2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

DOI: 10.26089/NumMet.v22r210 УДК: 004.2

Высокопроизводительные вычислительные платформы: текущий статус и тенденции развития

А. С. Антонов
Московский государственный университет им. М. В. Ломоносова, Научно-исследовательский вычислительный центр, Москва, Российская Федерация
ORCID: http://orcid.org/0000-0003-2820-7196, e-mail: asa@parallel.ru

И. В. Афанасьев
Московский государственный университет им. М. В. Ломоносова, Научно-исследовательский вычислительный центр, Москва, Российская Федерация
ORCID: http://orcid.org/0000-0002-0202-1548, e-mail: afanasiev_ilya@icloud.com

Вл. В. Воеводин
Московский государственный университет им. М. В. Ломоносова, Научно-исследовательский вычислительный центр, Москва, Российская Федерация
ORCID: http://orcid.org/0000-0001-6036-5106, e-mail: voevodin@parallel.ru

Аннотация: В данной статье представлен обзор современного состояния суперкомпьютерной техники. Обзор сделан с разных точек зрения — начиная от особенностей построения современных вычислительных устройств до особенностей архитектуры больших суперкомпьютерных комплексов. В данный обзор вошли описания самых мощных суперкомпьютеров мира и России по состоянию на начало 2021 г., а также некоторых менее мощных систем, интересных с других точек зрения. Также делается акцент на тенденциях развития суперкомпьютерной отрасли и описываются наиболее известные проекты построения будущих экзафлопсных суперкомпьютеров.

Ключевые слова: суперкомпьютер, производительность, эффективность, процессор, Top500, параллелизм, ускоритель, коммуникационная сеть.

Благодарности: Раздел 3 работы выполнен при поддержке РФФИ, грант 18–29–03230 мк. Разделы 2 и 4 выполнены в рамках госбюджетной темы “Разработка программных средств поддержки жизненного цикла и обеспечения эффективности суперкомпьютерных приложений, систем и центров” регистрационный номер АААА–А21–121011690003–6).

Для цитирования: Антонов А.С., Афанасьев И.В., Воеводин Вл.В. Высокопроизводительные вычислительные платформы: текущий статус и тенденции развития // Вычислительные методы и программирование. 2021. **22**, 138–181. doi 10.26089/NumMet.v22r210

High-performance computing platforms: current status and development trends

A. S. Antonov
Lomonosov Moscow State University, Research Computing Center, Moscow, Russia
ORCID: http://orcid.org/0000-0003-2820-7196, e-mail: asa@parallel.ru

I. V. Afanasiev
Lomonosov Moscow State University, Research Computing Center, Moscow, Russia
ORCID: http://orcid.org/0000-0002-0202-1548, e-mail: afanasiev_ilya@icloud.com

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

Vl. V. Voevodin
Lomonosov Moscow State University, Research Computing Center, Moscow, Russia
ORCID: http://orcid.org/0000-0001-6036-5106, e-mail: voevodin@parallel.ru

Abstract: This paper provides an overview of the current state of supercomputer technology. The review is done from different points of view — from the construction features of modern computing devices to the features of the architecture of large supercomputer complexes. This review includes descriptions of the most powerful supercomputers in the world and Russia since the early of 2021 as well as some less powerful systems that are interesting from other points of view. It also focuses on the development trends of the supercomputer industry and describes the most famous projects for building future exascale supercomputers.

Keywords: supercomputer, performance, efficiency, processor, Top500, parallelism, accelerator, interconnect.

Acknowledgements: Section 3 of the work was supported by the Russian Foundation for Basic Research, grant 18-29-03230 mk. Sections 2 and 4 were prepared as part of the state-sponsored topic "Development of Software Tools for Life Cycle Support and Efficiency Assurance of Supercomputer Applications, Systems and Centers" (registration number AAAA-A21-121011690003-6).

For citation: A. S. Antonov, I. V. Afanasyev, and Vl. V. Voevodin, "High-Performance Computing Platforms: Current Status and Development Trends," Numerical Methods and Programming. **22** (2), 138–181 (2021). doi: 10.26089/NumMet.v22r210.

СОДЕРЖАНИЕ

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

1. Введение. Высокопроизводительные вычислительные системы сегодня востребованы исключительно широко, что определяется растущим потенциалом предсказательного моделирования и технологий искусственного интеллекта для решения задач науки, промышленности, приоритетных задач государства. В настоящее время компаниями и научными группами предлагается целое множество вариантов построения вычислительных систем, и в каждом случае делается акцент на тех или иных особенностях архитектуры, которые помогают увеличивать реальную производительность компьютеров. В предлагаемом обзоре проведен разносторонний анализ современного состояния высокопроизводительных вычислительных систем, рассмотрены их основные классы, особенности и нюансы организации архитектуры, позволяющие достигать высоких скоростей работы, описаны тенденции, характерные для данной области.

Сегодня разнообразие высокопроизводительных вычислительных систем велико, поэтому и обзор построен в виде множества срезов, позволяющих оценить компьютеры данного класса с разных точек зрения.

Данный обзор посвящен суперкомпьютерной технике, и поэтому в нем упоминаются только компоненты и технологии, которые используются в современных суперкомпьютерах, признаком чего является вхождение в список Top500 [1] наиболее мощных компьютеров мира или же официально анонсированные суперкомпьютерные проекты.

Оценка производительности суперкомпьютеров не ограничивается только списком Top500. Есть большое число других суперкомпьютерных рейтингов (например, Graph500 [2], Green500 [3], HPCG [4] и др.), оценивающих высокопроизводительные вычислительные системы с других точек зрения.

Отметим также проект Algo500 [5], в рамках которого предлагается возможность использования любой реализации любого вычислительного алгоритма в качестве бенчмарка, используемого для построения нового суперкомпьютерного рейтинга. Это позволяет не ограничиваться только несколькими стандартными метриками, а получить полноценную оценку эффективности суперкомпьютера на множестве различных алгоритмов.

Мы не рассматриваем организацию распределенных вычислительных сред, хотя с помощью таких технологий можно набирать вычислительные ресурсы с огромной производительностью (например, в апреле 2020 метакомпьютерный проект Folding@Home [6] обошел по суммарной производительности объединенных ресурсов сумму всех суперкомпьютеров из последней редакции Top500).

В данном обзоре мы также не рассматриваем и организацию дисковых подсистем, хотя это является неотъемлемой частью области высокопроизводительных вычислений.

2. Классификации суперкомпьютеров и высокопроизводительных многопроцессорных вычислительных систем. Разновидностей архитектур высокопроизводительных вычислительных систем существует огромное множество. Чтобы его как-то упорядочить, в разное время предлагались различные способы их классификации [7–9].

Один из первых вариантов классификаций параллельных вычислительных систем был предложен в 1966 г. М. Флинном [10, 11]. В данной простой классификации, ставшей со временем классической, все множество вычислительных систем разбивается на четыре класса в зависимости от особенностей потоков команд и потоков данных.

В класс SISD (single instruction stream / single data stream; одиночный поток команд и одиночный поток данных) входят классические последовательные компьютеры. В таких системах все команды обрабатываются последовательно и каждая инициирует одну операцию с одним потоком данных.

В класс SIMD (single instruction stream / multiple data stream; одиночный поток команд и множественный поток данных) входят в первую очередь векторные компьютеры, в которых одна арифметическая операция может выполняться над всеми элементами вектора.

Класс MISD (multiple instruction stream / single data stream; множественный поток команд и одиночный поток данных) оказался вырожденным и пустым, поскольку не появилось архитектур компьютеров, предусматривающих обработку разными вычислительными устройствами одного и того же потока данных.

Наконец, в класс MIMD (multiple instruction stream / multiple data stream; множественный поток команд и множественный поток данных) входит большинство современных компьютеров, поскольку практически все из них предполагают наличие нескольких устройств обработки команд, каждое из которых работает со своим потоком данных.

Несмотря на частые упоминания, классификация Флинна не представляется особенно интересной. Это определяется крайней перегруженностью одного класса (MIMD), притом что остальные классы имеют

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

в первую очередь теоретическое или историческое значение. Существует также большое количество других классификаций вычислительных систем (Фенга, Хокни, Хендлера, Шнайдера, Скилликорна и т.д.) [7], все они могут представлять определенный интерес, но обладают и некоторыми недостатками, не позволяющими широко использовать их для классификации существующих высокопроизводительных вычислительных систем. В рамках данной обзорной статьи мы их рассматривать не будем.

На практике часто пользуются разбиением всех вычислительных систем на два основных класса — на системы с общей памятью (мультипроцессоры, SMP-системы) и системы с распределенной памятью (мультикомпьютеры) [8]. Первые из них обычно проще и эффективнее программируются, а вторые позволяют получать более масштабные вычислительные комплексы. Все чаще большие суперкомпьютерные системы строятся на основе гибридной архитектуры, когда на базе систем с общей памятью (вычислительных узлов) при помощи высокоскоростной коммуникационной сети строят более масштабные системы с распределенной памятью. Кроме того, в состав вычислительных узлов стали включать ускорители, у которых обычно есть своя локальная память, что приводит к тому, что и вычислительный узел может не обладать в полной мере свойствами компьютера с общей памятью.

3. Архитектура современных суперкомпьютеров и высокопроизводительных многопроцессорных вычислительных систем.

3.1. Процессорная основа. Иной взгляд на множество суперкомпьютеров мира можно получить, если проанализировать их процессорную основу. В данном разделе проанализированы основные современные многоядерные процессоры, графические ускорители и векторные процессоры, их основные аппаратные характеристики и особенности построения, а также преимущества и недостатки с точки зрения реализации различных классов параллельных приложений. Особый интерес представляют различные способы векторной обработки, присутствующие в архитектуре и многоядерных процессоров от Intel, AMD, ARM и графических ускорителей NVIDIA, и современных векторных процессоров компании NEC. Специальное внимание в обзоре уделяется различным модификациям архитектуры ARM, активно проникающей в область высокопроизводительных вычислений за счет сочетания высокого показателя энергоэффективности и производительности, сравнимой с процессорами компаний Intel и AMD, что и вызывает большой интерес со стороны компаний-производителей, например, Fujitsu, Huawei, Hewlett-Packard Enterprise и других.

§ 3.1.1. Особенности построения современных вычислительных устройств.

В данном разделе будут рассмотрены основные понятия, связанные с основными принципами и особенностями построения современных параллельных вычислительных систем: используемые системы команд, подходы к векторной обработке данных, виды параллелизма, реализация суперскалярности и др. Данная часть обзора позволяет достаточно подробно описать основные характерные свойства, присущие большинству современных процессоров и графических ускорителей, значительно упростив последующее описание конкретных моделей процессоров, которые будут рассматриваться впоследствии.

Системы команд. Система команд — это комбинация инструкций, доступных для выполнения процессору, и регистров, доступных для обработки. Использование различных наборов команд может как значительно облегчить, так и усложнить разработку процессора с более высокой производительностью или более низким энергопотреблением, так как система команд в значительной мере определяет аппаратный дизайн процессора.

x86 — это набор команд и одноименная архитектура, впервые реализованные в процессорах компании Intel в 1978 г. Название образовано от двух цифр, которыми заканчивались названия процессоров Intel ранних моделей. За время своего существования набор команд постоянно расширялся, сохраняя, однако, совместимость с предыдущими поколениями. Помимо Intel, архитектура также была реализована в процессорах других производителей: AMD, VIA, Transmeta, IDT и др. В настоящее время для 32-разрядной версии архитектуры существует еще одно название — IA-32 (Intel Architecture — 32). Архитектура x86 относится к типу архитектур CISC (Complex Instruction Set Computer), вследствие чего размер инструкции не фиксирован, а в архитектуре реализовано достаточно большое число сложных инструкций, выполняющих различные задачи. Одним из основных преимуществ CISC является то, что компилятор выполняет мало работы для перевода CISC-ассемблера в язык высокого уровня. Поскольку длина машинного кода относительно мала, для хранения инструкций требуется малый объем оперативной памяти, что в прошлом являлось важным преимуществом. На сегодняшний день в силу развития компиляторов и удешевления памяти, данные преимущества не являются столь принципиальными. Доступ к памяти происходит по

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

"словам", которые размещаются в памяти по принципу little-endian (от младшего к старшему разряду). Современные процессоры архитектуры x86 включают в себя декодеры сложных CISC-команд для преобразования их в упрощенный внутренний формат (так называемые микрооперации), с последующим конвейерным, суперскалярным и внеочередным выполнением. Вследствие этого данные микрооперации имеют ряд общих свойств с определенными типами RISC (reduced instruction set computer) инструкций, благодаря чему данный подход иногда называют "RISC-ядро" или "RISC translation", в том числе в интересах маркетинга. Существует ряд расширений архитектуры x86, разрабатываемых различными производителями микропроцессоров, которые, зачастую, присутствуют в одних моделях и отсутствуют в других. Примеры данных расширений: MMX — "мультимедийный" набор инструкций, выполняющих за одну машинную инструкцию несколько действий, характерных для процессов кодирования или декодирования потоковых аудио- и видеоданных, SSE, AVX — наборы векторных (SIMD) инструкций, и др.

В 2001 г. компании Intel и Hewlett Packard разработали архитектуру IA-64 для замены 32-разрядной x86, которая достаточно долгое время использовалась в сегменте серверов в процессорах Itanium. Архитектура IA-64 кардинально отличалась от IA-32, а для совместимости со старым программным обеспечением IA-64 использовала режим эмуляции, имевший очень низкую производительность. В результате в 2019 г. IA-64 была окончательно вытеснена 64-разрядной архитектурой x86-64, предложенной компанией AMD.

VLIW (very long instruction word, "очень длинное командное слово") — архитектура процессоров с несколькими вычислительными устройствами. Данная архитектура характеризуется тем, что одна инструкция процессора содержит несколько операций, которые должны выполняться параллельно. Команда VLIW-процессора состоит из набора полей, каждое из которых отвечает за свою операцию. Если какая-то часть процессора на данном этапе не востребована, то соответствующее поле команды не задействуется. В процессорах VLIW задача распределения операций по исполняющим устройствам решается во время компиляции, в результате чего в инструкциях явно указано, какое вычислительное устройство должно выполнять какую команду. Компилятор сам выявляет параллелизм в программе и явно сообщает аппаратуре, какие операции не зависят друг от друга. Код для VLIW-процессоров содержит точный план того, как процессор будет выполнять программу: когда будет выполнена каждая операция, какие функциональные устройства будут работать, какие регистры какие операнды будут содержать и т.д. Подход VLIW сильно упрощает архитектуру процессора, перекладывая задачу распределения вычислительных устройств и распараллеливания вычислений на компилятор. В этом и заключаются основные преимущества VLIW-процессоров: более простая вычислительная логика позволяет поместить большое количество вычислительных устройств в процессор, а поскольку в процессоре отсутствуют большие и сложные узлы, то сильно снижается и энергопотребление. При этом на уровне компилятора (который анализирует всю программу) зачастую значительно проще определять доступный ресурс параллелизма на уровне инструкций в суперскалярных процессорах, что позволяет создавать более эффективные программы. В то же время код для VLIW обладает невысокой плотностью. В случае, если компилятор не может выделить в программе достаточное число параллельных инструкций, он помещает явные пустые инструкции на простаивающие устройства, в результате чего программы для VLIW-процессоров могут быть гораздо длиннее, чем аналогичные программы для традиционных архитектур. Кроме того, невозможен перенос программ между различными поколениями VLIW-процессоров, так как скомпилированная программа работает корректно только на процессоре с тем же числом вычислительных устройств и латентностью инструкций [13].

RISC (Reduced Instruction Set Computer) — это архитектура процессора, в которой быстродействие увеличивается за счет упрощения инструкций, чтобы их декодирование было более простым, а время выполнения — меньшим. Термин "сокращенный" (reduced) в названии описывает тот факт, что сокращен объем работы (число тактов), за которое производится выполнение каждой инструкции, а не число инструкций, в то время как сложные инструкции CISC-процессоров могут требовать сотен циклов доступа к памяти для своего выполнения. Важным преимуществом RISC-процессоров является то, что модуль по декодированию команд занимает на чипе меньше места, вследствие чего значительно повышается энергоэффективность вычислений. На сегодняшний день ARM-процессоры используют упрощенный набор инструкций — RISC, и эта аббревиатура расшифровывается как Advanced RISC Machines (продвинутые RISC-машины).

Параллелизм. Параллельные вычислительные системы — это физические компьютерные, а также программные системы, реализующие тем или иным способом параллельную обработку данных за счет различного ресурса аппаратного параллелизма. В современных суперкомпьютерах присутствуют различ-

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

ные уровни параллелизма. Так, существует параллелизм на уровне узлов суперкомпьютера. Внутри каждого из узлов параллельно могут работать различные процессоры (сокеты), а параллельно с ними могут работать и сопроцессоры (ускорители). Внутри процессоров и графических ускорителей параллельно работают различные ядра, причем каждое из ядер зачастую может осуществлять параллельную обработку данных за счет использования векторной обработки данных, конвейерности и/или суперскалярности. Вследствие столь большого разнообразия способов аппаратной поддержки параллельной обработки данных, в данном обзоре большое внимание будет уделено как описанию базовых принципов параллельной обработки данных, так и их реализации в конкретных современных высокопроизводительных платформах и архитектурах.

Конвейерность. Идея конвейерной обработки заключается в выделении отдельных этапов в выполнении некоторой общей команды. Этот принцип подразумевает, что в каждый момент времени процессор работает над различными стадиями выполнения нескольких команд, причем на выполнение каждой стадии выделяются отдельные аппаратные ресурсы. На каждом такте каждая команда в конвейере продвигается на следующую стадию обработки, выполненная команда покидает конвейер, а новая поступает в него. Для примера рассмотрим случай, когда одному конвейерному устройству необходимо обработать 100 команд, каждая из которых может быть разбита на 5 микроопераций (этапов), выполняемых за 1 такт. В случае использования обычного устройства все данные команды будут обработаны за 500 тактов, а в случае использования конвейерного устройства — за $5 + 99 = 104$ такта. В результате ускорение по сравнению с последовательным устройством — почти в пять раз (по числу ступеней конвейера). Здесь важно отметить, что при дублировании всего вычислительного устройства в число раз, равное числу ступеней конвейера, можно получить идентичное ускорение, однако такой подход приводит к существенно большему как объему аппаратуры, так и ее стоимости. Поэтому на сегодняшний день конвейерная обработка данных реализуется в большинстве современных архитектур в виде параллелизма на уровне инструкций. К примеру, уже был упомянут конвейер RISC-процессора, состоящий из пяти ступеней: 1) получение инструкции, 2) декодирование инструкции, 3) выполнение, 4) доступ к памяти, 5) запись в регистр.

Векторная обработка. Векторная обработка (или векторизация) — это вид распараллеливания программы, при котором скалярные операции, обрабатывающие по паре операндов, заменяются на операции над векторами из данных, обрабатывающие все либо несколько элементов входных векторов в каждый момент времени. Вектор — это упорядоченный набор однотипных данных, размещенных на специальных векторных регистрах. Таким образом, векторная обработка данных позволяет выполнять векторные инструкции за время, близкое ко времени выполнения скалярных инструкций, благодаря чему в ряде случае получается достигать ускорения, пропорционального длине используемого вектора.

Традиционно векторная обработка используется при работе с регулярными структурами данных, например строками или столбцами матриц, линейными массивами и т.д. Однако также существуют и подходы к векторной обработке таких нерегулярных структур данных, как графы с использованием современных векторных вычислительных систем [14]. Кандидатами для векторизации обычно являются самые внутренние циклы программы, причем большинство современных компиляторов поддерживают автоматическую векторизацию при указании соответствующих опций либо директив.

С архитектурной точки зрения, векторная обработка данных может быть реализована различными способами. Первый способ основан на использовании “векторных расширений инструкций” — добавлении специальных векторных устройств в ядра центрального процессора. Примерами векторных расширений могут служить наборы инструкций MMX, SSE, AltiVec, AVX-512 и многие другие, присутствующие в большинстве современных многоядерных центральных процессоров. К примеру, векторное расширение AVX-512, представленное в процессорах Intel Xeon и Xeon Phi, позволяет обрабатывать векторы длиной 512 бит. В AVX-512 представлены различные операции над векторами: арифметические, логические, маскирования, битовые, gather/scatter и др. Важно отметить, что разные типы операций имеют разную латентность: в то время как арифметические операции над регистрами могут быть выполнены за 1 такт, время выполнения векторных операций работы с памятью (load/store, gather/scatter) определяется особенностями подсистемы памяти, вследствие чего данные операции всегда выполняются за значительно большее число тактов, что далеко не всегда способствует ускорению относительно скалярного аналога программы. Второй способ — это векторно-конвейерная обработка данных, основанная на обработке длинных векторов на конвейере из векторных устройств меньшей длины, как это реализовано в векторной системе NEC SX-Aurora TSUBASA.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

Суперскалярность. Под суперскалярностью понимается способность процессора использовать параллелизм на уровне инструкций, то есть исполнять несколько инструкций различных типов одновременно (на каждом такте работы процессора). Подавляющее большинство современных процессоров оборудованы функциональными устройствами различных типов, нацеленных на обработку различных операций: целочисленных сложений и умножений, целочисленных делений, обработки чисел с плавающей точкой, устройств работы с подсистемой памяти, векторными устройствами и др. Как правило, высокая производительность достигается, если компилятор может упорядочить программные инструкции для максимального использования имеющихся функциональных устройств таким образом, чтобы на каждом такте планировщик процессора мог предоставить различным функциональным устройствам как можно большее число инструкций различных типов.

Одновременная многопоточность (гипертрединг). Технология одновременной многопоточности позволяет одному физическому ядру обрабатывать одновременно несколько (обычно два или больше) логических потока. Первой реализацией одновременной многопоточности на практике стала технология гипертрединг [15, 16] компании Intel. На сегодняшний день аналоги технологии гипертрединг представлены и в других архитектурах: IBM Power (SMT) и ARM.

