

Соснин В.В., Балакшин П.В., Шилко Д.С. Введение в параллельные вычисления. – СПб: Университет ИТМО, 2020. – 51 с. **НЕТ!!!!**

В пособии излагаются основные понятия и определения теории параллельных вычислений. Рассматриваются основные принципы построения программ на языке «Си» для многоядерных и многопроцессорных вычислительных комплексов с общей памятью. Предлагается набор заданий для проведения лабораторных и практических занятий.

Учебное пособие предназначено для студентов, обучающихся по магистерским программам направления «09.04.04 – Программная инженерия», «09.04.01 – Информатика и вычислительная техника», и может быть использовано выпускниками (бакалаврами и магистрантами) при написании выпускных квалификационных работ, связанных с проектированием и исследованием многоядерных и многопроцессорных вычислительных комплексов.



Университет ИТМО – ведущий вуз России в области информационных и фотонных технологий, один из немногих российских вузов, получивших в 2009 году статус национального исследовательского университета. С 2013 года Университет ИТМО – участник программы повышения конкурентоспособности российских университетов среди ведущих мировых научно-образовательных центров, известной как проект «5 в 100». Цель Университета ИТМО – становление исследовательского университета мирового уровня, предпринимательского по типу, ориентированного на интернационализацию всех направлений деятельности.

© Университет ИТМО, 2020  
© Соснин В.В., Балакшин П.В., Шилко Д.С., 2020

# Содержание

<b>Introduction</b>	<b>5</b>
<b>1 Теоретические основы параллельных вычислений</b>	<b>7</b>
1.1 History of parallel computing . . . . .	7
1.2 Terms and Definitions . . . . .	9
1.3 Classification of parallel systems (architectures) . . . . .	14
1.4 Synchronization methods in parallel programs . . . . .	17
1.5 Automatic parallelization of programs . . . . .	20
1.6 The main approaches to parallelization . . . . .	22
1.7 Atomic operations in a multithreaded program . . . . .	23
1.8 Lock-free структуры данных . . . . .	25
<b>2 Показатели эффективности параллельной программы</b>	<b>30</b>
2.1 Параллельное ускорение и параллельная эффективность . .	30
2.2 Метод Амдала . . . . .	33
2.3 Метод Густавсона-Барсиса . . . . .	36
2.4 Модификация закона Амдала (по проф. Бухановскому) . . .	37
2.5 Измерение времени выполнения параллельных программ .	38
<b>3 Практические аспекты параллельного программирования</b>	<b>43</b>
3.1 Отладка параллельных программ . . . . .	43
3.2 Менеджеры управления памятью для параллельных программ . . . . .	44
3.3 Технология OpenMP . . . . .	45
3.4 Технология OpenCL . . . . .	56
3.5 Ошибки в многопоточных приложениях . . . . .	64
<b>4 Лабораторная работа №1. «Автоматическое распараллеливание программ»</b>	<b>72</b>
4.1 Порядок выполнения работы . . . . .	72
4.2 Состав отчета . . . . .	74
4.3 Подготовка к защите . . . . .	75
4.4 Варианты заданий . . . . .	75
<b>5 Лабораторная работа №2. «Исследование эффективности параллельных библиотек для C-программ»</b>	<b>78</b>
5.1 Порядок выполнения работы . . . . .	78
5.2 Состав отчета . . . . .	79

5.3	Подготовка к защите . . . . .	80
<b>6</b>	<b>Лабораторная работа №3. «Распараллеливание циклов с помощью технологии OpenMP»</b>	<b>81</b>
6.1	Порядок выполнения работы . . . . .	81
6.2	Состав отчета . . . . .	82
6.3	Подготовка к защите . . . . .	83
<b>7</b>	<b>Лабораторная работа №4. «Метод доверительных интервалов при измерении времени выполнения параллельной OpenMP-программы»</b>	<b>84</b>
7.1	Порядок выполнения работы . . . . .	84
7.2	Состав отчета . . . . .	85
7.3	Подготовка к защите . . . . .	86
<b>8</b>	<b>Лабораторная работа №5. «Параллельное программирование с использованием стандарта POSIX Threads»</b>	<b>87</b>
8.1	Порядок выполнения работы . . . . .	87
8.2	Состав отчета . . . . .	88
8.3	Подготовка к защите . . . . .	88
<b>9</b>	<b>Лабораторная работа №6. «Изучение технологии OpenCL»</b>	<b>89</b>
9.1	Порядок выполнения работы . . . . .	89
9.2	Состав отчета . . . . .	89
9.3	Подготовка к защите . . . . .	90
	<b>Список используемой литературы</b>	<b>91</b>

## Introduction

Currently, most microprocessors are multi-core. This applies not only to desktop computers, but also to mobile phones and tablets (so far, only embedded computing systems are an exception). To fully realize the potential of a multi-core system, a programmer needs to use special methods of parallel programming, which are becoming increasingly popular in industrial programming. However, parallel programming methods are significantly more difficult to master than traditional sequential program writing methods.

The purpose of this study book is to describe practical tasks (laboratory work) that can be used to consolidate the theoretical knowledge gained as part of a lecture course on parallel programming technologies. In addition, the book summarizes the basic principles of parallel programming.

While programming multi-threaded applications, you have to resolve conflicts that arise when simultaneously accessing the shared memory of several threads. The following three conceptually different approaches are currently used to synchronize simultaneous access to shared memory:

1. **Explicit use of blocking primitives** (mutexes, semaphores, condition variables). This approach historically appeared first and is now the most common and supported in most programming languages. The disadvantage of this method is a rather high entry threshold, since the programmer is required to manage blocking primitives in the "manual mode", tracking conflict situations when accessing shared memory.
2. **Software Transactional Memory (STM)**. This method is easier to learn and use than the previous one, however it still has limited support in compilers, and it will also be able to fully manifest itself with the wider distribution of processors with hardware support for STM.
3. **Non-blocking algorithms** (lockless, lock-free, wait-free algorithms). This method implies a complete rejection of the use of blocking primitives with the help of complex algorithmic tricks. Moreover, for the correct functioning of the non-blocking algorithm, it is required that the processor supports special atomic (conflict-free) operations of the form "compare and exchange" (cmpxchg, "compare and swap"). Currently, most processors have this type of operation as part of the instruction system (with rare exceptions, for example: "SPARC 32").

The methodological manual proposed is devoted to the first of the listed methods, since he received the greatest coverage in literature and the greatest

application in industrial programming. Two other methods may be the subject of in-depth training courses on parallel computing.

# 1 Теоретические основы параллельных вычислений

## 1.1 History of parallel computing

Conversation about the development of parallel programming is usually begun with the history of the development of supercomputers. However, the world's first supercomputer CDC6600, created in 1963, had only one central processor, so it can hardly be considered a full-fledged SMP system.

The third-ever CDC8600 supercomputer was designed to use four processors with shared memory, which suggests the first use of SMP, but the CDC8600 was never released since its development was discontinued in 1972.

Only in 1983 was it possible to create a working supercomputer (Cray X-MP), which used two central processors that used shared memory. It is worth noting that a little earlier (in 1980) the first Russian multiprocessor computer Elbrus-1 appeared, however, it was significantly inferior in performance to supercomputers of that time.

Already in 1994 it was possible to freely buy a desktop computer with two processors, when ASUS released its first motherboard with two sockets - connectors for installing processors.

The next step in the development of SMP-systems was the emergence of multi-core processors. The first multi-core processor for mass use was POWER4, released by IBM in 2001. But the truly widespread multi-core architecture received only in 2005, when AMD and Intel released their first dual-core processors.

The figure 1 shows how much CPU with a different number of cores occupied when creating supercomputers at different times (according to the materials of the site (<http://top500.org>)). Закрашенные области помечены цифрами 1, 2, 4, 6, 8, 10, 12, 16 для обозначения количества ядер. Ширина области по вертикали равна относительной частоте использования процессоров соответствующего типа в рассматриваемом году.



*Рис. 1: Frequency of using processors with different number of cores when creating supercomputers*

As you can see, the active use of dual-core processors in supercomputers began already in 2002, and by about 2005 completely disappeared, whereas in desktop computers their use was only just beginning in 2005. Based on this, you can make a simple forecast of the prevalence of multi-core "desktop" processors by the desired year, if we assume that they in general outline repeat the development of multi-core architectures of supercomputers.

## 1.2 Terms and Definitions

**Parallel computing** – A method of organizing calculations in which the program is a set of interacting modules that work simultaneously. Typically, the concept of concurrency may include:

1. **Instruction level parallelism** - one processor core can execute several instructions simultaneously. For example, this technology is implemented in Intel Pentium 4 processors.
2. **Hypertreading** - one processor core is designed so that it can do the work of two threads at once. Implemented in the Intel Core i7 series processors. When performing laboratory work, it is important to disable this in the BIOS (if the processor is compatible with this technology), since it significantly affects the performance of parallel acceleration and efficiency.
3. **Multi core programming** - a method for solving computational problems with simultaneous execution of program parts on different physical computing cores. All cores have shared memory banks, as they are located on the same computer. It includes the case of a multi-core processor architecture and a multi-processor architecture of a system in which there are several processors, since in both cases the program runs on several cores of one or many processors, but on one physical computer.
4. **Distributed computing** - a way to solve time-consuming computing problems using several computers, most often combined into a parallel computing system. Different parts of the program can run on different computers.

Usually the last two concepts are physically implemented using architectures. *SMP* и *MMP*. More details about these architectures can be found in the next section.



It is important to see difference between the concepts of parallel computing and parallel technologies. Let analyze the following concepts, which, although they are parallel technologies (at the core or internuclear interaction level), however, are not parallel calculations, but often textbf by mistake are assigned to them:

- *Pipeline data processing (superscalarity)* represents the simultaneous processing by the processor of several instructions, in which at one time for each of the instructions a different execution step is performed. For example, if any processor can simultaneously receive, decode, and execute an instruction, then when it receives the first instruction, it can decode the second and execute the third (figure ??). This way of organizing calculations is not parallel computing, because the instructions are still executed sequentially, and only one core is involved.

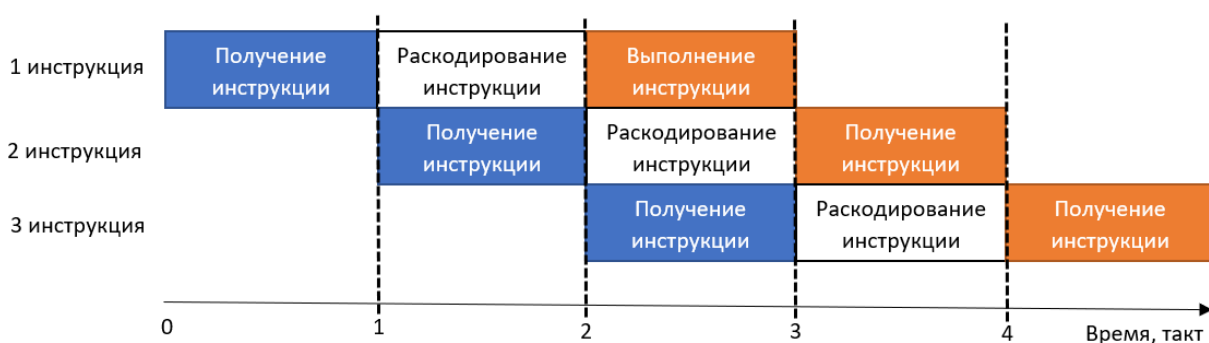


Рис. 2: Instruction pipelining

- *SIMD-extensions (MMX, SSE)* provide concurrency at the data level. For example, a processor can simultaneously multiply 4 numbers instead of one using the SSE instruction. However, the command flow still remains single, i.e. one program instruction is executed in a period of time, which is not the case of parallel computing.
- *Preemptive Multitasking* organized by the operating system. Several processes are in the execution queue and the OS decides how to manage the processor time between them. If the first thread is given a higher priority than the second, then the OS will allocate more time to execute the first thread, only one thread will be executed at a time, therefore, preemptive multitasking is also not included in the concept of parallel computing.

Various parallelization technologies are used to organize parallel computing:

- **Process** - the most heavyweight mechanism used for parallelization. Each process has its own independent address space, so data synchronization between processes is long and complicated. May include several threads of execution.
- **Thread** It runs independently of other threads, but has a common address space with other threads in the same process. At this level, data synchronization mechanisms are used (will be discussed later).
- **Fiber** - lightweight thread of execution. Like threads, fiber has a common address space, however, it uses joint multitasking instead of preemptive one. The OS does not switch the context from one thread to another, instead, the main thread itself allocates time for the child fiber to work, or is blocked logically (that is, the programmer controls the fiber life cycle). Also, all fibers work on one core, unlike threads, which can work on different kernels.

For a better understanding of threads, we will schematically consider its life cycle. Figure 3 shows that the flow can be in three states - readiness, expectation and fulfillment. After creating the stream, it is in a ready state. Then, the OS decides to change its state (clarifying multitasking). For fiber, the life cycle is the same, but the programmer or synchronization mechanisms control the transitions between them.

Different standards of programming languages can add new states to the life cycle of threads, for example, blocking a thread, interrupting a thread, and others, but the general scheme of work remains the same.

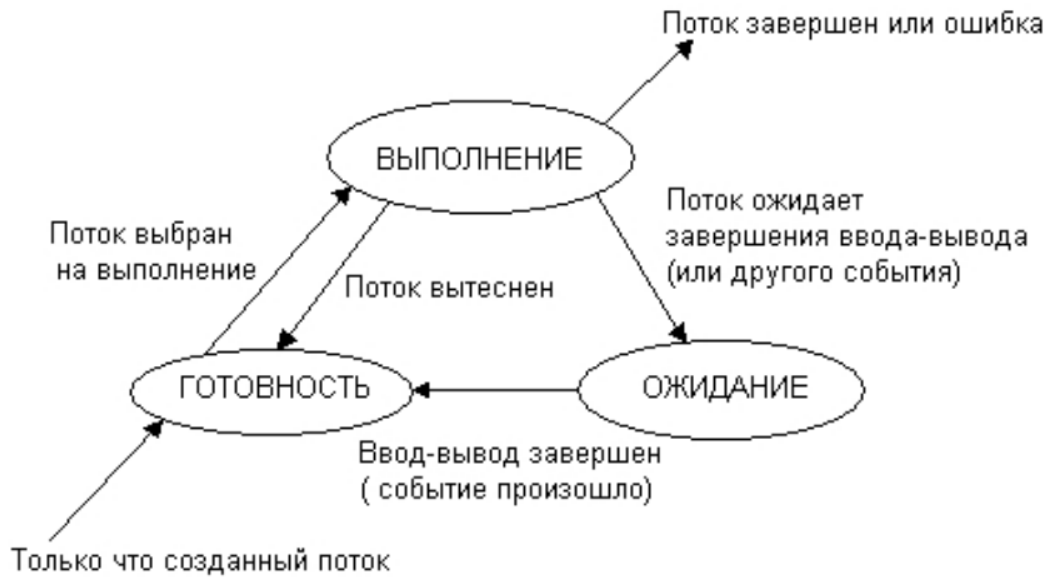


Рис. 3: Thread lifecycle

Among programmers there are concepts of **thread-safe** и **reentrant** functions, however, they may have different meanings in different communities. Definitions are written from various sources in table 1.

Таблица 1: Definitions of thread-safe and reentrant functions

Источник определения	Thread-safe	Reentrant
Qt	Внутри функции обращение ко всем общим переменным осуществляется строго последовательно, а не параллельно (Thread-safe является reentrant, но не наоборот)	При вызове функции одновременно несколькими потоками гарантируется правильная работа, только если потоки не используют общие данные
Linux	Функция показывает правильные результаты, даже если вызвана несколькими тредами одновременно	Функция показывает правильные результаты, даже если повторно вызвана изнутри себя
POSIX	?	Функция показывает правильные результаты, даже если вызвана несколькими тредами одновременно

Consider examples of functions that fit the definition of the Linux

community.

```
1  int t;
2  void swap(int *x, int *y) {
3      t = *x;
4      *x = *y;
5      // hardware interrupt
6      *y = t;
7  }
8  void interrupt_handler() {
9      int x = 1, y = 2;
10     swap(&x, &y);
11 }
```

This function is not thread safe, not reentrant, because all threads calling it will use the common variable `t`. If you call the function inside itself, then the value of `t` is overwritten and the parent function will not work correctly. Let's try to fix these errors by declaring a variable of type `t` `__threadint`.

```
1  __threadint t;
2  void swap(int *x, int *y) {
3      t = *x;
4      *x = *y;
5      // hardware interrupt
6      *y = t;
7  }
8  void interrupt_handler() {
9      int x = 1, y = 2;
10     swap(&x, &y);
11 }
```

Now the compiler will create a copy of the variable for each thread `t` and the function will become thread safe, however, it is still not reentrant for the same reason. We will save the value of the global variable `t` at the beginning of the function and restore it at the end.

```
1  int t;
2  void swap(int *x, int *y) {
3      int s;
4      s = t; // save global variable
5      t = *x;
6      *x = *y;
7      // hardware interrupt
8      *y = t;
9      t = s; // restore global variable
10 }
11 void interrupt_handler() {
12     int x = 1, y = 2;
13     swap(&x, &y);
14 }
```

The new function is reentrant, but again non thread-safe. Finally, we give an example of a standard and proper implementation of swap(), which is thread safe and reentrant:

```
1 void swap(int *x, int *y) {  
2     int t = *x;  
3     *x = *y;  
4     // hardware interrupt  
5     *y = t;  
6 }  
7 void interrupt_handler() {  
8     int x = 1, y = 2;  
9     swap(&x, &y);  
10 }
```

### 1.3 Classification of parallel systems (architectures)

By physical architecture, parallel systems can be divided into 2 types:

1. **SMP** (Shared Memory Parallelism, Symmetric MultiProcessor system) – multiprocessing, multicore, GPGPU.
2. **MPP** (Massively Parallel Processing) – cluster systems, GRID (distributed computing).

Next, we consider these two architectures in more detail.

**SMP** - architecture of multiprocessor systems in which two or more identical processors of comparable performance are connected in the same way to the shared memory (and peripheral devices) and perform the same functions (why, in fact, the system is called symmetrical). SMP systems are also called tightly coupled multiprocessors, since in this class of systems processors are closely connected to each other via a common bus and have equal access to all resources of the computing system (memory and input-output devices) and are controlled by all one copy of the operating system. In this architecture, all processors are located on the same physical machine, so they have common memory banks. There are two types of connecting processors to shared memory:

- The connection on the system bus is shown in the figure. 4. In this case, only one processor can access memory at any given moment, which imposes a significant limitation on the number of processors supported in such systems. The more processors, the greater the load on the shared bus, the longer each processor must wait until the bus is free to access the memory. A decrease in the overall

performance of such a system with an increase in the number of processors occurs very quickly, therefore, usually in such systems the number of processors does not exceed 2-4. An example of SMP machines with this method of connecting processors is any entry-level multiprocessor server.



*Puc. 4: SMP architecture. Processor connection via system bus*

- The crossbar switch is shown in the figure. 5. With this connection, the entire shared memory is divided into memory banks, each memory bank has its own bus, and the processors are connected to all buses, having access to any of the memory banks through them. Such a connection is technologically more complex, but it allows processors to access shared memory at the same time. This allows you to increase the number of processors in the system to 8-16 without a noticeable decrease in overall performance.



*Рис. 5: SMP architecture. Processor connection via dial-up connection*

The advantages of this approach are the high speed of data exchange between processors and the relative simplicity of software development. However, there may be problems with the scalability of the system (if there are only 2 sockets on the motherboard, you can't put 3 processors anymore).

**MMP** - architecture of multiprocessor systems, in which the memory between the processors is physically divided. On such systems, distributed computing is performed. The system is built from separate nodes containing a processor, a local memory bank, communication processors or network adapters, sometimes hard drives and other input-output devices. Access to the memory bank of this node is available only to processors from the same node. The nodes are connected by special communication channels (figure 6).



*Рис. 6: MMP Architecture*

The advantages of this approach are its good scalability (if necessary, to increase system performance it is enough to simply add more nodes). However,

the speed of interprocessor exchange is significantly reduced, since memory banks are now physically separated. Also, the cost of the software that distributes the calculations is very high.

## 1.4 Synchronization methods in parallel programs

In parallel programs, the developer often faces the problem of synchronization between threads. As a rule, problems arise when accessing memory and at the same time executing some critical sections of code - critical sections.

**Critical area** is the section of the program, which must be executed with the exclusive right of access to shared data referenced in this program. A process preparing to enter a critical area may be delayed if any other process is in progress at that time in a similar critical area.

This section will discuss in detail thread synchronization mechanisms at the program level.

The following methods are available for solving thread synchronization problems:

- **Atomic operations** are operations that are performed in their entirety or not at all. For example, a transaction to a database is an atomic operation. When two threads try to increment the same memory cell out of sync, the value can increase by 2, or maybe 1, depending on the behavior of the threads, since the increment operation is at least 3 assembler instructions. To avoid this, it is worth declaring the data type atomic (if there is one in this programming language / library). A special case of atomic operations is read-modify-write operations: compare-and-swap, test-and-set, fetch-and-add. The problem of the implementation of atomic operations will be raised in more detail in the section.

