1.
Optimizing software in C++
An optimization guide for Windows, Linux and Mac platforms

By Agner Fog. Copenhagen University College of Engineering.
Copyright © 2004 - 2011. Last updated 2011-06-08.

Contents

2

1 Introduction

This manual is for advanced programmers and software developers who want to make their software faster. It is assumed that the reader has a good knowledge of the C++ programming language and a basic understanding of how compilers work. The C++ language is chosen as the basis for this manual for reasons explained on page 8 below.

This manual is based mainly on my study of how compilers and microprocessors work. The recommendations are based on the x86 family of microprocessors from Intel, AMD and VIA including the 64-bit versions. The x86 processors are used in the most common platforms with Windows, Linux, BSD and Mac OS X operating systems, though these operating systems can also be used with other microprocessors. Many of the advices may apply to other platforms and other compiled programming languages as well.

This is the first in a series of five manuals:

1. Optimizing software in C++: An optimization guide for Windows, Linux and Mac platforms.

2. Optimizing subroutines in assembly language: An optimization guide for x86 platforms.

3. The microarchitecture of Intel, AMD and VIA CPUs: An optimization guide for assembly programmers and compiler makers.

4. Instruction tables: Lists of instruction latencies, throughputs and micro-operation breakdowns for Intel, AMD and VIA CPUs.

5. Calling conventions for different C++ compilers and operating systems.

The latest versions of these manuals are always available from www.agner.org/optimize. Copyright conditions are listed on page 165 below.

Those who are satisfied with making software in a high-level language need only read this first manual. The subsequent manuals are for those who want to go deeper into the technical details of instruction timing, assembly language programming, compiler technology, and microprocessor microarchitecture. A higher level of optimization can

sometimes be obtained by the use of assembly language for CPU-intensive code, as described in the subsequent manuals.

Please note that my optimization manuals are used by thousands of people. I simply don't have the time to answer questions from everybody. So please don't send your programming questions to me. You will not get any answer. Beginners are advised to seek information elsewhere and get a good deal of programming experience before trying the techniques in the present manual. There are various discussion forums on the Internet where you can get answers to your programming questions if you cannot find the answers in the relevant books and manuals.

I want to thank the many people who have sent me corrections and suggestions for my optimization manuals. I am always happy to receive new relevant information.

1.1 The costs of optimizing

University courses in programming nowadays stress the importance of structured and object-oriented programming, modularity, reusability and systematization of the software development process. These requirements are often conflicting with the requirements of optimizing the software for speed or size.

Today, it is not uncommon for software teachers to recommend that no function or method should be longer than a few lines. A few decades ago, the recommendation was the opposite: Don't put something in a separate subroutine if it is only called once. The reasons for this shift in software writing style are that software projects have become bigger and more complex, that there is more focus on the costs of software development, and that computers have become more powerful.

The high priority of structured software development and the low priority of program efficiency is reflected, first and foremost, in the choice of programming language and interface frameworks. This is often a disadvantage for the end user who has to invest in ever more powerful computers to keep up with the ever bigger software packages and who is still frustrated by unacceptably long response times, even for simple tasks.

Sometimes it is necessary to compromise on the advanced principles of software development in order to make software packages faster and smaller. This manual discusses how to make a sensible balance between these considerations. It is discussed how to identify and isolate the most critical part of a program and concentrate the optimization effort on that particular part. It is discussed how to overcome the dangers of a relatively primitive programming style that doesn't automatically check for array bounds violations, invalid pointers, etc. And it is discussed which of the advanced programming constructs are costly and which are cheap, in relation to execution time.

2 Choosing the optimal platform

2.1 Choice of hardware platform

The choice of hardware platform has become less important than it used to be. The distinctions between RISC and CISC processors, between PC's and mainframes, and between simple processors and vector processors are becoming increasingly blurred as the standard PC processors with CISC instruction sets have got RISC cores, vector processing instructions, multiple cores, and a processing speed exceeding that of yesterday's big mainframe computers.

Today, the choice of hardware platform for a given task is often determined by considerations such as price, compatibility, second source, and the availability of good

development tools, rather than by the processing power. Connecting several standard PC's in a network may be both cheaper and more efficient than investing in a big mainframe computer. Big supercomputers with massively parallel vector processing capabilities still have a niche in scientific computing, but for most purposes the standard PC processors are preferred because of their superior performance/price ratio.

The CISC instruction set (called x86) of the standard PC processors is not optimal from a technological point of view. This instruction set is maintained for the sake of backwards compatibility with a lineage of software that dates back to around 1980 where RAM memory and disk space were scarce resources. However, the CISC instruction set is better than its reputation. The compactness of the code makes caching more efficient today where cache size is a limited resource. The CISC instruction set may actually be better than RISC in situations where code caching is critical. The worst problem of the x86 instruction set is the scarcity of registers. This problem has been alleviated in the 64-bit extension to the x86 instruction set where the number of registers has been doubled.

Thin clients that depend on network resources are not recommended for critical applications because the response times for network resources cannot be controlled.

Small hand-held devices are becoming more popular and used for an increasing number of purposes such as email and web browsing that previously required a PC. Similarly, we are seeing an increasing number of devices and machines with embedded microcontrollers. I am not making any specific recommendation about which platforms and operating systems are most efficient for such applications, but it is important to realize that such devices typically have much less memory and computing power than PCs. Therefore, it is even more important to economize the resource use on such systems than it is on a PC platform. However, with a well optimized software design, it is possible to get a good performance for many applications even on such small devices, as discussed on page 159.

This manual is based on the standard PC platform with an Intel, AMD or VIA processor and a Windows, Linux, BSD or Mac operating system running in 32-bit or 64-bit mode. Much of the advice given here may apply to other platforms as well, but the examples have been tested only on PC platforms.

Graphics accelerators

The choice of platform is obviously influenced by the requirements of the task in question. For example, a heavy graphics application is preferably implemented on a platform with a graphics coprocessor or graphics accelerator card. Some systems also have a dedicated physics processor for calculating the physical movements of objects in a computer game or animation.

It is possible in some cases to use the high processing power of the processors on a graphics accelerator card for other purposes than rendering graphics on the screen. However, such applications are highly system dependent and therefore not recommended if portability is important. This manual does not cover graphics processors.

Programmable logic devices

A programmable logic device is a chip that can be programmed in a hardware definition language, such as VHDL or Verilog. Common devices are CPLDs and FPGAs. The difference between a software programming language, e.g. C++, and a hardware definition language is that the software programming language defines an algorithm of sequential instructions, where a hardware definition language defines hardware circuits consisting of digital building blocks such as gates, flip-flops, multiplexers, arithmetic units, etc. and the wires that connect them. The hardware definition language is inherently parallel because it defines electrical connections rather than sequences of operations.

A complex digital operation can often be executed faster in a programmable logic device than in a microprocessor because the hardware can be wired for a specific purpose.

It is possible to implement a microprocessor in an FPGA as a so-called soft processor. Such a soft processor is much slower than a dedicated microprocessor and therefore not advantageous by itself. But a solution where a soft processor activates critical application-specific instructions that are coded in a hardware definition language in the same chip can be a very efficient solution in some cases. An even more powerful solution is the combination of a dedicated microprocessor core and an FPGA in the same chip. Such hybrid solutions are now used in some embedded systems.

A look in my crystal ball reveals that similar solutions may some day be implemented in PC processors. The application program will be able to define application-specific instructions that can be coded in a hardware definition language. Such a processor will have an extra cache for the hardware definition code in addition to the code cache and the data cache.

2.2 Choice of microprocessor

The benchmark performance of competing brands of microprocessors are very similar thanks to heavy competition. Processors with multiple cores are advantageous for applications that can be divided into multiple threads that run in parallel. Small lightweight processors with low power consumption are actually quite powerful and may be sufficient for less intensive applications.

Some systems have a graphics processing unit, either on a graphics card or integrated in the CPU chip. Such units can be used as coprocessors to take care of some of the heavy graphics calculations. In some cases it is possible to utilize the computational power of the graphics processing unit for other purposes than it is intended for. Some systems also have a physics processing unit intended for calculating the movements of objects in computer games. Such a coprocessor might also be used for other purposes. The use of coprocessors is beyond the scope of this manual.

2.3 Choice of operating system

All newer microprocessors in the x86 family can run in both 16-bit, 32-bit and 64-bit mode.

16-bit mode is used in the old operating systems DOS and Windows 3.x. These systems use segmentation of the memory if the size of program or data exceeds 64 kbytes. This is quite inefficient. The modern microprocessors are not optimized for 16-bit mode and some operating systems are not backwards compatible with 16-bit programs. It is not recommended to make 16-bit programs, except for small embedded systems.

Today (2011) both 32-bit and 64-bit operating systems are common, and there is no big difference in performance between the systems. There is no heavy marketing of 64-bit software yet, but it is quite certain that the 64-bit systems will dominate in the future.

The 64-bit systems can improve the performance by 5-10% for some CPU-intensive applications with many function calls. If the bottleneck is elsewhere then there is no difference in performance between 32-bit and 64-bit systems. Applications that use large amounts of memory will benefit from the larger address space of the 64-bit systems.

A software developer may choose to make memory-hungry software in two versions. A 32-bit version for the sake of compatibility with existing systems and a 64-bit version for best performance.

The Windows and Linux operating systems give almost identical performance for 32-bit software because the two operating systems are using the same function calling

conventions. FreeBSD and Open BSD are identical to Linux in almost all respects relevant to software optimization. Everything that is said here about Linux also applies to BSD systems.

The Intel-based Mac OS X operating system is based on BSD, but the compiler uses position-independent code and lazy binding by default, which makes it less efficient. The performance can be improved by using static linking and by not using position-independent code (option `-fno-pic`).

64 bit systems have several advantages over 32 bit systems:

- The number of registers is doubled. This makes it possible to store intermediate data and local variables in registers rather than in memory.

- Function parameters are transferred in registers rather than on the stack. This makes function calls more efficient.

- The size of the integer registers is extended to 64 bits. This is only an advantage in applications that can take advantage of 64-bit integers.

- The allocation and deallocation of big memory blocks is more efficient.

- The SSE2 instruction set is supported on all 64-bit CPUs and operating systems.

- The 64 bit instruction set supports self-relative addressing of data. This makes position-independent code more efficient.

64 bit systems have the following disadvantages compared to 32 bit systems:

- Pointers, references, and stack entries use 64 bits rather than 32 bits. This makes data caching less efficient.

- Access to static or global arrays require a few extra instructions for address calculation in 64 bit mode if the image base is not guaranteed to be less than 2^{31}. This extra cost is seen in 64 bit Windows and Mac programs but rarely in Linux.

- Address calculation is more complicated in a large memory model where the combined size of code and data can exceed 2 Gbytes. This large memory model is hardly ever used, though.

- Some instructions are one byte longer in 64 bit mode than in 32 bit mode.

- Some 64-bit compilers are inferior to their 32-bit counterparts.

In general, you can expect 64-bit programs to run a little faster than 32-bit programs if there are many function calls, if there are many allocations of large memory blocks, or if the program can take advantage of 64-bit integer calculations. It is necessary to use 64-bit systems if the program uses more than 2 gigabytes of data.

The similarity between the operating systems disappears when running in 64-bit mode because the function calling conventions are different. 64-bit Windows allows only four function parameters to be transferred in registers, whereas 64-bit Linux, BSD and Mac allow up to fourteen parameters to be transferred in registers (6 integer and 8 floating point). There are also other details that make function calling more efficient in 64-bit Linux than in 64-bit Windows (See page 50 and manual 5: "Calling conventions for different C++ compilers and operating systems"). An application with many function calls may run slightly faster in 64-bit Linux than in 64-bit Windows. The disadvantage of 64-bit Windows may be

mitigated by making critical functions inline or static or by using a compiler that can do whole program optimization.

2.4 Choice of programming language

Before starting a new software project, it is important to decide which programming language is best suited for the project at hand. Low-level languages are good for optimizing execution speed or program size, while high-level languages are good for making clear and well-structured code and for fast and easy development of user interfaces and interfaces to network resources, databases, etc.

The efficiency of the final application depends on the way the programming language is implemented. The highest efficiency is obtained when the code is compiled and distributed as binary executable code. Most implementations of C++, Pascal and Fortran are based on compilers.

Several other programming languages are implemented with interpretation. The program code is distributed as it is and interpreted line by line when it is run. Examples include JavaScript, PHP, ASP and UNIX shell script. Interpreted code is very inefficient because the body of a loop is interpreted again and again for every iteration of the loop.

Some implementations use just-in-time compilation. The program code is distributed and stored as it is, and is compiled when it is executed. An example is Perl.

Several modern programming languages use an intermediate code (byte code). The source code is compiled into an intermediate code, which is the code that is distributed. The intermediate code cannot be executed as it is, but must go through a second step of interpretation or compilation before it can run. Some implementations of Java are based on an interpreter which interprets the intermediate code by emulating the so-called Java virtual machine. The best Java machines use just-in-time compilation of the most used parts of the code. C#, managed C++, and other languages in Microsoft's .NET framework are based on just-in-time compilation of an intermediate code.

The reason for using an intermediate code is that it is intended to be platform-independent and compact. The biggest disadvantage of using an intermediate code is that the user must install a large runtime framework for interpreting or compiling the intermediate code. This framework typically uses much more resources than the code itself.

Another disadvantage of intermediate code is that it adds an extra level of abstraction which makes detailed optimization more difficult. On the other hand, a just-in-time compiler can optimize specifically for the CPU it is running on, while it is more complicated to make CPU-specific optimizations in precompiled code.

The history of programming languages and their implementations reveal a zigzag course that reflects the conflicting considerations of efficiency, platform independence, and easy development. For example, the first PC's had an interpreter for Basic. A compiler for Basic soon became available because the interpreted version of Basic was too slow. Today, the most popular version of Basic is Visual Basic .NET, which is implemented with an intermediate code and just-in-time compilation. Some early implementations of Pascal used an intermediate code like the one that is used for Java today. But this language gained remarkably in popularity when a genuine compiler became available.

It should be clear from this discussion that the choice of programming language is a compromise between efficiency, portability and development time. Interpreted languages are out of the question when efficiency is important. A language based on intermediate code and just-in-time compilation may be a viable compromise when portability and ease of development are more important than speed. This includes languages such as C#, Visual Basic .NET and the best Java implementations. However, these languages have the

disadvantage of a very large runtime framework that must be loaded every time the program is run. The time it takes to load the framework and compile the program are often much more than the time it takes to execute the program, and the runtime framework may use more resources than the program itself when running. Programs using such a framework sometimes have unacceptably long response times for simple tasks like pressing a button or moving the mouse. The .NET framework should definitely be avoided when speed is critical.

The fastest execution is no doubt obtained with a fully compiled code. Compiled languages include C, C++, D, Pascal, Fortran and several other less well-known languages. My preference is for C++ for several reasons. C++ is supported by some very good compilers and optimized function libraries. C++ is an advanced high-level language with a wealth of advanced features rarely found in other languages. But the C++ language also includes the low-level C language as a subset, giving access to low-level optimizations. Most C++ compilers are able to generate an assembly language output, which is useful for checking how well the compiler optimizes a piece of code. Furthermore, most C++ compilers allow assembly-like intrinsic functions, inline assembly or easy linking to assembly language modules when the highest level of optimization is needed. The C++ language is portable in the sense that C++ compilers exist for all major platforms. Pascal has many of the advantages of C++ but is not quite as versatile. Fortran is also quite efficient, but the syntax is very old-fashioned.

Development in C++ is quite efficient thanks to the availability of powerful development tools. One popular development tool is Microsoft Visual Studio. This tool can make two different implementations of C++, directly compiled code and intermediate code for the common language runtime of the .NET framework. Obviously, the directly compiled version is preferred when speed is important.

An important disadvantage of C++ relates to security. There are no checks for array bounds violation, integer overflow, and invalid pointers. The absence of such checks makes the code execute faster than other languages that do have such checks. But it is the responsibility of the programmer to make explicit checks for such errors in cases where they cannot be ruled out by the program logic. Some guidelines are provided below, on page 15.

C++ is definitely the preferred programming language when the optimization of performance has high priority. The gain in performance over other programming languages can be quite substantial. This gain in performance can easily justify a possible minor increase in development time when performance is important to the end user.

There may be situations where a high level framework based on intermediate code is needed for other reasons, but part of the code still needs careful optimization. A mixed implementation can be a viable solution in such cases. The most critical part of the code can be implemented in compiled C++ or assembly language and the rest of the code, including user interface etc., can be implemented in the high level framework. The optimized part of the code can possibly be compiled as a dynamic link library (DLL) which is called by the rest of the code. This is not an optimal solution because the high level framework still consumes a lot of resources, and the transitions between the two kinds of code gives an extra overhead which consumes CPU time. But this solution can still give a considerable improvement in performance if the time-critical part of the code can be completely contained in a DLL.

Another alternative worth considering is the D language. D has many of the features of Java and C# and avoids many of the drawbacks of C++. Yet, D is compiled to binary code and can be linked together with C or C++ code. Compilers and IDE's for D are not yet as well developed as C++ compilers.

2.5 Choice of compiler

There are several different C++ compilers to choose between. It is difficult to predict which compiler will do the best job optimizing a particular piece of code. Each compiler does some things very smart and other things very stupid. Some common compilers are mentioned below.

Microsoft Visual Studio

This is a very user friendly compiler with many features, but also very expensive. A limited "express" edition is available for free. Visual Studio can build code for the .NET framework as well as directly compiled code. (Compile *without* the Common Language Runtime, CLR, to produce binary code). Supports 32-bit and 64-bit Windows. The integrated development environment (IDE) supports multiple programming languages, profiling and debugging. A command-line version of the C++ compiler is available for free in the Microsoft platform software development kit (SDK or PSDK). Supports the OpenMP directives for multi-core processing. Visual Studio optimizes reasonably well, but it is not the best optimizer.

Borland/CodeGear/Embarcadero C++ builder

Has an IDE with many of the same features as the Microsoft compiler. Supports only 32-bit Windows. Does not support the SSE and later instruction sets. Does not optimize as good as the Microsoft, Intel, Gnu and PathScale compilers.

Intel C++ compiler (parallel composer)

This compiler does not have its own IDE. It is intended as a plug-in to Microsoft Visual Studio when compiling for Windows and to Eclipse when compiling for Linux. It can also be used as a stand alone compiler when called from a command line or a make utility. It supports 32-bit and 64-bit Windows and 32-bit and 64-bit Linux as well as Intel-based Mac OS and Itanium systems.

The Intel compiler has a number of important optimization features:

- Very good support for vector operations using the single-instruction-multiple-data capabilities of the latest Intel instruction sets. This compiler can change simple code to vector code automatically (see page 106).

- Very good support for parallel processing on systems with multiple processors or multi-core processors. Can do automatic parallelization or explicit parallelization using the OpenMP directives.

- Supports CPU dispatch to make multiple code versions for different CPUs. (See page 132 for how to make this work on non-Intel processors).

- Comes with optimized math function libraries.

- Excellent support for vector intrinsic functions (see page 109)

- Excellent support for inline assembly on all platforms and the possibility of using the same inline assembly syntax in both Windows and Linux.

The most important disadvantage of the Intel compiler is that the compiled code may run with reduced speed or not at all on AMD and VIA processors. It is possible to avoid this problem by bypassing the so-called CPU-dispatcher that checks whether the code is running on an Intel CPU. See page 132 for details).

The Intel compiler is a good choice for code that can benefit from its many optimization features and for code that is ported to multiple operating systems.

Gnu

This is one of the best optimizing compilers available, though less user friendly. It is free and open source. It comes with most distributions of Linux, BSD and Mac OS X, 32-bit and 64-bit. Supports OpenMP. Supports vector intrinsics and automatic vectorization (see page 106). The Gnu function libraries are not fully optimized yet. Supports both AMD and Intel vector math libraries. The Gnu C++ compiler is available for many platforms, including 32-bit and 64-bit Windows. The Windows version is not always fully up to date. The Gnu compiler is the first choice for all Unix platforms.

PathScale

C++ compiler for 32- and 64-bit Linux. Has many good optimization options. Supports parallel processing, OpenMP and automatic vectorization. It is possible to insert optimization hints as pragmas in the code to tell the compiler e.g. how often a part of the code is executed. Optimizes very well. This compiler is a good choice for Linux platforms if the bias of the Intel compiler in favor of Intel CPUs cannot be tolerated.

PGI

C++ compiler for 32- and 64-bit Windows, Linux and Mac. Supports parallel processing, OpenMP and automatic vectorization. Optimizes reasonably well. Very poor performance for vector intrinsics.

Digital Mars

This is a cheap compiler for 32-bit Windows, including an IDE. Does not optimize well.

Open Watcom

Another open source compiler for 32-bit Windows. Does not, by default, conform to the standard calling conventions. Optimizes reasonably well.

Codeplay VectorC

A commercial compiler for 32-bit Windows. Integrates into the Microsoft Visual Studio IDE. Has not been updated since 2004. Can do automatic vectorization. Optimizes moderately well. Supports three different object file formats.

Comments

All of these compilers can be used as command-line versions without an IDE. Free trial versions are available for the commercial compilers.

Mixing object files from different compilers is generally possible on Linux platforms, and in some cases on Windows platforms. The Microsoft and Intel compilers for Windows are fully compatible on the object file level, and the Digital Mars compiler is mostly compatible with these. The CodeGear, Codeplay and Watcom compilers are not compatible with other compilers at the object file level.

My recommendation for good code performance is to use the Gnu, PathScale or Intel compiler for Unix applications and the Intel or Microsoft compiler for Windows applications.

The choice of compiler may in some cases be determined by the requirements of compatibility with legacy code, specific preferences for the IDE, for debugging facilities, easy GUI development, database integration, web application integration, mixed language programming, etc. In cases where the chosen compiler doesn't provide the best optimization it may be useful to make the most critical modules with a different compiler. Object files generated by the Intel and PathScale compilers can in most cases be linked into projects made with Microsoft or Gnu compilers without problems if the necessary library files are also included. Combining the Borland compiler with other compilers or function libraries is more difficult. The functions must have `extern "C"` declaration and the object files need to be

converted to OMF format. Alternatively, make a DLL with the best compiler and call it from a project built with another compiler.

2.6 Choice of function libraries

Some applications spend most of their execution time on executing library functions. Time-consuming library functions often belong to one of these categories:

- File input/output

- Graphics and sound processing

- Memory and string manipulation

- Mathematical functions

- Encryption, decryption, data compression

Most compilers include standard libraries for many of these purposes. Unfortunately, the standard libraries are not always fully optimized.

Library functions are typically small pieces of code that are used by many users in many different applications. Therefore, it is worthwhile to invest more efforts in optimizing library functions than in optimizing application-specific code. The best function libraries are highly optimized, using assembly language and automatic CPU-dispatching (see page 125) for the latest instruction set extensions.

If a profiling (see page 16) shows that a particular application uses a lot of CPU-time in library functions, or if this is obvious, then it may be possible to improve the performance significantly simply by using a different function library. If the application uses most of its time in library functions then it may not be necessary to optimize anything else than finding the most efficient library and economize the library function calls. It is recommended to try different libraries and see which one works best.

Some common function libraries are discussed below. Many libraries for special purposes are also available.

Microsoft

Comes with Microsoft compiler. Some functions are optimized well, others are not. Supports 32-bit and 64-bit Windows.

Borland / CodeGear / Embarcadero

Comes with the Borland C++ builder. Not optimized for SSE2 and later instruction sets. Supports only 32-bit Windows.

Gnu

Comes with the Gnu compiler. Not optimized as good as the compiler itself is. The 64-bit version is better than the 32-bit version. The Gnu compiler often inserts built-in code instead of the most common memory and string instructions. The built-in code is not optimal. Use option `-fno-builtin` to get library versions instead. The Gnu libraries support 32-bit and 64-bit Linux and BSD. The Windows version is currently not up to date.

Mac

The libraries included with the Gnu compiler for Mac OS X (Darwin) are part of the Xnu project. Some of the most important functions are included in the operating system kernel in the so-called commpage. These functions are highly optimized for the Intel Core and later

Intel processors. AMD processors and earlier Intel processors are not supported at all. Can only run on Mac platform.

Intel

The Intel compiler includes standard function libraries. Several special purpose libraries are also available, such as the "Intel Math Kernel Library" and "Integrated Performance Primitives". These function libraries are highly optimized for large data sets. However, the Intel libraries do not always work well on AMD and VIA processors. See page 132 for an explanation and possible workaround. Supports all x86 and x86-64 platforms.

AMD

AMD Math core library contains optimized mathematical functions. It also works on Intel processors. The performance is inferior to the Intel libraries. Supports 32- and 64-bit Windows and Linux.

Asmlib

My own function library made for demonstration purposes. Available from www.agner.org/optimize/asmlib.zip. Currently includes optimized versions of the functions `memcpy`, `memmove`, `memset`, `strlen`, `strcpy`, `strcat` and random number generators. Faster than most other libraries when running on the newest processors. Supports all x86 and x86-64 platforms.

Comparison of function libraries

Test	Processor	Microsoft	CodeGear	Intel	Mac	Gnu 32-bit	Gnu 32-bit -fno-builtin	Gnu 64 bit -fno-builtin	Asmlib
`memcpy` 16kB aligned operands	Intel Core 2	0.12	0.18	0.12	0.11	0.18	0.18	0.18	0.11
`memcpy` 16kB unaligned op.	Intel Core 2	0.63	0.75	0.18	0.11	1.21	0.57	0.44	0.12
`memcpy` 16kB aligned operands	AMD Opteron K8	0.24	0.25	0.24	n.a.	1.00	0.25	0.28	0.22
`memcpy` 16kB unaligned op.	AMD Opteron K8	0.38	0.44	0.40	n.a.	1.00	0.35	0.29	0.28
`strlen` 128 bytes	Intel Core 2	0.77	0.89	0.40	0.30	4.5	0.82	0.59	0.27
`strlen` 128 bytes	AMD Opteron K8	1.09	1.25	1.61	n.a.	2.23	0.95	0.6	1.19

Table 2.1. Comparing performance of different function libraries.
Numbers in the table are core clock cycles per byte of data (low numbers mean good performance). Aligned operands means that source and destination both have addresses divisible by 16.
Library versions tested:
Microsoft Visual studio 2008, v. 9.0
CodeGear Borland bcc, v. 5.5
Mac: Darwin8 g++ v 4.0.1.
Gnu: Glibc v. 2.7, 2.8.
Asmlib: v. 2.00.
Intel C++ compiler, v. 10.1.020. Functions `_intel_fast_memcpy` and `__intel_new_strlen` in library `libircmt.lib`. Function names are undocumented.

The best optimized libraries use the largest available registers for moving data. These work faster on processors with 128-bit internal data paths (e.g. Intel Core 2, AMD K10) than on older processors with 64-bit internal data paths (e.g. Intel Pentium 4, AMD Opteron K8). Tests on AMD K10 have not been made yet.

2.7 Choice of user interface framework

Most of the code in a typical software project goes to the user interface. Applications that are not computationally intensive may very well spend more CPU time on the user interface than on the essential task of the program.

Application programmers rarely program their own graphical user interfaces from scratch. This would not only be a waste of the programmers' time, but also inconvenient to the end user. Menus, buttons, dialog boxes, etc. should be as standardized as possible for usability reasons. The programmer can use standard user interface elements that come with the operating system or libraries that come with compilers and development tools.

A popular user interface library for Windows and C++ is Microsoft Foundation Classes (MFC). A competing product is Borland's now discontinued Object Windows Library (OWL). Several graphical interface frameworks are available for Linux systems. The user interface library can be linked either as a runtime DLL or a static library. A runtime DLL takes more memory resources than a static library, except when several applications use the same DLL at the same time.

A user interface library may be bigger than the application itself and take more time to load. A light-weight alternative is the Windows Template Library (WTL). A WTL application is generally faster and more compact than an MFC application. The development time for WTL applications can be expected to be higher due to poor documentation and lack of advanced development tools.

The simplest possible user interface is obtained by dropping the graphical user interface and use a console mode program. The inputs for a console mode program are typically specified on a command line or an input file. The output goes to the console or to an output file. A console mode program is fast, compact, and simple to develop. It is easy to port to different platforms because it doesn't depend on system-specific graphical interface calls. The usability may be poor because it lacks the self-explaining menus of a graphical user interface. A console mode program is useful for calling from other applications such as a make utility.

The conclusion is that the choice of user interface framework must be a compromise between development time, usability, program compactness, and execution time. No universal solution is best for all applications.

2.8 Overcoming the drawbacks of the C++ language

While C++ has many advantages when it comes to optimization, it does have some disadvantages that make developers choose other programming languages. This section discusses how to overcome these disadvantages when C++ is chosen for the sake of optimization.

Portability

C++ is fully portable in the sense that the syntax is fully standardized and supported on all major platforms. However, C++ is also a language that allows direct access to hardware interfaces and system calls. These are of course system-specific. In order to facilitate porting between platforms, it is recommended to put the user interface and other system-specific parts of the code in a separate module, and put the task-specific part of the code, which supposedly is system-independent, in another module.

The size of integers and other hardware-related details depend on the hardware platform and operating system. See page 30 for details.

Development time

Some developers feel that a particular programming language and development tool is faster to use than others. While some of the difference is simply a matter of habit, it is true that some development tools have powerful facilities that do much of the trivial programming work automatically. The development time and maintainability of C++ projects can be improved by consistent modularity and reusable classes.

Security

The most serious problem with the C++ language relates to security. Standard C++ implementations have no checking for array bounds violations and invalid pointers. This is a frequent source of errors in C++ programs and also a possible point of attack for hackers. It is necessary to adhere to certain programming principles in order to prevent such errors in programs where security matters.

Problems with invalid pointers can be avoided by using references instead of pointers, by initializing pointers to zero, by setting pointers to zero whenever the objects they point to become invalid, and by avoiding pointer arithmetics and pointer type casting. Linked lists and other data structures that typically use pointers may be replaced by more efficient container class templates, see page 92. Avoid the function `scanf`.

Violation of array bounds is probably the most common cause of errors in C++ programs. Writing past the end of an array can cause other variables to be overwritten, and even worse, it can overwrite the return address of the function in which the array is defined. This can cause all kinds of strange and unexpected behaviors. Arrays are often used as buffers for storing text or input data. A missing check for buffer overflow on input data is a common error that hackers often have exploited.

A good way to prevent such errors is to replace arrays by well-tested container classes. The standard template library (STL) is a useful source of such container classes. Unfortunately, many standard container classes use dynamic memory allocation in an inefficient way. See page and 90 for examples of how to avoid dynamic memory allocation. See page 92 for discussion of efficient container classes. An appendix to this manual at www.agner.org/optimize/cppexamples.zip contains examples of arrays with bounds checking and various efficient container classes.

Text strings are particularly problematic because there may be no certain limit to the lengths. The old C-style method of storing strings in character arrays is fast and efficient, but not safe unless the length of each string is checked before storing. The standard solution to this problem is to use string classes, such as `string` or `CString`. This is safe and flexible, but quite inefficient in large applications. The string classes allocate a new memory block every time a string is created or modified. This can cause the memory to be fragmented and involve a high overhead cost of heap management and garbage collection. A more efficient solution that doesn't compromise safety is to store all strings in one memory pool. See the examples in the appendix at www.agner.org/optimize/cppexamples.zip for how to store strings in a memory pool.

You may deviate from the above security advices in critical parts of the code where speed is important. This can be permissible only if the unsafe code is limited to well-tested functions, classes, templates or modules with a well-defined interface to the rest of the program.

3 Finding the biggest time consumers

3.1 How much is a clock cycle?

In this manual, I am using CPU clock cycles rather than seconds or microseconds as a time measure. This is because computers have very different speeds. If I write that something takes 10 μs today, then it may take only 5 μs on the next generation of computers and my manual will soon be obsolete. But if I write that something takes 10 clock cycles then it will still take 10 clock cycles even if the CPU clock frequency is doubled.

The length of a clock cycle is the reciprocal of the clock frequency. For example, if the clock frequency is 2 GHz then the length of a clock cycle is

$$\frac{1}{2\text{GHz}} = 0.5\text{ns}.$$

A clock cycle on one computer is not always comparable to a clock cycle on another computer. The Pentium 4 (NetBurst) CPU is designed for a higher clock frequency than other CPUs, but it uses more clock cycles than other CPUs for executing the same piece of code in general.

Assume that a loop in a program repeats 1000 times and that there are 100 floating point operations (addition, multiplication, etc.) inside the loop. If each floating point operation takes 5 clock cycles, then we can roughly estimate that the loop will take 1000 * 100 * 5 * 0.5 ns = 250 μs on a 2 GHz CPU. Should we try to optimize this loop? Certainly not! 250 μs is less than 1/50 of the time it takes to refresh the screen. There is no way the user can see the delay. But if the loop is inside another loop that also repeats 1000 times then we have an estimated calculation time of 250 ms. This delay is just long enough to be noticeable but not long enough to be annoying. We may decide to do some measurements to see if our estimate is correct or if the calculation time is actually more than 250 ms. If the response time is so long that the user actually has to wait for a result then we will consider if there is something that can be improved.

3.2 Use a profiler to find hot spots

Before you start to optimize anything, you have to identify the critical parts of the program. In some programs, more than 99% of the time is spent in the innermost loop doing mathematical calculations. In other programs, 99% of the time is spent on reading and writing data files while less than 1% goes to actually doing something on these data. It is very important to optimize the parts of the code that matters rather than the parts of the code that use only a small fraction of the total time. Optimizing less critical parts of the code will not only be a waste of time, it also makes the code less clear and more difficult to debug and maintain.

Most compiler packages include a profiler that can tell how many times each function is called and how much time it uses. There are also third-party profilers such as AQtime, Intel VTune and AMD CodeAnalyst.

There are several different profiling methods:

- Instrumentation: The compiler inserts extra code at each function call to count how many times the function is called and how much time it takes.

- Debugging. The profiler inserts temporary debug breakpoints at every function or every code line.

- Time-based sampling: The profiler tells the operating system to generate an interrupt, e.g. every millisecond. The profiler counts how many times an interrupt occurs in each

part of the program. This requires no modification of the program under test, but is less reliable.

- Event-based sampling: The profiler tells the CPU to generate interrupts at certain events, for example every time a thousand cache misses have occurred. This makes it possible to see which part of the program has most cache misses, branch mispredictions, floating point exceptions, etc. Event-based sampling requires a CPU-specific profiler. For Intel CPUs use Intel VTune, for AMD CPUs use AMD CodeAnalyst.

Unfortunately, profilers are often unreliable. They sometimes give misleading results or fail completely because of technical problems.

Some common problems with profilers are:

- Coarse time measurement. If time is measured with millisecond resolution and the critical functions take microseconds to execute then measurements can become imprecise or simply zero.

- Execution time too small or too long. If the program under test finishes in a short time then the sampling generates too little data for analysis. If the program takes too long time to execute then the profiler may sample more data than it can handle.

- Waiting for user input. Many programs spend most of their time waiting for user input or network resources. This time is included in the profile. It may be necessary to modify the program to use a set of test data instead of user input in order to make profiling feasible.

- Interference from other processes. The profiler measures not only the time spent in the program under test but also the time used by all other processes running on the same computer, including the profiler itself.

- Function addresses are obscured in optimized programs. The profiler identifies any hot spots in the program by their address and attempts to translate these addresses to function names. But a highly optimized program is often reorganized in such a way that there is no clear correspondence between function names and code addresses. The names of inlined functions may not be visible at all to the profiler. The result will be misleading reports of which functions take most time.

- Uses debug version of the code. Some profilers require that the code you are testing contains debug information in order to identify individual functions or code lines. The debug version of the code is not optimized.

- Jumps between CPU cores. A process or thread does not necessarily stay in the same processor core on multi-core CPUs, but event-counters do. This results in meaningless event counts for threads that jump between multiple CPU cores. You may need to lock a thread to a specific CPU core by setting a thread affinity mask.

- Poor reproducibility. Delays in program execution may be caused by random events that are not reproducible. Such events as task switches and garbage collection can occur at random times and make parts of the program appear to take longer time than normally.

There are various alternatives to using a profiler. A simple alternative is to run the program in a debugger and press break while the program is running. If there is a hot spot that uses 90% of the CPU time then there is a 90% chance that the break will occur in this hot spot. Repeating the break a few times may be enough to identify a hot spot. Use the call stack in the debugger to identify the circumstances around the hot spot.

Sometimes, the best way to identify performance bottlenecks is to put measurement instruments into the code rather than using a ready-made profiler. This does not solve all the problems associated with profiling, but it often gives more reliable results. If you are not satisfied with the way a profiler works then you may put the desired measurement instruments into the program itself. You may add counter variables that count how many times each part of the program is executed. Furthermore, you may read the time before and after each of the most important or critical parts of the program to measure how much time each part takes.

Your measurement code should have `#if` directives around it so that it can be disabled in the final version of the code. Inserting your own profiling instruments in the code itself is a very useful way to keep track of the performance during the development of a program.

The time measurements may require a very high resolution if time intervals are short. In Windows, you can use the `GetTickCount` or `QueryPerformanceCounter` functions for millisecond resolution. A much higher resolution can be obtained with the time stamp counter in the CPU, which counts at the CPU clock frequency. The `ReadTSC` function in the function library at www.agner.org/optimize/asmlib.zip gives easy access to this counter.

The time stamp counter becomes invalid if a thread jumps between different CPU cores. You may have to fix the thread to a specific CPU core during time measurements to avoid this. (In Windows, `SetThreadAffinityMask`, in Linux, `sched_setaffinity`).

The program should be tested with a realistic set of test data. The test data should contain a typical degree of randomness in order to get a realistic number of cache misses and branch mispredictions.

When the most time-consuming parts of the program have been found, then it is important to focus the optimization efforts on the time consuming parts only.

If a library function or any other small piece of code is particularly critical then it may be useful to measure the number of cache misses, branch mispredictions, floating point exceptions, etc. in this piece of code. My test programs at www.agner.org/optimize/testp.zip are useful for this purpose. See page 155 for details.

A profiler is most useful for finding problems that relate to CPU-intensive code. But many programs use more time loading files or accessing databases, network and other resources than doing arithmetic operations. The most common time-consumers are discussed in the following sections.

3.3 Program installation

The time it takes to install a program package is not traditionally considered a software optimization issue. But it is certainly something that can steal the user's time. The time it takes to install a software package and make it work cannot be ignored if the goal of software optimization is to save time for the user. With the high complexity of modern software, it is not unusual for the installation process to take more than an hour. Neither is it unusual that a user has to reinstall a software package several times in order to find and resolve compatibility problems.

Software developers should take installation time and compatibility problems into account when deciding whether to base a software package on a complex framework requiring many files to be installed.

The installation process should always use standardized installation tools. It should be possible to select all installation options at the start so that the rest of the installation

process can proceed unattended. Uninstallation should also proceed in a standardized manner.

3.4 Automatic updates

Many software programs automatically download updates through the Internet at regular time intervals. Some programs search for updates every time the computer starts up, even if the program is never used. A computer with many such programs installed can take several minutes to start up, which is a total waste of the user's time. Other programs use time searching for updates each time the program starts. The user may not need the updates if the current version satisfies the user's needs. The search for updates should be optional and off by default unless there is a compelling security reason for updating. The update process should run in a low priority thread, and only if the program is actually used. No program should leave a background process running when it is not in use. The installation of downloaded program updates should be postponed until the program is shut down and restarted anyway.

3.5 Program loading

Often, it takes more time to load a program than to execute it. The load time can be annoyingly high for programs that are based on big runtime frameworks, intermediate code, interpreters, just-in-time compilers, etc., as is commonly the case with programs written in Java, C#, Visual Basic, etc.

But program loading can be a time-consumer even for programs implemented in compiled C++. This typically happens if the program uses a lot of runtime DLL's (dynamically linked libraries, also called shared objects), resource files, configuration files, help files and databases. The operating system may not load all the modules of a big program when the program starts up. Some modules may be loaded only when they are needed, or they may be swapped to the hard disk if the RAM size is insufficient.

The user expects immediate responses to simple actions like a key press or mouse move. It is unacceptable to the user if such a response is delayed for several seconds because it requires the loading of modules or resource files from disk. Memory-hungry applications force the operating system to swap memory to disk. Memory swapping is a frequent cause of unacceptably long response times to simple things like a mouse move or key press.

Avoid an excessive number of DLLs, configuration files, resource files, help files etc. scattered around on the hard disk. A few files, preferably in the same directory as the `.exe` file, is acceptable.

3.6 Dynamic linking and position-independent code

Function libraries can be implemented either as static link libraries (`*.lib`, `*.a`) or dynamic link libraries, also called shared objects (`*.dll`, `*.so`). The mechanism of static linking is that the linker extracts the functions that are needed from the library file and copies them into the executable file. Only the executable file needs to be distributed to the end user.

Dynamic linking works differently. The link to a function in a dynamic library is resolved at run time. Therefore, both the executable file and one or more DLLs are loaded into memory when the program is run. Both the executable file and all the DLLs need to be distributed to the end user.

The advantages of using static linking rather than dynamic linking are:

- Static linking includes only the part of the library that is actually needed by the application, while dynamic linking makes the entire library (or at least a large part of it)

load into memory even when just a single function is needed.

- All the code is included in a single executable file when static linking is used. Dynamic linking makes it necessary to load several files when the program is started.

- It takes longer time to call a function in a dynamic link library than in a static link library because it needs an extra jump through a pointer in an import table.

- The memory space becomes more fragmented when the code is distributed between multiple DLLs. The DLLs are always loaded at round memory addresses divisible by the memory page size (4096). This will make all DLLs contend for the same cache lines. This makes code caching and data caching less efficient.

- DLLs are less efficient in some systems because of the needs of position-independent code, see below.

- Installing a second application that uses a newer version of the same DLL can change the behavior of the first application if dynamic linking is used, but not if static linking is used.

The advantages of dynamic linking are:

- Multiple applications running simultaneously can share the same DLL without the need to load more than one instance of the DLL into memory. This is useful on servers that run many processes simultaneously.

- A DLL can be updated to a new version without the need to update the program that calls it.

- A DLL can be called from programming languages that do not support static linking.

- A DLL can be useful for making plug-ins that add functionality to an existing program.

Weighing the above advantages of each method, it is clear that static linking is preferable in critical applications except for parts of the code that are rarely used.

Many function libraries are available in both static and dynamic versions. It is recommended to use the static version if a performance gain can be expected.

Some systems allow lazy binding of function calls. The principle of lazy binding is that the address of a linked function is not resolved when the program is loaded, but waits until the first time the function is called. Lazy binding can be useful if there is a low probability that the function will be called in a single session. But lazy binding definitely degrades performance if the function is called. A considerable delay comes when the function is called for the first time. This delay is caused by the need to load the dynamic linker, and in some cases from purging the code cache because of self-modifying code.

The delay on lazy binding leads to a usability problem in interactive programs because the response time to e.g. a mouse click becomes inconsistent and sometimes unacceptably long. I will not recommend lazy binding where it can be avoided.

The memory address at which a DLL is loaded cannot be determined in advance, because a fixed address might clash with another DLL requiring the same address. This problem is solved in Linux, BSD and Mac systems by using position-independent code. Unfortunately, this leads to a considerable extra overhead. There are two reasons for this overhead. The first reason is that the 32-bit instruction set has no support for self-relative addressing of data. The code needs to use an extra function call for setting up a register that points to the

data. This is done in every function that needs access to global, static or constant data. The result is that a register is wasted and that an extra overhead is incurred after almost every function call. The second reason is that Linux systems access all static data through pointers stored in a global offset table when position-independent code is used. This leads to extra instructions for loading a pointer from the global offset table every time a global, static or constant data item is accessed.

The 64-bit instruction set supports self-relative addressing of data. Therefore, there is less extra cost to position-independent code in 64-bit systems. Unfortunately, current compilers for 64-bit Linux still use a global offset table for position independent code although it would be possible to access the data without such a table.

The Windows system uses a different method for allowing a DLL to be loaded at an arbitrary address. All addresses in the DLL are simply modified when the DLL is loaded if the memory address is different from a predefined value. This process is known as relocation. A DLL may contain thousands of addresses to relocate, but the relocation process is nevertheless very fast because the address calculations are simple and are done only once.

The conclusion is that there is a significant runtime cost for position-independent code when using dynamic linking (shared objects) in Linux, BSD and Mac systems. Critical CPU-intensive code should never be stored in a shared object on these systems.

Note that many compilers for Mac OS X make position-independent code and lazy binding by default, even when shared objects are not used. Be sure to specify static linking and no position-independent code (`-fno-pic`) for best performance when building code for the Mac system.

3.7 File access

Reading or writing a file on a hard disk often takes much more time than processing the data in the file, especially if the user has a virus scanner that scans all files on access.

Sequential forward access to a file is faster than random access. Reading or writing big blocks is faster than reading or writing a small bit at a time. Do not read or write less than a few kilobytes at a time.

You may mirror the entire file in a memory buffer and read or write it in one operation rather than reading or writing small bits in a non-sequential manner.

It is usually much faster to access a file that has been accessed recently than to access it the first time. This is because the file has been copied to the disk cache.

Files on remote or removable media such as floppy disks and USB sticks may not be cached. This can have quite dramatic consequences. I once made a Windows program that created a file by calling `WritePrivateProfileString`, which opens and closes the file for each line written. This worked sufficiently fast on a hard disk because of disk caching, but it took several minutes to write the file to a floppy disk.

A big file containing numerical data is more compact and efficient if the data are stored in binary form than if the data are stored in ASCII form. A disadvantage of binary data storage is that it is not human readable and not easily ported to systems with big-endian storage.

Optimizing file access is more important than optimizing CPU use in programs that have many file input/output operations. It can be advantageous to put file access in a separate thread if there is other work that the processor can do while waiting for disk operations to finish.

3.8 System database

It can take several seconds to access the system database in Windows. It is more efficient to store application-specific information in a separate file than in the big registration database in the Windows system. Note that the system may store the information in the database anyway if you are using functions such as `GetPrivateProfileString` and `WritePrivateProfileString` to read and write configuration files (`*.ini` files).