Использование одновременной многопоточности мотивируется тем фактом, что современное процессорное ядро состоит из большого числа разнотипных блоков, выполняющих инструкции различных типов, причем на большинстве реальных задач загрузка ядра процессора далека от 100% по причине использования в каждый момент времени блока определенного типа. Значительную роль здесь играет и так называемая “стена памяти” — постоянно растущая разница в быстроте работы процессорных ядер и подсистемы памяти, из-за которой многие реальные приложения загружают ресурсы ЦПУ на 5–10%, большую часть времени ожидая получения данных из памяти. Технология одновременной многопоточности позволяет незанятым блокам процессора обрабатывать команды другой нити. Удвоения производительности обычно не происходит по понятным причинам — очень часто получается так, что двум задачам нужен один и тот же вычислительный блок в процессоре (к примеру, в memory-bound задачах большую часть времени работы будет ожидаемо занимать работа с блоком, отвечающим за обработку обращений к памяти), в результате чего происходит простой виртуальных нитей. Однако в ряде случаев одновременная многопоточность все же позволяет ускорить определенные классы приложений на 20–30% [17].

FMA-операции. В суперкомпьютерных приложениях широко распространена операция умножения двух чисел с последующим сложением результата с третьим числом: `a = a + b*c`. В случае, если данная операция выполняется над числами с плавающей запятой, то может выполняться либо двойное округление, либо однократное, когда сложение происходит с более точным внутренним представлением произведения. Версия с однократным округлением носит название fused multiply-add (FMA). Большинство современных процессоров имеют специализированный блок умножения-сложения (Multiply-ACcumulate, MAC), состоящий из умножителя, реализующего комбинационную логику, и сумматора, а также аккумулятора, в котором сохраняется результат. FMA-операции реализуются в том числе и внутри векторных расширений, например AVX-512, ARM Neon и др., а также внутри вычислительных ядер графических ускорителей. Использование FMA-операций позволяет в ряде случаев удвоить производительность приложений, выполняющих большой объем вычислений с плавающей точкой, поскольку операции сложения-умножения занимают один такт вместо двух.

Пиковая производительность. Оценка теоретической пиковой производительности центрального процессора традиционно сводится к простому умножению тактовой частоты на количество инструкций с плавающей точкой, которое процессор может выполнить за один такт. Однако наличие в современных центральных процессорах векторных расширений, FMA устройств, одновременной многопоточности, режима “boost” и других особенностей существенно затрудняет вычисление того, сколько конкретный процессор может выполнить операций с плавающей точкой за 1 такт. В итоге пиковая производительность может быть подсчитана по следующей формуле:

$$\text{peak\_performance} = \text{frequency} \times \text{SIMD\_width} \times \text{FMA} \times \text{superscalar} \times \text{cores},$$

где frequency — тактовая частота процессора (иногда с учетом режима boost, а иногда без, так как процессор не способен долго выполнять вычисления в данном режиме), FMA — множитель 1 или 2 в зависимости от того, включает ли набор инструкций процессора FMA-операции, SIMD_width — количество чисел одинарной или двойной точности, которое может быть обработано векторными элементами процес-

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

сора, superscalar — дополнительный множитель, в случае если процессор поддерживает суперскалярность (к примеру, в некоторых процессорах Intel установлено по два функциональных устройства с 512-битными операндами на ядро). Вследствие наличия множителя SIMD_width пиковая производительность при работе с одинарной точностью в 2 раза выше, чем с двойной (кроме случаев, когда в процессорах отсутствует векторная обработка данных в явном виде).

Поддержка работы с данными различной точности. Одной из важнейших характеристик любых вычислений является точность получаемых результатов. Традиционно, вычисления с плавающей точкой проводятся либо с одинарной, либо с двойной точностью, когда для представления вещественных чисел выделяется 4 или 8 байт соответственно. При этом скорость обработки чисел различной точности существенно зависит как от свойств целевой архитектуры, так и самой программы. К примеру, в архитектуре x86 скалярные операции над вещественными числами как двойной, так и одинарной точности реализуются путем приведения к 80-битным числам, вследствие чего производительность в ряде случаев может оказаться идентичной. Однако для векторизированных программ производительность работы с одинарной точностью может быть до двух раз выше вследствие обработки векторными инструкциями в два раза более длинных векторов. Дополнительный вклад вносит и работа с подсистемой памяти: в случае, если приложение относится к классу data-intensive, работа с одинарной точностью позволяет в два раза сократить объем загружаемых из памяти данных. Другой пример — архитектура NVIDIA GPU, в которой для работы с одинарной и двойной точностью предусмотрены CUDA-ядра разных типов, причем ядер для работы с одинарной точностью в два раза больше, чем с двойной.

Другой важной тенденцией является поддержка вычислений с половинной (а иногда и меньшей) точностью. К примеру, графические ускорители NVIDIA Tesla P100 поддерживают специальные векторные FMA-инструкции над парами из чисел половинной точности (2 байта), благодаря чему производительность графического ускорителя P100 при вычислениях с половинной точностью (18.7 Tflops) в два раза выше производительности с одинарной точностью (9.3 Tflops) и в четыре раза выше, чем с двойной (4.7 Tflops).

В последних поколениях графических ускорителей NVIDIA (начиная с Volta) дополнительно добавлены специализированные тензорные ядра, позволяющие работать с половинной точностью еще более эффективно, о которых будет рассказано более подробно в следующем разделе.

Tensor Cores. Тензорные ядра были впервые представлены компанией NVIDIA в архитектуре Volta GPU. Каждое из ядер данного типа выполняет операцию $D = AB + C$, где A, B, C и D — это матрицы размера 4×4. Тензорные ядра устанавливаются в современные графические ускорители NVIDIA в дополнение к традиционным ядрам, выполняющим обработку скалярных инструкций: к примеру, потоковый мультипроцессор графического ускорителя V100 оборудован восемью тензорными ядрами, в то время как ядер других типов существенно больше — 160. Тензорные ядра позволяют работать с различной точностью — TF32 (числа с плавающей точкой, использующие 10-битную мантиссу и 8-битную экспоненту, что позволяет им поддерживать достаточный для задач глубокого обучения числовой диапазон [18]), FP16 (числа с плавающей точкой половинной точности), INT8 и INT4 (8-битные и 4-битные целые числа), выполняя по 64 FMA-операции смешанной точности за один такт, что позволяет достигать крайне высокой производительности на задачах глубокого обучения.

§ 3.1.2. Центральные процессоры.

В данном разделе будут рассмотрены некоторые семейства и конкретные представители центральных процессоров, наиболее активно использующиеся в современных высокопроизводительных вычислениях. Двумя основными критериями выбора процессоров для рассмотрения будут: 1) их использование для построения наиболее производительных систем из списка Top500 и 2) использование нетрадиционных решений при их построении (например, архитектура ARM, векторно-конвейерная обработка данных и др.). Дополнительное внимание также будет уделено российским разработкам.

Семейство процессоров Intel. Процессоры Intel имеют наибольшую долю в современных высокопроизводительных вычислениях: так, 459 из 500 систем из списка Top500 (за 2020 г.) используют решения данной компании [19]. Однако в то же время среди первых десяти систем в редакции списка за 2020 г. только четыре используют процессоры Intel: Tianhe-2A, HPC5, Frontera и Dammam-7. Последними серверными микроархитектурами процессоров Intel на сегодняшний день являются: Skylake-SP (2015 г.), Cascade Lake-SP (2018 г.), Cooper Lake-SP (2019 г.) и Ice Lake-SP (2020 г.) [20]. Далее аппаратные характеристики будут приведены на примере процессора Intel Xeon Gold 6248 микроархитекту-

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

ры Cascade Lake-SP. Выбор обусловлен тем, что процессоры данной микроархитектуры Xeon Gold 6248 установлены в системе DAMMAM-7 и являются новейшими из присутствующих среди первых десяти систем списка Top500. Техпроцесс — 14 nm. Данные процессоры имеют 20 ядер, каждое с тактовой частотой 2.5 GHz (3.9 GHz в режиме boost) и возможностью запускать до двух нитей на каждое ядро при помощи технологии гипертрединг. Каждое из ядер может работать с векторными инструкциями AVX-512 длины 512 бит, вследствие чего пиковая производительность всего процессора составляет 3.2 Tflops (2.5(freq) × 16(simd) × 20(cores) × 2(fma) × 2(vector units)) на одинарной и 1.6 Tflops на двойной точности. Иерархия памяти Intel Xeon Gold 6248 состоит из кэша L1 объема 640 KB, кэша L2 объема 20 MB и кэша L3 объема 27.5 MB. Кэши L1, L2 — приватные для каждого из ядер, в то время как кэш L3 — разделяемый. Данный процессор использует 6-канальную DDR4 память с пиковой пропускной способностью до 140 GB/s и максимальным объемом до 1 TB [21], поддерживает интерконнект PCIe версии 3.0 с 48 каналами, а также многосокетные конфигурации.

Семейство процессоров AMD. Начиная с 2017 г. компания AMD выпускает процессоры семейства с кодовым названием EPYC. Процессоры AMD EPYC, так же как и процессоры Intel, основаны на архитектуре и одноименном семействе команд x86. На сегодняшний день существуют два поколения данных процессоров — Naples (выпущенные в 2017 г.) и Rome (выпущенные в 2019 г.), в то время как в 2021 г. ожидается запуск в серийное производство поколения Milan. Процессоры AMD EPYC Rome также часто используются в современных суперкомпьютерных вычислениях: на их основе собраны 5-я и 16-я (на ноябрь 2020 г.) системы SELENE и HAWK списка Top500. Всего в списке Top500 на момент 2020 г. установлена 21 система на основе процессоров компании AMD, что является вторым показателем после Intel. Установленные в двух вышеупомянутых системах процессоры имеют модель AMD EPYC 7742 [22]. Данные процессоры оборудованы 64 ядрами, каждое с тактовой частотой 2.25 GHz (3.4 GHz в режиме boost) и возможностью запускать до двух потоков в режиме SMP. Кроме того, каждое из ядер может использовать векторные инструкции AVX2 длины 256 бит, в том числе FMA, благодаря чему пиковая производительность процессора составляет 2.3 Tflops (2.25(freq) × 8(simd) × 2(fma) × 64(cores)) на одинарной точности и 1.15 Tflops на двойной точности. Техпроцесс AMD EPYC 7742 — 14 nm. Иерархия кэш-памяти состоит из трех уровней L1, L2 и L3 объема 4 MB, 32 MB и 256 MB соответственно. Данные процессоры используют 8-канальную DDR память с пиковой пропускной способностью до 204 GB/s и объемом до 4 TB на сокет, а также PCIe 4.0 с 128 каналами. Системы с несколькими сокетами строятся за счет использования межчиповых соединений Infinity Fabric [23].

Семейство процессоров IBM Power. Процессоры Power компании IBM являются третьей группой процессоров, активно использующихся в современных высокопроизводительных вычислениях. На сегодняшний день наиболее распространены процессоры поколения Power9 (2017 г. выпуска). В 2020 г. было начато производство процессоров Power10, которые, однако, еще не были установлены ни в одну из систем списка Top500, в то время как процессоры Power9 установлены в системы Summit и Sierra, расположенные на второй и третьей позициях рейтинга за 2020 г. Процессоры Power9 выпускаются в двух основных вариантах: с 12 ядрами, каждое из которых может запускать 8 потоков на ядро (режим SMT8), и с 24 ядрами, каждое из которых может запускать 4 потока на ядро (SMT4). Тактовая частота каждого ядра составляет 2.3 GHz (3.8 GHz в режиме boost), каждое ядро может работать с векторными инструкциями Altivec длины 128 бит (выполнять 8 FMA операций с одинарной точность за такт), вследствие чего пиковая производительность всего процессора составляет 1.7 Tflops (2.3(freq) × 24(cores) × 4(SIMD) × 2(fma) × 2(vector units)) на одинарной и 0.85 Tflops на двойной точности. Техпроцесс IBM Power9 — 14 nm. Иерархия кэш-памяти состоит из кэшей L1 (приватный для каждого ядра) объема 32 KB, L2 (разделяемого каждыми двумя ядрами) объема 512 KB, а также разделяемого всеми ядрами кэша L3 объема 120 MB. Подсистема памяти Power9 также существенно зависит от конфигурации, которых доступно две: scale out и scale up. Вариант scale out предназначен для традиционных кластеров и суперкомпьютеров, использующих однопроцессорные либо двухсокетные конфигурации. В конфигурации scale out поддерживается до 8 каналов DDR4 памяти с суммарным объемом памяти до 4 TB и пропускной способностью до 120 GB/s. Вариант scale up разработан для NUMA-серверов с четырьмя или более сокетами, в которых необходима поддержка большого объемы памяти с высокой пропускной способностью. В конфигурации scale up Power9 имеет два контроллера памяти, способных управлять четырьмя каналами Direct Media Interface (DMI). Каждый из каналов DMI подключается к одному выделенному чипу Centaur буфера памяти, который, в свою очередь, соединяется еще с четырьмя каналами DDR4, что позволяет процессору Power9 работать

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

с 32 каналами DDR4 памяти, тем самым обеспечивая дополнительные 128 MB кэш-памяти четвертого уровня [24]. Процессоры Power9 используют интерконнект PCIe версии 4.0 с 48 каналами, а также поддерживают NVLINK 2.0.

Процессоры ARM. ARM (Advanced RISC Machine) — это система команд, семейство описаний, а также готовых топологий 32-битных и 64-битных микропроцессорных ядер, разрабатываемых компанией ARM Limited. Многие производители процессоров, являющиеся лицензиатами ARM, проектируют собственные топологии ядер на базе системы команд ARM, некоторые из которых будут описаны далее в разделе. Одно из важнейших преимуществ процессоров на основе ARM — высокая энергоэффективность (об этом более подробно далее в разделе 4.1), благодаря чему решения на основе архитектуры ARM все более часто используются в высокопроизводительных и серверных вычислениях. Подтверждением этого является суперкомпьютер Fugaku — номер 1 в списке Top500 за 2020 г., в основе которого лежат ARM-процессоры A64FX компании Fujitsu. Наиболее современной архитектурой ARM-процессоров является ARMv8 (иногда используется название AArch64). Данная архитектура получила 64-битный набор инструкций и, как следствие, возможность работать с большим объемом оперативной памяти (4 GB и больше). Архитектура команд ARMv8-A поддерживается основными дистрибутивами Linux (SLES, OpenSuSE, RHEL и CentOS), и, кроме того, для ARMv8-A доступен также ряд компиляторов и библиотек для высокопроизводительных вычислений.

- Процессор A64FX (основа узла Fugaku) был разработан компанией Fujitsu в 2019 г. A64FX оборудован 48 вычислительными ядрами, работающими на тактовой частоте 1.8, 2.0 или 2.2 GHz (в зависимости от модификации), а также 2 либо 4 вспомогательными ядрами, необходимыми для работы операционной системы, ввода/вывода, асинхронных MPI коммуникаций и других вспомогательных задач. Это первый в мире процессор, имеющий архитектуру ARMv8.2-A SVE (Scalable Vector Extenstion), которая позволят работать с векторами длины 512 бит. Благодаря этому пиковая производительность всего процессора в максимальной конфигурации составляет 6.8 Tflops на одинарной точности и 3.4 Tflops (2.2(freq) × 48(cores) × 16(simd) × 2(fma) × 2(vector units)) на двойной. Набор векторных инструкций ARMv8.2-A SVE дополнительно позволяет производить вычисления с половинной точностью, благодаря чему производительность процессора в данном режиме увеличивается до 13.6 Tflops. Техпроцесс A64FX — 7 nm. Каждое ядро имеет приватный L1 кэш размера 64 KB. Ядра процессора группируются в четыре группы Core Memory Group (CMG), каждая из которых состоит из 13 ядер (12 вычислительных и одного вспомогательного), которые, в свою очередь, имеют общий кэш L2 размера 8 MB и 8 GB быстрой HBM памяти. Таким образом, весь процессор имеет кэш L2 общего объема 32 MB и оборудован 32 GB HBM памяти с пропускной способностью 1024 GB/s. Для соединения с другими узлами A64FX использует две линии внешнего интерфейса Tofu с пропускной способностью 28 Gb/s [25].

- Российская компания Baikal Electronics разрабатывает процессоры Baikal-M на основе архитектуры ARM, которые поступили в производство в 2019 г. Данные процессоры имеют 8 ядер ARM Cortex-A57 с тактовой частотой в 1.5 GHz (архитектура ARMv8-A). Baikal-M поддерживает векторные инструкции ARM NEON длины 128 бит, благодаря чему пиковая производительность всего процессора составляет 96 Gflops (1.5(freq) × 8(cores) × 4(simd) × 2(fma)) на одинарной точности и 48 Gflops на двойной. Техпроцесс Baikal-M — 28 nm. Каждое из ядер имеет приватный L1 кэш объема 32 KB. Ядра Baikal-M сгруппированы в 4 кластера по 2 ядра, причем каждый из кластеров имеет разделяемый его ядрами кэш L2 размера 1 MB (итого 4 MB на весь процессор), в то время как кэш L3 — общий для всех кластеров и имеет объем 8 MB. Процессор Baikal-M оборудован модулями памяти Crucial 8 GB DDR4 с пропускной способностью до 20 GB/s и объемом до 128 GB [26]. Большое количество высокоскоростных интерфейсов и производительность, сравнимая с десктопными процессорами Atom E3940 и Core i3-7300T от Intel, делают этот процессор достаточно конкурентоспособным в данном сегменте. Дополнительно поддерживается установка в виде сопроцессора специализированного восьмиядерного графического ускорителя Mali-T628 MP8.

- Китайская компания Huawei также в 2019 г. начала разработку процессоров Kunpeng на базе ARMv8. Процессоры Kunpeng имеют несколько модификаций, разрабатываемых как для высокопроизводительных вычислений, так и настольных компьютеров. Число ядер варьируется от 24 до 64 с тактовой частотой в диапазоне от 2.4 до 3 GHz. Наиболее производительная модификация — 64-ядерные про-

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

цессоры Kunpeng 920-6426, работающие на тактовой частоте в 2.6 GHz. Kunpeng 920-6426 поддерживает векторные инструкции ARM NEON длины 128 бит, благодаря чему пиковая производительность всего процессора составляет 1.3 Tflops (2.6(freq) × 64(cores) × 4(simd) × 2(fma)) на одинарной точности и 0.65 Tflops на двойной. Техпроцесс Kunpeng — 7 nm. Иерархия кэш-памяти состоит из трех уровней: приватных для каждого из ядер кэшей L1 объема 64 KB и L2 объема 512 KB, а также разделяемого всеми ядрами кэша L3 объема до 69 MB. Процессоры Kunpeng 920-6426 используют 8-канальную DDR4 память с пиковой пропускной способностью в 204 GB/s и максимальным объемом до 2 TB. Дополнительно данные процессоры поддерживают интерконнект PCIe 4.0 с 40 каналами [27].

- ThunderX — серверные процессоры на базе архитектуры ARM, выпускаемые компанией Cavium, которую позднее поглотила компания Marvell. На момент написания данного обзора самыми современными семействами являются ThunderX2 (2018 г. выпуска), а также анонсированными, но еще не поступившими в производство процессорами Marvell ThunderX3 (2021). Данные процессоры в основном нацелены на применение в облачных системах с поддержкой ARM-приложений, включая запуск приложений Android. ThunderX2 основаны на микроархитектуре Vulcan [28] и используют 32 ядра с тактовой частотой 2.5 GHz, соединяемых интерконнектом с кольцевой топологией. Каждое из ядер — суперскалярное, имеет поддержку out-of-order вычислений, позволяет запускать до четырех потоков в SMP режиме, а также использует по 2 устройства выполнения векторных инструкций ARM NEON длины 128 бит. Таким образом, пиковая производительность всего процессора составляет 1.2 Tflops (2.5(freq) × 32(cores) × 4(simd) × 2(fma) × 2(vector units)) на одинарной точности и 0.6 Tflops на двойной [29]. Техпроцесс ThunderX2 — 16 nm. Иерархия памяти состоит из приватных для каждого ядра кэшей L1 и L2 размера 32 KB и 256 KB на ядро соответственно, разделяемого кэша L3 размера 30 MB, а также 8-канальной DDR4 памяти с пропускной способностью 171 GB/s. Также поддерживается 56-канальный интерконнект PCIe 3.0.

  В процессорах ThunderX3 планируется увеличение числа ядер до 60 в односокетном варианте и 96 в двухсокетном, тактовой частоты до 3.1 GHz, и, кроме того, будет изменена топология интерконнекта ядер на кольцо с тремя подкольцами. Кроме того, размер кэша L3 будет существенно увеличен до 90 MB [30]. Самый главный недостаток микроархитектуры TX2 по сравнению с Intel Cascade Lake — работа с векторами (векторные расширения ARM NEON позволяют обрабатывать лишь 4 числа одинарной точности за такт, в то время как AVX-512 — 16 таких чисел, вследствие чего планируется увеличение числа векторных устройств в каждом из ядер до четырех). Ожидаемые TX3 будут иметь до 96 ядер и станут конкурировать с Fujitsu A64FX и Intel Cascade Lake второго поколения.

- В 2020 г. калифорнийской компанией Ampere был выпущен процессор Ampere Altra на базе архитектуры ARMv8, оборудованный 80 ядрами и ориентированный на использование в дата-центрах. Ядра имеют тактовую частоту 3 GHz, каждое из которых имеет 2 векторных устройства ARM NEON для работы с векторами длины 128 бит. Таким образом, пиковая производительность всего процессора составляет 3.8 Tflops (3(freq) × 80(cores) × 4(simd) × 2(fma) × 2(vector units)) на одинарной точности и 1.9 Tflops на двойной. Данный процессор предназначен для таких приложений, как аналитика данных, задачи искусственного интеллекта, работы с базами данных, веб-хостинга и облачных приложений. Специально для задач машинного обучения на аппаратном уровне реализована поддержка форматов данных FP16 (половинная точность) и INT8. Иерархия памяти состоит из приватных для каждого ядра кэша L1 объема 64 KB и L2 объема 1 MB, разделяемого всеми ядрами кэша L3 объема 32 MB, а также 8-канальной DDR4-3200 памяти с максимальной пропускной способностью до 200 GB/s и объемом до 1 TB. Данными процессорами также поддерживается 42-канальный интерконнект PCIe 3.0 [31].