### *1.7 Atomic operations in a multithreaded program.*

- **Semaphore** - an object that limits the number of threads that can enter this area of code. Typically, this number is set when the semaphore is initialized. Then, when capturing a semaphore by a thread, the number of threads that captured the semaphore is checked. If the maximum number of threads is reached, then the thread will wait until some of the threads that entered the code area release it. Often the use of semaphores is unjustified, since the overhead of creating and maintaining a semaphore is large. You should also avoid the "semaphore leak" a situation in which the thread does not exit the



semaphore when the code area completes execution if the programmer forgot to free the resource.

- **Reader/writer semaphore** gives the threads *only* read or write permissions, and while writing data to one thread, the rest of the threads do not have access to the resource. However, in such semaphores there may be a problem *resource starvation*, in which while streams will read data, other streams will not be able to write data for a long period of time or vice versa. A particular solution to this problem with equal priority of threads can be sequential access of threads in the queue to access and write.
- **Mutex** is a special case of a semaphore, in which only one thread can capture a given area of code. If the mutex serves several critical sections, only one thread can be in any of the critical sections. It is often used in the organization of critical section management, since it is "lighter" than the classical semaphore (it is enough to store one boolean variable instead of a counter), but unlike it, it is assumed that the same stream will capture and release the mutex. It should be noted that in the C++11, in addition to the standard mutex, there are various modifications of it: *recursive\_mutex* - mutex that allow re-entry into the critical section by the same stream. *timed\_mutex* - mutex with capture timer, and *recursive\_timed\_mutex*, combining the advantages of both versions.
- **Spinlock** is a lock in which a thread in a loop waits for a resource to free. It is not always the optimal solution, since the waiting thread works while waiting. Inside the code section, thread interruptions must be avoided to avoid deadlock.
- **Seqlock** - synchronization mechanism designed to quickly record a variable in multiple threads. The Linux kernel works as follows: the thread waits until the critical section is released (spinlock); when entering the section, the counter increments, the stream does its work. When exiting a section, the flow checks the counter value. If the counter value has not changed, it means that no one has written data at the moment and the stream exits the critical section, otherwise it reads the value of the variable again.
- **Knuth-Bendix completion algorithm** - One of the solutions to synchronization problems is the Knuth-Bendix algorithm from the course of discrete mathematics. With it, you can go from a sequential

program to a cascading one. However, this algorithm does not work for all programs; sometimes it can go into an infinite loop or fail.

- **Barrier** - a part of code in which the state of threads is synchronized. For example, if a function in the main thread requires all child threads to finish their work, you can put a barrier in front of it. Then it will wait for the completion of the child threads, after which all threads will continue to work. An example of a barrier implementation may be a critical section, the code of which is allowed to be executed only by the last thread that requested execution. The rest of the threads should expect it. To do this, you need to know how many threads should come to the barrier.
- **Non-blocking algorithms**. It is often useful not to use standard locking techniques, but to make the algorithm non-blocking. In this case, the programmer must independently guarantee that critical sections of the code will not be executed simultaneously and the integrity of the shared memory. Another advantage of such algorithms is the safe handling of interrupts. Other synchronization technologies are often used to implement such algorithms: read-modify-write, CAS (see section 1.7) and etc.
- **RCU (read-copy-update)** - is an algorithm that allows threads to efficiently read data, leaving the data updated at the end of the algorithm, while guaranteeing relevant data. Only one thread can write data, but multiple threads can read data at once. This is achieved, for example, by atomic pointer substitution (CAS). Old versions of the data are stored for past hits, as long as they have at least one pointer. There are newer tools for replacing the pointer: a separate deadlock for writers or the membarrier mechanism used in recent versions of Linux. RCU can be useful in organizing data structures without explicit locks.
- **Monitor** - An object encapsulating a mutex and utility variables to provide secure access to a method or variable by multiple threads. The monitor characterizes that at one moment only one thread can execute any of its methods. For example, if we have a class (in C++ terms) Account with add methods `_money()`, `sub_money()`, it makes sense to make it a monitor so that there are no conflicts when conducting operations with the account.

However, it is not necessary to organize parallel computations using synchronization or locks. Some technologies offer an alternative approach to parallel computing:

- **Program Transactional Memory** - a memory model in which operations performed on memory cells are atomic. Advantages of use: ease of use (enclosing code blocks in a transaction block), lack of locks, however, if used incorrectly, performance may drop, as well as the inability to use operations that cannot be undone inside a transaction block. In the compiler, GCC is supported since version 4.7 as follows:
  1. `__transaction_atomic { ... }` — an indication that the code block is a transaction;
  2. `__transaction_relaxed { ... }` — an indication that the unsafe code inside the block does not lead to side effects;
  3. `__transaction_cancel` — explicit transaction cancellation;
  4. `attribute((transaction_safe))` — indication of a transaction-safe function;
  5. `attribute((transaction_pure))` — indication of function without side effects.
- **Actor model** - a mathematical model of parallel computing, in which the program is an actor objects that interact with each other and can create new actors, send and send messages to each other. The parallelism of computations within one actor is assumed. Each actor has an address to which you can send a message. Each actor works in a separate thread. The actor model is used to organize email, some SOAP web services, and etc.

Despite the large number of synchronization methods, it is most often necessary to proceed from the problem being solved. For example, if we want to make a general incremental integer variable for several threads, it makes no sense to create a mutex or semaphore, it is more optimal to make the variable atomic. Always consider the overhead of creating locks and development time.

## 1.5 Automatic parallelization of programs

Parallel programming is a rather complicated manual process, so it seems obvious that it needs to be automated using a compiler. Such attempts

are made, however, the efficiency of auto-parallelization is not yet sufficient, because good indicators of parallel acceleration are achieved only for a limited set of simple for-cycles in which there are no data dependencies between iterations and the number of iterations cannot change after the start of the cycle. But even if the two indicated conditions are satisfied in a certain for-cycle, but it has a complex non-obvious structure, then its parallelization will not be performed. Types of automatic parallelization:

- *Fully automatic:* participation of the programmer is not required, all actions are performed by the compiler.
- *Semi-automatic:* programmer gives instructions to the compiler in the form of special keys that allow you to adjust some aspects of parallelization.

Weaknesses of automatic parallelization:

- erroneous change in program logic
- speed reduction instead of increase.
- lack of manual parallelization flexibility.
- only cycles are efficiently parallelized.
- inability to parallelize programs with a complex algorithm of work.

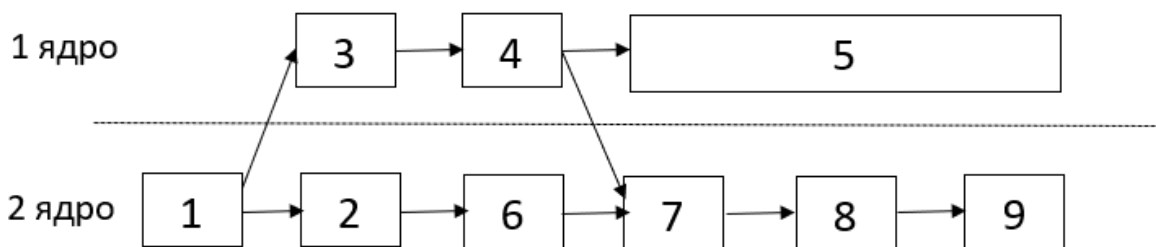
Here are examples of how the c-program in the src.c file can be automatically parallelized using some popular compilers:

- Compiler GNU Compiler Collection: `gcc -O3 -floop-parallelize-all -ftree-parallelize-loops=K -fdump-tree-parloops-details src.c`. In this case, the programmer can choose the value of the parameter K, which is recommended to be set equal to the number of cores (processors). Features of the auto-parallelization implementation in gcc are dedicated to an independent project:  
<https://gcc.gnu.org/wiki/AutoParInGCC>.
- Intel compiler: `icc -c -parallel -par-report file.cc`
- Oracle compiler: `solarisstudio -cc -O3 -xautopar -xloopinfo src.c`

## 1.6 The main approaches to parallelization

In practice, there are a large number of parallel programming patterns. However, all these patterns basically use three basic approaches to parallelization:

- **Data parallelization:** The programmer finds in the program an array of data whose elements the program sequentially processes in some func function. Then the programmer tries to break this data array into blocks that can be processed in func independently of each other. Then the programmer starts several threads at once, each of which executes func, but at the same time processes data blocks different from other flows in this function.
- **Instruction parallelization:** The programmer finds in the program sequentially called functions, the process of which does not affect each other (such functions do not change common global variables, and the results of one are not used by the other). Then the programmer starts these functions in parallel threads.
- **Parallelization of information flows:** A program is a set of functions that can be performed, and several functions can expect the result of the previous ones. In this case, each core performs the function for which the data is already ready. Consider this method as an example of an abstract dual-core processor, as the most difficult to understand. The structural algorithm shown in the figure 7 consists of 9 functions, some of which use the result of the previous function in their work. We assume that function 3 uses the result of the function 1, and function 7 uses the result of functions 4 and 6, etc., and also function 5 is executed in time approximately as much as functions 7, 8, and 9 combined. Then, on a dual-core machine, this parallelization method will be the optimal solution.



*Рис. 7: An example of the operation of a structural algorithm on a dual-core processor*

The three methods described are easier to understand by analogy from everyday life. Let two students in the construction team be given the task of sweeping the street and painting the fence. If students decide to use data parallelization, he will first sweep the street together, and then paint the fence together. If they decide to use parallelization according to the instructions, then one student will completely sweep the street, and the other will paint the entire fence at this time. This situation cannot be parallelized over the information flows, since these two actions are in no way dependent on each other. If we assume that they both need tools for work, then one of them must first go after them, and then both of them will begin to do their work.

In more cases, the decision to use the method is obvious due to the internal features of the parallelized program. The choice of method is determined by which one loads flows more evenly. Ideally, all threads should finish the work allocated to them approximately at the same time in order to optimally load the kernels (processors) and so that the threads that have finished their work do not stand idle while waiting for the completion of work by neighboring threads.

## **1.7 Atomic operations in a multithreaded program**

The main problem with parallel programming is the need to resolve conflicts while accessing the shared memory of several threads. To solve this problem, they usually try to streamline the access of streams to shared data using special tools - synchronization primitives. However, the question arises whether there are such atomic operations, the execution of which by several threads simultaneously does not require synchronization of actions, because these operations would be performed by the processor in one fell swoop, or, as they say, atomically (i.e., no other operation can push the previous atomic operation out of the processor before it is completed).

Almost all assembler instructions are such operations, since at a low level they use only those operations that are present in the processor command system, which means that they can be performed atomically (uninterruptedly). However, when compiling a C program, C language commands are usually translated into several assembler instructions. In this regard, the question arises about the possible existence of C-commands that are compiled into one assembler instruction. Such commands could not be "protected" by synchronization primitives (mutexes) in parallel programming.

However, there are very few such operations, and some of them can behave both atomically and non-atomically depending on the hardware platform for which the C program is compiled. Consider the simplest increment command

of an integer variable (int type) in C: "w++". You can easily verify (for example, using the gcc compiler's S ' key) that this command will be translated into three assembler instructions (take from memory, increase, put back):

```
1  movl  w, %ecx
2  addl  $1, %ecx
3  movl  %ecx, w
```

This means that it is not safe to perform the increment operation of a variable in several threads at the same time, because when executing assembler instruction 2, the thread can be interrupted and the processor transferred to another thread, which will receive an incorrect value of the underreduced variable.

It is logical to assume that assignment operations should not have the described problem. Indeed, in Assembler there is a separate instruction for writing the value of a variable to the specified address. Unfortunately, this assumption is not completely true: when assigning a variable of type char, this operation will be performed by a single assembler instruction. However, with other data types this cannot be said for sure. The general rule of thumb can be roughly stated as follows: "the atomicity of the assignment operation is guaranteed only for operations with data whose bit capacity does not exceed the processor bit capacity "

For example, when assigning an int variable to a 32-bit processor, one assembler instruction will be generated. However, when compiling the same operation on a 16-bit computer, two assembler instructions will be generated to write the low and high bits independently.

The formulated rule works with the assignment of variables and expressions, however, it cannot always be satisfied with the assignment of constants. Consider an example of C-code in which a 64-bit variable s (type uint64 \_t) is assigned a large number, obviously exceeding the 32-bit value:

```
1  uint64_t s;
2  s = 999999999999999L;
```

This code will be translated into the following assembler code on a 64-bit processor:

```
1  movabsq $999999999999999, %rsi
2  movq  %rsi, s
```

As you can see, the assignment operation was translated into two assembler instructions, which makes it impossible to safely parallelize such an operation.

This rule applies not only to the assignment operation, but also to the operation of reading a variable from memory, so any of these operations in a thread-safe environment will have to be protected by mutexes or critical sections.

A special case of atomic data changes is structure changes. To do this, we need to use a CAS operation with a pointer to this structure. Performing such an operation, the processor will create a second data structure with the specified fields and compare it with the old version of the structure. If the value of at least one field has changed, then it will atomically replace the pointer. There is overhead in this: even a simple change of one field of the structure requires the creation of a full copy of the structure, then to replace the pointer.

## 1.8 Lock-free data structures

In multi-threaded programs, problems with thread collaboration usually occur when accessing shared resources. In addition to the blocking approach using synchronization primitives, they also use the non-blocking approach. To avoid race conditions, you can use special non-blocking data structures. This approach is based on the use of atomic variables and lock-free or wait-free objects.

A shared object is called a lock-free object if it guarantees that some thread completes the operation on the object in a finite number of steps, regardless of the results of other threads.

An object is wait-free if each thread completes an operation on an object in a finite number of steps.

The question may arise: why are non-blocking data structures needed, if synchronization primitives can be used to access the usual data structure. Lock-free structures have several advantages over blocking data structures. So, in terms of bandwidth, they exceed blocking ones by 1.5–3 times, however, both blocking and non-blocking queues have poor scalability with respect to the number of threads. In terms of the delay of elements in the queue, non-blocking queues also have better characteristics, but their advantage is quite small. Also, the use of synchronization primitives can lead to deadlock, and errors can also occur associated with forgetting to capture or release primitives.

Lock-free data structures do not contain locks and remain in a consistent state regardless of the number of threads accessing it at the same time. Such data structures can be organized using RMW - read, modify and write operations that occur atomically.

An example of RMW operation is CAS. In the C++ library, there are two options for implementing this operation: weak and strong (figure ??).



Weak version may return false in case when the read value was equal to the expected one. Strong always returns the correct value.

```
bool
compare_exchange_weak(_Tp& __e, _Tp __i, memory_order __s,
                      memory_order __f) noexcept

bool
compare_exchange_strong(_Tp& __e, _Tp __i, memory_order __s,
                       memory_order __f) noexcept
```

*Puc. 8: Signatures of CAS operations in the C++ library*

An alternative to CAS operations is a pair of LL/SC operations in ARM processors. The load-link operation loads a value from memory, and store-conditional sets a new value, but only if the memory area has not changed. To implement LL/SC operations, we had to change the cache structure so that a LINK flag is added to each cache line. The flag is set during LL operation and is reset during SC or cache line preemption. LL/SC operations are not subject to the ABA problem, however, false sharing may occur due to hardware implementation. In modern processors, the cache line length is 64 - 128 bytes, therefore, several variables can be in the same cache line. When working with multiple variables on the same line, LL/SC operations will have a common LINK flag, which can lead to incorrect operation. In order to avoid this problem, one should place one variable in a line.

```
1 struct data {
2     atomic_int nShared1;
3     /* padding for cache line = 64 - sizeof(atomic_int) = 60 byte */
4     char _padding1[60];
5     atomic_int nShared2;
6     /* padding for cache line=60 byte */
7     char _padding2[60];
8 };
```

CAS operation can be quite easily implemented using LL / SC operations:

```
1 bool CAS(int *pAddr, int nExpected, int nNew) {
2     if (LL(pAddr) == nExpected)
3         return SC(pAddr, nNew);
4     return false;
5 }
```

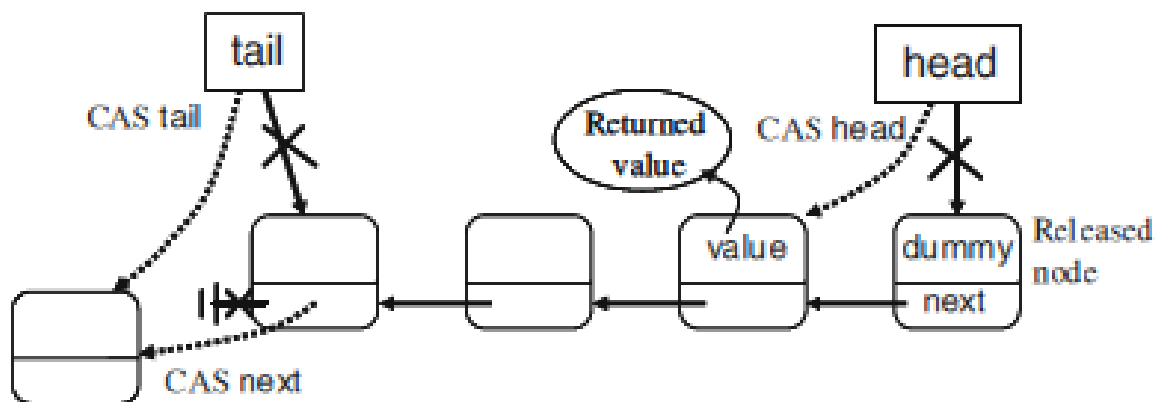
It is also important to understand that lock-free algorithms are sensitive to reordering machine instructions in their code. To avoid this, memory barriers

are used. The memory barrier  $X\_Y$  ensures that all X-operations before the barrier are executed before the Y-operations after the barrier begin to be executed. In theory, there are 4 types of barriers - LoadLoad, LoadStore, StoreLoad, StoreStore, but not all of them are implemented in all architectures. There are 4 processor memory models:

- **Relaxed model** – reordering of any memory access instructions is possible, even depending on the data (DEC Alpha).
- **Weak model** – reordering of any read and write instructions is possible, except for those that have data dependencies (ARM, PowerPC, Intel Itanium).
- **Strong model** – only reordering read to write is possible (x86).
- **Sequential consistency model** – any reordering is prohibited.

There are various lock-free data structures: queues (with strict and weakened order), stack, linked lists, hash tables. In C++, data structure data can be used by connecting various libraries. For example, Boost contains the implementation of the queue and stack, and Libcds contains all of the above.

An example of a lock-free data structure is the Michael-Scott queue. This queue is implemented on the basis of a singly linked list and two pointers, one of which points to the head of the list (dummy node), and the other to the tail (рисунок 9).



*Рис. 9: Michael-Scott queue*

Consider the simplified queue code from the libcds library. Below is the enqueue function - adding to the queue. First, the passed value is put in node. Then we try to put it in the tail of the line. After receiving the current

tail, the pointer advances until it reaches the actual tail. Then the value is put at the end of the queue and the value of the inserted element is assigned to the tail.

```
1  bool enqueue (value_type& val) {
2
3      node_type * pNew = node_traits::to_node_ptr(val);
4      node_type * t = m_pTail;
5
6      while (true) {
7          //продвижение хвоста
8          node_type * pNext = t->m_pNext.load();
9          if (pNext != nullptr) {
10             m_pTail.compare_exchange_weak(t, pNext);
11             continue;
12         }
13
14         //фактическая вставка нового элемента
15         node_type * tmp = nullptr;
16         if (t->m_pNext.compare_exchange_strong(tmp, pNew))
17             break;
18     }
19
20     //попытка продвинуть хвост
21     //в случае неудачи это сделает позже другой поток
22     m_pTail.compare_exchange_strong(t, pNew);
23
24     return true;
25 }
```

In order to get an element from the queue (dequeue function), so that the queue is not empty, and also that the tail and voices are advanced. The code is below.

```

1 value_type * dequeue() {
2
3     node_type * pNext;
4     node_type * h;
5
6     while (true) {
7         h = m_pHead;
8         pNext = h->m_pNext;
9
10        // кто-то успел изменить связь между головой и следующим узлом
11        if (m_pHead.load() != h)
12            continue;
13
14        //очередь пуста, голова всегда dummy node
15        if (pNext == nullptr)
16            return nullptr;
17
18        //хвост оказался не продвинут, пытаемся продвинуть
19        node_type * t = m_pTail.load();
20        if (h == t) {
21            m_pTail.compare_exchange_strong(t, pNext);
22            continue;
23        }
24
25        //продвигаем голову
26        if (m_pHead.compare_exchange_strong(h, pNext))
27            break;
28    }
29
30    return pNext;
31 }

```

## 2 Показатели эффективности параллельной программы

### 2.1 Параллельное ускорение и параллельная эффективность

Для оценки эффективности параллельной программы принято сравнивать показатели скорости исполнения этой программы при её запуске на нескольких идентичных вычислительных системах, которые различаются только количеством центральных процессоров (или ядер). На практике, однако, редко используют для этой цели несколько независимых аппаратных платформ, т.к. обеспечить их полную идентичность по всем параметрам достаточно сложно. Вместо этого, измерения проводятся на одной многопроцессорной (многоядерной) вычислительной системе, в которой искусственно ограничивается количество процессоров (ядер), задействованных в вычислениях. Это обычно достигается одним из следующих способов:

- Установка аффинности процессоров (ядер).
- Виртуализация процессоров (ядер).
- Управление количеством потоков выполнения.