3.9 Other databases

Many software applications use a database for storing user data. A database can consume a lot of CPU time, RAM and disk space. It may be possible to replace a database by a plain old data file in simple cases. Database queries can often be optimized by using indexes, working with sets rather than loops, etc. Optimizing database queries is beyond the scope of this manual, but you should be aware that there is often a lot to gain by optimizing database access.

3.10 Graphics

A graphical user interface can use a lot of computing resources. Typically, a specific graphics framework is used. The operating system may supply such a framework in its API. In some cases, there is an extra layer of a third-party graphics framework between the operating system API and the application software. Such an extra framework can consume a lot of extra resources.

Each graphics operation in the application software is implemented as a function call to a graphics library or API function which then calls a device driver. A call to a graphics function is time consuming because it may go through multiple layers and it needs to switch to protected mode and back again. Obviously, it is more efficient to make a single call to a graphics function that draws a whole polygon or bitmap than to draw each pixel or line separately through multiple function calls.

The calculation of graphics objects in computer games and animations is of course also time consuming, especially if there is no graphics processing unit.

Various graphics function libraries and drivers differ a lot in performance. I have no specific recommendation of which is best.

3.11 Other system resources

Writes to a printer or other device should preferably be done in big blocks rather than a small piece at a time because each call to a driver involves the overhead of switching to protected mode and back again.

Accessing system devices and using advanced facilities of the operating system can be time consuming because it may involve the loading of several drivers, configuration files and system modules.

3.12 Network access

Some application programs use internet or intranet for automatic updates, remote help files, data base access, etc. The problem here is that access times cannot be controlled. The network access may be fast in a simple test setup but slow or completely absent in a use situation where the network is overloaded or the user is far from the server.

These problems should be taken into account when deciding whether to store help files and other resources locally or remotely. If frequent updates are necessary then it may be optimal to mirror the remote data locally.

Access to remote databases usually requires log on with a password. The log on process is known to be an annoying time consumer to many hard working software users. In some cases, the log on process may take more than a minute if the network or database is heavily loaded.

3.13 Memory access

Accessing data from RAM memory can take quite a long time compared to the time it takes to do calculations on the data. This is the reason why all modern computers have memory caches. Typically, there is a level-1 data cache of 8 - 64 Kbytes and a level-2 cache of 256 Kbytes to 2 Mbytes. There may also be a level-3 cache.

If the combined size of all data in a program is bigger than the level-2 cache and the data are scattered around in memory or accessed in a non-sequential manner then it is likely that memory access is the biggest time-consumer in the program. Reading or writing to a variable in memory takes only 2-3 clock cycles if it is cached, but several hundred clock cycles if it is not cached. See page 27 about data storage and page 87 about memory caching.

3.14 Context switches

A context switch is a switch between different tasks in a multitasking environment, between different threads in a multithreaded program, or between different parts of a big program. Frequent context switches can reduce the performance because the contents of data cache, code cache, branch target buffer, branch pattern history, etc. may have to be renewed.

Context switches are more frequent if the time slices allocated to each task or thread are smaller. The lengths of the time slices is determined by the operating system, not by the application program.

The number of context switches is smaller in a computer with multiple CPUs or a CPU with multiple cores.

3.15 Dependency chains

Modern microprocessors can do out-of-order execution. This means that if a piece of software specifies the calculation of A and then B, and the calculation of A is slow, then the microprocessor can begin the calculation of B before the calculation of A is finished. Obviously, this is only possible if the value of A is not needed for the calculation of B.

In order to take advantage of out-of-order execution, you have to avoid long dependency chains. A dependency chain is a series of calculations, where each calculation depends on the result of the preceding one. This prevents the CPU from doing multiple calculations simultaneously or out of order. See page 103 for examples of how to break a dependency chain.

3.16 Execution unit throughput

There is an important distinction between the latency and the throughput of an execution unit. For example, it may take three clock cycles to do a floating point addition on a modern CPU. But it is possible to start a new floating point addition every clock cycle. This means that if each addition depends on the result of the preceding addition then you will have only

one addition every three clock cycles. But if all the additions are independent then you can have one addition every clock cycle.

The highest performance that can possibly be obtained in a computationally intensive program is achieved when none of the time-consumers mentioned in the above sections are dominating and there are no long dependency chains. In this case, the performance is limited by the throughput of the execution units rather than by the latency or by memory access.

The execution core of modern microprocessors is split between several execution units. Typically, there are two or more integer units, one floating point addition unit, and one floating point multiplication unit. This means that it is possible to do an integer addition, a floating point addition, and a floating point multiplication at the same time.

A code that does floating point calculations should therefore preferably have a balanced mix of additions and multiplications. Subtractions use the same unit as additions. Divisions take longer time and use the multiplication unit. It is possible to do integer operations in-between the floating point operations without reducing the performance because the integer operations use different execution units. For example, a loop that does floating point calculations will typically use integer operations for incrementing a loop counter, comparing the loop counter with its limit, etc. In most cases, you can assume that these integer operations do not add to the total computation time.

4 Performance and usability

A better performing software product is one that saves time for the user. Time is a precious resource for many computer users and much time is wasted on software that is slow, difficult to use, incompatible or error prone. All these problems are usability issues, and I believe that software performance should be seen in the broader perspective of usability. A list of literature on usability is given on page 165.

This is not a manual on usability, but I think that it is necessary here to draw the attention of software programmers to some of the most common obstacles to efficient use of software. The following list points out some typical sources of frustration and waste of time for software users as well as important usability problems that software developers should be aware of.

- Big runtime frameworks. The .NET framework and the Java virtual machine are frameworks that typically take much more resources than the programs they are running. Such frameworks are frequent sources of resource problems and compatibility problems and they waste a lot of time both during installation of the framework itself, during installation of the program that runs under the framework, during start of the program, and while the program is running. The main reason why such runtime frameworks are used at all is for the sake of cross-platform portability. Unfortunately, the cross-platform compatibility is not always as good as expected. I believe that the portability could be achieved more efficiently by better standardization of programming languages, operating systems, and API's.

- Memory swapping. Software developers typically have more powerful computers with more RAM than end users have. The developers may therefore fail to see the excessive memory swapping and other resource problems that cause the resource-hungry applications to perform poorly for the end user.

- Installation problems. The procedures for installation and uninstallation of programs should be standardized and done by the operating system rather than by individual

installation tools.

- Automatic updates. Automatic updating of software can cause problems if the network is unstable or if the new version causes problem that were not present in the old version. Updating mechanisms often disturb the users with nagging pop-up messages saying please install this important new update or even telling the user to restart the computer while he or she is busy concentrating on important work. The updating mechanism should never interrupt the user but only show a discrete icon signaling the availability of an update, or update automatically when the computer is restarted anyway. Software distributors are often abusing the update mechanism to advertise new versions of their software. This is annoying to the user.

- Compatibility problems. All software should be tested on different platforms, different screen resolutions, different system color settings and different user access rights. Software should use standard API calls rather than self-styled hacks and direct hardware access. Available protocols and standardized file formats should be used. Web systems should be tested in different browsers, different platforms, different screen resolutions, etc. Accessibility guidelines should be obeyed (See literature page 165).

- Copy protection. Some copy protection schemes are based on hacks that violate or circumvent operating system standards. Such schemes are frequent sources of compatibility problems and system breakdown. Many copy protection schemes are based on hardware identification. Such schemes cause problems when the hardware is updated. Most copy protection schemes are annoying to the user and prevent legitimate backup copying without effectively preventing illegitimate copying. The benefits of a copy protection scheme should be weighed against the costs in terms of usability problems and necessary support.

- Hardware updating. The change of a hard disk or other hardware often requires that all software be reinstalled and user settings are lost. It is not unusual for the reinstallation work to take a whole workday or more. Many software applications need better backup features, and current operating systems need better support for hard disk copying.

- Security. The vulnerability of software with network access to virus attacks and other abuse is extremely costly to many users.

- Background services. Many services that run in the background are unnecessary for the user and a waste of resources. Consider running the services only when activated by the user.

- Take user feedback seriously. User complaints should be regarded as a valuable source of information about bugs, compatibility problems, usability problems and desired new features. User feedback should be handled in a systematic manner to make sure the information is utilized appropriately. Users should get a reply about investigation of the problems and planned solutions. Patches should be easily available from a website.

5 Choosing the optimal algorithm

The first thing to do when you want to optimize a piece of CPU-intensive software is to find the best algorithm. The choice of algorithm is very important for tasks such as sorting, searching, and mathematical calculations. In such cases, you can obtain much more by choosing the best algorithm than by optimizing the first algorithm that comes to mind. In some cases you may have to test several different algorithms in order to find the one that works best on a typical set of test data.

That being said, I must warn against overkill. Don't use an advanced and complicated algorithm if a simple algorithm can do the job fast enough. For example, some programmers use a hash table for even the smallest list of data. A hash table can improve search times dramatically for very large data bases, but there is no reason to use it for lists that are so small that a binary search, or even a linear search, is fast enough. A hash table increases the size of the program as well as the size of data files. This can actually reduce speed if the bottleneck is file access or cache access rather than CPU time. Another disadvantage of complicated algorithms is that it makes program development more expensive and more error prone.

A discussion of different algorithms for different purposes is beyond the scope of this manual. You have to consult the general literature on algorithms and data structures for standard tasks such as sorting and searching, or the specific literature for more complicated mathematical tasks.

Before you start to code, you may consider whether others have done the job before you. Optimized function libraries for many standard tasks are available from a number of sources. For example, the Boost collection contains well-tested libraries for many common purposes (www.boost.org). The "Intel Math Kernel Library" contains many functions for common mathematical calculations including linear algebra and statistics, and the "Intel Performance Primitives" library contains many functions for audio and video processing, signal processing, data compression and cryptography (www.intel.com). If you are using an Intel function library then make sure it works well on non-Intel processors, as explained on page 132.

It is often easier said than done to choose the optimal algorithm before you start to program. Many programmers have discovered that there are smarter ways of doing things only after they have put the whole software project together and tested it. The insight you gain by testing and analyzing program performance and studying the bottlenecks can lead to a better understanding of the whole structure of the problem. This new insight can lead to a complete redesign of the program, for example when you discover that there are smarter ways of organizing the data.

A complete redesign of a program that already works is of course a considerable job, but it may be quite a good investment. A redesign can not only improve the performance, it is also likely to lead to a more well-structured program that is easier to maintain. The time you spend on redesigning a program may in fact be less than the time you would have spent fighting with the problems of the original, poorly designed program.

6 Development process

There is a considerable debate about which software development process and software engineering principles to use. I am not going to recommend any specific model. Instead, I will make a few comments about how the development process can influence the performance of the final product.

It is good to do a thorough analysis of the data structure, data flow and algorithms in the planning phase in order to predict which resources are most critical. However, there may be so many unknown factors in the early planning stage that a detailed overview of the problem cannot easily be obtained. In the latter case, you may view the software development work as a learning process where the main feedback comes from testing. Here, you should be prepared for several iterations of redesign.

Some software development models have a strict formalism that requires several layers of abstraction in the logical architecture of the software. You should be aware that there are

inherent performance costs to such a formalism. The splitting of software into an excessive number of separate layers of abstraction is a common cause of reduced performance.

Since most development methods are incremental or iterative in nature, it is important to have a strategy for saving a backup copy of every intermediate version. For one-man projects, it is sufficient to make a zip file of every version. For team projects, it is recommended to use a version control tool.

7 The efficiency of different C++ constructs

Most programmers have little or no idea how a piece of program code is translated into machine code and how the microprocessor handles this code. For example, many programmers do not know that double precision calculations are just as fast as single precision. And who would know that a template class is more efficient than a polymorphous class?

This chapter is aiming at explaining the relative efficiency of different C++ language elements in order to help the programmer choosing the most efficient alternative. The theoretical background is further explained in the other volumes in this series of manuals.

7.1 Different kinds of variable storage

Variables and objects are stored in different parts of the memory, depending on how they are declared in a C++ program. This has influence on the efficiency of the data cache (see page 87). Data caching is poor if data are scattered randomly around in the memory. It is therefore important to understand how variables are stored. The storage principles are the same for simple variables, arrays and objects.

Storage on the stack

Variables declared with the keyword `auto` are stored on the stack. The keyword `auto` is practically never used because automatic storage is the default for all variables and objects that are declared inside any function.

The stack is a part of memory that is organized in a first-in-last-out fashion. It is used for storing function return addresses (i.e. where the function was called from), function parameters, local variables, and for saving registers that have to be restored before the function returns. Every time a function is called, it allocates the required amount of space on the stack for all these purposes. This memory space is freed when the function returns. The next time a function is called, it can use the same space for the parameters of the new function.

The stack is the most efficient place to store data because the same range of memory addresses is reused again and again. If there are no big arrays, then it is almost certain that this part of the memory is mirrored in the level-1 data cache, where it is accessed quite fast.

The lesson we can learn from this is that all variables and objects should preferably be declared inside the function in which they are used.

It is possible to make the scope of a variable even smaller by declaring it inside { } brackets. However, most compilers do not free the memory used by a variable until the function returns even though it could free the memory when exiting the { } brackets in which the variable is declared. If the variable is stored in a register (see below) then it may be freed before the function returns.

Global or static storage

Variables that are declared outside of any function are called global variables. They can be accessed from any function. Global variables are stored in a static part of the memory. The static memory is also used for variables declared with the `static` keyword, for floating point constants, string constants, array initializer lists, `switch` statement jump tables, and virtual function tables.

The static data area is usually divided into three parts: one for constants that are never modified by the program, one for initialized variables that may be modified by the program, and one for uninitialized variables that may be modified by the program.

The advantage of static data is that they can be initialized to desired values before the program starts. The disadvantage is that the memory space is occupied throughout the whole program execution, even if the variable is only used in a small part of the program. This makes data caching less efficient.

Do not make variables global if you can avoid it. Global variables may be needed for communication between different threads, but that's about the only situation where they are unavoidable. It may be useful to make a variable global if it is accessed by several different functions and you want to avoid the overhead of transferring the variable as function parameter. But it may be a better solution to make the functions that access the saved variable members of the same class and store the shared variable inside the class. Which solution you prefer is a matter of programming style.

It is often preferable to make a lookup-table static. Example:

```
// Example 7.1
float SomeFunction (int x) {
   static float list[] = {1.1, 0.3, -2.0, 4.4, 2.5};
   return list[x];
}
```

The advantage of using `static` here is that the list does not need to be initialized when the function is called. The values are simply put there when the program is loaded into memory. If the word `static` is removed from the above example, then all five values have to be put into the list every time the function is called. This is done by copying the entire list from static memory to stack memory. Copying constant data from static memory to the stack is a waste of time in most cases, but it may be optimal in special cases where the data are used many times in a loop where almost the entire level-1 cache is used in a number of arrays that you want to keep together on the stack.

String constants and floating point constants are stored in static memory. Example:

```
// Example 7.2
a = b * 3.5;
c = d + 3.5;
```

Here, the constant `3.5` will be stored in static memory. Most compilers will recognize that the two constants are identical so that only one constant needs to be stored. All identical constants in the entire program will be joined together in order to minimize the amount of cache space used for constants.

Integer constants are usually included as part of the instruction code. You can assume that there are no caching problems for integer constants.

Register storage

A limited number of variables can be stored in registers instead of main memory. A register is a small piece of memory inside the CPU used for temporary storage. Variables that are

stored in registers are accessed very fast. All optimizing compilers will automatically choose the most often used variables in a function for register storage. The same register can be used for multiple variables as long as their uses (live ranges) don't overlap.

The number of registers is very limited. There are approximately six integer registers available for general purposes in 32-bit operating systems and fourteen integer registers in 64-bit systems.

Floating point variables use a different kind of registers. There are eight floating point registers available in 32-bit operating systems and sixteen in 64-bit operating systems. Some compilers have difficulties making floating point register variables in 32-bit mode unless the SSE2 instruction set is enabled.

Volatile

The `volatile` keyword specifies that a variable cannot be stored in a register, not even temporarily. This is necessary for variables that are accessed by more than one thread. Volatile storage prevents the compiler from doing any kind of optimization on the variable. It is sometimes used for turning off optimization of a particular variable.

Thread-local storage

Most compilers can make thread-local storage of static and global variables by using the keyword `__thread` or `__declspec(thread)`. Such variables have one instance for each thread. Thread-local storage is inefficient because it is accessed through a pointer stored in a thread environment block. Thread-local storage should be avoided, if possible, and replaced by storage on the stack (see above, p. 27). Variables stored on the stack always belong to the thread in which they are created.

Far

Systems with segmented memory, such as DOS and 16-bit Windows, allow variables to be stored in a far data segment by using the keyword `far` (arrays can also be `huge`). Far storage, far pointers, and far procedures are inefficient. If a program has too much data for one segment then it is recommended to use a different operating systems that allows bigger segments (32-bit or 64-bit systems).

Dynamic memory allocation

Dynamic memory allocation is done with the operators `new` and `delete` or with the functions `malloc` and `free`. These operators and functions consume a significant amount of time. A part of memory called the heap is reserved for dynamic allocation. The heap can easily become fragmented when objects of different sizes are allocated and deallocated in random order. The heap manager can spend a lot of time cleaning up spaces that are no longer used and searching for vacant spaces. This is called garbage collection. Objects that are allocated in sequence are not necessarily stored sequentially in memory. They may be scattered around at different places when the heap has become fragmented. This makes data caching inefficient.

Dynamic memory allocation also tends to make the code more complicated and error-prone. The program has to keep pointers to all allocated objects and keep track of when they are no longer used. It is important that all allocated objects are also deallocated in all possible cases of program flow. Failure to do so is a common source of error known as memory leak. An even worse kind of error is to access an object after it has been deallocated. The program logic may need extra overhead to prevent such errors.

See page 90 for a further discussion of the advantages and drawbacks of using dynamic memory allocation.

Some programming languages, such as Java, use dynamic memory allocation for all objects. This is of course inefficient.

Variables declared inside a class

Variables declared inside a class are stored in the order in which they appear in the class declaration. The type of storage is determined where the object of the class is declared. An object of a class, structure or union can use any of the storage methods mentioned above. An object cannot be stored in a register except in the simplest cases, but its data members can be copied into registers.

A class member variable with the `static` modifier will be stored in static memory and will have one and only one instance. Non-static members of the same class will be stored with each instance of the class.

Storing variables in a class or structure is a good way of making sure that variables that are used in the same part of the program are also stored near each other. See page 51 for the pros and cons of using classes.

7.2 Integers variables and operators

Integer sizes

Integers can be different sizes, and they can be signed or unsigned. The following table summarizes the different integer types available.

declaration	size, bits	minimum value	maximum value	in stdint.h
`char`	8	-128	127	`int8_t`
`short int` in 16-bit systems: `int`	16	-32768	32767	`int16_t`
`int` in 16-bit systems: `long int`	32	-2^{31}	$2^{31}-1$	`int32_t`
`long long` MS compiler: `__int64` 64-bit Linux: `long int`	64	-2^{63}	$2^{63}-1$	`int64_t`
`unsigned char`	8	0	255	`uint8_t`
`unsigned short int` in 16-bit systems: `unsigned int`	16	0	65535	`uint16_t`
`unsigned int` in 16-bit systems: `unsigned long`	32	0	$2^{32}-1$	`uint32_t`
`unsigned long long` MS compiler: `unsigned __int64` 64-bit Linux: `unsigned long int`	64	0	$2^{64}-1$	`uint64_t`
Table 7.1. Sizes of different integer types				

Unfortunately, the way of declaring an integer of a specific size is different for different platforms as shown in the above table. If the standard header file `stdint.h` or `inttypes.h` is available then it is recommended to use that for a portable way of defining integer types of a specific size.

Integer operations are fast in most cases, regardless of the size. However, it is inefficient to use an integer size that is larger than the largest available register size. In other words, it is inefficient to use 32-bit integers in 16-bit systems or 64-bit integers in 32-bit systems, especially if the code involves multiplication or division.

The compiler will always select the most efficient integer size if you declare an `int`, without specifying the size. Integers of smaller sizes (`char`, `short int`) are only slightly less efficient. In many cases, the compiler will convert these types to integers of the default size when doing calculations, and then use only the lower 8 or 16 bits of the result. You can assume that the type conversion takes zero or one clock cycle. In 64-bit systems, there is only a minimal difference between the efficiency of 32-bit integers and 64-bit integers, as long as you are not doing divisions.

It is recommended to use the default integer size in cases where the size doesn't matter and there is no risk of overflow, such as simple variables, loop counters, etc. In large arrays, it may be preferred to use the smallest integer size that is big enough for the specific purpose in order to make better use of the data cache. Bit-fields of sizes other than 8, 16, 32 and 64 bits are less efficient. In 64-bit systems, you may use 64-bit integers if the application can make use of the extra bits.

When considering whether a particular integer size is big enough for a specific purpose, you must consider if intermediate calculations can cause overflow. For example, in the expression `a = (b+c)/2`, it can happen that `(b+c)` overflows, even if `a`, `b` and `c` would all be below the maximum value. There is no automatic check for integer overflow.

Signed versus unsigned integers

In most cases, there is no difference in speed between using signed and unsigned integers. But there are a few cases where it matters:

- Division by a constant: Unsigned is faster than signed when you divide an integer with a constant (see page 143). This also applies to the modulo operator `%`.

- Conversion to floating point is faster with signed than with unsigned integers (see page 148).

- Overflow behaves differently on signed and unsigned variables. A positive overflow of a signed variable produces a negative result. An overflow of an unsigned variable produces a low positive result.

The conversion between signed and unsigned integers is costless. It is simply a matter of interpreting the same bits differently. A negative integer will be interpreted as a very large positive number when converted to unsigned.

```
// Example 7.3. Signed and unsigned integers
int a, b;
double c;
b = (unsigned int)a / 10;    // Convert to unsigned for fast division
c = a * 2.5;                 // Use signed when converting to double
```

In example 7.3 we are converting `a` to unsigned in order to make the division faster. Of course, this works only if it is certain that `a` will never be negative. The last line is implicitly converting `a` to `double` before multiplying with the constant `2.5`, which is `double`. Here we prefer `a` to be signed.

Be sure not to mix signed and unsigned integers in comparisons, such as `<`. The result of comparing signed with unsigned integers is ambiguous and may produce undesired results.

Integer operators

Integer operations are generally very fast. Simple integer operations such as addition, subtraction, comparison, bit operations and shift operations take only one clock cycle on most microprocessors.

Multiplication and division take longer time. Integer multiplication takes 11 clock cycles on Pentium 4 processors, and 3 - 4 clock cycles on most other microprocessors. Integer division takes 40 - 80 clock cycles, depending on the microprocessor. Integer division is faster the smaller the integer size on AMD processors, but not on Intel processors. Details about instruction latencies are listed in manual 4: "Instruction tables". Tips about how to speed up multiplications and divisions are given on page 142 and 143, respectively.

Increment and decrement operators

The pre-increment operator `++i` and the post-increment operator `i++` are as fast as additions. When used simply to increment an integer variable, it makes no difference whether you use pre-increment or post-increment. The effect is simply identical. For example,
`for (i=0; i<n; i++)` is the same as `for (i=0; i<n; ++i)`. But when the result of the expression is used, then there may be a difference in efficiency. For example,
`x = array[i++]` is more efficient than `x = array[++i]` because in the latter case, the calculation of the address of the array element has to wait for the new value of `i` which will delay the availability of `x` for approximately two clock cycles. Obviously, the initial value of `i` must be adjusted if you change pre-increment to post-increment.

There are also situations where pre-increment is more efficient than post-increment. For example, in the case `a = ++b;` the compiler will recognize that the values of `a` and `b` are the same after this statement so that it can use the same register for both, while the expression `a = b++;` will make the values of `a` and `b` different so that they cannot use the same register.

Everything that is said here about increment operators also applies to decrement operators on integer variables.

7.3 Floating point variables and operators

Modern microprocessors in the x86 family have two different types of floating point registers and correspondingly two different types of floating point instructions. Each type has advantages and disadvantages.

The original method of doing floating point operations involves eight floating point registers organized as a register stack. These registers have long double precision (80 bits). The advantages of using the register stack are:

- All calculations are done with long double precision.

- Conversions between different precisions take no extra time.

- There are intrinsic instructions for mathematical functions such as logarithms and trigonometric functions.

- The code is compact and takes little space in the code cache.

The register stack also has disadvantages:

- It is difficult for the compiler to make register variables because of the way the register stack is organized.

- Floating point comparisons are slow unless the Pentium-II or later instruction set is enabled.

- Conversions between integers and floating point numbers is inefficient.

- Division, square root and mathematical functions take more time to calculate when long double precision is used.

A newer method of doing floating point operations involves eight or sixteen vector registers (XMM or YMM) which can be used for multiple purposes. Floating point operations are done with single or double precision, and intermediate results are always calculated with the same precision as the operands. The advantages of using the vector registers are:

- It is easy to make floating point register variables.

- Vector operations are available for doing parallel calculations on vectors of two double precision or four single precision variables in the XMM registers (see page 105). If the AVX instruction set is available then each vector can hold four double precision or eight single precision variables in the YMM registers.

Disadvantages are:

- Long double precision is not supported.

- The calculation of expressions where operands have mixed precision require precision conversion instructions which can be quite time-consuming (see page 146).

- Mathematical functions must use a function library, but this is often faster than the intrinsic hardware functions.

The floating point stack registers are available in all systems that have floating point capabilities (except in device drivers for 64-bit Windows). The XMM vector registers are available in 64-bit systems and in 32-bit systems when the SSE2 or later instruction set is enabled (single precision requires only SSE). The YMM registers are available if the AVX instruction set is supported by the processor and the operating system. See page 125 for how to test for the availability of these instruction sets.

Most compilers will use the XMM registers for floating point calculations whenever they are available, i.e. in 64-bit mode or when the SSE2 instruction set is enabled. Few compilers are able to mix the two types of floating point operations and choose the type that is optimal for each calculation.

In most cases, double precision calculations take no more time than single precision. When the floating point registers are used, there is simply no difference in speed between single and double precision. Long double precision takes only slightly more time. Single precision division, square root and mathematical functions are calculated faster than double precision when the XMM registers are used, while the speed of addition, subtraction, multiplication, etc. is still the same regardless of precision on most processors (when vector operations are not used).

You may use double precision without worrying too much about the costs if it is good for the application. You may use single precision if you have big arrays and want to get as much data as possible into the data cache. Single precision is good if you can take advantage of vector operations, as explained on page 105.

Floating point addition takes 3 - 6 clock cycles, depending on the microprocessor. Multiplication takes 4 - 8 clock cycles. Division takes 14 - 45 clock cycles. Floating point comparisons are inefficient when the floating point stack registers are used. Conversions of float or double to integer takes a long time when the floating point stack registers are used.

Do not mix single and double precision when the XMM registers are used. See page 146.

Avoid conversions between integers and floating point variables, if possible. See page 146.

Applications that generate floating point underflow in XMM registers can benefit from setting the flush-to-zero mode rather than generating denormal numbers in case of underflow:

```
// Example 7.4. Set flush-to-zero mode (SSE):
#include <xmmintrin.h>
_MM_SET_FLUSH_ZERO_MODE(_MM_FLUSH_ZERO_ON);
```

It is strongly recommended to set the flush-to-zero mode unless you have special reasons to use denormal numbers. You may, in addition, set the denormals-are-zero mode if SSE2 is available:

```
// Example 7.5. Set flush-to-zero and denormals-are-zero mode (SSE2):
#include <xmmintrin.h>
_mm_setcsr(_mm_getcsr() | 0x8040);
```

See page 151 and 121 for more information about mathematical functions.

7.4 Enums

An `enum` is simply an integer in disguise. Enums are exactly as efficient as integers.

7.5 Booleans

The order of Boolean operands

The operands of the Boolean operators `&&` and `||` are evaluated in the following way. If the first operand of `&&` is false, then the second operand is not evaluated at all because the result is known to be false regardless of the value of the second operand. Likewise, if the first operand of `||` is true, then the second operand is not evaluated, because the result is known to be true anyway.

It may be advantageous to put the operand that is most often true last in an `&&` expression, or first in an `||` expression. Assume, for example, that `a` is true 50% of the time and `b` is true 10% of the time. The expression `a && b` needs to evaluate `b` when `a` is true, which is 50% of the cases. The equivalent expression `b && a` needs to evaluate `a` only when `b` is true, which is only 10% of the time. This is faster if `a` and `b` take the same time to evaluate and are equally likely to be predicted by the branch prediction mechanism. See page 44 for an explanation of branch prediction.

If one operand is more predictable than the other, then put the most predictable operand first.

If one operand is faster to calculate than the other then put the operand that is calculated the fastest first.

However, you must be careful when swapping the order of Boolean operands. You cannot swap the operands if the evaluation of the operands has side effects or if the first operand determines whether the second operand is valid. For example:

```
// Example 7.6
unsigned int i;  const int ARRAYSIZE = 100;  float list[ARRAYSIZE];
if (i < ARRAYSIZE && list[i] > 1.0) { ...
```

Here, you cannot swap the order of the operands because the expression `list[i]` is invalid when `i` is not less than `ARRAYSIZE`. Another example:

```
// Example 7.7
if (handle != INVALID_HANDLE_VALUE && WriteFile(handle, ...)) { ...
```

Here you cannot swap the order of the Boolean operands because you should not call `WriteFile` if the handle is invalid.

Boolean variables are overdetermined

Boolean variables are stored as 8-bit integers with the value 0 for false and 1 for true.

Boolean variables are overdetermined in the sense that all operators that have Boolean variables as input check if the inputs have any other value than 0 or 1, but operators that have Booleans as output can produce no other value than 0 or 1. This makes operations with Boolean variables as input less efficient than necessary. Take the example:

```
// Example 7.8a
bool a, b, c, d;
c = a && b;
d = a || b;
```

This is typically implemented by the compiler in the following way:

```
bool a, b, c, d;
if (a != 0) {
    if (b != 0) {
        c = 1;
    }
    else {
        goto CFALSE;
    }
}
else {
    CFALSE:
    c = 0;
}
if (a == 0) {
    if (b == 0) {
        d = 0;
    }
    else {
        goto DTRUE;
    }
}
else {
    DTRUE:
    d = 1;
}
```

This is of course far from optimal. The branches may take a long time in case of mispredictions (see page 44). The Boolean operations can be made much more efficient if it is known with certainty that the operands have no other values than 0 and 1. The reason why the compiler doesn't make such an assumption is that the variables might have other values if they are uninitialized or come from unknown sources. The above code can be optimized if `a` and `b` have been initialized to valid values or if they come from operators that produce Boolean output. The optimized code looks like this:

```
// Example 7.8b
char a = 0, b = 0, c, d;
c = a & b;
```

35

```
    d = a | b;
```

Here, I have used `char` (or `int`) instead of `bool` in order to make it possible to use the bitwise operators (`&` and `|`) instead of the Boolean operators (`&&` and `||`). The bitwise operators are single instructions that take only one clock cycle. The OR operator (`|`) works even if `a` and `b` have other values than 0 or 1. The AND operator (`&`) and the EXCLUSIVE OR operator (`^`) may give inconsistent results if the operands have other values than 0 and 1.

Note that there are a few pitfalls here. You cannot use `~` for NOT. Instead, you can make a Boolean NOT on a variable which is known to be 0 or 1 by XOR'ing it with 1:

```
// Example 7.9a
bool a, b;
b = !a;
```

can be optimized to:

```
// Example 7.9b
char a = 0, b;
b = a ^ 1;
```

You cannot replace `a && b` with `a & b` if `b` is an expression that should not be evaluated if `a` is false. Likewise, you cannot replace `a || b` with `a | b` if `b` is an expression that should not be evaluated if `a` is true.

The trick of using bitwise operators is more advantageous if the operands are variables than if the operands are comparisons, etc. For example:

```
// Example 7.10
bool a; float x, y, z;
a = x > y && z != 0;
```

This is optimal in most cases. Don't change `&&` to `&` unless you expect the `&&` expression to generate many branch mispredictions.

Boolean vector operations

An integer may be used as a Boolean vector. For example, if `a` and `b` are 32-bit integers, then the expression `y = a & b;` will make 32 AND-operations in just one clock cycle. The operators `&`, `|`, `^`, `~` are useful for Boolean vector operations.

7.6 Pointers and references

Pointers versus references

Pointers and references are equally efficient because they are in fact doing the same thing. Example:

```
// Example 7.11
void FuncA (int * p) {
    *p = *p + 2;
}

void FuncB (int & r) {
    r = r + 2;
}
```

These two functions are doing the same thing and if you look at the code generated by the compiler you will notice that the code is exactly identical for the two functions. The difference is simply a matter of programming style. The advantages of using pointers rather than references are:

- When you look at the function bodies above, it is clear that `p` is a pointer, but it is not clear whether `r` is a reference or a simple variable. Using pointers makes it more clear to the reader what is happening.

- It is possible to do things with pointers that are impossible with references. You can change what a pointer points to and you can do arithmetic operations with pointers.

The advantages of using references rather than pointers are:

- The syntax is simpler when using references.

- References are safer to use than pointers because in most cases they are sure to point to a valid address. Pointers can be invalid and cause fatal errors if they are uninitialized, if pointer arithmetic calculations go outside the bounds of valid addresses, or if pointers are type-casted to a wrong type.

- References are useful for copy constructors and overloaded operators.

- Function parameters that are declared as constant references accept expressions as arguments while pointers and non-constant references require a variable.

Efficiency

Accessing a variable or object through a pointer or reference may be just as fast as accessing it directly. The reason for this efficiency lies in the way microprocessors are constructed. All non-static variables and objects declared inside a function are stored on the stack and are in fact addressed relative to the stack pointer. Likewise, all non-static variables and objects declared in a class are accessed through the implicit pointer known in C++ as '`this`'. We can therefore conclude that most variables in a well-structured C++ program are in fact accessed through pointers in one way or another. Therefore, micro-processors have to be designed so as to make pointers efficient, and that's what they are.

However, there are disadvantages of using pointers and references. Most importantly, it requires an extra register to hold the value of the pointer or reference. Registers are a scarce resource, especially in 32-bit mode. If there are not enough registers then the pointer has to be loaded from memory each time it is used and this will make the program slower. Another disadvantage is that the value of the pointer is needed a few clock cycles before the time the variable pointed to can be accessed.

Pointer arithmetic

A pointer is in fact an integer that holds a memory address. Pointer arithmetic operations are therefore as fast as integer arithmetic operations. When an integer is added to a pointer then its value is multiplied by the size of the object pointed to. For example:

```
// Example 7.12
struct abc {int a; int b; int c;};
abc * p; int i;
p = p + i;
```

Here, the value that is added to `p` is not `i` but `i*12`, because the size of `abc` is 12 bytes. The time it takes to add `i` to `p` is therefore equal to the time it takes to make a multiplication and an addition. If the size of `abc` is a power of 2 then the multiplication can be replaced by

a shift operation which is much faster. In the above example, the size of `abc` can be increased to 16 bytes by adding one more integer to the structure.

Incrementing or decrementing a pointer does not require a multiplication but only an addition. Comparing two pointers requires only an integer comparison, which is fast. Calculating the difference between two pointers requires a division, which is slow unless the size of the type of object pointed to is a power of 2 (See page 143 about division).

The object pointed to can be accessed approximately two clock cycles after the value of the pointer has been calculated. Therefore, it is recommended to calculate the value of a pointer well before the pointer is used. For example, `x = *(p++)` is more efficient than `x = *(++p)` because in the latter case the reading of `x` must wait until a few clock cycles after the pointer `p` has been incremented, while in the former case `x` can be read before `p` is incremented. See page 32 for more discussion of the increment and decrement operators.

7.7 Function pointers

Calling a function through a function pointer typically takes a few clock cycles more than calling the function directly if the target address can be predicted. The target address is predicted if the value of the function pointer is the same as last time the statement was executed. If the value of the function pointer has changed then the target address is likely to be mispredicted, which causes a long delay. See page 44 about branch prediction. A Pentium M processor may be able to predict the target if the changes of the function pointer follows a simple regular pattern, while Pentium 4 and AMD processors are sure to make a misprediction every time the function pointer has changed.

7.8 Member pointers

In simple cases, a data member pointer simply stores the offset of a data member relative to the beginning of the object, and a member function pointer is simply the address of the member function. But there are special cases such as multiple inheritance where a much more complicated implementation is needed. These complicated cases should definitely be avoided.

A compiler has to use the most complicated implementation of member pointers if it has incomplete information about the class that the member pointer refers to. For example:

```
// Example 7.13
class c1;
int c1::*MemberPointer;
```

Here, the compiler has no information about the class `c1` other than its name at the time `MemberPointer` is declared. Therefore, it has to assume the worst possible case and make a complicated implementation of the member pointer. This can be avoided by making the full declaration of `c1` before `MemberPointer` is declared. Avoid multiple inheritance, virtual functions, and other complications that make member pointers less efficient.

Most C++ compilers have various options to control the way member pointers are implemented. Use the option that gives the simplest possible implementation if possible, and make sure you are using the same compiler option for all modules that use the same member pointer.

7.9 Smart pointers

A smart pointer is an object that behaves like a pointer. It has the special feature that the object it points to is deleted when the pointer is deleted. Smart pointers are used only for objects stored in dynamically allocated memory, using `new`. The purpose of using smart

pointers is to make sure the object is deleted properly and the memory released when the object is no longer used. A smart pointer may be considered a container that contains only a single element.

The most common implementations of smart pointers are `auto_ptr` and `shared_ptr`. `auto_ptr` has the feature that there is always one, and only one, `auto_ptr` that owns the allocated object, and ownership is transferred from one `auto_ptr` to another by assignment. `shared_ptr` allows multiple pointers to the same object.

There is no extra cost to accessing an object through a smart pointer. Accessing an object by `*p` or `p->member` is equally fast whether `p` is a simple pointer or a smart pointer. But there is an extra cost whenever a smart pointer is created, deleted, copied or transferred from one function to another. These costs are higher for `shared_ptr` than for `auto_ptr`.

Smart pointers can be useful in the situation where the logic structure of a program dictates that an object must be dynamically created by one function and later deleted by another function and these two functions are unrelated to each other (not member of the same class). If the same function or class is responsible for creating and deleting the object then you don't need a smart pointer.

If a program uses many small dynamically allocated objects with each their smart pointer then you may consider if the cost of this solution is too high. It may be more efficient to pool all the objects together into a single container, preferably with contiguous memory. See the discussion of container classes on page 92.

7.10 Arrays

An array is implemented simply by storing the elements consecutively in memory. No information about the dimensions of the array is stored. This makes the use of arrays in C and C++ faster than in other programming languages, but also less safe. This safety problem can be overcome by defining a container class that behaves like an array with bounds checking, as illustrated in this example:

```
// Example 7.14a. Array with bounds checking
template <typename T, unsigned int N> class SafeArray {
protected:
    T a[N];                          // Array with N elements of type T
public:
    SafeArray() {                    // Constructor
        memset(a, 0, sizeof(a));     // Initialize to zero
    }
    int Size() {                     // Return the size of the array
        return N;
    }
    T & operator[] (unsigned int i) { // Safe [] array index operator
        if (i >= N) {
            // Index out of range. The next line provokes an error.
            // You may insert any other error reporting here:
            return *(T*)0;   // Return a null reference to provoke error
        }
        // No error
        return a[i];      // Return reference to a[i]
    }
};
```

More examples of container classes are given in www.agner.org/optimize/cppexamples.zip.

An array using the above template class is declared by specifying the type and size as template parameters, as example 7.14b below shows. It is accessed with a square brackets

index, just as a normal array. The constructor sets all elements to zero. You may remove the `memset` line if you don't want this initialization, or if the type `T` is a class with a default constructor that does the necessary initialization. The compiler may report that `memset` is deprecated. This is because it can cause errors if the size parameter is wrong, but it is still the fastest way to set an array to zero. The `[]` operator will detect an error if the index is out of range (see page 140 on bounds checking). An error message is provoked here in a rather unconventional manner by returning a null reference. This will provoke an error message in a protected operating system if the array element is accessed, and this error is easy to trace with a debugger. You may replace this line by any other form of error reporting. For example, in Windows, you may write `FatalAppExitA(0,"Array index out of range");` or better, make your own error message function.

The following example illustrates how to use `SafeArray`:

```
// Example 7.14b
SafeArray <float, 100> list;          // Make array of 100 floats
for (int i = 0; i < list.Size(); i++) { // Loop through array
   cout << list[i] << endl;           // Output array element
}
```

An array initialized by a list should preferably be static, as explained on page 28. An array can be initialized to zero by using `memset`:

```
// Example 7.15
float list[100];
memset(list, 0, sizeof(list));
```

A multidimensional array should be organized so that the last index changes fastest:

```
// Example 7.16
const int rows = 20, columns = 50;
float matrix[rows][columns];
int i, j;  float x;
   for (i = 0; i < rows; i++)
      for (j = 0; j < columns; j++)
         matrix[i][j] += x;
```

This makes sure that the elements are accessed sequentially. The opposite order of the two loops would make the access non-sequential which makes the data caching less efficient.

The size of all but the first dimension may preferably be a power of 2 if the rows are indexed in a non-sequential order in order to make the address calculation more efficient:

```
// Example 7.17
int FuncRow(int);  int FuncCol(int);
const int rows = 20, columns = 32;
float matrix[rows][columns];
int i;  float x;
for (i = 0; i < 100; i++)
   matrix[FuncRow(i)][FuncCol(i)] += x;
```

Here, the code must compute `(FuncRow(i)*columns + FuncCol(i)) * sizeof(float)` in order to find the address of the matrix element. The multiplication by `columns` in this case is faster when columns is a power of two. In the preceding example, this is not an issue because an optimizing compiler can see that the rows are accessed consecutively and can calculate the address of each row by adding the length of a row to the address of the preceding row.

The same advice applies to arrays of structure or class objects. The size (in bytes) of the objects should preferably be a power of 2 if the elements are accessed in a non-sequential order.

The advice of making the number of columns a power of 2 does not always apply to arrays that are bigger than the level-1 data cache and accessed non-sequentially because it may cause cache contentions. See page 87 for a discussion of this problem.

7.11 Type conversions
The C++ syntax has several different ways of doing type conversions:

```
// Example 7.18
int i;  float f;
f = i;                     // Implicit type conversion
f = (float)i;              // C-style type casting
f = float(i);              // Constructor-style type casting
f = static_cast<float>(i); // C++ casting operator
```

These different methods have exactly the same effect. Which method you use is a matter of programming style. The time consumption of different type conversions is discussed below.

Signed / unsigned conversion

```
// Example 7.19
int i;
if ((unsigned int)i < 10) { ...
```

Conversions between signed and unsigned integers simply makes the compiler interpret the bits of the integer in a different way. There is no checking for overflow, and the code takes no extra time. These conversions can be used freely without any cost in performance.

Integer size conversion

```
// Example 7.20
int i;  short int s;
i = s;
```

An integer is converted to a longer size by extending the sign-bit if the integer is signed, or by extending with zero-bits if unsigned. This typically takes one clock cycle if the source is an arithmetic expression. The size conversion often takes no extra time if it is done in connection with reading the value from a variable in memory, as in example 7.21.

```
// Example 7.21
short int a[100];  int i, sum = 0;
for (i=0; i<100; i++) sum += a[i];
```

Converting an integer to a smaller size is done simply by ignoring the higher bits. There is no check for overflow. Example:

```
// Example 7.22
int i;  short int s;
s = (short int)i;
```

This conversion takes no extra time. It simply stores the lower 16 bits of the 32-bit integer.

Floating point precision conversion

Conversions between `float`, `double` and `long double` take no extra time when the floating point register stack is used. It takes between 2 and 15 clock cycles (depending on the processor) when the XMM registers are used. See page 32 for an explanation of register stack versus XMM registers. Example:

```
// Example 7.23
float a; double b;
a += b;
```

In this example, the conversion is costly if XMM registers are used. `a` and `b` should be of the same type to avoid this. See page 146 for further discussion.

Integer to float conversion

Conversion of a signed integer to a `float` or `double` takes 4 - 16 clock cycles, depending on the processor and the type of registers used. Conversion of an unsigned integer takes longer time. It is faster to first convert the unsigned integer to a signed integer if there is no risk of overflow:

```
// Example 7.24
unsigned int u;  double d;
d = (double)(signed int)u;  // Faster, but risk of overflow
```

Integer to float conversions can sometimes be avoided by replacing an integer variable by a float variable. Example:

```
// Example 7.25a
float a[100];  int i;
for (i = 0;  i < 100;  i++) a[i] = 2 * i;
```

The conversion of `i` to float in this example can be avoided by making an additional floating point variable:

```
// Example 7.25b
float a[100];  int i;  float i2;
for (i = 0, i2 = 0;  i < 100;  i++, i2 += 2.0f) a[i] = i2;
```

Float to integer conversion

Conversion of a floating point number to an integer takes a very long time unless the SSE2 or later instruction set is enabled. Typically, the conversion takes 50 - 100 clock cycles. The reason is that the C/C++ standard specifies truncation so the floating point rounding mode has to be changed to truncation and back again.

If there are floating point-to-integer conversions in the critical part of a code then it is important to do something about it. Possible solutions are:

Avoid the conversions by using different types of variables.

Move the conversions out of the innermost loop by storing intermediate results as floating point.

Use 64-bit mode or enable the SSE2 instruction set (requires a microprocessor that supports this).

Use rounding instead of truncation and make a round function using assembly language. See page 147 for details about rounding.

Pointer type conversion

A pointer can be converted to a pointer of a different type. Likewise, a pointer can be converted to an integer, or an integer can be converted to a pointer. It is important that the integer has enough bits for holding the pointer.

These conversions do not produce any extra code. It is simply a matter of interpreting the same bits in a different way or bypassing syntax checks.

These conversions are not safe, of course. It is the responsibility of the programmer to make sure the result is valid.

Re-interpreting the type of an object

It is possible to make the compiler treat a variable or object as if it had a different type by type-casting its address:

```
// Example 7.26
float x;
*(int*)&x |= 0x80000000;   // Set sign bit of x
```

The syntax may seem a little odd here. The address of `x` is type-casted to a pointer to an integer, and this pointer is then de-referenced in order to access `x` as an integer. The compiler does not produce any extra code for actually making a pointer. The pointer is simply optimized away and the result is that `x` is treated as an integer. But the `&` operator forces the compiler to store `x` in memory rather than in a register. The above example sets the sign bit of `x` by using the `|` operator which otherwise can only be applied to integers. It is faster than `x = -abs(x);`.