Другие разработки. На сегодняшний день также разрабатываются процессоры на базе архитектур, отличных от x86 и ARM, некоторые из которых будут перечислены далее в разделе. Некоторые из них являются процессорами общего назначения, другие — специализированными (например, нейропроцессорами).

- Эльбрус-8С (и Эльбрус-8СВ) — процессоры российской компании МЦСТ, разработанные в 2018 г. и запущенные в серийное производство в 2020 г. Эльбрус-8СВ является сильно модифицированной и более высокопроизводительной версией процессора Эльбрус-8С. Эльбрус-8СВ оборудован 8 ядрами

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

с тактовой частотой 1.3 GHz и является VLIW-процессором, что позволяет эффективно задействовать параллелизм на уровне инструкций. Планирование исполнения команд отводится компилятору, благодаря чему процессор может выполнять на каждом ядре за один машинный такт до 50 операций в векторном режиме (над упакованными 32-разрядными данными). За счет исключения из конструкции VLIW-процессора ряда блоков, выполняющих планирование операций, значительно сокращается потребляемая мощность (для Эльбрус-8СВ это 75–90 W). Это существенно меньше аналогичных решений на базе x86 архитектуры и позволяет повысить соотношение производительности и потребляемой мощности серверных систем. Пиковая производительность процессора Эльбрус-8СВ составляет 520 Gflops (1.3(freq) × 8(cores) × 50(ops_per_cycle)) на одинарной точности и 260 Gflops на двойной. Техпроцесс — 28 nm. Иерархия памяти состоит из приватных кэшей L1 и L2 объема 64 KB и 512 KB на каждое ядро соответственно, а также разделяемого всеми ядрами кэша L3 объема 16 MB. В качестве оперативной памяти используется 4-канальная DDR3-1600 с пиковой пропускной способностью до 51 GB/s и максимальным объемом 64 GB. Посредством PCIe каналов межпроцессорного обмена с пропускной способностью 8 GB/s возможно объединение до четырех процессоров Эльбрус-8СВ в ccNUMA систему с общей когерентной памятью [32].

- Sunway. Национальный центр по проектированию высокопроизводительных интегральных микросхем в Китае в 2016 г. разработал процессоры четвертого поколения Sunway SW26010, установленные в систему Sunway TaihuLight (первое место в Top500 с 2016 по 2018 г.). Данные процессоры имеют 260 ядер, работающих с тактовой частотой 1.45 GHz. Sunway SW26010 имеет RISC-архитектуру с поддержкой SIMD-инструкций длины 256 бит и внеочередным исполнением команд. Таким образом, пиковая производительность процессора составляет 3 Tflops (1.45(freq) × 260(cores) × 8(simd)) на одинарной точности и 1.5 Tflops на двойной. Техпроцесс — 65 nm. Вычислительные ядра процессора сгруппированы в 4 кластера (Computing Processing Element, CPE) по 64 ядра, соединенные решеткой-массивом 8 × 8. Дополнительно к каждому CPE установлено одно ядро общего назначения — процессорный элемент управления (Management Processing Element, MPE). Ядра MPE, занимая небольшую часть площади процессора и потребляя малую долю электроэнергии, предназначены для повышения общей производительности при выполнении последовательных скалярных вычислений, в то время как в CPE построена упрощенная микроархитектура, позволяющая получить высокую производительность для параллельных вычислений с плавающей точкой. Для программиста независимость MPE от CPE означает, что можно разрабатывать программы для выполнения как на MPE, так и на кластере CPE.

  Иерархия памяти разделяется между MPE и CPE, а когерентность кэша достигается только между MPE. В MPE имеется двухуровневый кэш: кэш L1 объема 32 KB на один MPE и общий для всех MPE кэш L2 объема 256 KB (на весь процессор). В каждом CPE имеется кэш L1 объема 16 KB и сверхоперативная память SPM (scratch pad memory) объема 64 KB (на один CPE). Кроме того, в кластере CPE предусмотрен еще общий кэш второго уровня для команд. В процессоры Sunway SW26010 может быть установлена DDR3 память объема до 32 GB с пропускной способностью до 136 GB/s [33].

- SPARC64 — семейство микропроцессоров, разрабатываемых компанией HAL Computer Systems и производимых компанией Fujitsu, которые являются предшественником описанных ранее процессоров A64FX. Наиболее современной модификацией на момент написания данного обзора являются процессоры SPARC64 XIfx, установленные в систему FUJITSU PRIMEHPC FX100 (27 место в Top500 за 2015 г.). Процессоры SPARC64 XIfx имеют 34 ядра, 32 из которых вычислительные, а 2 — вспомогательные, используемые для работы операционной системы, асинхронных MPI коммуникаций и других системных служб. Ядра работают на тактовой частоте 2.2 GHz и имеют векторные устройства, работающие с векторами длины 256 бит. Таким образом, пиковая производительность процессора SPARC64 XIfx составляет 1.1 Tflops (2.2(freq) × 32(cores) × 8(sinmd) × 2(fma)) на одинарной точности и 0.55 Tflops на двойной. Техпроцесс SPARC64 XIfx — 40 nm. Ядра процессора разбиты на две группы, каждая из которых состоит из 16 вычислительных ядер и 1 вспомогательного ядра, которые совместно используют общий кэш L2 объема 12 MB (итого 24 MB на весь процессор). В процессор устанавливается специализированная HMC (Hybrid Memory Cube) память в виде восьми модулей по 4 GB каждый (суммарно 32 GB памяти на процессор). HMC память использует

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

трехмерную микросборку из нескольких (от 4 до 8) чипов DRAM-памяти, выполненную при помощи технологии сквозных межкремниевых соединений, что значительно повышает ее пропускную способность. Для процессора SPARC64 XIfx пропускная способность памяти составляет 480 GB/s (суммарная на чтение и запись). Дополнительно обе группы ядер соединяются с одним PCIe 3.0 контроллером и одним Tofu2 контроллером (интерконнект между процессорными узлами) [34].

- Нейропроцессор Alchip MN-Core был разработан японской компанией Preferred Networks в 2020 г. Данный нейропроцессор предназначен для задач глубинного обучения и оптимизирован для работы с плотными матрицами, часто встречаемыми в данном классе задач. Несмотря на то что базовым режимом работы MN-Core является половинная точность, в нем доступны вычисления с одинарной и двойной точностью, конечно, достигаемые ценой совместной работы нескольких вычислительных блоков и пропорционального снижения производительности. Пиковая производительность данного нейропроцессора составляет 524 Tflops на половинной точности, 131 Tflops на одинарной и 32.8 Tflops на двойной. Техпроцесс MN-Core — 12 nm. Архитектура MN-Core основана на матрично-арифметических блоках (MAU), которые используют принципы векторной обработки данных для работы с плотными матрицами [35]. MN-Core имеет 32 GB оперативной памяти и устанавливается в систему как сопроцессор на основе интерконнекта PCIe 3.0. Данный нейропроцессор имеет очень высокий показатель энергоэффективности — 1 Tflops/W при вычислениях с половинной точностью, благодаря чему на основе данного процессора был разработан суперкомпьютер MN-3, в 2020 г. занявший первую позицию в списке Green500 самых энергоэффективных систем мира.

- Cerebras Wafer Scale Engine — нейропроцессор компании Cerebras, представленный в 2020 г. (запуск в производство планируется в 2021), который содержит рекордные 1.2 трлн транзисторов. Размеры данной микросхемы составляют 203×229 mm, в нее входят 400000 ядер, 18 GB памяти с пиковой пропускной способностью до 9 PB/s, а энергопотребление составляет 15 kW. Техпроцесс Cerebras Wafer Scale Engine — 7 nm. Ядра данного процессора связаны ячеистой сетью с общей пропускной способностью 100 Pb/s. Иерархия памяти одноуровневая, кэш-памяти нет [36].

§ 3.1.3. Ускорители.

В данном разделе будут рассмотрены ускорители, часто применяющиеся при построении современных высокопроизводительных систем. Основной отличительной особенностью таких систем является установка в систему ускорителей в качестве сопроцессоров с использованием некоторого интерконнекта, вследствие чего модель использования ускорителей существенно отличается от таковой для центральных процессоров, а термины “ускоритель” и “сопроцессор” обычно взаимозаменяемы. Обычно наиболее вычислительно-затратные участки программ (подходящие для реализации на ускорителе данного типа) выгружаются на ускоритель, в то время как остальная часть программы выполняется на основном процессоре. Кроме того, ускоритель и центральный процессор обычно имеют различные типы памяти и, как следствие, различные адресные пространства, что требует явных пересылок данных между ними. Двумя наиболее известными представителями являются графические ускорители NVIDIA и AMD, позволяющие ускорять различные классы массивно-параллельных вычислительных задач. Кроме того, значительный интерес представляют векторные процессоры NEC SX-Aurora TSUBASA, которые также устанавливаются в систему в виде сопроцессора, и другие специализированные ускорители, предназначенные для решения еще более узкого класса задач.

NVIDIA. Графические ускорители (GPU, Graphics Processing Unit) NVIDIA широко применяются в современных высокопроизводительных вычислениях: в 2020 г. 6 из 10 первых систем списка Top500 использовали решения на их основе. Графические процессоры NVIDIA существенно отличаются от традиционных многоядерных центральных процессоров. Они устанавливаются в систему как сопроцессоры, подключаясь через шину PCI или NVLINK, и имеют тысячи легковесных вычислительных ядер с низкой тактовой частотой. Кроме того, GPU оборудованы быстрой HBM памятью с пропускной способностью около 1 TB/s. Вычислительные ядра графического ускорителя группируются в так называемые варпы — вычислительные единицы, в каждый момент времени выполняющие одну и ту же инструкцию над различными данными, что является одним из вариантов векторной обработки данных. В результате графические процессоры хорошо подходят для ускорения различных классов массивно-параллельных задач, в которых присутствует параллелизм по данным, притом как интенсивно использующих подсистему памяти

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

Таблица 1. Сравнение основных аппаратных характеристик ускорителей P100, V100 и A100

Table 1. Comparision of the main hardware characteristics of the P100, V100 and A100 accelerators

Характеристика Characteristic	P100	V100	A100
Число CUDA-ядер Number of CUDA cores	3584 (FP32) 1792 (FP64)	5376 (FP32) 2560 (FP64)	6912 (FP32) 3456 (FP32)
Число тензорных ядер Number of tensor cores	—	672	432
Тактовая частота, GHz Clock frequency, GHz	1.48	1.53	1.41
Производительность, одинарная точность, Tflops Performance, single precision,Tflops	10.6	15.7	19.5
Производительность, двойная точность, Tflops Performance, double precision,Tflops	5.3	7.8	9.7
Производительность, половинная точность, Tflops Performance, half precision,Tflops	21	31 (при использовании CUDA ядер / using CUDA cores) 125 (при использовании тензорных ядер / using tensor cores)	78 (при использовании CUDA ядер / using CUDA cores) 312 (при использовании тензорных ядер / using tensor cores)
Тип интерконнекта Interconnect type	PCI или NVLINK PCI or NVLINK	PCI или NVLINK 2.0 PCI or NVLINK 2.0	PCI или NVLINK 3.0 PCI or NVLINK 3.0
Пропускная способность интерконнекта, GB/s Interconnect bandwidth, GB/s	160 (NVLINK), 32 (PCI)	300 (NVLINK), 32 (PCI)	600 (NVLINK), 32 (PCI)
Объем памяти, GB Memory size, GB	$\leqslant 16$	$\leqslant 32$	$\leqslant 40$
Пропускная способность памяти, GB/s Memory bandwidth, GB/s	$\leqslant 720$	$\leqslant 900$	$\leqslant 1555$
Размер L1 кэша, KB L1 cache size, KB	64 KB	$\leqslant 96$ (конфигурируемый) (configurable)	$\leqslant 164$ (конфигурируемый) (configurable)
Размер L2 кэша, MB L2 cache size, MB	4	6	40

(за счет наличия быстрой HBM памяти), так и вычислительные устройства (за счет наличия большого числа ядер). Графические ускорители NVIDIA относятся к одному из трех семейств: Tesla, нацеленные на высокопроизводительные вычисления и машинное обучение, Quadro, нацеленные на работу с профессиональной графикой, и GeForce, использующиеся в основном в игровой индустрии. Далее в обзоре будут рассмотрены только ускорители семейства Tesla.

Наиболее часто используемыми представителями семейства Tesla являются P100, V100 и A100 трех последних архитектур Pascal, Volta и Ampere соответственно. Tesla P100 была представлена компанией в 2016 г., в то время как Tesla V100 — в 2017 г., а A100 — в 2020 г.

В табл. 1 представлено сравнение аппаратных характеристик графических ускорителей P100, V100 и A100. Общими тенденциями являются: увеличение в более новых поколениях числа ядер, производительности, объема и пропускной способности памяти, размера кэш-памяти, скорости интерконнекта, поддержка более производительной работы с арифметикой меньшей точности, а также наличие дополнительных программных и аппаратных нововведений. К примеру, в архитектуре Pascal (P100) была добавлена поддержка Unified памяти (единое адресное пространство памяти для CPU и GPU), встроенная

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

память с коррекцией ошибок (без накладных расходов), аппаратная поддержка работы с арифметикой половинной точности, а также повышенная производительность атомарных инструкций. В архитектуре Volta (V100) были добавлены такие нововведения, как обновленный интерфейс работы с кэш-памятью L1 и разделяемой памятью, а также тензорные ядра. В архитектуре Ampere (A100) была реализована технология Multi-Instance GPU (MIG), которая позволяет безопасно разделить графический процессор A100 Tensor Core на семь отдельных экземпляров, предоставляя нескольким пользователям возможность коллективно использовать один ускоритель. Кроме того, были добавлены поддержка API для явной загрузки определенных данных в кэш L2 (префетчинг), возможность асинхронной загрузки данных из глобальной в разделяемую память (в обход регистров), а также усовершенствованный механизм запуска расчетов на GPU.

AMD Instinct. В конце 2020 г. компания AMD представила AMD Instinct MI100 accelerator GPU с производительностью 11.5 Tflops при вычислениях с двойной точностью, что делает его быстрейшим в мире GPU для научных расчетов (для сравнения, NVIDIA A100 имеет производительность 9.7 Tflops при вычислениях с двойной точностью). Однако на других типах вычислений MI100 показывает меньшую в сравнении с A100 производительность: 23.1 Tflops при вычислениях с одинарной точностью и 184.6 Tflops при вычислениях с половинной. Для построения вычислительных устройств (Compute Units) MI100 использует Matrix Core Engines — гибкий процессор, который может выполнять новый тип инструкций wavefront, которые сгруппированы вместе подобно Matrix Fused Multiply-Add или MFMA. Данные инструкции могут работать с 4-битными целыми числами (INT4), 8-битными целыми числами (INT8), 16-битными числами с плавающей запятой половинной точности (FP16), 16-битными BFloat16 (bf16) и 32-битными числами с плавающей запятой одинарной точности (FP32).

Так же как и NVIDIA GPU, ускорители MI100 используют HBM2 память общим объемом до 32 GB и пиковой пропускной способностью до 1.23 TB/s. Для подключения используется стандарт PCI Express 4.0. Несколько GPU могут быть объединены в одну (до четырех карт) при помощи интерфейса Infinity fabric с пропускной способностью 267 GB/s, по аналогии с NVLINK. При этом компания AMD решила разделить архитектуру GPU на две линейки: RDNA для игровых видеокарт Radeon и CDNA для ускорителей Instinct [37]. Вследствие этого все компоненты, отвечающие за рендеринг, были убраны из MI100, в том числе функции растеризации, тесселяции, графические кэши, движок дисплея и т.д. Кэш последнего уровня в иерархии памяти MI100 — L2 объема 8 MB и пропускной способностью 6 TB/s.

NEC Vector Engine Processor. NEC SX-Aurora TSUBASA — новейшая суперкомпьютерная архитектура семейства SX [38, 39], анонсированная и выпущенная японской компанией NEC в 2018 г. В отличие от своих предшественников, архитектура системы SX-Aurora TSUBASA состоит из векторного устройства (vector engine, VE), являющегося основным векторным процессором, а также векторного хоста (vector host, VH), являющегося процессором архитектуры x86, соединенных через шину PCI. Таким образом, система SX-Aurora TSUBASA так же построена на основе модели “процессор + ускоритель”. Однако, в отличие от архитектуры NVIDIA GPU, VE используется в качестве основного процессора для выполнения приложений, в то время как VH используется в качестве сопроцессора для выполнения функций базовой операционной системы, а также части последовательных (невекторизуемых) вычислений, которые могут быть явно выгружены пользователем с векторного устройства.

Векторное устройство (VE) имеет восемь мощных векторных ядер, работающих с тактовой частотой 1.6 GHz, благодаря чему пиковая производительность всего VE достигает 4.3 Tflops. Каждое векторное ядро SX-Aurora состоит из блока скалярной обработки (scalar processing unit, SPU), блока векторной обработки (vector processing unit, VPU) и подсистемы памяти. Большинство вычислений выполняется при помощи VPU, в то время как SPU обеспечивают функциональность типичного центрального процессора. Поскольку SX-Aurora является не просто ускорителем, а самодостаточным процессором, SPU предназначен для обеспечения относительно высокой производительности при скалярных вычислениях. VPU каждого векторного ядра имеет свой собственный относительно простой конвейер команд, предназначенный для декодирования и переупорядочивания векторных команд, поступающих из SPU. Декодированные инструкции выполняются на векторно-параллельных конвейерах (vector parallel pipeline, VPP). Для хранения результатов промежуточных вычислений каждое векторное ядро оснащено 64 векторными регистрами с общей емкостью регистра, равной 128 KB, причем каждый регистр предназначен для хранения векторов из 256 элементов двойной точности.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

Векторная обработка внутри каждого из ядер VE основана на использовании 32 идентичных параллельных конвейеров, обрабатывающих расположенные на регистрах векторы частями по 32 элемента в соответствии с моделью SIMD. Таким образом, одна инструкция, оперирующая с векторами из 256 элементов, будет обработана за 8 тактов процессора при условии, что не происходило остановок, вызванных командами с высокой латентностью (например, загрузки данных из подсистемы памяти). Каждый VPP имеет 3 блока FMA (Fused Multiply-Add), 2 арифметико-логических блока (arithmetic-logic unit, ALU) и 1 блок, предназначенный для обработки команд с высокой латентностью (извлечение квадратного корня, деления и другие), а также блок работы с подсистемой памяти. В зависимости от структуры программы, необходимые данные перенаправляются между вычислительными блоками VPP, образуя векторный конвейер. Подсистема памяти векторного устройства состоит из шести модулей HBM-памяти, обеспечивающих пропускную способность памяти, равную 1.22 TB/s и кэша последнего уровня (LLC) объема 16 MB с пропускной способностью около 3 TB/s. Важно отметить, что кэши L1 и L2 также присутствуют в Vector Engine, однако используются только для работы SPU.

Matrix-2000. Процессоры Matrix-2000 в 2017 г. были разработаны специально для китайского суперкомпьютера Tianhe-2 (первое место в Top500 в июне 2013) с целью заменить устаревающие ускорители Knights Corner (KNC) компании Intel. Аналогично KNC, Matrix-2000 устанавливаются в систему в виде ускорителей на основе PCI Express 3.0. Matrix-2000 включает в себя 128 RISC-ядер, каждое из которых имеет тактовую частоту 1.2 GHz. Каждое ядро представляет собой компьютер с набором коротких команд (RISC), имеющий 8–12-этапный конвейер. Дополнительно ядра Matrix-200 включают в себя набор 256-битных векторных инструкций, которые обрабатываются двумя блоками векторной обработки (VPU). Таким образом, пиковая производительность процессора составляет 4.9 Tflops (1.2(freq) $\times$ 128(cores) $\times$ 8(simd) $\times$ 2(vector units) $\times$ 2(fma)) на одинарной точности и 2.4 Tflops на двойной. Ядра Matrix-2000 разбиты на четыре суперузла (по 32 ядра на суперузел), которые связаны между собой через коммуникационную сеть Fast Interconnect Transport (каждый с каждым) и при этом имеют когерентную кэш-память. Каждый из суперузлов, в свою очередь, состоит из 8 кластеров, а каждый кластер из коммутатора, блока управления каталогом (DCU), 4 ядер ЦП и общего кэша. Каждый из суперузлов соединяется с двумя контроллерами DDR4 памяти с максимальной пропускной способностью 153 GB/s [40].

Google TPU. В 2016 г. компания Google анонсировала специализированные тензорные процессоры Google (Google Tensor Processing Unit, TPU), разработанные специально для ускорения задач глубокого обучения. Поскольку основной задачей TPU является обработка плотных матриц (занимающая почти все время машинного обучения), Google TPU — это интегральная схема специального назначения (ASIC), в которой размещены тысячи множителей и сумматоров, соединенных физически в одну большую физическую матрицу (архитектура конвейерного массива). Современные TPU состоят из четырех 2-ядерных чипов в одном TPU. Внутри чипа работают скалярные, векторные и матричные блоки, выполняющие операции соответствующего типа [41]. Каждый матричный блок способен выполнять 16 тысяч FMA-операций за такт с пониженной точностью bfloat16. Bfloat16 — это специализированное 16-битное представление чисел с плавающей точкой, которое обеспечивает более быстрое обучение в сравнение с представлением половинной точности IEEE [42]. За счет того, что TPU лишены характерных для CPU и GPU компонент — регистров, кэшей, поддержки внеочередного исполнения и др., которые необходимы для вычислений общего назначения, TPU в 30–80 раз более энергоэффективны в сравнении с современными центральными процессорами и графическими ускорителями [43]. На сегодняшний день двумя последними поколениями TPU являются TPUv3 (2018 г.) и TPUv4 (2020 г.), которые имеют сравнимые с NVIDIA V100 и A100 GPU производительность при вычислениях с половинной точностью: 180 Tflops (TPUv3) против 120 Tflops (V100 GPU). Данные поколения отличаются от предшественников использованием HBM памяти (объема 32 GB для четвертого поколения). TPU могут объединяться в группы до четырех устройств при помощи скоростного интерконнекта, в то время как подключение к центральным процессорам осуществляется на основе шины PCIe.

