, 187
    null, 189
    null value, 189
    past-the-end, 190
    subscript operator, 191
    subtraction, 193
    type, 186
pointer type
    underlying type, 187
porting, 16
post-decrement operator, 53
post-increment operator, 53
postcondition
    of a function, 135
potential scope, 87
pre-decrement operator, 53
pre-increment operator, 53
precedence of operator, 47
precision
    of a floating point number system, 115
precondition
    of a function, 135
predicate, 70
primary expression, 30
primitive recursion, 218
private
    class member, 281
private:, 281
procedural programming, 144
processor, 17
production
    of a Lindenmayer system, 229
program
    const-correct, 29
    layout, 25
    valid, 23
Program Listing
    callpow.cpp, 135
    callpow2.cpp, 145
    callpow3.cpp, 146
    callpow4.cpp, 150
    choosing_numbers.cpp, 296
    clock.cpp, 299
    collatz.cpp, 91
    diff.cpp, 115
    dragon.cpp, 235
    eratosthenes.cpp, 161
    eratosthenes2.cpp, 166
    euler.cpp, 112
    fahrenheit.cpp, 46
    fill.cpp, 186
    fill2.cpp, 195
    fill3.cpp, 197
    fill4.cpp, 198
    fill5.cpp, 199
    fill6.cpp, 199
    harmonic.cpp, 123
    limits.cpp, 55
    lindenmayer.cpp, 232
    list_example.cpp, 319
    loaded_dice.cpp, 295
    loaded_dice.h, 293
    math.cpp, 149
    math.h, 149
    merge_sort.cpp, 229
    minimum_sort.cpp, 223
    node.cpp, 310
    node.h, 309
    perfect2.cpp, 143
    pow.cpp, 145
    power8.cpp, 22
    power8_condensed.cpp, 25
    power8_exact.cpp, 38
    power8_using.cpp, 40
    prime.cpp, 85
    prime2.cpp, 151
    random.cpp, 292

random.h, 292
rational.cpp, 269
scope.cpp, 88
set.cpp, 202
shortest_path.cpp, 179
snowflake.cpp, 234
string_matching.cpp, 169
sum_n.cpp, 82
userational.cpp, 248
userational2.cpp, 268
vector_iterators.cpp, 200
Program Listing (Solutions)
age_verification.cpp, 330
and.cpp, 347
average_speed.cpp, 337
babylonian.cpp, 371
badloop.cpp, 367
binary_search2.cpp, 419
bush.cpp, 428
celsius.cpp, 338
clock_solution.cpp, 455
cole.cpp, 357
copy_using_pointers.cpp, 403
crack.cpp, 416
create_rational_vector_3.cpp, 429
cross_sum.cpp, 354
cycles.cpp, 404
dec2bin.cpp, 354
dec2bin2.cpp, 356
dec2float.cpp, 367
dice.cpp, 357
double_integer.cpp, 368
euler_prime.cpp, 397
eval.cpp, 348
extended_int.cpp, 435
fak-1.cpp, 353
famous_last_digits.cpp, 361
first_occurrence.cpp, 405
fpsys.cpp, 372
fpsys2.cpp, 379
frequencies.cpp, 394
geometry.cpp, 438
hanoi.cpp, 423
harmonic.cpp, 454
inverse_matrix.cpp, 392
josephus.cpp, 340
k_composite.cpp, 391
last_occurrence.cpp, 405
lcm.cpp, 328
lindenmayer_a.cpp, 421
lindenmayer_b.cpp, 422
lindenmayer_c.cpp, 423
lineq.cpp, 442
lucas_lehmer.cpp, 425
mandelbrot.cpp, 375
mccarthy.cpp, 411
modulo7.cpp, 432
mr_prod_sum.cpp, 390
multhree.cpp, 329
n_queens.cpp, 360
normalize_rat.cpp, 441
not.cpp, 347
one.cpp, 348
or.cpp, 347
panini.cpp, 427
partition.cpp, 412
partition2.cpp, 414
partition3.cpp, 415
perfect.cpp, 355
perpetual_calendar.cpp, 383
pi1.cpp, 369
pi2.cpp, 369
power20.cpp, 329
power8_fast.cpp, 331
power8_slow.cpp, 331
power_cross_sums.cpp, 386
prime_all.cpp, 353
pythagoras.cpp, 342
random_triangle.cpp, 456
rational.cpp, 453
rational.h, 451
rational_ops.cpp, 434
read_array.cpp, 393
read_array2.cpp, 461
rec2it.cpp, 417
rec2it2.cpp, 417

round.cpp, 385
shortest_path_fast.cpp, 396
sinus.cpp, 370
solve_cubic_equation.cpp, 449
solve_quadratic_equation.cpp, 443
sort_array.cpp, 393
sort_array2.cpp, 401
square_product.cpp, 330
std_sort.cpp, 418
sudoku.cpp, 408
swap.cpp, 399
swap2.cpp, 440
threebin.cpp, 338
tribool.cpp, 431
twinprimes.cpp, 380
twinprimes2.cpp, 380
ulam_spiral.cpp, 358
unique.cpp, 400
vat.cpp, 339
visual_crypto.cpp, 460
xbm.cpp, 398
xbm_merge.cpp, 458
zero.cpp, 347
program state, 17
programming language, 10
functional, 26
promotion, 72
bool to int, 72
char to (unsigned) int, 166
float to double, 113
pseudorandom numbers, 290
public
class member, 281
public:, 281
Pythagorean triple, 67

qualified function call, 255
qualified name, 26
quantum computer, 20

RAM, 17
random access
in array, 162
random access memory (RAM), 17
random number, 290
random number generator, 290
ANSIC, 294
drand48, 293
knuth8, 290
range
of pointers, 194
valid, 194
rational numbers, 246
recurrence relation, 226
recursion
infinite, 214
primitive, 218
tail-end, 217
recursive call, 213
recursive function, 213
correctness, 216
termination, 216
refactoring, 97
reference, 263
assignment, 264
const, 271
implementation, 263
initialization, 263
Reference Guideline, 266
reference type, 263
relational operators, 71
relative error
in floating point computations, 119
reserved name, 29
return by reference, 265
return by value, 265
return statement, 36, 102
return type, 137
return value, 31
Reverse Polish Notation, 77
right associativity, 48
runtime error, 137
rvalue, 31