There are a number of dangers to be aware of when type-casting pointers:

- The trick violates the strict aliasing rule of standard C, specifying that two pointers of different types cannot point to the same object (except for char pointers). An optimizing compiler might store the floating point and integer representations in two different registers. You need to check if the compiler does what you want it to. It is safer to use a union, as in example 14.23 page 149.
- The trick will fail if the object is treated as bigger than it actually is. This above code will fail if an `int` uses more bits than a `float`. (Both use 32 bits in x86 systems).
- If you access part of a variable, for example 32 bits of a 64-bit double, then the code will not be portable to platforms that use big endian storage.
- If you access a variable in parts, for example if you write a 64-bit double 32 bits at a time, then the code is likely to execute slower than intended because of a store forwarding delay in the CPU (See manual 3: "The microarchitecture of Intel, AMD and VIA CPUs").

Const cast

The `const_cast` operator is used for relieving the `const` restriction from a pointer. It has some syntax checking and is therefore more safe than the C-style type-casting without adding any extra code. Example:

```
// Example 7.27
class c1 {
   const int x;        // constant data
   public:
   c1() : x(0) {};     // constructor initializes x to 0
   void xplus2() {     // this function can modify x
      *const_cast<int*>(&x) += 2;}  // add 2 to x
};
```

43

The effect of the `const_cast` operator here is to remove the `const` restriction on `x`. It is a way of relieving a syntax restriction, but it doesn't generate any extra code and doesn't take any extra time. This is a useful way of making sure that one function can modify `x`, while other functions can not.

Static cast

The `static_cast` operator does the same as the C-style type-casting. It is used, for example, to convert `float` to `int`.

Reinterpret cast

The `reinterpret_cast` operator is used for pointer conversions. It does the same as C-style type-casting with a little more syntax check. It does not produce any extra code.

Dynamic cast

The `dynamic_cast` operator is used for converting a pointer to one class to a pointer to another class. It makes a runtime check that the conversion is valid. For example, when a pointer to a base class is converted to a pointer to a derived class, it checks whether the original pointer actually points to an object of the derived class. This check makes `dynamic_cast` more time-consuming than a simple type casting, but also safer. It may catch programming errors that would otherwise go undetected.

Converting class objects

Conversions involving class objects (rather than pointers to objects) are possible only if the programmer has defined a constructor, an overloaded assignment operator, or an over-loaded type casting operator that specifies how to do the conversion. The constructor or overloaded operator is as efficient as a member function.

7.12 Branches and switch statements

The high speed of modern microprocessors is obtained by using a pipeline where instructions are fetched and decoded in several stages before they are executed. However, the pipeline structure has one big problem. Whenever the code has a branch (e.g. an if-else structure), the microprocessor doesn't know in advance which of the two branches to feed into the pipeline. If the wrong branch is fed into the pipeline then the error is not detected until 10 - 20 clock cycles later and the work it has done by fetching, decoding and perhaps speculatively executing instructions during this time has been wasted. The consequence is that the microprocessor wastes several clock cycles whenever it feeds a branch into the pipeline and later discovers that it has chosen the wrong branch.

Microprocessor designers have gone to great lengths to reduce this problem. The most important method that is used is branch prediction. Modern microprocessors are using advanced algorithms to predict which way a branch will go based on the past history of that branch and other nearby branches. The algorithms used for branch prediction are different for each type of microprocessor. These algorithms are described in detail in manual 3: "The microarchitecture of Intel, AMD and VIA CPUs".

A branch instruction takes typically 0 - 2 clock cycles in the case that the microprocessor has made the right prediction. The time it takes to recover from a branch misprediction is approximately 12 - 25 clock cycles, depending on the processor. This is called the branch misprediction penalty.

Branches are relatively cheap if they are predicted most of the time, but expensive if they are often mispredicted. A branch that always goes the same way is predicted well, of course. A branch that goes one way most of the time and rarely the other way is mispredicted only when it goes the other way. A branch that goes many times one way, then many times the other way is mispredicted only when it changes. A branch that follows

44

a simple periodic pattern can also be predicted quite well if it is inside a loop with few or no other branches. A simple periodic pattern can be, for example, to go one way two times and the other way three times. Then again two times the first way and three times the other way, etc. The worst case is a branch that goes randomly one way or the other with a 50-50 chance of going either way. Such a branch will be mispredicted 50% of the time.

A for-loop or while-loop is also a kind of branch. After each iteration it decides whether to repeat or to exit the loop. The loop-branch is usually predicted well if the repeat count is small and always the same. The maximum loop count that can be predicted perfectly varies between 9 and 64, depending on the processor. Nested loops are predicted well only on some processors. On many processors, a loop that contains several branches is not predicted well.

A switch statements is a kind of branch that can go more than two ways. Switch statements are most efficient if the case labels follow a sequence where each label is equal to the preceding label plus one, because it can be implemented as a table of jump targets. A switch statement with many labels that have values far from each other is inefficient because the compiler must convert it to a branch tree.

On older processors, a switch statement with sequential labels is simply predicted to go the same way as last time it was executed. It is therefore certain to be mispredicted whenever it goes another way than last time. Newer processors are sometimes able to predict a switch statement if it follows a simple periodic pattern or if it is correlated with preceding branches and the number of different targets is small.

The number of branches and switch statements should preferably be kept small in the critical part of a program, especially if the branches are poorly predictable. It may be useful to roll out a loop if this can eliminate branches, as explained in the next paragraph.

The target of branches and function calls are saved in a special cache called the branch target buffer. Contentions in the branch target buffer can occur if a program has many branches or function calls. The consequence of such contentions is that branches can be mispredicted even if they otherwise would be predicted well. Even function calls can be mispredicted for this reason. A program with many branches and function calls in the critical part of the code can therefore suffer from mispredictions.

In some cases it is possible to replace a poorly predictable branch by a table lookup. For example:

```
// Example 7.28a
float a;  bool b;
a = b ? 1.5f : 2.6f;
```

The `?:` operator here is a branch. If it is poorly predictable then replace it by a table lookup:

```
// Example 7.28b
float a;  bool b = 0;
static const float lookup[2] = {2.6f, 1.5f};
a = lookup[b];
```

If a `bool` is used as an array index then it is important to make sure it is initialized or comes from a reliable source so that it can have no other values than 0 or 1. See page 35.

The examples on page 140 and 141 show various ways of reducing the number of branches.

Manual 3: "The microarchitecture of Intel, AMD and VIA CPUs" gives more details on branch predictions in the different microprocessors.

7.13 Loops

The efficiency of a loop depends on how well the microprocessor can predict the loop control branch. See the preceding paragraph and manual 3: "The microarchitecture of Intel, AMD and VIA CPUs" for an explanation of branch prediction. A loop with a small and fixed repeat count and no branches inside can be predicted perfectly. As explained above, the maximum loop count that can be predicted depends on the processor. Nested loops are predicted well only on some processors that have a special loop predictor. On other processors, only the innermost loop is predicted well. A loop with a high repeat count is mispredicted only when it exits. For example, if a loop repeats a thousand times then the loop control branch is mispredicted only one time in thousand so the misprediction penalty is only a negligible contribution to the total execution time.

Loop unrolling

In some cases it can be an advantage to unroll a loop. Example:

```
// Example 7.29a
int i;
for (i = 0; i < 20; i++) {
    if (i % 2 == 0) {
        FuncA(i);
    }
    else {
        FuncB(i);
    }
    FuncC(i);
}
```

This loop repeats 20 times and calls alternately `FuncA` and `FuncB`, then `FuncC`. Unrolling the loop by two gives:

```
// Example 7.29b
int i;
for (i = 0; i < 20; i += 2) {
    FuncA(i);
    FuncC(i);
    FuncB(i+1);
    FuncC(i+1);
}
```

This has three advantages:

- The `i<20` loop control branch is executed 10 times rather than 20.

- The fact that the repeat count has been reduced from 20 to 10 means that it can be predicted perfectly on a Pentium 4.

- The `if` branch is eliminated.

Loop unrolling also has disadvantages:

- The unrolled loop takes up more space in the code cache or micro-op cache.

- The Core2 processor performs better on very small loops (less than 65 bytes of code).

- If the repeat count is odd and you unroll by two then there is an extra iteration that has to be done outside the loop. In general, you have this problem when the repeat count is not certain to be divisible by the unroll factor.

Loop unrolling should only be used if there are specific advantages that can be obtained. If a loop contains floating point calculations and the loop counter is an integer, then you can generally assume that the overall computation time is determined by the floating point code rather than by the loop control branch. There is nothing to gain by unrolling the loop in this case.

Loop unrolling should preferably be avoided on processors with a micro-op cache (e.g. Sandy Bridge) because it is important to economize the use of the micro-op cache.

Compilers will usually unroll a loop automatically if this appears to be profitable (see page 70). The programmer does not have to unroll a loop manually unless there is a specific advantage to obtain, such as eliminating the `if`-branch in example 7.29b.

The loop control condition

The most efficient loop control condition is a simple integer counter. A microprocessor with out-of-order capabilities (see page 103) will be able to evaluate the loop control statement several iterations ahead.

It is less efficient if the loop control branch depends on the calculations inside the loop. The following example converts a zero-terminated ASCII string to lower case:

```
// Example 7.30a
char string[100], *p = string;
while (*p != 0) *(p++) |= 0x20;
```

If the length of the string is already known then it is more efficient to use a loop counter:

```
// Example 7.30b
char string[100], *p = string;  int i, StringLength;
for (i = StringLength; i > 0; i--) *(p++) |= 0x20;
```

A common situation where the loop control branch depends on calculations inside the loop is in mathematical iterations such as Taylor expansions and Newton-Raphson iterations. Here the iteration is repeated until the residual error is lower than a certain tolerance. The time it takes to calculate the absolute value of the residual error and compare it to the tolerance may be so high that it is more efficient to determine the worst-case maximum repeat count and always use this number of iterations. The advantage of this method is that the microprocessor can execute the loop control branch ahead of time and resolve any branch misprediction long before the floating point calculations inside the loop are finished. This method is advantageous if the typical repeat count is near the maximum repeat count and the calculation of the residual error for each iteration is a significant contribution to the total calculation time.

A loop counter should preferably be an integer. If a loop needs a floating point counter then make an additional integer counter. Example:

```
// Example 7.31a
double x, n, factorial = 1.0;
for (x = 2.0; x <= n; x++) factorial *= x;
```

This can be improved by adding an integer counter and using the integer in the loop control condition:

```
// Example 7.31b
double x, n, factorial = 1.0;  int i;
for (i = (int)n - 2, x = 2.0; i >= 0; i--, x++) factorial *= x;
```

Note the difference between commas and semicolons in a loop with multiple counters, as in example 7.31b. A `for`-loop has three clauses: initialization, condition, and increment. The three clauses are separated by semicolons, while multiple statements within each clause are separated by commas. There should be only one statement in the condition clause.

Comparing an integer to zero is sometimes more efficient than comparing it to any other number. Therefore, it is slightly more efficient to make a loop count down to zero than making it count up to some positive value, `n`. But not if the loop counter is used as an array index. The data cache is optimized for accessing arrays forwards, not backwards.

Copying or clearing arrays

It may not be optimal to use a loop for trivial tasks such as copying an array or setting an array to all zeroes. Example:

```
// Example 7.32a
const int size = 1000;  int i;
float a[size], b[size];
// set a to zero
for (i = 0; i < size; i++) a[i] = 0.0;
// copy a to b
for (i = 0; i < size; i++) b[i] = a[i];
```

It is often faster to use the functions `memset` and `memcpy`:

```
// Example 7.32b
const int size = 1000;
float a[size], b[size];
// set a to zero
memset(a, 0, sizeof(a));
// copy a to b
memcpy(b, a, sizeof(b));
```

Most compilers will automatically replace such loops by calls to `memset` and `memcpy`, at least in simple cases. The explicit use of `memset` and `memcpy` is unsafe because serious errors can happen if the size parameter is bigger than the destination array. But the same errors can happen with the loops if the loop count is too big.

7.14 Functions

Function calls may slow down a program for the following reasons:

- The function call makes the microprocessor jump to a different code address and back again. This may take up to 4 clock cycles. In most cases the microprocessor is able to overlap the call and return operations with other calculations to save time.

- The code cache works less efficiently if the code is fragmented and scattered around in memory.

- Function parameters are stored on the stack in 32-bit mode. Storing the parameters on the stack and reading them again takes extra time. The delay is significant if a parameter is part of a critical dependency chain, especially on the Pentium 4 processor.

- Extra time is needed for setting up a stack frame, saving and restoring registers, and possibly save exception handling information.

- Each function call statement occupies a space in the branch target buffer (BTB). Contentions in the BTB can cause branch mispredictions if the critical part of a program has many calls and branches.

The following methods may be used for reducing the time spent on function calls in the critical part of a program.

Avoid unnecessary functions

Some programming textbooks recommend that every function that is longer than a few lines should be split up into multiple functions. I disagree with this rule. Splitting up a function into multiple smaller functions only makes the program less efficient. Splitting up a function just because it is long does not make the program more clear unless the function is doing multiple logically distinct tasks. A critical innermost loop should preferably be kept entirely inside one function, if possible.

Use inline functions

An inline function is expanded like a macro so that each statement that calls the function is replaced by the function body. A function is inlined if the `inline` keyword is used or if its body is defined inside a class definition. Inlining a function is advantageous if the function is small or if it is called only from one place in the program. Small functions are often inlined automatically by the compiler. On the other hand, the compiler may in some cases ignore a request for inlining a function if the inlining causes technical problems.

Avoid nested function calls in the innermost loop

A function that calls other functions is called a *frame function*, while a function that doesn't call any other function is called a *leaf function*. Leaf functions are more efficient than frame functions for reasons explained on page 62. If the critical innermost loop of a program contains calls to frame functions then the code can probably be improved by inlining the frame function or by turning the frame function into a leaf function by inlining all the functions that it calls.

Use macros instead of functions

A macro declared with `#define` is certain to be inlined. But beware that macro parameters are evaluated every time they are used. Example:

```
// Example 7.33a. Use macro as inline function
#define max(a,b) (a > b ? a : b)
y = max(sin(x), cos(x));
```

In this example, `sin(x)` or `cos(x)` is calculated twice because the macro is referencing it twice. You can avoid this by using an inline function instead of a macro. If you want the function to work with any type of parameters then make it a template:

```
// Example 7.33b. Replace macro by template
template <typename T>
static inline T max(T const & a, T const & b) {
   return a > b ? a : b;
}
```

Use fastcall functions

The keyword `__fastcall` changes the function calling method in 32-bit mode so that the first two (three on CodeGear compiler) integer parameters are transferred in registers rather than on the stack. This can improve the speed of functions with integer parameters.

Floating point parameters are not affected by `__fastcall`. The implicit '`this`' pointer in member functions is also treated like a parameter, so there may be only one free register left for transferring additional parameters. Therefore, make sure that the most critical integer parameter comes first when you are using `__fastcall`. Function parameters are

transferred in registers by default in 64-bit mode. Therefore, the `__fastcall` keyword is not recognized in 64-bit mode.

Make functions static

It is recommended to add the keyword `static` to all non-member functions that are not needed outside the module in which they are defined. The `static` declaration restricts the scope of the function to the current module (i.e. the current `.cpp` file). This enables the compiler to optimize across function calls. You cannot use this optimization for class member functions, because the keyword `static` has a different meaning for member functions.

Use whole program optimization

Some compilers have an option for whole program optimization or for combining multiple `.cpp` files into a single object file. This enables the compiler to optimize register allocation and parameter transfer across all `.cpp` modules that make up a program. Whole program optimization cannot be used for function libraries distributed as object or library files.

Use 64-bit mode

Parameter transfer is more efficient in 64-bit mode than in 32-bit mode, and more efficient in 64-bit Linux than in 64-bit Windows. In 64-bit Linux, the first six integer parameters and the first eight floating point parameters are transferred in registers, totaling up to fourteen register parameters. In 64-bit Windows, the first four parameters are transferred in registers, regardless of whether they are integers or floating point numbers. Therefore, 64-bit Linux is more efficient than 64-bit Windows if functions have more than four parameters. There is no difference between 32-bit Linux and 32-bit Windows in this respect.

7.15 Function parameters

Function parameters are transferred by value in most cases. This means that the value of the parameter is copied to a local variable. This is efficient for simple types such as `int`, `float`, `double`, `bool`, `enum` as well as pointers and references.

Arrays are always transferred as pointers unless they are wrapped into a class or structure.

The situation is more complex if the parameter has a composite type such as a structure or class. The transfer of a parameter of composite type is most efficient if all of the following conditions are met:

- the object is so small that it fits into a single register
- the object has no copy constructor and no destructor
- the object has no virtual member
- the object does not use runtime type identification (RTTI)

If any of these conditions is not met then it is usually faster to transfer a pointer or reference to the object. If the object is large then it obviously takes time to copy the entire object. Any copy constructor must be called when the object is copied to the parameter, and the destructor, if any, must be called before the function returns.

The preferred method for transferring composite objects to a function is by a const reference. A const reference makes sure that the original object is not modified. Unlike a pointer or a non-const reference, a const reference allows the function argument to be an expression or an anonymous object. The compiler can easily optimize away a const reference if the function is inlined.

An alternative solution is to make the function a member of the object's class or structure. This is equally efficient.

Simple function parameters are transferred on the stack in 32-bit systems, but in registers in 64-bit systems. The latter is more efficient. 64-bit Windows allows a maximum of four parameters to be transferred in registers. 64-bit Unix systems allow up to fourteen parameters to be transferred in registers (8 float or double plus 6 integer, pointer or reference parameters). The `this` pointer in member functions counts a one parameter. Further details are given in manual 5: "Calling conventions for different C++ compilers and operating systems".

7.16 Function return types

The return type of a function should preferably be a simple type, a pointer, a reference, or void. Returning objects of a composite type is more complex and often inefficient.

Objects of a composite type can be returned in registers only in the simplest cases. See manual 5: "Calling conventions for different C++ compilers and operating systems" for details on when objects can be returned in registers.

Except for the simplest cases, composite objects are returned by copying them into a place indicated by the caller through a hidden pointer. The copy constructor, if any, is usually called in the copying process, and the destructor is called when the original is destroyed. In simple cases, the compiler may be able to avoid the calls to the copy constructor and the destructor by constructing the object on its final destination, but don't count on it.

Instead of returning a composite object, you may consider the following alternatives:

- Make the function a constructor for the object.

- Make the function modify an existing object rather than making a new one. The existing object can be made available to the function through a pointer or reference, or the function could be a member of the object's class.

- Make the function return a pointer or reference to a static object defined inside the function. This is efficient, but risky. The returned pointer or reference is valid only until the next time the function is called and the local object is overwritten, possibly in a different thread. If you forget to make the local object static then it becomes invalid as soon as the function returns.

- Make the function construct an object with `new` and return a pointer to it. This is inefficient because of the costs of dynamic memory allocation. This method also involves the risk of memory leaks if you forget to delete the object.

7.17 Structures and classes

Nowadays, programming textbooks recommend object oriented programming as a means of making software more clear and modular. The so-called objects are instances of structures and classes. The object oriented programming style has both positive and negative impacts on program performance. The positive effects are:

- Variables that are used together are also stored together if they are members of the same structure or class. This makes data caching more efficient.

- Variables that are members of a class need not be passed as parameters to a class member function. The overhead of parameter transfer is avoided for these variables.

The negative effects of object oriented programming are:

51

- Non-static member functions have a 'this' pointer which is transferred as an implicit parameter to the function. The overhead of parameter transfer for 'this' is incurred on all non-static member functions.

- The 'this' pointer takes up one register. Registers are a scarce resource in 32-bit systems.

- Virtual member functions are less efficient (see page 54).

No general statement can be made about whether the positive or the negative effects of object oriented programming are dominating. At least, it can be said that the use of classes and member functions is not expensive. You may use an object oriented programming style if it is good for the logical structure and clarity of the program as long as you avoid an excessive number of function calls in the most critical part of the program. The use of structures (without member functions) has no negative effect on performance.

7.18 Class data members (properties)

The data members of a class or structure are stored consecutively in the order in which they are declared whenever an instance of the class or structure is created. There is no performance penalty for organizing data into classes or structures. Accessing a data member of a class or structure object takes no more time than accessing a simple variable.

Most compilers will align data members to round addresses in order to optimize access, as given in the following table.

Type	size, bytes	alignment, bytes
bool	1	1
char, signed or unsigned	1	1
short int, signed or unsigned	2	2
int, signed or unsigned	4	4
64-bit integer, signed or unsigned	8	8
pointer or reference, 32-bit mode	4	4
pointer or reference, 64-bit mode	8	8
float	4	4
double	8	8
long double	8, 10, 12 or 16	8 or 16

Table 7.2. Alignment of data members.

This alignment can cause holes of unused bytes in a structure or class with members of mixed sizes. For example:

```
// Example 7.34a
struct S1 {
   short int a;  // 2 bytes. first byte at  0, last byte at  1
                 // 6 unused bytes
   double b;     // 8 bytes. first byte at  8, last byte at 15
   int d;        // 4 bytes. first byte at 16, last byte at 19
                 // 4 unused bytes
};
S1 ArrayOfStructures[100];
```

Here, there are 6 unused bytes between a and b because b has to start at an address divisible by 8. There are also 4 unused bytes in the end. The reason for this is that the next instance of S1 in the array must begin at an address divisible by 8 in order to align its b member by 8. The number of unused bytes can be reduced to 2 by putting the smallest members last:

```
// Example 7.34b
struct S1 {
   double b;      // 8 bytes. first byte at  0, last byte at  7
   int d;         // 4 bytes. first byte at  8, last byte at 11
   short int a;   // 2 bytes. first byte at 12, last byte at 13
                  // 2 unused bytes
};
S1 ArrayOfStructures[100];
```

This reordering has made the structure 8 bytes smaller and the array 800 bytes smaller.

Structure and class objects can often be made smaller by reordering the data members. If the class has at least one virtual member functions then there is a pointer to a virtual table before the first data member or after the last member. This pointer is 4 bytes in 32-bit systems and 8 bytes in 64-bit systems. If you are in doubt how big a structure or each of its members are then you may make some tests with the `sizeof` operator.

The code for accessing a data member is more compact if the offset of the member relative to the beginning of the structure or class is less than 128. Example:

```
// Example 7.35
class S2 {
   public:
   int a[100];  // 400 bytes. first byte at   0, last byte at 399
   int b;       // 4 bytes.   first byte at 400, last byte at 403
   int ReadB() {return b;}
};
```

The offset of b is 400 here. Any code that accesses b through a pointer or a member function such as ReadB needs to add a 4-bytes offset to the pointer. If a and b are swapped then both can be accessed with a 1-byte signed integer as offset. This makes the code more compact so that the code cache is used more efficiently. It is therefore recommended that big arrays and other big objects come last in a structure or class declaration. If it is not possible to contain all data members within the first 128 bytes then put the most used members in the first 128 bytes.

7.19 Class member functions (methods)
Each time a new object of a class is declared or created it will generate a new instance of the data members. But each member function has only one instance. The function code is not copied because the same code can be applied to all instances of the class.

Calling a member function is as fast as calling a simple function with a pointer or reference to a structure. For example:

```
// Example 7.36
class S3 {
   public:
   int a;
   int b;
   int Sum1() {return a + b;}
};
int Sum2(S3 * p) {return p->a + p->b;}
int Sum3(S3 & r) {return  r.a +  r.b;}
```

The three functions Sum1, Sum2 and Sum3 are doing exactly the same thing and they are equally efficient. If you look at the code generated by the compiler, you will notice that some compilers will make exactly identical code for the three functions. Sum1 has an implicit 'this' pointer which does the same thing as p and r in Sum2 and Sum3. Whether you want

to make the function a member of the class or give it a pointer or reference to the class or structure is simply a matter of programming style. Some compilers make `Sum1` slightly more efficient than `Sum2` and `Sum3` in 32-bit Windows by transferring '`this`' in a register rather than on the stack.

A `static` member function cannot access any non-static data members or non-static member functions. A static member function is faster than a non-static member function because it doesn't need the '`this`' pointer. You may make member functions faster by making them static if they don't need any non-static access.

7.20 Virtual member functions

Virtual functions are used for implementing polymorphic classes. Each instance of a polymorphic class has a pointer to a table of pointers to the different versions of the virtual functions. This so-called virtual table is used for finding the right version of the virtual function at runtime. Polymorphism is one of the main reasons why object oriented programs can be less efficient than non-object oriented programs. If you can avoid virtual functions then you can obtain most of the advantages of object oriented programming without paying the performance costs.

The time it takes to call a virtual member function is a few clock cycles more than it takes to call a non-virtual member function, provided that the function call statement always calls the same version of the virtual function. If the version changes then you may get a misprediction penalty of 10 - 20 clock cycles. The rules for prediction and misprediction of virtual function calls is the same as for switch statements, as explained on page 45.

The dispatching mechanism can be bypassed when the virtual function is called on an object of known type, but you cannot always rely on the compiler bypassing the dispatch mechanism even when it would be obvious to do so. See page 73.

Runtime polymorphism is needed only if it cannot be known at compile time which version of a polymorphic member function is called. If virtual functions are used in a critical part of a program then you may consider whether it is possible to obtain the desired functionality without polymorphism or with compile-time polymorphism.

It is sometimes possible to obtain the desired polymorphism effect with templates instead of virtual functions. The template parameter should be a class containing the functions that have multiple versions. This method is faster because the template parameter is always resolved at compile time rather than at runtime. Example 7.42 on page 58 shows an example of how to do this. Unfortunately, the syntax is so kludgy that it may not be worth the effort.

7.21 Runtime type identification (RTTI)

Runtime type identification adds extra information to all class objects and is not efficient. If the compiler has an option for RTTI then turn it off and use alternative implementations.

7.22 Inheritance

An object of a derived class is implemented in the same way as an object of a simple class containing the members of both parent and child class. Members of parent and child class are accessed equally fast. In general, you can assume that there is hardly any performance penalty to using inheritance.

There may be a slight degradation in code caching for the following reasons:

- The size of the parent class data members is added to the offset of the child class members. The code that accesses data members with a total offset bigger than 127 bytes is slightly less compact. See page 53.

- The member functions of parent and child are typically stored in different modules. This may cause a lot of jumping around and less efficient code caching. This problem can be solved by making sure that functions which are called near each other are also stored near each other. See page 88 for details.

Inheritance from multiple parent classes in the same generation can cause complications with member pointers and virtual functions or when accessing an object of a derived class through a pointer to one of the base classes. You may avoid multiple inheritance by making objects inside the derived class:

```
// Example 7.37a. Multiple inheritance
class B1; class B2;
class D : public B1, public B2 {
public:
    int c;
};
```

Replace with:

```
// Example 7.37b. Alternative to multiple inheritance
class B1; class B2;
class D : public B1 {
public:
    B2 b2;
    int c;
};
```

7.23 Constructors and destructors

A constructor is implemented internally as a member function which returns a reference to the object. The allocation of memory for a new object is not necessarily done by the constructor itself. Constructors are therefore as efficient as any other member functions. This applies to default constructors, copy constructors, and any other constructors.

A class doesn't need a constructor. A default constructor is not needed if the object doesn't need initialization. A copy constructor is not needed if the object can be copied simply by copying all data members. A simple constructor may be inlined for improved performance.

A copy constructor may be called whenever an object is copied by assignment, as a function parameter, or as a function return value. The copy constructor can be a time consumer if it involves allocation of memory or other resources. There are various ways to avoid this wasteful copying of memory blocks, for example:
- Use a reference or pointer to the object instead of copying it
- Use a "move constructor" to transfer ownership of the memory block. This requires a compiler with C++0x support.
- Make a member function or friend function or operator that transfers ownership of the memory block from one object to another. The object that looses ownership of the memory block should have its pointer set to NULL. There should of course be a destructor that destroys any memory block that the object owns.

A destructor is as efficient as a member function. Do not make a destructor if it is not necessary. A virtual destructor is as efficient as a virtual member function. See page 54.

7.24 Unions

A union is a structure where data members share the same memory space. A union can be used for saving memory space by allowing two data members that are never used at the same time to share the same piece of memory. See page 89 for an example.

A union can also be used for accessing the same data in different ways. Example:

```
// Example 7.38
union {
    float f;
    int i;
} x;
x.f = 2.0f;
x.i |= 0x80000000;  // set sign bit of f
cout << x.f;        // will give -2.0
```

In this example, the sign bit of `f` is set by using the bitwise OR operator, which can only be applied to integers.

7.25 Bitfields

Bitfields may be useful for making data more compact. Accessing a member of a bitfield is less efficient than accessing a member of a structure. The extra time may be justified in case of large arrays if it can save cache space or make files smaller.

It is faster to compose a bitfield by the use of `<<` and `|` operations than to write the members individually. Example:

```
// Example 7.39a
struct Bitfield {
    int a:4;
    int b:2;
    int c:2;
};
Bitfield x;
int A, B, C;
x.a = A;
x.b = B;
x.c = C;
```

Assuming that the values of `A`, `B` and `C` are too small to cause overflow, this code can be improved in the following way:

```
// Example 7.39b
union Bitfield {
    struct {
        int a:4;
        int b:2;
        int c:2;
    };
    char abc;
};
Bitfield x;
int A, B, C;
x.abc = A | (B << 4) | (C << 6);
```

Or, if protection against overflow is needed:

```
// Example 7.39c
x.abc = (A & 0x0F) | ((B & 3) << 4) | ((C & 3) <<6 );
```

7.26 Overloaded functions

The different versions of an overloaded function are simply treated as different functions. There is no performance penalty for using overloaded functions.

7.27 Overloaded operators

An overloaded operator is equivalent to a function. Using an overloaded operator is exactly as efficient as using a function that does the same thing.

An expression with multiple overloaded operators will cause the creation of temporary objects for intermediate results, which may be undesired. Example:

```
// Example 7.40a
class vector {                                // 2-dimensional vector
public:
   float x, y;                                // x,y coordinates
   vector() {}                                // default constructor
   vector(float a, float b) {x = a; y = b;}   // constructor
   vector operator + (vector const & a) {     // sum operator
      return vector(x + a.x, y + a.y);}       // add elements
};

vector a, b, c, d;
a = b + c + d;              // makes intermediate object for (b + c)
```

The creation of a temporary object for the intermediate result `(b+c)` can be avoided by joining the operations:

```
// Example 7.40b
a.x = b.x + c.x + d.x;
a.y = b.y + c.y + d.y;
```

Fortunately, most compilers will do this optimization automatically in simple cases.

7.28 Templates

A template is similar to a macro in the sense that the template parameters are replaced by their values before compilation. The following example illustrates the difference between a function parameter and a template parameter:

```
// Example 7.41
int Multiply (int x, int m) {
   return x * m;}

template <int m>
int MultiplyBy (int x) {
   return x * m;}

int a, b;
a = Multiply(10,8);
b = MultiplyBy<8>(10);
```

`a` and `b` will both get the value 10 * 8 = 80. The difference lies in the way `m` is transferred to the function. In the simple function, `m` is transferred at runtime from the caller to the called function. But in the template function, `m` is replaced by its value at compile time so that the compiler sees the constant 8 rather than the variable `m`. The advantage of using a template parameter rather than a function parameter is that the overhead of parameter transfer is

avoided. The disadvantage is that the compiler needs to make a new instance of the template function for each different value of the template parameter. If `MultiplyBy` in this example is called with many different factors as template parameters then the code can become very big.

In the above example, the template function is faster than the simple function because the compiler knows that it can multiply by a power of 2 by using a shift operation. `x*8` is replaced by `x<<3`, which is faster. In the case of the simple function, the compiler doesn't know the value of `m` and therefore cannot do the optimization unless the function can be inlined. (In the above example, the compiler is able to inline and optimize both functions and simply put 80 into `a` and `b`. But in more complex cases it might not be able to do so).

A template parameter can also be a type. The example on page 39 shows how you can make arrays of different types with the same template.

Templates are efficient because the template parameters are always resolved at compile time. Templates make the source code more complex, but not the compiled code. In general, there is no cost in terms of execution speed to using templates.

Two or more template instances will be joined into one if the template parameters are exactly the same. If the template parameters differ then you will get one instance for each set of template parameters. A template with many instances makes the compiled code big and uses more cache space.

Excessive use of templates makes the code difficult to read. If a template has only one instance then you may as well use a `#define`, `const` or `typedef` instead of a template parameter.

Templates may be used for metaprogramming, as explained at page 152.

Using templates for polymorphism

A template class can be used for implementing a compile-time polymorphism, which is more efficient than the runtime polymorphism that is obtained with virtual member functions. The following example shows first the runtime polymorphism:

```
// Example 7.42a. Runtime polymorphism with virtual functions
class CHello {
public:
   void NotPolymorphic();    // Non-polymorphic functions go here
   virtual void Disp();      // Virtual function
   void Hello() {
      cout << "Hello ";
      Disp();                // Call to virtual function
   }
};

class C1 : public CHello {
   public:
   virtual void Disp() {
      cout << 1;
   }
};

class C2 : public CHello {
   public:
   virtual void Disp() {
      cout << 2;
   }
};
```

58

```
void test () {
   C1 Object1;  C2 Object2;
   CHello * p;
   p = &Object1;
   p->NotPolymorphic();       // Called directly
   p->Hello();                // Writes "Hello 1"
   p = &Object2;
   p->Hello();                // Writes "Hello 2"
}
```

The dispatching to `C1::Disp()` or `C2::Disp()` is done at runtime here if the compiler doesn't know what class of object `p` points to (see page 73). Current compilers are not very good at optimizing away `p` and inlining the call to `Object1.Hello()`, though future compilers may be able to do so.

If it is known at compile-time whether the object belongs to class `C1` or `C2`, then we can avoid the inefficient virtual function dispatch process. This can be done with a special trick which is used in the Active Template Library (ATL) and Windows Template Library (WTL):

```
// Example 7.42b. Compile-time polymorphism with templates

// Place non-polymorphic functions in the grandparent class:
class CGrandParent {
public:
   void NotPolymorphic();
};

// Any function that needs to call a polymorphic function goes in the
// parent class. The child class is given as a template parameter:
template <typename MyChild>
class CParent : public CGrandParent {
public:
   void Hello() {
      cout << "Hello ";
      // call polymorphic child function:
      (static_cast<MyChild*>(this))->Disp();
   }
};

// The child classes implement the functions that have multiple
// versions:
class CChild1 : public CParent<CChild1> {
   public:
   void Disp() {
      cout << 1;
   }
};

class CChild2 : public CParent<CChild2> {
   public:
   void Disp() {
      cout << 2;
   }
};

void test () {
   CChild1 Object1;  CChild2 Object2;
   CChild1 * p1;
   p1 = &Object1;
   p1->Hello();               // Writes "Hello 1"
   CChild2 * p2;
   p2 = &Object2;
   p2->Hello();               // Writes "Hello 2"
```

```
    }
```

Here `CParent` is a template class which gets information about its child class through a template parameter. It can call the polymorphic member of its child class by type-casting its '`this`' pointer to a pointer to its child class. This is only safe if it has the correct child class name as template parameter. In other words, you must make sure that the declaration

```
    class CChild1 : public CParent<CChild1> {
```

has the same name for the child class name and the template parameter.

The order of inheritance is now as follows. The first generation class (`CGrandParent`) contains any non-polymorphic member functions. The second generation class (`CParent<>`) contains any member functions that need to call a polymorphic function. The third generations classes contain the different versions of the polymorphic functions. The second generation class gets information about the third generation class through a template parameter.

No time is wasted on runtime dispatch to virtual member functions if the class of the object is known. This information is contained in `p1` and `p2` having different types. A disadvantage is that `CParent::Hello()` has multiple instances that take up cache space.

The syntax in example 7.42b is admittedly very kludgy. The few clock cycles that we may save by avoiding the virtual function dispatch mechanism is rarely enough to justify such a complicated code that is difficult to understand and therefore difficult to maintain. If the compiler is able to do the devirtualization (see page 73) automatically then it is certainly more convenient to rely on compiler optimization than to use this complicated template method.

7.29 Threads

Threads are used for doing two or more jobs simultaneously or seemingly simultaneously. If the computer has only one CPU core then it is not possible to do two jobs simultaneously. Each thread will get time slices of typically 30 ms for foreground jobs and 10 ms for background jobs. The context switches after each time slice are quite costly because all caches have to adapt to the new context. It is possible to reduce the number of context switches by making longer time slices. This will make applications run faster at the cost of longer response times for user input. (In Windows you can increase the time slices to 120 ms by selecting optimize performance for background services under advanced system performance options. I don't know if this is possible in Linux).

Threads are useful for assigning different priorities to different tasks. For example, in a word processor the user expects an immediate response to pressing a key or moving the mouse. This task must have a high priority. Other tasks such as spell-checking and repagination are running in other threads with lower priority. If the different tasks were not divided into threads with different priorities then the user might experience unacceptably long response times to keyboard and mouse inputs when the program is busy doing the spell checking.

Any task that takes a long time, such as heavy mathematical calculations, should be scheduled in a separate thread if the application has a graphical user interface. Otherwise the program will be unable to respond quickly to keyboard or mouse input.

It is possible to make a thread-like scheduling in an application program without invoking the overhead of the operating system thread scheduler. This can be accomplished by doing the heavy background calculations piece by piece in a function that is called from the message loop of a graphical user interface (`OnIdle` in Windows MFC). This method may be faster

than making a separate thread in systems with only one CPU core, but it requires that the background job can be divided into small pieces of a suitable duration.

The best way to fully utilize systems with multiple CPU cores is to divide the job into multiple threads. Each thread can then run on its own CPU core.

There are four kinds of costs to multithreading that we have to take into account when optimizing multithreaded applications:

- The cost of starting and stopping threads. Don't put a task into a separate thread if it is short in duration compared with the time it takes to start and stop the thread.

- The cost of task switching. This cost is minimized if the number of threads with the same priority is no more than the number of CPU cores.

- The cost of synchronizing and communicating between threads. The overhead of semaphores, mutexes, etc. is considerable. If two threads are often waiting for each other in order to get access to the same resource then it may be better to join them into one thread. A variable that is shared between multiple threads must be declared `volatile`. This prevents the compiler from doing optimizations on that variable.

- The different threads need separate storage. No function or class that is used by multiple threads should rely on static or global variables. (See thread-local storage p. 29) The threads have each their stack. This can cause cache contentions if the threads share the same cache.

Multithreaded programs must use thread-safe functions. A thread-safe function should never use static variables.

See chapter 10 page 101 for further discussion of the techniques of multithreading.

7.30 Exceptions and error handling

Exception handling is intended for detecting errors that seldom occur and recovering from error conditions in a graceful way. You may think that exception handling takes no extra time as long as the error doesn't occur, but unfortunately this is not always true. The program may have to do a lot of bookkeeping in order to know how to recover in the event of an exception. The costs of this bookkeeping depends very much on the compiler. Some compilers have efficient table-based methods with little or no overhead while other compilers have inefficient code-based methods or require runtime type identification (RTTI), which affects other parts of the code. See ISO/IEC TR18015 Technical Report on C++ Performance for further explanation.

The following example explains why bookkeeping is needed:

```
// Example 7.43
class C1 {
   public:
   ...
   ~C1();
};

void F1() {
   C1 x;
   ...
}

void F0() {
   try {
```

```
      F1();
   }
   catch (...) {
   ...
   }
}
```

The function `F1` is supposed to call the destructor for the object `x` when it returns. But what if an exception occurs somewhere in `F1`? Then we are breaking out of `F1` without returning. `F1` is prevented from cleaning up because it has been brutally interrupted. Now it is the responsibility of the exception handler to call the destructor of `x`. This is only possible if `F1` has saved all information about the destructor to call or any other cleanup that may be necessary. If `F1` calls another function which in turn calls another function, etc., and if an exception occurs in the innermost function, then the exception handler needs all information about the chain of function calls and it needs to follow the track backwards though the function calls to check for all the necessary cleanup jobs to do. This is called stack unwinding.

All functions have to save some information for the exception handler, even if no exception ever happens. This is the reason why exception handling may be expensive. If exception handling is not necessary for your application then you should disable it in order to make the code smaller and more efficient. You can disable exception handling for the whole program by turning off the exception handling option in the compiler. You can disable exception handling for a single function by adding `throw()` to the function prototype:

```
void F1() throw();
```

This tells the compiler not to save recovery information for function `F1`. It is recommended to add `throw()` to functions that are critical to program performance.

The compiler makes a distinction between *leaf functions* and *frame functions*. A frame function is a function that calls at least one other function. A leaf function is a function that doesn't call any other function. A leaf function is simpler than a frame function because the stack unwinding information can be left out if exceptions can be ruled out or if there is nothing to clean up in case of an exception. A frame function can be turned into a leaf function by inlining all the functions that it calls. The best performance is obtained if the critical innermost loop of a program contains no calls to frame functions.

In some cases, it is optimal to use exception handling even in the most critical part of a program. This is the case if alternative implementations are less efficient and you want to be able to recover from errors. The following example illustrates such a case:

```
// Example 7.44
// Portability note: This example is specific to Microsoft compilers.
#include <excpt.h>
#include <float.h>
#include <math.h>
#define EXCEPTION_FLT_OVERFLOW  0xC0000091L

void MathLoop() {
   const int arraysize = 1000;  unsigned int dummy;
   double  a[arraysize], b[arraysize], c[arraysize];

   // Enable exception for floating point overflow:
   _controlfp_s(&dummy, 0, _EM_OVERFLOW);
   // _controlfp(0, _EM_OVERFLOW); // if above line doesn't work

   int i = 0;   // Initialize loop counter outside both loops
   // The purpose of the while loop is to resume after exceptions:
   while (i < arraysize) {
```

```
            // Catch exceptions in this block:
            __try {
                // Main loop for calculations:
                for ( ; i < arraysize; i++) {

                    // Overflow may occur in multiplication here:
                    a[i] = log (b[i] * c[i]);
                }
            }
            // Catch floating point overflow but no other exceptions:
            __except (GetExceptionCode() == EXCEPTION_FLT_OVERFLOW
            ? EXCEPTION_EXECUTE_HANDLER : EXCEPTION_CONTINUE_SEARCH) {
                // Floating point overflow has occurred.
                // Reset floating point status:
                _fpreset();
                _controlfp_s(&dummy, 0, _EM_OVERFLOW);
                // _controlfp(0, _EM_OVERFLOW); // if above doesn't work

                // Re-do the calculation in a way that avoids overflow:
                a[i] = log(b[i]) + log(c[i]);

                // Increment loop counter and go back into the for-loop:
                i++;
            }
        }
    }
```

Assume that the numbers in `b[i]` and `c[i]` are so big that overflow can occur in the multiplication `b[i]*c[i]`, though this only happens rarely. The above code will catch an exception in case of overflow and redo the calculation in a way that takes more time but avoids the overflow. Taking the logarithm of each factor rather than the product makes sure that no overflow can occur, but the calculation time is doubled.

The time it takes to make support for the exception handling is negligible because there is no `try` block or function call (other than `log`) inside the critical innermost loop. `log` is a library function which we assume is optimized. We cannot change its possible exception handling support anyway. The exception is costly when it occurs, but this is not a problem since we are assuming that the occurrence is rare.

Testing for the overflow condition inside the loop does not cost anything here because we are relying on the microprocessor hardware for raising an exception in case of overflow. The exception is caught by the operating system which redirects it to the exception handler in the program if there is a `try` block.

There is a portability issue to catching hardware exceptions. The mechanism relies on non-standardized details in both compiler, operating system and CPU hardware. Porting such an application to a different platform is likely to require modifications in the code.

Let's look at the possible alternatives to exception handling in this example. We might check for overflow by checking if `b[i]` and `c[i]` are too big before multiplying them. This would require two floating point comparisons, which are relatively costly because they must be inside the innermost loop. Another possibility is to always use the safe formula `a[i] = log(b[i]) + log(c[i]);`. This would double the number of calls to `log`, and logarithms take a long time to calculate. If there is a way to check for overflow outside the loop without checking all the array elements then this might be a better solution. It might be possible to do such a check before the loop if all the factors are generated from the same few parameters. Or it might be possible to do the check after the loop if the results are combined by some formula into a single result. An uncaught overflow condition will generate the value infinity, and this value will propagate through the calculations so that the final result will be infinity or NAN (Not A Number) if an overflow or another error has occurred

anywhere in the calculations. The program can check the final result to see if it is a valid number (e.g. with `_finite()`) and redo the calculations in a safe way in case of error. The calculations may take more time than normal on some microprocessors when an operand is infinity or NAN.

Avoiding the cost of exception handling

Exception handling is not necessary when no attempt is made to recover from errors. If you just want the program to issue an error message and stop the program in case of an error then there is no reason to use `try`, `catch`, and `throw`. It is more efficient to define your own error-handling function that simply prints an appropriate error message and then calls `exit`.

Calling `exit` may not be safe if there are allocated resources that need to be cleaned up, as explained below. There are other possible ways of handling errors without using exceptions. The function that detects an error can return with an error code which the calling function can use for recovering or for issuing an error message.

It is recommended to use a systematic and well thought-through approach to error handling. You have to distinguish between recoverable and non-recoverable errors; make sure allocated resources are cleaned up in case of an error; and make appropriate error messages to the user.

Making exception-safe code

Assume that a function opens a file in exclusive mode, and an error condition terminates the program before the file is closed. The file will remain locked after the program is terminated and the user will be unable to access the file until the computer is rebooted. To prevent this kind of problems you must make your program exception safe. In other words, the program must clean up everything in case of an exception or other error condition. Things that may need to be cleaned up include:

- Memory allocated with `new` or `malloc`.

- Handles to windows, graphic brushes, etc.

- Locked mutexes.

- Open database connections.

- Open files and network connections.

- Temporary files that need to be deleted.

- User work that needs to be saved.

- Any other allocated resource.

The C++ way of handling cleanup jobs is to make a destructor. A function that reads or writes a file can be wrapped into a class with a destructor that makes sure the file is closed. The same method can be used for any other resource, such as dynamically allocated memory, windows, mutexes, database connections, etc.