NVIDIA DPU. После покупки компании Mellanox NVIDIA интегрировала ряд их решений в NVIDIA Data Processing Units (DPUs) [44, 45]. NVIDIA DPU — это система на кристалле, сочетающая в себе:

1. Высокопроизводительный многоядерный центральный процессор архитектуры ARM, тесно связанный с другими компонентами системы на кристалле;

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

2. Набор графических ускорителей, способных повышать производительность решения задач ИИ и машинного обучения, безопасности, телекоммуникаций и хранения, и др.;
3. Высокопроизводительный сетевой интерфейс, способный анализировать, обрабатывать и эффективно передавать данные на используемые в DPU центральный и графические процессоры.

Основная область использования DPU — дата-центры, где DPU выполняют задачи управления сетью, памятью и технологиями безопасности, на которые приходятся около 30% вычислительной производительности дата-центра [46], для чего используются специальные программируемые сетевые интерфейсы и интерфейсы хранения. Кроме того, DPU также разгружает, ускоряет и изолирует все важные службы безопасности центра обработки данных, что включает в себя поддержку межсетевых экранов следующего поколения, микросегментацию, встроенное шифрование данных, и др. В 2021 г. NVIDIA планирует выпустить на рынок две модели DPU: BlueField 2 и BlueField 2X. Данные DPU оснащены 8-ядерным процессором на основе архитектуры ARM A72 и сетевым интерфейсом ConnectX-6 Dx network, который может подключаться к 200 Gb/s Ethernet и Infiniband [47]. В BlueField 2X дополнительно добавляются графические ускорители Ampere.

3.2. Вычислительные узлы. Подход к построению подсистемы памяти является важным способом классификации современных суперкомпьютерных систем. Часто выделяют три класса систем: UMA, NUMA, ccNUMA. Данная классификация может рассматриваться в качестве дополнительной детализации для классификации Флинна, в которой большинство современных систем относятся к классу MIMD и, как следствие, требуют дополнительного разделения [48].

В UMA (Uniform Memory Access) системах все процессоры имеют доступ к любой области памяти на основе операций load/store в рамках единого адресного пространства, причем время доступа любого процессора к любой ячейке памяти одинаково. Обычно UMA системы строятся на основе использования общей шины памяти всеми процессорами, причем каждый процессор может использовать свою собственную кэш-память. Принципиальный недостаток UMA систем — их плохая масштабируемость, поскольку при увеличении числа процессоров увеличивается как длина шины (а значит, и латентность), так и пропускная способность, доступная каждому из процессоров [49]. Таким образом, общая шина становится узким местом, значительно ограничивая производительность приложений, интенсивно использующих память.

В результате часто используются NUMA (Non Uniform Memory Access) системы, в которых все процессоры имеют доступ ко всем частям основной памяти, однако время (латентность) доступа различных процессоров к памяти различается в зависимости от того, к какой области осуществляется доступ. Обычно у каждого из процессоров есть своя локальная память, доступ к которой может осуществляться значительно быстрее, чем к памяти других процессоров. Разница в латентности и пропускной способности между локальной и удаленной памятью может составлять, к примеру, от 200% до 700%, и это существенно влияет на производительность разрабатываемых для NUMA систем приложений.

Кроме того, как было показано в предыдущих разделах, почти все современные процессоры оборудованы достаточно сложной иерархией кэш-памяти. Для NUMA систем это означает, что необходимо поддерживать когерентность (согласованность) данных, расположенных в кэш-памяти различных процессоров, что обычно реализуется на аппаратном уровне межпроцессорных обменов между контроллерами кэш-памяти, а также при помощи специальных протоколов, в том числе MESIF [50]. В результате все коммерчески поставляемые NUMA-компьютеры классифицируются как кэш-когерентные системы (ccNUMA). При этом для построения ccNUMA систем используются различные интерконнекты, например Intel QuickPath (до Skylake) или Intel Ultra Path (после Skylake). Основная проблема ccNUMA систем — значительная сложность разработки программ для них, поскольку на программиста ложится обязанность обеспечения локальной работы с данными. При этом необходимо учитывать как неоднородность доступа к памяти, так и особенности работы механизмов согласования кэш-памяти. К примеру, в случае, когда различные процессоры обращаются к одному и тому же блоку памяти или одной переменной, для данных систем наблюдается значительное снижение производительности, обуславливаемое накладными расходами, направленными на поддержание когерентности кэш-памяти. Вследствие этого на сегодняшний день лишь часть из существующего программного обеспечения способна эффективно работать на NUMA системах.

3.3. Коммуникационная основа. Еще один важный компонент суперкомпьютерных систем (помимо подробно описанной ранее процессорной основы) — это коммуникационная основа (или интерконнект), обеспечивающая связь вычислительных узлов (различных процессоров) между собой. В начале

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

данного раздела будут рассмотрены примеры наиболее часто использующихся топологий современных коммуникационных сетей, после чего будут приведены основные характеристики для наиболее распространенных интерконнектов, таких как Infiniband, Ethernet и др.

§ 3.3.1. Особенности построения современных коммуникационных сетей.

Вначале дадим краткие определения основных характеристик коммуникационных сетей, которые будут часто использоваться далее в разделе: топологии, латентности, пропускной способности. Топология — это конфигурация графа сети, определяющего взаимное расположение узлов и каналов связи. Примеры различных топологий включают в себя кольцо, двумерную решетку, гиперкуб и другие, причем различные топологии могут быть более или менее эффективны для различных коммуникационных шаблонов, встречаемых в реальных суперкомпьютерных программах. Двумя важными характеристиками быстродействия сети являются латентность (latency) и пропускная способность (bandwidth). Под пропускной способностью сети обычно понимается количество информации, передаваемой между узлами сети в единицу времени, вследствие чего значения пропускной способности обычно выражаются в байтах в секунду. Пропускная способность бывает однонаправленная (uni-directional), равная максимальной скорости, с которой процесс на одном узле может передавать данные другому процессу на другом узле, и двунаправленная (bi-directional), равная максимальной скорости, с которой два процесса могут одновременно обмениваться данными по сети при одновременном приеме и получении.

Латентностью (задержкой) называется время, затрачиваемое программным обеспечением и устройствами сети на передачу пакета нулевой длины между двумя узлами сети. Полная латентность складывается из программной и аппаратной составляющих, а выражается латентность, обычно, в микросекундах. При этом латентность между различными парами узлов в топологии (например, соседними и удаленными) может существенно различаться, отчего возникают понятия минимальной и максимальной латентности.

§ 3.3.2. Топологии коммуникационных сетей.

Различных топологий коммуникационных сетей существует очень много. В данном обзоре будут рассмотрены лишь несколько из наиболее часто встречаемых топологий: “многомерный тор”, “толстое дерево”, “бабочка” и “стрекоза” с целью познакомить читателя обзора с основными принципами построения коммуникационных сетей, а также различными преимуществами и недостатками каждой из рассмотренных топологий.

XD-тор. В топологии N-мерный тор вычислительные узлы суперкомпьютера организованы в N-мерную сетку: каждый из узлов соединен со своими ближайшими соседями по каждому из измерений, причем граничные узлы также соединены между собой. Примеры топологии одномерного и двумерного тора приведены на рис. 1. Для данной топологии используются различные размерности. К примеру, компания Fujitsu использует модель 6-мерного тора под названием “Тофу” в суперкомпьютерах K Computer и Fugaku (самым производительным в списке за 2020 г.). Важные достоинства данной топологии — относительная дешевизна и эффективная поддержка коммуникативных шаблонов, встречающихся в сеточных задачах, однако данная сеть имеет достаточно высокую латентность для удаленных друг от друга узлов.

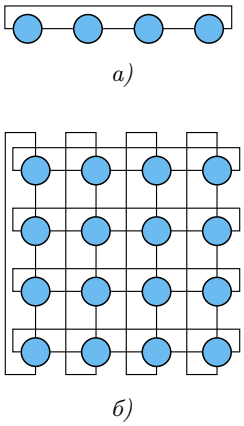

Рис. 1. Топология XD-тор [51]: *а)* одномерный тор; *б)* двумерный тор

Fig. 1. XD-torus topology [51]: *а)* one-dimensional torus; *б)* two-dimensional torus

Fat Tree (folded Clos). Сеть с топологией “толстого дерева” (fat tree) представляет собой дерево, листьями которого являются вычислительных узлы суперкомпьютера, а внутренними вершинами — коммутаторы сети. Существует также более общий вариант данной топологии — Сеть Клоза (Clos network) [53]. Толстое дерево является односторонней версией сети Клоза, в которой объединены входные и выходные коммутаторы, в результате чего именно под названием folded Clos эта топология часто фигурирует при описании сетей современных дата-центров. В изначальных версиях топологии толстого дерева пропускная способность связей между коммутаторами более высоких уровней была больше (иными словами, связи “толще”), отсюда и название данной топологии (рис. 2). Зачастую используется удвоение пропускной способности (или числа связей)

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

на каждом уровне. Чтобы передать сообщение с одного узла на другой, нужно использовать коммутаторы различных уровней. Сначала сообщение передается на несколько уровней вверх из первого процессора, пока не достигнет наименьшего общего предка данных вершин, потом на несколько уровней вниз, до второго процессора. Таким образом, сообщение проходит максимум $2\log_2(N)$ вершин. Суперкомпьютеры, использующие сеть на основе “толстого дерева”, крайне распространены: например, два самых быстрых по состоянию на декабрь 2018 г. суперкомпьютера Summit и Sierra, а также Tianhe-2. Топология “толстое дерево” имеет ряд преимуществ, такие как низкая латентность и эффективная поддержка большего числа коммуникационных профилей для различных приложений [52].

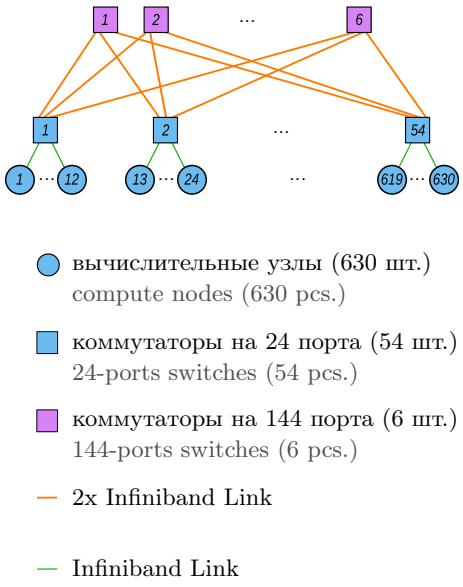

Рис. 2. Топология толстого дерева на примере суперкомпьютера СКИФ МГУ “Чебышев”

Fig. 2. Fat Tree topology using the SKIF MSU “Chebyshev” supercomputer

Топологии Butterfly и Flattened Butterfly. В топологии Butterfly число коммутаторов существенно превышает число используемых вычислительных узлов. Данная топология имеет несколько уровней (рис. 3). На последнем уровне соединение идет между соседними столбцами, на предпоследнем между столбцами на расстоянии 2, на следующем на расстоянии 4 и т.д. Таким образом, если сеть объединяет P вычислительных узлов, то количество коммутаторов должно быть равно $P(\log_2(P)+1)$. Вычислительные узлы соединяются непосредственно с коммутаторами нулевого уровня.

Топология Flattened Butterfly [55] была разработана с целью снизить стоимость топологии Butterfly. Идея топологии Flattened Butterfly состоит в том, чтобы объединить коммутаторы, находящиеся в одном столбце топологии Butterfly: коммутаторы R0, R1, R2 и R3 (рис. 4) объединены в один коммутатор R0’. В результате устраняются все внешние связи между коммутаторами одного столбца в топологии Butterfly, в то время как все остальные связи остаются. Таким образом, коммутаторы соединены каналами в $n' = n-1$ измерениях, соответствующих $n-1$ столбцу межрангового соединения в исходной топологии Butterfly. Например, для одномерной топологии Flattened Butterfly у каждого из N коммутаторов будет $N-1$ связей.

Топология Flattened Butterfly значительно уменьшает стоимость по сравнению с обычным Butterfly, а также позволяет быстрее строить нужные маршруты в обмене сообщениями по сравнению с топологией толстого дерева [55].

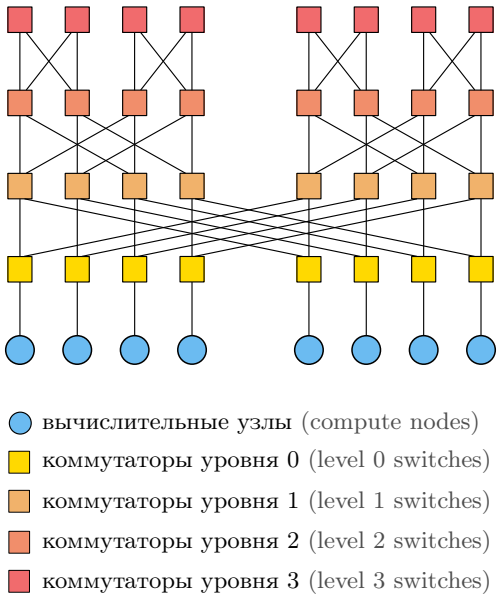

Рис. 3. Топология Butterfly [54]

Fig. 3. Butterfly topology [54]

Dragonfly. Топология Dragonfly является иерархической топологией. Несколько групп соединены вместе, используя связи “все ко всем”, т.е. каждая группа имеет по крайней мере одну связь непосредственно с другой группой. Топология внутри каждой группы может быть любой, например тот же Flattened Butterfly.

§ 3.3.3. Характеристики наиболее распространенных коммуникационных сетей.

Infiniband. По состоянию на 2019 г. Infiniband являлся наиболее популярной сетью для суперкомпьютеров, используется более чем в 30% систем из списка Top500. Контроллеры Infiniband и сетевые коммутаторы производятся компаниями Mellanox и Intel. Подобно многим современным шинам, в Infiniband используются дифференциальные пары для передачи последовательных сигналов. Две пары вместе состав-

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

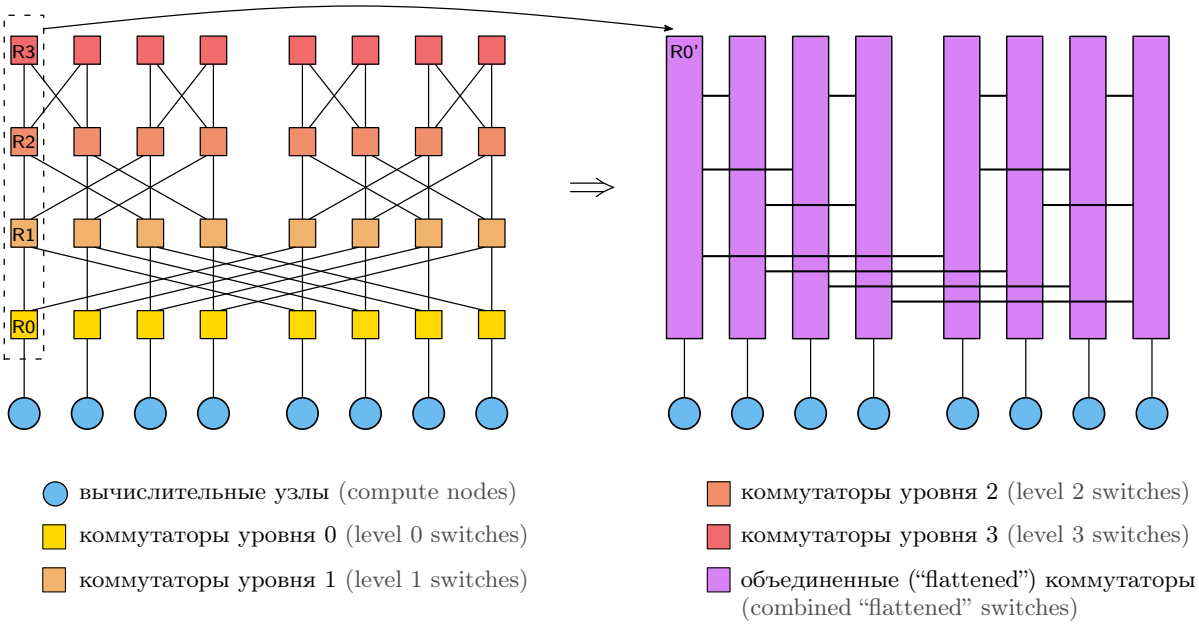

Рис. 4. Пример преобразования топологии Butterfly во Flattened Butterfly [55]

Fig. 4. An example of conversion from the Butterfly to Flattened Butterfly topology [55]

ляют одну базовую двунаправленную последовательную шину (lane), обозначаемую 1x. Порты Infiniband состоят из одной шины или агрегированных групп 4x или 12x базовых двунаправленных шин. Чаще всего применяются порты 4x. Два последних поколения Infiniband — QDR, FDR, EDR (2014 г.) и HDR (2017 г.). Теоретическая пиковая пропускная способность поколения EDR — от 100 до 300 Gb/s, поколения HDR — от 200 до 600 Gb/s для 12x, латентность — приблизительно 100–500 ns.

Intel Omni-Path. Intel Omni-Path — высокопроизводительная коммуникационная архитектура от компании Intel, представленная в 2015 г. и предназначенная для высокопроизводительных вычислительных кластеров. Первая реализация Omni-Path (100 Series) с пропускной способностью 100 Gb/s, согласно заявлениям Intel, обеспечивает меньший уровень задержки и более высокую практическую пропускную способность в сравнении с сетью Infiniband поколения EDR. Однако летом 2019 г. компания Intel отменила планы по выпуску устройств на базе технологии Omni-Path второго поколения (с планируемой пропускной способностью 200 Gb/s). Латентность сети Omni-Path первого поколения сопоставима с латентностью Infiniband поколения EDR (в среднем на 11% ниже [56]).

Ethernet. Ethernet — семейство технологий пакетной передачи данных между устройствами для компьютерных и промышленных сетей. Два наиболее часто используемых в HPC стандарта Ethernet — 10G Ethernet и 40/100G Ethernet, которые имеют пропускную способность в 100 Gb/s, что сопоставимо с Infiniband. Решения на основе стандарта Ethernet разрабатываются рядом компаний, например Cisco. Cisco Ethernet, включая Cisco Nexus 9000 и 3000 семейств, разработаны для обеспечения производительности и малой задержки, необходимых для поддержки HPC. Коммутаторы Nexus обеспечивают задержку менее микросекунды; например, коммутатор Nexus 3500 обеспечивает задержку менее 150 ns [57].

Tofu (K Computer). Tofu — это высокопроизводительная коммуникационная сеть для массивно-параллельных компьютеров, изначально разработанная компанией Fujitsu для создания K Computer, состоящего более чем 80000 узлов (поколение Tofu 1). На сегодняшний день поколение Tofu D [58] используется в суперкомпьютере Fugaku, являющимся первым в списке Top500 2020 г. Топология сети любого поколения Tofu представляет собой хорошо масштабируемый шестимерный тор [59]. Пропускная способность канала сети Tofu 1 составляет 40 Gb/s в каждом направлении, Tofu D — 54.4 Gb/s, однако в Tofu D число линий было увеличено с 4 до 6, благодаря чему суммарная полоса пропускания между узлами увеличилась в два раза с 160 Gb/s до 320 Gb/s. Латентность передачи сообщения между соседними узлами в сети Tofu 1 составляет от 910 до 1150 ns, а для сети Tofu D — от 490 до 540 ns (измерялась передача пакета длиной

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

8 байт между узлами одной стойки). Каждый узел может производить пересылки данных одновременно в четырех направлениях для Tofu 1 и шести для Tofu D. Сетевые интерфейсы и коммутатор Tofu интегрированы в разработанный компанией Fujitsu чип под названием InterConnect Controller (ICC).

Cray Aries. Сеть Cray Aries [60] — это высокоскоростное межсоединение, используемое в суперкомпьютерах Cray Theta, занимающих на момент 2020 г. 39-ю позицию в рейтинге Top500. Cray Aries использует трехуровневую топологию Dragonfly: первый уровень — это небольшое количество узлов, подключенных к одному коммутатору. Следующий уровень — это соединение двух стоек узлов в полную сетку. В последнем уровне соединяются все группы из двух стоек. Таким образом, суперкомпьютеры на основы сети Aries используют архитектуру системы, которая объединяет четыре контроллера сетевого интерфейса (NIC) и 48-портовый коммутатор в одно устройство, которое подключается к четырем вычислительным узлам XC через соединение 16x PCI-Express Gen3. Каждая сетевая карта подключена к двум портам на коммутаторе, в то время как для каналов между коммутаторами доступно 40 портов. Интерфейс PCIe обеспечивает скорость передачи, равную 8 GB/s в каждом направлении, что дает двунаправленную пиковую пропускную способность, равную 16 GB/s. Сетевые каналы имеют пропускную способность до 42 Gb/s в каждом направлении. Протокол маршрутизации Aries использует адаптивную маршрутизацию для выбора наилучшего пути в сети для каждого пакета. Этот метод маршрутизации позволяет избежать перегрузки, но также позволяет пакетам проходить по неминимальным маршрутам.

Латентность данной сети также существенно зависит от взаимного расположения узлов: максимально удаленные друг от друга узлы требуют до 5 переходов (hops) сообщений, в то время как каждый из переходов имеет латентность около 100 ns [61]. Согласно приведенным в работе [62] тестам, реальная латентность данной сети составляет 120–300 ns в зависимости от числа участвующих в обменах узлов и типа MPI операции (в зависимости от размера пакета).