**Установка аффинности.** Под аффинностью (processor affinity/pinning) понимается указание операционной системе запускать указанный поток/процесс на явно заданном процессоре (ядре). Установить аффинность можно либо с помощью специального системного вызова изнутри самой параллельной программы, либо некоторым образом извне параллельной программы (например, средствами "Диспетчера задач" или с помощью команды "start" с ключом "/AFFINITY" в ОС MS Windows, или команды "taskset" в ОС Linux). Недостатки этого метода:

- Необходимость модифицировать исследуемую параллельную программу (при использовании системного вызова изнутри самой программы).
- Невозможность управлять аффинностью на уровне потоков, т.к. обычно ОС позволяет устанавливать аффинность только для процессов (при установке аффинности внешними по отношению к параллельной программе средствами).

**Виртуализация процессоров (ядер).** При создании виртуальной ЭВМ в большинстве специализированных программ (например, VMWare,

VirtualBox) есть возможность "выделить" создаваемой виртуальной машине не все присутствующие в хост-системе процессоры (ядра), а только часть из них. Это можно использовать для имитации тестового окружения с заданным количеством ядер (процессоров). Например, на рисунке 10 показано, что для настраиваемой виртуальной машины из восьми доступных физических (и логических) процессоров доступными являются только три.

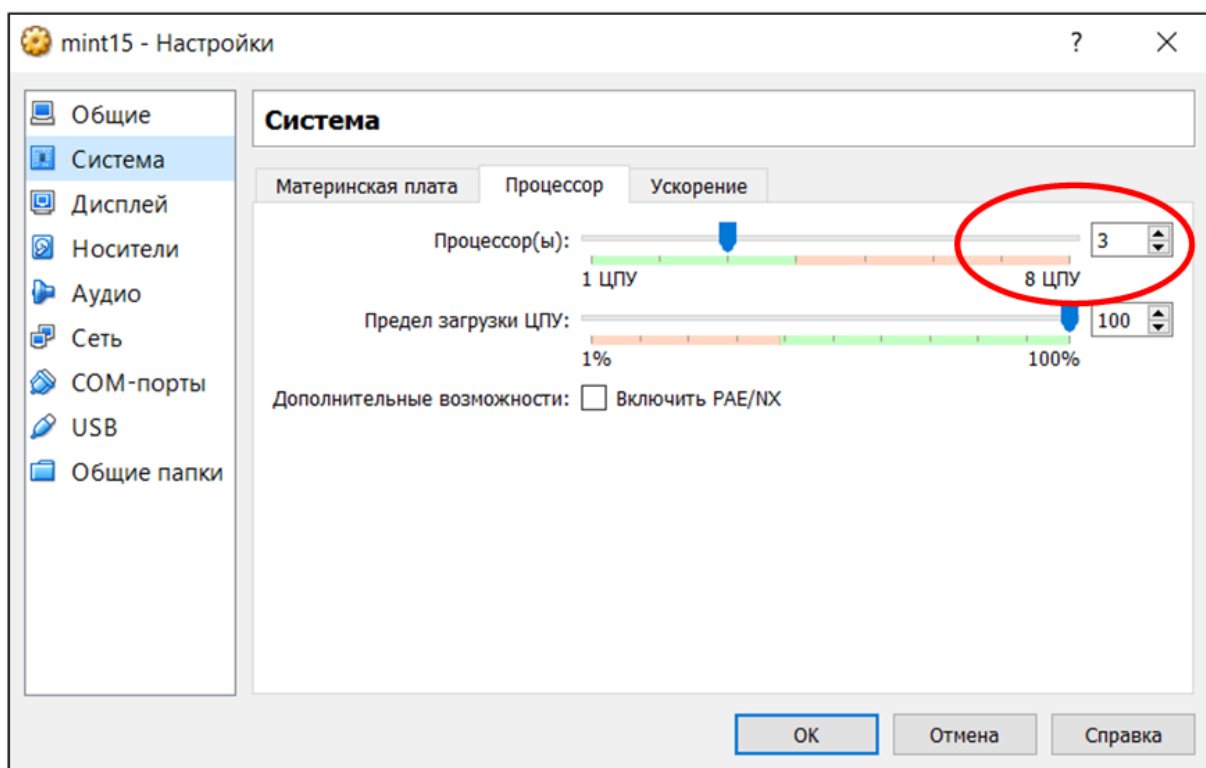


Рис. 10: Выбор количества виртуальных процессоров в Oracle VirtualBox

Недостатком описанного подхода являются накладные расходы виртуализации, которые непредсказуемым образом могут сказаться на результатах экспериментального измерения производительности параллельной программы. Достоинством виртуализации (по сравнению с управляемой аффинностью) является более естественное поведение тестируемой программы при использовании доступных процессоров, т.к. ОС не даёт жёстких указаний, что те или иные потоки всегда должны быть "привязаны" к заранее заданным процессорам (ядрам) – эта особенность позволяет более точно воспроизвести сценарий потенциального "живого" использования тестируемой программы, что повышает достоверность получаемых замеров производительности.

**Управление количеством потоков.** При создании параллельных программ достаточно часто количество создаваемых в процессе работы

программы потоков не задаётся в виде жёстко фиксированной величины. Напротив, оно является гибко конфигурируемой величиной  $p$ , выбор значения которой позволяет оптимальным образом использовать вычислительные ресурсы той аппаратной платформы, на которой запускается программа. Это позволяет программе "адаптироваться" под то количество процессоров (ядер), которое есть в наличии на конкретной ЭВМ.

Эту особенность параллельной программы можно использовать для экспериментального измерения её показателей эффективности, для чего параллельную программу запускают при значениях  $p = 1, 2, \dots, n$ , где  $n$  – это количество доступных процессоров (ядер) на используемой для тестирования многопроцессорной аппаратной платформе. Описанный подход позволяет искусственно ограничить количество используемых при работе программы процессоров (ядер), т.к. в любой момент времени параллельная программа может исполняться не более, чем на  $p$  вычислителях. Анализируя измерения скорости работы программы, полученные для различных  $p$ , можно рассчитать значения некоторых показателей эффективности распараллеливания (см. ниже).

**Параллельное ускорение (parallel speedup).** В отличие от применяемого в физике понятия величины ускорения как прироста скорости в единицу времени, в программировании под параллельным ускорением понимают безразмерную величину, отражающую прирост скорости выполнения параллельной программы на заданном количестве процессоров по сравнению с однопроцессорной системой, т.е.

$$S(p) = \frac{V(p)}{V(1)}, \quad (1)$$

где  $V(p)$  – средняя скорость выполнения программы на  $p$  процессорах (ядрах), выраженная в условных единицах работы в секунду (УЕР/с). Примерами УЕР могут быть количество просуммированных элементов матрицы, количество обработанных фильтром точек изображения, количество записанных в файл байт и т.п.

Считается, что значение  $S(p)$  никогда не может превысить  $p$ , что на интуитивном уровне звучит правдоподобно, ведь при увеличении количества работников, например, в четыре раза невозможно добиться выполнения работы в пять раз быстрее. Однако, как мы рассмотрим ниже, в экспериментах вполне может наблюдаться сверх-линейное параллельное ускорение при увеличении количества процессоров. Конечно, такой результат чаще всего означает ошибку экспериментатора, однако существуют ситуации, когда этот результат можно объяснить тем, что при увеличении количества процессоров не только кратно увеличивается их вы-

числительный ресурс, но так же кратно увеличивается объём кэш-памяти первого уровня, что позволяет в некоторых задачах существенно повысить процент кэш-попаданий и, как следствие, сократить время решения задачи.

**Параллельная эффективность (parallel efficiency).** Хотя величина параллельного ускорения является безразмерной, её анализ не всегда возможен без информации о значении  $p$ . Например, пусть в некотором эксперименте оказалось, что  $S(p) = 10$ . Не зная значение  $p$ , мы лишь можешь сказать, что при параллельном выполнении программа стала работать в 10 раз быстрее. Однако если при этом  $p = 1000$ , это ускорение нельзя считать хорошим достижением, т.к. в других условиях можно было добиться почти 1000 кратного прироста скорости работы и не тратить столь внушительные ресурсы на плохо распараллеливаемую задачу. Напротив, при значении  $p = 11$  можно было бы считать величину  $S(p) = 10$  вполне приемлемой.

Эта проблема привела к необходимости определить ещё один показатель эффективности параллельной программы, который бы позволил получить некоторую оценку эффективности распараллеливания с учётом количества процессоров (ядер). Этой величиной является **параллельная эффективность**

$$E(p) = \frac{S(p)}{p} = \frac{V(p)}{p \cdot V(1)} \quad (2)$$

Среднюю скорость выполнения программы  $V(p)$  можно измерить следующими двумя *неэквивалентными* методами:

- **Метод Амдала:** рассчитать  $V(p)$ , зафиксировав объём выполняемой работы (при этом изменяется время выполнения программы для различных  $p$ ).
- **Метод Густавсона-Барсиса:** рассчитать  $V(p)$ , зафиксировав время работы тестовой программы (при этом изменяется количество выполненной работы для различных  $p$ ).

Рассмотрим подробнее каждый из указанных методов в двух следующих подразделах.

## 2.2 Метод Амдала

При оценке эффективности распараллеливания некоторой программы, выполняющей фиксированный объём работы, скорость выполнения можно выразить следующим образом:  $V(p)|_{w=const} = \frac{w}{t(p)}$ , где  $w$  — это



общее количество УЕР, содержащихся в рассматриваемой программе,  $t(p)$  – время выполнения работы  $w$  при использовании  $p$  процессоров. Тогда выражение для параллельного ускорения примет вид:

$$S(p)|_{w=const} = \frac{V(p)}{V(1)} = \frac{w}{t(p)} = \frac{w}{t(1)} = \frac{t(1)}{t(p)}. \quad (3)$$

Запишем время  $t(1)$  следующим образом:

$$t(1) = t(1) + (k \cdot t(1) - k \cdot t(1)) = k \cdot t(1) + (1 - k) \cdot t(1), \quad (4)$$

где  $k \in [0, 1)$  - это коэффициент распараллеленности программы, которым мы обозначим долю времени, в течение которого выполняется идеально распараллеленный код внутри рассматриваемой программы. Такой код можно выполнить ровно в  $p$  раз быстрее, если количество процессоров увеличить в  $p$  раз. Заметим, что коэффициент  $k$  никогда не равен единице, т.к. в любой программе всегда присутствует нераспараллеливаемый код, который приходится выполнять последовательно на одном процессоре (ядре), даже если их доступно несколько. Если для некоторой программы  $k = 0$ , то при запуске этой программы на любом количестве процессоров  $p$  она будет решаться за одинаковое время.

Учитывая, что в методе Амдала количество работы остаётся неизменным при любом  $p$  (т.к.  $w = const$ ), можно утверждать, что значение  $k$  не изменяется в проводимых экспериментах, следовательно можем записать:

$$t(p) = \frac{k \cdot t(1)}{p} + (1 - k) \cdot t(1), \quad (5)$$

где первое слагаемое даёт время работы распараллеленного в  $p$  раз идеально распараллеливаемого кода, а второе слагаемое – время работы нераспараллеленного кода, которое не меняется при любом  $p$ . Подставив формулу (5) в (3), получим выражение

$$S(p)|_{w=const} = \frac{t(1)}{t(p)} = \frac{t(1)}{\frac{k \cdot t(1)}{p} + (1 - k) \cdot t(1)} = \frac{1}{\frac{k}{p} + 1 - k},$$

которое перепишем в виде

$$S(p)|_{w=const} = S_A(p) = \left( \frac{k}{p} + 1 - k \right)^{-1} \quad (6)$$

более известном как **закон Амдала** – по имени американского учёного Джина Амдала, предложившего это выражение в 1967 году. До сих пор

в специализированной литературе по параллельным вычислениям именно этот закон является основополагающим, т.к. позволяет получить теоретическое ограничение сверху для скорости выполнения некоторой заданной программы при распараллеливании.

График зависимости параллельного ускорения от количества ядер изображен на рисунке 11:

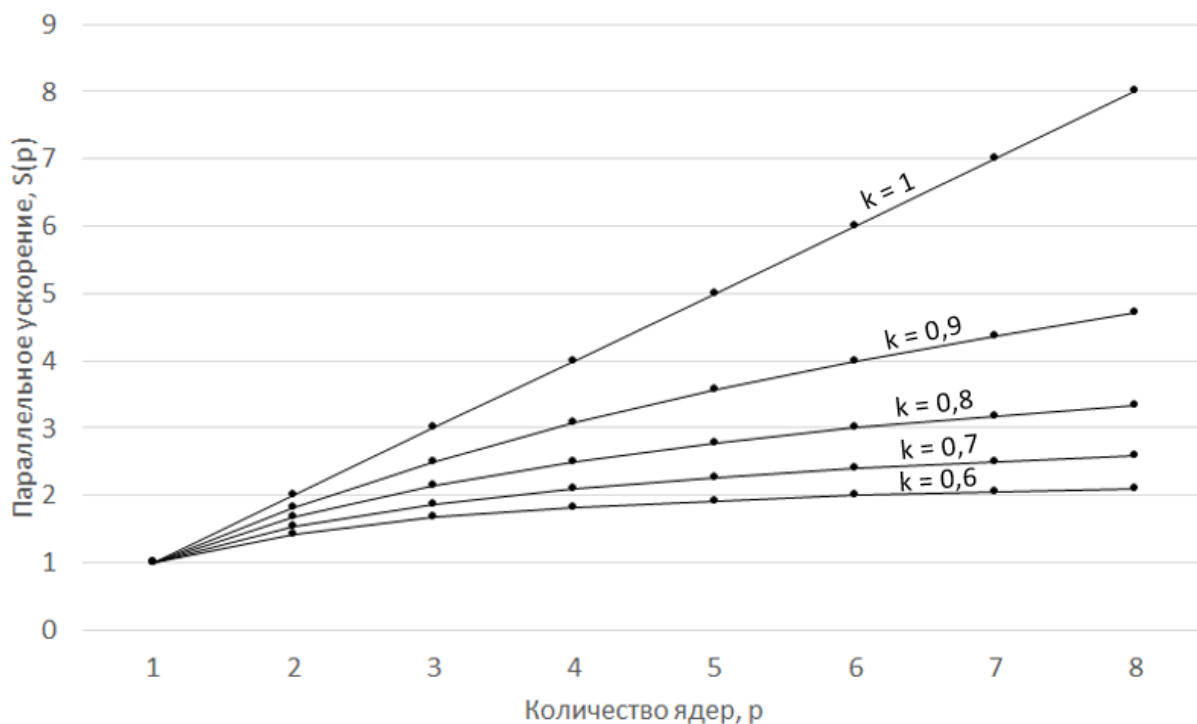


Рис. 11: График зависимости параллельного ускорения от количества ядер по Амдалу

Отметим, что выражение для расчёта параллельной эффективности при использовании метода Амдала можно получить, объединив формулы (2) и (6), а именно:

$$E_A(p) = (k + p - p \cdot k)^{-1} \quad (7)$$

Важным допущением закона Амдала является идеализация физического смысла величины  $k$ , состоящая в предположении, что идеально распараллеленный код будет давать линейный прирост скорости работы при изменении  $p$  от 0 до  $+\infty$ . При решении реальных задач приходится ограничивать этот интервал сверху некоторым конечным положительным значением  $p_{max}$  и/или исключать из этого интервала все значения, не кратные некоторой величине, обычно задающей размерность задачи.

Например, код программы, выполняющей конволюционное кодирование независимо для пяти равноразмерных файлов, может давать линейное ускорение при изменении  $p$  от 1 до 5, но уже при  $p = 6$  скорее всего покажет нулевой прирост скорости выполнения задачи (по сравнению с решением при  $p = 5$ ). Это объясняется тем, что конволюционное кодирование, также известно как "свёрточное", является принципиально нераспараллеливаемым при кодировании выбранного блока данных.

### 2.3 Метод Густавсона-Барсиса

При оценке эффективности распараллеливания некоторой программы, работающей фиксированное время, скорость выполнения можно выразить следующим образом:  $V(p)|_{t=const} = \frac{w(p)}{t}$ , где  $w(p)$  – это общее количество УЕР, которые программа успевает выполнить за время  $t$  при использовании  $p$  процессоров. Тогда выражение (1) для параллельного ускорения примет вид:

$$S(p)|_{t=const} = \frac{V(p)}{V(1)} = \frac{w(p)}{t} : \frac{w(1)}{t} = \frac{w(p)}{w(1)}. \quad (8)$$

Запишем количество работы  $w(1)$  следующим образом:

$$w(1) = w(1) + (k \cdot w(1) - k \cdot w(1)) = k \cdot w(1) + (1 - k) \cdot w(1), \quad (9)$$

где  $k \in [0, 1)$  – это уже упомянутый ранее коэффициент распараллеленности программы. Тогда первое слагаемое можно считать количеством работы, которая идеально распараллеливается, а второе – количество работы, которую распараллелить не удастся при добавлении процессоров (ядер).

При использовании  $p$  процессоров количество выполненной работы  $w(p)$  очевидно станет больше, при этом оно будет состоять из двух слагаемых:

- количество нераспараллеленных условных единиц работы  $(1 - k) \cdot w(1)$ , которое не изменится по сравнению с формулой (9).
- количество распараллеленных УЕР, объём которых увеличиться в  $p$  раз по сравнению с формулой (??), т.к. в работе будет задействовано  $p$  процессоров вместо одного.

Учитывая сказанное, получим следующее выражение для  $w(p)$ :

$w(p) = p \cdot k \cdot w(1) + (1 - k) \cdot w(1)$ , тогда с учетом формулы (8) получим:  $\frac{w(p)}{w(1)} = \frac{p \cdot k \cdot w(1) + (1 - k) \cdot w(1)}{w(1)}$ , что позволяет записать:

$$S(p)|_{t=const} = S_{GB}(p) = p \cdot k + 1 - k \quad (10)$$

Приведённое выражение называется **законом Густавсона-Барсиса**, который Джон Густавсон и Эдвин Барсис сформулировали в 1988 году.

## 2.4 Модификация закона Амдала (по проф. Бухановскому)

В реальных вычислительных системах ОС тратит ресурсы на создание и удаление новых потоков. Время, затраченное на эти операции не учитывается в законе Амдала. Параллельное ускорение  $S(p)$  зависит от количества ядер и доли распараллеливаемых операций, но не зависит от количества последних. Выведем формулу в которой количество операций для которых необходимо создать поток будет учитываться.

Пусть  $N$  – количество распараллеливаемых операций,  $M$  – количество нераспараллеливаемых операций,  $t_c$  – время выполнения одной операции,  $p$  – количество вычислителей(ядер),  $T_i$  – время выполнения программы при использовании  $i$  параллельных потоков на  $i$  вычислителях,  $\alpha$  – некий масштабирующий коэффициент, инкапсулирующий в себе количество времени, требуемого на создание, удаление потока и прочие накладные операции. По формуле (3),  $S(p) = \frac{T_1}{T_p}$ .

Найдем сначала  $T_1$ . Так как это код выполняется линейно, то время затраченное на его выполнение будет равно количеству операций помноженному на время выполнения одной операции:  $T_1 = t_c(N + M)$ .

Время выполнение распараллельной программы  $T_p$  включает в себя время на создание потока:  $t_c\alpha(p - 1)N$  (нужно создать  $(p - 1)$  новых потоков, так как главный поток уже создан и для каждого затратить какое-то время  $\alpha$ ), время работы распараллеливаемого кода на всех ядрах:  $\frac{t_cN}{p}$  и время работы нераспараллеливаемого кода  $t_cM$ . Итого, разделив  $T_1$  на  $T_p$ , получим формулу закона Амдала по проф. Бухановскому:

$$S(p, N) = \frac{T_1}{T_p} = \frac{N + M}{\alpha(p - 1)N + \frac{N}{p} + M} \quad (11)$$

Из формулы (11) видно, что с ростом количество ядер после определенного предела  $S(p, N)$  не будет расти как в законе Амдала, так как время будет тратиться много времени на создание новых потоков. На рисунке 12 наглядно видно, что  $S(p, N)$  уменьшается при большом количестве потоков и становится заметно меньше  $S(p)$  по Амдалу даже при небольшом значении  $\alpha$ .