The C++ exception handling system makes sure that all destructors for local objects are called. The program is exception safe if there are wrapper classes with destructors to take care of all cleanup of allocated resources.

If you make your own error handling system instead of using exception handling then you cannot be sure that all destructors are called and resources cleaned up. If an error handler calls `exit()`, `abort()`, `_endthread()`, etc. then there is no guarantee that all destructors are called. The safe way to handle an unrecoverable error without using exceptions is to return from the function. The function may return an error code if possible, or the error code may be stored in a global object. The calling function must then check for the error code. If the latter function also has something to clean up then it must return to its own caller, and so on.

7.31 Other cases of stack unwinding

The preceding paragraph described a mechanism called stack unwinding that is used by exception handlers for cleaning up and calling any necessary destructors after jumping out of a function in case of an exception without using the normal return route. This mechanism is also used in two other situations:

The stack unwinding mechanism may be used when a thread is terminated. The purpose is to detect if any objects declared in the thread have a destructor that needs to be called. It is recommended to return from functions that require cleanup before terminating a thread. You cannot be certain that a call to `_endthread()` cleans up the stack. This behaviour is implementation dependent.

The stack unwinding mechanism is also used when the function `longjmp` is used for jumping out of a function. Avoid the use of `longjmp` if possible. Don't rely on `longjmp` in time-critical code.

7.32 Preprocessing directives

Preprocessing directives (everything that begins with `#`) are costless in terms of program performance because they are resolved before the program is compiled.

`#if` directives are useful for supporting multiple platforms or multiple configurations with the same source code. `#if` is more efficient than `if` because `#if` is resolved at compile time while `if` is resolved at runtime.

`#define` directives are equivalent to `const` definitions when used for defining constants. For example, `#define ABC 123` and `static const int ABC = 123;` are equally efficient because, in most cases, an optimizing compiler can replace an integer constant with its value. However, the `const int` declaration may in some cases take memory space where a `#define` directive never takes memory space. A floating point constant always takes memory space, even when it has not been given a name.

`#define` directives when used as macros are sometimes more efficient than functions. See page 49 for a discussion.

7.33 Namespaces

There is no cost in terms of execution speed to using namespaces.

8 Optimizations in the compiler

8.1 How compilers optimize

Modern compilers can do a lot of modifications to the code in order to improve performance. It is useful for the programmer to know what the compiler can do and what it can not do. The following sections describe some of the compiler optimizations that it is relevant for the programmer to know about.

Function inlining

The compiler can replace a function call by the body of the called function. Example:

```
// Example 8.1a
float square (float a) {
   return a * a;}

float parabola (float x) {
   return square(x) + 1.0f;}
```

The compiler may replace the call to square by the code inside square:

```
// Example 8.1b
float parabola (float x) {
   return x * x + 1.0f;}
```

The advantages of function inlining are:
>   The overhead of call and return and parameter transfer are eliminated.
>   Code caching will be better because the code becomes contiguous.
>   The code becomes smaller if there is only one call to the inlined function.
>   Function inlining can open the possibility for other optimizations, as explained below.

The disadvantage of function inlining is that the code becomes bigger if there is more than one call to the inlined function and the function is big. The compiler is more likely to inline a function if it is small or if it is called from only one or a few places.

Constant folding and constant propagation

An expression or subexpression containing only constants will be replaced by the calculated result. Example:

```
// Example 8.2a
double a, b;
a = b + 2.0 / 3.0;
```

The compiler will replace this by

```
// Example 8.2b
a = b + 0.666666666666666666667;
```

This is actually quite convenient. It is easier to write `2.0/3.0` than to calculate the value and write it with many decimals. It is recommended to put a parenthesis around such a subexpression to make sure the compiler recognizes it as a subexpression. For example, `b*2.0/3.0` will be calculated as `(b*2.0)/3.0` rather than as `b*(2.0/3.0)` unless you put a parenthesis around the constant subexpression.

A constant can be propagated through a series of calculations:

```
// Example 8.3a
float parabola (float x) {
   return x * x + 1.0f;}
```

```
float a, b;
a = parabola (2.0f);
b = a + 1.0f;
```

The compiler may replace this by

```
// Example 8.3b
a = 5.0f;
b = 6.0f;
```

Constant folding and constant propagation is not possible if the expression contains a function which cannot be inlined or cannot be calculated at compile time. For example:

```
// Example 8.4
double a = sin(0.8);
```

The `sin` function is defined in a separate function library and you cannot expect the compiler to be able to inline this function and calculate it at compile time. Some compilers are able to calculate the most common math functions such as `sqrt` and `pow` at compile-time, but not the more complicated functions like `sin`.

Pointer elimination

A pointer or reference can be eliminated if the target pointed to is known. Example:

```
// Example 8.5a
void Plus2 (int * p) {
    *p = *p + 2;}

int a;
Plus2 (&a);
```

The compiler may replace this by

```
// Example 8.5b
a += 2;
```

Common subexpression elimination

If the same subexpression occurs more than once then the compiler may calculate it only once. Example:

```
// Example 8.6a
int a, b, c;
b = (a+1) * (a+1);
c = (a+1) / 4;
```

The compiler may replace this by

```
// Example 8.6b
int a, b, c, temp;
temp = a+1;
b = temp * temp;
c = temp / 4;
```

Register variables

The most commonly used variables are stored in registers (see page 28).

The maximum number of integer register variables is approximately six in 32-bit systems and fourteen in 64-bit systems.

The maximum number of floating point register variables is eight in 32-bit systems and sixteen in 64-bit systems. Some compilers have difficulties making floating point register variables in 32-bit systems unless the SSE2 (or later) instruction set is enabled.

The compiler will choose the variables that are used most for register variables. This includes pointers and references, which can be stored in integer registers. Typical candidates for register variables are temporary intermediates, loop counters, function parameters, pointers, references, 'this' pointer, common subexpressions, and induction variables (see below).

A variable cannot be stored in a register if its address is taken, i.e. if there is a pointer or reference to it. Therefore, you should avoid making any pointer or reference to a variable that could benefit from register storage.

Live range analysis

The live range of a variable is the range of code in which the variable is used. An optimizing compiler can use the same register for more than one variable if their live-ranges do not overlap or if they are sure to have the same value. This is useful when the number of available registers is limited. Example:

```
// Example 8.7
int SomeFunction (int a, int x[]) {
    int b, c;
    x[0] = a;
    b = a + 1;
    x[1] = b;
    c = b + 1;
    return c;
}
```

In this example, a, b and c can share the same register because their live ranges do not overlap. If `c = b + 1` is changed to `c = a + 2` then a and b cannot use the same register because their live ranges now overlap.

Compilers do not normally use this principle for objects stored in memory. It will not use the same memory area for different objects even when their live ranges do not overlap. See page 89 for an example of how to make different objects share the same memory area.

Join identical branches

The code can be made more compact by joining identical pieces of code. Example:

```
// Example 8.8a
double x, y, z;  bool b;
if (b) {
    y = sin(x);
    z = y + 1.;
}
else {
    y = cos(x);
    z = y + 1.;
}
```

The compiler may replace this by

```
// Example 8.8b
double x, y;  bool b;
if (b) {
    y = sin(x);
}
```

```
    else {
        y = cos(x);
    }
    z = y + 1.;
```

Eliminate jumps

Jumps can be avoided by copying the code that it jumps to. Example:

```
// Example 8.9a
int SomeFunction (int a, bool b) {
    if (b) {
        a = a * 2;
    }
    else {
        a = a * 3;
    }
    return a + 1;
}
```

This code has a jump from `a=a*2;` to `return a+1;`. The compiler can eliminate this jump by copying the return statement:

```
// Example 8.9b
int SomeFunction (int a, bool b) {
    if (b) {
        a = a * 2;
        return a + 1;
    }
    else {
        a = a * 3;
        return a + 1;
    }
}
```

A branch can be eliminated if the condition can be reduced to always true or always false:

```
// Example 8.10a
if (true) {
    a = b;
}
else {
    a = c;
}
```

Can be reduced to:

```
// Example 8.10b
a = b;
```

A branch can also be eliminated if the condition is known from a previous branch. Example:

```
// Example 8.11a
int SomeFunction (int a, bool b) {
    if (b) {
        a = a * 2;
    }
    else {
        a = a * 3;
    }
    if (b) {
        return a + 1;
```

```
        }
        else {
            return a - 1;
        }
}
```

The compiler may reduce this to:

```
// Example 8.11b
int SomeFunction (int a, bool b) {
    if (b) {
        a = a * 2;
        return a + 1;
    }
    else {
        a = a * 3;
        return a - 1;
    }
}
```

Loop unrolling

Some compilers will unroll loops if a high degree of optimization is requested. See page 46. This may be advantageous if the loop body is very small or if it opens the possibility for further optimizations. Loops with a very low repeat count may be completely unrolled to avoid the loop overhead. Example:

```
// Example 8.12a
int i, a[2];
for (i = 0; i < 2; i++) a[i] = i+1;
```

The compiler may reduce this to:

```
// Example 8.12b
int a[2];
a[0] = 1; a[1] = 2;
```

Unfortunately, some compilers unroll too much. Excessive loop unrolling is not optimal because it takes too much space in the code cache and it fills up the loop buffer that some microprocessors have. In some cases it can be useful to turn off the loop unroll option in the compiler.

Loop invariant code motion

A calculation may be moved out of a loop if it is independent of the loop counter. Example:

```
// Example 8.13a
int i, a[100], b;
for (i = 0; i < 100; i++) {
    a[i] = b * b + 1;
}
```

The compiler may change this to:

```
// Example 8.13b
int i, a[100], b, temp;
temp = b * b + 1;
for (i = 0; i < 100; i++) {
    a[i] = temp;
}
```

Induction variables

An expression that is a linear function of a loop counter can be calculated by adding a constant to the previous value. Example:

```
// Example 8.14a
int i, a[100];
for (i = 0; i < 100; i++) {
   a[i] = i * 9 + 3;
}
```

The compiler may avoid the multiplication by changing this to:

```
// Example 8.14b
int i, a[100], temp;
temp = 3;
for (i = 0; i < 100; i++) {
   a[i] = temp;
   temp += 9;
}
```

Induction variables are often used for calculating the addresses of array elements. Example:

```
// Example 8.15a
struct S1 {double a; double b;};
S1 list[100];   int i;
for (i = 0; i < 100; i++) {
   list[i].a = 1.0;
   list[i].b = 2.0;
}
```

In order to access an element in `list`, the compiler must calculate its address. The address of `list[i]` is equal to the address of the beginning of `list` plus `i*sizeof(S1)`. This is a linear function of `i` which can be calculated by an induction variable. The compiler can use the same induction variable for accessing `list[i].a` and `list[i].b`. It can also eliminate `i` and use the induction variable as loop counter when the final value of the induction variable can be calculated in advance. This reduces the code to:

```
// Example 8.15b
struct S1 {double a; double b;};
S1 list[100], *temp;
for (temp = &list[0]; temp < &list[100]; temp++) {
   temp->a = 1.0;
   temp->b = 2.0;
}
```

The factor `sizeof(S1)` = 16 is actually hidden behind the C++ syntax in example 8.15b. The integer representation of `&list[100]` is `(int)(&list[100]) = (int)(&list[0]) + 100*16`, and `temp++` actually adds 16 to the integer value of `temp`.

The compiler doesn't need induction variables to calculate the addresses of array elements of simple types because the CPU has hardware support for calculating the address of an array element if the address can be expressed as a base address plus a constant plus an index multiplied by a factor of 1, 2, 4 or 8, but not any other factor. If `a` and `b` in example 8.15a were `float` instead of `double`, then `sizeof(S1)` would be 8 and no induction variable would be needed because the CPU has hardware support for multiplying the index by 8.

The compilers I have studied do not make induction variables for floating point expressions or more complex integer expressions. See page 79 for an example of how to use induction variables for calculating a polynomial.

Scheduling

A compiler may reorder instructions for the sake of parallel execution. Example:

```
// Example 8.16
float a, b, c, d, e, f, x, y;
x = a + b + c;
y = d + e + f;
```

The compiler may interleave the two formulas in this example so that `a+b` is calculated first, then `d+e`, then `c` is added to the first sum, then `f` is added to the second sum, then the first result is stored in `x`, and last the second result is stored in `y`. The purpose of this is to help the CPU doing multiple calculations in parallel. Modern CPUs are actually able to reorder instructions without help of the compiler (see page 103), but the compiler can make this reordering easier for the CPU.

Algebraic reductions

Most compilers can reduce simple algebraic expressions using the fundamental laws of algebra. For example, a compiler may change the expression `-(-a)` to `a`.

I don't think that programmers write expressions like `-(-a)` very often, but such expressions may occur as a result of other optimizations such as function inlining. Reducible expressions also occur quite often as a result of macro expansions.

Programmers do, however, often write expressions that can be reduced. This may be because the non-reduced expression better explains the logic behind the program or because the programmer hasn't thought about the possibility of algebraic reduction. For example, a programmer may prefer to write `if(!a && !b)` rather than the equivalent `if(!(a || b))` even though the latter has one operator less. Fortunately, all compilers are able to do the reduction in this case.

You cannot expect a compiler to reduce complicated algebraic expressions. For example, only one of the compilers I have tested were able to reduce `(a*b*c)+(c*b*a)` to `a*b*c*2`. It is quite difficult to implement the many rules of algebra in a compiler. Some compilers can reduce some types of expressions and other compilers can reduce other types of expressions, but no compiler I have ever seen can reduce them all. In the case of Boolean algebra, it is possible to implement a universal algorithm (e.g. Quine–McCluskey or Espresso) that can reduce any expression, but none of the compilers I have tested seem to do so.

The compilers are better at reducing integer expressions than floating point expressions, even though the rules of algebra are the same in both cases. This is because algebraic manipulations of floating point expressions may have undesired effects. This effect can be illustrated by the following example:

```
// Example 8.17
char a = -100, b = 100, c = 100, y;
y = a + b + c;
```

Here, `y` will get the value -100+100+100 = 100. Now, according to the rules of algebra, we may write:

```
y = c + b + a;
```

This may be useful if the subexpression `c+b` can be reused elsewhere. In this example, we are using 8-bit integers which range from -128 to +127. An integer overflow will make the value wrap around. Adding 1 to 127 will generate -128, and subtracting 1 from -128 generates 127. The calculation of `c+b` will generate an overflow and give the result -56 rather than 200. Next, we are adding -100 to -56 which will generate an underflow and give the result 100 rather than -156. Surprisingly, we end up with the correct result because the overflow and underflow neutralize each other. This is the reason why it is safe to use algebraic manipulations on integer expressions (except for the <, <=, > and >= operators).

The same argument does not apply to floating point expressions. Floating point variables do not wrap around on overflow and underflow. The range of floating point variables is so large that we do not have to worry much about overflow and underflow except in special mathematical applications. But we do have to worry about loss of precision. Let's repeat the above example with floating point numbers:

```
// Example 8.18
float a = -1.0E8, b = 1.0E8, c = 1.23456, y;
y = a + b + c;
```

The calculation here gives `a+b=0`, and then `0+1.23456 = 1.23456`. But we will not get the same result if we change the order of the operands and add `b` and `c` first. `b+c = 100000001.23456`. The `float` type holds a precision of approximately seven significant digits, so the value of `b+c` will be rounded to `100000000`. When we add `a` to this number we get `0` rather than `1.23456`.

The conclusion to this argument is that the order of floating point operands cannot be changed without the risk of losing precision. The compilers will not do so unless you specify an option that allows less precise floating point calculations. Even with all relevant optimization options turned on, the compilers will not do such obvious reductions as `0/a = 0` because this would be invalid if `a` was zero or infinity or NAN (not a number). Different compilers behave differently because there are different opinions on which imprecisions should be allowed and which not.

You cannot rely on the compiler to do any algebraic reductions on floating point code and you can rely on only the most simple reductions on integer code. It is more safe to do the reductions manually. I have tested the capability to reduce various algebraic expressions on seven different compilers. The results are listed in table 8.1 below.

Devirtualization

An optimizing compiler can bypass the virtual table lookup for a virtual function call if it is known which version of the virtual function is needed. Example:

```
// Example 8.19. Devirtualization
class C0 {
   public:
   virtual void f();
};

class C1 : public C0 {
   public:
   virtual void f();
};

void g() {
   C1 obj1;
   C0 * p = & obj1;
   p->f();                 // Virtual call to C1::f
}
```

73

Without optimization, the compiler needs to look up in a virtual table to see whether the call `p->f()` goes to `C0::f` or `C1::f`. But an optimizing compiler will see that `p` always points to an object of class `C1`, so it can call `C1::f` directly without using the virtual table. Unfortunately, few compilers are able to do this optimization.

8.2 Comparison of different compilers

I have made a series of experiments on seven different brands of C++ compilers to see whether they were able to do different kinds of optimizations. The results are summarized in table 8.1. The table shows whether the different compilers succeeded in applying the various optimization methods and algebraic reductions in my test examples.

The table can give some indication of which optimizations you can expect a particular compiler to do and which optimizations you have to do manually.

It must be emphasized that the compilers may behave differently on different test examples. You cannot expect a compiler to always behave according to the table.

Optimization method	Microsoft	Borland	Intel	Gnu	PathScale	PGI	Digital Mars	Watcom	Codeplay
Function inlining	x	-	x	x	x	x	-	-	x
Constant folding	x	x	x	x	x	x	x	x	x
Constant propagation	x	-	x	x	x	x	-	-	x
Pointer elimination	x	x	x	x	x	x	x	x	x
Common subexpression elimin., integer	x	(x)	x	x	x	x	x	x	x
Common subexpression elimin., float	x	-	x	x	x	x	-	x	x
Register variables, integer	x	x	x	x	x	x	x	x	x
Register variables, float	x	-	x	x	x	x	-	x	x
Live range analysis	x	x	x	x	x	x	x	x	x
Join identical branches	x	-	-	x	-	-	-	x	-
Eliminate jumps	x	x	x	x	x	x	-	x	x
Eliminate branches	x	-	x	x	x	x	-	-	-
Remove branch that is always true/false	x	-	x	x	x	x	x	x	x
Loop unrolling	x	-	x	x	x	x	-	-	x
Loop invariant code motion	x	-	x	x	x	x	x	x	x
Induction variables for array elements	x	x	x	x	x	x	x	x	x
Induction variables for other integer expressions	x	-	x	x	x	-	x	x	x
Induction variables for float expressions	-	-	-	-	-	-	-	-	-
Automatic vectorization	-	-	x	x	x	x	-	-	x
Devirtualization	-	-	-	x	-	-	-	-	-
Profile-guided optimization	x	-	x	x	x	x	-	-	-
Whole program optimization	x	-	x	x	x	-	-	-	-
Integer algebra reductions:									
a+b = b+a	x	(x)	x	x	x	x	-	x	x
a*b = b*a	x	(x)	x	x	x	x	-	x	x
(a+b)+c = a+(b+c)	x	-	x	x	-	-	x	x	-
a+b+c = c+b+a	x	-	-	x	-	-	-	-	-
a+b+c+d = (a+b)+(c+d)	-	-	x	x	-	-	-	-	-
a*b+a*c = a*(b+c)	x	-	x	x	x	-	-	-	x
a*x*x*x + b*x*x + c*x + d =	x	-	x	x	x	-	-	-	x

74

((a*x+b)*x+c)*x+d								
x*x*x*x*x*x*x*x = $((x^2)^2)^2$	-	-	x	-	-	-	-	-
a+a+a+a = a*4	x	-	x	x	-	-	-	-	x
-(-a) = a	x	-	x	x	x	x	x	x	-
a-(-b) = a+b	x	-	x	x	x	x	-	x	-
a-a = 0	x	-	x	x	x	x	x	x	x
a+0 = a	x	x	x	x	x	x	x	x	x
a*0 = 0	x	x	x	x	x	x	x	-	x
a*1 = a	x	x	x	x	x	x	x	x	x
(-a)*(-b) = a*b	x	-	x	x	x	-	-	-	-
a/a = 1	-	-	-	-	x	-	-	-	x
a/1 = a	x	x	x	x	x	x	x	x	x
0/a = 0	-	-	-	x	-	-	-	x	x
(-a == -b) = (a == b)	-	-	-	x	x	-	-	-	-
(a+c == b+c) = (a == b)	-	-	-	-	x	-	-	-	-
!(a < b) = (a >= b)	x	x	x	x	x	x	x	x	x
(a<b && b<c && a<c) = (a<b && b<c)	-	-	-	-	-	-	-	-	-
Multiply by constant = shift and add	x	x	x	x	-	x	x	x	-
Divide by constant = multiply and shift	x	-	x	x	x	(-)	x	-	-
Floating point algebra reductions:									
a+b = b+a	x	-	x	x	x	x	-	-	x
a*b = b*a	x	-	x	x	x	x	-	-	x
a+b+c = a+(b+c)	x	-	x	x	-	-	-	-	-
(a+b)+c = a+(b+c)	-	-	x	x	-	-	-	-	-
a*b*c = a*(b*c)	x	-	-	x	-	-	-	-	-
a+b+c+d = (a+b)+(c+d)	-	-	-	x	-	-	-	-	-
a*b+a*c = a*(b+c)	x	-	-	-	x	-	-	-	x
a*x*x*x + b*x*x + c*x + d = ((a*x+b)*x+c)*x+d	x	-	x	x	x	-	-	-	-
x*x*x*x*x*x*x*x = $((x^2)^2)^2$	-	-	-	x	-	-	-	-	-
a+a+a+a = a*4	x	-	-	x	x	-	-	-	-
-(-a) = a	-	-	x	x	x	x	x	x	-
a-(-b) = a+b	-	-	-	x	x	x	-	x	-
a+0 = a	x	-	x	x	x	x	x	x	-
a*0 = 0	-	-	x	x	x	x	-	x	x
a*1 = a	x	-	x	x	x	x	x	-	x
(-a)*(-b) = a*b	-	-	-	x	x	x	-	-	-
a/a = 1	-	-	-	-	-	-	-	-	x
a/1 = a	x	-	x	x	x	-	x	-	-
0/a = 0	-	-	-	x	x	-	-	x	x
(-a == -b) = (a == b)	-	-	-	x	x	-	-	-	-
(-a > -b) = (a < b)	-	-	-	x	x	-	-	-	x
Divide by constant = multiply by reciprocal	x	x	-	x	x	-	-	x	-
Boolean algebra reductions:									
!(!a) = a	x	-	x	x	x	x	x	x	x
(a&&b) || (a&&c) = a&&(b||c)	x	-	x	x	x	-	-	-	-
!a && !b = !(a || b)	x	x	x	x	x	x	x	x	x
a && !a = false, a || !a = true	x	-	x	x	x	x	-	-	-
a && true = a, a || false = a	x	x	x	x	x	x	x	x	-
a && false = false, a || true = true	x	-	x	x	x	x	x	x	-
a && a = a	x	-	x	x	x	x	-	-	-
(a&&b) || (a&&!b) = a	x	-	-	x	x	-	-	-	-
(a&&b) || (!a&&c) = a ? b : c	x	-	x	x	-	-	-	-	-

(a&&b) || (!a&&c) || (b&&c) = a ? b : c	x	-	-	x	-	-	-	-	-
(a&&b) || (a&&b&&c) = a&&b	x	-	-	x	x	-	-	-	-
(a&&b) || (a&&c) || (a&&b&&c) = a&&(b||c)	x	-	-	x	x	-	-	-	-
(a&&!b) || (!a&&b) = a XOR b	-	-	-	-	-	-	-	-	-
Bit vector algebra reductions:									
~(~a) = a	x	-	x	x	x	x	x	-	-
(a&b)|(a&c) = a&(b|c)	x	-	x	x	x	x	-	-	x
(a|b)&(a|c) = a|(b&c)	x	-	x	x	x	x	-	-	x
~a & ~b = ~(a | b)	-	-	x	x	x	x	-	-	-
a & a = a	x	-	-	x	x	x	-	-	x
a & ~a = 0	-	-	x	x	x	x	-	-	-
a & -1 = a, a | 0 = a	x	-	x	x	x	x	x	x	x
a & 0 = 0, a | -1 = -1	x	-	x	x	x	x	x	x	x
(a&b) | (~a&c) | (b&c) = (a&b) | (~a&c)	-	-	-	-	-	-	-	-	-
a&b&c&d = (a&b)&(c&d)	-	-	-	x	-	-	-	-	-
a ^ 0 = a	x	x	x	x	x	-	x	x	x
a ^ -1 = ~a	x	-	x	x	x	-	x	x	-
a ^ a = 0	x	-	x	x	x	x	-	x	x
a ^ ~a = -1	-	-	-	x	x	x	-	-	-
(a&~b) | (~a&b) = a ^ b	-	-	-	-	-	-	-	-	-
~a ^ ~b = a ^ b	-	-	-	x	x	-	-	-	-
a<<b<<c = a<<(b+c)	x	-	x	x	x	-	-	x	x
Integer XMM (vector) reductions:									
Common subexpression elimination	x	n.a.	x	x	x	-	n.a.	n.a.	x
Constant folding	-	n.a.	-	x	-	-	n.a.	n.a.	-
a+b = b+a, a*b = b*a	-	n.a.	-	x	-	-	n.a.	n.a.	x
(a+b)+c = a+(b+c)	-	n.a.	-	-	-	-	n.a.	n.a.	-
a*b+a*c = a*(b+c)	-	n.a.	-	-	-	-	n.a.	n.a.	-
x*x*x*x*x*x*x*x = $((x^2)^2)^2$	-	n.a.	-	-	-	-	n.a.	n.a.	-
a+a+a+a = a*4	-	n.a.	-	-	-	-	n.a.	n.a.	-
-(-a) = a	-	n.a.	-	-	-	-	n.a.	n.a.	-
a-a = 0	-	n.a.	x	-	-	-	n.a.	n.a.	-
a+0 = a	-	n.a.	-	-	-	-	n.a.	n.a.	-
a*0 = 0	-	n.a.	-	x	-	-	n.a.	n.a.	-
a*1 = a	-	n.a.	-	x	-	-	n.a.	n.a.	-
(-a)*(-b) = a*b	-	n.a.	-	-	-	-	n.a.	n.a.	-
!(a < b) = (a >= b)	-	n.a.	-	-	-	-	n.a.	n.a.	-
Floating point XMM (vector) reductions:									
a+b = b+a, a*b = b*a	x	n.a.	-	x	-	-	n.a.	n.a.	x
a+b+c = a+(b+c)	-	n.a.	-	-	-	-	n.a.	n.a.	-
a*b+a*c = a*(b+c)	-	n.a.	-	-	-	-	n.a.	n.a.	-
-(-a) = a	-	n.a.	-	-	-	-	n.a.	n.a.	-
a-a = 0	-	n.a.	-	x	-	-	n.a.	n.a.	-
a+0 = a	-	n.a.	x	-	-	-	n.a.	n.a.	-
a*0 = 0	-	n.a.	x	-	-	-	n.a.	n.a.	-
a*1 = a	-	n.a.	-	x	-	-	n.a.	n.a.	-
a/1 = a	-	n.a.	-	x	-	-	n.a.	n.a.	-
0/a = 0	-	n.a.	x	x	-	-	n.a.	n.a.	-
Divide by constant = multiply by reciprocal	-	n.a.	-	-	-	-	n.a.	n.a.	-

Boolean XMM (vector) reductions:									
~(~a) = a	-	n.a.	-	-	-	-	n.a.	n.a.	-
(a&b)|(a&c) = a&(b|c)	-	n.a.	-	-	-	-	n.a.	n.a.	-
a & a = a, a	a = a	-	n.a.	x	x	-	-	n.a.	n.a.	-
a & ~a = 0	-	n.a.	-	x	-	-	n.a.	n.a.	-
a & -1 = a, a	0 = a	-	n.a.	-	-	-	-	n.a.	n.a.	-
a & 0 = 0	-	n.a.	-	x	-	-	n.a.	n.a.	-
a	-1 = -1	-	n.a.	-	-	-	-	n.a.	n.a.	-
a ^ a = 0	-	n.a.	x	x	-	-	n.a.	n.a.	-
andnot(a,a) = 0	-	n.a.	-	x	-	-	n.a.	n.a.	-
a<<b<<c = a<<(b+c)	-	n.a.	-	-	-	-	n.a.	n.a.	-

Table 8.1. Comparison of optimizations in different C++ compilers

The tests were carried out with all relevant optimization options turned on, including relaxed floating point precision. The following compiler versions were tested:
Microsoft C++ Compiler v. 14.00 for 80x86 / x64 (Visual Studio 2005).
Borland C++ 5.82 (Embarcadero/CodeGear/Borland C++ Builder 5, 2009).
Intel C++ Compiler v. 11.1 for IA-32/Intel64, 2009.
Gnu C++ v. 4.1.0, 2006 (Red Hat).
PathScale C++ v. 3.1, 2007.
PGI C++ v. 7.1-4, 2008.
Digital Mars Compiler v. 8.42n, 2004.
Open Watcom C/C++ v. 1.4, 2005.
Codeplay VectorC v. 2.1.7, 2004.
No differences were observed between the optimization capabilities for 32-bit and 64-bit code for the Microsoft, Intel, Gnu and PathScale compilers.

8.3 Obstacles to optimization by compiler

There are several factors that can prevent the compiler from doing the optimizations that we want it to do. It is important for the programmer to be aware of these obstacles and to know how to avoid them. Some important obstacles to optimization are discussed below.

Pointer aliasing

When accessing a variable through a pointer or reference, the compiler may not be able to completely rule out the possibility that the variable pointed to is identical to some other variable in the code. Example:

```
// Example 8.20
void Func1 (int a[], int * p) {
   int i;
   for (i = 0; i < 100; i++) {
      a[i] = *p + 2;
   }
}

void Func2() {
   int list[100];
   Func1(list, &list[8]);
}
```

Here, it is necessary to reload `*p` and calculate `*p+2` a hundred times because the value pointed to by `p` is identical to one of the elements in `a[]` which will change during the loop. It is not permissible to assume that `*p+2` is a loop-invariant code that can be moved out of the loop. Example 8.20 is indeed a very contrived example, but the point is that the compiler cannot rule out the theoretical possibility that such contrived examples exist. Therefore the compiler is prevented from assuming that `*p+2` is a loop-invariant expression that it can move outside the loop.

Most compilers have an option for assuming no pointer aliasing (`/Oa`). The easiest way to overcome the obstacle of possible pointer aliasing is to turn on this option. This requires that you analyze all pointers and references in the code carefully to make sure that no variable or object is accessed in more than one way in the same part of the code. It is also possible to tell the compiler that a specific pointer does not alias anything by using the keyword `__restrict` or `__restrict__`, if supported by the compiler.

We can never be sure that the compiler takes the hint about no pointer aliasing. The only way to make sure that the code is optimized is to do it explicitly. In example 8.20, you could calculate `*p+2` and store it in a temporary variable outside the loop if you are sure that the pointer does not alias any elements in the array. This method requires that you can predict where the obstacles to optimization are.

Cannot optimize across modules

The compiler doesn't have information about functions in other modules than the one it is compiling. This prevents it from making optimizations across function calls. Example:

```
// Example 8.21
module1.cpp
int Func1(int x) {
    return x*x + 1;
}

module2.cpp
int Func2() {
    int a = Func1(2);
    ...
}
```

If `Func1` and `Func2` were in the same module then the compiler would be able do function inlining and constant propagation and reduce `a` to the constant 5. But the compiler doesn't have the necessary information about `Func1` when compiling `module2.cpp`.

The simplest way to solve this problem is to combine the multiple `.cpp` modules into one by means of `#include` directives. This is sure to work on all compilers. Some compilers have a feature called whole program optimization, which will enable optimizations across modules (See page 81).

Pure functions

A pure function is a function that has no side-effects and its return value depends only on the values of its arguments. This closely follows the mathematical notion of a "function".

Multiple calls to a pure function with the same arguments are sure to produce the same result. A compiler can eliminate common subexpressions that contain pure function calls and it can move out loop-invariant code containing pure function calls. Unfortunately, the compiler cannot know that a function is pure if the function is defined in a different module or a function library.

Therefore, it is necessary to do optimizations such as common subexpression elimination, constant propagation, and loop-invariant code motion manually when it involves pure function calls.

The Gnu compiler and the Intel compiler for Linux have an attribute which can be applied to a function prototype to tell the compiler that this is a pure function. Example:

```
// Example 8.22
#ifdef __GNUC__
```

```
    #define pure_function   __attribute__((const))
    #else
    #define pure_function
    #endif

    double Func1(double) pure_function ;

    double Func2(double x) {
       return Func1(x) * Func1(x) + 1.;
    }
```

Here, the Gnu compiler will make only one call to `Func1`, while other compilers will make two.

Some other compilers (Microsoft, Intel) know that standard library functions like `sqrt`, `pow` and `log` are pure functions, but unfortunately there is no way to tell these compilers that a user-defined function is pure.

Algebraic reduction

Most compilers can do simple algebraic reductions such as -(-a) = a, but they are not able to do more complicated reductions. Algebraic reduction is a complicated process which is difficult to implement in a compiler.

Many algebraic reductions are not permissible for reasons of mathematical purity. In many cases it is possible to construct obscure examples where the reduction would cause overflow or loss of precision, especially in floating point expressions (see page 72). The compiler cannot rule out the possibility that a particular reduction would be invalid in a particular situation, but the programmer can. It is therefore necessary to do the algebraic reductions explicitly in many cases.

Integer expressions are less susceptible to problems of overflow and loss of precision for reasons explained on page 72. It is therefore possible for the compiler to do more reductions on integer expressions than on floating point expressions. Most reductions involving integer addition, subtraction and multiplication are permissible in all cases, while many reductions involving division and relational operators (e.g. '>') are not permissible for reasons of mathematical purity. For example, compilers cannot reduce the integer expression `-a > -b` to `a < b` because of a very obscure possibility of overflow.

Table 8.1 (page 77) shows which reductions the compilers are able to do, at least in some situations, and which reductions they cannot do. All the reductions that the compilers cannot do must be done manually by the programmer.

Floating point induction variables

Compilers cannot make floating point induction variables for the same reason that they cannot make algebraic reductions on floating point expressions. It is therefore necessary to do this manually. This principle is useful whenever a function of a loop counter can be calculated more efficiently from the previous value than from the loop counter. Any expression that is an n'th degree polynomial of the loop counter can be calculated by n additions and no multiplications. The following example shows the principle for a 2'nd order polynomial:

```
    // Example 8.23a. Loop to make table of polynomial
    const double A = 1.1, B = 2.2, C = 3.3; // Polynomial coefficients
    double Table[100];                       // Table
    int x;                                   // Loop counter
    for (x = 0; x < 100; x++) {
       Table[x] = A*x*x + B*x + C;           // Calculate polynomial
    }
```

The calculation of this polynomial can be done with just two additions by the use of two induction variables:

```cpp
// Example 8.23b. Calculate polynomial with induction variables
const double A = 1.1, B = 2.2, C = 3.3; // Polynomial coefficients
double Table[100];                      // Table
int x;                                  // Loop counter
const double A2 = A + A;                // = 2*A
double Y = C;                           // = A*x*x + B*x + C
double Z = A + B;                       // = Delta Y
for (x = 0; x < 100; x++) {
    Table[x] = Y;                       // Store result
    Y += Z;                             // Update induction variable Y
    Z += A2;                            // Update induction variable Z
}
```

The loop in example 8.23b has two loop-carried dependency chains, namely the two induction variables Y and Z. Each dependency chain has a latency which is the same as the latency of a floating point addition. This is small enough to justify the method. A longer loop-carried dependency chain would make the induction variable method unfavorable, unless the value is calculated from a value that is two or more iterations back.

The method of induction variables can also be vectorized if you take into account that each value is calculated from the value that lies r places back in the sequence, where r is the number of elements in a vector or the loop unroll factor. A little math is required for finding the right formula in each case.

Inlined functions have a non-inlined copy

Function inlining has the complication that the same function may be called from another module. The compiler has to make a non-inlined copy of the inlined function for the sake of the possibility that the function is also called from another module. This non-inlined copy is dead code if no other modules call the function. This fragmentation of the code makes caching less efficient.

There are various ways around this problem. If a function is not referenced from any other module then add the keyword `static` to the function definition. This tells the compiler that the function cannot be called from any other module. The `static` declaration makes it easier for the compiler to evaluate whether it is optimal to inline the function, and it prevents the compiler from making an unused copy of an inlined function. The `static` keyword also makes various other optimizations possible because the compiler doesn't have to obey any specific calling conventions for functions that are not accessible from other modules. You may add the `static` keyword to all local non-member functions.

Unfortunately, this method doesn't work for class member functions because the `static` keyword has a different meaning for member functions. You can force a member function to be inlined by declaring the function body inside the class definition. This will prevent the compiler from making a non-inlined copy of the function, but it has the disadvantage that the function is always inlined even when it is not optimal to do so (i.e. if the member function is big and is called from many different places).

Some compilers have an option (Windows: `/Gy`, Linux: `-ffunction-sections`) which allows the linker to remove unreferenced functions. It is recommended to turn on this option.

8.4 Obstacles to optimization by CPU

Modern CPUs can do a lot of optimization by executing instructions out of order. Long dependency chains in the code prevent the CPU from doing out-of-order execution, as explained on page 23.

Avoid long dependency chains, especially loop-carried dependency chains with long latencies.

8.5 Compiler optimization options

All C++ compilers have various optimization options that you can turn on and off. It is important to study the available options for the compiler you are using and turn on all relevant options.

Many optimization options are incompatible with debugging. A debugger can execute a code one line at a time and show the values of all variables. Obviously, this is not possible when parts of the code have been reordered, inlined, or optimized away. It is common to make two versions of a program executable: a debug version with full debugging support which is used during program development, and a release version with all relevant optimization options turned on. Most IDE's (Integrated Development Environments) have facilities for making a debug version and a release version of object files and executables. Make sure to distinguish these two versions and turn off debugging and profiling support in the optimized version of the executable.

Most compilers offer the choice between optimizing for size and optimizing for speed. Optimizing for size is relevant when the code is fast anyway and you want the executable to be as small as possible or when code caching is critical. Optimizing for speed is relevant when CPU access and memory access are critical time consumers. Choose the strongest optimization option available.

Some compilers offer profile-guided optimization. This works in the following way. First you compile the program with profiling support. Then you make a test run with a profiler which determines the program flow and the number of times each function and branch is executed. The compiler can then use this information to optimize the code and put the different functions in the optimal order.

Some compilers have support for whole program optimization. This works by compiling in two steps. All source files are first compiled to an intermediate file format instead of the usual object file format. The intermediate files are then linked together in the second step where the compilation is finished. Register allocation and function inlining is done at the second step. The intermediate file format is not standardized. It is not even compatible with different versions of the same compiler. It is therefore not possible to distribute function libraries in this format.

Other compilers offer the possibility of compiling multiple `.cpp` files into a single object file. This enables the compiler to do cross-module optimizations when interprocedural optimization is enabled. A more primitive, but efficient, way of doing whole program optimization is to join all source files into one by means of `#include` directives and declare all functions static or inline. This will enable the compiler to do interprocedural optimizations of the whole program.

During the history of CPU development, each new generation of CPUs increased the available instruction set. The newer instruction sets enable the compiler to make more efficient code, but this makes the code incompatible with old CPUs. The Pentium Pro instruction set makes floating point comparisons more efficient. This instruction set is supported by all modern CPUs. The SSE2 instruction set is particularly interesting because it makes floating point code more efficient in some cases and it makes it possible to use vector instructions (see page 105). Using the SSE2 instruction set is not always optimal, though. In some cases the SSE2 instruction set makes floating point code slower, especially when the code mixes float and double (see page 146). The SSE2 instruction set is supported by most CPUs and operating systems available today.

You may choose a newer instruction set when compatibility with old CPUs is not needed. Even better, you may make multiple versions of the most critical part of the code to support different CPUs. This method is explained on page 125.

The code becomes more efficient when there is no exception handling. It is recommended to turn off support for exception handling unless the code relies on structured exception handling and you want the code to be able to recover from exceptions. See page 61.

It is recommended to turn off support for runtime type identification (RTTI). See page 54.

It is recommended to enable fast floating point calculations or turn off requirements for strict floating point calculations unless the strictness is required. See page 73 and 72 for discussions.

Turn on the option for "function level linking" if available. See page 80 for an explanation of this option.

Use the option for "assume no pointer aliasing" if you are sure the code has no pointer aliasing. See page 77 for an explanation. (The Microsoft compiler supports this option only in the Professional and Enterprise editions).

Do not turn on correction for the "FDIV bug". The FDIV bug is a minor error in the oldest Pentium CPUs which may cause slight imprecision in some rare cases of floating point division. Correction for the FDIV bug causes floating point division to be slower.

Many compilers have an option for "standard stack frame" or "frame pointer". The standard stack frame is used for debugging and exception handling. Omitting the standard stack frame makes function calls faster and makes an extra register available for other purposes. This is advantageous because registers is a scarce resource. Do not use a stack frame unless your program relies on exception handling.

8.6 Optimization directives

Some compilers have many keywords and directives which are used for giving specific optimization instructions at specific places in the code. Many of these directives are compiler-specific. You cannot expect a directive for a Windows compiler to work on a Linux compiler, or vice versa. But most of the Microsoft directives work on the Intel compiler for Windows and the Gnu compiler for Windows, while most of the Gnu directives work on the PathScale and Intel compilers for Linux.

Keywords that work on all C++ compilers

The `register` keyword can be added to a variable declaration to tell the compiler that you want this to be a register variable. The register keyword is only a hint and the compiler may not take the hint, but it can be useful in situations where the compiler is unable to predict which variables will be used most.

The opposite of `register` is `volatile`. The `volatile` keyword makes sure that a variable is never stored in a register, not even temporarily. This is intended for variables that are shared between multiple threads, but it can also be used for turning off all optimizations of a variable for test purposes.

The `const` keyword tells that a variable is never changed. This will allow the compiler to optimize away the variable in many cases. For example:

```
// Example 8.24. Integer constant
const int ArraySize = 1000;
```

82

```
    int List[ArraySize];
    ...
    for (int i = 0; i < ArraySize; i++) List[i]++;
```

Here, the compiler can replace all occurrences of `ArraySize` by the value `1000`. The loop in example 8.24 can be implemented in a more efficient way if the value of the loop count (`ArraySize`) is constant and known to the compiler at compile time. No memory will be allocated for an integer constant, unless the address of it (`&ArraySize`) is taken.

A `const` pointer or `const` reference cannot change what it points to. A `const` member function cannot modify data members. It is recommended to use the `const` keyword wherever appropriate to give the compiler additional information about a variable, pointer or member function because this may improve the possibilities for optimization. For example, the compiler can safely assume that the value of a class data member is unchanged across a call to a `const` function that is member of the same class.

The `static` keyword has several meanings depending on the context. The keyword `static`, when applied to a non-member function, means that the function is not accessed by any other modules. This makes inlining more efficient and enables interprocedural optimizations. See page 80.

The keyword `static`, when applied to a global variable means that it is not accessed by any other modules. This enables interprocedural optimizations.

The keyword `static`, when applied to a local `const` variable means that it is initialized only the first time the function is called. Example:

```
// Example 8.25
void Func () {
    static const double log2 = log(2.0);
    ...
}
```

Here, `log(2.0)` is only calculated the first time `Func` is executed. Without `static`, the logarithm would be re-calculated every time `Func` is executed. This has the disadvantage that the function must check if it has been called before. This is faster than calculating the logarithm again, but it would be even faster to make `log2` a global `const` variable or replace it with the calculated value.

The keyword `static`, when applied to a class member function means that it cannot access any non-static data members or member functions. A static member function is called faster than a non-static member function because it doesn't need a 'this' pointer. It is recommended to make member functions `static` where appropriate.

Compiler-specific keywords

Fast function calling. `__fastcall` or `__attribute__((fastcall))`. The fastcall modifier can make function calls faster in 32-bit mode. The first two integer parameters are transferred in registers rather than on the stack (three parameters on CodeGear compiler). Fastcall functions are not compatible across compilers. Fastcall is not needed in 64-bit mode where the parameters are transferred in registers anyway.

Pure function. `__attribute__((const))` (Linux only). Specifies a function to be pure. This allows common subexpression elimination and loop-invariant code motion. See page 78.

Assume no pointer aliasing. `__declspec(noalias)` or `__restrict` or
`#pragma optimize("a",on)`. Specifies that pointer aliasing does not occur. See page 77 for an explanation. Note that these directives do not always work.

Data alignment. `__declspec(align(16))` or `__attribute__((aligned(16)))`.
Specifies alignment of arrays and structures. Useful for vector operations, see page 105.

8.7 Checking what the compiler does

It can be very useful to study the code that a compiler generates to see how well it optimizes the code. Sometimes the compiler does quite ingenious things to make the code more efficient, and sometimes it does incredibly stupid things. Looking at the compiler output can often reveal things that can be improved by modifications of the source code, as the example below shows.

The best way to check the code that the compiler generates is to use a compiler option for assembly language output. On most compilers you can do this by invoking the compiler from the command line with all the relevant optimization options and the options `-S` or `/Fa` for assembly output. The assembly output option is also available from the IDE on some systems. If the compiler doesn't have an assembly output option then use an object file disassembler.

Note that the Intel compiler has an option for source annotation in the assembly output (`/FAs` or `-fsource-asm`). This option makes the assembly output more readable but unfortunately it prevents certain optimizations. Do not use the source annotation option if you want to see the result of full optimization.

It is also possible to see the compiler-generated code in the disassembly window of a debugger. However, the code that you see in the debugger is not the optimized version because the debugging options prevent optimization. The debugger cannot set a breakpoint in the fully optimized code because it doesn't have the line number information. It is often possible to insert a fixed breakpoint in the code with an inline assembly instruction for interrupt 3. The code is `__asm int 3;` or `__asm ("int 3");` or `__debugbreak();`. If you run the optimized code (release version) in the debugger then it will break at the interrupt 3 breakpoint and show a disassembly, probably without information about function names and variable names. Remember to remove the interrupt 3 breakpoint again.