HPE Slingshot. HPE Slingshot — коммуникационная сеть от компании Cray, основанная на стандарте Ethernet. Данная сеть будет использоваться во всех трех экзафлопсных суперкомпьютерах США: Aurora, Frontier и El Captain, описанных в дальнейших разделах. Топология данной сети — Dragonfly, построена на основе 64-портовых коммутаторов ROSETTA, каждый порт с пропускной способностью 200 Gb/s в одном направлении. Такие коммутаторы обеспечивают масштабирование до сотен тысяч узлов, позволяя передавать сообщения между узлами суперкомпьютера не более чем за три перехода [63]. Каждый порт коммутатора имеет пропускную способность 200 Gb/s в каждом направлении. Периферийные порты подключаются к сетевому адаптеру Ethernet или внешним коммутаторам на уровне 100 Gb/s или 200 Gb/s. HPE Slingshot поддерживает адаптивную маршрутизацию и алгоритмы контроля перегрузки, которые динамически отправляют пакеты на основе поступающей в режиме реального времени глобальной информации о загрузке в сети, а также передовые механизмы управления перегрузками. Согласно [64], коммутатор ROSETTA имеет среднюю и медианную латентность, равные 300 и 400 ns соответственно.

Sunway. Sunway Network — интерконнект, разработанный для суперкомпьютера TaihuLight, который с июня 2016 по июнь 2018 г. являлся самым производительным суперкомпьютером в мире. Сеть Sunway состоит из трех разных уровней: верхний — с центральной коммутационной сетью, соединяющей различные суперузлы, средний — сеть, которая полностью соединяет все 256 узлов в каждом суперузле, и нижний, который соединяет вычислительные узлы с другими ресурсами, такими как устройства ввода-вывода. Sunway использует адаптер хост-канала (HCA) и микросхемы коммутатора от компании Mellanox. Пропускная способность сети около 96 Gb/s, в то время как латентность равна 1000 ns [33].

Ангара. Сеть “Ангара” [65] — российская высокоскоростная коммуникационная сеть на базе сверхбольших интегральных схем (СБИС). Используемые СБИС являются разработкой АО “НИЦЭВТ” и выпускаются по технологии 65 nm. Сеть “Ангара” использует топологию “многомерный тор” (возможны одномерные, двумерные, трехмерные и четырехмерные варианты) в силу ориентации на задачи математического моделирования. “Ангара” также поддерживает режим прямого доступа к памяти удаленных узлов RDMA, технологию GPUDirect и все стандартные средства программирования (библиотека MPI, технология OpenMP, библиотека SHMEM, стек протоколов TCP/IP). Коммуникационная сеть “Ангара” совместима с процессорами x86, “Эльбрус”, а также ускорителями GPU и FPGA. Сеть “Ангара” отличается высокой пропускной способностью линков и низкими задержками передачи, которые соответствуют мировому уровню. Так, латентность передачи данных между соседними узлами составляет 130 ns, а пропускная способность — 75 Gb/s на соединение в каждом направлении (каждый узел сети имеет до 8 соединений). На различных тестах и приложениях сеть “Ангара” позволяет достичь таких показателей

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

производительности и масштабируемости, которые не уступают соответствующим характеристикам на вычислительных системах с использованием сети Mellanox Infiniband 4xFDR [66] или превосходят их.

3.4. ПЛИС-технологии. FPGA (Field-Programmable Gate Array) — программируемая логическая интегральная схема (ПЛИС). В отличие от обычных цифровых микросхем, логика работы ПЛИС не определяется при изготовлении, а задается посредством программирования (проектирования). Крайняя гибкость FPGA и их потенциал производительности сделали их привлекательным выбором в широком диапазоне вычислительных областей, от быстрого прототипирования схем до высокопроизводительных вычислений. Увеличение доступности транзисторов на матрице позволило создавать FPGA с большим количеством вычислительных ресурсов и топологий межсоединений, что, в свою очередь, привело к созданию систем с широким спектром вариантов реализации. Раньше, учитывая ограниченную емкость устройств ПЛИС, они использовались, как правило, в качестве интерфейсных логических схем (glue-logic), теперь же возможность конфигурировать эти устройства позволяет инженерам-разработчикам решать множество различных задач.

Современные ПЛИС позволяют реализовать даже очень сложные функции со временем выполнения в один такт. Программируемость FPGA гарантирует, что они могут быть настроены в соответствии с конкретными потребностями приложения без затрат или задержек на разработку специального сопроцессора. FPGA может обеспечить сопроцессорную обработку с широкими возможностями настройки для широкого спектра приложений в одном кристалле. Наличие встроенной памяти в ПЛИС (в том числе стандарта HBM) также дает значительные преимущества в производительности. Кроме того, наличие памяти на кристалле означает, что пропускная способность логики сопроцессора для доступа к памяти не ограничивается количеством контактов ввода-вывода, которые имеет устройство. При этом память тесно связана с логикой алгоритма и снижает потребность во внешней высокоскоростной кэш-памяти. Это, в свою очередь, позволяет избежать проблем с доступом к кэш-памяти и когерентностью, потребляющими много энергии. Использование внутренней памяти также означает, что сопроцессору не требуются дополнительные контакты ввода-вывода для увеличения доступного объема памяти. Усовершенствования архитектуры, увеличенное количество логических ячеек и скорость способствуют увеличению производительности современных FPGA. Например, при повышении тактовой частоты в среднем на 25% для каждого поколения FPGA производительность логических вычислений улучшилась примерно в 92 раза за последнее десятилетие, в то время как стоимость FPGA снизилась на 90% за тот же период времени.

На момент написания данного обзора существуют два крупных производителя FPGA-чипов: Xilinx и Intel, которые контролируют 58% и 42% рынка соответственно [67]. Основатели Xilinx изобрели первый чип FPGA в 1985 г., в то время как Intel пришла на рынок недавно — в 2015 г., поглотив компанию Altera. Обе компании регулярно выпускают новые модели FPGA, к примеру: ZU11EG от Xilinx и Stratix 10 SX650 series от Altera (Intel), выпущенные в 2019 г. Сравнение различных FPGA обычно осуществляется на основе рассмотрения ресурсов, доступных на устройстве. К примеру, число логических элементов в Altera SX650 — 612 тыс. (+207 тыс.), в Xilinx ZU11EG — 653 тыс., то есть они имеют сравнимые характеристики [68].

Приложения, хорошо подходящие для ускорения с помощью FPGA, обычно имеют массивный ресурс параллелизма и работают с небольшими целочисленными типами данных либо с данными с плавающей точкой небольшой точности. Существенный прирост производительности при использовании FPGA был получен для задач криптографии, фильтрации сетевых пакетов, машинного обучения, оцифровки изображений, обработки ультразвука, финансовых расчетов, биоинформатики и др. Зачастую FPGA используется как сопроцессор к центральному процессору (аналогично графическим ускорителям): на FPGA выносятся все самые требовательные к вычислительной мощности операции, так как при конфигурации FPGA к конкретной задаче можно создать специализированный сопроцессор для каждого приложения HPC. Появление FPGA в HPC в последние годы в основном связано с существенным прогрессом в инструментах разработки FPGA и аппаратных технологиях: к примеру, использование OpenCL для программирования FPGA [69]. Можно выделить следующие примеры реальных суперкомпьютерных систем, реализованных с использованием FPGA: (1) EulerPrime HPC Intel Arria-10 GX1150, (2) Virtex-4 SX EasyPath FPGAs, (3) суперкомпьютеры MaxWell, (4) суперкомпьютер Cray XD1.

3.5. Архитектура и особенности построения наиболее мощных суперкомпьютеров мира. В данном разделе рассматривается архитектура наиболее мощных суперкомпьютеров мира на начало 2021 г. Наряду с ними рассматриваются также некоторые вычислительные системы, интересные с других точек зрения.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

§ 3.5.1. Самые мощные системы из списка Top500.

Список 500 наиболее мощных суперкомпьютеров мира Top500 [1] уже много лет традиционно используется для оценки уровня развития суперкомпьютерной отрасли. Не удивительно, что наибольшее внимание уделяется первой десятке списка как самым масштабным вычислительным системам своего времени. Рассмотрим первые десять суперкомпьютеров списка Top500 по состоянию на начало 2021 г.

Суперкомпьютер Fugaku. Суперкомпьютер Fugaku создан совместно RIKEN (RIKEN Center for Computational Science) и компанией Fujitsu и является сейчас самым мощным суперкомпьютером мира. Несмотря на то что полноценное использование Fugaku планировалось начать в 2021 г., дебют этой системы произошел в июне 2020 г. Fugaku занял первое место в списке Top500, превзойдя по производительности предыдущего лидера, суперкомпьютер Summit, приблизительно в три раза. Также стоит отметить, что это первый суперкомпьютер, основанный на архитектуре ARM, который занял первое место в списке Top500. Пиковая производительность суперкомпьютера Fugaku (ноябрь 2020 г.) составляет 537 Pflops, производительность на тесте Linpack 442 Pflops (размер матрицы 21288960), производительность на тесте HPCG — 16 Pflops. Энергопотребление суперкомпьютера Fugaku около 30 MW. Стоимость создания суперкомпьютера Fugaku вместе с созданием соответствующих технологий оценивается в 130 млрд йен (более 1 млрд долларов) [112].

Всего в Fugaku 158976 вычислительных узлов — 396 стоек по 384 узла и 36 стоек по 192 узла [70]. Каждая стойка имеет водяное охлаждение. В системе отсутствуют графические ускорители, на узлах есть только процессоры общего назначения, а именно процессоры Fujitsu A64FX. Они основаны на архитектуре ARM версии 8.2A и используют масштабируемые векторные расширения (Scalable Vector Extensions). На каждом узле установлено 32 GB оперативной памяти HBM2 со скоростью доступа 1024 GB/s.

В системе используется проприетарный интерконнект Tofu Interconnect D (28 Gb/s × 2 линии × 10 портов), созданный компанией Fujitsu. Эта коммуникационная сеть имеет топологию шестимерного тора, однако для использования в программном обеспечении абстрагирует это до трехмерного тора [71]. Для оптимальной работы такой топологии используется специальная реализация MPI. Для реализации ввода/вывода используется PCIe Gen3 x16.

В суперкомпьютере Fugaku используется операционная система Red Hat Enterprise Linux 8 и облегченное ядро операционной системы McKernel, поверх которых составлена оптимизированная экосистема из пакетов. Файловая система — FEFS (Fujitsu Exabyte File System) [72], для файлового ввода/вывода используется LLIO (Lightweight Layered IO-Accelerator) [73]. Создается оптимизированный под суперкомпьютер Fugaku программный стек. Компанией Fujitsu разрабатывается собственный компилятор для C/C++ Fujitsu Compiler. Также доступны компиляторы GCC и ARM Compiler. На системе также используются свои оптимизированные реализации MPI: Fujitsu MPI (основан на OpenMPI), RIKEN-MPICH (основан на MPICH). Для установки открытого ПО используется пакетный менеджер Spack, ведется учет программ, которые можно скомпилировать на Fugaku [74].

	Fugaku
Производительность Performance	
пиковая (peak)	537 Pflops
Linpack	442 Pflops
HPCG	16004 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	1 (2020)
текущее (current)	1 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	30 MW
Энергоэффективность Energy efficiency	15 Gflops/W
Страна установки Country	Япония Japan
Год установки Year	2020

Суперкомпьютер Summit. Суперкомпьютеры Summit и Sierra созданы в рамках проекта CORAL (Collaboration of Oak Ridge, Argonne, and Livermore) [75]. Суперкомпьютер Summit [76] разработан компанией IBM для Oak Ridge National Laboratory. Суперкомпьютер был введен в строй в июне 2018 г., заменив Titan, став на тот момент самым мощным суперкомпьютером мира.

Суперкомпьютер Summit объединяет 4608 вычислительных узлов, которые являются двухсокетными узлами IBM POWER9 (AC922), каждый такой узел содержит 6 графических ускорителей NVIDIA Tesla V100.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

В качестве коммуникационной сети используется сеть от Mellanox: Enhanced Data Rate (EDR) InfiniBand с пропускной способностью 100 Gb/s. Топология сети Fat Tree. Основной файловой системой IBM Summit является параллельная файловая система IBM Spectrum Scale.

Суперкомпьютер IBM Summit используется для исследований в областях энергетики, национальной безопасности, генетики, медицины, климатическом моделировании и многих других.

Суперкомпьютер Sierra. Суперкомпьютер IBM Sierra [77] установлен в Lawrence Livermore National Laboratory и входит в состав Livermore Computing Complex.

Архитектура суперкомпьютера Sierra практически аналогична суперкомпьютеру Summit за тем исключением, что в вычислительный узел входит не 6, а 4 графических ускорителя NVIDIA Tesla V100 (Volta), а общее число вычислительных узлов — 4320.

Суперкомпьютер Sierra используется, главным образом, для построения прогнозов в приложении к задачам управления ядерным арсеналом, помогая обеспечить безопасность, надежность и эффективность ядерного оружия США.

Суперкомпьютер Sunway TaihuLight. Суперкомпьютер Sunway TaihuLight [78], разработанный в National Research Center of Parallel Computer Engineering & Technology и установленный в National Supercomputing Center in Wuxi, в 2016 г. стал первым в мире суперкомпьютером с пиковой производительностью свыше 100 Pflops. Sunway TaihuLight построен на базе 260-ядерных китайских процессоров SW26010.

Базовым элементом вычислительной системы является однопроцессорный вычислительный узел. 256 вычислительных узлов объединяются в так называемый суперузел, четыре суперузла входят в кабинет, а весь суперкомпьютер Sunway TaihuLight состоит из 40 кабинетов.

В вычислительный узел Sunway TaihuLight входит 32 GB оперативной памяти, а также контроллер управления узлом, источник питания, интерфейсные схемы и т.д. Вычислительные узлы в рамках одного суперузла объединены при помощи матричного коммутатора. Для передачи сообщений между суперузлами используется центральная коммуникационная сеть. Бисекционная пропускная способность составляет 70 TB/s, пропускная способность канала сети — 16 GB/s, а диаметр сети (максимальное число переходов между вычислительными узлами) равен семи.

Система хранения суперкомпьютера включает дисковый массив объемом 20 PB. Для охлаждения вычислителя используется непрямое водяное охлаждение, а для периферийных систем применяется смешанное воздушно-водяное охлаждение.

В Sunway TaihuLight реализована Linux-подобная операционная система Raise OS 2.0.5. Установлены компиляторы с языков Фортран, Си/C++, OpenACC 2.0 с расширениями, OpenMP, для распараллеливания между узлами используется MPI. В статье [78] рассматривается ряд приложений, реализованных на суперкомпьютере Sunway TaihuLight.

	Summit
Производительность Performance	
пиковая (peak)	201 Pflops
Linpack	149 Pflops
HPCG	2926 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	1 (2018)
текущее (current)	2 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	10 MW
Энергоэффективность Energy efficiency	15 Gflops/W
Страна установки Country	США USA
Год установки Year	2018

	Sierra
Производительность Performance	
пиковая (peak)	126 Pflops
Linpack	95 Pflops
HPCG	1796 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	2 (2018)
текущее (current)	3 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	7.4 MW
Энергоэффективность Energy efficiency	12.7 Gflops/W
Страна установки Country	США USA
Год установки Year	2018

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

	Sunway TaihuLight
Производительность Performance	
пиковая (peak)	125 Pflops
Linpack	93 Pflops
HPCG	481 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	1 (2016)
текущее (current)	4 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	15 MW
Энергоэффективность Energy efficiency	6 Gflops/W
Страна установки Country	Китай China
Год установки Year	2016

	Selene
Производительность Performance	
пиковая (peak)	79 Pflops
Linpack	63 Pflops
HPCG	1623 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	5 (2020)
текущее (current)	5 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	2.6 MW
Энергоэффективность Energy efficiency	24 Gflops/W
Страна установки Country	США USA
Год установки Year	2020

Суперкомпьютер Selene. Суперкомпьютер NVIDIA Selene был изначально построен в стандартном центре обработки данных всего за три недели [79] по сравнению с 9–12 месяцами, которые обычно требуются для создания типичной суперкомпьютерной установки. Такое быстрое развертывание стало возможным благодаря системе NVIDIA Plug and Play DGX. В дальнейшем мощность системы была удвоена [80].

NVIDIA Selene построен по архитектуре DGX SuperPOD. Суперкомпьютер Selene состоит из 560 вычислительных узлов, каждый из которых включает систему DGX A100. В системе в качестве основных процессоров используются AMD EPYC 7742 64C, графические ускорители — NVIDIA A100. В совокупности это дает 555520 вычислительных ядер и 1120 TB оперативной памяти. Используемый интерконнект — Mellanox HDR Infiniband с пропускной способностью 200 Gb/s. Охлаждается суперкомпьютер воздушным путем.

NVIDIA Selene работает под управлением операционной системы Ubuntu 20.04.1 LTS. Программное обеспечение включает в себя средства компиляции NVCC, библиотеки NVIDIA CUDA и CUDA-X, а также средства компиляции Intel Composer, библиотеку Intel MKL.

На суперкомпьютере Selene решаются задачи в области искусственного интеллекта, например: обучение самоуправляемых автомобилей, совершенствование разговорного искусственного интеллекта. Ресурсы суперкомпьютера активно используются для борьбы с пандемией коронавируса в рамках Folding@home Initiative.

Суперкомпьютер Tianhe-2A. Суперкомпьютер Tianhe-2A (TH-2A, иногда “Milkyway”) — вычислительная система, расположенная в National Supercomputer Center (Гуанчжоу, Китай) [40]. Была спроектирована в 2013 г. National University of Defense Technology (NUDT) и компанией Inspur, представляет собой модернизацию системы Tianhe-2 (TH-2). Система занимала первое место в списке Top500 с июня 2013 г. по ноябрь 2015 г.

Ключевым отличием модернизированной версии вычислительной системы Tianhe-2A от ее старой версии являет-

	Tianhe-2A
Производительность Performance	
пиковая (peak)	101 Pflops
Linpack	61 Pflops
Место в Top500 Top500 rank	
наивысшее (the highest)	1 (2013)
текущее (current)	6 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	18 MW
Энергоэффективность Energy efficiency	3.3 Gflops/W
Страна установки Country	Китай China
Год установки Year	2013

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

ся замена ускорителей Xeon Phi на сопроцессоры собственного производства — Matrix-2000. Связано это изменение с запретом властей США на поставку процессоров Intel китайским суперкомпьютерным центрам. Всего Tianhe-2 состоит из 17792 вычислительных узлов, в каждый из которых входят два процессора Intel Ivy Bridge и два 128-ядерных ускорителя Matrix-2000. Каждый узел имеет 192 GB памяти. Вся сетевая логика была реализована в двух специализированных интегральных схемах (ASIC): сетевой карте (NIC) и чипе сетевого коммутатора (NRC). Сетевая карта содержит интерфейс x16 PCI Express 3.0 и подключается к портам, имеющим 8-полосный интерфейс SerDes 14 GB/s. Пропускная способность одного чипа NRC составляет 5.37 TB/s. Коммуникационная сеть, обозначенная как TH Express-2, имеет иерархическую топологию: каждые 32 вычислительных узла входят в один вычислительный фрейм и подключаются с помощью коммутатора 32×32, а вычислительные фреймы подключаются друг к другу через 24 576-портовых коммутатора верхнего уровня.

Модернизация TH-2A потребовала разработки и реализации стека программного обеспечения для ускорителя Matrix-2000. Этот программный стек предоставляет среду компиляции и выполнения для технологий OpenMP и OpenCL. В режиме ядра имеется облегченная операционная система на основе Linux со встроенным драйвером ускорителя, работающая на Matrix-2000, которая обеспечивает управление ресурсами устройства и обмен данными с центральным процессором через соединение PCI Express. Операционная система управляет вычислительными ядрами с помощью механизма пула потоков, который позволяет планировать задачи с низкими издержками и высокой эффективностью.

Суперкомпьютер JUWELS Booster Module. Суперкомпьютер JUWELS Booster Module [81] производства компании Atos установлен в Forschungszentrum Jülich (FZJ) в Германии. На данный момент это самый мощный суперкомпьютер Европы. Он является частью модульной системной архитектуры, второй модуль которой на процессорах Intel Xeon занял в Top500 позицию 44.

JUWELS Booster Module построен по архитектуре BullSequana. Он состоит из 936 вычислительных узлов, в состав которых входит 2 процессора AMD EPYC Rome 7402 и 4 графических ускорителя NVIDIA A100. В качестве коммуникационной сети используется Mellanox InfiniBand HDR топологии DragonFly.

Из больших проектов, в которых будет использоваться JUWELS Booster Module, называются Human Brain Project, а также проект по изучению климатических изменений.

Суперкомпьютер HPC5. Суперкомпьютер HPC5 (расшифровывается как High Performance Computing — layer 5) [82] установлен в компании Eni, занимающейся разработкой источников энергии в нефтегазовой отрасли. На данный момент HPC5 является самым производительным суперкомпьютером в мире из публично известных коммерческих систем.

HPC5 состоит из 1820 узлов Dell EMC PowerEdge C4140, на каждом узле установлено по два процессора Intel Xeon Gold 6252 (24 ядра, 48 потоков, архитектура Cascade Lake) и 4 графических ускорителя NVIDIA V100. Общий объем оператив-

JUWELS Booster Module	
Производительность Performance	
пиковая (peak)	71 Pflops
Linpack	44 Pflops
HPCG	1275 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	7 (2020)
текущее (current)	7 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	1.8 MW
Энергоэффективность Energy efficiency	24 Gflops/W
Страна установки Country	Германия Germany
Год установки Year	2020

HPC5	
Производительность Performance	
пиковая (peak)	52 Pflops
Linpack	35 Pflops
HPCG	860 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	6 (2020)
текущее (current)	8 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	2.3 MW
Энергоэффективность Energy efficiency	15.7 Gflops/W
Страна установки Country	Италия Italy
Год установки Year	2020

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

ной памяти системы 349440 GB. В качестве коммуникационной сети используется Mellanox HDR Infiniband с пропускной способностью 200 Gb/s.