Sarrus' rule, 181
scope

global, 87
local, 87
namespace, 87
of a declaration, 87
of a function declaration, 141
scope operator, 201
segmentation fault, 190
selection, 80
selection statements, 81
self-assignment, 318
semantical value range
of a struct, 252
semantics, 23
sentinel, 173
separate compilation, 145
sequence point, 63
set, 201
set iterator, 202
Sheffer stroke, 73
Sheffer, Henry M. (1883–1964), 73
short circuit evaluation, 73
short int, 62
shortest path problem, 172
side effect, 31
Sieve of Eratosthenes, 160
sign
of a floating point number, 115
signature of a function, 152
signed char, 62
significand
of a floating point number, 115
Single Modification Rule, 64
singly-linked list, 304
sizeof operator, 164
sorting algorithm
bubble-sort, 224
insert-sort, 224
merge-sort, 224
minimum-sort, 220
sourcecode, 16
availability, 147
spaghetti code, 144
special character, 166
standard conversion, 286
standard error, 34
standard input, 33
standard library, 26
algorithm, 196
mathematical functions, 152
std::cerr, 34
std::cin, 33
std::cout, 34
std::pow, 150
std::sqrt, 150
standard output, 34
statement, 35
break, 93
compound, 85
continue, 94
control, 81
declaration, 35
do, 92
execution, 35
expression, 35
for, 82
goto, 102
if, 80
if-else, 81
iteration, 81
jump, 93
null, 35
return, 36, 102
selection, 81
switch, 99
while, 90
static array, 161
static storage duration, 90
static variable, 90, 102
staticctype, 320
std::cerr, 34
std::cin, 33
std::complex, 276
std::cout, 34
std::list, 322
std::sqrt, 150
std::string, 167

storage duration, 88
    automatic, 89
    dynamic, 305
    static, 90
streaming algorithm, 168
string, 166
    initialization from string literal, 167
string literal, 166
string matching, 167
    Knuth-Morris-Pratt algorithm, 167
    obvious algorithm, 167
struct, 248
    assignment, 251
    data member, 249
    declaration, 250
    default initialization, 252
    definition, 249
    functionality, 251
    initialization, 251
    member access, 251
    member specification, 250
    underlying type, 249
    value range, 249
      semantical, 252
      syntactical, 252
subscript
    of array element, 162
subscript operator
    array, 162
    pointer, 191
subtraction assignment operator, 54
subtraction operator, 50
Sudoku, 210
swapping
    of memory, 20
switch statement, 99
syntactical value range
    of a struct, 252
syntax, 23
syntax error, 23

tail-end recursion, 217
template
    class, 199
    function, 198
temporary object, 265
    reference to, 265
ternary operator, 101
topological sorting, 49
Towers of Hanoi, 241
traversal
    of a container, 201
true, 70
Turing machine, 100
Turing, Alan (1912–1954), 100
turtle graphics, 230
two's complement, 59
type, 27
    arithmetic, 47
    bool, 70
    built-in, 27
    char, 166
    class, 289
    const-qualified, 28
    double, 110
    dynamic, 320
    float, 110
    floating point, 109
    functionality of, 27
    fundamental, 27
    incomplete, 250
    incomplete array, 162
    int, 47
    integral, 59
    long double, 126
    long int, 62
    name of, 27
    of a variable, 27
    of an expression, 30
    of formal argument, 137
    pointer, 186
    reference, 263
    short int, 62
    signed char, 62
    static, 320
    underlying a struct, 249

underlying an array, 161
unsigned char, 62
unsigned int, 55
unsigned long int, 62
unsigned short int, 62
value range of, 27
void, 139
typedef, 201

Ulam spiral, 107
Ulam, Stanislaw (1909–1984), 107
unary minus operator, 52
unary operator, 32
unary plus operator, 52
undecidable problem, 100
undefined behavior, 24
underflow
of value range, 55
underlying type
of a pointer type, 187
Unix, 18
unnamed object, 30
unqualified function call, 255
unqualified name, 26
unsigned char, 62
unsigned int, 55
unsigned long int, 62
unsigned short int, 62
unspecified behavior, 24
user-defined assignment, 317
user-defined conversion, 286
user-defined initialization, 314
using directive, 39

valid program, 23
valid range, 194
value
of a variable, 27
of an expression, 30
value (semantical term), 27
value range
of a struct, 249
of a type, 27
of type bool, 70
of type double, 120
of type float, 120
of type int, 54
of type unsigned int, 56
overflow, 55
semantical, 252
syntactical, 252
underflow, 55
variable, 27
address of, 27
control, 83
global, 140
name of, 27
static, 90, 102
type of, 27
value of, 27
variable declaration, 27
variable definition, 28
vector, 165
iterator, 200
visibility
of name, 87
visual cryptography, 300
void, 139
void function, 139
von Neumann computer, 17

while statement, 90
body, 90
whitespace, 170
Windows, 18
wireframe model, 260

XBM graphics format, 183

zero-initialization, 102
zero-sum game, 299