The following example shows what the assembly output of a compiler can look like and how you can use it for improving the code.

```
// Example 8.26a
void Func(int a[], int & r) {
   int i;
   for (i = 0; i < 100; i++) {
      a[i] = r + i/2;
   }
}
```

The Intel compiler generates the following assembly code from example 8.26a (32-bit mode):

```
; Example 8.26a compiled to assembly:
ALIGN     4                              ; align by 4
PUBLIC ?Func@@YAXQAHAAH@Z                ; mangled function name
?Func@@YAXQAHAAH@Z   PROC NEAR           ; start of Func
; parameter 1: 8 + esp                   ; a
; parameter 2: 12 + esp                  ; r
$B1$1:                                   ; unused label
        push      ebx                    ; save ebx on stack
```

```
        mov         ecx, DWORD PTR [esp+8]    ; ecx = a
        xor         eax, eax                  ; eax = i = 0
        mov         edx, DWORD PTR [esp+12]   ; edx = r
  $B1$2:                                       ; top of loop
        mov         ebx, eax                  ; compute i/2 in ebx
        shr         ebx, 31                   ; shift down sign bit of i
        add         ebx, eax                  ; i + sign(i)
        sar         ebx, 1                    ; shift right = divide by 2
        add         ebx, DWORD PTR [edx]      ; add what r points to
        mov         DWORD PTR[ecx+eax*4],ebx  ; store result in array
        add         eax, 1                    ; i++
        cmp         eax, 100                  ; check if i < 100
        jl          $B1$2                     ; repeat loop if true
  $B1$3:                                       ; unused label
        pop         ebx                       ; restore ebx from stack
        ret                                   ; return
        ALIGN       4                         ; align
  ?Func@@YAXQAHAAH@Z ENDP                      ; mark end of procedure
```

Most of the comments generated by the compiler have been replaced by my comments, in green. It takes some experience to get used to read and understand compiler-generated assembly code. Let me explain the above code in details. The funny looking name `?Func@@YAXQAHAAH@Z` is the name of `Func` with a lot of added information about the function type and its parameters. This is called name mangling. The characters '`?`', '`@`' and '`$`' are allowed in assembly names. The details about name mangling are explained in manual 5: "Calling conventions for different C++ compilers and operating systems". The parameters `a` and `r` are transferred on the stack at address `esp+8` and `esp+12` and loaded into `ecx` and `edx`, respectively. (In 64-bit mode, the parameters would be transferred in registers rather than on the stack). `ecx` now contains the address of the first element of the array `a` and `edx` contains the address of the variable that `r` points to. A reference is the same as a pointer in assembly code. Register `ebx` is pushed on the stack before it is used and popped from the stack before the function returns. This is because the register usage convention says that a function is not allowed to change the value of `ebx`. Only the registers `eax`, `ecx` and `edx` can be changed freely. The loop counter `i` is stored as a register variable in `eax`. The loop initialisation `i=0;` has been translated to the instruction `xor eax,eax`. This is a common way of setting a register to zero that is more efficient than `mov eax,0`. The loop body begins at the label `$B1$2:`. This is just an arbitrary name that the compiler has chosen for the label. It uses `ebx` as a temporary register for computing `i/2+r`. The instructions `mov ebx,eax` / `shr ebx,31` copies the sign bit of `i` into the least significant bit of `ebx`. The next two instructions `add ebx, eax` / `sar ebx,1` adds this to `i` and shifts one place to the right in order to divide `i` by 2. The instruction `add ebx, DWORD PTR [edx]` adds, not `edx` but the variable whose address is in `edx`, to `ebx`. The square brackets mean use the value in `edx` as a memory pointer. This is the variable that `r` points to. Now `ebx` contains `i/2+r`. The next instruction `mov DWORD PTR [ecx+eax*4],ebx` stores this result in `a[i]`. Note how efficient the calculation of the array address is. `ecx` contains the address of the beginning of the array. `eax` holds the index, `i`. This index must be multiplied by the size (in bytes) of each array element in order to calculate the address of element number `i`. The size of an `int` is 4. So the address of array element `a[i]` is `ecx+eax*4`. The result `ebx` is then stored at address `[ecx+eax*4]`. This is all done in a single instruction. The CPU supports this kind of instructions for fast access to array elements. The instruction `add eax,1` is the loop increment `i++`. `cmp eax, 100` / `jl $B1$2` is the loop condition `i < 100`. It compares `eax` with `100` and jumps back to the `$B1$2` label if `i < 100`. `pop ebx` restores the value of `ebx` that was saved in the beginning. `ret` returns from the function.

The assembly listing reveals three things that can be optimized further. The first thing we notice is that it does some funny things with the sign bit of `i` in order to divide `i` by 2. The

compiler has not noticed that `i` can never be negative so that we don't have to care about the sign bit. We can tell it this by making `i` an `unsigned int` or by type-casting `i` to `unsigned int` before dividing by `2` (See page 143).

The second thing we notice is that the value pointed to by `r` is re-loaded from memory a hundred times. This is because we forgot to tell the compiler to assume no pointer aliasing (see page 77). Adding the compiler option "assume no pointer aliasing" (if valid) can possibly improve the code.

The third thing that can be improved is that `r+i/2` could be calculated by an induction variable because it is a staircase function of the loop index. The integer division prevents the compiler from making an induction variable unless the loop is rolled out by 2. (See page 71).

The conclusion is that we can help the compiler optimize example 8.26a by rolling out the loop by two and making an explicit induction variable. (This eliminates the need for the first two suggested improvements).

```
// Example 8.26b
void Func(int a[], int & r) {
   int i;
   int Induction = r;
   for (i = 0; i < 100; i += 2) {
      a[i] = Induction;
      a[i+1] = Induction;
      Induction++;
   }
}
```

The compiler generates the following assembly code from example 8.26b:

```
; Example 8.26b compiled to assembly:
ALIGN     4                              ; align by 4
PUBLIC ?Func@@YAXQAHAAH@Z                ; mangled function name
?Func@@YAXQAHAAH@Z   PROC NEAR           ; start of Func
; parameter 1: 4 + esp                   ; a
; parameter 2: 8 + esp                   ; r
$B1$1:                                   ; unused label
        mov       eax, DWORD PTR [esp+4] ; eax = address of a
        mov       edx, DWORD PTR [esp+8] ; edx = address in r
        mov       ecx, DWORD PTR [edx]   ; ecx = Induction
        lea       edx, DWORD PTR [eax+400] ; edx = point to end of a
$B2$2:                                   ; top of loop
        mov       DWORD PTR [eax], ecx   ; a[i] = Induction;
        mov       DWORD PTR [eax+4], ecx ; a[i+1] = Induction;
        add       ecx, 1                 ; Induction++;
        add       eax, 8                 ; point to a[i+2]
        cmp       edx, eax               ; compare with end of array
        ja        $B2$2                  ; jump to top of loop
$B2$3:                                   ; unused label
        ret                              ; return from Func
        ALIGN     4
; mark_end;
?Func2@@YAXQAHAAH@Z ENDP
```

This solution is clearly better. The loop body now contains only six instructions rather than nine, even though it is doing two iterations in one. The compiler has replaced `i` by a second induction variable (`eax`) which contains the address of the current array element. Rather than comparing `i` with `100` in the loop control it compares the array pointer `eax` to the address of the end of the array, which it has calculated in advance and stored in `edx`.

Furthermore, this solution is using one register less so that it doesn't have to push and pop `ebx`.

9 Optimizing memory access

9.1 Caching of code and data

A cache is a proxy for the main memory in a computer. The proxy is smaller and closer to the CPU than the main memory and therefore it is accessed much faster. There may be two or three levels of cache for the sake of fastest possible access to the most used data.

The speed of CPUs is increasing faster than the speed of RAM memory. Efficient caching is therefore becoming more and more important.

9.2 Cache organization

It is useful to know how a cache is organized if you are making programs that have big data structures with non-sequential access and you want to prevent cache contention. You may skip this section if you are satisfied with more heuristic guidelines.

Most caches are organized into lines and sets. Let me explain this with an example. My example is a cache of 8 kb size with a line size of 64 bytes. Each line covers 64 consecutive bytes of memory. One kilobyte is 1024 bytes, so we can calculate that the number of lines is 8*1024/64 = 128. These lines are organized as 32 sets $\times$ 4 ways. This means that a particular memory address cannot be loaded into an arbitrary cache line. Only one of the 32 sets can be used, but any of the 4 lines in the set can be used. We can calculate which set of cache lines to use for a particular memory address by the formula: (set) = ($memory$ $address$) / ($line$ $size$) % ($number$ of $sets$). Here, / means integer division with truncation, and % means modulo. For example, if we want to read from memory address a = 10000, then we have (set) = (10000 / 64) % 32 = 28. This means that a must be read into one of the four cache lines in set number 28. The calculation becomes easier if we use hexadecimal numbers because all the numbers are powers of 2. Using hexadecimal numbers, we have a = 0x2710 and (set) = (0x2710 / 0x40) % 0x20 = 0x1C. Reading or writing a variable from address 0x2710 will cause the cache to load the entire 64 or 0x40 bytes from address 0x2700 to 0x273F into one of the four cache lines from set 0x1C. If the program afterwards reads or writes to any other address in this range then the value is already in the cache so we don't have to wait for another memory access.

Assume that a program reads from address 0x2710 and later reads from addresses 0x2F00, 0x3700, 0x3F00 and 0x4700. These addresses all belong to set number 0x1C. There are only four cache lines in each set. If the cache always chooses the least recently used cache line then the line that covered the address range from 0x2700 to 0x273F will be evicted when we read from 0x4700. Reading again from address 0x2710 will cause a cache miss. But if the program had read from different addresses with different set values then the line containing the address range from 0x2700 to 0x273F would still be in the cache. The problem only occurs because the addresses are spaced a multiple of 0x800 apart. I will call this distance the critical stride. Variables whose distance in memory is a multiple of the critical stride will contend for the same cache lines. The critical stride can be calculated as ($critical$ $stride$) = ($number$ of $sets$) $\times$ ($line$ $size$) = ($total$ $cache$ $size$) / ($number$ of $ways$).

If a program contains many variables and objects that are scattered around in memory then there is a risk that several variables happen to be spaced by a multiple of the critical stride and cause contentions in the data cache. The same can happen in the code cache if there are many functions scattered around in program memory. If several functions that are used in the same part of the program happen to be spaced by a multiple of the critical stride then

this can cause contentions in the code cache. The subsequent sections describe various ways to avoid these problems.

More details about how caches work can be found in Wikipedia under CPU cache (en.wikipedia.org/wiki/L2_cache).

The details of cache organization for different processors are covered in manual 3: "The microarchitecture of Intel, AMD and VIA CPUs".

9.3 Functions that are used together should be stored together

The code cache works most efficiently if functions that are used near each other are also stored near each other in the code memory. The functions are usually stored in the order in which they appear in the source code. It is therefore a good idea to collect the functions that are used in the most critical part of the code together near each other in the same source file. Keep often used functions separate from seldom used functions, and put seldom used branches such as error handling in the end of a function or in a separate function.

Sometimes, functions are kept in different source files for the sake of modularity. For example, it may be convenient to have the member functions of a parent class in one source file and the derived class in another source file. If the member functions of parent class and derived class are called from the same critical part of the program then it can be advantageous to keep the two modules contiguous in program memory. This can be done by controlling the order in which the modules are linked together. The link order is usually the order in which the modules appear in the project window or makefile. You can check the order of functions in memory by requesting a map file from the linker. The map file tells the address of each function relative to the beginning of the program. The map file includes the addresses of library functions linked from static libraries (`.lib` or `.a`), but not dynamic libraries (`.dll` or `.so`). There is no easy way to control the addresses of dynamically linked library functions.

9.4 Variables that are used together should be stored together

Cache misses are very expensive. A variable can be fetched from the cache in just a few clock cycles, but it can take more than a hundred clock cycles to fetch the variable from RAM memory if it is not in the cache.

The cache works most efficiently if pieces of data that are used together are stored near each other in memory. Variables and objects should preferably be declared in the function in which they are used. Such variables and objects will be stored on the stack, which is very likely to be in the level-1 cache. The different kinds of variable storage are explained on page 27. Avoid global and static variables if possible, and avoid dynamic memory allocation (`new` and `delete`).

Object oriented programming can be an efficient way of keeping data together. Data members of a class (also called properties) are always stored together in an object of the class. Data members of a parent class and a derived class are stored together in an object of the derived class (see page 51).

The order in which data are stored can be important if you have big data structures. For example, if a program has two arrays, `a` and `b`, and the elements are accessed in the order `a[0]`, `b[0]`, `a[1]`, `b[1]`, ... then you may improve the performance by organizing the data as an array of structures:

```
// Example 9.1a
int Func(int);
const int size = 1024;
```

```
    int a[size], b[size], i;
    ...
    for (i = 0; i < size; i++) {
        b[i] = Func(a[i]);
    }
```

The data in this example can be accessed sequentially in memory if organized as follows:

```
    // Example 9.1b
    int Func(int);
    const int size = 1024;
    struct Sab {int a; int b;};
    Sab ab[size];
    int i;
    ...
    for (i = 0; i < size; i++) {
        ab[i].b = Func(ab[i].a);
    }
```

There will be no extra overhead in the program code for making the structure in example 9.1b. On the contrary, the code becomes simpler because it needs only calculate element addresses for one array rather than two.

Some compilers will use different memory spaces for different arrays even if they are never used at the same time. Example:

```
    // Example 9.2a
    void F1(int   x[]);
    void F2(float x[]);

    void F3(bool y) {
        if (y) {
            int a[1000];
            F1(a);
        }
        else {
            float b[1000];
            F2(b);
        }
    }
```

Here it is possible to use the same memory area for a and b because their live ranges do not overlap. You can save a lot of cache space by joining a and b in a union:

```
    // Example 9.2b
    void F3(bool y) {
        union {
            int   a[1000];
            float b[1000];
        };
        if (y) {
            F1(a);
        }
        else {
            F2(b);
        }
    }
```

Using a union is not a safe programming practice, of course, because you will get no warning from the compiler if the uses of a and b overlap. You should use this method only for big objects that take a lot of cache space. Putting simple variables into a union is not optimal because it prevents the use of register variables.

9.5 Alignment of data

A variable is accessed most efficiently if it is stored at a memory address which is divisible by the size of the variable. For example, a `double` takes 8 bytes of storage space. It should therefore preferably be stored at an address divisible by 8. The size should always be a power of 2. Objects bigger than 16 bytes should be stored at an address divisible by 16. You can generally assume that the compiler takes care of this alignment automatically.

The alignment of structure and class members may cause a waste of cache space, as explained in example 7.34 page 52.

You may choose to align large objects and arrays by the cache line size, which is typically 64 bytes. This makes sure that the beginning of the object or array coincides with the beginning of a cache line. Some compilers will align large static arrays automatically but you may as well specify the alignment explicitly by writing:

```
__declspec(align(64)) int BigArray[1024]; // Windows syntax
```

or

```
int BigArray[1024] __attribute__((aligned(64))); // Linux syntax
```

See page 95 and 124 for discussion of aligning dynamically allocated memory.

9.6 Dynamic memory allocation

Objects and arrays can be allocated dynamically with `new` and `delete`, or `malloc` and `free`. This can be useful when the amount of memory required is not known at compile time. Four typical uses of dynamic memory allocation can be mentioned here:

- A large array can be allocated dynamically when the size of the array is not known at compile time.

- A variable number of objects can be allocated dynamically when the total number of objects is not known at compile time.

- Text strings and similar objects of variable size can be allocated dynamically.

- Arrays that are too large for the stack can be allocated dynamically.

The advantages of dynamic memory allocation are:

- Gives a more clear program structure in some cases.

- Does not allocate more space than needed. This makes data caching more efficient than when a fixed-size array is made very big in order to cover the worst case situation of the maximum possible memory requirement.

- Useful when no reasonable upper limit to the required amount of memory space can be given in advance.

The disadvantages of dynamic memory allocation are:

- The process of dynamic allocation and deallocation of memory takes much more time than other kinds of storage. See page 27.

- The heap space becomes fragmented when objects of different sizes are allocated and deallocated in random order. This makes data caching inefficient.

- An allocated array may need to be resized in the event that it becomes full. This may require that a new bigger memory block is allocated and the entire contents copied to the new block. Any pointers to data in the old block then become invalid.

- The heap manager will start garbage collection when the heap space has become too fragmented. This garbage collection may start at unpredictable times and cause delays in the program flow at inconvenient times when a user is waiting for response.

- It is the responsibility of the programmer to make sure that everything that has been allocated is also deallocated. Failure to do so will cause the heap to be filled up. This is a common programming error known as memory leaks.

- It is the responsibility of the programmer to make sure that no object is accessed after it has been deallocated. Failure to do so is also a common programming error.

- The allocated memory may not be optimally aligned. See page 124 for how to align dynamically allocated memory.

- A matrix or multidimensional array is less efficient when the row length is not known at compile time because of the extra work needed for calculating row addresses at each access. The compiler may not be able to optimize this with induction variables.

It is important to weigh the advantages over the disadvantages when deciding whether to use dynamic memory allocation. There is no reason to use dynamic memory allocation when the size of an array or the number of objects is known at compile time or a reasonable upper limit can be defined.

The cost of dynamic memory allocation is negligible when the number of allocations is limited. Dynamic memory allocation can therefore be advantageous when a program has one or a few arrays of variable size. The alternative solution of making the arrays very big to cover the worst case situation is a waste of cache space. A situation where a program has several large arrays and where the size of each array is a multiple of the critical stride (see above, page 87) is likely to cause contentions in the data cache.

If the number of elements in an array grows during program execution then it is preferable to allocate the final array size right from the beginning rather than allocating more space step by step. In most systems, you cannot increase the size of a memory block that has already been allocated. If the final size cannot be predicted or if the prediction turns out to be too small, then it is necessary to allocate a new bigger memory block and copy the contents of the old memory block into the beginning of the new bigger memory block. This is inefficient, of course, and causes the heap space to become fragmented. An alternative is to keep multiple memory blocks, either in the form of a linked list or with an index of memory blocks. A method with multiple memory blocks makes the access to individual array elements more complicated and time consuming.

A collection of a variable number of objects is often implemented as a linked list. Each element in a linked list has its own memory block and a pointer to the next block. A linked list is less efficient than a linear array for the following reasons:

- Each object is allocated separately. The allocation, deallocation and garbage collection takes a considerable amount of time.

- The objects are not stored contiguously in the memory. This makes data caching less efficient.

- Extra memory space is used for the link pointers and for information stored by the heap manager for each allocated block.

- Walking through a linked list takes more time than looping through a linear array. No link pointer can be loaded until the previous link pointer has been loaded. This makes a critical dependency chain which prevents out-of-order execution.

It is often more efficient to allocate one big block of memory for all the objects (memory pooling) than to allocate a small block for each object.

A little-known alternative to using `new` and `delete` is to allocate variable-size arrays with `alloca`. This is a function that allocates memory on the stack rather than the heap. The space is automatically deallocated when returning from the function in which `alloca` was called. There is no need to deallocate the space explicitly when `alloca` is used. The advantages of `alloca` over `new` and `delete` or `malloc` and `free` are:

- There is very little overhead to the allocation process because the microprocessor has hardware support for the stack.

- The memory space never becomes fragmented thanks to the first-in-last-out nature of the stack.

- Deallocation has no cost because it goes automatically when the function returns. There is no need for garbage collection.

- The allocated memory is contiguous with other objects on the stack, which makes data caching very efficient.

The following example shows how to make a variable-size array with `alloca`:

```
// Example 9.3
#include <malloc.h>

void SomeFunction (int n) {
   if (n > 0) {
      // Make dynamic array of n floats:
      float * DynamicArray = (float *)alloca(n * sizeof(float));
      // (Some compilers use the name _alloca)
      for (int i = 0; i < n; i++) {
         DynamicArray[i] = WhateverFunction(i);
         // ...
      }
   }
}
```

Obviously, a function should never return any pointer or reference to anything it has allocated with `alloca`, because it is deallocated when the function returns. `alloca` may not be compatible with structured exception handling. See the manual for your compiler for restrictions on using `alloca`.

9.7 Container classes

Whenever dynamic memory allocation is used, it is recommended to wrap the allocated memory into a container class. The container class must have a destructor to make sure

everything that is allocated is also de-allocated. This is the best way to prevent memory leaks and other common programming errors associated with dynamic memory allocation.

Container classes can also be convenient for adding bounds-checking to an array and for more advanced data structures with First-In-First-Out or First-In-Last-Out access, sort and search facilities, binary trees, hash maps etc.

It is common to make container classes in the form of templates where the type of objects they contain is provided as a template parameter. There is no performance cost to using templates.

Ready made container class templates are available for many different purposes. The most commonly used set of containers is the Standard Template Library (STL) which comes with most modern C++ compilers. The advantage of using ready made containers is that you don't have to reinvent the wheel. The containers in the STL are universal, flexible, well tested, and very useful for many different purposes.

However, the STL is designed for generality and flexibility, while execution speed, memory economy, cache efficiency and code size have got low priority. Especially the memory allocation is unnecessarily wasteful in the STL. Some STL templates, such as `list`, `set` and `map` are prone to even allocate more memory blocks than there are objects in the container. STL `deque` (doubly ended queue) allocates one memory block for every four objects. STL `vector` stores all the objects in the same memory block, but this memory block is re-allocated every time it is filled up, which happens quite often because the block size grows by only 50% or less each time. An experiment where 10 elements were inserted, one by one, into an STL `vector` turned up to cause seven memory allocations of sizes 1, 2, 3, 4, 6, 9 and 13 objects, respectively (MS Visual Studio 2008 version). This wasteful behavior can be prevented by calling `vector::reserve` with a prediction or estimate of the final size needed before adding the first object to the `vector`. The other STL containers do not have such a feature for reserving memory in advance.

The frequent allocation and de-allocation of memory with `new` and `delete` (or `malloc` and `free`) causes the memory to become fragmented and caching becomes inefficient. There is a large overhead cost to memory management and garbage collection, as mentioned above.

The generality of the STL also costs in terms of code size. In fact, the STL has been criticized for code bloat and complexity (en.wikipedia.org/wiki/Standard_Template_Library). The objects stored in an STL container are allowed to have constructors and destructors. The copy constructors and destructors of each object are called every time an object is moved, which may happen quite often. This is necessary if the objects stored are containers themselves. But implementing a matrix in STL as a `vector` of `vector`s, as is often seen, is certainly not the most efficient solution.

Many containers use linked lists. A linked list is a convenient way of making the container expandable, but it is very inefficient. Linear arrays are faster than linked lists in most cases.

The so-called iterators that are used in STL for accessing container elements are cumbersome to use for many programmers and they are not necessary if you can use a linear list with a simple index. A good compiler can optimize away the extra overhead of the iterator in some cases, but not all.

Fortunately, there are more efficient alternatives that can be used where execution speed, memory economy and small code size has higher priority than code generality. The most important remedy is memory pooling. It is more efficient to store many objects together in one big memory block than to store each object in its own allocated memory block. A large

block containing many objects can be copied or moved with a single call to `memcpy` rather than moving each object separately if there are no copy constructors and destructors to call.

I have implemented a collection of example container classes that use these methods to improve efficiency. These are available as an appendix to this manual at www.agner.org/optimize/cppexamples.zip containing container classes and templates for several different purposes. All these examples are optimized for execution speed and for minimizing memory fragmentation. Bounds checking is included for the sake of security, but may be removed after debugging if required for performance reasons. Use these example containers in cases where the performance of the STL is not satisfactory.

The following considerations should be taken into account when choosing a container for a specific purpose:

- Contain one or multiple elements? If the container is to hold only one element then use a smart pointer (see page 38).

- Is the size known at compile time? If the number of elements is known at compile time or a not-too-big upper limit can be set then the optimal solution is a fixed size array or container without dynamic memory allocation. Dynamic memory allocation may be needed, however, if the array or container is too big for the stack.

- Is the size known before the first element is stored? If the total number of elements to store is known before the first element is stored (or if a reasonable estimate can be made) then it is preferred to use a container that allows you to reserve the amount of memory needed in advance rather than allocating piecewise or re-allocating when a memory block turns out to be too small.

- Are objects numbered consecutively? If objects are identified by consecutive indices or by keys within a limited range then a simple array is the most efficient solution.

- Is a multidimensional structure needed? A matrix or multidimensional array should be stored in one contiguous memory block. Do not use one container for each row or column. The access is faster if the number of elements per row is a constant known at compile time.

- Are objects accessed in a FIFO manner? If objects are accessed on a First-In-First-Out (FIFO) basis then use a queue. It is more efficient to implement a queue as a circular buffer than as a linked list.

- Are objects accessed in a FILO manner? If objects are accessed on a First-In-Last-Out (FILO) basis then use a linear array with a top-of-stack index.

- Are objects identified by a key? If the key values are confined to a narrow range then a simple array can be used. If the number of objects is high then the most efficient solution may be a binary tree or a hash map.

- Do objects have a natural ordering? If you need to do searches of the kind: "what is the nearest element to x?" or "how many elements are there between x and y?" then you may use a sorted list or a binary tree.

- Is searching needed after all objects have been added? If search facilities are needed, but only after all objects have been stored in the container, then a linear array will be an efficient solution. Sort the array after all elements have been added and then use binary search for finding elements. A hash map may also be an efficient solution.

- <u>Is searching needed before all objects have been added?</u> If search facilities are needed, and new objects can be added at any time, then the solution is more complicated. If the total number of elements is small then a sorted list is the most efficient solution because of its simplicity. But a sorted list can be very inefficient if the list is large because the insertion of a new element in the list causes all subsequent elements in the sequence to be moved. A binary tree or a hash map is needed in this case. A binary tree may be used if elements have a natural order and there are search requests for elements in a specific interval. A hash map can be used if elements have no specific order but are identified by a unique key.

- <u>Do objects have mixed types or sizes?</u> It is possible to store objects of different types or strings of different lengths in the same memory pool. See www.agner.org/optimize/cppexamples.zip. If the number and types of elements is known at compile time then there is no need to use a container or memory pool.

- <u>Alignment?</u> Some applications require that data are aligned at round addresses. Especially the use of intrinsic vectors requires alignment to addresses divisible by 16. Alignment of data structures to addresses divisible by the cache line size (typically 64) can improve performance in some cases.

- <u>Multiple threads?</u> Container classes are generally not thread safe if multiple threads can add, remove or modify objects simultaneously. In multithreaded applications it is much more efficient to have separate containers for each thread than to temporarily lock a container for exclusive access by each thread.

- <u>Pointers to contained objects?</u> It may not be safe to make a pointer to a contained object because the container may move the object in case memory re-allocation is needed. Objects inside containers should be identified by their index or key in the container rather than by pointers or references. It is OK, however, to pass a pointer or reference to such an object to a function that doesn't add or remove any objects if no other threads have access to the container.

- <u>Can the container be recycled?</u> There is a large cost to creating and deleting containers. If the program logic allows it, it may be more efficient to re-use a container than to delete it and create a new one.

I have provided several examples of suitable containers class templates in www.agner.org/optimize/cppexamples.zip. These may be used as alternatives to the standard template library (STL) if the full generality and flexibility of the STL containers is not needed. You may write your own container classes or modify the ones that are available to fit specific needs.

9.8 Strings

Text strings typically have variable lengths that are not known at compile time. The storage of text strings in classes like `string`, `wstring` or `CString` uses `new` and `delete` to allocate a new memory block every time a string is created or modified. This can be quite inefficient if a program creates or modifies many strings.

In most cases, the fastest way to handle strings is the old fashioned C style with character arrays. Strings can be manipulated with C functions such as `strcpy`, `strcat`, `strlen`, `sprintf`, etc. But beware that these functions have no check for overflow of the arrays. An array overflow can cause unpredictable errors elsewhere in the program that are very difficult to diagnose. It is the responsibility of the programmer to make sure the arrays are sufficiently large to handle the strings including the terminating zero and to make overflow checks where necessary.

If you want to improve speed without jeopardizing safety, you may store all strings in a memory pool, as explained above. Examples are provided in an appendix to this manual at www.agner.org/optimize/cppexamples.zip.

9.9 Access data sequentially

A cache works most efficiently when the data are accessed sequentially. It works somewhat less efficiently when data are accessed backwards and much less efficiently when data are accessed in a random manner. This applies to reading as well as writing data.

Multidimensional arrays should be accessed with the last index changing in the innermost loop. This reflects the order in which the elements are stored in memory. Example:

```
// Example 9.4
const int NUMROWS = 100, NUMCOLUMNS = 100;
int matrix[NUMROWS][NUMCOLUMNS];
int row, column;
for (row = 0; row < NUMROWS; row++)
   for (column = 0; column < NUMCOLUMNS; column++)
      matrix[row][column] = row + column;
```

Do not swap the order of the two loops (except in Fortran where the storage order is opposite).

9.10 Cache contentions in large data structures

It is not always possible to access a multidimensional array sequentially. Some applications (e.g. in linear algebra) require other access patterns. This can cause severe delays if the distance between rows in a big matrix happen to be equal to the critical stride, as explained on page 87. This will happen if the size of a matrix line (in bytes) is a high power of 2.

The following example illustrates this. My example is a function which transposes a quadratic matrix, i.e. each element matrix[r][c] is swapped with element matrix[c][r].

```
// Example 9.5a
const int SIZE = 64;              // number of rows/columns in matrix

void transpose(double a[SIZE][SIZE]) { // function to transpose matrix
   // define a macro to swap two array elements:
   #define swapd(x,y) {temp=x; x=y; y=temp;}

   int r, c;  double temp;
   for (r = 1; r < SIZE; r++) {        // loop through rows
      for (c = 0; c < r; c++) {        // loop columns below diagonal
         swapd(a[r][c], a[c][r]);      // swap elements
      }
   }
}

void test () {
   __declspec(__align(64))        // align by cache line size
   double matrix[SIZE][SIZE];     // define matrix
   transpose(matrix);             // call transpose function
}
```

Transposing a matrix is the same as reflecting it at the diagonal. Each element matrix[r][c] below the diagonal is swapped with element matrix[c][r] at its mirror position above the diagonal. The c loop in example 9.5a goes from the leftmost column to the diagonal. The elements at the diagonal remain unchanged.

The problem with this code is that if the elements `matrix[r][c]` below the diagonal are accessed row-wise, then the mirror elements `matrix[c][r]` above the diagonal are accessed column-wise.

Assume now that we are running this code with a 64×64 matrix on a Pentium 4 computer where the level-1 data cache is 8 kb = 8192 bytes, 4 ways, with a line size of 64. Each cache line can hold 8 `double`'s of 8 bytes each. The critical stride is 8192 / 4 = 2048 bytes = 4 rows.

Let's look at what happens inside the loop, for example when `r` = 28. We take the elements from row 28 below the diagonal and swap these elements with column 28 above the diagonal. The first eight elements in row 28 share the same cache line. But these eight elements will go into eight different cache lines in column 28 because the cache lines follow the rows, not the columns. Every fourth of these cache lines belong to the same set in the cache. When we reach element number 16 in column 28, the cache will evict the cache line that was used by element 0 in this column. Number 17 will evict number 1. Number 18 will evict number 2, etc. This means that all the cache lines we used above the diagonal have been lost at the time we are swapping column 29 with line 29. Each cache line has to be reloaded eight times because it is evicted before we need the next element. I have confirmed this by measuring the time it takes to transpose a matrix using example 9.5a on a Pentium 4 with different matrix sizes. The results of my experiment are given below. The time unit is clock cycles per array element.

Matrix size	Total kilobytes	Time per element
63×63	31	11.6
64×64	32	16.4
65×65	33	11.8
127×127	126	12.2
128×128	128	17.4
129×129	130	14.4
511×511	2040	38.7
512×512	2048	230.7
513×513	2056	38.1

Table 9.1. Time for transposition of different size matrices, clock cycles per element.

The table shows that it takes 40% more time to transpose the matrix when the size of the matrix is a multiple of the level-1 cache size. This is because the critical stride is a multiple of the size of a matrix line. The delay is less than the time it takes to reload the level-1 cache from the level-2 cache because the out-of-order execution mechanism can prefetch the data.

The effect is much more dramatic when contentions occur in the level-2 cache. The level-2 cache is 512 kb, 8 ways. The critical stride for the level-2 cache is 512 kb / 8 = 64 kb. This corresponds to 16 lines in a 512×512 matrix. My experimental results in table 9.1 show that it takes six times as long time to transpose a matrix when contentions occur in the level-2 cache as when contentions do not occur. The reason why this effect is so much stronger for level-2 cache contentions than for level-1 cache contentions is that the level-2 cache cannot prefetch more than one line at a time.

A simple way of solving the problem is to make the rows in the matrix longer than needed in order to avoid that the critical stride is a multiple of the matrix line size. I tried to make the matrix 512×520 and leave the last 8 columns unused. This removed the contentions and the time consumption was down to 36.

There may be cases where it is not possible to add unused columns to a matrix. For example, a library of math functions should work efficiently on all sizes of matrices. An efficient solution in this case is to divide the matrix into smaller squares and handle one square at a time. This is called square blocking or tiling. This technique is illustrated in example 9.5b.

```
// Example 9.5b
void transpose(double a[SIZE][SIZE]) {
   // Define macro to swap two elements:
   #define swapd(x,y) {temp=x; x=y; y=temp;}
   // Check if level-2 cache contentions will occur:
   if (SIZE > 256 && SIZE % 128 == 0) {
      // Cache contentions expected. Use square blocking:
      int r1, r2, c1, c2; double temp;
      // Define size of squares:
      const int TILESIZE = 8;   // SIZE must be divisible by TILESIZE
      // Loop r1 and c1 for all squares:
      for (r1 = 0; r1 < SIZE; r1 += TILESIZE) {
         for (c1 = 0; c1 < r1; c1 += TILESIZE) {
            // Loop r2 and c2 for elements inside sqaure:
            for (r2 = r1; r2 < r1+TILESIZE; r2++) {
               for (c2 = c1; c2 < c1+TILESIZE; c2++) {
                  swapd(a[r2][c2],a[c2][r2]);
               }
            }
         }
         // At the diagonal there is only half a square.
         // This triangle is handled separately:
         for (r2 = r1+1; r2 < r1+TILESIZE; r2++) {
            for (c2 = r1; c2 < r2; c2++) {
               swapd(a[r2][c2],a[c2][r2]);
            }
         }
      }
   }
   else {
      // No cache contentions. Use simple method.
      // This is the code from example 9.5a:
      int r, c;  double temp;
      for (r = 1; r < SIZE; r++) {    // loop through rows
         for (c = 0; c < r; c++) {    // loop columns below diagonal
            swapd(a[r][c], a[c][r]);  // swap elements
         }
      }
   }
}
```

This code took 50 clock cycles per element for a 512×512 matrix in my experiments.

Contentions in the level-2 cache are so expensive that it is very important to do something about them. You should therefore be aware of situations where the number of columns in a matrix is a high power of 2. Contentions in the level-1 cache are less expensive. Using complicated techniques like square blocking for the level-1 cache may not be worth the effort.

Square blocking and similar methods are further described in the book "Performance Optimization of Numerically Intensive Codes", by S. Goedecker and A. Hoisie, SIAM 2001.

9.11 Explicit cache control

Microprocessors with the SSE and SSE2 instruction sets have certain instructions that allow you to manipulate the data cache. These instructions are accessible from compilers that have support for intrinsic functions (i.e. Microsoft, Intel and Gnu). Other compilers need assembly code to access these instructions.

Function	Assembly name	Intrinsic function name	Instruction set
Prefetch	PREFETCH	_mm_prefetch	SSE
Store 4 bytes without cache	MOVNTI	_mm_stream_si32	SSE2
Store 8 bytes without cache	MOVNTQ	_mm_stream_pi	SSE
Store 16 bytes without cache	MOVNTPS	_mm_stream_ps	SSE
Store 16 bytes without cache	MOVNTPD	_mm_stream_pd	SSE2
Store 16 bytes without cache	MOVNTDQ	_mm_stream_si128	SSE2

Table 9.2. Cache control instructions.

There are other cache control instructions than the ones mentioned in table 9.2, such as flush and fence instructions, but these are hardly relevant to optimization.

Prefetching data

The prefetch instruction can be used for fetching a cache line that we expect to use later in the program flow. However, this did not improve the execution speed in any of the examples I have tested on a Pentium 4 processor. The reason is that modern processors prefetch data automatically thanks to out-of-order execution and advanced prediction mechanisms. It can be expected that future microprocessors will be able to automatically prefetch data for regular access patterns containing one or two streams with different strides. Therefore, you don't have to prefetch data explicitly if data access can be arranged in regular patterns with fixed strides.

Uncached memory store

An uncached write is more expensive than an uncached read because the write causes an entire cache line to be read and written back.

The so-called nontemporal write instructions (MOVNT) are designed to solve this problem. These instructions write directly to memory without loading a cache line. This is advantageous in cases where we are writing to uncached memory and we do not expect to read from the same or a nearby address again before the cache line would be evicted. Don't mix nontemporal writes with normal writes or reads to the same memory area.

The nontemporal write instructions are not suitable for example 9.5 because we are reading and writing from the same address so a cache line will be loaded anyway. If we modify example 9.5 so that it writes only, then the effect of nontemporal write instructions becomes noticeable. The following example transposes a matrix and stores the result in a different array.

```
// Example 9.6a
const int SIZE = 512; // number of rows and columns in matrix

// function to transpose and copy matrix
void TransposeCopy(double a[SIZE][SIZE], double b[SIZE][SIZE]) {
   int r, c;
   for (r = 0; r < SIZE; r++) {
      for (c = 0; c < SIZE; c++) {
         a[c][r] = b[r][c];
      }
   }
```

```
    }
```

This function writes to matrix `a` in a column-wise manner where the critical stride causes all writes to load a new cache line in both the level-1 and the level-2 cache. Using the nontemporal write instruction prevents the level-2 cache from loading any cache lines for matrix `a`:

```
// Example 9.6b.
#include "xmmintrin.h"  // header for intrinsic functions

// This function stores a double without loading a cache line:
static inline void StoreNTD(double * dest, double const & source) {
    _mm_stream_pi((__m64*)dest, *(__m64*)&source);  // MOVNTQ
    _mm_empty();                                     // EMMS
}

const int SIZE = 512; // number of rows and columns in matrix
// function to transpose and copy matrix
void TransposeCopy(double a[SIZE][SIZE], double b[SIZE][SIZE]) {
    int r, c;
    for (r = 0; r < SIZE; r++) {
        for (c = 0; c < SIZE; c++) {
            StoreNTD(&a[c][r], b[r][c]);
        }
    }
}
```

The execution times per matrix cell for different matrix sizes were measured on a Pentium 4 computer. The measured results were as follows:

Matrix size	Time per element Example 9.6a	Time per element Example 9.6b
64×64	14.0	80.8
65×65	13.6	80.9
512×512	378.7	168.5
513×513	58.7	168.3

Table 9.3. Time for transposing and copying different size matrices, clock cycles per element.

As table 9.3 shows, the method of storing data without caching is advantageous if, and only if, a level-2 cache miss can be expected. The 64×64 matrix size causes misses in the level-1 cache. This has hardly any effect on the total execution time because the cache miss on a store operation doesn't delay the subsequent instructions. The 512×512 matrix size causes misses in the level-2 cache. This has a very dramatic effect on the execution time because the memory bus is saturated. This can be ameliorated by using nontemporal writes. If the cache contentions can be prevented in other ways, as explained in chapter 9.10, then the nontemporal write instructions are not optimal.

There are certain restrictions on using the instructions listed in table 9.2. All these instructions require that the microprocessor has the SSE or SSE2 instruction set, as listed in the table. The 16-byte instructions `MOVNTPS`, `MOVNTPD` and `MOVNTDQ` require that the operating system has support for XMM registers; see page 125.

The Intel compiler can insert nontemporal writes automatically in vectorized code when the `#pragma vector nontemporal` is used. However, this does not work in example 9.6b.

The `MOVNTQ` instruction must be followed by an `EMMS` instruction before any floating point instructions. This is coded as `_mm_empty()` as shown in example 9.6b. The `MOVNTQ` instruction cannot be used in 64-bit device drivers for Windows.

10 Multithreading

The clock frequency of the CPU is limited by physical factors. The way to increase the throughput of CPU-intensive programs when the clock frequency is limited is to do multiple things at the same time. There are three ways to do things in parallel:

- Using multiple CPUs or multi-core CPUs, as described in this chapter.

- Using the out-of-order capabilities of modern CPUs, as described in chapter 11.

- Using the vector operations of modern CPUs, as described in chapter 12.

It is important to distinguish between coarse-grained parallelism and fine-grained parallelism when deciding whether it is advantageous to do things in parallel. Coarse-grained parallelism refers to the situation where a long sequence of operations can be carried out independently of other tasks that are running in parallel. Fine-grained parallelism is the situation where a task can be divided into many small subtasks, but it is not possible to work for very long on a particular subtask before coordination with other subtasks is necessary.

Most modern CPUs have two or more cores, and it can be expected that the number of cores will grow in the future. The use of multiple CPU cores is useful for coarse-grained parallelism but not for fine-grained parallelism because communication and synchronization between the different cores is slow. The methods described in chapter 11 and 12 are more useful for fine-grained parallelism.

The way to use multiple CPU cores is to divide the work into multiple threads. The use of threads is discussed on page 60. We should preferably have no more threads with the same priority than the number of cores or logical processors available in the system. The number of logical processors available can be determined by a system call (e.g. `GetProcessAffinityMask` in Windows).

There are several ways to divide the workload between multiple CPU cores:

- Define multiple threads and put an equal amount of work into each thread. This method works with all compilers.

- Use automatic parallelization. The Gnu, Intel and PathScale compilers can automatically detect opportunities for parallelization in the code and divide it into multiple threads.

- Use OpenMP directives. OpenMP is a standard for specifying parallel processing in C++ and Fortran. These directives are supported by Microsoft, Intel, PathScale and Gnu compilers. See www.openmp.org and the compiler manual for details.

- Use function libraries with internal multi-threading, e.g. Intel Math Kernel Library.

The multiple CPU cores or logical processors usually share the same cache, at least at the last cache level, and possibly even the same level-1 cache. The advantage of sharing the same cache is that communication between threads becomes faster and that threads can share the same code and read-only data. The disadvantage is that the cache will be filled

up if the threads use different memory areas, and there will be cache contentions if the threads write to the same memory areas.

A CPU with multiple cores or a system with multiple CPUs is very useful for running multiple tasks in parallel. A CPU-intensive program with possibilities for coarse-grained parallelism can benefit very much from dividing the calculations into multiple threads. The workload should preferably be divided evenly between the threads. Multithreading may not be advantageous for fine-grained parallelism because the communication and synchronization between threads consume a considerable amount of extra resources.

Data that are read only can be shared between multiple threads, while data that are modified should be separate for each thread. Avoid having two threads writing to the same cache line, because they will invalidate each other's caches and cause large delays. The easiest way to make thread-specific data is to declare it locally in the thread function so that it is stored on the stack. Each thread has its own stack. Alternatively, you may define a structure or class for containing thread-specific data and make one instance for each thread. This structure or class should be aligned by at least the cache line size in order to avoid multiple threads writing to the same cache line. The cache line size is typically 64 bytes on contemporary processors. The cache line size may possibly be more (128 or 256 bytes) on future processors.

Running multiple threads on a system with only one logical processor is not an advantage if the threads are competing for the same resources. But it can be a good idea to put time-consuming calculations into a separate thread with lower priority than the user interface. It is also useful to put file access and network access in separate threads so that one thread can do calculations while another thread is waiting for response from a hard disk or network.

Various development tools for supporting multi-threaded software are available from Intel. See Intel Technology Journal Vol. 11, Iss. 4 (www.intel.com/technology/itj/).

10.1 Hyperthreading

Some versions of Intel microprocessors are able to run two threads in each core. For example, a Core i7 processor with four cores can run eight threads simultaneously. This processor has four physical processors but eight logical processors.

Hyperthreading is Intel's term for running multiple threads in the same processor core. Two threads running in the same core will always compete for the same resources, such as cache and execution units. If any of the shared resources are limiting factors for the performance then there is no advantage to using hyperthreading. On the contrary, each thread may run at less than half speed because of cache evictions and other resource conflicts. But if a large fraction of the time goes to cache misses, branch misprediction, or long dependency chains then each thread will run at more than half the single-thread speed. In this case there is an advantage to using hyperthreading, but the performance is not doubled. A thread that shares the resources of the core with another thread will always run slower than a thread that runs alone in the core.

It is often necessary to do experiments in order to determine whether it is advantageous to use hyperthreading or not in a particular application.

If hyperthreading is not advantageous then it is necessary to query certain operating system functions (e.g. `GetLogicalProcessorInformation` in Windows) to determine if the processor has hyperthreading. If so, then you can avoid hyperthreading by using only the even-numbered logical processors (0, 2, 4, etc.). Older operating systems lack the necessary functions for distinguishing between the number of physical processors and the number of logical processors.

There is no way to tell a hyperthreading processor to give higher priority to one thread than another. Therefore, it can often happen that a low-priority thread steals resources from a higher-priority thread running in the same core. It is the responsibility of the operating system to avoid running two threads with widely different priority in the same processor core. Unfortunately, contemporary operating systems are not always avoiding this.

The Intel compiler is capable of making two threads where one thread is used for prefetching data for the other thread. However, in most cases you can rely on automatic prefetching so this feature is rarely needed.

11 Out of order execution

Modern x86 CPUs can execute instructions out of order or do more than one thing at the same time. The following example shows how to take advantage of this capability:

```
// Example 11.1a
float a, b, c, d, y;
y = a + b + c + d;
```

This expression is calculated as `((a+b)+c)+d`. This is a dependency chain where each addition has to wait for the result of the preceding one. You can improve this by writing:

```
// Example 11.1b
float a, b, c, d, y;
y = (a + b) + (c + d);
```

Now the two parentheses can be calculated independently. The CPU will start to calculate `(c+d)` before it has finished the calculation of `(a+b)`. This can save several clock cycles. You cannot assume that an optimizing compiler will change the code in example 11.1a to 11.1b automatically, although it appears to be an obvious thing to do. The reason why compilers do not make this kind of optimizations is that it may cause a loss of precision, as explained on page 72. You have to set the parentheses manually.