Основным применением суперкомпьютера является построение моделей подземных резервуаров нефти и газа, а также поиск новых месторождений по данным сейсмической разведки. Примерами приложений являются ECHELON от компании Stone Ridge для моделирования резервуаров и tNavigator от Rock Flow Dynamics. Планируется также использование суперкомпьютера для расчета задач, связанных с переходом к возобновляемым источникам энергии, например для расчета параметров сверхпроводящих магнитов для удержания плазмы при ядерном синтезе и для запуска климатических моделей Арктики. Также заявлен совместный проект с IBM по обучению моделей искусственного интеллекта.

Суперкомпьютер Frontera. Суперкомпьютер Frontera [83] установлен в Texas Advanced Computing Center в Остине, штат Техас. На создание суперкомпьютера Национальный научный фонд США выделил 60 млн долларов. Главное предназначение системы — предоставление вычислительной мощности для научных исследований в области астрономии, физике высоких энергий и других областях.

Frontera — это суперкомпьютер с традиционной архитектурой, основанной на узлах Dell PowerEdge C6420, содержащих по два процессора Intel Xeon Platinum 8280 (56 ядер, без гипертредининга) с тактовой частотой 2.7 GHz. Каждый узел содержит 128 GB оперативной памяти. Есть быстрое локальное SSD-хранилище объемом 144 GB. Всего в суперкомпьютере изначально насчитывалось 8008 узлов. В 2021 г. к суперкомпьютеру были добавлены 396 узлов Dell R640 [84], имеющих те же процессоры Intel Xeon 8280 и память. Расширение было вызвано необходимостью увеличить компьютерные ресурсы для преодоления пандемии Covid-19, а также для предсказания скорости ветра в сезон ураганов. Новые узлы пока что не учтены в списке Top500.

Основное хранилище данных от Data Direct Networks состоит из 50 PB на жестких дисках (300 GB/s) и 3 PB флэш-памяти (1.5 TB/s). Также суперкомпьютер предоставляет отдельные узлы хранения данных, содержащие 4 процессора Intel Xeon Platinum 8280M, 2.1 TB быстрой памяти Intel Optane и еще дополнительно 3.2 TB постоянной памяти. Всего таких узлов 16, по остальным характеристикам они не отличаются от вычислительных.

В суперкомпьютере Frontera используется коммуникационная сеть Mellanox InfiniBand HDR топологии Fat Tree, с линками 100 Gb/s от узлов к коммутаторам и 200 Gb/s от крайних коммутаторов до корневых. Шесть корневых коммутаторов соединяют вычислительные узлы с файловыми системами. В каждой стойке стоит по два 40-портовых коммутатора, в каждый из них подключаются 44 узла, объединяясь в пары двумя линками по 100 Gb/s, занимая 22 порта коммутатора. Остальные 18 портов уходят к корневым коммутаторам.

На суперкомпьютере поддерживаются многие прикладные пакеты, такие как OpenFOAM, Matlab, TensorFlow и другие. Помимо этого, присутствует поддержка Intel MKL, которая включает оптимизированные версии BLAS и LAPACK, а также Intel MPI.

Суперкомпьютер Dammam-7. Суперкомпьютер Dammam-7 [85], установленный в нефтяной компании Saudi Aramco, назван в честь первой коммерческой нефтяной скважины, открытой в Саудовской Аравии в 1938 г. Dammam-7 создан компанией HPE Cray (совместно с Dhahran Techno Valley и Solutions). Система CS-Storm построена на базе процессоров Intel Xeon Gold 6248 с использованием графических ускорителей NVIDIA Tesla V100. В качестве коммуникационной сети используется NVIDIA InfiniBand HDR с пропускной способностью 100 Gb/s. Общий объем оперативной памяти составляет 506 TB.

	Frontera
Производительность Performance	
пиковая (peak)	39 Pflops
Linpack	24 Pflops
Место в Top500 Top500 rank	
наивысшее (the highest)	5 (2019)
текущее (current)	9 (2021)
Страна установки Country	США USA
Год установки Year	2019

	Dammam-7
Производительность Performance	
пиковая (peak)	55 Pflops
Linpack	22 Pflops
HPCG	881 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	10 (2020)
текущее (current)	10 (2021)
Страна установки Country	Саудовская Аравия Saudi Arabia
Год установки Year	2020

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

Dammam-7 будет использоваться для поиска ископаемого топлива, обрабатывая огромные наборы геофизических данных.

§ 3.5.2. Лидеры Top500 по энергопотреблению.

Поскольку энергопотребление становится одним из ключевых факторов при создании мощных суперкомпьютеров, отдельное внимание уделяется созданию систем с рекордно низким энергопотреблением. Рассмотрим системы, возглавляющие список Green500 [3], характеризующий системы по отношению производительности на тесте Linpack к энергопотреблению (на начало 2021 г.).

Суперкомпьютер NVIDIA DGX SuperPOD. Самым энергоэффективным суперкомпьютером мира в настоящее время является одна из реализаций технологии NVIDIA DGX SuperPOD [86], находящаяся в списке Top500 на 170-ом месте. Система построена на базе центральных процессоров AMD EPYC 7742 64C и графических ускорителей NVIDIA A100. В качестве коммуникационной сети используется Mellanox HDR Infiniband с пропускной способностью 200 Gb/s.

Экстраполяция значения энергоэффективности NVIDIA DGX SuperPOD, равного 26.2 Gflops/W, линейно на уровень экзафлопсной производительности приведет к потребляемой мощности в 38 MW (без учета дополнительного оборудования, необходимого для масштабирования). Это показывает необходимый уровень энергопотребления для построения экзафлопсных систем с использованием самых энергоэффективных на данный момент технологий построения суперкомпьютеров.

Суперкомпьютер MN-3. В июне 2020 г. был представлен суперкомпьютер MN-3 с пиковой производительностью 3.9 Pflops, обеспечивший энергоэффективность 21.1 Gflops/W, что сделало его самым энергоэффективным суперкомпьютером в мире. Система была создана Preferred Networks, японским стартапом в области искусственного интеллекта, который использовал собственный ускоритель MN-Core для достижения рекордной эффективности MN-3. С Preferred Networks сотрудничал производитель модульных систем Supermicro, который в своем отчете подробно рассказал об оборудовании и процессах, стоящих за лидером в рейтингах [87]. К ноябрю 2020 г. количество вычислительных узлов системы было уменьшено, пиковая производительность снизилась до 3.1 Pflops, что позволило увеличить энергоэффективность до 26 Gflops/W [88].

Один вычислительный узел MN-3 объединяет два процессора Intel Xeon Platinum, четыре платы MN-Core, до 6 TB оперативной памяти DDR4 и модули постоянной памяти Intel Optane. Четыре платы MN-Core сервера подключены к слотам PCIe x16 на материнской плате Supermicro и к плате MN-Core Direct Connect, которая обеспечивает высокоскоростную связь. В текущей конфигурации суперкомпьютер MN-3 объединяет 32 таких вычислительных узла.

Preferred Networks не планирует продавать свою технологию энергоэффективных компьютеров, поэтому MN-3 и его собственное оборудование доступны только для собственных исследований и разработок компании.

§ 3.5.3. Российские системы в списке Top500.

Представительство российских систем в списке наиболее мощных суперкомпьютеров мира Top500 не слишком велико.

NVIDIA DGX SuperPOD	
Производительность Performance	
пиковая (peak)	2.8 Pflops
Linpack	2.4 Pflops
HPCG	63 Tflops
Место в Top500 Top500 rank	
наивысшее (the highest)	170 (2020)
текущее (current)	170 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	90 KW
Энергоэффективность Energy efficiency	26.2 Gflops/W
Страна установки Country	США USA
Год установки Year	2020

	MN-3
Производительность Performance	
пиковая (peak)	3.1 Pflops
Linpack	1.7 Pflops
Место в Top500 Top500 rank	
наивысшее (the highest)	330 (2020)
текущее (current)	330 (2021)
Энергопотребление (Linpack) Energy consumption (Linpack)	62 KW
Энергоэффективность Energy efficiency	26 Gflops/W
Страна установки Country	Япония Japan
Год установки Year	2020

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

Максимального значения оно достигало в 2011 г. — 12 систем. Наиболее высокое место, занимаемое российским суперкомпьютером в списке Top500, — 12 (суперкомпьютер "Ломоносов" в 2009 г.). Рассмотрим две российские системы, входящие в список Top500 на начало 2021 г.

Суперкомпьютер "Кристофари". "Кристофари" [89] — суперкомпьютер, разработанный дочерней компанией Сбербанка SberCloud совместно с компанией NVIDIA. Он получил название в честь Николая Кристофари — первого клиента Сбербанка, открывшего в нем сберегательную книжку. Суперкомпьютер построен на основе высокопроизводительных узлов NVIDIA DGX-2, каждый из которых оснащен шестнадцатью вычислительными ускорителями Tesla V100 и двумя процессорами Intel Xeon Platinum 8168 24C. Каждый V100 содержит 32 GB памяти HBM2, что дает 512 GB. В качестве коммуникационной сети используется Mellanox InfiniBand EDR со скоростью 100 Gb/s.

Компьютер предоставляется в аренду на коммерческой основе. "Кристофари" доступен клиентам облачного сервиса компании SberCloud с 12 декабря 2019 г. Суперкомпьютер "Кристофари" создан специально для работы с алгоритмами искусственного интеллекта. Его мощности позволяют обучать программные модели, основанные на сложных нейронных сетях, в рекордно короткие сроки. Он предназначен для научно-исследовательских, коммерческих и государственных организаций, работающих в различных отраслях экономики: нефтегазовой, электроэнергетике, тяжелой промышленности, медицине, телекоммуникациях, ритейле и финансовом секторе. Суперкомпьютер "Кристофари" аттестовали для работы с персональными данными [90]. Сбербанк использует "Кристофари" для распознавания речи при анализе обращений клиентов в колл-центр банка, а также для автоматизированного робота-оператора. Суперкомпьютер также тестируют клиенты облачной платформы Сбербанка.

	Кристофари **Christofari**
Производительность Performance	
пиковая (peak)	8.8 Pflops
Linpack	6.7 Pflops
Место в Top500 Top500 rank	
наивысшее (the highest)	29 (2019)
текущее (current)	40 (2021)
Страна установки Country	Россия Russia
Год установки Year	2019

Суперкомпьютер "Ломоносов-2". Первая очередь суперкомпьютера "Ломоносов-2" [91, 92] была запущена в Московском государственном университете имени М. В. Ломоносова в 2014 г. Суперкомпьютер поставлен российской компанией "Т-Платформы" и долгое время был самым мощным суперкомпьютером России и Восточной Европы.

Первая очередь суперкомпьютера "Ломоносов-2" состояла из 1280 вычислительных узлов (5 стоек) на платформе процессоров Intel Haswell-EP E5-2697v3 и ускорителей NVidia Tesla K40M, в дальнейшем число узлов этого типа было увеличено до 1519 [93]. Далее к суперкомпьютеру добавились 160 вычислительных узлов на базе процессоров Intel Xeon Gold 6126 с двумя графическими ускорителями NVidia P100, 4 вычислительных узла на базе Intel Xeon Phi 7230, 16 вычислительных узлов на базе процессоров Intel Xeon Gold 6142 с двумя графическими ускорителями NVidia Tesla V100 и 18 вычислительных узлов на базе процессоров Intel Xeon Gold 6240 с двумя графическими ускорителями NVidia Tesla V100.

	Ломоносов-2 **Lomonosov-2**
Производительность Performance	
пиковая (peak)	5.5 Pflops
Linpack	2.5 Pflops
Место в Top500 Top500 rank	
наивысшее (the highest)	23 (2014)
текущее (current)	156 (2021)
Страна установки Country	Россия Russia
Год установки Year	2014

В системе две независимые управляющие сети стандарта Ethernet и две сети FDR InfiniBand. Одна из них используется для MPI-трафика и имеет современную топологию Flattened Butterfly, которая не только лучше масштабируется на системах большого размера, но и позволяет снизить количество используемых сетевых коммутаторов, сокращая стоимость сетевой инфраструктуры до 40% по сравнению с традиционными топологиями. Вторая сеть InfiniBand используется для доступа к данным и имеет стандартную топологию Fat Tree. Система хранения данных включает хранилище домашних каталогов объемом 290 TB, хранилище рабочих файлов на 1400 TB и служебное хранилище на 246 TB.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

Вычислительная часть системы построена на базе суперкомпьютерных платформ A-Class с высокой плотностью вычислений. A-Class имеет жидкостную систему охлаждения, в которой теплоносителем выступает вода с высоким тепловым потенциалом.

Возможностями суперкомпьютерного комплекса МГУ для выполнения фундаментальных исследований пользуются более 2200 ученых, специалистов и преподавателей из 20 подразделений университета, более 200 научных и учебных организаций России. Это настоящий центр коллективного пользования, обеспечивающий суперкомпьютерными ресурсами все вычислительное сообщество Российской Федерации. Каждый день на суперкомпьютере "Ломоносов-2" выполняются около 1000 вычислительно сложных задач, отвечающих всем приоритетным направлениям развития науки, техники и технологий, определенных Стратегией научно-технологического развития Российской Федерации. Востребованность суперкомпьютерного комплекса МГУ исключительно велика, определяя постоянную 100%-ю загрузку ресурсов ЦКП (очередь заданий, ждущих освобождения ресурсов, постоянно держится на уровне 200–300 заданий). На основе выполнения сотен проектов по изучению математических и физических принципов разработки суперкомпьютерных технологий ведется создание сверхмасштабируемых алгоритмов, пакетов и комплексов программ, реализующих высокоточные вычислительные модели и методы предсказательного моделирования, а также методики их внедрения в технологический цикл российских промышленных и научных организаций. Спектр исследований, поддерживаемых Суперкомпьютерным комплексом МГУ, исключительно широк: индустрия наносистем и новые материалы, живые системы, информационно-телекоммуникационные системы, энергетика и энергосбережение, транспортные, авиационные и космические системы, рациональное природопользование, перспективные вооружения, военная и специальная техника, безопасность и противодействие терроризму, цифровые технологии, роботизированные системы, персонализированная медицина и высокотехнологичное здравоохранение, и многие другие.

§ 3.5.4. Суперкомпьютеры, не входящие в список Top500.

Не все высокопроизводительные суперкомпьютеры можно встретить в рейтингах типа Top500. Ряд систем являются закрытыми, обычно про них очень мало публичной информации. Некоторые организации (например, National Center for Supercomputing Applications, где установлен суперкомпьютер Blue Waters) принципиально отказываются участвовать в рейтингах. Наконец, существуют специализированные системы, созданные под конкретные задачи в соответствии с принципами суперкомпьютерного кодизайна (например, Anton 2). Такие системы зачастую не выполняют операции с данными типа double precision, что необходимо для теста Linpack, но могут успешно использоваться для того класса задач, для которых они предназначены.

Суперкомпьютер Blue Waters. Один из самых амбициозных проектов по созданию суперкомпьютера Blue Waters [94] стартовал в США в 2007 г. Используя грант NSF, специалисты National Center for Supercomputing Applications (NCSA) и Университета штат Иллинойс построили National Petascale Computing Facility (NPCF), в котором разместился суперкомпьютер Blue Waters. Изначально поставщиком суперкомпьютера должна была стать компания IBM, но она вышла из проекта по причине финансовых разногласий. В качестве нового поставщика суперкомпьютера была выбрана компания Cray. В 2013 г. суперкомпьютер Blue Waters был введен в эксплуатацию.

Пиковая производительность суперкомпьютера Blue Waters составляет 13.34 Pflops. Система состоит из 22636 вычислительных узлов Cray XE6 и 4228 вычислительных узлов Cray XK7. В основу суперкомпьютера легли многоядерные серверные процессоры AMD Opteron серии 6200 и графические карты nVidia Tesla архитектуры Kepler. Объем оперативной памяти системы составляет 1.6 PB. Коммуникационная сеть суперкомпьютера Blue Waters — Cray Gemini топологии трехмерный тор $24 \times 24 \times 24$, пропускная способность канала — 9.6 Gb/s.

За годы эксплуатации доступ к суперкомпьютеру Blue Waters получили участники множества исследовательских проектов из самых разных областей науки.

Суперкомпьютер Anton 2. Суперкомпьютер Anton 2 [95] является хорошим примером специализированного высокопроизводительного компьютера. Он установлен в Pittsburgh Supercomputing Center и назван в честь основоположника научной микроскопии Антони ван Левенгука. Система разработки лаборатории D.E. Shaw Research предназначена специально для моделирования молекулярной динамики белков, нуклеиновых кислот, липидов и других видов молекул.

Суперкомпьютер Anton 2 состоит из 512 вычислительных узлов на основе интегральных схем специального назначения (ASIC), разработанных для численной реализации алгоритма классической моле-

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

кулярной динамики в моделях с межчастичными потенциалами, ориентированными на биомолекулярные задачи. Вычислительные узлы соединены коммуникационной сетью с топологией трехмерного тора $8 \times 8 \times 8$ с пропускной способностью каждого канала 224 Gb/s. При создании микропроцессоров ASIC и коммуникационной сети разработчики преследовали цель реализации максимально возможного уровня распараллеливания молекулярно-динамического алгоритма. ASIC содержит 16 подсистем ("flex"), каждая из которых имеет 4 встроенных процессора, называемых геометрическими ядрами. Геометрические ядра представляют собой 32-разрядные процессорные ядра, реализующие арифметические операции с фиксированной точкой. Для их программирования используется компилятор GNU Compiler Collection C++ с использованием специальных инструкций для пересылки данных.

Суперкомпьютер Anton 2 является первой платформой, обеспечивающей скорость моделирования в несколько микросекунд физического времени в день для биологических систем с миллионами атомов.

4. Тенденции развития суперкомпьютеров и высокопроизводительных многопроцессорных вычислительных систем.

4.1. Повышение энергоэффективности. Энергоэффективность — это первоочередная задача при развертывании любых компьютерных систем. Для любых систем — от мобильных устройств с батарейным питанием до дата-центров и суперкомпьютеров — энергопотребление ограничивает производительность, которую могут обеспечить вычислительные системы. В настоящее время инженеры и исследователи ищут альтернативу современным суперкомпьютерам, которые в большинстве своем основаны на низкоэнергоэффективных процессорах.

Аналогично списку Top500 из наиболее производительных суперкомпьютерных систем мира, два раза в год составляется список Green500 [3] из наиболее энергоэффективных суперкомпьютеров, в котором системы ранжируются на основе метрики Flops/W. Данная метрика вычисляется как производительность, полученная на основе теста Linpack, деленная на энергопотребление системы в ходе выполнения данного теста. Далее будут рассмотрены основные тенденции списка Green500 на основе анализа его первых 10 систем и их основных аппаратных составляющих.

Среди первых десяти систем присутствуют семь систем на основе графических ускорителей NVIDIA, а также две системы на основе процессоров архитектуры ARM A64FX, в которых отсутствует сопроцессор. Кроме того, четыре системы используют центральные процессоры AMD EPYC. Далее будут описаны основные причины высокой энергоэффективности данных систем, которая по большей части определяется высокой энергоэффективностью их процессорной основы или используемого ускорителя.

Основная причина более высокой энергоэффективности процессоров на архитектуре ARM по сравнению с процессорами Intel на основе архитектуры x86 — это использование набора команд RISC, в то время как каждый процессор на базе архитектуры x86 имеет набор команд, подобный CISC. Из-за совокупности причин, в том числе необходимости в обратной совместимости со старыми наборами инструкций, нефиксированной длины инструкции и др., блок декодирования CISC-процессоров имеет гораздо более сложное устройство (например, наличие сложного суперскалярного конвейера), вследствие чего процессоры на основе архитектуры x86 потребляют существенно больше энергии [96]. К примеру, максимальное энергопотребление процессора A64FX (основа узла суперкомпьютера Fugaku) составляет около 170 W, вследствие чего при производительности в 3.47 Tflops (на двойной точности) энергоэффективность A64FX составляет 19.8 Gflops/W. Для сравнения: процессоры Intel KNL со сравнимой пиковой производительностью в 3.46 Tflops потребляют приблизительно 245 W, вследствие чего их энергоэффективность составляет 14.4 Gflops/W [98].

Современные графические ускорители NVIDIA так же имеют высокие показатели энергоэффективности. Достигается это за счет высокой пиковой производительности современных GPU, сильно упрощенной подсистемы памяти (всего лишь 2 уровня иерархии кэш-памяти и наличие программного кэша, что сильно снижает энергопотребление, необходимое на перемещение данных), а также сильно упрощенной вычислительной логики легковесных ядер GPU. Так, пиковая производительность V100 GPU составляет 7 Tflops, а максимальное энергопотребление — 250 W, вследствие чего энергоэффективность GPU составляет 28 Gflops/W.

Наконец, современные процессоры AMD Epyc тоже демонстрируют высокие показатели энергоэффективности. Данные процессоры, во-первых, используют специальный кэш инструкций для их хранения после того, как инструкции были декодированы в микрооперации. За счет этого процессор экономит время (около двух тактов) и энергию за счет пропуска этапа повторного декодирования ранее встречавшихся инструкций. Кроме того, данные процессоры используют несложную нейронную сеть для повышения

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

точности предсказания ветвлений, что также приводит к более высокой производительности и меньшим потерям энергии. Наконец, иерархия памяти AMD Epyc нацелена на уменьшение объема перемещаемых данных: кэш L3 имеет большой объем (256 MB), что помогает значительно сократить число "дорогих" обращений к оперативной памяти, кэш L1 использует политику обратной записи, а копирование содержимого регистров реализовано посредством простого переименования целевого регистра, а не комбинацией загрузки и сохранения [97].