При  $N = 100$ ,  $M = 20$ ,  $\alpha = 0.05$

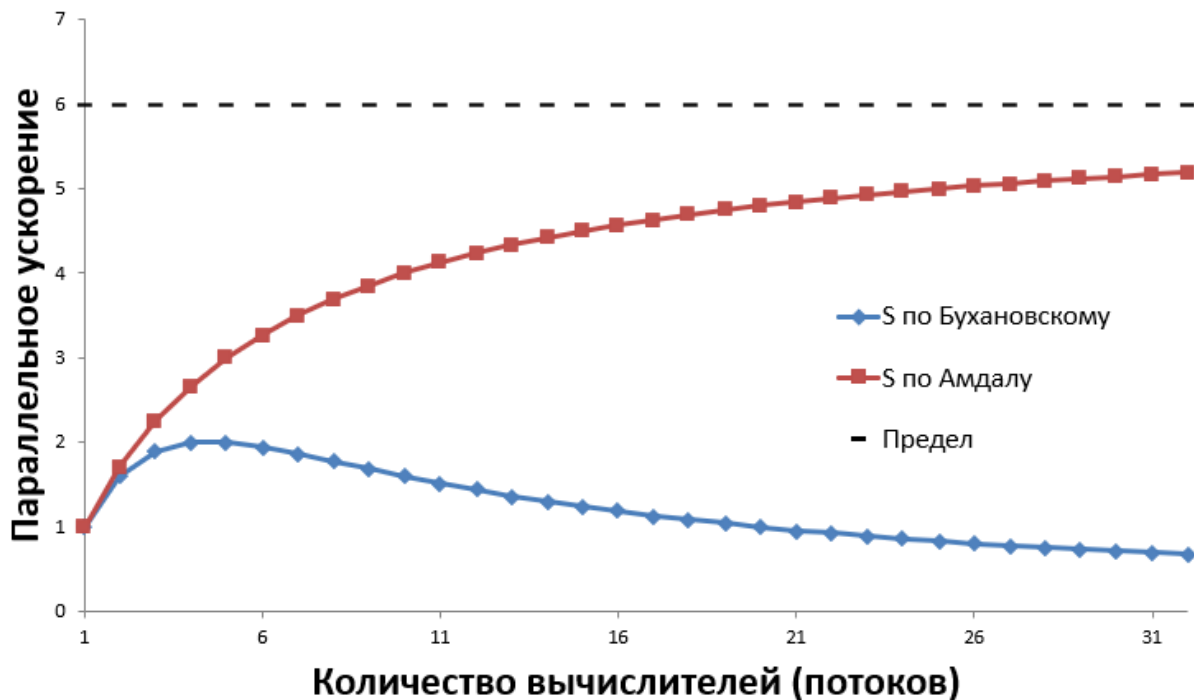


Рис. 12: График зависимости параллельного ускорения от количества потоков

## 2.5 Измерение времени выполнения параллельных программ

**Инструменты измерения времени.** Измерение времени работы программы в языке С не является сложной проблемой, однако при параллельном программировании возникает ряд специфических сложностей при выполнении этой операции. Далеко не все функции, пригодные для измерения времени работы последовательной программы, подойдут для измерения времени работы многопоточной программы.

Например, если в однопоточной программе для измерения времени работы участка кода использовать функции `ctime` или `localtime`, то они успешно справятся с поставленной задачей. Однако после распараллеливания этого участка кода возможно возникновение трудноидентифицируемых проблем с неправильным измерением времени, т.к. обе указанные функции имеют внутреннюю `static`-переменную, которая при попытке изменить её одновременно несколькими потоками может принять непредсказуемое значение.

С целью решить описанную проблему в некоторых С-компиляторах (например, `gcc`) были реализованы потокобезопасные (`thread-safe`, `re-entrant`) версии этих функций: `ctime_r` и `localtime_r`. К сожалению, эти функции до-

ступны не во всех компиляторах. Например, в компиляторе Visual Studio аналогичную проблему решили использованием функций с совсем иными именами и API: GetTickCount, GetLocalTime, GetSystemTime. Перечислим для полноты изложения некоторые другие gcc-функции, которые также позволяют измерять время: time, getrusage, gmtime, gettimeofday.

Ещё одна стандартная C-функция clock также не может быть использована для измерения времени выполнения многопоточных программ. Однако причина этого не в отсутствии реэнтерабельности, а в особенностях способа, которым эта функция рассчитывает прошедшее время: clock возвращает количество тиков процессора, которые были выполнены при работе программы суммарно всеми её потоками. Очевидно, что это количество остается почти неизменным при выполнении программы разным количеством потоков ("почти", т.к. накладные расходы на создание, удаление и управление потоками предлагается в целях упрощения изложения считать несущественными).

В итоге оказалось, что удовлетворительного *кросс-платформенного* решения для потокобезопасного измерения времени с высокой точностью (до микросекунд) средствами чистого языка C пока не существует. Проблему, однако, можно решить, используя сторонние библиотеки, выбирая те из них, которые имеют реализацию на целевых платформах.

Выгодно выделяется среди таких библиотек система OpenMP, которая реализована в абсолютном большинстве современных компиляторов для всех современных операционных систем. В OpenMP есть две функции для измерения времени: `omp_get_wtime` и `omp_get_wtick`, которые можно использовать в C-программах, если подключить заголовочный файл `omp.h` и при компиляции указать нужный ключ (например, в gcc это ключ `"-fopenmp"`).

**Погрешность измерения времени.** Другим интересным моментом при измерении времени работы параллельной программы является способ, с помощью которого исследователь исключает из замеров различные случайные погрешности, неизбежно возникающие при эксперименте в работающей операционной системе, которая может начать процесс обновления или оптимизации, не уведомляя пользователя. Общепринятыми является способ, при котором исследователь проводит не один, а сразу  $N$  экспериментов с параллельной программой, не меняя исходные данные. Получается  $N$  замеров времени, которые в общем случае будут различными вследствие различных случайных факторов, влияющих на проводимый эксперимент. Далее чаще всего используется один из следующих методов:

1. *Расчёт доверительного интервала:* с учётом всех  $N$  измерений

рассчитывается доверительный интервал, например, с помощью метода Стьюдента.

2. *Поиск минимального замера:* среди  $N$  измерений выбирается наименьшее и именно оно используется в качестве окончательного результата.

Первый метод даёт корректный результат, только если ошибки замеров распределены по нормальному закону. Чаще всего это так, поэтому применение метода оправдано и позволяет получить дополнительную информацию о возможном применении тестируемой программы в живых условиях работающей ОС.

Второй метод не предъявляет требований к виду закона распределения ошибки измерений и этим выгодно отличается от предыдущего. Кроме того, при больших  $N$  выбор минимального замера позволит с большой вероятностью исключить из эксперимента все фоновые влияния операционной системы и получить в качестве результата точное измерение времени работы программы в идеальных условиях.

**Практический пример.** Сравним на примере описанные выше методы избавления от погрешности экспериментальных замеров времени. Будем измерять накладные расходы OpenMP на создание и удаление потоков следующим образом:

```
1  for (i = 1; i < 382; i++) {
2      omp_set_num_threads(i);
3      double T1 = omp_get_wtime();
4      #pragma omp parallel // parallel section start
5      #pragma omp master
6          s++; // parallel section end
7      double T2 = omp_get_wtime();
8      print_delta(T1, T2);
9  }
```

В строке 3 мы даём OpenMP указание, чтобы при входе в параллельную область, расположенную далее в программе, было создано  $i$  потоков. Если не давать этого указания, OpenMP создаст количество потоков по количеству доступных в системе вычислителей (ядер или логических процессоров). В строке 4 мы запускаем параллельную область программы, OpenMP создаёт  $i$  потоков. В строке 5 мы даём указание выполнять последующую простейшую инструкцию лишь в одном потоке (остальные потоки не будут делать никакой работы. Это нужно, чтобы в замеряемое время работы попали только расходы на создание/удаление потоков, а все

прочие расходы терялись бы на их фоне. В строке 6 заканчивается параллельная область, OpenMP удаляет из памяти  $i$  потоков. Более подробное описание использованных команд OpenMP можно найти в разделе 3.3 "Технология OpenMP" данного учебного пособия.

Эксперименты с приведённой программой проводились на компьютере с процессором Intel Core i5 (4 логических процессора) с 8 гигабайт ОЗУ в операционной системе Debian Wheezy. Опытным путём было выявлено, что использованная операционная система на доступной аппаратной платформе не может создать более 381 потока в OpenMP-программе (этим объясняется значение в строке 1). Было проведено в общей сложности  $N=100$  экспериментов, результаты которых обрабатывались каждым из двух описанных методов. Полученные результаты приведены на рисунке 13.

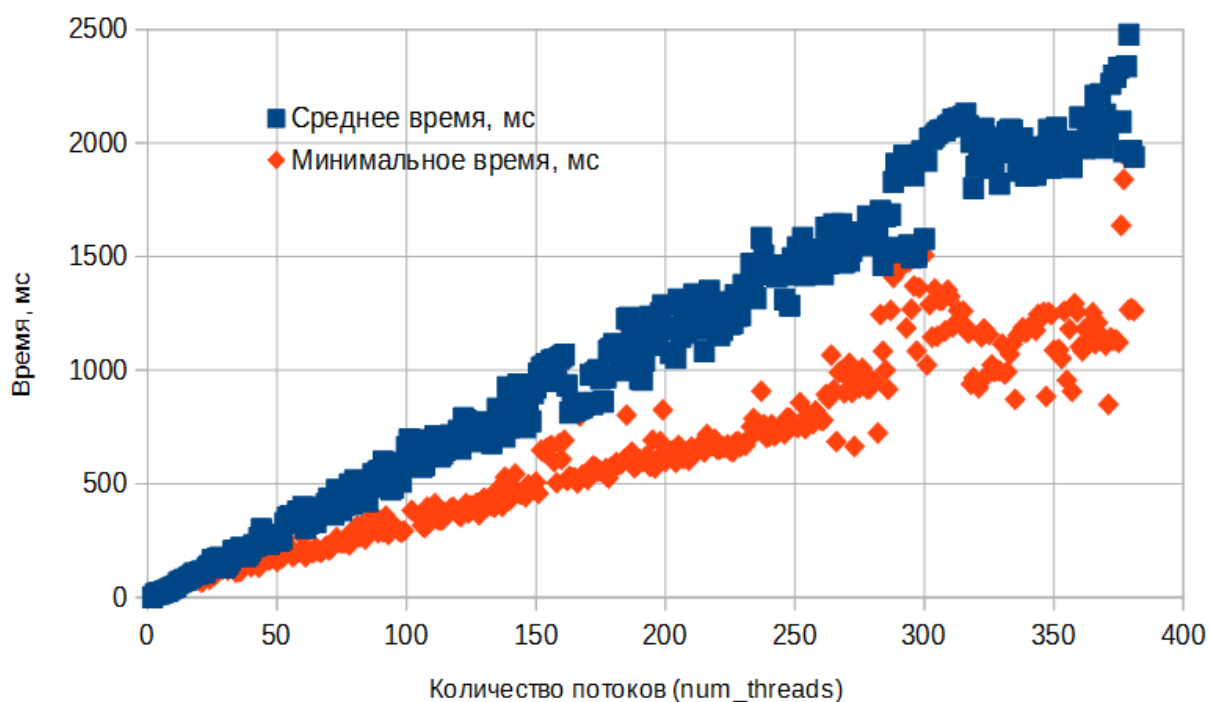


Рис. 13: Результаты измерения накладных расходов OpenMP при создании и удалении потоков

По оси ординат откладывается измеренная величина ( $T_2 - T_1$ ) в миллисекундах, по оси абсцисс – значения переменной  $i$ , означающие количество создаваемых потоков. Верхний график, состоящий из синих квадратов, показывает усреднённую величину ( $T_2 - T_1$ ) по 100 проведённым экспериментам. Доверительный интервал при этом не показан, т.к. он загромождал бы график, не добавляя информативности, однако ширина доверительного интервала с уровнем доверия 90% приблизительно



соответствует разбросу по вертикали квадратов верхнего графика для соседних значений  $i$ .

Нижний график, состоящий из ромбов, представляет собой минимальные из 100 проведённых замеров величины  $(T_2 - T_1)$  для указанных на оси абсцисс значений  $i$ . Видим, что даже большого количества экспериментов оказалось недостаточно, чтобы нижний график имел бы гладкую непрерывную структуру без заметных флуктуаций.

## 3 Практические аспекты параллельного программирования

### 3.1 Отладка параллельных программ

Средства отладки параллельных программ встроены в большинство популярных интегрированных сред разработки (IDE), например: Visual Studio, Eclipse CDT, Intel Parallel Studio и т.п. Эти средства включают в себя удобную визуализацию временных диаграмм исполнения потоков, автоматический поиск подозрительных участков программы, в которых могут наблюдаться гонки данных и взаимоблокировки.

Несмотря на эффективность существующих инструментов отладки, при работе в дебаггере (debugger) с параллельной программой возникают существенные затруднения, т.к. для своего корректного функционирования отладчик добавляет в машинный код исходной параллельной программы дополнительные инструкции, которые изменяют временную диаграмму выполнения потоков по отношению друг к другу. Это может приводить к ситуациям, когда при тестировании программы в отладчике не наблюдаются гонки данных и взаимоблокировки, которые при запуске Release-версии программы проявятся в полной мере.

Также при отладке многопоточной программы следует иметь в виду, что её поведение (как при штатной работе, так и при отладке) может существенным образом различаться при использовании одноядерного и многоядерного процессора. При запуске нескольких потоков на одноядерной машине они будут выполняться в режиме деления времени, т.е. последовательно. Значит, в этом случае не будут наблюдаться многие проблемы с совместным доступом к памяти и обеспечением когерентности кэшей, присущие многоядерным системам. Кроме того, при отладке программы на одноядерной системе программист может использовать неявные приёмы обеспечения последовательности выполнения операций.

Например, программист может некорректно предполагать, что при выполнении высокоприоритетного потока низкоприоритетный поток не может завладеть процессором. Это предположение корректно только в одноядерной системе, ведь при наличии нескольких ядер и малом количестве высокоприоритетных потоков вполне может наблюдаться ситуация, когда низкоприоритетный поток завладеет одним из ядер, при одновременной работе высокоприоритетного потока на соседнем ядре.

### 3.2 Менеджеры управления памятью для параллельных программ

При вызове функций `malloc/free` в однопоточной программе не возникает проблем даже при довольно высокой интенсивности вызовов одной из них. Однако в параллельных программах эти функции могут стать узким местом, т.к. при их одновременном использовании из нескольких потоков происходит блокировка общего ресурса (менеджера управления памятью), что может привести к существенной деградации скорости работы многопоточной программы.

Получается, что несмотря на формальную потокобезопасность стандартных функций работы с памятью, они могут стать потоко неэффективными при очень интенсивной работе с памятью нескольких параллельно работающих потоков.

Для решения этой проблемы существует ряд сторонних программ, называемых "Менеджер управления памятью (МУП)" (Memory Allocator), как платных, так и бесплатных с открытым исходным кодом. Каждое из них обладает своими достоинствами и недостатками, которые следует учитывать при выборе. Перечислим наиболее распространённые МУП с указанием ссылок на официальные сайты:

- `tcmalloc`: <http://goog-perftools.sourceforge.net/doc/tcmalloc.html>
- `ptmalloc`: <http://www.malloc.de/malloc/ptmalloc3-current.tar.gz>
- `dmalloc`: <http://dmalloc.com/>
- `HOARD`: <http://www.hoard.org/>
- `nedmalloc`: <http://www.nedprod.com/programs/portable/nedmalloc/>
- `jemalloc`: <http://jemalloc.net/>
- `mimalloc`: <https://github.com/microsoft/mimalloc>

Перечисленные МУП разработаны таким образом, что ими можно "незаметно" для параллельной программы подменить стандартные МУП библиотеки `libc` языка C. Это значит, что выбор конкретного МУП никак не влияет на исходный код программы, поэтому общая практика использования сторонних МУП такова: параллельная программа изначально создаётся с использованием МУП `libc`, затем проводится профилирование

работающей программы, затем при обнаружении узкого места (bottleneck) в функциях malloc/free принимается решение заменить стандартный МУП одним из перечисленных.

Также стоит отметить, что некоторые технологии распараллеливания (например, Intel TBB) уже имеют в своём составе специализированный МУП, оптимизированный для выполнения в многопоточном режиме.

### 3.3 Технология OpenMP

**Краткая характеристика технологии.** Первая версия стандарта OpenMP появилась в 1997 году при поддержке крупнейших IT-компаний мира (Intel, IBM, AMD, HP, Nvidia и др.). Целью нового стандарта было предложить кроссплатформенный инструмент для распараллеливания, который был бы более высокоуровневый, чем API управления потоками, предлагаемые операционной системой. На данный момент OpenMP стандартизована для трёх языков программирования: C, C++ и Фортран.

**Поддержка компиляторами.** Абсолютное большинство существующих современных компиляторов C/C++ поддерживают OpenMP версии 2.0 (например, как gcc, так и Visual Studio). Однако лишь немногие компиляторы поддерживают более новую версию OpenMP 4.0, поэтому далее при изложении материала будет в качестве "общего знаменателя" использоваться технология OpenMP 2.0.

OpenMP определяет набор директив препроцессору, которые дают указание компилятору заменить следующий за ними исходный код на его параллельную версию с помощью доступных компилятору средств, например с помощью POSIX Threads в Linux или Windows Threads в операционных системах Microsoft. Для корректной трансляции директив необходимо при компиляции указать специальный ключ, значение которого зависит от компилятора (примеры приведены в таблице 2).

Таблица 2: Ключи компиляторов для запуска OpenMP

Название компилятора	Ключ компилятору для включения OpenMP
Gcc	-fopenmp
icc (Intel C/C++ compiler)	-openmp
Sun C/C++ compiler	-xopenmp
Visual Studio C/C++ compiler	/openmp
PGI (Nvidia C/C++ compiler)	-mp

Помимо препроцессорных директив, OpenMP определяет набор библиотечных функций, для вызова которых в исходном коде потребуется подключить заголовочный файл OpenMP:

```
1 #include <omp.h>
```

**Отличительные особенности.** Среди прочих технологий распараллеливания OpenMP выделяется следующими важными и характеристиками:

- Инкрементное распараллеливание.
- Обратная совместимость.
- Высокий уровень абстракций.
- Низкий коэффициент трансформации.
- Поддержка крупнейшими IT-гигантами.
- Автоматическое масштабирование.

*Инкрементное распараллеливание.* OpenMP позволяет распараллеливать существующую последовательную программу с помощью небольших итераций-правок, на каждой из которых будет достигаться всё больший коэффициент распараллеленности программы. Эта особенность является уникальной, т.к. большинство других технологий предполагают существенное изменение структуры распараллеливаемой программы уже на первом этапе процесса распараллеливания, т.е. первая работоспособная параллельная версия программы появляется после длительного процесса отладки и программирования новых компонентов, которые неизбежно добавляются при распараллеливании. OpenMP лишён этого недостатка.

*Обратная совместимость.* Большинство программных технологий развиваются с обеспечением обратной совместимости (backward compatibility), когда более новая версия программы поддерживает работоспособность старых файлов. Термин "*прямая совместимость*" (forward compatibility) имеет противоположный смысл: файлы, созданные в программе новой версии, остаются работоспособными при использовании старой версии программы. В случае OpenMP это проявляется в том, что распараллеленная программа будет корректно скомпилирована в однопоточном режиме даже на старом компиляторе, который не поддерживает OpenMP. Важно отметить, что прямая совместимость обеспечивается, если при распараллеливании не используются библиотечные функции OpenMP, а присутствуют только препроцессорные директивы. При

наличии библиотечных функций для обеспечения обратной совместимости потребуется написать функции-заглушки в файле "omp.h" (некоторые компиляторы умеют генерировать эти заглушки при использовании специального ключа).

*Высокий уровень абстракций.* Одна единственная препроцессорная директива OpenMP после обработки компилятором приводит к существенной трансформации исходной программы с добавлением большого количества новой логики, отвечающей за определение доступного в системе количества процессоров, за запуск и уничтожение потоков, за распределение работы между потоками и т.п. Все эти операции OpenMP берёт на себя, взамен программист получает набор очень высокоуровневых инструментов распараллеливания. У высокоуровневых языков есть и традиционный недостаток: в OpenMP отсутствует возможность изменить некоторые внутренние детали работы с потоками (например, нельзя установить аффинность потоков или уменьшить накладные расходы на создание/удаление потоков).

*Низкий коэффициент параллельной трансформации (КПТ).* При распараллеливании существующей последовательной программы приходится вносить в неё достаточно большое количество изменений. Пусть КПТ – это отношение строк нового программного кода, который добавился в результате распараллеливания, к общему количеству строк кода в программе. В OpenMP КПТ обычно существенно ниже, чем у большинства других технологий распараллеливания. Это объясняется высоким уровнем абстракции языка OpenMP (см. предыдущий пункт).