The effect of dependency chains is stronger when they are long. This is often the case in loops. Consider the following example, which calculates the sum of 100 numbers:

```
// Example 11.2a
const int size = 100;
float list[size], sum = 0;  int i;
for (i = 0; i < size; i++) sum += list[i];
```

This has a long dependency chain. If a floating point addition takes 5 clock cycles, then this loop will take approximately 500 clock cycles. You can improve the performance dramatically by unrolling the loop and splitting the dependency chain in two:

```
// Example 11.2b
const int size = 100;
float list[size], sum1 = 0, sum2 = 0;  int i;
for (i = 0; i < size; i += 2) {
   sum1 += list[i];
   sum2 += list[i+1];}
sum1 += sum2;
```

If the microprocessor is doing an addition to `sum1` from time T to T+5, then it can do another addition to `sum2` from time T+1 to T+6, and the whole loop will take only 256 clock cycles.

Calculations in a loop where each iteration needs the result of the preceding one is called a loop-carried dependency chain. Such dependency chains can be very long and very time-

consuming. There is a lot to gain if such dependency chains can be broken up. The two summation variables `sum1` and `sum2` are called accumulators. The optimal number of accumulators for floating point addition and multiplication is three or four. Current CPUs have only one floating point addition unit, but this unit is pipelined, as explained above, so that it can start a new addition before the preceding addition is finished.

Unrolling a loop becomes a little more complicated if the number of iterations is not divisible by the unroll factor. For example, if the number of elements in `list` in example 11.2b was an odd number then we would have to add the last element outside the loop or add an extra dummy element to `list` and make this extra element zero.

It is not necessary to unroll a loop and use multiple accumulators if there is no loop-carried dependency chain. A microprocessor with out-of-order capabilities can overlap the iterations and start the calculation of one iteration before the preceding iteration is finished. Example:

```
// Example 11.3
const int size = 100;  int i;
float a[size], b[size], c[size];
float register temp;
for (i = 0; i < size; i++) {
   temp = a[i] + b[i];
   c[i] = temp * temp;
}
```

Microprocessors with out-of-order capabilities are very smart. They can detect that the value of register `temp` in one iteration of the loop in example 11.3 is independent of the value in the previous iteration. This allows it to begin calculating a new value of `temp` before it is finished using the previous value. It does this by assigning a new physical register to `temp` even though the logical register that appears in the machine code is the same. This is called register renaming. The CPU can hold many renamed instances of the same logical register.

This advantage comes automatically. There is no reason to unroll the loop and have a `temp1` and `temp2`. All modern CPUs are capable of register renaming and doing multiple calculations in parallel if certain conditions are satisfied. The conditions that make it possible for the CPU to overlap the calculations of loop iterations are:

- No loop-carried dependency chain. Nothing in the calculation of one iteration should depend on the result of the previous iteration (except for the loop counter, which is calculated fast if it is an integer).

- All intermediate results should be saved in registers, not in memory. The renaming mechanism works only on registers, not on variables in memory or cache. Most compilers will make `temp` a register variable in example 11.3 even without the `register` keyword. The CodeGear compiler cannot make floating point register variables, but will save `temp` in memory. This prevents the CPU from overlapping calculations.

- The loop branch should be predicted. This is no problem if the repeat count is large or constant. If the loop count is small and changing then the CPU may occasionally predict that the loop exits, when in fact it does not, and therefore fail to start the next calculation. However, the out-of-order mechanism allows the CPU to increment the loop counter ahead of time so that it may detect the misprediction before it is too late. You should therefore not be too worried about this condition.

In general, the out-of-order execution mechanism works automatically. However, there are a couple of things that the programmer can do to take maximum advantage of out-of-order execution. The most important thing is to avoid long dependency chains. Another thing that you can do is to mix different kinds of operations in order to divide the work evenly between

the different execution units in the CPU. It can be advantageous to mix integer and floating point calculations as long as you don't need conversions between integers and floating point numbers. It can also be advantageous to mix floating point addition with floating point multiplication, to mix simple integer with vector integer operations, and to mix mathematical calculations with memory access.

12 Using vector operations

Today's microprocessors have vector instructions that make it possible to do operations on all elements of a vector simultaneously. This is also called Single-Instruction-Multiple-Data (SIMD) operations or MMX, XMM and YMM instructions.

Vector operations are useful when doing calculations on large data sets where the same operation is performed on multiple data elements and the program logic allows parallel calculations. Examples are image processing, sound processing, and mathematical operations on vectors and matrixes. Algorithms that are inherently serial, such as most sorting algorithms, are not suited for vector operations. Algorithms that rely heavily on table lookup or require a lot of data shuffling, such as many encryption algorithms, cannot easily be implemented as vector operations.

The vector operations use a set of special vector registers. The size of each vector register is 128 bits (XMM) if the SSE2 instruction set is available, or 256 bits (YMM) if the AVX instruction set is supported by the microprocessor and the operating system.

A 128-bit XMM register can be organized as a vector of sixteen 8-bit integers, eight 16-bit integers, four 32-bit integers, two 64-bit integers, four `float`'s or two `double`'s. For example, you can add two vectors of four `float`'s each in one operation. This is four additions in one. A 256-bit YMM register can hold eight `float`'s or four `double`'s, but not integers. The older MMX registers that are 64 bits wide may be considered obsolete.

In most cases the XMM vector operations require that the SSE2 or later instruction set is enabled in both the compiler, the CPU, and the operating system. See page 125 for how to check this. Vectors must be aligned by 16, i.e. stored at a memory address that is divisible by 16 (see below).

Vector operations are particularly fast on newer processors. These processors can calculate a vector just as fast as a scalar (Scalar means not a vector). Earlier processors are using a 64-bit execution unit twice for calculating a 128-bit vector. There is hardly any advantage in making a vector of two double precision floats on processors with 64-bit units when the time it takes to calculate a vector is double the time it takes to calculate each of the two scalars.

The use of vector operations is more advantageous the smaller the data elements are. For example, it takes the same time to add two vectors of four `float`'s and two vectors of two `double`'s. With `float`'s you get four additions in the same time that it takes to do two additions with `double`'s. It is almost always advantageous to use vector operations on contemporary CPUs if the data fit nicely into the vector registers. It may not be advantageous if a lot of data manipulation is required for putting the right data into the right vector elements.

12.1 AVX instruction set and YMM registers

The 128-bit XMM registers are extended to 256-bit registers named YMM in the AVX instruction set. The main advantage of the AVX instruction set is that it allows larger floating point vectors. There are also other advantages that may improve the performance somewhat.

Code that is compiled for the AVX instruction set can run only if supported by the operating system. AVX is supported in Windows 7 and Windows Server 2008 R2 as well as in Linux kernel version 2.6.30 and later.

The AVX instruction set is supported in the latest compilers from Microsoft, Intel and Gnu.

There is a problem when mixing code compiled with and without AVX support. There is a performance penalty when going from AVX code to non-AVX code because of a change in the YMM register state. This penalty should be avoided by calling the intrinsic function `_mm256_zeroupper()` before any transition from AVX code to non-AVX code. This can be necessary in the following cases:

- If part of a program is compiled with AVX support and another part of the program is compiled without AVX support then call `_mm256_zeroupper()` before leaving the AVX part.

- If a function is compiled in multiple versions with and without AVX using CPU dispatching then call `_mm256_zeroupper()` before leaving the AVX part.

- If a piece of code compiled with AVX support calls a function in a library other than the library that comes with the compiler, and the library has no AVX support, then call `_mm256_zeroupper()` before calling the library function.

12.2 Automatic vectorization
The best compilers can use vector operations automatically in cases where the parallelism is obvious. See the compiler documentation for detailed instructions.

Example:

```
// Example 12.1a
// Use Intel compiler f. Windows with options /arch:SSE2
// or  Intel compiler f. Linux with options -mSSE2

const int size = 100;

void AddTwo(float aa[size], float bb[size]) {
   #pragma vector aligned      // assume aa and bb are aligned by 16
   #pragma ivdep               // assume no pointer aliasing
   for (int i = 0; i < size; i++) {
      aa[i] = bb[i] + 2.0f;
   }
}

void CallAddTwo() {
   __declspec (align(16))      // arrays must be aligned by 16
   float a[size], b[size];     // define two arrays
   AddTwo(a, b);               // call AddTwo function
}
```

The vector operations require that arrays `a` and `b` be aligned by 16, i.e. stored at an address divisible by 16. Here, the `#pragma vector aligned` tells the Intel compiler that `aa` and `bb` are properly aligned. The compiler cannot assume this because the function `AddTwo` could be called with unaligned arrays as parameters. The `#pragma ivdep` tells the Intel compiler to assume that there is no pointer aliasing (see page 77). The syntax for the pragmas may be different for different compilers. An alternative way of giving the compiler this information is to use the `__restrict__` keyword to tell the compiler that there is no

pointer aliasing, and use the `__assume_aligned` directive to tell the compiler that the arrays are aligned by 16:

```
// Example 12.1b
void AddTwo(float * __restrict__ aa, float * __restrict__ bb) {
    // Tell Intel compiler that pointers are aligned:
    __assume_aligned(aa,16); __assume_aligned(bb,16);
    for (int i = 0; i < size; i++) {
        aa[i] = bb[i] + 2.0f;
    }
}
```

Example 12.1a and 12.1b are equivalent. Here, the compiler will use the vector instructions to do four additions at a time. This will increase the execution speed by a factor two or more.

The automatic vectorization works optimally if the following conditions are satisfied:

1. Use a compiler that supports automatic vectorization, such as Gnu, Intel or PathScale.

2. Use appropriate compiler options to enable the desired instruction set (/arch:SSE2, /arch:AVX etc. for Windows, -mSSE2, -mAVX, etc. for Linux)

3. Align arrays and structures by 16 for SSE2, preferably 32 for AVX.

4. If the alignment is not visible in the scope of the function where you want vectorization then it is necessary to tell the compiler that data are aligned.

5. If the arrays or structures are accessed through pointers or references then tell the compiler explicitly that pointers do not alias, if appropriate.

6. The loop count should preferably be a constant that is divisible by the number of elements in a vector.

The compiler can vectorize the code if only conditions 1 - 3 are satisfied, but the code will be more efficient if conditions 4 - 6 are also satisfied. If conditions 4 or 5 are not satisfied then the compiler will generate two versions of the code and check the data addresses at runtime. If data addresses are misaligned or overlapping then it will use the non-vectorized version of the code. If condition 6 is not satisfied then the compiler will generate extra code to take care of the possible remaining data (The number of remaining data is the remainder of the loop count divided by the number of elements in a vector). These extra runtime checks is a waste of time and the extra code is a waste of cache space. You should therefore make sure that condition 4 - 6 are satisfied if possible.

The Intel compiler tells when it has vectorized a loop. The output says "remark: LOOP WAS VECTORIZED" with an indication of the line number. The compiler does not tell if it has inserted extra runtime checks for condition 4 - 6, not even with the maximum optimization report level. You have to look at the assembly output listing to see this (see page 84).

The alignment of the data you want to vectorize is important. If you tell the compiler to assume that data are aligned when in fact they are not then the program will crash with a general protection exception.

The compiler can also use vector operations where there is no loop if the same operation is performed on a sequence of consecutive variables. Example:

```
// Example 12.2
__declspec(align(16))     // Make all instances of S1 aligned
```

```
struct S1 {                  // Structure of 4 floats
   float a, b, c, d;
};

void Func() {
   S1 x, y;
   ...
   #pragma vector always  // Tell compiler to vectorize
   x.a = y.a + 1.;
   x.b = y.b + 2.;
   x.c = y.c + 3.;
   x.d = y.d + 4.;
};
```

A structure of four `float`'s fits into a 128-bit XMM register. In example 12.2, the compiler will store the constant vector `(1.,2.,3.,4.)` in static memory. When `Func` is called, it will load the structure `y` into an XMM register, add the constant vector `(1.,2.,3.,4.)`, and store the result in `x`. The `#pragma vector always` tells the compiler to override the normal heuristics and vectorize the code regardless of whether it is judged to be advantageous or not. The compiler is not always able to predict correctly whether vectorization will be advantageous or not. You can use the `#pragma vector always` to tell the compiler to vectorize, or `#pragma novector` to tell the compiler not to vectorize. The pragmas must be placed immediately before the loop or the series of statements that you want them to apply to.

It is recommended to use the smallest data size that fits the application. Example:

```
// Example 12.3
const int size = 1000;
__declspec(align(16))          // align a and b by 16
int a[size], b[size];
...
for (int i = 0; i < size, i++) {
   a[i] = b[i] * 2 + 1;
}
```

The Intel compiler will vectorize this loop, which takes approximately 1100 clock cycles on a Pentium 4. If the values in `a` and `b` can fit into 16-bit integers rather than the default 32-bit integers without overflow, then the execution speed can be almost doubled by changing the type of `a` and `b` to `short int`. If 8-bit integers are sufficient then the speed can be increased further by changing the type of `a` and `b` to `char`. Without vectorization, the loop takes 4000 clock cycles regardless of integer size.

The vector instructions cannot multiply integers of any size other than 16 bits, unless the SSE4.1 instruction set is available. The multiplication by 2 in example 12.3 is done by adding the value to itself. It is not possible to make integer division in vectors.

12.3 Explicit vectorization

It is difficult to predict whether the compiler will vectorize a loop or not. Some loops are vectorized automatically when the elements are floating point, but not if they are integers. And some loops are vectorized in 32-bit mode but not in 64-bit mode. The following example shows a code that the Intel compiler does not vectorize automatically. The code has a branch that chooses between two expressions for every element in the arrays:

```
// Example 12.4a
const int size = 128;         // Array size
short int aa[size], bb[size], cc[size], dd[size];

void SelectAndAdd(short int g) {
```

```
        for (int i = 0; i < size; i++) {
            aa[i] = (bb[i] > 0) ? (cc[i] + 2) : (dd[i] + g);
        }
    }
```

Using intrinsic functions

It is possible to tell the compiler which vector operations to use by using the so-called intrinsic functions. This is useful in situations like example 12.4a where the compiler doesn't vectorize the code automatically. It is also useful for compilers that cannot vectorize the code automatically.

Intrinsic functions are primitive operations in the sense that each intrinsic function call is translated to just one or a few machine instructions. Intrinsic functions are supported by the Intel, Gnu, Microsoft and PathScale compilers. (The PGI compiler supports intrinsic functions, but in a very inefficient way. The Codeplay compiler has some support for intrinsic functions, but the function names are not compatible with the other compilers).

The best performance is obtained with the Gnu or the Intel compiler. Other compilers do not optimize code containing intrinsic functions as good as the Gnu and Intel compilers do.

We want to vectorize the loop in example 12.4a so that we can handle eight elements at a time in vectors of eight 16-bit integers. The branch inside the loop can be implemented in various ways depending on the available instruction set. The most compatible way is to make a bit-mask which is all 1's when bb[i] > 0 is true, and all 0's when false. The expression cc[i]+2 is AND'ed with this mask, and dd[i]+g is AND'ed with the inverted mask. The expression that is AND'ed with all 1's is unchanged, while the expression that is AND'ed with all 0's gives zero. An OR combination of these two gives the chosen expression. (Intel CPUs with the SSE4.1 instruction set can use _mm_blendv_epi8 instead of the AND/OR operations. AMD CPUs with the XOP instruction set can use the VPCMOV instruction, for which an intrinsic function name has not yet been defined).

Example 12.4b shows how this can be implemented with intrinsic functions:

```
// Example 12.4b
// SSE2 instruction set must be enabled
#include <emmintrin.h>         // Define intrinsic functions

const int size = 128;          // Array size
__declspec(align(16))          // Align arrays by 16
short int aa[size], bb[size], cc[size], dd[size];

// Macro to type-cast an array address to a vector address:
#define ToVectorAddress(x) ((__m128i*)&(x))

void SelectAndAdd(short int g) {
    // Define local vector variables:
    __m128i a, b, c, d, mask, zero, two, g_broadcast;
    // Set zero to a vector of (0,0,0,0,0,0,0,0)
    zero = _mm_set1_epi16(0);
    // Set two to a vector of (2,2,2,2,2,2,2,2)
    two  = _mm_set1_epi16(2);
    // Set g_broadcast to a vector of (g,g,g,g,g,g,g,g)
    g_broadcast = _mm_set1_epi16(g);

    // Roll out loop by eight to fit the eight-element vectors:
    for (int i = 0; i < size; i += 8) {
        // Load eight consecutive elements from bb into vector b:
        b = _mm_load_si128(ToVectorAddress(bb[i]));
        // Load eight consecutive elements from cc into vector c:
        c = _mm_load_si128(ToVectorAddress(cc[i]));
        // Load eight consecutive elements from dd into vector d:
```

```
        d = _mm_load_si128(ToVectorAddress(dd[i]));
        // Add 2 to each element in vector c
        c = _mm_add_epi16(c, two);
        // Add g to each element in vector d
        d = _mm_add_epi16(d, g_broadcast);
        // Compare each element in b to 0 and generate a bit-mask:
        mask = _mm_cmpgt_epi16(b, zero);
        // AND each element in vector c with the corresponding bit-mask:
        a = _mm_and_si128(c, mask);
        // AND each element in vector d with the inverted bit-mask:
        mask = _mm_andnot_si128(mask, d);
        // OR the results of the two AND operations:
        a = _mm_or_si128(a, mask);
        // Store the result vector in eight consecutive elements in aa:
        _mm_store_si128(ToVectorAddress(aa[i]), a);
    }
}
```

The resulting code will be very efficient because it handles eight elements at a time and it avoids the branch inside the loop. Example 12.4b executes three to seven times faster than example 12.4a, depending on how predictable the branch inside the loop is.

The intrinsic vector functions have names that begin with _mm_. These functions are listed in the header file emmintrin.h and explained in the documentation for the Intel C++ compiler and in "IA-32 Intel Architecture Software Developer's Manual" Volume 2A and 2B under "Instruction set reference".

Using vector classes

Programming in the way of example 12.4b is quite tedious indeed, and the program is not very readable. It is possible to write the same in a more human-intelligible way by wrapping the vectors into classes. A set of vector classes for this purpose is provided in various header files that come with the Intel compiler. See the documentation for the Intel C++ compiler for details.

The following table lists the available vector classes. Including dvec.h will give you access to all of these.

Vector class	Size of each element, bits	Number of elements	Type of elements	Total size of vector, bits	Header file
Is8vec8	8	8	char	64	ivec.h
Iu8vec8	8	8	unsigned char	64	ivec.h
Is16vec4	16	4	short int	64	ivec.h
Iu16vec4	16	4	unsigned short int	64	ivec.h
Is32vec2	32	2	int	64	ivec.h
Iu32vec2	32	2	unsigned int	64	ivec.h
I64vec1	64	1	__int64	64	ivec.h
Is8vec16	8	16	char	128	dvec.h
Iu8vec16	8	16	unsigned char	128	dvec.h
Is16vec8	16	8	short int	128	dvec.h
Iu16vec8	16	8	unsigned short int	128	dvec.h
Is32vec4	32	4	int	128	dvec.h
Iu32vec4	32	4	unsigned int	128	dvec.h
I64vec2	64	2	__int64	128	dvec.h
F32vec4	32	4	float	128	fvec.h
F64vec2	64	2	double	128	dvec.h
F32vec8	32	8	float	256	dvec.h
F64vec4	64	4	double	256	dvec.h

Table 12.1. Vector classes

It is not recommended to use the 64-bit vectors in ivec.h because these are incompatible with floating point code. If you do use the 64-bit vectors then you have to execute _mm_empty() after the 64-bit vector operations and before any floating point code. The 128-bit vectors do not have this problem. The 256-bit vectors require that the AVX instruction set is supported by the microprocessor and the operating system.

Rewriting example 12.4b with the use of vector classes makes the code look simpler:

```
// Example 12.4c
// SSE2 instruction set must be enabled
#include <dvec.h>              // Define vector classes

const int size = 128;         // Array size
__declspec(align(16))         // Align arrays by 16
short int aa[size], bb[size], cc[size], dd[size];

// Macro to type-cast an array to a vector:
#define ToVector(x) (*(Is16vec8*)&(x))

void SelectAndAdd(short int g) {
   // Define local vector class variables:
   Is16vec8 a, b, c, d, zero;

   // Set zero to a vector of (0,0,0,0,0,0,0,0)
   zero = Is16vec8(0,0,0,0,0,0,0,0);

   // Roll out loop by eight to fit the eight-element vectors:
   for (int i = 0; i < size; i += 8) {
```

```
        b = ToVector(bb[i]);        // Load eight elements from bb[i]
        c = ToVector(cc[i]);        // Load eight elements from cc[i]
        d = ToVector(dd[i]);        // Load eight elements from dd[i]
        d += _mm_set1_epi16(g);     // Make eight copies of g and add
        c += _mm_set1_epi16(2);     // Make eight copies of 2 and add
        a = select_gt(b,zero,c,d);  // a = b > 0 ? c : d;
        ToVector(aa[i]) = a;        // Store eight elements in aa[i]
    }
}
```

Example 12.4c looks simpler than example 12.4b because we can use the operator + or +=
and the predefined function `select_gt` on the vector class objects.

We still haven't got rid of all the intrinsic functions and the ugly type-casting macro
`ToVector`. We can make the code even more readable by defining our own vector class
and operators, as explained below.

Redefining a vector class

It is possible to define your own vector classes with associated operators and functions.
This is done most easily by inheriting from the predefined vector classes. You can define
new member functions and operators for the derived class, but you cannot add any data
members without destroying the vectorization. The vector classes have a size that fits into
the vector registers. Adding any data members or virtual member functions will make the
class objects bigger so they don't fit into the vector registers any more.

The following example shows how example 12.4c can be made even more readable by
redefining the vector classes.

```
// Example 12.4d
// SSE2 instruction set must be enabled
#include <dvec.h>               // Define vector classes

const int size = 128;           // Array size
__declspec(align(16))           // Align arrays by 16
short int aa[size], bb[size], cc[size], dd[size];

// Redefine class for vector of eight 16-bit integers:
class Int16Vector8 : public Is16vec8 {
   public:
   // Default constructor:
   Int16Vector8() {};
   // Constructor to convert from parent type:
   Int16Vector8(Is16vec8 const & x) : Is16vec8(x) { }
   // Constructor to convert from type __m128i used in intrinsics:
   Int16Vector8(__m128i const & x) : Is16vec8(x) { }
   // Constructor to broadcast integer into all elements of vector:
   Int16Vector8(short int const n) {*this = _mm_set1_epi16(n);}
   // Member function to load eight values from array:
   void load(short int const * p) {
      *this = *(Int16Vector8 const*)p;}
   // Member function to store eight values into array:
   void store(short int * p) {*(Int16Vector8*)p = *this;}
};

// Define another class for a vector of eight 16-bit masks,
// Each mask should only be 0xFFFF for true or 0x0000 for false.
class Bool16Vector8 : public M128 {
   public:
   // Default constructor:
   Bool16Vector8() {};
   // Constructor to convert from type __m128i used in intrinsics:
```

```
        Bool16Vector8(__m128i const x) : M128(x) { }
    };

    // Operator + adds the same integer to all elements of vector
    static inline Int16Vector8 operator + (Int16Vector8 const & a,
    short int const b) {
        return a + Int16Vector8(b);}

    // Operator > compares two vectors and returns a vector of masks.
    // Each mask is 0xFFFF for true, 0x0000 for false.
    static inline Bool16Vector8 operator > (Int16Vector8 const & a,
    Int16Vector8 const & b) {
        return cmpgt(a, b);}

    // Operator > compares all elements of vector with the same integer
    // and returns a vector of masks.
    static inline Bool16Vector8 operator > (Int16Vector8 const & a,
    short int const b) {
        return a > Int16Vector8(b);}

    // Function select does the same as the ?: operator,
    // element by element:
    // return_value[i] = s[i] ? a[i] : b[i]
    // s is a vector of 8 masks, each mask is
    // 0xFFFF for true, 0x0000 for false.
    static inline Int16Vector8 select (Bool16Vector8 const & s,
    Int16Vector8 const & a, Int16Vector8 const & b) {
        return (s & a) | andnot(s, b);}

    // And here comes our revised SelectAndAdd function
    void SelectAndAdd(short int g) {
        int i;
        // Define local vector class variables:
        Int16Vector8 a, b, c, d;

        // Roll out loop by eight to fit the eight-element vectors:
        for (i = 0; i < size; i += 8) {
            b.load(bb + i);            // Load eight elements from bb[i]
            c.load(cc + i);            // Load eight elements from cc[i]
            d.load(dd + i);            // Load eight elements from dd[i]
            a = select(b > 0, c + 2, d + g); // Select c+2 or d+g
            a.store(aa + i);        // Store eight elements in aa[i]
        }
    }
```

Example 12.4d does exactly the same as example 12.4b and c. The `SelectAndAdd` function looks simpler now because all the intrinsic functions and type conversions have been wrapped into the new class definitions. Defining the new class `Int16Vector8` which inherits from class `Is16vec8` allows us to redefine the constructors and operators to suit our needs. The constructor that converts from `Is16vec8` to `Int16Vector8` makes it possible to use all the operators that are defined for `Is16vec8` on vectors of class `Int16Vector8`. For example, it would be legal to write `a = b * c + d`.

The constructor that converts from `__m128i` to `Int16Vector8` makes it possible to use the intrinsic functions that return `__m128i`.

The constructor that converts from `short int` to `Int16Vector8` makes it easy to make a vector with the same value in all elements.

The load member function makes it easy to make a vector out of eight consecutive elements in an array. Note that the first element must be properly aligned. For example `a = aa` in the above example would be legal, `a = aa + 1` would cause an error, `a = aa + 8`

would be legal. The `store` member function makes it easy to store a vector into eight consecutive elements of an array. The first element must be properly aligned.

The `operator +` is already defined, but I have made an overloaded version for adding an integer to all elements in an array. Without this operator we would have to replace `c+2` by `c+Int16Vector8(2)`. We may define similar operators for `-` and `*`.

The `operator >` is not predefined, so we have to define it. I have made two overloaded versions of the `operator >`, one where the second operand is a vector and one where the second operand is an integer.

The `operator >` returns a vector of eight 16-bit masks that are used as Booleans. I have defined a separate class, `Bool16Vector8`, for this vector of Booleans. We could have used `Int16Vector8` for this purpose, but I think it is safer to define a separate class for a vector of elements that can have no other values than `0x0000` and `0xFFFF`. This allows us to make sure that the `select` function is called with only this type as the `s` parameter, and not just any vector.

Note that I have declared all the functions and operators `static` and `inline` in order to make it easier for the compiler to optimize the code. The vector parameters are declared as references (`&`) in order to make the parameter transfer efficient in case the function or operator is not inlined. The reference parameters have the `const` modifier in order to allow an expression rather than a variable as input.

It is, admittedly, not easier to make the code like example 12.4d than the previous examples because of the extra class definitions and operators when these are used only once. The idea is to put these definitions into a reusable header file. Once these definitions are made and tested, the only thing that has to be done to make vectorized code is to write the code like the `SelectAndAdd` function in example 12.4d.

The three versions 12.4b, c and d generate almost exactly the same code and they are equally efficient. It may come as a surprise that the complicated syntax with inheritance in many levels doesn't generate a complicated code. We can thank the compiler for that. The Intel compiler does an excellent job of optimizing this code, and so does the Gnu compiler. The functions, operators and constructors are all inlined and the references and parameter transfer overheads eliminated. Furthermore, it uses the methods of loop-invariant code motion. The ability of the compilers to do constant propagation on vector code is very limited, however.

Table 12.2 below shows measured execution times for the loop in example 12.4a, b and c on different microprocessors. The execution time for example 12.4a depends on the predictability of the `bb[i] > 0` branch. The values listed as best case apply when the branch always goes the same way. The values listed as worst case apply when the branch goes randomly one way or the other. The timing results for example 12.4b and c do not depend on branch predictability.

Code example	Branch predicta-bility	Time per element AMD64	Time per element Pentium 4	Time per element Pentium M	Time per element Core 2
12.4a	best case	8.0	4.6	9.0	8.0
12.4a	worst case	9.2	9.8	10.9	10.4
12.4b	worst case	1.0	1.6	1.3	0.5
12.4c	worst case	1.0	1.6	1.3	0.7

Table 12.2. Loop execution times for example 12.4 for different microprocessors. Clock cycles per element.

Defining missing operators

It can be quite complicated to do operations that are not supported by the vector class operators or intrinsic functions. For example, the SSE2 instruction set has no function for multiplying vectors of 32 bit integers. There is a $32 \times 32 \rightarrow 64$ bit multiply operation that we can use, but it requires a lot of data shuffling to get the data into the right places:

```
// Example 12.5
static inline Is32vec4 operator * (Is32vec4 const & a,
Is32vec4 const & b) {
    __m128i a13, b13, prod02, prod13, prod01, prod23, prod0123;
    a13 = _mm_shuffle_epi32(a, 0xF5);             // (-,a3,-,a1)
    b13 = _mm_shuffle_epi32(b, 0xF5);             // (-,b3,-,b1)
    prod02 = _mm_mul_epu32(a, b);                 // (-,a2*b2,-,a0*b0)
    prod13 = _mm_mul_epu32(a13, b13);             // (-,a3*b3,-,a1*b1)
    prod01 = _mm_unpacklo_epi32(prod02,prod13);   // (-,-,a1*b1,a0*b0)
    prod23 = _mm_unpackhi_epi32(prod02,prod13);   // (-,-,a3*b3,a2*b2)
    prod0123 = _mm_unpacklo_epi64(prod01,prod23); // (ab3,ab2,ab1,ab0)
    return prod0123;
}
```

This, of course, takes much longer time than 16 bit multiplications, which have intrinsic support. Multiplications by a power of 2 can be done by shifting.

There is no way to do integer division in vectors. Integer division by a constant can sometimes be done by multiplication and shifting. Otherwise, you have to convert to floating point, as example 12.6 below shows, or do the entire calculation with floating point vectors.

Mixed vector types

The following examples show how to vectorize a loop with mixed types:

```
// Example 12.6a
const int size = 128;  // size of arrays

void AddAndDivide(int aa[], int bb[], int cc[], float ff[]) {
    for (int i = 0; i < size; i++) {
        aa[i] = (bb[i] + cc[i]) / ff[i];
    }
}
```

Example 12.6a is not vectorized automatically by the compiler. We can use the same method as in example 12.4d to get the following code:

```
// Example 12.6b
// SSE2 instruction set must be enabled
#include <dvec.h>
const int size = 128;  // size of arrays

// Redefine class for vector of four 32-bit integers:
class Int32Vector4 : public Is32vec4 {
  public:
  // Default constructor:
  Int32Vector4() {};
  // Constructor to convert from parent type:
  Int32Vector4(Is32vec4 const & x) : Is32vec4(x) { }
  // Constructor to convert from type __m128i used in intrinsics:
  Int32Vector4(__m128i const & x) : Is32vec4(x) { }
  // Constructor to broadcast integer into all elements of vector:
  Int32Vector4(int const n) {*this = _mm_set1_epi32(n);}
  // Member function to load four values from array:
  void load (int const * p) {*this = *(Int32Vector4 const*)p;}
  // Member function to store four values into array:
```

```
        void store(int * p) {*(Int32Vector4*)p = *this;}
    };

    // Redefine class for vector of four 32-bit floats:
    class FloatVector4 : public F32vec4 {
        public:
        // Default constructor:
        FloatVector4() {};
        // Constructor to convert from parent type:
        FloatVector4(F32vec4 const & x) : F32vec4(x) { }
        // Constructor to convert from type __m128 used in intrinsics:
        FloatVector4(__m128 const & x) : F32vec4(x) { }
        // Constructor to broadcast float into all elements of vector:
        FloatVector4(float const x) : F32vec4(x) { }
        // Constructor to convert from vector of integers:
        FloatVector4(Int32Vector4 const & x) {*this = _mm_cvtepi32_ps(x);}
        // Member function to load four values from array:
        void load (float const * p) {*this = *(FloatVector4 const*)p;}
        // Member function to store four values into array:
        void store(float * p) {*(FloatVector4*)p = *this;}
    };

    // Function to convert vector of floats to vector of integers,
    // using rounding:
    static inline Int32Vector4 round(FloatVector4 const & x) {
        return _mm_cvtps_epi32(x);
    }

    // Function to convert vector of floats to vector of integers,
    // using truncation:
    static inline Int32Vector4 truncate(FloatVector4 const & x) {
        return _mm_cvttps_epi32(x);}

    // Vectorized function:
    void AddAndDivide(int aa[], int bb[], int cc[], float ff[]) {
        Int32Vector4 a, b, c;
        FloatVector4 f, t;

        // Roll out loop by four to fit the four-element vectors:
        for (int i = 0; i < size; i += 4) {
            b.load(bb + i);           // Load four values from bb
            c.load(cc + i);           // Load four values from cc
            f.load(ff + i);           // Load four values from ff
            t = FloatVector4(b + c);  // Add integers, convert to floats
            t = t / f;                // Divide by floats
            a = truncate(t);          // Convert back to integers
            a.store(aa + i);          // Store four values in aa
        }
    }
```

Note that there are two different ways of converting an integer vector to a float vector. The `_mm_cvtepi32_ps` function converts each integer element to the floating point representation of the same value. The `_mm_castsi128_ps` function changes the type of the vector but preserves all bits unchanged. The latter function is used when an integer vector is used as a mask to select floating point values by means of AND and OR operations, as in example 12.4. The mask for true, which is 0xFFFFFFFF would be converted to the floating point representation of -1.0 (= 0xBF800000) by `_mm_cvtepi32_ps`, but remain unchanged by `_mm_castsi128_ps`. The opposite conversion is done with `_mm_cvttps_epi32` and `_mm_castps_si128` respectively.

Mixing vectors with different number of elements

Vectorization of loops becomes more difficult when mixing different types of vectors with different numbers of elements. If the integers in example 12.6b are changed to 16-bit integers then we have integer vectors with eight elements and floating point vectors with four elements. The loop must be rolled out by eight and make two floating point vectors for every integer vector:

```
// Example 12.6c
// SSE2 instruction set must be enabled
#include <dvec.h>
const int size = 128;  // Size of arrays

class Int32Vector4; // Define as in example 12.6b

class Int16Vector8; // Define as in example 12.4d

class FloatVector4; // Define as in example 12.6b

// Define as in example 12.6b:
static inline Int32Vector4 truncate(FloatVector4 const & x);

// Unpack a vector of eight 16-bit integers into two vectors of
// four 32-bit integers each, using sign-extension:
static inline void unpack (Int32Vector4 & uLo, Int32Vector4 & uHi,
Int16Vector8 const & pk) {
   __m128i aLo, aHi;
   aLo = _mm_unpacklo_epi16(pk, pk);  // Unpack low  four integers
   aHi = _mm_unpackhi_epi16(pk, pk);  // Unpack high four integers
   uLo = _mm_srai_epi32(aLo, 16);     // Sign-extend to 32 bits
   uHi = _mm_srai_epi32(aHi, 16);     // Sign-extend to 32 bits
}

// Pack two vectors of four 32-bit integers each into one
// vector of eight 16-bit integers using signed saturation:
static inline void pack (Int16Vector8 & pk, Int32Vector4 const & uLo,
Int32Vector4 const & uHi) {
   pk = _mm_packs_epi32(uLo, uHi);
}

void AddAndDivide (short int *aa, short int *bb, short int *cc,
float *ff) {
   // Define temporary vectors:
   Int16Vector8 a, b, c;          // Vectors of eight 16-bit integers
   Int32Vector4 p1, p2;           // Vectors of four 32-bit integers
   FloatVector4 f1, f2, t1, t2;   // Vectors of four floats

   // Roll out loop by eight to fit the vector with most elements
   for (int i = 0; i < size; i += 8) {
       b .load(bb + i);           // Load eight values from bb
       c .load(cc + i);           // Load eight values from cc
       f1.load(ff + i);           // Load first four values from ff
       f2.load(ff + i + 4);       // Load next  four values from ff
       a = b + c;                 // Add eight integers
       unpack (p1, p2, a);        // Unpack into 2 times 4 integers
       t1 = FloatVector4(p1) / f1; // Divide with first 4 floats
       t2 = FloatVector4(p2) / f2; // Divide with next  4 floats
       p1 = truncate(t1);         // Convert first 4 back to int
       p2 = truncate(t2);         // Convert next  4 back to int
       pack (a, p1, p2);          // Pack 2 times 4 int into 8 int
       a.store(aa + i);           // Store eight values into aa
   }
}
```

Example 12.6c does not give the same result as example 12.6b in case of overflow because it uses saturation. Example 12.6c is less efficient than example 12.6b because of the extra time needed for packing and unpacking the vectors with different numbers of elements. Mixing vectors with different element sizes should be avoided if possible.

Making your own vector classes

In the above examples, I have defined my own vector classes by inheriting from the vector classes that are already defined in `dvec.h` in order to inherit the functions and operators that are already defined. In some cases it is preferable to make our own vector classes from scratch rather than basing our classes on inheritance from predefined classes. This can be useful if you don't have the Intel header files, if you don't want to use them, or if you want the operators to work differently.

It is a bad idea to inherit from the predefined vector classes if the algebra for your vectors is different from the simple element-by-element operations defined in `dvec.h`. Let's consider the example of complex numbers. The product of two complex numbers is not computed just by multiplying the real parts and the imaginary parts. The formula is more complicated:

$$(a + ib) \cdot (c + id) = ((ac - bd) + i(ad + bc))$$

The formula for division is even more complicated. It is possible to define a class `complex` for complex numbers by inheriting from `F64vec2` (a vector of two `double`'s) and then redefining the `*` and `/` operators for class `complex`. The `+` and `-` operators do not have to be redefined. But there is a pitfall here. The expression `y = (a + b) * (c + d)` would use the `*` operator for the base class `F64vec2`, not the `*` operator for the derived class `complex`, because both parentheses have type `F64vec2`. We can avoid this by also redefining the `+` operator, but the method is still somewhat risky because there may be some other operator, such as `*=`, that we have forgotten to redefine. It is safer to define a new base class rather than redefining an existing class. This prevents the inadvertent use of operators that we have forgotten to redefine.

The following example shows how to define a vector class from scratch and define operators for complex number algebra.

```
// Example 12.7
#include <emmintrin.h>

// Define class complexd for vector of two doubles.
// The low element represents the real part and the high element
// represents the imaginary part of a complex number:

__declspec(align(16))       // must be aligned
class complexd {
  protected:
  __m128d vec;               // Vector of two double's
  public:

  // Default constructor:
  complexd() {};

  // Constructor to make from real and imaginary part:
  complexd(double const re, double const im) {
    vec = _mm_set_pd(im, re);}

  // Constructor to convert from real number. Set imag. part to 0:
  complexd(double const x) {
    vec = _mm_set_pd(0., x);}

  // Constructor to convert from type __m128d used in intrinsics:
```

118

```cpp
    complexd(__m128d const & x) {vec = x;}

    // Operator to convert to __m128d used in intrinsics
    operator __m128d() const {return vec;}

    // Member function to extract real part:
    double Re() const {
       return *(double const*)&vec;}

    // Member function to extract Imaginary part:
    double Im() const {
       return *(double const*)&_mm_shuffle_pd(vec, vec, 1);}
};

// Define operators for class complexd:

// complex + real: add only to real part:
static inline complexd operator + (complexd const &a, double const b){
    return _mm_add_sd(a, _mm_load_sd(&b));}

// complex - real: subtract only from real part:
static inline complexd operator - (complexd const &a, double const b){
    return _mm_sub_sd(a, _mm_load_sd(&b));}

// complex * real: multiply both real and imaginary part:
static inline complexd operator * (complexd const &a, double const b){
    return _mm_mul_pd(a, _mm_set1_pd(b));}

// complex / real: multiply both real and imag. part by reciprocal b:
static inline complexd operator / (complexd const &a, double const b){
    return _mm_mul_pd(a, _mm_set1_pd(1. / b));}

// complex + complex: add both parts:
static inline complexd operator + (complexd const & a,
complexd const & b) {
    return _mm_add_pd(a, b);}

// complex - complex: subtract both parts:
static inline complexd operator - (complexd const & a,
complexd const & b) {
    return _mm_sub_pd(a, b);}

// complex * complex: (a.re*b.re-a.im*b.im, a.re*b.im+b.re*a.im).
// This version is for SSE2 instruction set. It is more efficient
// to use _mm_addsub_pd and _mm_hadd_pd if SSE3 is available.
static inline complexd operator * (complexd const & a,
complexd const & b) {
    __m128d a_flipped;                   // (a.im,a.re)
    __m128d b_re;                        // (b.re,b.re)
    __m128d b_im;                        // (-b.im,b.im)
    static const union {                 // (signbit,0)
       int i[4];   __m128d v;
    } signbitlow = {0,0x80000000,0,0};

    b_im = _mm_shuffle_pd(b,b,3);      // Imag. part of b in both
    b_re = _mm_shuffle_pd(b,b,0);      // Real  part of b in both
    a_flipped = _mm_shuffle_pd(a,a,1); // Swap real and imag parts of a
    b_im = _mm_xor_pd(b_im, signbitlow.v); // Change sign of low

    // Multiply and add:
    return (complexd)_mm_mul_pd(a, b_re) +
           (complexd)_mm_mul_pd(a_flipped, b_im);
}

// complex / complex:
```

```
    // (a.re*b.re+a.im*b.im, b.re*a.im-a.re*b.im)/(b.re*b.re+b.im*b.im).
    // This version is for SSE2 instruction set. It is more efficient
    // to use _mm_addsub_pd and _mm_hadd_pd if SSE3 is available.
    static inline complexd operator / (complexd const & a,
    complexd const & b) {
        __m128d a_flipped;                 // (a.im,a.re)
        __m128d b_re;                      // (b.re,b.re)
        __m128d b_im;                      // (-b.im,b.im)
        __m128d ab_conj;                   // a * conjugate(b)
        __m128d abs_b_square;              // (b.re*b.re,b.im*b.im)
        double b_abs_square_recip;         // 1/(abs(b*b))

        abs_b_square = _mm_mul_pd(b, b);   // (b.re*b.re,b.im*b.im)
        // Reciprocal of horizontal add:
        b_abs_square_recip = 1. /
            *(double const*)& _mm_add_sd(abs_b_square,
            _mm_shuffle_pd(abs_b_square, abs_b_square, 1));

        // The following code is made as similar to the operator * as
        // possible, to enable common subexpression elimination in code
        // that contains both operator * and operator / with the same
        // second operand:
        static const union {               // (signbit,0)
            int i[4];  __m128d v;
        } signbitlow = {0,0x80000000,0,0};
        b_im = _mm_shuffle_pd(b,b,3);      // Imag. part of b in both
        b_re = _mm_shuffle_pd(b,b,0);      // Real  part of b in both
        a_flipped = _mm_shuffle_pd(a,a,1); // Swap real and imag parts of a
        b_im = _mm_xor_pd(b_im, signbitlow.v); // Change sign of re

        // Multiply and subtract:
        ab_conj = (complexd)_mm_mul_pd(a, b_re) -
                  (complexd)_mm_mul_pd(a_flipped, b_im);
        // Multiply by b_abs_square_recip:
        return (complexd)ab_conj * b_abs_square_recip;
    }

    // - complex: (-a.re, -a.im):
    static inline complexd operator - (complexd const & a) {
        static const union {               // (signbit,signbit)
            int i[4];  __m128d v;
        } signbits = {0,0x80000000,0,0x80000000};
        return _mm_xor_pd(a, signbits.v);  // Change sign of both elements
    }

    // complex conjugate: (a.re, -a.im)
    static inline complexd operator ~ (complexd const & a) {
        static const union {               // (signbit,signbit)
            int i[4];  __m128d v;
        } signbithigh = {0,0,0,0x80000000};
        return _mm_xor_pd(a, signbithigh.v); // Change sign of imag. part
    }


    // Example of use:
    // Second order complex polynomial:
    complexd polynomial(complexd x) {
        const complexd a(1.,2.), b(3.,4.), c(5.,6.);
        return (a * x + b) * x + c;
    }
```

Example 12.7 shows how to define your own vector class from scratch and make appropriate operators for it. The Intel and Gnu compilers optimize this code quite well. The code may be further improved by using SSE3 instructions, if available. (The SSE3

instructions `_mm_addsub_pd` and `_mm_movedup_pd` are designed specifically for complex math. `_mm_hadd_pd` is useful for horizontal addition).

Another question is whether it is optimal to use vector operations at all in this case. The `*` and `/` operators involve a lot of shuffling in order to get the data elements into the right places. An implementation without vectorization does not require these shuffle operations. The vectorized code is in fact faster than the non-vectorized code on microprocessors with 128 bit execution units, but not always on older microprocessors with 64 bit execution units.

Vectorized table lookup

Lookup tables can be useful for optimizing code, as explained on page 138. Unfortunately, table lookup is often an obstacle to vectorization. The newest instruction sets include a few instructions that may be used for vectorized table lookup with small tables. These instructions are summarized below. Microprocessors with these instructions are not available yet (2009).

Intrinsic function	Max. number of elements in table	Size of each table element	Number of simultaneous lookups	Instruction set needed
_mm_perm_epi8	32	1 byte = char	16	XOP, AMD only
_mm_permutevar_ps	4	4 bytes = float or int	4	AVX
_mm256_permutevar_ps	4	4 bytes = float or int	8	AVX
Table 12.3. Intrinsic functions for vectorized table lookup				

12.4 Mathematical functions for vectors

There are various function libraries for computing mathematical functions such as logarithms, exponential functions, trigonometric functions, etc. in vector operands. These function libraries are useful for vectorizing mathematical code.

The following example is a piece of code that can benefit from vectorization.

```
// Example 12.8a
#include <math.h>

const int size = 128;
float a[size], b[size];

void Func() {
   for (int i = 0; i < size; i++) {
      a[i] = 0.5f * exp(b[i]);
   }
}
```

The exponential function is available in a vector version calculating four exponential function values in one function call. This function is found in the Intel library called "Short Vector Math Library" (SVML) which comes with the Intel C++ compiler. The Intel compiler has intrinsic names for the SVML functions in the header file `ia32intrin.h`. For other compilers, you must write the function prototypes manually based on a library listing.