4.2. Использование в задачах искусственного интеллекта. Очень важный современный тренд — эффективная поддержка вычислительными платформами задач искусственного интеллекта. На сегодняшний день проектируются специализированные процессоры, выпускаются модификации уже существующих платформ, которые ориентированы именно на данный очень быстро развивающийся сегмент, например NVIDIA, Google, Fujitsu, IBM, Huawei и другие, позволяющие быстро выполнять специальные операции над данными различной точности: двойной, одинарной и половинной. Дополнительно разрабатываются уже упоминавшиеся ранее специализированные форматы, такие как BFLOAT и TF32 [18], иллюстрация которых приведена на рис. 5. Основной идеей данных специализированных форматов является уменьшение размера, используемого для хранения чисел с плавающей точкой, с одновременным сохранением либо диапазона значений, либо точности — в сравнении с другими форматами. К примеру, в формате TF32 в архитектуре NVIDIA Ampere GPU используется 1 бит для хранения знака, 8 бит для хранения экспоненты и 10 бит для хранения мантиссы, а в формате BFLOAT16 для хранения мантиссы используется 7 бит. Благодаря тому, что данные форматы, так же как и FP32, используют для хранения экспоненты 8 бит, значительно упрощается процесс их преобразования в FP32 (в отличие, к примеру, от FP16, где под экспоненту отводится 5 бит). Благодаря этому тензорные ядра NVIDIA GPU могут производить вычисление матричных произведений в формате TF32, получая входные данные и аккумулируя результаты в формате FP32.

4.3. Экзафлопсные инициативы. Термин "экзафлопсные вычисления" (exascale computing; 1 Exaflops = 10^{18} арифметических операций над вещественными числами с плавающей точкой в секунду) стал активно употребляться после появления в 2008 г. первого суперкомпьютера петафлопсного уровня IBM Roadrunner [99]. По первоначальным прогнозам, первый экзафлопсный компьютер должен был появиться в мире примерно в 2018–2020 гг. [100]. Позднее эти прогнозы корректировались в ту или иную сторону. К моменту написания данного обзора появление первого экзафлопсного компьютера намечено на конец 2021 г., скорее всего это будет американский суперкомпьютер Frontier. Если же не ограничи-

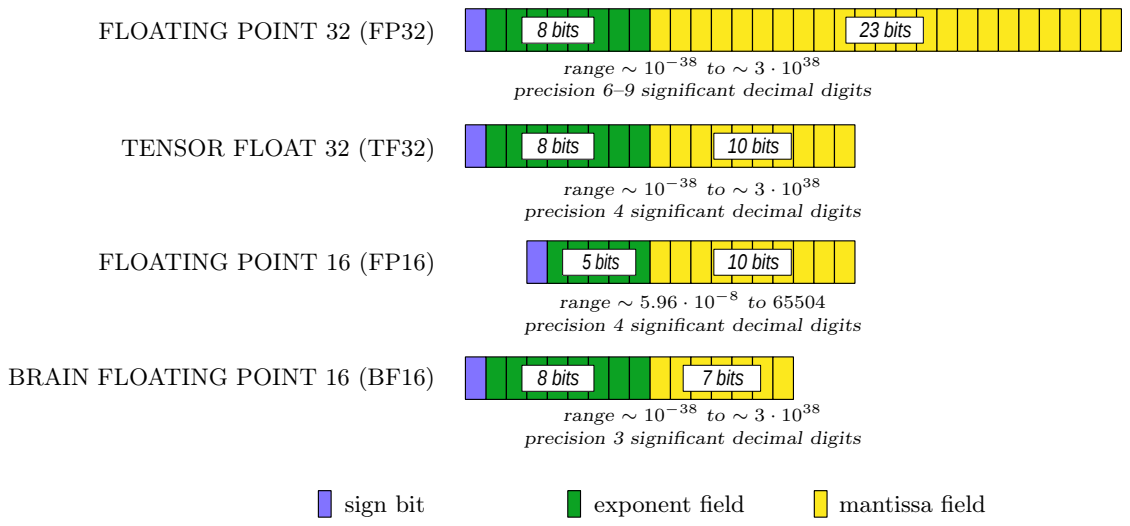

Рис. 5. Форматы представления чисел с плавающей точкой, использующиеся в задачах глубокого обучения [18]

Fig. 5. Floating-point number representation formats used in deep learning tasks [18]

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

ваться только суперкомпьютерами, то барьер экзафлопсной производительности был преодолен в 2020 г. метакомпьютерным проектом Folding@home при решении задачи по поиску лекарства от коронавирусной инфекции COVID–19 [101].

Создание суперкомпьютеров экзафлопсного уровня стало важным вызовом, стимулирующим развитие суперкомпьютерных технологий. В ряде стран этот вызов был принят на уровне правительственных программ. Так, например, в США правительство выделило 126 млн долларов из федерального бюджета на создание экзафлопсных суперкомпьютеров [102].

§ 4.3.1. Экзафлопсные проекты США.

В США объявлены три проекта по созданию суперкомпьютеров экзафлопсного уровня. Сравнение основных характеристик планируемых суперкомпьютеров приводится в табл. 2.

Frontier. Национальная лаборатория Ок-Риджа (Oak Ridge National Laboratory) имеет богатый многолетний опыт развертывания, запуска и поддержания функционирования сверхмощных суперкомпьютерных систем. Начиная с 2005 г. в ORNL введены в эксплуатацию три крупнейшие суперкомпьютерные системы — Jaguar, Titan и Summit. Однако в ближайшее время к запуску запланирована еще одна вычислительная система, которая должна стать первым экзафлопсным суперкомпьютером в мире.

Суперкомпьютер Frontier [103] будет введен в эксплуатацию в США в 2021 г. Пиковая производительность решения обещает превысить 1.5 Eflops. В денежном выражении это может стать крупнейшим в истории контрактом Cray, стоимостью свыше 600 млн долларов. В суперкомпьютере Frontier будут установлены специальные версии процессоров AMD EPYC и графические процессоры Radeon Instinct, соединенные шинами Infinity Fabric. Каждый узел будет иметь один сетевой порт HPE Slingshot [106] для каждого графического процессора с оптимизированной связью между графическими процессорами и сетью, чтобы обеспечить оптимальную производительность для высокопроизводительных вычислений и рабочих нагрузок. Каждый вычислительный узел будет содержать один центральный процессор AMD EPYC и четыре графических ускорителя Radeon Instinct с высокоскоростными связями и когерентной памятью между ними в узле. Интерфейс HPE Slingshot сможет передавать данные со скоростью до 200 Gb/s на порт. Всего площадь устройства займет 678 м^2, а уровень потребления энергии составит 40 MW. Используя суперкомпьютер Frontier, ученые смогут проводить больше вычислений, выявлять новые закономерности в данных и разрабатывать инновационные методы анализа данных для ускорения темпов научных открытий.

Таблица 2. Сравнение экзафлопсных проектов США

Table 2. Comparison of US exascale projects

Суперкомпьютер Supercomputer	Frontier	Aurora	El Capitan
Архитектура CPU CPU architecture	AMD EPYC (Future Zen)	Intel Xeon Scalable “Sapphire Rapids”	AMD EPYC “Genoa” (Zen 4)
Архитектура GPU GPU architecture	Radeon Instinct	Intel Xe “Ponte Vecchio”	Radeon Instinct
Интерконнект Interconnect	HPE Slingshot	HPE Slingshot	HPE Slingshot
Пиковая производительность Peak performance	1.5 Eflops	1 Eflops	2 Eflops
Энергопотребление Energy consumption	30–40 MW	$\leqslant$ 60 MW	$\leqslant$ 40 MW
Стоимость Cost	\$600 млн	\$500 млн	\$600 млн
Место установки Site	Oak Ridge	Argonne	Lawrence Livermore
Производитель Manufacturer	HPE (Cray)	HPE (Cray), Intel	HPE (Cray)
Год установки Year	2021	2021	2023

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

Aurora. Представители Министерства энергетики США (U.S. Department of Energy), которое выступает заказчиком, в марте 2019 г. официально подтвердили, что создаваемый компаниями Intel и Cray суперкомпьютер Aurora [104], способный "обеспечить устойчивую производительность порядка одного экзафлопса", будет запущен в строй в Аргоннской национальной лаборатории (Argonne National Laboratory) ближе к концу 2021 г. Сумма заключенного контракта оценивается в 500 млн долларов. Изначально именно суперкомпьютер Aurora должен был стать первым американским экзафлопсным суперкомпьютером [105], однако произошедшее смещение сроков привело к тому, что сейчас его запуск запланирован позже запуска суперкомпьютера Frontier.

Конструкция суперкомпьютера Aurora базируется на базе унифицированных кластерных систем Cray Shasta, объединенных коммуникационной сетью HPE Slingshot [106] и программным стеком Shasta. Каждая система Shasta базируется на процессорах Intel Xeon Scalable нового поколения, вычислительной архитектуре графических ускорителей Intel Xe, энергонезависимой памяти Intel Optane Datacenter Persistent Memory нового поколения, с использованием программного стека Intel One API. Каждый вычислительный узел суперкомпьютера будет содержать два процессора Intel Xeon Scalable третьего поколения ("Sapphire Rapids") и 6 графических процессоров Intel Xe "Ponte Vecchio".

Суперкомпьютер Aurora будет использовать "новые технологии Intel, разработанные специально для конвергенции искусственного интеллекта и высокопроизводительных вычислений в экстремальных масштабах". К ним относятся в том числе решения на вычислительной архитектуре Intel Xe. По словам представителей партнеров проекта (Министерства энергетики США, Аргоннской лаборатории, Intel и Cray), суть проекта не только в достижении "горизонта экзаскейла", основная роль проекта Aurora — в ускорении сближения высокопроизводительных вычислений с традиционными методиками моделирования с применением анализа данных и искусственного интеллекта.

El Capitan. Сотрудничество компании Cray (сейчас принадлежит HPE) и Ливерморской национальной лаборатории (Lawrence Livermore National Laboratory) ведется с 1978 г., в те годы лаборатория была одним из первых мест в мире, где был установлен суперкомпьютер Cray-1. В настоящее время Cray и Ливерморская лаборатория начали новый проект по созданию на базе лаборатории экзафлопсного суперкомпьютера. Новый суперкомпьютер, El Capitan [107], архитектурно будет похож на суперкомпьютер Frontier. Он будет построен на основе платформы Cray Shasta. Системы Shasta будут соединены интерконнектом HPE Slingshot, и в качестве системы хранения будет использоваться ClusterStor. Каждый вычислительный узел будет содержать один центральный процессор AMD EPYC и четыре графических ускорителя Radeon Instinct с высокоскоростными связями и когерентной памятью между ними в узле.

Ожидается, что производительность суперкомпьютера El Capitan будет порядка 2 Eflops (первоначально планировалась пиковая производительность 1.5 Eflops). Суперкомпьютер будет потреблять порядка 40 MW энергии. Основная сфера применения суперкомпьютера — моделирование ядерного вооружения, другие предполагаемые области применения — поиск лекарства от рака и причин мутаций белков RAS человека, вызывающих онкологические заболевания. Запуск данного суперкомпьютера запланирован на начало 2023 г. Ориентировочная стоимость суперкомпьютера составляет 600 млн долларов.

§ 4.3.2. Экзафлопсные проекты Китая.

Правительство Китая финансирует сразу три отличающихся по своей архитектуре проекта по созданию сверхмощного суперкомпьютера. Речь идет о создании суперкомпьютеров на базе трех разных организаций. Данные проекты были объявлены достаточно давно, для них были разработаны прототипы суперкомпьютеров, но новой информации по ним весьма мало, текущий статус этих проектов не до конца понятен, изначально объявленные сроки не соблюдены.

Tianhe. Tianhe — линейка суперкомпьютеров, спроектированных совместно National University of Defense Technology (NUDT) и компанией Inspur. На данный момент линейка насчитывает два суперкомпьютера: Tianhe-1 и Tianhe-2. Последний в 2013 г. занял первое место в списке Top500. На тесте Linpack он достиг 33.8 Pflops, пиковая же производительность составляет 55 Pflops.

Разработка нового суперкомпьютера Tianhe-3 [108] началась в 2016 г. В 2018 г. прототип Tianhe-3 был успешно построен и протестирован, а запуск самого Tianhe-3 изначально был запланирован на середину 2020 г.

Предполагается использование 64-ядерных центральных процессоров (предположительно архитектуры ARM, возможно, Phytium Xiaomi) с пиковой производительностью 2 Tflops. В качестве ускорителя предполагается использовать Matrix-3000, следующую версию китайского ускорителя Matrix-2000, приме-

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

ненного в суперкомпьютере Tianhe-2A. Предполагается, что каждый такой ускоритель будет иметь, как минимум, 96 вычислительных ядер и достигать 10 Tflops пиковой производительности. В каждый вычислительный узел будет входить 8 центральных процессоров и 8 ускорителей, что в сумме дает 96 Tflops.

Tianhe-3 будет состоять из 100 стоек по 128 вычислительных узлов в каждой, что в сумме дает пиковую производительность 1.29 Eflops. В качестве интерконнекта предполагается использование коммуникационной сети собственной разработки топологии 3D butterfly с пропускной способностью 400 Gb/s.

Sunway (ShenWei). Второй проект, который тоже уже был представлен прототипом [109], принадлежит компании Sunway. Прототип является предшественником системы Sunway exascale, которая будет установлена в National Supercomputing Center в Цзинане. Ожидается, что машина Sunway будет использовать процессоры ShenWei, последняя версия которых применяется для суперкомпьютера Sunway Taihulight с производительностью на тесте Linpack 93 Pflops. Однако никаких подробностей о деталях реализации процессоров, которые будут использоваться для нового экзафлопсного компьютера, пока не обнаружено. В качестве интерконнекта предполагается использовать коммуникационную сеть китайского производства с пропускной способностью 200 Gb/s. В качестве сроков создания суперкомпьютера указывались конец 2020 — начало 2021 г. Объявлялась стоимость реализации проекта создания экзафлопсного суперкомпьютера в 3 млрд юаней (470.6 млн долларов).

Sugon. То же самое можно сказать и про третий прототип экзафлопсного суперкомпьютера, анонсированного в 2018 г. Суперкомпьютер с экзафлопсной производительностью Shuguang [110], как ожидается, будет опираться на процессоры архитектуры x86 китайского производства. Китайский производитель процессоров Hygon имеет такую технологию в виде процессора Zen server. Hygon уже поставляет свои первые такие чипы местного производства на внутренний рынок, и эти чипы, вероятно, и станут основой для всей суперкомпьютерной экзафлопсной системы Shuguang.

Планируется, что каждый вычислительный узел суперкомпьютера Shuguang будет содержать два центральных процессора Hygon и два ускорителя также производства Hygon. Все вычислительные узлы будут соединены коммуникационной сетью топологии 6D-тор с пропускной способностью 200 Gb/s. Объявлялось об использовании в данном суперкомпьютере погружного жидкостного охлаждения.

§ 4.3.3. Экзафлопсные проекты Японии.

Fugaku. Суперкомпьютер K производства Fujitsu некоторое время (2011–2012 гг.) был первым в списке Top500, его производительность на тесте Linpack превышала 10 Pflops. 30 августа 2019 г. было объявлено об остановке работы суперкомпьютера K.

В 2014 г. в рамках проекта FLAGPSHIP 2020 начались совместные работы Riken и Fujitsu по созданию суперкомпьютера Fugaku (изначально Post-K) на базе процессоров ARM. К 2020 г. пиковая производительность суперкомпьютера Fugaku достигла 537 Pflops, и он стал самым мощным суперкомпьютером мира, заняв 1 место в списке Top500.

Пиковая производительность Fugaku заявлялась в 100 раз выше производительности системы Fujitsu K, то есть должна была превысить 1 Eflops. При этом потребляемая мощность должна была вырасти всего в три раза.

При работе с числами половинной точности пиковая производительность суперкомпьютера Fugaku составляет 2.15 Eflops. На бенчмарке HPC-AI [111] с данными смешанной точности на Fugaku достигнута производительность 2.0 Eflops — таким образом, Fugaku является первым суперкомпьютером, достигшим производительности выше одного экзафлопса любой точности на любом типе оборудования.

Суперкомпьютер Fugaku официально будет принят в эксплуатацию в 2021 г. Создатели суперкомпьютера рассматривают варианты создания суперкомпьютеров следующих поколений [112].

§ 4.3.4. Экзафлопсные проекты Европы.

Еще с 2011 г. в Европейском союзе развивались три проекта по созданию аппаратных и программных технологий для экзафлопсных суперкомпьютеров:

- CRESTA (Collaborative Research into Exascale Systemware, Tools and Applications) [113],
- DEEP (Dynamical ExaScale Entry Platform) [114],
- Mont-Blanc [115].

EuroHPC. 28 ноября 2018 г. был объявлен тендер о создании и закупке Евросоюзом трех суперкомпьютерных систем, которые должны были дать толчок к созданию экзафлопсных систем в рамках проекта

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

EuroHPC [116]. Эти системы называются “pre-exascale computing systems”, что подчеркивает намерение консорциума вплотную подобраться к экзафлопсной производительности. Эти три суперкомпьютера:

- LUMI (Финляндия), ожидаемая пиковая производительность — 550 Pflops [117],
- MareNostrum 5 (Испания), ожидаемая пиковая производительность — 200 Pflops [118],
- Leonardo (Италия), ожидаемая пиковая производительность — 200+ Pflops [119].

У проекта EuroHPC объявлены две основные цели:

- развитие европейской суперкомпьютерной инфраструктуры — покупка и развертывание в ЕС не менее двух суперкомпьютеров, которые будут среди Top5 в мире (ориентировочно к 2022/2023 г.), и не менее двух других, которые сегодня входят в Top25,
- поддержка научно-исследовательской и инновационной деятельности — развитие европейской суперкомпьютерной экосистемы, стимулирование индустрии поставок технологий и предоставление суперкомпьютерных ресурсов во многих прикладных областях в распоряжение большого числа государственных и частных пользователей, включая малые и средние предприятия.

Возможно, базой для создания европейских экзафлопсных суперкомпьютеров могут стать процессоры, разработанные в рамках European Processor Initiative [12]. Объявлялось, что общий бюджет проекта EuroHPC составляет около 1 млрд евро.

5. Заключение. В данной статье приведен обзор архитектур высокопроизводительных вычислительных платформ в мире на начало 2021 г. Суперкомпьютерный мир меняется стремительно, что постоянно подтверждается появлением новых систем и выводом из эксплуатации отслуживших свое. Появляется новая процессорная основа, возникают новые варианты интерконнекта; наконец, время от времени происходят и изменения в архитектуре суперкомпьютеров. Каким будет суперкомпьютерный мир через несколько лет, можно только предполагать, но очевидно, что высокопроизводительные вычисления становятся необходимы всем, и востребованность в больших суперкомпьютерных системах будет только возрастать.

Динамичное развитие суперкомпьютерной области приводит к тому, что, помимо описанного в рамках данного обзора, постоянно появляется что-то интересное и заслуживающее внимания (например, весной 2021 г. NVIDIA объявила о выпуске нового процессорного модуля BlueField-3 [120]). Поэтому авторы планируют развитие данного обзора, в которое будут включать описания новых суперкомпьютеров, архитектурных особенностей и тенденций развития, появляющихся в суперкомпьютерном мире. Авторы будут благодарны читателям за любые пожелания и рекомендации того, что еще стоило бы включить в следующие редакции подобного обзора.