*Поддержка крупнейшими IT-гигантами.* Уже при разработке OpenMP о его поддержке заявили крупнейшие игроки IT-мира. Это обеспечило не только высокое качество разработки стандарта, но и наличие готовых реализаций стандарта в популярных компиляторах. Несмотря на прошедшие два десятка лет OpenMP не растерял приверженцев и поддержка новейших версий OpenMP с достаточно малой задержкой появляется в компиляторах. Например, при текущей версии стандарта OpenMP 4.5 наиболее популярные компиляторы уже поддерживают версию OpenMP 4.0. Исключением является только фирма Microsoft. Их компилятор вот уже несколько версий неизменно поддерживает только OpenMP 2.0.

*Автоматическое масштабирование.* Низкоуровневые технологии распараллеливания (POSIX Threads, OpenCL) предлагают программисту вручную управлять количеством создаваемых потоков при выполнении параллельной работы. Это обеспечивает возможность гибко управлять и настраивать процесс создания потоков в зависимости от количества до-

ступных системе процессоров (ядер), но при этом требует от программиста большое количество неавтоматизируемой работы. В OpenMP управление масштабированием происходит в автоматическом режиме, т.е. OpenMP сам запрашивает у операционной системы количество доступных процессоров и выбирает количество создаваемых потоков. Но при необходимости OpenMP оставляет возможность устанавливать требуемое количество потоков вручную.

**Примеры OpenMP-программ.** Рассмотрим ниже простейшие примеры работающих параллельных программ, начиная с традиционного для программирования примера "Hello, World":

```
1 #pragma omp parallel
2 printf("Hello, world!");
```

Результатом работы будет выведенное несколько раз в консоль сообщение. Количество сообщений определяется количеством логических процессоров, доступных системе (например, при использовании технологии HyperThreading при двух ядрах количество логических процессоров будет равно четырём).

Действие директивы `pragma` распространяется на следующий за ней исполняемый блок. В данном случае это вызов функции `printf`, но можно было бы заключить произвольное количество операций в фигурные скобки, чтобы расширить исполняемый блок:

```
1 int i = 1;
2 #pragma omp parallel
3 {
4     printf("Hello, world!");
5     #pragma omp atomic
6     i++;
7 }
```

В этой программе заключенный в фигурные скобки блок операций выполняется одновременно на нескольких ядрах. При этом в строке 5 процессору даётся указание выполнить операцию "i++" атомарно, т.е. не параллельно, а последовательно каждым из потоков.

С одной стороны, это приводит к тому, что операция инкремента перестаёт быть распараллеленной, что снижает скорость многоядерного выполнения. С другой стороны, директива `atomic` в данном случае необходима, т.к. иначе могла бы возникнуть сложно обнаруживаемая проблема с гонкой данных, проявляющаяся в конфликте при записи данных в общую

область памяти одновременно несколькими потоками в переменную `i`. Заметим, что директива `atomic` может применяться только для однострочных простых команд присваивания.

Для изоляции более сложных составных команд с возможным вызовом пользовательских и системных функций следует использовать директиву `critical`, которая допускает (в отличие от директивы `atomic`) возможность расширения своей области действия на блок операций, заключённый в фигурные скобки? при этом каждая `critical`-секция может иметь имя, позволяющее сгруппировать разные критические секции по этому имени, чтобы предотвратить появление единой распределённой по всей программе критической секции:

```
1  int i = 1;
2  #pragma omp parallel
3  {
4      printf("Hello, world!");
5      #pragma omp critical
6      {
7          i++;
8          printf("i=%d\n", i);
9      }
10 }
```

В этом случае функция `printf` в строке 4 выполняется всеми потоками параллельно, что может привести к перемешиванию выводимых символов. Напротив, функция `printf` в строке 8 выполняется потоками строго по очереди, что предотвращает возможные конфликты между ними, однако замедляет выполнение программы из-за искусственного ограничения коэффициента распараллеленности.

Приведём пример распараллеливания программы, содержащей последовательный вызов функций `run_function1` и `run_function2`, которые не зависят друг от друга (т.е. не используют общих данных и результаты работы одной не влияют на результаты работы другой) и поэтому допускающих удобное *распараллеливание по инструкциям* в чистом виде:

```
1  #pragma omp parallel sections
2  {
3      #pragma omp section
4      run_function1();
5      #pragma omp section
6      run_function2();
7  }
```

Рассмотрим пример распараллеливания цикла с использованием



OpenMP. Пусть в каждую ячейку одномерного массива нужно записать индекс этой ячейки, возведённый в шестую степень:

```
1 int i; int a[10];
2 #pragma omp parallel for
3 for (i = 0; i < 10; ++i) {
4     a[i] = i*i*i*i*i*i;
5 }
```

Пусть указанная программа выполняется на двухъядерном процессоре. Тогда первый процессор рассчитает значения с  $a[0]$  по  $a[4]$ , второй процессор – значения с  $a[5]$  по  $a[9]$ . Видимо, что при записи в массив процессору не мешают друг другу, т.к. работают с разными частями массива. Попробуем оптимизировать предыдущий вариант, сократив количество операций умножения для возведения в шестую степень:

```
1 int i, tmp;
2 #pragma omp parallel for
3 for (i = 0; i < 10; ++i) {
4     tmp = i*i*i;          /* attempt to optimize */
5     a[i] = tmp*tmp;       /* error */
6 }
```

В указанном случае программа будет корректно работать только при наличии одного процессора (ядра). При наличии нескольких ядер будет наблюдаться состояние гонки данных при одновременной записи нового значения в переменную  $tmp$  (строка 4) несколькими потоками, в результате массив будет заполнен некорректно. Например, пусть первый поток, выполняющий итерацию  $i=2$  записал в  $tmp$  число 8. Теперь при вычислении  $a[2]$  поток попытается записать число  $8*8$ , однако если до начала строки 5 успеет вклиниться второй поток, работающей с итерацией  $i=7$ , то значение  $tmp$  превратится в  $7*7*7$ , а значение  $a[2]$ , рассчитываемое первым потоком, превратится в  $7^6$ , вместо положенных 64. Исправим допущенную ошибку следующим образом:

```
1 int i, tmp;
2 #pragma omp parallel for private(tmp)
3 for (i = 0; i < 10; ++i) {
4     tmp = i*i*i;
5     a[i] = tmp*tmp;
6 }
```

В директиве препроцессору появился новый элемент: `private`. Этот элемент задаёт через запятую перечень локальных (приватных) для каждого потока переменных. В данном случае такая переменная одна:  $tmp$ .

Другой равноценный способ исправить ошибку – это перенести объявление переменной "int tmp" внутрь параллельной области, что заставит OpenMP считать эту переменную локальной для каждого потока. Может возникнуть вопрос, почему в перечень локальных переменных не добавлена  $i$ . Ответ не очевиден: OpenMP по умолчанию считает переменную распараллеливаемого цикла локальной.

Любая переменная, объявленная внутри параллельной области, считается в OpenMP локальной, поэтому такие переменные не нужно указывать в списке. Любая переменная, объявленная вне этой области является глобальной (в нашем случае глобальной переменной является указатель на массив). Но если требуется явным образом указать на глобальность переменной, следует рядом с командой `private` использовать команду `shared(x, ...)`, где  $x$  задаёт список глобальных переменных.

Рассмотрим пример, в котором нужно рассчитать сумму и для дальнейшего исполнения формировать массив элементов следующего ряда:  $\{ 1^i, 2^i, 3^i, 4^i, 5^i \}$  для различных значений  $i$ , например:  $i = 1, 2, 3$ . Приведём ниже решение поставленной задачи, но умышленно допустим в ней ошибку:

```
1  int i, j, sum[3], tmp[5];
2  #pragma omp parallel for private(tmp)                      /* ошибка!!
   */
3  for (i = 0; i < 3; ++i) {
4      for (j = 1; j <= 5; ++j) {
5          tmp[j] = pow(j, i);
6          /* ошибка!! */
7          sum[i] = calculate_sum(tmp, 5);
8      }
}
```

В строке 2 происходит запуск параллельной области, но программист забывает указать, что переменные  $j$  и массив `tmp` должны быть локальными для каждого треда. Действительно, в строке 4 происходит инкремент общей для потоков переменной  $j$ , который выполняется всеми потоками одновременно. В этой ситуации потоки могут мешать друг другу, переписав чужое значение  $j$ . Исправим обе ошибки следующим образом:

```

1  int i, j, sum[3];
2  #pragma omp parallel for private(j) // OK
3  for (i = 0; i < 3; ++i) {
4      int tmp[5];
5      for (j = 1; j <= 5; ++j) {
6          tmp[j] = pow(j, i);           // OK
7          sum[i] = calculate_sum(tmp, 5);
8      }
9  }

```

Видим, что теперь переменная `j` явным образом обозначена локальной (`private`). С массивом `tmp` решение другое – он весь помещается внутрь параллельной области (т.е. у каждого потока будет свой собственный не зависящий от других экземпляров массива `tmp`). Почему же нельзя было просто указать переменную `tmp` в перечне команды `private`, как это было сделано для `j`? Ответ связан со спецификой языка C: переменная `tmp` является указателем, который при работе цикла не меняется, но меняется содержимое памяти, на которое указывает `tmp`. Это значит, что указывание `tmp` в качестве `private`-переменной не решило бы проблему с гонками данных, т.к. все потоки получили бы один и тот же адрес `tmp` и мешали бы друг другу, записывая новые значения по этому адресу.

Рассмотрим ещё одну типичную для параллельного программирования ошибку. Следующая программа считает сумму чисел от 1 до 100:

```

1  int i, sum = 0;
2  #pragma omp parallel for
3  for (i = 0; i < 100; ++i) /* error */
4      sum += i;

```

Переменная `sum` является глобальной, поэтому при попытке записать в неё новое значение потоки будут мешать друг другу. Чтобы исправить ошибку, нам придётся использовать локальную для каждого потока сумму, а затем потребуется сложить все эти локальные суммы:

```

1  int i, sum = 0, sum_private = 0;
2  #pragma omp parallel private (sum_private)
3  {
4      sum_private = 0;           /* repeated initialization! */
5      #pragma omp for
6      for (i = 0; i < 100; ++i)
7          sum_private += i;
8      #pragma omp atomic
9      sum += sum_private;
10 }

```

Видим начало параллельной области в строке 2 – именно в этом месте OpenMP создаёт несколько потоков. В строке 6 новые потоки не

создаются (т.к. отсутствует ключевое слово `parallel`), но входящие в цикл потоке делят итерации между собой, а не выполняют каждый все итерации целиком. В строке 8 рассчитавший свою частичную сумму поток пытается прибавить эту сумму к общей сумме. Это приходится делать с помощью директивы `atomic`, которая гарантирует, что потоки не будут мешать друг другу при перезаписи `sum`.

Ещё один сложный момент – это повторная инициализация переменной `sum_private` в строке 4: необходимость в этом возникает, т.к. OpenMP не инициализирует локальные переменные, даже если есть глобальные переменные с идентичными именами. Подобное решение призвано уменьшить накладные расходы на копирование переменных.

Описанный подход является работоспособным, однако он почти не используется на практике, т.к. стандарт OpenMP для целого класса подобных задач предлагает более высокоуровневое и простое решение. Оно состоит в использовании команды `reduction`:

```
1  int i, sum = 0;
2  #pragma omp parallel for reduction (+:sum)
3  for (i = 0; i < 100; ++i)
4      sum += i;
```

Команда `reduction` помечает перечисленные переменные как локальные, а в конце параллельной области все локальные переменные объединяет (агрегирует) в одну глобальную переменную с тем же именем, используя указанную операцию. В нашем случае операцией является суммирование. Но OpenMP допускает вместо знака `+` использовать `*`, `-`, `/`. Важно, что `reduction` кроме прочего выполняет инициализацию переменных не значениями исходных глобальных переменных, а наиболее соответствующими логики агрегации значениями: например, при суммировании переменная инициализируется нулём, а при умножении – единицей.

При распараллеливании цикла может оказаться, что итерации неравноразмерны по количеству выполняемой работы между собой. Это может привести к тому, что один поток справится с выделенной частью итераций намного быстрее второго потока и будет простаивать. Для решения этой проблемы OpenMP предлагает четыре разных способа распределения итераций по потокам.

- *Способ по умолчанию:* при этом итерации делятся на количество частей, равное количеству потоков; каждый поток выполняет после этого свою часть и не может взять чужую работу.

- *Статическое распределение (static)*: итерации разбиваются на части указанного пользователем размера; затем ещё до начала работы каждый поток получает фиксированное количество частей и выполняет только их без возможности переключиться на другие.
- *Динамическое распределение (dynamic)*: итерации разбиваются на части указанного пользователем размера; затем сразу начинается работа цикла и каждый поток получает новую часть итераций по мере завершения работы над предыдущей.
- *Управляемое распределение (guided)*: компилятор разбивает итерации на количество частей, равное удвоенному количеству потоков; затем сразу начинается работа цикла и каждый поток получает новую часть итераций по мере завершения работы над предыдущей, при этом размер нововыданной части уменьшается по сравнению с предыдущим разом, но не может стать меньше указанного пользователем константного значения.

Упомянутый в каждом из методов пользовательский параметр называется `chunk_size`. Каждый из указанных методов имеет свою область применения, в которой он может обеспечить максимальное параллельное ускорение. Отметим, что режимы `dynamic` и `guided` несмотря на свою логичность имеют и свои недостатки: они требуют существенных накладных расходов во время работы цикла по сравнению со `static`. Также важно понимать, что при выборе числа `chunk_size` необходимо учитывать особенности работы механизма кеширования.

Рассмотрим пример статического распределения итераций:

```

1  int i; double sum = 0;
2  #pragma omp parallel for reduction (+:sum) schedule(static,1)
3  for (i = 1; i < 100; ++i)
4      sum += 1.0/i;

```

При наличии трёх ядер OpenMP создаст три потока. Первому потоку достанутся итерации  $i = 1, 4, 7, \dots, 97$  второму – итерации  $i = 2, 5, 8, \dots, 98$ , третьему – итерации  $i = 3, 6, 9, \dots, 99$ . Обратим внимание, что выбор малого значения параметра `chunk_size = 1` в данном случае не имеет каких-либо негативных эффектов. Однако если бы  $i$  использовалась в качестве индекса при обращении к массиву, то предложенный вариант разбиения привёл бы к обращению в память не подряд

по последовательным адресам, а разреженно с шагом 3, что ухудшило бы показатели cache hit при использовании кэширования.

Рассмотрим ещё один пример:

```
1 double result1, result2, result3;
2 #pragma omp parallel num_threads(3)
3 {
4     #pragma omp for reduction (+:result1) nowait
5     for (i = 0; i < 100; ++i) result1 += i;
6     #pragma omp sections
7     {
8         #pragma omp section
9         result2 = calculate_pi();
10        #pragma omp section
11        result3 = calculate_e();
12    }
13 }
14 use_results(result1, result2, result3);
```

Здесь приводится пример, как можно указать OpenMP количество создаваемых потоков с помощью опции `num_threads` (строка 2), не ориентируясь на реально доступное количество ядер (процессоров) на компьютере. Далее три созданных потока делят между собой 100 итераций уже знакомым нам способом. Однако опция `nowait` позволяет первому справившемуся с работой потоку не дожидаться остальных, а перейти к следующей за циклом работе. За циклом в параллельном режиме выполняются две функции (строки 9 и 11). Каждая из функций заключена в секцию (`section`), которые должны иметь родительский элемент `sections`. В итоге первый освободившийся после цикла поток займётся вычислением функции в строке 9. Второй освободившийся поток вычислит функцию в строке 11. Третьему потоку не достанется работы помимо своей доли итераций в первом цикле. Общим требованием OpenMP к распараллеливаемым циклам является их *каноничность*. Цикл `for` называется *каноническим*, если можно при его начале заранее рассчитать количество предстоящих итераций. Это возможно, если одновременно выполняются следующие условия:

- внутри цикла нет операций `break` и `return`;
- внутри цикла нет операции `goto`, ведущей вовне цикла;
- переменная цикла (итератор) не изменяется внутри цикла;

При этом запись цикла должна иметь вид `"for (i = A; i < B; i+=C)"`, где числа A, B, C не должны меняться во время работы цикла. Второй параметр цикла может использовать не только знак `"<"`, но и `">"`, `">="`, `"<="`.

Третий параметр цикла может не только инкрементировать, но декрементировать переменную цикла (допускается краткая форма записи "i++").

Если итерация *k* влияет на результаты итерации *m*, то цикл нельзя распараллеливать, т.к. нельзя заранее предсказать порядок завершения итераций несколькими потоками. Ответственность за обнаружение таких конфликтов лежит на программисте. Например, OpenMP не обнаружит взаимозависимость итераций и скомпилирует следующую программу:

```
1  #pragma omp parallel for num_threads(2)
2  for(i = 1; i < 20; i++)
3      a[i] = 2*a[i - 1];
```

В этой программе поток 0 скорее всего не успеет заполнить элемент *a*[9] к тому моменту, когда поток 1 будет вычислять значение *a*[10] = 2\**a*[9].

### 3.4 Технология OpenCL

**Краткая характеристика технологии.** OpenCL — фреймворк для написания компьютерных программ, связанных с параллельными вычислениями на различных графических и центральных процессорах, а также FPGA. В OpenCL входят язык программирования, который основан на стандарте языка программирования Си C99, и интерфейс программирования приложений. OpenCL обеспечивает параллелизм на уровне инструкций и на уровне данных и является осуществлением техники GPGPU. OpenCL является полностью открытым стандартом, его использование не облагается лицензионными отчислениями. С помощью этой технологии можно производить гетерогенные параллельные вычисления (распределять задачи между разными устройствами).

Как мы уже знаем, можно распараллелить программу по задачам между небольшим числом производительных ядер (процессоры современных ПК) или по данным между тысячами простых медленных ядер (вычислительные ядра современных GPU). Именно для задач, решаемых с помощью распараллеливания по данным используется OpenCL.

**Архитектура технологии OpenCL.** В OpenCL разделяют два вида устройств: *host*, который управляет общей логикой и *device*, которые выполняют вычисления. В роли *хоста* обычно выступает центральный процессор, а в роли *device* - GPU и другие устройства. *Device* делится на вычислительные модули *computer units*, которые в свою очередь состоят из обрабатывающих элементов (*processing elements*) (рисунок 14).



Непосредственно вычисления производятся в обрабатывающих элементах устройства.

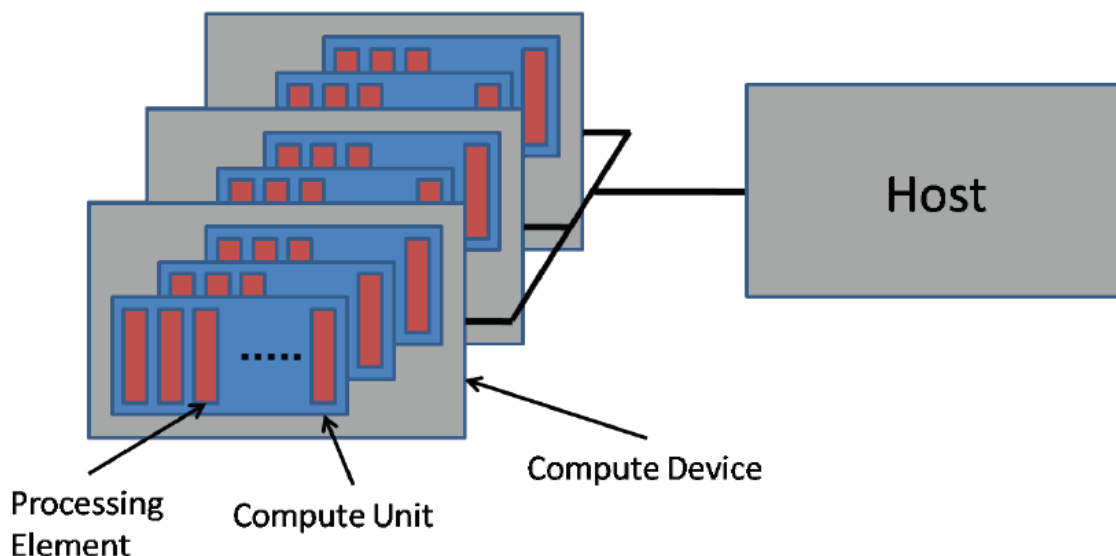


Рис. 14: Архитектура OpenCL

Физически *computer unit* представляет собой *work-group*, который состоит из ячеек *work-item*, которые и выполняют вычисления (рисунок 15).