The same vectorized math functions are available with different names in the AMD math core library and with the PathScale compiler. The vectorized code looks like this:

```
// Example 12.8b
// Compile with SSE2 enabled.
// Link with appropriate library, e.g. svml_dispmt.lib

// Define vector classes:
#include <dvec.h>

// Function prototype for exponential function on vector of 4 floats
extern "C" __m128 __svml_expf4(__m128); // for Intel SVML v. 10.3

const int size = 128;
__declspec(align(16))
float a[size], b[size];

void Func() {
    // Define vector of 4 floats for temporary storage:
    F32vec4 temp;

    // Roll out loop by four to fit vectors of four floats:
    for (int i = 0; i < size; i += 4) {

        // Load four values from b into vector:
        temp = *(F32vec4 const *)&b[i];

        // Call exponential founction on vector and multiply by 0.5:
        temp = __svml_expf4(temp) * F32vec4(0.5f);

        // Store result in four consecutive elements of a:
        *(F32vec4 *)&a[i] = temp;
    }
}
```

The code in example 12.8a is automatically converted to 12.8b when compiled with an Intel, Gnu or PathScale compiler, with the appropriate compiler switches for automatic vectorization. The explicit call to the SVML library is useful if another compiler is used or if the compiler cannot vectorize the code automatically. The AMD math core library may be used instead.

The vector math functions have different names in different libraries. For example, the exponential function on a vector of 4 floats has the following names:

Library	Function name
Intel SVML v.10.2 & earlier	`vmlsExp4`
Intel SVML v.10.3 & later	`__svml_expf4`
Intel SVML + ia32intrin.h	`_mm_exp_ps`
AMD Core Math Library	`__vrs4_expf`

The SVML has different variants of the functions with different performance. Functions with suffix `_ep` for "Extended Performance" are faster but less accurate, e.g. `__svml_expf4_ep`. Functions with suffix `_ha` have higher accuracy, but are slower. There are also versions with 256 bit vectors for processors with AVX, e.g. `__svml_expf8`. See page 132 for how to use the Intel function libraries on non-Intel processors.

An alternative is to use the Intel Vector Math Library (VML) which is part of Intel's Math kernel Library (MKL) (www.intel.com). The VML functions take the entire arrays as input and output. This library contains mathematical functions for large vectors. The following code shows how to use the VML Library:

```
// Example 12.8c
// Link with mkl_c.lib from Intel Math Kernel Library
```

```
// Obtain this header file from Intel Math Kernel Library:
#include <mkl_vml_functions.h>

// Obtain header file for vector classes from Intel C++ compiler:
#include <dvec.h>

const int size = 128;
__declspec(align(16))
float a[size], b[size];

void Func() {
    // Define vector of 4 floats for temporary storage:
    F32vec4 temp;

    // Intel Math Kernel library function vsExp calculates 'size'
    // exponentials. The loop is inside this library function:
    vsExp(size, b, a);

    // Multiply all elements in a by 0.5:
    for (int i = 0; i < size; i += 4) {

        // Multiply four elements by 0.5:
        *(F32vec4 *)&a[i] *= F32vec4(0.5f);
    }
}
```

Note that the Intel function libraries detect whether the code is running on an Intel CPU or not. If the CPU is not an Intel then the library may not use the optimal instruction set or it may fail completely. See page 132 for how to overcome this limitation.

The following table compares the measured computation time in clock cycles per element on various CPUs for example 12.8a, b and c.

Code example	Compiler used	Time per element AMD64	Time per element Pentium 4	Time per element Pentium M	Time per element Core 2
12.8a	Microsoft	59	311	121	173
12.8a	Intel 9.1	80	45	33	49
12.8a*	Intel 9.1	60	45	33	49
12.8b	Intel 9.1	25	17	18	12
12.8b*	Intel 9.1	18	17	18	12
12.8c	Intel 9.1	16	12	14	13
12.8c*	Intel 9.1	11	12	14	13

Table 12.4. Loop execution times for example 12.8 for different microprocessors. Clock cycles per element (* = CPU detection overridden).

These tests show that it is very advantageous to use the math libraries. The Vector Math Library (example 12.8c) is a little faster than the Short Vector Math Library (example 12.8b) on some processors, but the Vector Math Library has a much larger footprint in the code cache and a very large initialization routine, which is called the first time the library function is called. The initialization time is not included in the above time measurements. The performance of the Core 2 processor is lower than expected due to the fact that the Intel function libraries are not yet optimized for this processor and that the performance is limited by the predecoder rather than the execution units in this processor.

12.5 Aligning dynamically allocated memory

Memory allocated with `new` or `malloc` is typically aligned by 8 rather than by 16. This is a problem with vector operations, which require alignment by 16. The Intel compiler has solved this problem by defining `_mm_malloc` and `_mm_free`.

A more convenient method is to wrap the allocated array into a container class that takes care of the alignment. See www.agner.org/optimize/cppexamples.zip for examples of how to make aligned arrays with vector access.

12.6 Aligning RGB video or 3-dimensional vectors

RGB image data have three values per point. This does not fit into a vector of e.g. four floats. The same applies to 3-dimensional geometry and other odd-sized vector data. The data have to be aligned by the vector size for the sake of efficiency. Using unaligned reads and writes will slow down the execution to the point where it is not advantageous to use vector operations. You may choose one of the following solutions, depending on what fits best into the algorithm in question:

- Put in an unused fourth value to make the data fit into the vector. This is a simple solution, but it increases the amount of memory used. You may avoid this method if memory access is a bottleneck.

- Organize the data into groups of four (or eight) points with the four R value in one vector, the four G values in the next vector, and the four B value in the last vector.

- Organize the data with all the R values first, then all the G values, and last all the B values.

The choice of which method to use depends on what fits best into the algorithm in question. You may choose the method that gives the simplest code.

If the number of points is not divisible by the vector size then add a few unused points in the end in order to get an integral number of vectors.

12.7 Conclusion

Vectorized code often contains a lot of extra instructions for converting the data to the right format and getting them into the right positions. The amount of extra data conversion and shuffling that is needed determines whether it is profitable to use vectorized code or not.

The code in example 12.7 is slower than non-vectorized code on older processors, but faster on processors with 128 bit execution units. The code in example 12.6b and 12.6c is faster than the non-vectorized code on all processors despite the extra data conversion, packing and unpacking. This is because the bottleneck here is not data conversion and packing, but division. Division is very time-consuming and there is a lot to save by doing division in single precision vectors. The code in example 12.8b and c benefit a lot from vectorization.

I will conclude this section by summing up the factors that decide whether vectorization is optimal or not.

Factors that make vectorization favorable:

- Small data types: char, short int, float.
- Similar operations on all data in large arrays.
- Array size divisible by vector size.

- Unpredictable branches that select between two simple expressions.
- Operations that are only available with vector operands: minimum, maximum, saturated addition, fast approximate reciprocal, fast approximate reciprocal square root, RGB color difference.
- Mathematical vector function libraries.
- Use Gnu or Intel compiler.
- Use CPUs with execution units same size as vector register.

Factors that make vectorization less favorable:

- Larger data types: __int64, double.
- Misaligned data.
- Extra data conversion, shuffling, packing, unpacking needed.
- Predictable branches that can skip large expressions when not selected.
- Compiler has insufficient information about pointer alignment and aliasing.
- Operations that are missing in the instruction set for the appropriate type of vector, such as 32-bit integer multiplication prior to SSE4.1 and integer division.
- Older CPUs with execution units smaller than the vector register size.

Vectorized code is more difficult for the programmer to make and therefore more error prone. The vectorized code should therefore preferably be put away in reusable and well-tested library modules and header files.

13 Making critical code in multiple versions for different CPUs

Microprocessor producers keep adding new instructions to the instruction set. These new instructions can make certain kinds of code execute faster. The most important addition to the instruction set is the vector operations mentioned in chapter 12.

A disadvantage of using the newest instruction set is that the compatibility with older microprocessors is lost. This dilemma can be solved by making the most critical parts of the code in multiple versions for different CPUs. This is called CPU dispatching. For example, you may want to make one version that takes advantage of the AVX instruction set, another version for CPUs with only the SSE2 instruction set, and a generic version that is compatible with old microprocessors without any of these instruction sets. The program should automatically detect which instruction set is supported by the CPU and the operating system and choose the appropriate version of the subroutine for the critical innermost loops.

13.1 CPU dispatch strategies

It is quite expensive - in terms of development, testing and maintenance - to make a piece of code in multiple versions, each carefully optimized and fine-tuned for a particular set of CPUs. These costs can be justified for general function libraries that are used in multiple applications, but rarely for application-specific code. If you consider making highly optimized code with CPU dispatching, then it is advisable to make it in the form of a re-usable library if possible. This also makes testing and maintenance easier.

I have done a good deal of research on CPU dispatching and discovered that many common programs use inappropriate CPU dispatch methods.

The most common pitfalls of CPU dispatching are:

- Optimizing for present processors rather than future processors. Consider the time it takes to develop and publish a function library with CPU dispatching. Add to this the

time it takes before the application programmer gets the new version of the library. Add to this the time it takes to develop and market the application program. Add to this the time before the end user gets the latest version of the application program. All in all, it will often take several years before your code is running in the majority of end user's computers. At this time, any processor that you optimized for is likely to be obsolete. Programmers very often underestimate this time lag.

- Thinking in terms of specific processor models rather than processor features. The programmer typically thinks "what works best on processor X?" rather than "what works best on processors with this instruction set?". A list of which code branch to use for each processor model is going to be very long and difficult to maintain. And it is unlikely that the end user will have an up-to-date version. The CPU dispatcher should not look at CPU brand names and model numbers, but on what instruction sets and other features it has.

- Assuming that processor model numbers form a logical sequence. If you know that processor model N supports a particular instruction set, then you cannot assume that model N+1 supports at least the same instruction set. Neither can you assume that model N-1 is inferior. A model with a higher number is not necessarily newer. The CPU family and model numbers are not always sequential, and you cannot make any assumption about an unknown CPU based on its family and model number.

- Failure to handle unknown processors properly. Many CPU dispatchers are designed to handle only known processors. Other brands or models that were unknown at the time of programming will typically get the generic branch, which is the one that gives the worst possible performance. We must bear in mind, that many users will prefer to run a speed-critical program on the newest CPU model, which quite likely is a model that was unknown at the time of programming. The CPU dispatcher should give a CPU of unknown brand or model the best possible branch if it has an instruction set that is compatible with that branch. The common excuse that "we don't support processor X" is simply not appropriate here. It reveals a fundamentally flawed approach to CPU dispatching.

- Underestimating the cost of keeping a CPU dispatcher updated. It is tempting to fine-tune the code to a specific CPU model and then think that you can make an update when the next new model comes on the market. But the cost of fine-tuning, testing, verifying and maintaining a new branch of code is so high that it is unrealistic that you can do this every time a new processor enters the market for many years to come. Even big software companies often fail to keep their CPU dispatchers up to date. A more realistic goal is to make a new branch only when a new instruction set opens the possibility for significant improvements.

- Making too many branches. If you are making branches that are fine-tuned for specific CPU brands or specific models then you will soon get a lot of branches that take up cache space and are difficult to maintain. Any specific bottleneck or any particularly slow instruction that you are dealing with in a particular CPU model is likely to be irrelevant within a year or two. Often, it is sufficient to have just two branches: one for the latest instruction set and one that is compatible with CPUs that are up to five or ten years old. The CPU market is developing so fast that what is brand new today will be mainstream next year.

- Ignoring virtualization. The time when the CPUID instruction was certain to truly represent a known CPU model is over. Virtualization is becoming increasingly important. A virtual processor may have a reduced number of cores in order to reserve resources for other virtual processors on the same machine. The virtual processor may be given a false model number to reflect this or for the sake of

compatibility with some legacy software. It may even have a false vendor string. In the future we may also see emulated processors and FPGA soft cores that do not correspond to any known hardware CPU. These virtual processors can have any brand name and model number. The only CPUID information that we can rely on is the feature information, such as supported instruction sets and cache sizes.

Fortunately, the solution to these problems is quite simple in most cases: The CPU dispatcher should have as few branches as possible, and the dispatching should be based on which instruction sets the CPU supports, rather than its brand, family and model number.

I have seen many examples of poor CPU dispatching. For example, the latest version of Mathcad (v. 15.0) is using a six years old version of Intel's Math Kernel Library (MKL v. 7.2). This library has a CPU dispatcher that doesn't handle current CPUs optimally. The speed for certain tasks on current Intel CPUs can be increased by more than 33% when the CPUID is artificially changed to the old Pentium 4. The reason is that the CPU dispatcher in the MKL relies on the CPU family number, which is 15 on the old Pentium 4, while all newer Intel CPUs have family number 6! The speed on non-Intel CPUs was more than doubled for this task when the CPUID was manipulated to indicate an Intel Pentium 4. Even worse, many software products fail to recognize VIA processors because this brand was less popular at the time the software was developed.

A CPU dispatch mechanism that treats different brands of CPUs unequally can become a serious legal issue, as you can read about in [my blog](#). Here, you can also find more examples of bad CPU dispatching.

Obviously, you should apply CPU dispatching only to the most critical part of the program - preferably isolated into a separate function library. The radical solution of making the entire program in multiple versions should be used only when instruction sets are mutually incompatible. A function library with a well-defined functionality and a well-defined interface to the calling program is more manageable and easier to test, maintain and verify than a program where the dispatch branches are scattered everywhere in the source files.

13.2 Difficult cases

In most cases, the optimal branch can be chosen based on the CPUID information about supported instruction sets, cache size, etc. There are some cases, however, where there are different ways of doing the same thing and the CPUID instruction doesn't give the necessary information about which implementation is best. These cases are usually dealt with in assembly language. Here are some examples:

- `strlen` function. The string length function scans a string of bytes to find the first byte of zero. A good implementation uses XMM registers to test 16 bytes at a time and afterwards a `BSF` (bit scan forward) instruction to localize the first byte of zero within a block of 16 bytes. Some CPUs have particularly slow implementations of the bit scan instruction. Programmers that have tested the `strlen` function in isolation have been unsatisfied with the performance of this function on CPUs with a slow bit scan instruction and have implemented a separate version for specific CPU models. However, we must consider that the bit scan instruction is executed only once for each function call so that you have to call the function billions of time before the performance even matters, which few programs do. Hence, it is hardly worth the effort to make special versions of the `strlen` function for CPUs with slow bit scan instructions. My recommendation is to use the bit scan instruction and expect this to be the fastest solution on future CPUs.

- Half size execution units. The size of vector registers has been increased from 64-bit MMX to 128-bit XMM and now 256-bit YMM registers. The first processors that supported 128-bit vector registers had in fact only 64-bit execution units. Each 128-

bit operation was split into two 64-bit operations so that there was hardly any speed advantage in using the larger vector size. Later models had the full 128-bit execution units and hence higher speed. In the same way, we may see some processors that handle 256-bit operations as two 128-bit operations, and similarly for further expansions of the register size in the future. There are situations where a vector implementation is optimal only on CPUs with full-size execution units. The problem is that it is difficult for the CPU dispatcher to know whether the largest vector register size is handled at half speed or full speed. I would recommend to use the largest available register size anyway, expecting that future models will handle the large registers at full speed. If this solution is not satisfactory then you may make a list of CPUs on which it is optimal to use the smaller register size and then use the larger register size for all processors that are not on this list. The advantage of making a list of processors that need the inferior implementation rather than a list of processors that need the best implementation is that it is less likely that the list has to be updated when new processors are introduced on the market.

- High precision math. Libraries for high precision math allow addition of integers with a very large number of bits. This is usually done in a loop of ADC (add with carry) instructions where the carry bit must be saved from one iteration to the next. The carry bit can be saved either in the carry flag or in a register. If the carry bit is kept in the carry flag then the loop must rely on instructions that use the zero flag only (e.g. DEC, JNZ). This solution can incur a large delay due to the so-called partial flags stall because the CPU has problems separating the flags register into the carry and zero flags and joining them again, especially on Intel CPUs (See manual 3: The microarchitecture of Intel, AMD and VIA CPUs). This is one of the few cases where it makes sense to dispatch by CPU brand. The version that works best on the newest CPU of a particular brand is likely to be the optimal choice for future models of the same brand.

  Memory copying. There are several different ways of copying blocks of memory. These methods are discussed in manual 2: "Optimizing subroutines in assembly language", section 17.9: "Moving blocks of data", where it is also discussed which method is fastest on different processors. In a C++ program, you should choose an up-to-date function library that has a good implementation of the memcpy function. There are so many different cases for different microprocessors, different alignments and different sizes of the data block to copy that the only reasonable solution is to have a standard function library that takes care of the CPU dispatching. This function is so important and generally used that most function libraries have CPU dispatching for this function, though not all libraries have the best and most up-to-date solution. The compiler is likely to use the memcpy function implicitly when copying a large object, unless there is a copy constructor specifying otherwise.

In difficult cases like these, it is important to remember that your code is likely to run most of its time on processors that were unknown at the time it was programmed. Therefore, it is important to consider which method is likely to work best on future processors, and choose this method for all unknown processors that support the necessary instruction set. It may not be worth the effort to make a CPU dispatcher based on complicated criteria or lists of specific CPU models if the problem is likely to go away in the future due to general improvements in microprocessor hardware design.

The ultimate solution would be to include a performance test that measures the speed of each version of the critical code to see which solution is optimal on the actual processor. However, this involves the problems that the clock frequency may vary dynamically and that measurements are unstable due to interrupts and task switches; so that it is necessary to test the different versions alternatingly several times in order to make a reliable decision.

13.3 Test and maintenance

There are two things to test when software uses CPU dispatching:

1. How much you gain in speed by using a particular code version.

2. Check that all code versions work correctly.

The speed test should preferably be done on the type of CPU that each particular branch of code is intended for. In other words, you need to test on several different CPUs if you want to optimize for several different CPUs.

On the other hand, it is not necessary to have many different CPUs to verify that all code branches works correctly. A code branch for a low instruction set can still run on a CPU with a higher instruction set. Therefore, you only need a CPU with the highest instruction set in order to test all branches for correctness. It is therefore recommended to put a test feature into the code that allows you to override the CPU dispatching and run any code branch for test purposes.

If the code is implemented as a function library or a separate module then it is convenient to make a test program that can call all code branches separately and test their functionality. This will be very helpful for later maintenance. However, this is not a textbook on test theory. Advice on how to test a software module for correctness must be found elsewhere.

13.4 Implementation

The CPU dispatch mechanism can be implemented in different places making the dispatch decision at different times:

- Dispatch on every call. A branch tree or switch statement leads to the appropriate version of the critical function. The branching is done every time the critical function is called. This has the disadvantage that the branching takes time.

- Dispatch on first call. The function is called through a function pointer which initially points to a dispatcher. The dispatcher changes the function pointer and makes it point to the right version of the function. This has the advantage that it doesn't spend time on deciding which version to use in case the function is never called. This method is illustrated in example 13.1 below.

- Make pointer at initialization. The program or library has an initialization routine that is called before the first call to the critical function. The initialization routine sets a function pointer to the right version of the function. This has the advantage that the response time is consistent for the function call.

- Load library at initialization. Each code version is implemented in a separate dynamic link library (`*.dll` or `*.so`). The program has an initialization routine that loads the appropriate version of the library. This method is useful if the library is very large or if different versions have to be compiled with different compilers.

- Dispatch at load time. The program uses a procedure linkage table (PLT) that is initialized when the program is loaded. This method requires OS support. It is planned to be available in Linux and perhaps Mac OS. See page 131 below.

- Dispatch at installation time. Each code version is implemented in a separate dynamic link library (`*.dll` or `*.so`). The installation program makes a symbolic link to the appropriate version of the library. The application program loads the library through the symbolic link.

- Use different executables. This method can be used if instruction sets are mutually incompatible. You may make separate executables for 32-bit and 64-bit systems. The appropriate version of the program may be selected during the installation process or by an .exe file stub.

If different versions of the critical code are compiled with different compilers then it is recommended to specify static linking for any library functions called by the critical code so that you don't have to distribute all the dynamic libraries (`*.dll` or `*.so`) that belong to each compiler with the application.

The availability of various instruction sets can be determined with system calls (e.g. `IsProcessorFeaturePresent` in Windows). Alternatively, you may call the CPUID instruction directly, or use the CPU detection function that I have supplied in the library www.agner.org/optimize/asmlib.zip. The name of this function is `InstructionSet()`. The following example shows how to implement the *dispatch on first call* method using `InstructionSet()`:

```
// Example 13.1
// CPU dispatching on first call

// Header file for InstructionSet()
#include "asmlib.h"

// Define function type with desired parameters
typedef int CriticalFunctionType(int parm1, int parm2);

// Function prototype
CriticalFunctionType CriticalFunction_Dispatch;

// Function pointer serves as entry point.
// After first call it will point to the appropriate function version
CriticalFunctionType * CriticalFunction = &CriticalFunction_Dispatch;

// Lowest version
int CriticalFunction_386(int parm1, int parm2) {...}

// SSE2 version
int CriticalFunction_SSE2(int parm1, int parm2) {...}

// AVX version
int CriticalFunction_AVX(int parm1, int parm2) {...}

// Dispatcher. Will be called only first time
int CriticalFunction_Dispatch(int parm1, int parm2)
{
   // Get supported instruction set, using asmlib library
   int level = InstructionSet();

   // Set pointer to the appropriate version (May use a table
   // of function pointers if there are many branches):
   if (level >= 11)
   {  // AVX supported
      CriticalFunction = &CriticalFunction_AVX;
   }
   else if (level >= 4)
   {  // SSE2 supported
      CriticalFunction = &CriticalFunction_SSE2;
   }
   else
   {  // Generic version
      CriticalFunction = &CriticalFunction_386;
   }
```

```
    // Now call the chosen version
    return (*CriticalFunction)(parm1, parm2);
}

int main()
{
    int a, b, c;
    ...

    // Call critical function through function pointer
    a = (*CriticalFunction)(b, c);


    ...
    return 0;
}
```

The `InstructionSet()` function is available in the function library [asmlib](#), which is available in different versions for different compilers. This function is OS independent and checks both the CPU and the operating system for support of the different instructions sets. The different versions of `CriticalFunction` in example 13.1 can be placed in separate modules if necessary, each compiled for the specific instruction set.

13.5 CPU dispatching in Gnu compiler

An "indirect function" feature has been announced for Linux and supported by the Gnu C/C++ compiler. This feature is intended for CPU dispatching and is planned to be used in the Gnu C library. It requires support from both compiler, linker and loader (binutils version 2.20, glibc version 2.11 ifunc branch).

The syntax for the Gnu compiler is illustrated in the following example with the same functionality as example 13.1:

```
    // Example 13.2. CPU dispatching in Gnu compiler

    // Header file for InstructionSet()
    #include "asmlib.h"

    // Lowest version
    int CriticalFunction_386(int parm1, int parm2) {...}

    // SSE2 version
    int CriticalFunction_SSE2(int parm1, int parm2) {...}

    // AVX version
    int CriticalFunction_AVX(int parm1, int parm2) {...}

    // Dispatcher. Will be called only once
    int __attribute__ ((ifunc)) CriticalFunction (int parm1, int parm2)
    {
        // This "ifunc" is not a function with two integer parameters
        // returning an integer, but a function without parameters
        // returning a pointer to such a function.

        // Get supported instruction set, using asmlib library
        int level = InstructionSet();

        // Set pointer to the appropriate version (May use a table
        // of function pointers if there are many branches):
        if (level >= 11)
        {   // AVX supported
            return &CriticalFunction_AVX;
```

```
   }
   if (level >= 4)
   {  // SSE2 supported
      return &CriticalFunction_SSE2;
   }
   // Default version
   return &CriticalFunction_386;
}

int main()
{
   int a, b, c;
   ...

   // Call critical function
   a = CriticalFunction(b, c);

   ...
   return 0;
}
```

All calls to `CriticalFunction` go through a procedure linkage table (PLT). The "`ifunc`" CPU dispatcher is called only once, setting the PLT to the desired version of the function. This is done either on the first call (for static linking) or at load time (for shared objects).

Note that this information is preliminary. I have not yet found a compiler version where it works (June 2011). Source: http://gcc.gnu.org/bugzilla/show_bug.cgi?id=40528.

13.6 CPU dispatching in Intel compiler

Intel compilers have a feature for making multiple versions of a function for different Intel CPUs. It uses the *dispatch on every call* method. When the function is called, a dispatch is made to the right version of the function. The automatic dispatching can be made for all suitable functions in a module by compiling the module with, e.g. the option `/QaxAVX` or `-axAVX`. This will make multiple versions even of functions that are not critical. It is possible to do the dispatching only for speed-critical functions by using the directive `__declspec(cpu_dispatch(...))`. See the Intel C++ Compiler Documentation for details. Note that the CPU dispatch mechanism in the Intel compiler works only for Intel CPUs, not for other brands of CPUs such as AMD and VIA. The next section (page 134) shows a way to work around this limitation and other flaws in the CPU detection mechanism.

The CPU dispatch mechanism in the Intel compiler is less efficient than the Gnu compiler mechanism because it makes dispatching on every call of the critical function. The Intel mechanism executes a series of branches every time the function is called, while the Gnu mechanism stores a pointer to the desired version in a procedure linkage table. If a dispatched function calls another dispatched function then the dispatch branch of the latter is executed even though the CPU-type is already known at this place. This can be avoided by inlining the latter function, but it may be better to do the CPU dispatching explicitly as in example 13.1 page 130.

The Intel compilers and function libraries have features for automatic CPU dispatching. Many Intel library functions have two or more versions for different processors and instruction sets. Likewise, the compiler can automatically generate multiple versions of the user-written code with automatic CPU dispatching.

Unfortunately, the CPU detection mechanism in Intel compilers has several flaws:

- The best possible version of the code is chosen only when running on an Intel processor. The CPU dispatcher checks whether the processor is an Intel before it checks which instruction set it supports. An inferior version of the code is selected if the processor is not an Intel, even if the processor is compatible with a better version of the code. This can lead to a dramatic degradation of performance on AMD and VIA processors.

- Explicit CPU dispatching works only with Intel processors. A non-Intel processor makes the dispatcher signal an error simply by performing an illegal operation that crashes the program.

- The CPU dispatcher does not check if XMM registers are supported by the operating system. It will crash on old operating systems that do not support SSE.

Several function libraries published by Intel have similar CPU dispatch mechanisms, and some of these also treat non-Intel CPUs in a suboptimal way.

The fact that the Intel CPU dispatcher treats non-Intel CPUs in a suboptimal way has become a serious legal issue. See my blog for details.

The behavior of the Intel compiler puts the programmer in a bad dilemma. You may prefer to use the Intel compiler because it has many advanced optimizing features, and you may want to use the well optimized Intel function libraries, but who would like to put a tag on his program saying that it doesn't work well on non-Intel machines?

Possible solutions to this problem are the following:

- Compile for a specific instruction set, e.g. `/arch:SSE2`. The compiler will produce the optimal code for this instruction set and insert only the SSE2 version of most library functions without CPU dispatching. Test if the program runs satisfactorily on a non-Intel CPU. If not, then it may be necessary to replace the CPU detection function as described below. The program will not be compatible with old microprocessors that do not have SSE2.

- Make two or more versions of the most critical part of the code and compile them separately with the appropriate instruction set specified. Insert an explicit CPU dispatching in the code to call the version that fits the microprocessor it is running on.

- Replace the CPU detection function of the Intel compiler with another function with the same name. This method is described below.

- Make calls directly to the CPU-specific versions of the library functions. The CPU-specific functions have names with the following suffixes: generic `.A`, SSE2 `.J`, SSE3 `.L`, SSE4.1 `.N`, AVX `.R`. The dot in the function name is not allowed in C++ so you need to use objconv or a similar utility for modifying the name in the object file.

- The ideal solution would be an open source library of well-optimized functions with a performance that can compete with Intel's libraries and with support for multiple platforms and multiple instruction sets. I have no knowledge of any such library.

The performance on non-Intel processors can be improved by using one or more of the above methods if the most time-consuming part of the program contains automatic CPU dispatching or memory-intensive functions such as `memcpy`, `memmove`, `memset`, or mathematical functions such as `pow`, `log`, `exp`, `sin`, etc.

Overriding the Intel CPU detection function

It is possible to replace the CPU detection function in Intel function libraries and compiler-generated code by making another function that has exactly the same name. If the static version of the function libraries is used then a function in a C or C++ source file will take precedence over a library function with the same name. The code examples below show how to override the Intel CPU detection functions in code generated by the Intel compiler as well as code that uses Intel function libraries. This works even for code that uses Intel function libraries with a different compiler, for example the Intel short vector math library (SVML) with the Gnu compiler. These code examples can be used for improving performance and compatibility with non-Intel microprocessors, microcode and old operating systems.

These examples are ready to copy and paste into your code. Remember to use static linking (e.g. option `/MT`). A function in a dynamic library (`*.dll`, `*.so`) cannot be changed without disassembling and modifying the library. These code examples below use the `InstructionSet` function from my asmlib library (www.agner.org/optimize/asmlib.zip).

Note that these examples are based on my own research, not on publicly available information. They have worked well in all tests on compiler versions 7 through 12, but of course we cannot be sure that they will work with future versions of the compiler and libraries. The examples are intended to work in both Windows and Linux, 32-bit and 64-bit. They have not been tested in Mac systems.

CPU dispatcher for Intel compiler and libraries

Example 13.3 overrides the Intel CPU dispatcher for code generated by the Intel compiler, as well as for code that uses the general Intel function libraries (e.g. `libirc`), the math library (`libm`), and the short vector math library (`svml`).

```
// Example 13.3. Dispatcher-patch for Intel compiler
#include "asmlib.h" // Header file for asmlib library

#ifdef __cplusplus
extern "C" {        // Avoid C++ name mangling
#endif

// Global variable indicating cpu
int __intel_cpu_indicator = 0;

// CPU dispatcher function
void __intel_cpu_indicator_init() {
   // Get CPU level from asmlib library function
   switch (InstructionSet()) {
   case 0:  // No special instruction set supported
      __intel_cpu_indicator = 1;
      break;
   case 1: case 2: // MMX supported
      __intel_cpu_indicator = 8;
      break;
   case 3:  // SSE supported
      __intel_cpu_indicator = 0x80;
      break;
   case 4:  // SSE2 supported
      __intel_cpu_indicator = 0x200;
      break;
   case 5:  // SSE3 supported
      __intel_cpu_indicator = 0x800;
      break;
   case 6: case 7:  // Supplementary-SSE3 supported
      __intel_cpu_indicator = 0x1000;
      break;
```

```
        case 8: case 9:  // SSE4.1 supported
            __intel_cpu_indicator = 0x2000;
            break;
        case 10: case 11: // SSE4.2 and POPCNT supported
            __intel_cpu_indicator = 0x8000;
            break;
        case 12: // AVX, PCLMUL and AES supported
        default:
            __intel_cpu_indicator = 0x20000;
            break;
    }
#ifdef __cplusplus
}   // End of extern "C"
#endif
```

This applies to Intel C++ compiler version 12 and earlier and to the libraries that are included with this compiler on all x86 and x86-64 platforms. It may or may not work in future versions.

Programs compiled with the Intel compiler with the `/Qax` option will run sub-optimally on non-Intel processors unless the above patch is included. Programs compiled with the `/Qx` option will not run at all on non-Intel processors unless the above patch is included. Programs compiled with the `/arch` option only will use the SSE2 version of most library functions without CPU dispatching, even when a higher instruction set is specified.

It is recommended to check with a debugger that the function `__intel_cpu_indicator_init()` is called at least once and that the code works correctly on both Intel and non-Intel processors. A complete list of the values of `__intel_cpu_indicator` is given in manual 5: "Calling conventions" in table 19.

CPU dispatcher for Intel Math Kernel Library

Example 13.4 overrides the Intel CPU dispatcher for the Intel math kernel library (`MKL`), except the vector math library (`VML`). This override may not be needed in 32-bit mode.

```
// Example 13.4.
// Dispatcher-patch for Intel math kernel library v. 10.x and later
#include "asmlib.h" // Header file for asmlib

#ifdef __cplusplus
extern "C" {         // Avoid C++ name mangling
#endif

// CPU dispatcher function for MKL
int mkl_serv_cpu_detect(void) {
    int n = InstructionSet(); // Get instruction set
    if (n < 4) return 0; // SSE2 not supported

    if (sizeof(void*) == 4) {
        // 32 bit mode
        switch (n) {
        case 4:  // SSE2 supported
            n = 2; break;
        case 5:  // SSE3 supported
            n = 3; break;
        case 6: case 7: case 8: case 9: // Supplementary-SSE3 supported
            n = 4; break;
        case 10: // SSE4.2 supported
            n = 5; break;
        default: // AVX
            n = 6; // 6 if library supports AVX (experimental in 10.2)
            // n = 5; // 5 for earlier versions
```

```
            break;
        }
    }
    else {
        // 64 bit mode
        switch (n) {
        case 4:  // SSE2 supported
            n = 0; break;
        case 5:  // SSE3 supported
            n = 1; break;
        case 6: case 7: case 8: case 9: // Supplementary-SSE3 supported
            n = 2; break;
        case 10: // SSE4.2 supported
            n = 3; break;
        default: // AVX
            n = 4; // 4 if MKL library supports AVX (version 10.3)
            // n = 3; // 3 for earlier versions
            break;
        }
    }
    return n;
}
#ifdef __cplusplus
}  // End of extern "C"
#endif
```

This code doesn't work if the return value is higher than the latest instruction set supported by the library. You may need to limit the return value if an earlier version of MKL is used. Later versions are likely to allow higher values.

CPU dispatcher for Intel Vector Math Library

Example 13.5 overrides the Intel CPU dispatcher for the Intel Vector Math Library (VML) and some other sub-libraries.

```
// Example 13.5.
// Dispatcher-patch for Intel VML v. 10.x and later
#include "asmlib.h" // Header file for asmlib

#ifdef __cplusplus
extern "C" {         // Avoid C++ name mangling
#endif

// CPU dispatcher function for VML etc.
int mkl_vml_serv_cpu_detect(void) {
    int n = InstructionSet();
    if (n < 4) return 0; // SSE2 not supported

    if (sizeof(void*) == 4) {
        // 32 bit mode
        switch (n) {
        case 4:  // SSE2 supported
            n = 2; break;
        case 5:  // SSE3 supported
            n = 3; break;
        case 6: case 7: // Supplementary-SSE3 supported
            n = 4; break;
        case 8: case 9: // SSE4.1 supported
            n = 5; break;
        case 10:  // SSE4.2 supported
            n = 6; break;
        default:  // AVX supported
            n = 7; // 7 if VML library supports AVX (version 10.3)
            // n = 6; // 6 for earlier versions
```

136

```
            break;
        }
    }
    else {
        // 64 bit mode
        switch (n) {
        case 4: case 5: // SSE2 supported
            n = 1; break;
        case 6: case 7: // Supplementary-SSE3 supported
            n = 2; break;
        case 8: case 9: // SSE4.1 supported
            n = 3; break;
        case 10:  // SSE4.2 supported
            n = 4; break;
        default:  // AVX supported
            n = 5; // 5 if VML library supports AVX (version 10.3)
            // n = 4; // 4 for earlier versions
            break;
        }
    }
    return n;
}

// Earlier versions use this name
int mkl_vml_service_cpu_detect(void) {
    return mkl_vml_serv_cpu_detect();
}

#ifdef __cplusplus
}  // End of extern "C"
#endif
```

This code has been tested with VML in MKL version 10.2 and 10.3 beta. It doesn't work if the return value is higher than the latest instruction set supported by the library. You may need to limit the return value if an earlier version of the library is used. Later versions are likely to allow higher values.

The Short Vector Math Library (SVML) needs the code in example 13.3, not example 13.5.

CPU dispatching in Intel Performance Primitives Library

The CPU dispatching in the Intel Performance Primitives library (IPP) version 6.1 appears to work well with non-Intel processors. No override appears to be needed.

Future versions may be checked by testing the external variable ippJumpIndexForMergedLibs after calling ippStaticInit(). This variable should have the following values: 0: < SSE2, 1: SSE2, 2: SSE3, 3: SSSE3, 4: SSE4.1, 5: Intel Atom. You may do the test in example 13.6 on a non-Intel processor.

```
// Example 13.6. Dispatcher test for Intel IPP
#include <ipp.h>
#include <stdio.h>

extern "C" {
    extern int ippJumpIndexForMergedLibs;
}

int main() {
    ippStaticInit();
    printf("\nJump index = %i\n", ippJumpIndexForMergedLibs);
    return 0;
}
```

14 Specific optimization tips

14.1 Use lookup tables

Reading a value from a table of constants is very fast if the table is cached. Usually it takes only a few clock cycles to read from a table in the level-1 cache. We can take advantage of this fact by replacing a function call with a table lookup if the function has only a limited number of possible inputs.

Let's take the integer factorial function (n!) as an example. The only allowed inputs are the integers from 0 to 12. Higher inputs give overflow and negative inputs give infinity. A typical implementation of the factorial function looks like this:

```
// Example 14.1a
int factorial (int n) {          // n!
   int i, f = 1;
   for (i = 2; i <= n; i++) f *= i;
   return f;
}
```

This calculation requires `n-1` multiplications, which can take quite a long time. It is more efficient to use a lookup table:

```
// Example 14.1b
int factorial (int n) {          // n!
   // Table of factorials:
   static const int FactorialTable[13] = {1, 1, 2, 6, 24, 120, 720,
       5040, 40320, 362880, 3628800, 39916800, 479001600};
   if ((unsigned int)n < 13) {   // Bounds checking (see page 140)
      return FactorialTable[n];  // Table lookup
   }
   else {
      return 0;                   // return 0 if out of range
   }
}
```

This implementation uses a lookup table instead of calculating the value each time the function is called. I have added a bounds check on `n` here because the consequence of `n` being out of range is possibly more serious when `n` is an array index than when `n` is a loop count. The method of bounds checking is explained below on page 140.

The table should be declared `const` in order to enable constant propagation and other optimizations. In most cases it is also recommended to declare the table `static`. This makes sure that the table is initialized when the program is loaded rather than each time the function is called. You may declare the function `inline`.

Replacing a function with a lookup table is advantageous in most cases where the number of possible inputs is limited and there are no cache problems. It is not advantageous to use a lookup table if you expect the table to be evicted from the cache between each call, and the time it takes to calculate the function is less than the time it takes to reload the value from memory plus the costs to other parts of the program of occupying a cache line.

Table lookup cannot be vectorized with the current instruction set. Do not use lookup tables if this prevents a faster vectorized code.

Storing something in static memory can cause caching problems because static data are likely to be scattered around at different memory addresses. If caching is a problem then it

may be useful to copy the table from static memory to stack memory outside the innermost loop. This is done by declaring the table inside a function but outside the innermost loop and without the `static` keyword:

```
// Example 14.1c
void CriticalInnerFunction () {
   // Table of factorials:
   const int FactorialTable[13] = {1, 1, 2, 6, 24, 120, 720,
      5040, 40320, 362880, 3628800, 39916800, 479001600};
   ...
   int i, a, b;
   // Critical innermost loop:
   for (i = 0; i < 1000; i++) {
      ...
      a = FactorialTable[b];
      ...
   }
}
```

The `FactorialTable` in example 14.1c is copied from static memory to the stack when `CriticalInnerFunction` is called. The compiler will store the table in static memory and insert a code that copies the table to stack memory at the start of the function. Copying the table takes extra time, of course, but this is permissible when it is outside the critical innermost loop. The loop will use the copy of the table that is stored in stack memory which is contiguous with other local variables and therefore likely to be cached more efficiently than static memory.

If you don't care to calculate the table values by hand and insert the values in the code then you may of course make the program do the calculations. The time it takes to calculate the table is not significant as long as it is done only once. One may argue that it is safer to calculate the table in the program than to type in the values because a typo in a hand-written table may go undetected.

The principle of table lookup can be used in any situation where a program chooses between two or more constants. For example, a branch that chooses between two constants can be replaced by a table with two entries. This may improve the performance if the branch is poorly predictable. For example:

```
// Example 14.2a
float a;  int b;
a = (b == 0) ? 1.0f : 2.5f;
```

If we assume that `b` is always 0 or 1 and that the value is poorly predictable, then it is advantageous to replace the branch by a table lookup:

```
// Example 14.2b
float a;  int b;
static const float OneOrTwo5[2] = {1.0f, 2.5f};
a = OneOrTwo5[b & 1];
```

Here, I have AND'ed `b` with `1` for the sake of security. `b & 1` is certain to have no other value than `0` or `1` (see page 141). This extra check on `b` can be omitted, of course, if the value of `b` is guaranteed to be `0` or `1`. Writing `a = OneOrTwo5[b!=0];` will also work, although slightly less efficiently. This method is inefficient, however, when `b` is a `float` or `double` because all the compilers I have tested implement `OneOrTwo5[b!=0]` as `OneOrTwo5[(b!=0) ? 1 : 0]` in this case so we don't get rid of the branch. It may seem illogical that the compiler uses a different implementation when `b` is floating point. The reason is, I guess, that compiler makers assume that floating point comparisons are more predictable than integer comparisons. The solution `a = 1.0f + b * 1.5f;` is efficient

when `b` is a `float`, but not if `b` is an integer because the integer-to-float conversion takes more time than the table lookup.

Lookup tables are particular advantageous as replacements for `switch` statements because `switch` statements often suffer from poor branch prediction. Example:

```
// Example 14.3a
int n;
switch (n) {
case 0:
   printf("Alpha");  break;
case 1:
   printf("Beta");   break;
case 2:
   printf("Gamma");  break;
case 3:
   printf("Delta");  break;
}
```

This can be improved by using a lookup table:

```
// Example 14.3b
int n;
static char const * const Greek[4] = {
    "Alpha", "Beta", "Gamma", "Delta"
};
if ((unsigned int)n < 4) { // Check that index is not out of range
   printf(Greek[n]);
}
```

The declaration of the table has `const` twice because both the pointers and the texts they point to are constant.

14.2 Bounds checking

In C++, it is often necessary to check if an array index is out of range. This may typically look like this:

```
// Example 14.4a
const int size = 16; int i;
float list[size];
...
if (i < 0 || i >= size) {
   cout << "Error: Index out of range";
}
else {
   list[i] += 1.0f;
}
```

The two comparisons `i < 0` and `i >= size` can be replaced by a single comparison:

```
// Example 14.4b
if ((unsigned int)i >= (unsigned int)size) {
   cout << "Error: Index out of range";
}
else {
   list[i] += 1.0f;
}
```

A possible negative value of `i` will appear as a large positive number when `i` is interpreted as an unsigned integer and this will trigger the error condition. Replacing two

comparisons by one makes the code faster because testing a condition is relatively expensive, while the type conversion generates no extra code at all.

This method can be extended to the general case where you want to check whether an integer is within a certain interval:

```
// Example 14.5a
const int min = 100, max = 110;  int i;
...
if (i >= min && i <= max) { ...
```

can be changed to:

```
// Example 14.5b
if ((unsigned int)(i - min) <= (unsigned int)(max - min)) { ...
```

There is an even faster way to limit the range of an integer if the length of the desired interval is a power of 2. Example:

```
// Example 14.6
float list[16]; int i;
...
list[i & 15] += 1.0f;
```

This needs a little explanation. The value of `i&15` is guaranteed to be in the interval from 0 to 15. If `i` is outside this interval, for example `i` = 18, then the `&` operator (bitwise and) will cut off the binary value of `i` to four bits, and the result will be 2. The result is the same as `i` modulo 16. This method is useful for preventing program errors in case the array index is out of range and we don't need an error message if it is. It is important to note that this method works only for powers of 2 (i.e. 2, 4, 8, 16, 32, 64, ...). We can make sure that a value is less than 2^n and not negative by AND'ing it with 2^n-1. The bitwise AND operation isolates the least significant n bits of the number and sets all other bits to zero.

14.3 Use bitwise operators for checking multiple values at once

The bitwise operators `&`, `|`, `^`, `~`, `<<`, `>>` can test or manipulate all the bits of an integer in one operation. For example, if each bit of a 32-bit integer has a particular meaning, then you can set multiple bits in a single operation using the `|` operator; you can clear or mask out multiple bits with the `&` operator; and you can toggle multiple bits with the `^` operator.

The `&` operator is also useful for testing multiple conditions in a single operation. Example:

```
// Example 14.7a. Testing multiple conditions
enum Weekdays {
    Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
};
Weekdays Day;
if (Day == Tuesday || Day == Wednesday || Day == Friday) {
    DoThisThreeTimesAWeek();
}
```

The `if` statement in this example has three conditions which are implemented as three branches. They can be joined into a single branch if the constants `Sunday`, `Monday`, etc. are defined as powers of 2:

```
// Example 14.7b. Testing multiple conditions using &
enum Weekdays {
    Sunday = 1, Monday = 2, Tuesday = 4, Wednesday = 8,
    Thursday = 0x10, Friday = 0x20, Saturday = 0x40
```

141

```
    };
    Weekdays Day;
    if (Day & (Tuesday | Wednesday | Friday)) {
        DoThisThreeTimesAWeek();
    }
```

By giving each constant a value that is a power of 2 in example 14.7b, we are in fact using each bit in `Day` for signifying one of the weekdays. The maximum number of constants we can define in this way is equal to the number of bits in an integer, usually 32. In 64-bit systems we can use 64-bit integers with hardly any loss of efficiency.

The expression `(Tuesday | Wednesday | Friday)` in example 14.7b is converted by the compiler to the value `0x2C` so that the `if` condition can be calculated by a single `&` operation, which is very fast. The result of the `&` operation will be non-zero, and therefore count as true, if any of the bits for `Tuesday`, `Wednesday` or `Friday` is set in the variable `Day`.

Note the difference between the Boolean operators `&&`, `||`, `!` and the corresponding bitwise operators `&`, `|`, `~`. The Boolean operators produce a single result, true (1) or false (0); and the second operand is evaluated only when needed. The bitwise operators produce 32 results when applied to 32-bit integers, and they always evaluate both operands. Nevertheless, the bitwise operators are calculated much faster than the Boolean operators because they do not use branches, provided that the operands are integer expressions rather than Boolean expressions.

There are lots of things you can do with bitwise operators using integers as Boolean vectors, and these operations are very fast. This can be useful in programs with many Boolean expressions. Whether the constants are defined with `enum`, `const`, or `#define` makes no difference for the performance.

14.4 Integer multiplication

Integer multiplication takes longer time than addition and subtraction (3 - 10 clock cycles, depending on the processor). Optimizing compilers will often replace integer multiplication by a constant with a combination of additions and shift operations. Multiplying by a power of 2 is faster than multiplying by other constants because it can be done as a shift operation. For example, `a * 16` is calculated as `a << 4`, and `a * 17` is calculated as `(a << 4) + a`.

You can take advantage of this by preferably using powers of 2 when multiplying with a constant. The compilers also have fast ways of multiplying by 3, 5 and 9.

Multiplications are done implicitly when calculating the address of an array element. In some cases this multiplication will be faster when the factor is a power of 2. Example:

```
// Example 14.8
const int rows = 10, columns = 8;
float matrix[rows][columns];
int i, j;
int order(int x);
...
for (i = 0; i < rows; i++) {
    j = order(i);
    matrix[j][0] = i;
}
```

Here, the address of `matrix[j][0]` is calculated internally as `(int)&matrix[0][0] + j * (columns * sizeof(float))`.

Now, the factor to multiply `j` by is `(columns * sizeof(float)) = 8 * 4 = 32`. This is a power of 2, so the compiler can replace `j * 32` with `j << 5`. If `columns` had not been a power of 2 then the multiplication would take longer time. It can therefore be advantageous to make the number of columns in a matrix a power of 2 if the rows are accessed in a non-sequential order.

The same applies to an array of structure or class elements. The size of each object should preferably be a power of 2 if the objects are accessed in a non-sequential order. Example:

```
// Example 14.9
struct S1 {
    int a;
    int b;
    int c;
    int UnusedFiller;
};
int order(int x);
const int size = 100;
S1 list[size];  int i, j;
...
for (i = 0; i < size; i++) {
    j = order(i);
    list[j].a = list[j].b + list[j].c;
}
```

Here, we have inserted `UnusedFiller` in the structure to make sure its size is a power of 2 in order to make the address calculation faster.

The advantage of using powers of 2 applies only when elements are accessed in non-sequential order. If the code in example 14.8 and 14.9 is changed so that it has `i` instead of `j` as index then the compiler can see that the addresses are accessed in sequential order and it can calculate each address by adding a constant to the preceding one (see page 71). In this case it doesn't matter if the size is a power of 2 or not.

The advise of using powers of 2 does not apply to very big data structures. On the contrary, you should by all means avoid powers of 2 if a matrix is so big that caching becomes a problem. If the number of columns in a matrix is a power of 2 and the matrix is bigger than the cache then you can get very expensive cache contentions, as explained on page 96.

14.5 Integer division

Integer division takes much longer time than addition, subtraction and multiplication (27 - 80 clock cycles for 32-bit integers, depending on the processor).

Integer division by a power of 2 can be done with a shift operation, which is much faster.

Division by a constant is faster than division by a variable because optimizing compilers can compute a / b as $a * (2^n / b) >> n$ with a suitable choice of n. The constant $(2^n / b)$ is calculated in advance and the multiplication is done with an extended number of bits. The method is somewhat more complicated because various corrections for sign and rounding errors must be added. This method is described in more detail in manual 2: "Optimizing subroutines in assembly language". The method is faster if the dividend is unsigned.

The following guidelines can be used for improving code that contains integer division:

* Integer division by a constant is faster than division by a variable

* Integer division by a constant is faster if the constant is a power of 2

- Integer division by a constant is faster if the dividend is unsigned

Examples:

```
// Example 14.10
int a, b, c;
a = b / c;                // This is slow
a = b / 10;               // Division by a constant is faster
a = (unsigned int)b / 10; // Still faster if unsigned
a = b / 16;               // Faster if divisor is a power of 2
a = (unsigned int)b / 16; // Still faster if unsigned
```

The same rules apply to modulo calculations:

```
// Example 14.11
int a, b, c;
a = b % c;                // This is slow
a = b % 10;               // Modulo by a constant is faster
a = (unsigned int)b % 10; // Still faster if unsigned
a = b % 16;               // Faster if divisor is a power of 2
a = (unsigned int)b % 16; // Still faster if unsigned
```

You can take advantage of these guidelines by using a constant divisor that is a power of 2 if possible and by changing the dividend to unsigned if you are sure that it will not be negative.

Division of a loop counter by a constant can be avoided by rolling out the loop by the same constant. Example:

```
// Example 14.12a
int list[300];
int i;
for (i = 0;  i < 300;  i++) {
   list[i] += i / 3;
}
```

This can be replaced with:

```
// Example 14.12b
int list[300];
int i, i_div_3;
for (i = i_div_3 = 0;  i < 300;  i += 3, i_div_3++) {
   list[i]   += i_div_3;
   list[i+1] += i_div_3;
   list[i+2] += i_div_3;
}
```

A similar method can be used to avoid modulo operations:

```
// Example 14.13a
int list[300];
int i;
for (i = 0;  i < 300;  i++) {
   list[i] = i % 3;
}
```

This can be replaced with:

```
// Example 14.13b
int list[300];
int i;
for (i = 0;  i < 300;  i += 3) {
```

```
    list[i]   = 0;
    list[i+1] = 1;
    list[i+2] = 2;
}
```

The loop unrolling in example 14.12b and 14.13b works only if the loop count is divisible by the unroll factor. If not, then you must do the extra operations outside the loop:

```
// Example 14.13c
int list[301];
int i;
for (i = 0;  i < 301;  i += 3) {
    list[i]   = 0;
    list[i+1] = 1;
    list[i+2] = 2;
}
list[300] = 0;
```

14.6 Floating point division

Floating point division takes much longer time than addition, subtraction and multiplication (20 - 45 clock cycles).

Floating point division by a constant should be done by multiplying with the reciprocal:

```
// Example 14.14a
double a, b;
a = b / 1.2345;
```

Change this to:

```
// Example 14.14b
double a, b;
a = b * (1. / 1.2345);
```

The compiler will calculate `(1./1.2345)` at compile time and insert the reciprocal in the code, so you will never spend time doing the division. Some compilers will replace the code in example 14.14a with 14.14b automatically but only if certain options are set to relax floating point precision (see page 73). It is therefore safer to do this optimization explicitly.

Divisions can sometimes be eliminated completely. For example:

```
// Example 14.15a
if (a > b / c)
```

can sometimes be replaced by

```
// Example 14.15b
if (a * c > b)
```

But beware of the pitfalls here: The inequality sign must be reversed if $c < 0$. The division is inexact if b and c are integers, while the multiplication is exact.

Multiple divisions can be combined. For example:

```
// Example 14.16a
double y, a1, a2, b1, b2;
y = a1/b1 + a2/b2;
```

Here we can eliminate one division by making a common denominator:

```
// Example 14.16b
double y, a1, a2, b1, b2;
y = (a1*b2 + a2*b1) / (b1*b2);
```

The trick of using a common denominator can even be used on completely independent
divisions. Example:

```
// Example 14.17a
double a1, a2, b1, b2, y1, y2;
y1 = a1 / b1;
y2 = a2 / b2;
```

This can be changed to:

```
// Example 14.17b
double a1, a2, b1, b2, y1, y2, reciprocal_divisor;
reciprocal_divisor = 1. / (b1 * b2);
y1 = a1 * b2 * reciprocal_divisor;
y2 = a2 * b1 * reciprocal_divisor;
```

14.7 Don't mix float and double

Floating point calculations usually take the same time regardless of whether you are using
single precision or double precision, but there is a penalty for mixing single and double
precision in programs compiled for 64-bit operating systems and programs compiled for the
instruction set SSE2 or later. Example:

```
// Example 14.18a
float a, b;
a = b * 1.2;        // Mixing float and double is bad
```

The C/C++ standard specifies that all floating point constants are double precision by
default, so `1.2` in this example is a double precision constant. It is therefore necessary for
the compiler to convert `b` from single precision to double precision before multiplying with
the double precision constant and then convert the result back to single precision. These
conversions take a lot of time. You can avoid the conversions and make the code up to 5
times faster either by making the constant single precision or by making `a` and `b` double
precision:

```
// Example 14.18b
float a, b;
a = b * 1.2f;      // everything is float

// Example 14.18c
double a, b;
a = b * 1.2;       // everything is double
```

There is no penalty for mixing different floating point precisions when the code is compiled
for old processors without the SSE2 instruction set, but it may be preferable to keep the
same precision in all operands in case the code is later ported to another platform.

14.8 Conversions between floating point numbers and integers

Conversion from floating point to integer

According to the standards for the C++ language, all conversions from floating point
numbers to integers use truncation towards zero, rather than rounding. This is unfortunate
because truncation takes much longer time than rounding unless the SSE2 instruction set is

146

used. It is recommended to enable the SSE2 instruction set if possible. SSE2 is always enabled in 64-bit mode.

A conversion from floating point to integer without SSE2 typically takes 40 clock cycles. If you cannot avoid conversions from `float` or `double` to `int` in the critical part of the code, then you may improve efficiency by using rounding instead of truncation. This is approximately three times faster. The logic of the program may need modification to compensate for the difference between rounding and truncation.

Efficient conversion from `float` or `double` to integer can be done with the functions `lrintf` and `lrint`. Unfortunately, these functions are missing in many commercial compilers due to controversies over the C99 standard. An implementation of the `lrint` function is given in example 14.19 below. The function rounds a floating point number to the nearest integer. If two integers are equally near then the even integer is returned. There is no check for overflow. This function is intended for 32-bit Windows and 32-bit Linux with Microsoft, Intel and Gnu compilers.

```
// Example 14.19
static inline int lrint (double const x) { // Round to nearest integer
    int n;
#if defined(__unix__) || defined(__GNUC__)
    // 32-bit Linux, Gnu/AT&T syntax:
    __asm ("fldl %1 \n fistpl %0 " : "=m"(n) : "m"(x) : "memory" );
#else
    // 32-bit Windows, Intel/MASM syntax:
    __asm fld qword ptr x;
    __asm fistp dword ptr n;
#endif
    return n;}
```

This code will work only on Intel/x86-compatible microprocessors. The function is also available in the function library at www.agner.org/optimize/asmlib.zip.

The following example shows how to use the `lrint` function:

```
// Example 14.20
double d = 1.6;
int a, b;
a = (int)d;        // Truncation is slow. Value of a will be 1
b = lrint(d);      // Rounding is fast. Value of b will be 2
```

In 64-bit mode or when the SSE2 instruction set is enabled there is no difference in speed between rounding and truncation. The missing functions can be implemented as follows in 64-bit mode or when the SSE2 instruction set is enabled:

```
// Example 14.21.  // Only for SSE2 or x64
#include <emmintrin.h>

static inline int lrintf (float const x) {
    return _mm_cvtss_si32(_mm_load_ss(&x));}

static inline int lrint (double const x) {
    return _mm_cvtsd_si32(_mm_load_sd(&x));}
```

The code in example 14.21 is faster than other methods of rounding, but neither faster nor slower than truncation when the SSE2 instruction set is enabled.

Conversion from integer to floating point

Conversion of integers to floating point is faster than from floating point to integer. The conversion time is typically between 5 and 20 clock cycles. It may in some cases be

advantageous to do simple integer calculations in floating point variables in order to avoid conversions from integer to floating point.

Conversion of unsigned integers to floating point numbers is less efficient than signed integers. It is more efficient to convert unsigned integers to signed integers before conversion to floating point if the conversion to signed integer doesn't cause overflow. Example:

```
// Example 14.22a
unsigned int u;  double d;
d = u;
```

If you are certain that $u < 2^{31}$ then convert it to signed before converting to floating point:

```
// Example 14.22b
unsigned int u;  double d;
d = (double)(signed int)u;
```

14.9 Using integer operations for manipulating floating point variables

Floating point numbers are stored in a binary representation according to the IEEE standard 754 (1985). This standard is used in almost all modern microprocessors and operating systems (but not in some very old DOS compilers).

The representation of `float`, `double` and `long double` reflects the floating point value written as $\pm 2^{eee} \cdot 1.\text{fffff}$, where $\pm$ is the sign, `eee` is the exponent, and `fffff` is the binary decimals of the fraction. The sign is stored as a single bit which is 0 for positive and 1 for negative numbers. The exponent is stored as a biased binary integer, and the fraction is stored as the binary digits. The exponent is always normalized, if possible, so that the value before the decimal point is 1. This '1' is not included in the representation, except in the `long double` format. The formats can be expressed as follows:

```
struct Sfloat {
   unsigned int fraction : 23; // fractional part
   unsigned int exponent :  8; // exponent + 0x7F
   unsigned int sign     :  1; // sign bit
};

struct Sdouble {
   unsigned int fraction : 52; // fractional part
   unsigned int exponent : 11; // exponent + 0x3FF
   unsigned int sign     :  1; // sign bit
};

struct Slongdouble {
   unsigned int fraction : 63; // fractional part
   unsigned int one      :  1; // always 1 if nonzero and normal
   unsigned int exponent : 15; // exponent + 0x3FFF
   unsigned int sign     :  1; // sign bit
};
```

The values of nonzero floating point numbers can be calculated as follows:

$$floatvalue = (-1)^{sign} \cdot 2^{exponent-127} \cdot \left(1 + fraction \cdot 2^{-23}\right),$$
$$doublevalue = (-1)^{sign} \cdot 2^{exponent-1023} \cdot \left(1 + fraction \cdot 2^{-52}\right),$$
$$longdoublevalue = (-1)^{sign} \cdot 2^{exponent-16383} \cdot \left(one + fraction \cdot 2^{-63}\right).$$

The value is zero if all bits except the sign bit are zero. Zero can be represented with or without the sign bit.

The fact that the floating point format is standardized allows us to manipulate the different parts of the floating point representation directly with the use of integer operations. This can be an advantage because integer operations are faster than floating point operations. You should use such methods only if you are sure you know what you are doing. See the end of this section for some caveats.

We can change the sign of a floating point number simply by inverting the sign bit:

```
// Example 14.23
union {
    float f;
    int i;
} u;
u.i ^= 0x80000000; // flip sign bit of u.f
```

We can take the absolute value by setting the sign bit to zero:

```
// Example 14.24
union {
    float f;
    int i;
} u;
u.i &= 0x7FFFFFFF; // set sign bit to zero
```

We can check if a floating point number is zero by testing all bits except the sign bit:

```
// Example 14.25
union {
    float f;
    int i;
} u;
if (u.i & 0x7FFFFFFF) { // test bits 0 - 30
    // f is nonzero
}
else {
    // f is zero
}
```

We can multiply a nonzero floating point number by 2^n by adding n to the exponent:

```
// Example 14.26
union {
    float f;
    int i;
} u;
int n;
if (u.i & 0x7FFFFFFF) { // check if nonzero
    u.i += n << 23;      // add n to exponent
}
```

Example 14.26 does not check for overflow and works only for positive n. You can divide by 2^n by subtracting n from the exponent if there is no risk of underflow.

The fact that the representation of the exponent is biased allows us to compare two positive floating point numbers simply by comparing them as integers:

```
// Example 14.27
union {
```

```
    float f;
    int i;
} u, v;
if (u.i > v.i) {
    // u.f > v.f if both positive
}
```

Example 14.27 assumes that we know that `u.f` and `v.f` are both positive. It will fail if both are negative or if one is 0 and the other is -0 (zero with sign bit set).

We can shift out the sign bit to compare absolute values:

```
// Example 14.28
union {
    float f;
    unsigned int i;
} u, v;
if (u.i * 2 > v.i * 2) {
    // abs(u.f) > abs(v.f)
}
```

The multiplication by 2 in example 14.28 will shift out the sign bit so that the remaining bits represent a monotonically increasing function of the absolute value of the floating point number.

We can convert an integer in the interval $0 <= n < 2^{23}$ to a floating point number in the interval [1.0, 2.0) by setting the fraction bits:

```
// Example 14.29
union {
    float f;
    int i;
} u;
int n;
u.i = (n & 0x7FFFFF) | 0x3F800000; // Now 1.0 <= u.f < 2.0
```

This method is useful for random number generators.

In general, it is faster to access a floating point variable as an integer if it is stored in memory, but not if it is a register variable. The union forces the variable to be stored in memory, at least temporarily. Using the methods in the above examples will therefore be a disadvantage if other nearby parts of the code could benefit from using registers for the same variables.

In these examples we are using unions rather than type casting of pointers because this method is safer. Type casting of pointers may not work on compilers that rely on the strict aliasing rule of standard C, specifying that pointers of different types cannot point to the same object, except for `char` pointers.

The above examples all use single precision. Using double precision in 32-bit systems gives rise to some extra complications. A double is represented with 64 bits, but 32-bit systems do not have inherent support for 64-bit integers. Many 32-bit systems allow you to define 64-bit integers, but they are in fact represented as two 32-bit integers, which is less efficient. You may use the upper 32 bits of a `double` which gives access to the sign bit, the exponent, and the most significant part of the fraction. For example, to test the sign of a double:

```
// Example 14.23b
union {
    double d;
    int i[2];
```

```
    } u;
    if (u.i[1] < 0) {   // test sign bit
        // u.d is negative or -0
    }
```

It is not recommended to modify a double by modifying only half of it, for example if you want to flip the sign bit in the above example with `u.i[1] ^= 0x80000000;` because this is likely to generate a store forwarding delay in the CPU (See manual 3: "The microarchitecture of Intel, AMD and VIA CPUs"). This can be avoided in 64-bit systems by using a 64-bit integer rather than two 32-bit integers to alias upon the double.

Another problem with accessing 32 bits of a 64-bit double is that it is not portable to systems with big-endian storage. Example 14.23b and 14.30 will therefore need modification if implemented on other platforms with big-endian storage. All x86 platforms (Windows, Linux, BSD, Intel-based Mac OS, etc.) have little-endian storage, but other systems may have big endian storage (e.g. PowerPC).

We can make an approximate comparison of doubles by comparing bits 32-62. This can be useful for finding the numerically largest element in a matrix for use as pivot in a Gauss elimination. The method in example 14.28 can be implemented like this in a pivot search:

```
// Example 14.30
const int size = 100;
// Array of 100 doubles:
union {double d; unsigned int u[2]} a[size];
unsigned int absvalue, largest_abs = 0;
int i, largest_index = 0;
for (i = 0; i < size; i++) {
    // Get upper 32 bits of a[i] and shift out sign bit:
    absvalue = a[i].u[1] * 2;
    // Find numerically largest element (approximately):
    if (absvalue > largest_abs) {
        largest_abs = absvalue;
        largest_index = i;
    }
}
```

Example 14.30 finds the numerically largest element in an array, or approximately so. It may fail to distinguish elements with a relative difference less than 2^{-20}, but this is sufficiently accurate for the purpose of finding a suitable pivot element. The integer comparison is likely to be faster than a floating point comparison. On big endian systems you have to replace `u[1]` by `u[0]`.

14.10 Mathematical functions

The most common mathematical functions such as logarithms, exponential functions, trigonometric functions, etc. are implemented in hardware in the x86 CPUs. However, a software implementation is faster than the hardware implementation in most cases when the SSE2 instruction set is available. The best compilers use the software implementation if the SSE2 instruction set is enabled.

The advantage of using a software implementation rather than a hardware implementation of these functions is higher for single precision than for double precision. But the software implementation is faster than the hardware implementation in most cases, even for double precision.

You may use the Intel math function library with a different compiler by including the library `libmmt.lib` and the header file `mathimf.h` that come with the Intel C++ compiler. This library contains many useful mathematical functions. A lot of advanced mathematical

functions are supplied in Intel's Math Kernel Library, available from www.intel.com. (See also page 121). The AMD math core library contains similar functions.

Note that the Intel function libraries do not use the best possible instruction set when running on AMD processors (see page 132 for how to overcome this limitation), while the AMD math core library does not discriminate against Intel processors.

15 Metaprogramming

Metaprogramming means to make code that makes code. For example, in interpreted script languages, it is often possible to make a piece of code that produces a string and then interpret that string as code.

Metaprogramming can be useful in compiled languages such as C++ for doing some calculations at compile time rather than at runtime if all the inputs to the calculations are available at compile time. (Of course there is no such advantage in interpreted languages where everything happens at runtime).

The following techniques can be considered metaprogramming in C++:

- Preprocessor directives. For example use `#if` instead of `if`. This is a very efficient way of removing superfluous code, but there are serious limitations to what the preprocessor can do because it comes before the compiler and it understands only the simplest expressions and operators.

- Make a C++ program that produces another C++ program (or part of it). This can be useful in some cases, for example to produce tables of mathematical functions that you want as static arrays in the final program. This requires, of course, that you compile twice.

- An optimizing compiler may try to do as much as possible at compile time. For example, most compilers will reduce `int x = 2 * 5;` to `int x = 10;`

- Templates are instantiated at compile time. A template instance has its parameters replaced by their actual values before it is compiled. This is the reason why there is virtually no cost to using templates (see p. 57). It is possible to express any algorithm with template metaprogramming, but this method is extremely complicated and clumsy, as you will see shortly.

The following examples explain how metaprogramming can be used to speed up the calculation of the power function when the exponent is an integer known at compile time.

```
// Example 15.1a. Calculate x to the power of 10
double xpow10(double x) {
   return pow(x,10);
}
```

The `pow` function uses logarithms in the general case, but in this case it will recognize that 10 is an integer, so that the result can be calculated using multiplications only. The following algorithm is used inside the `pow` function when the exponent is a positive integer:

```
// Example 15.1b. Calculate integer power using loop
double ipow (double x, unsigned int n) {
   double y = 1.0;                    // used for multiplication
   while (n != 0) {                   // loop for each bit in nn
       if (n & 1) y *= x;             // multiply if bit = 1
       x *= x;                        // square x
```

152

```
        n >>= 1;                      // get next bit of n
    }
    return y;                         // return y = pow(x,n)
}


double xpow10(double x) {
    return ipow(x,10);                // ipow faster than pow
}
```

The method used in example 15.1b is easier to understand when we roll out the loop and reorganize:

```
// Example 15.1c. Calculate integer power, loop unrolled
double xpow10(double x) {
    double x2  = x * x;          // x^2
    double x4  = x2*x2;          // x^4
    double x8  = x4*x4;          // x^8
    double x10 = x8*x2;          // x^10
    return x10;                  // return x^10
}
```

As we can see, it is possible to calculate `pow(x,10)` with only four multiplications. How was it possible to come from example 15.1b to 15.1c? We took advantage of the fact that `n` is known at compile time to eliminate everything that depends only on `n`, including the `while` loop, the `if` statement and all the integer calculations. The code in example 15.1c is faster than 15.1b, and in this case it may be smaller as well.

The conversion from example 15.1b to 15.1c was done by me manually, but if we want to generate a piece of code that works for any compile-time constant `n`, then we need metaprogramming. None of the compilers I have tested can convert example 15.1a to 15.1c automatically, and only the Gnu compiler will convert example 15.1b to 15.1c. We can only hope that future compilers will do such optimizations automatically, but as long as this is not the case we may need metaprogramming.

The next example shows this calculation implemented with template metaprogramming. Don't panic if you don't understand it. I am giving this example only to show how tortuous and convoluted template metaprogramming is.

```
// Example 15.1d. Integer power using template metaprogramming

// Template for pow(x,N) where N is a positive integer constant.
// General case, N is not a power of 2:
template <bool IsPowerOf2, int N>
class powN {
public:
    static double p(double x) {
    // Remove right-most 1-bit in binary representation of N:
    #define N1 (N & (N-1))
        return powN<(N1&(N1-1))==0,N1>::p(x) * powN<true,N-N1>::p(x);
    #undef N1
    }
};

// Partial template specialization for N a power of 2
template <int N>
class powN<true,N> {
public:
    static double p(double x) {
        return powN<true,N/2>::p(x) * powN<true,N/2>::p(x);
    }
};
```

```cpp
    // Full template specialization for N = 1
    template<>
    class powN<true,1> {
    public:
        static double p(double x) {
            return x;
        }
    };

    // Full template specialization for N = 0
    // This is used only for avoiding infinite loop if powN is
    // erroneously called with IsPowerOf2 = false where it should be true.
    template<>
    class powN<true,0> {
    public:
        static double p(double x) {
            return 1.0;
        }
    };

    // Function template for x to the power of N
    template <int N>
    static inline double IntegerPower (double x) {
        // (N & N-1)==0 if N is a power of 2
        return powN<(N & N-1)==0,N>::p(x);
    }

    // Use template to get x to the power of 10
    double xpow10(double x) {
        return IntegerPower<10>(x);
    }
```

If you want to know how this works, here's an explanation. Please skip the following explanation if you are not sure you need it.

In C++ template metaprogramming, loops are implemented as recursive templates. The `powN` template is calling itself in order to emulate the `while` loop in example 15.1b. Branches are implemented by (partial) template specialization. This is how the `if` branch in example 15.1b is implemented. The `powN` template is a class template rather than a function template because partial template specialization is allowed only for classes. The splitting of `N` into the individual bits of its binary representation is particularly tricky. I have used the trick that `N1 = N&(N-1)` gives the value of `N` with the rightmost 1-bit removed. If `N` is a power of 2 then `N&(N-1)` is 0. The constant `N1` could have been defined in other ways than by a macro, but the method used here is the only one that works on all the compilers I have tried.

The Microsoft, Intel and Gnu compilers are actually reducing example 15.1d to 15.1c as intended, while the Borland and Digital Mars compilers produce less optimal code because they fail to eliminate common sub-expressions.

Why is template metaprogramming so complicated? Because the C++ template feature was never designed for this purpose. It just happened to be possible. Template meta-programming is so complicated that I consider it very unwise to use it. This is nothing but a playground for eggheads. Complicated code is a risk factor in itself, and the cost of verifying, debugging and maintaining such code is simply so high that it doesn't justify the relatively small gain in performance.

The D language allows compile-time `if` statements (called `static if`), but no compile-time loops. We can only hope that such a feature becomes available in the future. If a future version of C++ should allow compile-time `if` and compile-time `while` loops, then the

transformation of example 15.1b to metaprogramming would be straightforward. The MASM assembly language has full metaprogramming features, including the ability to define function names and variable names from string functions. A metaprogramming implementation analogous to example 15.1b and d in assembly language is provided as an example in the "Macro loops" chapter in manual 2: "Optimizing subroutines in assembly language".

While we are waiting for better metaprogramming tools to be available, we may choose the compilers that are best at doing metaprogramming at their own initiative whenever it is possible. A compiler that automatically reduces example 15.1a to 15.1c would of course be the easiest and the most reliable solution. (In my tests, the Intel compiler reduced 15.1a to an inlined 15.1b and the Gnu compiler reduced 15.1b to 15.1c, but none of the compilers reduced 15.1a to 15.1c).

16 Testing speed

Testing the speed of a program is an important part of the optimization job. You have to check if your modifications actually increase the speed or not.

There are various profilers available which are useful for finding the hot spots and measuring the overall performance of a program. The profilers are not always accurate, however, and it may be difficult to measure exactly what you want when the program spends most of its time waiting for user input or reading disk files. See page 16 for a discussion of profiling.

When a hot spot has been identified, then it may be useful to isolate the hot spot and make measurements on this part of the code only. This can be done very accurately by using the so-called time stamp counter. This is a counter which measures the number of clock pulses since the CPU was started. The length of a clock cycle is the reciprocal of the clock frequency, as explained on page 16. If you read the value of the time stamp counter before and after executing a critical piece of code then you can get the exact time consumption as the difference between the two clock counts.

The value of the time stamp counter can be obtained with the function `ReadTSC` listed below in example 16.1. This code works only for compilers that support intrinsics. Alternatively, you can get `ReadTSC` as a library function available in www.agner.org/optimize/asmlib.zip.

```
// Example 16.1
#include <intrin.h>          // Or #include <ia32intrin.h> etc.

__int64 ReadTSC() {          // Returns time stamp counter
   int dummy[4];             // For unused returns
   volatile int DontSkip;    // Volatile to prevent optimizing
   __int64 clock;            // Time
   __cpuid(dummy, 0);        // Serialize
   DontSkip = dummy[0];      // Prevent optimizing away cpuid
   clock = __rdtsc();        // Read time
   __cpuid(dummy, 0);        // Serialize again
   DontSkip = dummy[0];      // Prevent optimizing away cpuid
   return clock;
}
```

You can use this function to measure the clock count before and after executing the critical code. A test setup may look like this:

```
// Example 16.2
// Link with appropriate version of asmlib
```

```
#include <stdio.h>
#include <asmlib.h>          // Use ReadTSC() from library asmlib..
                             // or from example 16.1
void CriticalFunction();   // This is the function we want to measure

const int NumberOfTests = 10;        // Number of times to test
int i; __int64 time1;
__int64 timediff[NumberOfTests];     // Time difference for each test
for (i = 0; i < NumberOfTests; i++) { // Repeat NumberOfTests times
    time1 = ReadTSC();                // Time before test

    CriticalFunction();               // Critical function to test

    timediff[i] = ReadTSC() - time1;  // (time after) - (time before)
}
printf("\nResults:");                 // Print heading
for (i = 0; i < NumberOfTests; i++) { // Loop to print out results
    printf("\n%2i  %10I64i", i, timediff[i]);
}
```

The code in example 16.2 calls the critical function ten times and stores the time consumption of each run in an array. The values are then output after the test loop. The time that is measured in this way includes the time it takes to call the ReadTSC function. You can subtract this value from the counts. It is measured simply by removing the call to CriticalFunction in example 16.2.

The measured time is interpreted in the following way. The first count is always higher than the subsequent counts. This is the time it takes to execute CriticalFunction when code and data are not cached. The subsequent counts give the execution time when code and data are cached as good as possible. The first count and the subsequent counts represent the "worst case" and "best case" values. Which of these two values is closest to the truth depends on whether CriticalFunction is called once or multiple times in the final program and whether there is other code that uses the cache in between the calls to CriticalFunction. If your optimization effort is concentrated on CPU efficiency then it is the "best case" counts that you should look at to see if a certain modification is profitable. On the other hand, if your optimization effort is concentrated on arranging data in order to improve cache efficiency, then you may also look at the "worst case" counts. In any event, the clock counts should be multiplied by the clock period and by the number of times CriticalFunction is called in a typical application to calculate the time delay that the end user is likely to experience.

Occasionally, the clock counts that you measure are much higher than normal. This happens when a task switch occurs during execution of CriticalFunction. You cannot avoid this in a protected operating system, but you can reduce the problem by increasing the thread priority before the test and setting the priority back to normal afterwards.

An alternative to the test setup in example 16.2 is to use one of the test tools that I have made available at www.agner.org/optimize/testp.zip. These test tools are based on the same principle as example 16.2, but they can give additional performance counts for cache misses, misaligned memory references, branch mispredictions, floating point underflows, etc. This information is based on the performance monitor counters that are built into the CPU hardware. A test is performed by inserting the critical function into the test program and compiling it. My test tools are intended for testing small critical pieces of code, not for testing whole programs.

The time stamp counter is a little inaccurate on microprocessors that can change the clock frequency (Intel SpeedStep® technology). A more accurate measurement can be obtained

with a performance monitor counter for "core clock cycles", using the test program mentioned above.

16.1 The pitfalls of unit-testing

It is common practice to test each function or class separately in software development. This unit-testing is necessary for verifying the functionality of an optimized function, but unfortunately the unit-test doesn't give the full information about the performance of the function in terms of speed.

Assume that you have two different versions of a critical function and you want to find out which one is fastest. The typical way to test this is to make a small test program that calls the critical function many times with a suitable set of test data and measure how long time it takes. The version that performs best under this unit-test may have a larger memory footprint than the alternative version. The penalty of cache misses is not seen in the unit-test because the total amount of code and data memory used by the test program is likely to be less than the cache size.

When the critical function is inserted in the final program, it is very likely that code cache and data cache are critical resources. Modern CPUs are so fast that the clock cycles spent on executing instructions are less likely to be a bottleneck than memory access and cache size. If this is the case then the optimal version of the critical function may be the one that takes longer time in the unit-test but has a smaller memory footprint.

If, for example, you want to find out whether it is advantageous to roll out a big loop then you cannot rely on a unit-test without taking cache effects into account.

You can calculate how much memory a function uses by looking at a link map or an assembly listing. Use the "generate map file" option for the linker. Both code cache use and data cache use can be critical. The branch target buffer is also a cache that can be critical. Therefore, the number of jumps, calls and branches in a function should also be considered.

A realistic performance test should include not only a single function or hot spot but also the innermost loop that includes the critical functions and hot spots. The test should be performed with a realistic set of data in order to get reliable results for branch mispredictions. The performance measurement should not include any part of the program that waits for user input. The time used for file input and output should be measured separately.

The fallacy of measuring performance by unit-testing is unfortunately very common. Even some of the best optimized function libraries available use excessive loop unrolling so that the memory footprint is unreasonably large.

16.2 Worst-case testing

Most performance tests are done under the best-case conditions. All disturbing influences are removed, all resources are sufficient, and the caching conditions are optimal. Best-case testing is useful because it gives more reliable and reproducible results. If you want to compare the performance of two different implementations of the same algorithm, then you need to remove all disturbing influences in order to make the measurements as accurate and reproducible as possible.

However, there are cases where it is more relevant to test the performance under the worst-case conditions. For example, if you want to make sure that the response time to user input never exceeds one second, then you should test the response time under worst-case conditions.

Programs that produce streaming audio or video should also be tested under worst-case conditions in order to make sure that they always keep up with the expected real-time speed. If the computer system is too slow, then there will be occasional delays or glitches in the output, which is usually unacceptable.

Each of the following methods could possibly be relevant when testing worst-case performance:

- The first time you activate a particular part of the program, it is likely to be slower than the subsequent times because of lazy loading of the code, cache misses and branch mispredictions.

- Test the whole software package, including all runtime libraries and frameworks, rather than isolating a single function. Switch between different parts of the software package in order to increase the likelihood that certain parts of the program code are uncached or even swapped to disk.

- Software that relies on network resources and servers should be tested on a network with heavy traffic and a server in full use rather than a dedicated test server.

- Use large data files and databases with lots of data.

- Use an old computer with a slow CPU, an insufficient amount of RAM, a lot of irrelevant software installed, a lot of background processes running, and a fragmented hard disk.

- Test with different brands of CPUs, different types of graphics cards, etc.

- Use an antivirus program that scans all files on access.

- Run multiple processes or threads simultaneously. If the microprocessor has hyperthreading, then try to run two threads in each processor core.

- Try to allocate more RAM than there is, in order to force the swapping of memory to disk.

- Provoke cache misses by making the code size or data used in the innermost loop bigger than the cache size. Alternatively, you may actively invalidate the cache. The operating system may have a function for this purpose, or you may use the `_mm_clflush` intrinsic function.

- Provoke branch mispredictions by making the data more random than normal.

17 Optimization in embedded systems

Microcontrollers used in small embedded applications have less computing resources than standard PCs. The clock frequency may be a hundred or even a thousand times lower; and the amount of RAM memory may even be a million times less than in a PC. Nevertheless, it is possible to make software that runs quite fast on such small devices if you avoid the large graphics frameworks, interpreters, just-in-time compilers, system database, and other extra software layers and frameworks typically used on bigger systems.

The smaller the system, the more important it is to choose a software framework that uses few resources. On the smallest devices, you don't even have an operating system.

The best performance is obtained by choosing a programming language that can be cross-compiled on a PC and then transferred as machine code to the device. Any language that requires compilation or interpretation on the device itself is a big waste of resources. For these reasons, the preferred language will often be C or C++. Critical device drivers may need assembly language.

C++ takes only slightly more resources than C if you follow the guidelines below. You may choose either C or C++ based on what is most appropriate for the desired program structure.

It is important to economize the use of RAM memory. Big arrays should be declared inside the function they are used in so that they are deallocated when the function returns. Alternatively, you may reuse the same array for multiple purposes.

All dynamic memory allocation using `new`/`delete` or `malloc`/`free` should be avoided because of the large overhead of managing a memory heap. The heap manager has a garbage collector which is likely to consume time at unpredictable intervals which may interfere with real time applications.

Remember that container classes in the STL (Standard Template Library) and other container class libraries use dynamic memory allocation with `new` and `delete`, and often excessively so. These containers should definitely be avoided unless you have ample resources. For example, a FIFO queue should be implemented as a circular buffer with fixed size to avoid dynamic memory allocation. Do not use a linked list (see page 92).

All common implementations of string classes use dynamic memory allocation. You should avoid these and handle text strings in the old fashioned C style as character arrays. Note that the C style string functions have no check for overflow of the arrays. It is the responsibility of the programmer to make sure the arrays are sufficiently large to handle the strings including the terminating zero and to make overflow checks where necessary (see page 95).

Virtual functions in C++ take more resources than non-virtual functions. Avoid virtual functions if possible.

Smaller microprocessors have no native floating point execution units. Any floating point operation on such processors requires a big floating point library which is very time consuming. Therefore, you should avoid any use of floating point expressions. For example, `a = b * 2.5` may be changed to `a = b * 5 / 2` (be aware of possible overflow on the intermediate expression `b * 5`). As soon as you have even a single constant with a decimal point in your program, you will be loading the entire floating point library. If you want a number to be calculated with two decimals, for example, you should multiply it by 100 so that it can be represented as an integer.

Integer variables can be 8, 16 or 32 bits (rarely 64). You may save RAM space, if necessary, by using the smallest integer size that doesn't cause overflow in the particular application. The integer size is not standardized across platforms. See the compiler documentation for the size of each integer type.

In this chapter, I have described some of the considerations that are particularly important on small devices with limited resources. Most of the advice in the rest of the present manual is also relevant to small devices, but there are some differences due to the design of small microcontrollers:

- Smaller microcontrollers have no branch prediction . There is no need to take branch prediction into account in the software.

- Smaller microcontrollers have no cache . There is no need to organize data to optimize caching.

- Smaller microcontrollers have no out-of-order execution. There is no need to break down dependency chains .

18 Overview of compiler options

Table 18.1. Command line options relevant to optimization				
	MS compiler Windows	**Gnu compiler Linux**	**Intel compiler Windows**	**Intel compiler Linux**
Optimize for speed	`/O2 or /Ox`	`-O3`	`/O3`	`-O3`
Interprocedural optimization	`/Og`			
Whole program optimization	`/GL`	`--combine -fwhole- program`	`/Qipo`	`-ipo`
No exception handling	`/EHs-`			
No stack frame	`/Oy`	`-fomit- frame- pointer`		`-fomit- frame- pointer`
No runtime type identification (RTTI)	`/GR-`	`-fno-rtti`	`/GR-`	`-fno-rtti`
Assume no pointer aliasing	`/Oa`			`-fno-alias`
Non-strict floating point		`-ffast-math`	`/fp:fast /fp:fast=2`	`-fp-model fast, -fp- model fast=2`
Simple member pointers	`/vms`			
Fastcall functions	`/Gr`			
Function level linking (remove unreferen- ced functions)	`/Gy`	`-ffunction- sections`	`/Gy`	`-ffunction- sections`
SSE instruction set (128 bit float vectors)	`/arch:SSE`	`-msse`	`/arch:SSE`	`-msse`
SSE2 instruction set (128 vectors of inte- ger or double)	`/arch:SSE2`	`-msse2`	`/arch:SSE2`	`-msse2`
SSE3 instruction set		`-msse3`	`/arch:SSE3`	`-msse3`
Suppl. SSE3 instr. set		`-mssse3`	`/arch:SSSE2`	`-mssse3`
SSE4.1 instr. set		`-msse4.1`	`/arch:SSE4.1`	`-msse4.1`
AVX instr. set	`/arch:AVX`	`-mAVX`	`/arch:AVX`	`-mAVX`
Automatic CPU dispatch			`/QaxSSE3`, etc. (Intel CPU only)	`-axSSE3`, etc. (Intel CPU only)
Automatic vectorization		`-mveclibabi`	(requires no specific option)	
Automatic paralleli- zation by multiple threads			`/Qparallel`	`-parallel`
Parallelization by OpenMP directives	`/openmp`	`-fopenmp`	`/Qopenmp`	`-openmp`
32 bit code		`-m32`		
64 bit code		`-m64`		
Static linking (multithreaded)	`/MT`	`-static`	`/MT`	`-static`
Generate assembly listing	`/FA`	`-S - masm=intel`	`/FA`	`-S`
Generate map file	`/Fm`			

			/Qopt-report	-opt-report
Generate optimization report				

Table 18.2. Compiler directives and keywords relevant to optimization

	MS compiler Windows	Gnu compiler Linux	Intel compiler Windows	Intel compiler Linux
Align by 16	`__declspec( align(16))`	`__attribute(( aligned(16)))`	`__declspec( align(16))`	`__attribute(( aligned(16)))`
Assume pointer is aligned			`#pragma vector aligned`	`#pragma vector aligned`
Assume pointer not aliased	`#pragma optimize("a", on)` `__restrict`	`__restrict`	`__declspec( noalias)` `__restrict` `#pragma ivdep`	`__restrict` `#pragma ivdep`
Assume function is pure		`__attribute(( const))`		`__attribute(( const))`
Assume function does not throw exceptions	`throw()`	`throw()`	`throw()`	`throw()`
Assume function called only from same module	`static`	`static`	`static`	`static`
Assume member function called only from same module		`__attribute__ ((visibility ("internal")))`		`__attribute__ ((visibility ("internal")))`
Vectorize			`#pragma vector always`	`#pragma vector always`
Optimize function	`#pragma optimize(...)`			
Fastcall function	`__fastcall`	`__attribute(( fastcall))`	`__fastcall`	
Noncached write			`#pragma vector nontemporal`	`#pragma vector nontemporal`

Table 18.3. Predefined macros

	MS compiler Windows	Gnu compiler Linux	Intel compiler Windows	Intel compiler Linux
Compiler identification	`_MSC_VER` and not `__INTEL_COMPILER`	`__GNUC__` and not `__INTEL_COMPILER`	`__INTEL_COMPILER`	`__INTEL_COMPILER`
16 bit platform	not `_WIN32`	n.a.	n.a.	n.a.
32 bit platform	not `_WIN64`		not `_WIN64`	
64 bit	`_WIN64`	`_LP64`	`_WIN64`	`_LP64`

162

platform				
Windows platform	`_WIN32`		`_WIN32`	
Linux platform	n.a.	`__unix__` `__linux__`		`__unix__` `__linux__`
x86 platform	`_M_IX86`		`_M_IX86`	
x86-64 platform	`_M_IX86` and `_WIN64`		`_M_X64`	`_M_X64`

19 Literature

Other manuals by Agner Fog
The present manual is number one in a series of five manuals. See page 3 for a list of titles.

Literature on code optimization
Intel: "Intel 64 and IA-32 Architectures Optimization Reference Manual".
developer.intel.com.
> Many advices on optimization of C++ and assembly code for Intel CPUs. New versions are produced regularly.

AMD: "Software Optimization Guide for AMD Family 10h Processors". www.amd.com.
> Advices on optimization of C++ and assembly code for AMD CPUs. New versions are produced regularly.

Intel: "Intel® C++ Compiler Documentation". Included with Intel C++ compiler, which is available from www.intel.com.
> Manual on using the optimization features of Intel C++ compilers.

Wikipedia article on compiler optimization. en.wikipedia.org/wiki/Compiler_optimization.

ISO/IEC TR 18015, "Technical Report on C++ Performance". www.open-std.org/jtc1/sc22/wg21/docs/TR18015.pdf.

OpenMP. www.openmp.org. Documentation of the OpenMP directives for parallel processing.

Scott Meyers: "Effective C++". Addison-Wesley. Third Edition, 2005; and "More Effective C++". Addison-Wesley, 1996.
> These two books contain many tips on advanced C++ programming, how to avoid hard-to-find errors, and some tips on improving performance.

Stefan Goedecker and Adolfy Hoisie: "Performance Optimization of Numerically Intensive Codes", SIAM 2001.
> Advanced book on optimization of C++ and Fortran code. The main focus is on mathematical applications with large data sets. Covers PC's, workstations and scientific vector processors.

Henry S. Warren, Jr.: "Hacker's Delight". Addison-Wesley, 2003.
> Contains many bit manipulation tricks

Michael Abrash: "Zen of code optimization", Coriolis group books 1994.
> Mostly obsolete.

Rick Booth: "Inner Loops: A sourcebook for fast 32-bit software development", Addison-Wesley 1997.
> Mostly obsolete.

Microprocessor documentation
Intel: "IA-32 Intel Architecture Software Developer's Manual", Volume 1, 2A, 2B, and 3A and 3B. developer.intel.com.

AMD: "AMD64 Architecture Programmer's Manual", Volume 1 - 5. www.amd.com.

<u>Literature on usability</u>

Jenny Preece (ed): "Human-Computer interaction". Addison-Wesley 1994.
      University-level textbook. Theoretical but easy to understand, with many illustrations
      and exercises.

Ben Shneiderman & Catherine Plaisant: "Designing the User Interface: Strategies for
Effective Human-Computer Interaction". 4th ed. Addison Wesley 2004.
      Comprehensive university-level textbook on the design of human/computer interface.

Ernest McCormick: "Human factor in engineering and design". McGraw-Hill 1976
      Theoretical textbook about input/output devices, ergonomics, cognition, psychology.

W. M. Newman & M. G. Lamming: "Interactive System Design". Addison-Wesley 1995.
      Textbook with the main focus on user psychology and cognition, including user
      study, modeling user activity, and systems analysis.

Jef Raskin: "The Humane Interface: New Directions for Designing Interactive Systems".
Addison-Wesley 2000.
      Theoretical book on human/computer interface and cognitive psychology with
      detailed discussion of commands, displays, cursors, icons, menus, etc.

W3C: "Web Content Accessibility Guidelines 1.0". 1999. www.w3.org/TR/WAI-WEBCONTENT/.
      Guidelines for handicap-friendly web user interfaces. Some of the advices are also
      applicable to other software interfaces.

<u>Internet forums</u>

Several internet forums and newsgroups contain useful discussions about code optimization. See www.agner.org/optimize and the FAQ for the newsgroup comp.lang.asm.x86 for some links.

20 Copyright notice