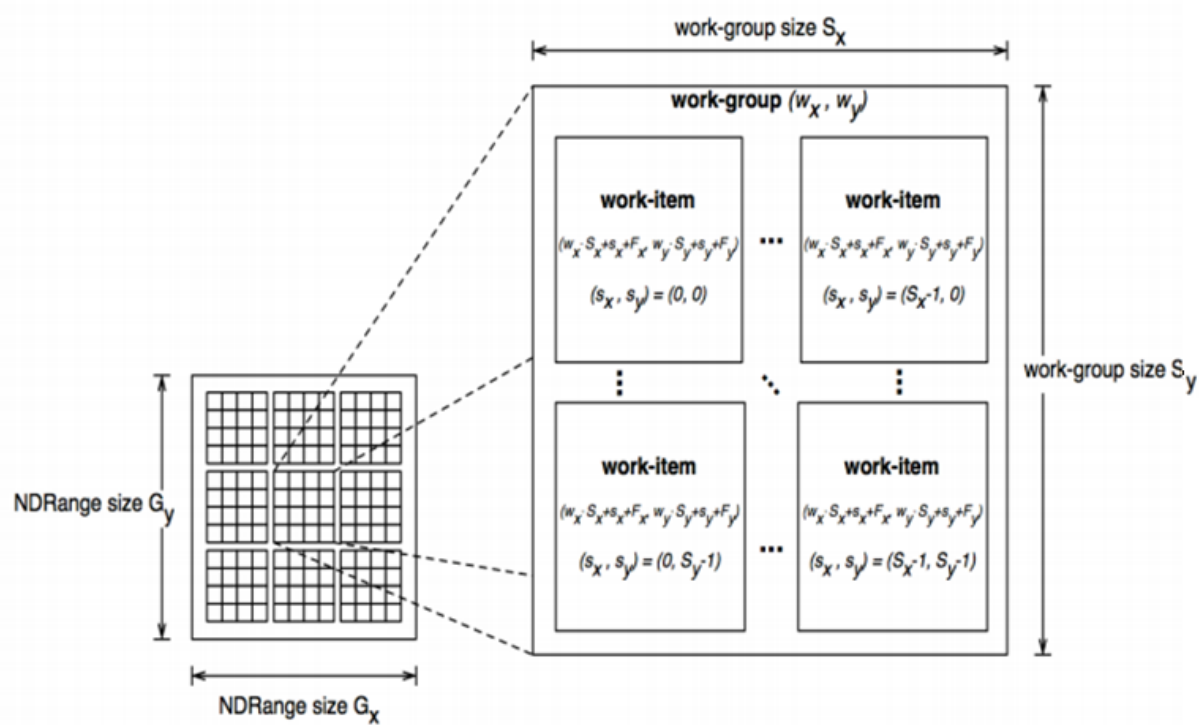


Рис. 15: Архитектура OpenCL - строение work-group элемента



Команды в OpenCL, образуют очередь. *Host* направляет команды на устройства. Эти команды становятся в очередь аналогичных команд. Можно реализовать очередь с соблюдением порядка и без соблюдения. Функции работы с получением ID *work-group* и *work-item* приведены на рисунке 16.

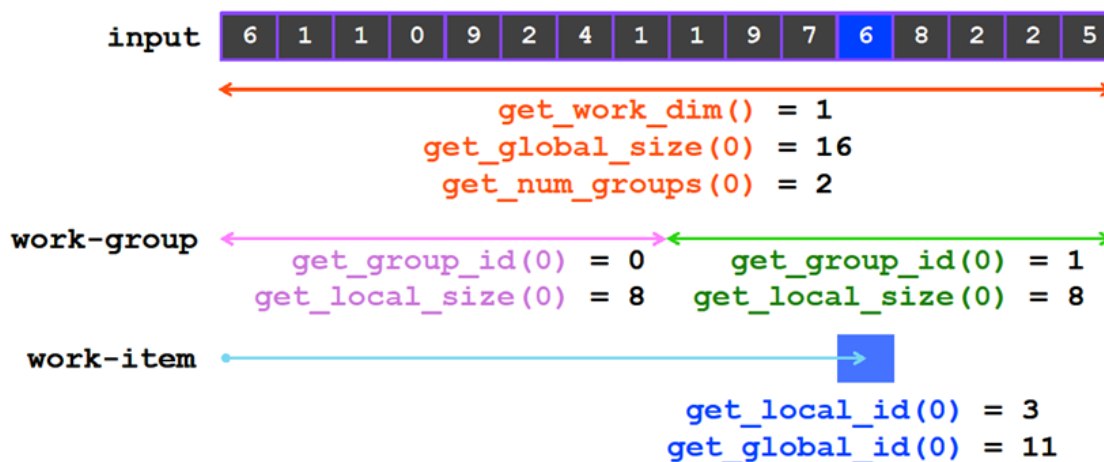


Рис. 16: OpenCL. Работа с *work-group* и *work-item*

**Виды памяти в OpenCL-устройствах.** Для взаимодействия с данными программист может использовать разные уровни памяти. На рисунке 17 видно, что существуют следующие виды памяти:

- Частная память (*private*). Самая быстрая из всех видов. Эксклюзивна для каждого элемента работы.
- Локальная память (*local memory*). Может быть использована компилятором при большом количестве локальных переменных в какой-либо функции. По скоростным характеристикам локальная память значительно медленнее, чем регистровая. Доступ из элементов работе в одной *work-group*.
- Константная память (*constant memory*). Достаточно быстрая из доступных GPU. Есть возможность записи данных с хоста, но при этом в пределах всего GPU возможно лишь чтение. Динамическое выделение в отличие от глобальной памяти в константной не поддерживается.
- Глобальная память (*global memory*). Самый медленный тип памяти, из доступных GPU. Глобальные переменные можно выде-

лить с помощью спецификатора, а также динамически. Глобальная память в основном служит для хранения больших объемов данных, поступивших на device с host'а. В алгоритмах, требующих высокой производительности, количество операций с глобальной памятью необходимо свести к минимуму.

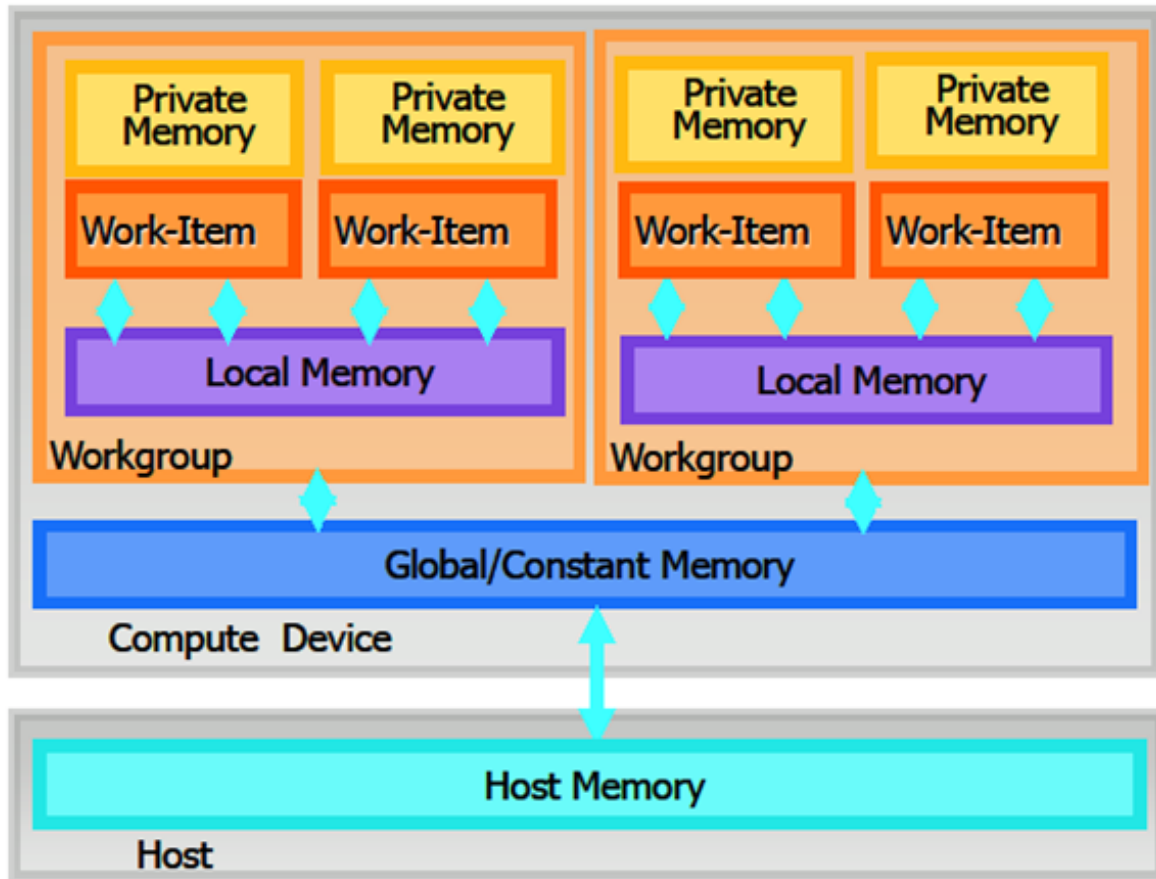


Рис. 17: Виды памяти в OpenCL-устройствах

Программист должен явно отдавать копирования между разной памятью.

Программа на OpenCL может включать в себя следующую последовательность действий:

1. **Выбор платформы:** `clGetPlatformIDs`, `clGetPlatformInfo`
2. **Выбор устройства:** `clGetDeviceIDs`, `clGetDeviceInfo`
3. **Создание вычислительного контекста:** `clCreateContextFromType`
4. **Создание очереди команд:** `clCreateCommandQueueWithProperties`
5. **Выделение памяти в виде буферов:** `clCreateBuffer`

6. **Создание объекта «программа»:** `clCreateProgramWithSource`
7. **Компиляция кода:** `clBuildProgram`
8. **Создание «ядра» (объект `kernel`):** `clCreateKernel`
9. **Работа с `Work-Group`:** `clGetKernelWorkGroupInfo`
10. **Выполнение ядра:** `clEnqueueNDRangeKernel`
11. **Ожидание выполнения ядра:** `clWaitForEvents`
12. **Profiling:** `clGetEventProfilingInfo`

Далее рассмотрим некоторые из этих действий подробнее.

**Выбор платформы, устройства и создание контекста.** Контекст (*context*) служит для управления объектами и ресурсами OpenCL. Все ресурсы OpenCL привязаны к контексту. С контекстом ассоциированы следующие данные (рисунок 18):

- устройства
- объекты программ
- ядра
- объекты памяти
- очереди команд.

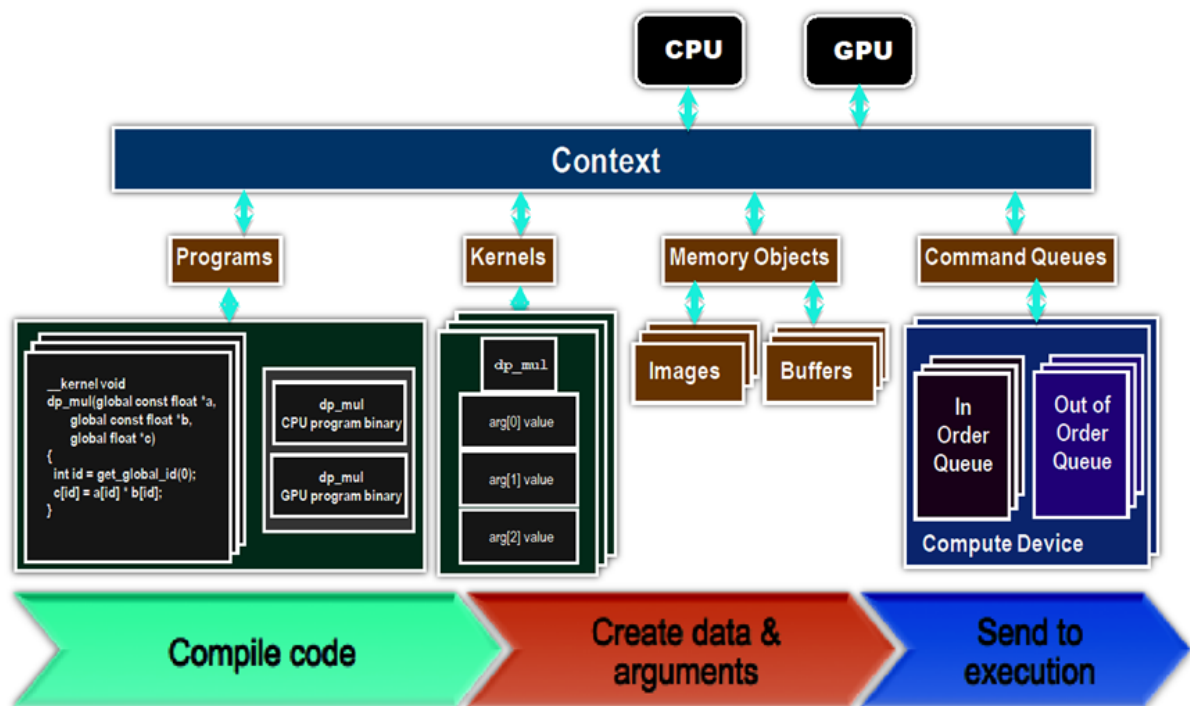


Рис. 18: Архитектура OpenCL - контекст

Можно получить информацию о платформе и вычислительных ядрах с помощью специальных функций, чтобы затем создать контекст:

- *clGetPlatformInfo()* - содержит информацию о платформе, на которой работает программа
- *clGetDeviceDs()* - содержит информацию о подключенных устройствах
- *clGetDeviceInfo()* - содержит информацию о данном девайсе: его тип, совместимость и тд.

Контекст можно создать при помощи функции *clCreateContext()*. Вот пример его создания:

```

/*1*/    //Get the platform ID
/*2*/    cl_platform_id platform;
/*3*/    clGetPlatformIDs(1, &platform, NULL);
/*4*/
/*5*/    //Get the first GPU device associated with the platform
/*6*/    cl_device_id device;
/*7*/    clGetDeviceIDs(platform, CL_DEVICE_TYPE_GPU, 1, &device, NULL);
/*8*/
/*9*/    //Create an OpenCL context for the GPU device
/*10*/    cl_context context;
/*11*/    context = clCreateContext(NULL, 1, &device, NULL, NULL, NULL);

```

*Рис. 19: Архитектура OpenCL - контекст*

В строку 3 мы получаем ID платформы, в строке 7 ID первого GPU на этой платформе, в строке 11 создаем контекст для этого девайса. Подробнее про аргументы, принимаемые этими функциями можно прочитать в документации. Есть также функция *clCreateContextFromType()* для создания контекста, ассоциированного с устройствами определенного типа.

**Ядро.** Ядром называется функция, являющаяся частью программы и параллельно исполняющаяся на устройстве. Ядро является аналогом потоковой функции. Часть, выполняющаяся на устройстве, состоит из набора ядер, объявленных с квалификатором **\_\_kernel**. Компилирование ядер может осуществляться во время исполнения программы с помощью функций API. Работа в рамках одной work group выполняется одновременно всеми work items.

При написании ядра можно использовать следующие квалификаторы для переменных:

- **\_\_global** или **global** – данные в глобальной памяти.
- **\_\_constant** или **constant** – данные в константной памяти.
- **\_\_local** или **local** – данные в локальной памяти.
- **\_\_private** или **private** – данные в частной памяти.
- **\_\_read\_only** и **\_\_write\_only** – квалификаторы режима доступа.


Скомпилировать код ядра можно с помощью функций *clCreateProgramWithSource()*, *clBuildProgram()* и *clCreateKernel()*. Пример компиляции и запуска программы по перемножению двух массивов приведен на рисунке 20.

```

// Build program object and set up kernel arguments
const char* source = "__kernel void dp_mul(__global const float *a, \n"
    "                                __global const float *b, \n"
    "                                __global float *c, \n"
    "                                int N) \n"
    "{ \n"
    "    int id = get_global_id (0); \n"
    "    if (id < N) \n"
    "        c[id] = a[id] * b[id]; \n"
    "} \n";

cl_program program = clCreateProgramWithSource(context, 1, &source, NULL, NULL);
clBuildProgram(program, 0, NULL, NULL, NULL, NULL);
cl_kernel kernel = clCreateKernel(program, "dp_mul", NULL);
clSetKernelArg(kernel, 0, sizeof(cl_mem), (void*)&d_buffer);
clSetKernelArg(kernel, 1, sizeof(int), (void*)&N);
// Set number of work-items in a work-group
size_t localWorkSize = 256;
int numWorkGroups = (N + localWorkSize - 1) / localWorkSize; // round up
size_t globalWorkSize = numWorkGroups * localWorkSize; // must be evenly divisible by localWorkSize
clEnqueueNDRangeKernel(cmd_queue, kernel, 1, NULL, &globalWorkSize, &localWorkSize, 0, NULL, NULL);

```



*Рис. 20: OpenCL - компиляция и запуск ядра*

Подробнее об остальных особенностях технологии OpenCL можно прочесть на ресурсе <http://docplayer.ru/37490743-Programmirovanie-na-opencl.html> и в официальной документации

1. OpenCL – официальный сайт: <http://www.khronos.org/opencl/>
2. Intel OpenCL: <http://software.intel.com/en-us/articles/intel-opencl-sdk/>
3. NVIDIA OpenCL: [http://www.nvidia.ru/object/cuda\\_opengl\\_new\\_ru.html](http://www.nvidia.ru/object/cuda_opengl_new_ru.html)
4. AMD OpenCL: <http://www.amd.com/us/products/technologies/stream-technology/opencl/Pages/opencl.aspx>

### 3.5 Ошибки в многопоточных приложениях

Помимо привычных для программиста ошибок, встречающихся в компьютерных программах, существует ряд ошибок, специфичных для параллельного программирования. Эти ошибки обусловлены следующими особенностями параллельных программ:

- **Синхронизация потоков.** Программист должен обеспечить корректную последовательность выполняемых разными потоками операций. В общем случае невозможно точно сказать, в какой последовательности будут выполняться команды потоков, т.к. операционная система может в произвольный момент времени приостановить выполнение потока.
- **Взаимодействие потоков.** Также программист не должен допускать конфликтов при обращении к общим для потоков областям памяти.
- **Балансировка нагрузки.** Если в распараллеленной программе один из потоков выполняет 99% работы, то даже на 64-ядерной системе параллельное ускорение едва ли превысит значение 1.01.
- **Масштабируемость.** В идеале параллельная программа должна одинаково хорошо распараллеливать выполняемую работу на любом доступном количестве процессоров. Однако добиться этого нелегко и это часто приводит к трудно обнаруживаемым ошибкам.

Рассмотрим далее подробнее следующий неполный перечень типовых ошибок, возникающих в параллельных программах независимо от используемой технологии распараллеливания:

- Потеря точности операций с плавающей точкой.
- Взаимные блокировки (deadlock).
- Состояния гонки (race conditions).
- Проблема АВА.
- Инверсия приоритетов.
- Голодание (starvation).

- False Sharing.

**Потеря точности.** Если параллельная программа используется для проведения операций с плавающей точкой при работе с вещественными переменными, расположенными в общей для потоков памяти, то при каждом запуске программы может получаться разный результат вещественных расчётов. Это объясняется тем, что при работе нескольких потоков невозможно точно предсказать, в каком порядке операционная система предоставит этим потокам процессор, т.к. в любой момент любой поток может быть временно приостановлен по усмотрению ОС. Это в свою очередь приводит к неопределённой последовательности выполнения операций с плавающей точкой, результат которых, как известно, может зависеть от порядка.

Рассмотрим пример, иллюстрирующий сказанное:

```
1  int i;  
2  float s = 0;  
3  #pragma omp parallel for reduction (+:s) num_threads(8)  
4  for (i = 1; i < 1000000; ++i) {  
5      s += 1.0/i;  
6  }  
7  printf("s=%f\n", s);
```

Здесь в переменную *s* суммируются результаты вещественных вычислений восемью потоками. В результат получается *s*=14.393189. Однако если эту же программу выполняет всего один поток (для этого нужно в строке 3 установить значение параметра *num\_threads* в 1), то результат получится иным: *s*=14.357357. Различие между двумя приведёнными значениями составляет примерно 0.25%.

Получается, что параллельная программа может давать разный результат при запуске на разных платформах. Это следует учитывать, проводя верификацию параллельных программ с использованием однопоточных их нераспараллеленных аналогов.

**Взаимные блокировки.** Одним из часто используемых примитивов синхронизации является мьютекс, позволяющий нескольким потокам согласованно и последовательно выполнять критические области кода, расположенные внутри параллельных секций кода. Критические секции замедляют работу программы, т.к. в каждый момент времени только один поток может находиться внутри критической секции. С помощью мьютексов, например, реализуются функции *omp\_set\_lock* и *omp\_unset\_lock* в OpenMP. При обрамлении этими функциями некоторого участка кода можно сделать из него критическую секцию, вход в которую контролируется условным программным замком (*lock*). В сложных программах может



использовать несколько замков. Это может привести к тому, что два потока, захватывающие несколько замков, застопорят выполнение друг друга без всякой возможности выйти из состояния ожидания друг друга. Такая ситуация называется deadlock (взаимная блокировка).

Простейшим примером взаимной блокировки является работа двух потоков, первый из которых захватывает сначала блокировку 1, потом блокировку 2, а второй сначала захватывает блокировку 2, потом блокировку 1. В результате, возникнет deadlock, если операции будут выполняться в следующем порядке:

- поток1 захватил блокировку 1;
- поток2 захватил блокировку 2;
- поток1 бесконечно ждёт освобождения блокировки 2;
- поток2 бесконечно ждёт освобождения блокировки 1.

Одна из неприятных сторон описанной ситуации заключается в том, что далеко не всегда взаимная блокировка происходит при отладке программы, когда её можно было бы легко выявить и исправить, т.к. вероятность наложение событий нужным образом может быть очень мала. В результате работающая и сданная заказчику программа может в случайные моменты времени "зависать" по якобы непонятным причинам. Рассмотрим пример искусственно реализованной взаимоблокировки, в котором можно рассчитать вероятность её возникновения при многократном запуске.

В приведённой ниже программе в строке 7 создаётся поток, который в бесконечном цикле захватывает замок1, замок2 и инкрементирует переменную s, освобождая после этого оба замка. В строке 13 создаётся поток, который тоже бесконечно инкрементирует s, однако захватывает замки в другом порядке: замок2, замок1. В строке 19 создаётся поток, который следит за состоянием s, опрашивая эту переменную каждые 10 мс. Если последний поток обнаруживает, что переменная s перестала изменяться, он печатает сообщение о возникшей взаимоблокировке и завершает программу.

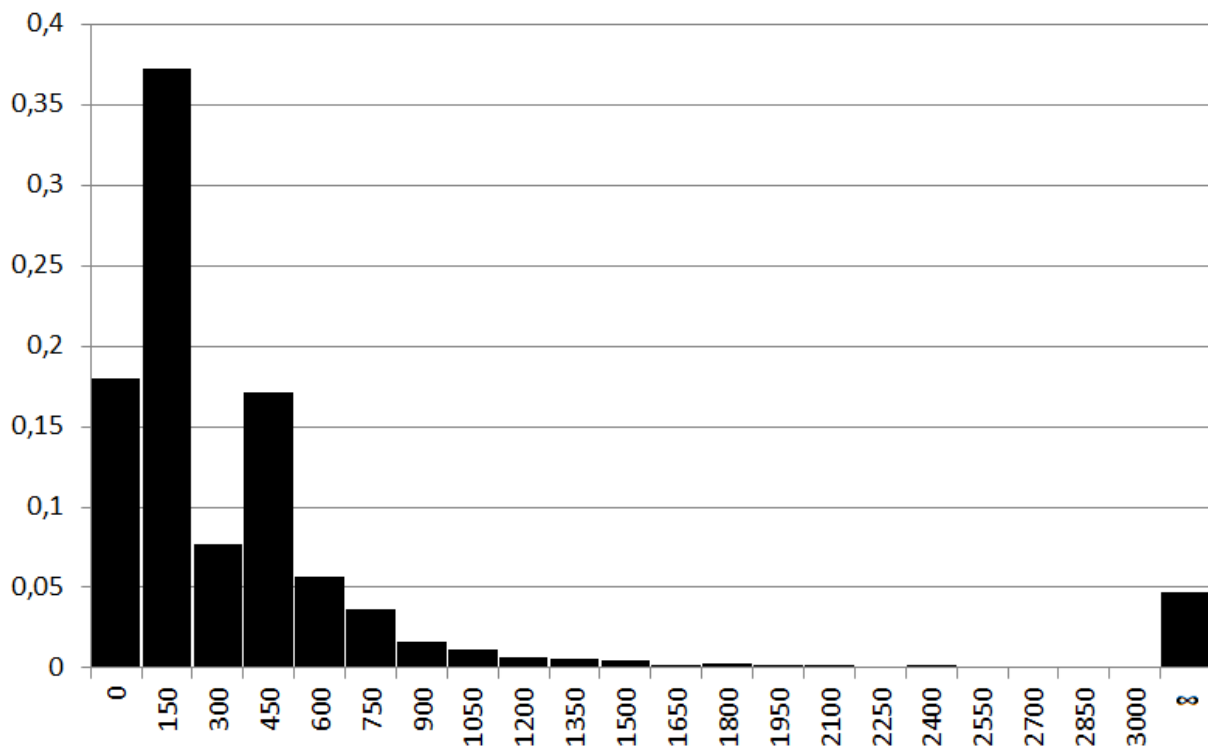
```

1  int old_s, s = 0;
2  omp_lock_t lock1, lock2;
3  omp_init_lock(&lock1);
4  omp_init_lock(&lock2);
5  #pragma omp parallel sections
6  {
7      #pragma omp section
8      for (;;) {
9          omp_set_lock(&lock1);    omp_set_lock(&lock2);
10         s++;
11         omp_unset_lock(&lock2);    omp_unset_lock(&lock1);
12     }
13     #pragma omp section
14     for (;;) {
15         omp_set_lock(&lock2);    omp_set_lock(&lock1);
16         s++;
17         omp_unset_lock(&lock1);    omp_unset_lock(&lock2);
18     }
19     #pragma omp section
20     {
21         for(old_s = !s; old_s != s; old_s = s)
22             usleep(10000);
23         printf("Deadlock with s=%i\n", s);
24         omp_destroy_lock(&lock1);    omp_destroy_lock(&lock2);
25         exit(0);
26     }
27 }

```

Эксперименты с приведённой программой проводились на компьютере с процессором Intel Core i5 (4 логических процессора) с 8 гигабайт ОЗУ в операционной системе Debian Wheezy. Программа была запущена 10000 раз, и было получено 10000 значений переменной *s* на момент возникновения взаимоблокировки. Результаты этих измерения приведены на рисунке 21 в виде гистограммы плотности распределения *s*.

На приведённом рисунке на оси абсцисс подписаны правые границы столбиков. Последний столбик содержит все попадания от 3000 до бесконечности. Среднее значение *s* в описанном случае оказалось равным 2445, т.е. два потока успевают примерно 1222.5 раза захватить и отпустить замки в заведомо неверном опасном порядке без возникновения взаимоблокировки.



*Рис. 21: Гистограмма распределения количества запусков параллельной программ до возникновения взаимоблокировки*

Для исправления описанной ошибки нужно сделать порядок захвата замков одинаковым во всех потоках. Иногда советуют во всей программе установить некоторое общее правило захвата замков, например, можно захватывать замки в алфавитном порядке.

Помимо описанной ситуации с неверным порядком захвата мьютексов, существуют и другие причины взаимоблокировок. Например, повторный захват мьютекса (замка). Неверно написанная программа может попытаться повторно захватить уже захваченный ею замок, предварительно его не освободив. При этом повторная попытка захвата полностью останавливает работу потока. Если логика программы требует повторный захват мьютекса (например, для организации рекурсии), следует использовать специальный подвид замков: рекурсивные мьютексы.

**Состояние гонки (race conditions)** - ошибка в параллельной программе, при которой результат вычисления зависит от порядка в котором выполняется код (при каждом запуске параллельной программы он может быть разным). Например, рассмотрим следующую ситуацию. Один поток изменяет значение глобальной переменной. В это время второй поток выводит это значение на печать. Если второй поток напечатает значение раньше, чем его изменит первый, то программа отработает правильно, однако если код выполнится позже, то выведется новое значение, присво-

енное в первом потоке.

```
1  int a = 0;
2  #pragma omp parallel num_threads(2) shared(a)
3  {
4      if (omp_get_thread_num() == 0) //first thread
5          a = 2; //changes global variable value
6      else { //second thread
7          printf("%d",a); //print var value: 0 or 2?
8      }
9  }
```

С данной проблемой можно столкнуться даже в тех программах, в которых многопоточное программирование не используется явно, но используются какие-то разделяемые ресурсы. Например, если программа копирует текст из поля ввода в буфер обмена и затем тут же вставляется текст в другое поле, то, если она будет запускаться на компьютере одна, то всегда будет работать правильно. Однако если одновременно с ней будет работать программа, также использующая буфер обмена, она может перезаписать значение буфера обмена, даже если команды копирования и вставки будут расположены строго друг за другом. Использование общих ресурсов, даже на очень короткий срок может привести к ошибке.

Такое явление получило название "гейзенбаг" или "плавающая ошибка". Чтобы избежать этой ситуации надо блокировать запись нового значения переменной в первой потоке, пока второй поток не закончит работу. Например, в технологии OpenMP эту проблему решить следующим образом (сохранить старое значение в другой переменной):

```
1  int a = 0;
2  int old_a = a;
3  #pragma omp parallel num_threads(2) shared(a)
4  {
5      if (omp_get_thread_num() == 0) //first thread
6          a = 2; //changes global variable value
7      else { //second thread
8          printf("%d",old_a); //print old variable value
9      }
10 }
```

**Проблема АВА** - проблема, при которой поток два раза читая одинаковое значение "думает", что данные не изменились. Например, первый поток присвоил переменной значение А. Второй поток присвоил ей значение В, а потом снова А. Когда первый поток снова читает эту переменную, она равна А, и он "думает", что ничего не изменилось. Более практичный пример из программирования: в переменной хранится адрес, указывающий на начало массива. Второй поток освобождает память для нового

массива функцией `free` и создает его функцией `malloc`, которая выделила память в том же месте, так как эта область памяти уже свободна. Когда первый поток сравнивает значения указателя на массив до и после, он видит, что они равны и решает, что массив не изменился, хотя на его месте уже хранятся новые данные. Чтобы решить эту проблему можно хранить признак того, что массив был изменен.

**Инверсия приоритетов.** Представим ситуацию, в которой существует 3 потока с приоритетами: высокий, средний и низкий соответственно, причем потоки с высоким и низким приоритетом захватывают общий мьютекс. Пусть поток с низким приоритетом захватил мьютекс и начал свое выполнение, но его прервал поток со средним приоритетом. Теперь, если поток с высоким приоритетом перехватит мьютекс начнет выполняться, он будет ждать освобождения мьютекса, но поток с низким приоритетом не может его освободить, так как его вытеснил поток со средним приоритетом. Эта проблема решается заданием всем потокам одного приоритета на время удержания мьютекса.

**Голодание(*starvation*)** возникает, когда поток с низким приоритетом долго находится в состоянии готовности и простаивает. Такое голодание вызвано нехваткой процессорного времени, существует также голодание, вызванное невозможностью работы с данными (запрет на чтение и/или запись). В современных ОС эта проблема решается следующим образом: даже если у потока очень низкий приоритет, он все равно вызывается на исполнение через определенное количество времени. В своих программах следует разумно разделять задачи между тreads, чтобы поток, выполняющий более важную и долгую задачу имел более высокий приоритет.

**False sharing** - ситуация, возникающая с системами, поддерживающими когерентность памяти(кешей), при которой производятся лишние (ненужные в этом месте программы) операции для передачи данных между потоками. *Когерентность памяти(кешей)* - свойство памяти, при котором при изменении значения ячейки памяти одним процессом, эти изменения становятся видны в остальных процессах. На организацию такой памяти тратятся большие ресурсы, так как при каждом изменении значения одним потоком, нужно извещать остальные. Рассмотрим следующий пример:

```

1  struct str {
2      char a;
3      char b;
4  };
5  int n = 10000;
6  struct str array[n];
7
8  void fprint_a()
9  {
10     for (int i = 0; i < n; ++i)
11         str[i].a = 'a';
12 }
13
14 void fprint_b()
15 {
16     for (int i = 0; i < n; ++i)
17         str[i].b = 'b';
18 }

```

Если запустить функции `fprint_a` и `fprint_b` в двух разных потоках, то из-за постоянной синхронизации памяти между потоками, программа будет работать медленно, так как `a` и `b` находятся в одной строке кеша (обычно 64 байта). Более разумно будет распараллелить каждый цикл между потоками (например, с помощью директивы препроцессора `#pragma omp parallel for` в OpenMP).

## 4 Лабораторная работа №1. «Автоматическое распараллеливание программ»

### 4.1 Порядок выполнения работы

1. На компьютере с многоядерным процессором установить ОС Linux и компилятор GCC версии не ниже 4.7.2. При невозможности установить Linux или отсутствии компьютера с многоядерным процессором можно выполнять лабораторную работу на виртуальной машине.
2. На языке Си написать консольную программу lab1.c, решающую задачу, указанную в п.IV (см. ниже). В программе нельзя использовать библиотечные функции сортировки, выполнения матричных операций и расчёта статистических величин. В программе нельзя использовать библиотечные функции, отсутствующие в стандартных заголовочных файлах stdio.h, stdlib.h, math.h, sys/time.h. Задача должна решаться 50 раз с разными начальными значениями генератора случайных чисел (ГСЧ). Структура программы примерно следующая:

```
1  #include <stdio.h>
2  #include <stdlib.h>
3  #include <sys/time.h>
4  int main(int argc, char* argv[])
5  {
6      int i, N;
7      struct timeval T1, T2;
8      long delta_ms;
9      N = atoi(argv[1]); /* N равен первому параметру командной строки */
10     gettimeofday(&T1, NULL); /* запомнить текущее время T1 */
11     for (i=0; i<50; i++) /* 50 экспериментов */
12     {
13         srand(i); /* инициализировать начальное значение ГСЧ */
14         /* Заполнить массив исходных данных размером N */
15         /* Решить поставленную задачу, заполнить массив с результатами
16         */
17         /* Отсортировать массив с результатами указанным методом */
18         gettimeofday(&T2, NULL); /* запомнить текущее время T2 */
19         delta_ms = 1000*(T2.tv_sec - T1.tv_sec) + (T2.tv_usec - T1.tv_usec)
20         /1000;
21         printf("\nN=%d. Milliseconds passed: %ld\n", N, delta_ms); /* T2 -
22         T1 */
23         return 0;
24     }
```

3. Скомпилировать написанную программу без использования авто-

матического распараллеливания с помощью следующей команды:  
`/home/user/gcc -O3 -Wall -Werror -o lab1-seq lab1.c`

4. Скомпилировать написанную программу, используя встроенное в gcc средство автоматического распараллеливания Graphite с помощью следующей команды `"/home/user/gcc -O3 -Wall -Werror -floop-parallelize-all -ftree-parallelize-loops=K lab1.c -o lab1-par-K"` (переменной K поочерёдно присвоить хотя бы 4 различных целых значений, выбор обосновать).
5. В результате получится одна нераспараллеленная программа и четыре или более распараллеленных.
6. Закрыть все работающие в операционной системе прикладные программы (включая Winamp, uTorrent, браузеры и Skype), чтобы они не влияли на результаты последующих экспериментов.
7. Запускать файл `lab1-seq` из командной строки, увеличивая значения N до значения N1, при котором время выполнения превысит 0.01 с. Подобным образом найти значение  $N=N2$ , при котором время выполнения превысит 2 с.
8. Используя найденные значения N1 и N2, выполнить следующие эксперименты (для автоматизации проведения экспериментов рекомендуется написать скрипт):
  - запускать `lab1-seq` для значений  $N = N1, N1 + \Delta, N1 + 2\Delta, N1 + 3\Delta, \dots, N2$  и записывать получающиеся значения времени `delta_ms(N)` в функцию `seq(N)`;
  - запускать `lab1-par-K` для значений  $N = N1, N1 + \Delta, N1 + 2\Delta, N1 + 3\Delta, \dots, N2$  и записывать получающиеся значения времени `delta_ms(N)` в функцию `par - K(N)`;
  - значение  $\Delta$  выбрать так:  $\Delta = (N2 - N1)/10$ .
9. Написать отчёт о проделанной работе.
10. Подготовиться к устным вопросам на защите.
11. Найти вычислительную сложность алгоритма до и после распараллеливания, сравнить полученные результаты.



12. **Необязательное задание №1 (для получения оценки «четыре» и «пять»).** Провести аналогичные описанным эксперименты, используя вместо gcc компилятор Solaris Studio (или любой другой на своё усмотрение). При компиляции следует использовать следующие опции для автоматического распараллеливания: `«solarisstudio -cc -O3 -xautopar -xloopinfo lab1.c»`.
13. **Необязательное задание №2 (для получения оценки «пять»).** Это задание выполняется только после выполнения предыдущего пункта. Провести аналогичные описанным эксперименты, используя вместо gcc компилятор Intel ICC (или любой другой на своё усмотрение). В ICC следует при компиляции использовать следующие опции для автоматического распараллеливания: `«icc -parallel -par-report -par-threshold K -o lab1-icc-par-K lab1.c»`.
- Если ключ `«-par-report»` не работает в вашей версии компилятора, то желательно использовать более актуальный ключ `«-qopt-report-phase=par»`.

## 4.2 Состав отчета

1. Титульный лист с названием вуза, ФИО студента и названием работы.
2. Содержание отчета (с указанием номера страниц и т.п.).
3. Описание решаемой задачи (взять из п.I и п.IV).
4. Краткая характеристика использованного для проведения экспериментов процессора, операционной системы и компилятора GCC (официальное название, номер версии/модели, разрядность, количество ядер, ёмкость ОЗУ и т.п.).
5. Полный текст программы lab1.c в виде отдельного файла.
6. Таблицы значений и графики функций  $seq(N)$ ,  $par-K(N)$  с указанием величины параллельного ускорения.
7. Подробные выводы с анализом приведённых графиков и полученных результатов.
8. Отчёт предоставляется на бумажном носителе или на флешке.

### 4.3 Подготовка к защите

1. Уметь объяснить каждую строку программы, представленной в отчёте.
2. Знать о назначении и основных особенностях GCC, а также о назначении всех использованных в работе ключей компиляции GCC.
3. Знать материал лекции №1.
4. Взять с собой все нужные файлы для демонстрации работы программы.

### 4.4 Варианты заданий

Вариант задания выбирается в соответствии с приведёнными ниже описанием этапов, учитывая, что число  $A = \Phi * И * О$ , где  $\Phi$ ,  $И$ ,  $О$  означают количество букв в фамилии, имени и отчестве студента. Номер варианта в соответствующих таблицах выбирается по формуле  $X = 1 + ((A \bmod 47) \bmod B)$ , где  $B$  – количество элементов в соответствующей таблице, а операция  $\bmod$  означает остаток от деления. Например, при  $A = 476$  и  $B = 5$ , получим  $X = 1 + ((470+6) \bmod 47) \bmod 5 = 1 + (6 \bmod 5) = 2$ . Порядок вычислений должен быть следующим:

1. **Этап Generate.** Сформировать массив  $M1$  размерностью  $N$ , заполнив его с помощью функции `rand_r` (нельзя использовать `rand`) случайными вещественными числами, имеющими равномерный закон распределения в диапазоне от 1 до  $A$  (включительно). Аналогично сформировать массив  $M2$  размерностью  $N/2$  со случайными вещественными числами в диапазоне от  $A$  до  $10*A$ .
2. **Этап Map.** В массиве  $M1$  к каждому элементу применить операцию из таблицы:

Номер варианта	Операция
1	Гиперболический синус с последующим возведением в квадрат
2	Гиперболический косинус с последующим увеличением на 1
3	Гиперболический тангенс с последующим уменьшением на 1
4	Гиперболический котангенс корня числа
5	Деление на Пи с последующим возведением в третью степень
6	Кубический корень после деления на число e
7	Экспонента квадратного корня (т.е. $M1[i] = \exp(\sqrt{M1[i]})$ )

Затем в массиве M2 каждый элемент поочерёдно сложить с предыдущим (для этого вам понадобится копия массива M2, из которого нужно будет брать операнды), а к результату сложения применить операцию из таблицы (считать, что для начального элемента массива предыдущий элемент равен нулю):

Номер варианта	Операция
1	Модуль синуса (т.е. $M2[i] =  \sin(M2[i] + M2[i-1]) $ )
2	Модуль косинуса
3	Модуль тангенса
4	Модуль котангенса
5	Натуральный логарифм модуля тангенса
6	Десятичный логарифм, возведенный в степень e
7	Кубический корень после умножения на число Пи
8	Квадратный корень после умножения на e

3. **Этап Merge.** В массивах M1 и M2 ко всем элементам с одинаковыми индексами попарно применить операцию из таблицы (результат записать в M2):

Номер варианта	Операция
1	Возведение в степень (т.е. $M2[i] = M1[i]^{M2[i]}$ )
2	Деление (т.е. $M2[i] = M1[i]/M2[i]$ )
3	Умножение
4	Выбор большего (т.е. $M2[i] = \max(M1[i], M2[i])$ )
5	Выбор меньшего
6	Модуль разности

4. **Этап Sort.** Полученный массив необходимо отсортировать методом, указанным в таблице (для этого нельзя использовать библиотечные функции; можно взять реализацию в виде свободно доступного исходного кода):

Номер варианта	Операция
1	Сортировка выбором (Selection sort).
2	Сортировка расчёской (Comb sort).
3	Пирамидальная сортировка (сортировка кучи, Heapsort).
4	Глупая сортировка (Stupid sort).
5	Гномья сортировка (Gnome sort).
6	Сортировка вставками (Insertion sort).
7	Сортировка выбором (Selection sort).

5. **Этап Reduce.** Рассчитать сумму синусов тех элементов массива  $M2$ , которые при делении на минимальный ненулевой элемент массива  $M2$  дают чётное число (при определении чётности учитывать только целую часть числа). Результатом работы программы по окончании пятого этапа должно стать одно число  $X$ , которое следует использовать для верификации программы после внесения в неё изменений (например, до и после распараллеливания итоговое число  $X$  не должно измениться в пределах погрешности). Значение числа  $X$  следует привести в отчёте для различных значений  $N$ .

## 5 Лабораторная работа №2. «Исследование эффективности параллельных библиотек для С-программ»

### 5.1 Порядок выполнения работы

1. В исходном коде программы, полученной в результате выполнения лабораторной работы №1, нужно на этапах Map и Merge все циклы с вызовами математических функций заменить их векторными аналогами из библиотеки «AMD Framewave» (<http://framewave.sourceforge.net>). При выборе конкретной Framewave-функции необходимо убедиться, что она помечена как MT (Multi-Threaded), т.е. распараллеленная. Полный перечень доступных функций находится по ссылке: [http://framewave.sourceforge.net/Manual/fw\\_section\\_060.html#fw\\_section\\_060](http://framewave.sourceforge.net/Manual/fw_section_060.html#fw_section_060). Например, Framewave-функция min в списке поддерживаемых технологий имеет только SSE2, но не MT.

*Примечание:* выбор библиотеки Framewave не является обязательным, можно использовать любую другую параллельную библиотеку, если в ней нужные функции распараллелены, так, например, можно использовать ATLAS (для этой библиотеки необходимо выключить троттлинг и энергосбережение, а также разобраться с механизмом изменения числа потоков) или Intel Integrated Performance Primitives.

2. Добавить в начало программы вызов Framewave-функции SetNumThreads(M) для установки количества создаваемых параллельной библиотекой потоков, задействуемых при выполнении распараллеленных Framewave-функций. Нужно число M следует устанавливать из параметра командной строки (argv) для удобства автоматизации экспериментов.
3. Скомпилировать программу, не применяя опции автоматического распараллеливания, использованные в лабораторной работе №1. Провести эксперименты с полученной программой для тех же значений  $N_1$  и  $N_2$ , которые использовались в лабораторной работе №1, при  $M = 1, 2, \dots, K$ , где  $K$  – количество процессоров (ядер) на экспериментальном стенде.
4. Сравнить полученные результаты с результатами лабораторной работы №1: на графиках показать, как изменилось время выполне-

ния программы, параллельное ускорение и параллельная эффективность.

5. Написать отчёт о проделанной работе.
6. Подготовиться к устным вопросам на защите.
7. **Необязательное задание №1** (для получения оценки «четыре» и «пять»). Исследовать параллельное ускорение для различных значений  $M > K$ , т.е. оценить накладные расходы при создании чрезмерного большого количества потоков. Для иллюстрации того, что программа действительно распараллелилась, привести график загрузки процессора (ядер) во время выполнения программы при  $N = N_2$  для всех использованных  $M$ . Для получения графика можно как написать скрипт, так и просто сделать скриншот диспетчера задач, указав на скриншоте моменты начала и окончания эксперимента (в отчёте нужно привести текст скрипта или название использованного диспетчера).
8. **Необязательное задание №2** (для получения оценки «пять»). Это задание выполняется только после выполнения предыдущего пункта. Используя закон Амдала, рассчитать коэффициент распараллеливания для всех экспериментов и привести его на графиках. Прокомментировать полученные результаты.

## 5.2 Состав отчета

1. Титульный лист с названием вуза, ФИО студентов и названием работы.
2. Содержание отчета (с указанием номера страниц и т.п.).
3. Краткая характеристика использованного для проведения экспериментов процессора, операционной системы и компилятора (официальное название, номер версии/модели, разрядность, количество ядер, ёмкость ОЗУ и т.п.).
4. Описание особенностей конфигурации использованной параллельной библиотеки, включая описание последовательности шагов, принятых для установки библиотеки, и использованных опций компиляции.

5. Полный текст полученной параллельной программы, а также текст всех скриптов, использованных для компилирования программы и проведения экспериментов.
6. Графики функций времени выполнения использованных программ, а также графики параллельного ускорения и параллельной эффективности для разных  $N$  и  $M$  (допускается совмещать несколько графиков в одной системе координат).
7. Подробные выводы с анализом приведённых графиков и полученных результатов. Отчёт предоставляется на бумажном носителе или на флешке.

### **5.3 Подготовка к защите**

1. Уметь объяснить каждую строку программы, представленной в отчёте.
2. Знать о назначении всех использованных в работе ключей компиляции.
3. Знать материал лекции №2.
4. Взять с собой все нужные файлы для демонстрации работы программы.

## 6 Лабораторная работа №3. «Распараллеливание циклов с помощью технологии OpenMP»

### 6.1 Порядок выполнения работы

1. Добавить во все for-циклы в программе из ЛР №1 следующую директиву OpenMP:  
”#pragma omp parallel for default(none) private(...) shared(...)”. Наличие всех перечисленных параметров в указанной директиве является обязательным.
2. Проверить все for-циклы на внутренние зависимости по данным между итерациями. Если зависимости обнаружались, использовать для защиты критических секций директиву ”#pragma omp critical” или ”#pragma omp atomic” (если операция атомарна), или параметр reduction (предпочтительнее) или вообще отказаться от распараллеливания цикла (свой выбор необходимо обосновать).
3. Убедиться, что получившаяся программа обладает свойством прямой совместимости с компиляторами, не поддерживающими OpenMP (для проверки этого можно скомпилировать программу без опции ”-fopenmp”, в результате не должно быть сообщений об ошибках, а программа должна корректно работать).
4. Провести эксперименты, измеряя параллельное ускорение. Привести сравнение графиков параллельного ускорения с ЛР №1 и ЛР №2.
5. Провести эксперименты, добавив параметр ”schedule” и варьируя в экспериментах тип расписания. Исследование нужно провести для всех возможных расписаний: static, dynamic, guided. Способ варьирования параметра chunk\_size выбрать самостоятельно (но должно быть не менее 5 точек варьирования). Привести сравнение параллельного ускорения при различных расписаниях с результатами п.4.
6. Выбрать из рассмотренных в п.4 и п.5 наилучший вариант при различных  $N$ . Сформулировать условия, при которых наилучшие результаты получились бы при использовании других типов расписания.



7. Найти вычислительную сложность алгоритма до и после распараллеливания, сравнить полученные результаты.
8. Написать отчёт о проделанной работе.
9. Подготовиться к устным вопросам на защите.
10. **Необязательное задание №1** (для получения оценки «четыре» и «пять»). Для иллюстрации того, что программа действительно распараллелилась, привести график загрузки процессора (ядер) от времени при выполнении программы при  $N = N_1$  для лучшего варианта распараллеливания. Для получения графика можно как написать скрипт так и просто сделать скриншот диспетчера задач, указав на скриншоте моменты начала и окончания эксперимента (в отчёте нужно привести текст скрипта или название использованного диспетчера). Недостаточно привести однократное моментальное измерение загрузки утилитой htop, т.к. требуется привести график изменения загрузки за всё время выполнения программы.
11. **Необязательное задание №2** (для получения оценки «пять»). Построить график параллельного ускорения для точек  $N < N_1$  и найти значения  $N$ , при которых накладные расходы на распараллеливание превышают выигрыш от распараллеливания (независимо для различных типов расписания).

## 6.2 Состав отчета

1. Титульный лист с названием вуза, ФИО студентов и названием работы.
2. Содержание отчета (с указанием номера страниц и т.п.).
3. Краткое описание решаемой задачи.
4. Характеристика использованного для проведения экспериментов процессора, операционной системы и компилятора GCC (точное название, номер версии/модели, разрядность, количество ядер и т.п.).
5. Полный текст программы с использованием параметра schedule.
6. Подробные выводы с анализом каждого из приведённых графиков.

7. Отчёт предоставляется в бумажном или электронном виде вместе с полным текстом программы. По требованию преподавателя нужно быть готовыми скомпилировать и запустить этот файл на компьютере в учебной аудитории (или своём ноутбуке).

### **6.3 Подготовка к защите**

1. Уметь объяснить каждую строку программы, представленной в отчёте.
2. Уметь объяснить выводы, полученные в результате работы.
3. Знать назначение каждой директивы OpenMP, использованной в программе.
4. Повторить материал лекции №3, прочитать главу про OpenMP в методическом пособии
5. Прочитать материал о директиве `#pragma_omp_for` в книге Антонова «Параллельное программирование с использованием технологии OpenMP: Учебное пособие», знать ответы на вопросы в конце соответствующей главы.

## 7 Лабораторная работа №4. «Метод доверительных интервалов при измерении времени выполнения параллельной OpenMP-программы»

### 7.1 Порядок выполнения работы

1. В программе, полученной в результате выполнения ЛР-3, так изменить этап Generate, чтобы генерируемый набор случайных чисел не зависел от количества потоков, выполняющих программу. Например, на каждой итерации  $i$  перед вызовом `rand_r` можно вызывать функцию `srand(f(i))`, где  $f$  – произвольно выбранная функция. Можно придумать и использовать любой другой способ.
2. Заменить вызовы функции `gettimeofday` на `omp_get_wtime`.
3. Распараллелить вычисления на этапе Sort, для чего выполнить сортировку в два этапа:
  - Отсортировать первую и вторую половину массива в двух независимых нитях (можно использовать OpenMP-директиву `"parallel sections"`);
  - Объединить отсортированные половины в единый массив.
4. Написать функцию, которая один раз в секунду выводит в консоль сообщение о текущем проценте завершения работы программы. Указанную функцию необходимо запустить в отдельном потоке, параллельно работающем с основным вычислительным циклом.
5. Обеспечить прямую совместимость (forward compatibility) написанной параллельной программы. Для этого все вызываемые функции вида `«omp_»` можно условно переопределить в препроцессорных директивах, например, так:

```
1 #ifdef _OPENMP
2     #include "omp.h"
3 #else
4     int omp_get_num_procs() { return 1; }
5 #endif
```

6. Провести эксперименты, варьируя  $N$  от  $\min(\frac{N_x}{2}, N_1)$  до  $N_2$ , где значения  $N_1$  и  $N_2$  взять из ЛР-1, а  $N_x$  – это такое значение  $N$ ,

при котором накладные расходы на распараллеливание превышают выигрыш от распараллеливания. Написать отчёт о проделанной работе. Подготовиться к устным вопросам на защите.

7. **Необязательное задание на «четвёрку» и «пятёрку».** Уменьшить количество итераций основного цикла с 100 до 10 и провести эксперименты, измеряя время выполнения следующими методами:

- Использование минимального из десяти полученных замеров;
- Расчёт по десяти измерениям доверительного интервала с уровнем доверия 95%.

Привести графики параллельного ускорения для обоих методов в одной системе координат, при этом нижнюю и верхнюю границу доверительного интервала следует привести двумя независимыми графиками.

8. **Необязательное задание на «пятёрку»:** в п.3 задания на этапе Sort выполнить параллельную сортировку не двух частей массива, а  $k$  частей в  $k$  нитях (тредах), где  $k$  – это количество процессоров (ядер) в системе, которое становится известным только на этапе выполнения программы с помощью команды «`k = omp_get_num_procs()`».

## 7.2 Состав отчета

1. Титульный лист с названием вуза, ФИО студентов и названием работы.
2. Содержание отчета (с указанием номера страниц и т.п.).
3. Краткое описание решаемой задачи.
4. Характеристика использованного для проведения экспериментов процессора, операционной системы и компилятора GCC (точное название, номер версии/модели, разрядность, количество ядер и т.п.).
5. Полный текст программы и использованных в процессе работы скриптов и инструментов с указанием параметров запуска.
6. Подробные выводы с анализом каждого из приведённых графиков.

Отчёт предоставляется в бумажном или электронном виде вместе с полным текстом программы. По требованию преподавателя нужно быть готовыми скомпилировать и запустить этот файл на компьютере в учебной аудитории (или своём ноутбуке).

### **7.3 Подготовка к защите**

1. Уметь объяснить каждую строку программы, представленной в отчёте.
2. Уметь объяснить выводы, полученные в результате работы.
3. Знать назначение каждой директивы OpenMP, использованной в программе.
4. Повторить материал лекции №4, прочитать главу про OpenMP в методическом пособии.
5. Знать ответы на вопросы из разделов «Задание» книги Антонова «Параллельное программирование с использованием технологии OpenMP: Учебное пособие» (см. страницы 12, 28, 35, 54).

## 8 Лабораторная работа №5. «Параллельное программирование с использованием стандарта POSIX Threads»

### 8.1 Порядок выполнения работы

1. Взять в качестве исходной OpenMP-программу из ЛР-5, в которой распараллелены все этапы вычисления. Убедиться, что в этой программе корректно реализован одновременный доступ к общей переменной, используемой для вывода в консоль процента завершения программы.
2. Изменить исходную программу так, чтобы вместо OpenMP-директив применялся стандарт «POSIX Threads»:
  - для получения оценки «3» достаточно изменить только один этап (Generate, Map, Merge, Sort), который является узким местом (bottle neck), а также функцию вывода в консоль процента завершения программы;
  - для получения оценки «4» и «5» необходимо изменить всю программу, но допускается в качестве расписания циклов использовать «schedule static»;
  - для получения оценки «5» необходимо хотя бы один цикл распараллелить, реализовав вручную расписание «schedule dynamic» или «schedule guided».
3. Провести эксперименты и по результатам выполнить сравнение работы двух параллельных программ («OpenMP» и «POSIX Threads»), которое должно описывать следующие аспекты работы обеих программ (для различных  $N$ ):
  - полное время решения задачи;
  - параллельное ускорение;
  - доля времени, проводимого на каждом этапе вычисления («нормированная диаграмма с областями и накоплением»);
  - количество строк кода, добавленных при распараллеливании, а также грубая оценка времени, потраченного на распараллеливание (накладные расходы программиста);
  - остальные аспекты, которые вы выяснили самостоятельно (**Обязательный пункт**);

## **8.2 Состав отчета**

1. Титульный лист с названием вуза, ФИО студентов и названием работы. Содержание отчета (с указанием номера страниц и т.п.).
2. Содержание отчета (с указанием номера страниц и т.п.).
3. Краткое описание решаемой задачи.
4. Характеристика использованного для проведения экспериментов процессора, операционной системы и компилятора GCC (точное название, номер версии/модели, разрядность, количество ядер и т.п.).
5. Полный текст программ («OpenMP» и «POSIX Threads») и использованных в процессе работы скриптов, и инструментов с указанием параметров запуска.
6. Подробные выводы с анализом каждого из приведённых графиков.

Отчёт предоставляется в бумажном или электронном виде вместе с полным текстом программы. По требованию преподавателя нужно быть готовыми скомпилировать и запустить этот файл на компьютере в учебной аудитории (или своём ноутбуке).

## **8.3 Подготовка к защите**

1. Уметь объяснить каждую строку программы, представленной в отчёте.
2. Уметь объяснить выводы, полученные в результате работы.
3. Подготовиться к ответам на вопросы по материалам лекции №5.

## **9 Лабораторная работа №6. «Изучение технологии OpenCL»**

### **9.1 Порядок выполнения работы**

1. Вам необходимо реализовать один (для оценки 3) или два (для оценки 4) этапа вашей программы из предыдущих лабораторных работ. При этом вычисления можно проводить как на CPU, так и на GPU (на своё усмотрение, но GPU предпочтительнее).
2. **Дополнительное задание (оценка 5).**
  - Выполнение заданий для оценки 3 и 4.
  - Расчёт доверительного интервала.
  - Посчитать время 2 способами: с помощью profiling и с помощью обычного замера (как в предыдущих заданиях).
  - Оценить накладные расходы.
  - \* Необязательное задание для магистрантов с большим количеством свободного времени: Проводить вычисления совместно на GPU и CPU (т.е. итерации в некоторой обоснованной пропорции делятся между GPU и CPU, и параллельно на них выполняются).
3. При желании данную лабораторную работу можно написать на CUDA.

### **9.2 Состав отчета**

1. Титульный лист с названием вуза, ФИО студентов и названием работы. Содержание отчета (с указанием номера страниц и т.п.).
2. Краткое описание решаемой задачи.
3. Характеристика использованного для проведения экспериментов процессора, операционной системы и компилятора GCC (точное название, номер версии/модели, разрядность, количество ядер и т.п.).
4. Полный текст распараллеленной программы (для п.2 и п.3).
5. Подробные выводы.



Отчёт предоставляется в бумажном или электронном виде вместе с полным текстом программы. По требованию преподавателя нужно быть готовыми скомпилировать и запустить этот файл на компьютере в учебной аудитории (или своём ноутбуке).

### **9.3 Подготовка к защите**

1. Уметь объяснить каждую строку программы, представленной в отчёте. Уметь объяснить выводы, полученные в результате работы.
2. Прочитать раздел методички 3.4 *Технология OpenCL*.
3. Изучить материал лекции по технологии OpenCL.

## Список используемой литературы

1. Соснин В.В., Балакшин П.В. Введение в параллельные вычисления. – СПб: Университет ИТМО, 2016. – 51 с.
2. Top 500 supercomputers list. URL: <https://www.top500.org/> (дата обращения: 12.02.19).
3. Википедия. Симметричная мультипроцессорность. URL: [https://ru.wikipedia.org/wiki/Симметричная\\_многопроцессорность](https://ru.wikipedia.org/wiki/Симметричная_многопроцессорность) (дата обращения: 12.02.19).
4. Википедия. Массово-параллельная архитектура. URL: [https://ru.wikipedia.org/wiki/Массово-параллельная\\_архитектура](https://ru.wikipedia.org/wiki/Массово-параллельная_архитектура) (дата обращения: 12.02.19).
5. Википедия. OpenCL. URL: <https://ru.wikipedia.org/wiki/OpenCL> (дата обращения: 13.02.19).
6. Введение в GPU-вычисления - CUDA/OpenCL. URL: <http://my-it-notes.com/2013/06/gpu-processing-intro-cuda-opengl/> (дата обращения: 13.02.19).
7. Бастраков С.И. Программирование на OpenCL. - Нижний новгород: ННГУ, 2011.
8. OpenCL – официальный сайт URL: <http://www.khronos.org/opengl/> (дата обращения: 13.02.19).
9. Антонов А.С. Параллельное программирование с использованием технологии MPI. - Москва: МГУ, 2004. - 72 с.
10. Антонов А.С. Параллельное программирование с использованием технологии OpenMP. - Москва: МГУ, 2009.- 78 с.
11. Википедия. Параллельные вычисления. URL: [https://ru.wikipedia.org/wiki/Параллельные\\_вычисления](https://ru.wikipedia.org/wiki/Параллельные_вычисления) (дата обращения: 14.02.19).
12. Википедия. Распределенные вычисления. URL: [https://ru.wikipedia.org/wiki/Распределенные\\_вычисления](https://ru.wikipedia.org/wiki/Распределенные_вычисления) (дата обращения: 14.02.19).

13. Стандарт языка C++. ISO/IEC 14882:2011. URL: <https://www.iso.org/standard/50372.html> (дата обращения: 14.02.19).
14. Википедия. Проблема АВА. URL: [https://ru.wikipedia.org/wiki/Проблема\\_АВА](https://ru.wikipedia.org/wiki/Проблема_АВА) (дата обращения: 14.02.19).
15. Транзакционная память: история и развитие. URL: <https://habr.com/ru/post/221667/> (дата обращения: 14.02.19).
16. Википедия. Модель акторов. URL: [https://ru.wikipedia.org/wiki/Модель\\_акторов](https://ru.wikipedia.org/wiki/Модель_акторов) (дата обращения: 14.02.19).
17. Кормен, Томас Х. и др. Алгоритмы: построение и анализ, 3-е изд. : Пер. с англ. - М. : ООО "И. Д. Вильямс", 2013. - 1328 с. : ил. - Парал. тит. англ.
18. Lock-free структуры данных. Очередной трактат. URL: <https://habr.com/ru/post/219201/> (дата обращения: 14.09.19)
19. False sharing в многопоточном приложении на Java. URL: <https://habr.com/ru/post/187752/> (дата обращения: 14.09.19)
20. Load-link/store-conditional. URL: <https://en.wikipedia.org/wiki/Load-link/store-conditional> (дата обращения: 14.09.19)
21. Concurrent Data Structures (libcdfs). URL: <http://libcdfs.sourceforge.net/> (дата обращения: 14.09.19)
22. Очередь Майкла и Скотта. URL: [https://neerc.ifmo.ru/wiki/index.php?title=Очередь\\_Майкла\\_и\\_Скотта](https://neerc.ifmo.ru/wiki/index.php?title=Очередь_Майкла_и_Скотта) (дата обращения: 14.09.19)
23. Баер Ж.-Л., Барлоу Р., Вудворд М., и др. Системы параллельной обработки: Пер. с англ. / Под. ред. Д. Ивенса. - М.: Мир, 1985. - 416 с.: ил.
24. Валях Е. Последовательно-параллельные вычисления. Пер. с англ. - М.: Мир, 1985. - 456 с.: ил.
25. Воеводин В.В., Воеводин Вл.В. Параллельные вычисления. - СПб.: БХВ-Петербург, 2002. - 608 с.: ил.