Список литературы

1. Home: TOP500. https://www.top500.org.
2. Graph 500: Large-Scale Benchmarks. https://graph500.org.
3. Green500: TOP500. https://www.top500.org/lists/green500.
4. HPCG: TOP500. https://www.top500.org/lists/hpcg.
5. *Antonov A.S., Nikitenko D.A., Voevodin V.V.* Algo500 — a new approach to the joint analysis of algorithms and computers // Lobachevskii J. Math. 2020. **41**, N 8. 1435–1443. doi 10.1134/S1995080220080041.
6. Together We Are Powerful. Folding@home. https://foldingathome.org.
7. *Воеводин Вл.В., Капитонова А.П.* Методы описания и классификации архитектур вычислительных систем. М.: Изд-во МГУ, 1994.
8. *Воеводин В.В., Воеводин Вл.В.* Параллельные вычисления. СПб.: БХВ–Петербург, 2002.
9. Классификации архитектур вычислительных систем | PARALLEL.RU — Информационно-аналитический центр по параллельным вычислениям. https://parallel.ru/computers/taxonomy.
10. *Flynn M.J.* Very high-speed computing systems // Proc. IEEE. 1966. **54**, N 12. 1901–1909.
11. *Flynn M.J.* Some computer organizations and their effectiveness // IEEE Trans. Comput. 1972. **21**, N 9. 948–960.
12. Home: European Processor Initiative. https://www.european-processor-initiative.eu.
13. *Fisher J.A., Faraboschi P., Young C.* VLIW Processors // Encyclopedia of Parallel Computing. Boston: Springer, 2011. 2135–2142. doi 10.1007/978-0-387-09766-4_471.
14. *Afanasyev I.V., Voevodin V.V., Komatsu K., Kobayashi H.* VGL: a high-performance graph processing framework for the NEC SX-Aurora TSUBASA vector architecture // The Journal of Supercomputing. 2021. doi 10.1007/s11227-020-03564-9.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

15. Технология Intel® Hyper-Threading. https://www.intel.ru/content/www/ru/ru/architecture-and-technology/hyper-threading/hyper-threading-technology.html.
16. *Okonta O.E. et al.* Performance evaluation of hyper threading technology architecture using Microsoft operating system platform // West African Journal of Industrial and Academic Research. 2015. **15**, N 1. 52–67.
17. Intel Hyper Threading Performance with A Core i7 On Ubuntu 18.04 LTS. https://www.phoronix.com/scan.php?page=article&item=intel-ht-2018&num=4.
18. What is the TensorFloat-32 Precision Format? | NVIDIA Blog. https://blogs.nvidia.com/blog/2020/05/14/tensorfloat-32-precision-format/.
19. AMD Chips Away at Intel in World's Top 500 Supercomputers as GPU War Looms. https://www.crn.com/news/components-peripherals/amd-chips-away-at-intel-in-world-s-top-500-supercomputers-as-gpu-war-looms.
20. Intel Server Roadmap: 14nm Cooper Lake in 2019, 10nm Ice Lake in 2020. https://www.anandtech.com/show/13194/intel-shows-xeon-2018-2019-roadmap-cooper-lakesp-and-ice-lakesp-confirmed.
21. Intel® Xeon® Gold 6248 Processor (27.5M Cache, 2.50 GHz) Product Specifications. https://ark.intel.com/content/www/us/en/ark/products/192446/intel-xeon-gold-6248-processor-27-5m-cache-2-50-ghz.html.
22. EPYC 7742-AMD. https://en.wikichip.org/wiki/amd/epyc/7742.
23. Epyc-Wikipedia. https://en.wikipedia.org/wiki/Epyc.
24. IBM Power9-WikiChip. https://en.wikichip.org/wiki/ibm/microarchitectures/power9#Memory_Hierarchy.
25. FUJITSU Processor A64FX. https://www.fujitsu.com/downloads/SUPER/a64fx/a64fx_datasheet.pdf.
26. Baikal-M—Baikal Electronics—WikiChip. https://en.wikichip.org/wiki/baikal/baikal-m.
27. *Wang Y.C. et al.* An empirical study of HPC workloads on Huawei Kunpeng 916 processor // IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). New York: IEEE Press, 2019. 360–367.
28. Vulcan–Microarchitectures–Cavium–WikiChip. https://en.wikichip.org/wiki/cavium/microarchitectures/vulcan.
29. *Calore E. et al.* ThunderX2 performance and energy-efficiency for HPC workloads // Computation **8**, N 1. 2020. 1–17. doi 10.3390/computation8010020.
30. Hot Chips 2020: Marvell Details ThunderX3 CPUs–Up to 60 Cores Per Die, 96 Dual-Die in 2021. https://www.anandtech.com/show/15995/hot-chips-2020-marvell-details-thunderx3.
31. eMAG–Ampere–WikiChip. https://en.wikichip.org/wiki/ampere_computing/emag.
32. *Альфонсо Д.М. и др.* Микроархитектура восьмиядерного универсального микропроцессора "Эльбрус 8С" // Вопросы радиоэлектроники. № 3. 6–13. 2016.
33. Китайский процессорно-суперкомпьютерный путь. https://www.osp.ru/os/2017/01/13051592.
34. SPARC64 XIfx: Fujitsu's Next Generation Processor for HPC. https://www.fujitsu.com/global/Images/20140811hotchips26_tcm100-1156766.pdf.
35. MN-Core: Preferred Networks. https://projects.preferred.jp/mn-core/en.
36. Cerebras Wafer Scale Engine. https://www.ixbt.com/news/2019/08/20/cerebras-wafer-scale-engine-ipad-pro-1-tdp-15.html.
37. Первый GPU на архитектуре CDNA. https://www.hardwareluxx.ru/index.php/news/hardware/grafikkarten/50677-pervyj-gpu-na-arkhitekture-cdna-amd-predstavila-instinct-mi100.html.
38. *Komatsu K. et al.* An approach to the highest efficiency of the HPCG benchmark on the SX-ACE supercomputer // Proceedings of the Conference for High Performance Computing, Networking, Storage, and Analysis (SC15). 2015. http://sc15.supercomputing.org/sites/all/themes/SC15images/tech_poster/poster_files/post277s2-file3.pdf.
39. *Komatsu K. et al.* Performance evaluation of a vector supercomputer SX-Aurora TSUBASA // SC18: International Conference for High Performance Computing, Networking, Storage, and Analysis. Piscataway: IEEE Press, 2018. 685–696.
40. Dongarra J. Report on the Tianhe-2A system. Tech Report No. ICL-UT-17-04. Knoxville: Univ. Tennessee, 2017. https://www.dropbox.com/s/0jyh5qlgok73t1f/TH-2A-report.pdf?dl=0.
41. Hardwareluxx: Третье поколение Google TPU. https://www.hardwareluxx.ru/index.php/news/hardware/prozessoren/44682-googles-tpu.html.
42. System Architecture | Cloud TPU. https://cloud.google.com/tpu/docs/system-architecture.
43. Google Cloud Blog. https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu.
44. What Is a Data Processing Unit (DPU). https://www.forbes.com/sites/janakirammsv/2020/10/11/what-is-a-data-processing-unit-dpu-and-why-is-nvidia-betting-on-it.
45. What Is a DPU? | NVIDIA Blog. https://blogs.nvidia.com/blog/2020/05/20/whats-a-dpu-data-processing-unit/.
46. Hardwareluxx: NVIDIA + Mellanox. https://www.hardwareluxx.ru/index.php/news/hardware/grafikkarten/50439-nvidia-mellanox-konvergentsiya-tekhnologij-v-budushchikh-dpu.html.
47. NVIDIA BLUEFIELD-2 DPU. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-2-dpu.pdf.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

48. ИНТУИТ: Лекция. https://intuit.ru/studies/courses/1156/190/lecture/4942?page=4.
49. NUMA Deep Dive. Part 1. https://frankdenneman.nl/2016/07/07/numa-deep-dive-part-1-uma-numa.
50. *Kanter D.* The common system interface: Intel's future interconnect // Real World Technologies. 2007. https://www.realworldtech.com/common-system-interface/
51. *Wang T. et al.* NovaCube: A low latency torus-based network architecture for data centers // IEEE Global Communications Conference. New York: IEEE Press, 2014. 2252–2257.
52. Fat Tree — Википедия. https://ru.wikipedia.org/wiki/Fat_Tree.
53. *X. Yuan.* On Nonblocking Folded-Clos Networks in Computer Communication Environments // IEEE International Parallel & Distributed Processing Symposium. IEEE Press, 2011. 188–196. doi 10.1109/IPDPS.2011.27.
54. Сеть бабочек—Butterfly network—qaz.wiki. https://ru.qaz.wiki/wiki/Butterfly_network.
55. *Kim J., Dally W.J., Abts D.* Flattened butterfly: a cost-efficient topology for high-radix networks // ACM SIGARCH Computer Architecture News. 2007. **35**, N 2. doi 10.1145/1273440.1250679.
56. Intel® Omni-Path Architecture Performance Tested for HPC. https://www.intel.ru/content/www/ru/ru/high-performance-computing-fabrics/omni-path-architecture-performance-overview.html.
57. Low-Latency Ethernet Solutions for High-Performance Computing. https://www.cisco.com/c/dam/en_us/solutions/industries/docs/education/ethernet-solutions-high-performance-computing-education.pdf.
58. *Ajima Y. et al.* The Tofu interconnect d // Proc. IEEE Int. Conf. on Cluster Computing. New York: IEEE Press, 2018. 646–654.
59. *Ajima Y. et al.* Tofu: Interconnect for the K computer // Fujitsu Sci. Tech. J. 2012. **48**, N 3. 280–285.
60. Aries network on Theta. Argonne leadership computing facility. https://www.alcf.anl.gov/support-center/theta/aries-network-theta.
61. *Parker S., Chunduri S., Harms K., Kandalla K.* Performance evaluation of MPI on Cray XC40 Xeon Phi systems. 2018. https://cug.org/proceedings/cug2018_proceedings/includes/files/pap131s2-file1.pdf.
62. *Harms K., Leggett T., Allen B., Coghlan S., Fahey M., Holohan C., McPheeters G., Rich P.* Theta: rapid installation and acceptance of an XC40 KNL system // Concurrency and Computation: Practice and Experience. 2018. **30**, N 1. doi 10.1002/cpe.4336.
63. HPE Cray. https://buy.hpe.com/ru/ru/servers/cray-systems/cray-supercomputer/cray-supercomputer/hpe-cray-supercomputers/p/1012927320.
64. *Sensi D., Girolamo S., et al.* An in-depth analysis of the slingshot interconnect // SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. New York: IEEE Press, 2020. 481–494.
65. *Исмагилов Т.Ф., Семенов А.С., Симонов А.С.* Результаты оценочного тестирования отечественной высокоскоростной коммуникационной сети Ангара // Суперкомпьютерные дни в России. М.: Изд-во МГУ, 2016. 626–639.
66. *Симонов А.С., Жабин И.А., Куштанов Е.Р., Семенов А.С. и др.* Высокоскоростная сеть Ангара: архитектура и результаты применения // Вопросы кибербезопасности. 2019. № 4. 46–53.
67. Xilinx's Project Everest Looks Like Bad News for Intel. https://www.fool.com/investing/2018/04/17/xilinxs-project-everest-looks-like-bad-news-for-in.aspx.
68. Xilinx vs. Intel High-End FPGA Series Comparison. https://hardwarebee.com/xilinx-vs-intel-high-end-fpga-series-comparison.
69. *Craven S., Athanas P.* Examining the viability of FPGA supercomputing // EURASIP Journal on Embedded systems. 2007. doi 10.1155/2007/93652.
70. Specifications: Supercomputer Fugaku. https://www.fujitsu.com/global/about/innovation/fugaku/specifications.
71. *Ajima Y., Sumimoto S., Shimizu T.* Tofu: A 6D mesh/torus interconnect for exascale computers // Computer. 2009. **42**, N 11. 36–40.
72. FEFS: Scalable Cluster File System. https://www.fujitsu.com/downloads/TC/sc11/fefs-sc11.pdf.
73. Next Generation File System Design. http://oss-tsukuba.org/wp-content/uploads/2018/09/2018-GFarmWS-Fujitsu.pdf.
74. About Fugaku: RIKEN Center for Computational Science. https://www.r-ccs.riken.jp/en/fugaku/about.
75. Fact Sheet: Collaboration of Oak Ridge, Argonne, and Livermore (CORAL). https://www.energy.gov/downloads/fact-sheet-collaboration-oak-ridge-argonne-and-livermore-coral.
76. Summit: Oak Ridge Leadership Computing Facility. https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit.
77. Sierra: High Performance Computing. https://hpc.llnl.gov/hardware/platforms/sierra.
78. *Fu H., Liao J., Yang J., et al.* The Sunway TaihuLight supercomputer: system and applications // Sci. China Inf. Sci. 2016. **59**, N 7. doi 10.1007/s11432-016-5588-7.
79. AI of the Storm: How We Built the Most Powerful Industrial Computer in the U.S. in Three Weeks During a Pandemic. https://blogs.nvidia.com/blog/2020/08/14/making-selene-pandemic-ai.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

80. Role of the New Machine: Amid Shutdown, NVIDIA's Selene Supercomputer Busier Than Ever. https://blogs.nvidia.com/blog/2020/12/18/nvidia-selene-busy.
81. Forschungszentrum Jülich: JUWELS. https://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUWELS/JUWELS_node.html.
82. HPC5: the supercomputer working for energy. https://www.eni.com/en-IT/operations/green-data-center-hpc5.html.
83. Frontera. https://frontera-portal.tacc.utexas.edu.
84. Texas Advanced Computing Center: TACC LAUNCHES EXPANDED FRONTERA SUPERCOMPUTER TO SUPPORT URGENT COMPUTING. https://www.tacc.utexas.edu/-/tacc-launches-expanded-frontera-supercomputer-to-support-urgent-computing.
85. Aramco and STC unveil Dammam 7 Supercomputer. https://www.aramco.com/en/news-media/news/2021/aramco-and-stc-unveil-dammam-7-supercomputer.
86. NVIDIA DGX SuperPOD for Enterprise. https://www.nvidia.com/en-us/data-center/dgx-superpod.
87. #1 Green500 Supercomputer Delivers the World's Best Performance-Per-Watt | Supermicro. https://www.supermicro.com/en/success-story/green500-pfn-number1.
88. Preferred Networks' MN-3 Supercomputer. https://www.preferred.jp/en/news/pr20201117.
89. SberCloud: Christofari. https://sbercloud.ru/ru/christofari.
90. Суперкомпьютер "Кристофари" первым в РФ аттестовали для работы с персональными данными. https://tass.ru/ekonomika/8121173.
91. *Voevodin V., Antonov A., Nikitenko D., Shvets P., Sobolev S., Stefanov K., Voevodin Vad., Zhumatiy S., Brechalov A., Naumov A.* Lomonosov-2: Petascale supercomputing at Lomonosov Moscow State University // Contemporary High Performance Computing: from Petascale toward Exascale. Vol. 3. Boca Raton: CRC Press, 2019. 305–330.
92. *Voevodin V.V., Antonov A.S., Nikitenko D.A., Shvets P.A., Sobolev S.I., Sidorov I.Yu., Stefanov K.S., Voevodin Vad.V., Zhumatiy S.A.* Supercomputer Lomonosov-2: large scale, deep monitoring and fine analytics for the user community // Supercomputing Frontiers and Innovations. 2019. **6**, N 2. 4–11. doi 10.14529/jsfi190201.
93. PARALLEL.RU. Суперкомпьютер "ЛОМОНОСОВ-2" | PARALLEL.RU — Информационно-аналитический центр по параллельным вычислениям. https://parallel.ru/cluster/lomonosov2.html.
94. Blue Waters User Portal. https://bluewaters.ncsa.illinois.edu.
95. *Shaw D.E., Grossman J.P., Bank J.A., Batson B., Butts J.A., Chao J.C., Deneroff M.M., Dror R.O., Even A., et al.* Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer // SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis. Piscataway: IEEE Press, 2014. 41–53.
96. ARM's secret recipe for power efficient processing. https://www.androidauthority.com/arms-secret-recipe-for-power-efficient-processing-409850.
97. EPYC: A Study in Energy Efficient CPU Design. https://www.amd.com/system/files/documents/The-Energy-Efficient-AMD-EPYC-Design.pdf.
98. Arm Supercomputer Captures The Energy Efficiency Crown. https://www.nextplatform.com/2019/11/22/arm-supercomputer-captures-the-energy-efficiency-crown.
99. *Barker K.J., Davis K., Hoisie A., Kerbyson D., Lang M., Pakin S., Sancho J.* Entering the petaflop era: the architecture and performance of roadrunner // Proc. 2008 ACM/IEEE Conference on Supercomputing. Austin: IEEE Press, 2008. 1–11. doi 10.1109/SC.2008.5217926.
100. *Thibodeau P.* Scientists, IT community await exascale computers. 2009. https://www.computerworld.com/article/2550451/scientists--it-community-await-exascale-computers.html.
101. Folding@Home Network Breaks the ExaFLOP Barrier in Fight Against Coronavirus. https://www.tomshardware.com/news/folding-at-home-breaks-exaflop-barrier-fight-coronavirus-covid-19.
102. *Thibodeau P.* Obama sets $126M for next-gen supercomputing. 2011. https://www.computerworld.com/article/2513219/obama-sets--126m-for-next-gen-supercomputing.html.
103. Frontier. https://www.olcf.ornl.gov/frontier.
104. Aurora: Argonne Leadership Computing Facility. https://alcf.anl.gov/aurora.
105. *Zarley B.D.* America's first exascale supercomputer to be built by 2021. https://www.theverge.com/2019/3/18/18271328/supercomputer-build-date-exascale-intel-argonne-national-laboratory-energy.
106. HPE Slingshot Interconnect: High Performance Network for HPE Cray Supercomputers. https://www.hpe.com/us/en/compute/hpc/slingshot-interconnect.html.
107. El Capitan Supercomputer at Lawrence Livermore National Lab. https://www.hpe.com/us/en/compute/hpc/cray/doe-el-capitan-press-release.html.
108. China fleshes out exascale design for Tianhe-3 supercomputer. https://www.nextplatform.com/2019/05/02/china-fleshes-out-exascale-design-for-tianhe-3.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

109. China launches exascale supercomputer prototype. http://www.xinhuanet.com/english/2018-08/06/c_137369865.htm.
110. China launches third prototype exascale computer. http://www.xinhuanet.com/english/2018-10/22/c_137550589.htm.
111. HPC AI500: A Benchmark Suite for HPC AI Systems. https://www.benchcouncil.org/HPCAI500.
112. Japan's Fugaku supercomputer. https://www.japantimes.co.jp/news/2021/01/07/business/tech/japans-fugaku-supercomputer.
113. CRESTA: Developing techniques and solutions. http://www.cresta-project.eu.
114. DEEP Projects. https://www.deep-projects.eu.
115. Home: Mont-Blanc. https://www.montblanc-project.eu.
116. Home: European High Performance Computer Joint Undertaking. https://eurohpc-ju.europa.eu.
117. LUMI Front Page. https://www.lumi-supercomputer.eu.
118. MareNostrum. https://www.bsc.es/marenostrum.
119. Leonardo pre-exascale supercomputer. https://www.cineca.it/en/hot-topics/Leonardo.
120. NVIDIA Extends Data Center Infrastructure Processing Roadmap with BlueField-3 | NVIDIA Newsroom. https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3.

Поступила в редакцию
6 апреля 2021

Принята к публикации
2 июня 2021

Информация об авторах

Александр Сергеевич Антонов — к.ф.-м.н., вед. научн. сотр., Московский государственный университет им. М. В. Ломоносова, Научно-исследовательский вычислительный центр, Ленинские горы, 1, стр. 4, 119991, Москва, Российская Федерация.

Илья Викторович Афанасьев — к.ф.-м.н., техник, Московский государственный университет им. М. В. Ломоносова, Научно-исследовательский вычислительный центр, Ленинские горы, 1, стр. 4, 119991, Москва, Российская Федерация.

Владимир Валентинович Воеводин — д.ф.-м.н., член.-корр. РАН, профессор, директор, Московский государственный университет им. М. В. Ломоносова, Научно-исследовательский вычислительный центр, Ленинские горы, 1, стр. 4, 119991, Москва, Российская Федерация.

References

1. Home: TOP500. https://www.top500.org. Cited April 25, 2021.
2. Graph 500: Large-Scale Benchmarks. https://graph500.org. Cited April 25, 2021.
3. Green500: TOP500. https://www.top500.org/lists/green500. Cited April 25, 2021.
4. HPCG: TOP500. https://www.top500.org/lists/hpcg. Cited April 25, 2021.
5. A. S. Antonov, D. A. Nikitenko, and V. V. Voevodin, "Algo500 — A New Approach to the Joint Analysis of Algorithms and Computers," Lobachevskii J. Math. **41** (8), 1435–1443 (2020). doi 10.1134/S1995080220080041.
6. Together We Are Powerful. Folding@home. https://foldingathome.org. Cited April 25, 2021.
7. Vl. V. Voevodin and A. P. Kapitonova, *Methods for Describing and Classifying Computing Systems Architectures* (Mosk. Gos. Univ., Moscow, 1994) [in Russian].
8. V. V. Voevodin and Vl. V. Voevodin, *The Parallel Computing* (BHV-Petersburg, St. Petersburg, 2002) [in Russian].
9. Classification of Computing Systems Architectures — PARALLEL.RU — Information and Analytical Center for Parallel Computing. https://parallel.ru/computers/taxonomy. Cited April 25, 2021.
10. M. J. Flynn, "Very High-Speed Computing Systems," Proc. IEEE **54** (12), 1901–1909 (1966). doi 10.1109/PROC.1966.5273.
11. M. J. Flynn, "Some Computer Organizations and Their Effectiveness," IEEE Trans. Comput. **21** (9), 948–960 (1972). doi 10.1109/TC.1972.5009071.
12. Home: European Processor Initiative. https://www.european-processor-initiative.eu. Cited April 25, 2021.
13. J. A. Fisher, P. Faraboschi, and C. Young, "VLIW Processors," in *Encyclopedia of Parallel Computing* (Springer, Boston, 2011), pp. 2135–2142. doi 10.1007/978-0-387-09766-4_471.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

14. I. V. Afanasyev, V. V. Voevodin, K. Komatsu, and H. Kobayashi, "VGL: A High-Performance Graph Processing Framework for the NEC SX-Aurora TSUBASA Vector Architecture," J. Supercomput. (2021). doi 10.1007/s11227-020-03564-9.
15. Technology Intel® Hyper-Threading. https://www.intel.ru/content/www/ru/ru/architecture-and-technology/hyper-threading/hyper-threading-technology.html. Cited April 25, 2021 [in Russian].
16. O. E. Okonta, D. Ajani, A. A. Owolabi, et al., "Performance Evaluation of Hyper Threading Technology Architecture Using Microsoft Operating System Platform," West Afr. J. Ind. Acad. Res. **15** (1), 52–67 (2015).
17. Intel Hyper Threading Performance with a Core i7 on Ubuntu 18.04 LTS. https://www.phoronix.com/scan.php?page=article&item=intel-ht-2018&num=4. Cited April 25, 2021.
18. What is the TensorFloat-32 Precision Format? — NVIDIA Blog. https://blogs.nvidia.com/blog/2020/05/14/tensorfloat-32-precision-format. Cited April 25, 2021.
19. AMD Chips Away at Intel in World's Top 500 Supercomputers as GPU War Looms. https://www.crn.com/news/components-peripherals/amd-chips-away-at-intel-in-world-s-top-500-supercomputers-as-gpu-war-looms. Cited April 25, 2021.
20. Intel Server Roadmap: 14nm Cooper Lake in 2019, 10nm Ice Lake in 2020. https://www.anandtech.com/show/13194/intel-shows-xeon-2018-2019-roadmap-cooper-lakesp-and-ice-lakesp-confirmed. Cited April 25, 2021.
21. Intel® Xeon® Gold 6248 Processor (27.5M Cache, 2.50 GHz) Product Specifications. https://ark.intel.com/content/www/us/en/ark/products/192446/intel-xeon-gold-6248-processor-27-5m-cache-2-50-ghz.html. Cited April 25, 2021.
22. EPYC 7742-AMD. https://en.wikichip.org/wiki/amd/epyc/7742. Cited April 25, 2021.
23. Epyc-Wikipedia. https://en.wikipedia.org/wiki/Epyc. Cited April 25, 2021.
24. IBM Power9-WikiChip. https://en.wikichip.org/wiki/ibm/microarchitectures/power9/#Memory_Hierarchy. Cited April 25, 2021.
25. FUJITSU Processor A64FX. https://www.fujitsu.com/downloads/SUPER/a64fx/a64fx_datasheet.pdf. Cited April 25, 2021.
26. Baikal-M—Baikal Electronics—WikiChip. https://en.wikichip.org/wiki/baikal/baikal-m. Cited April 25, 2021.
27. Y.-C. Wang, J.-K. Chen, B.-R. Li, "An Empirical Study of HPC Workloads on Huawei Kunpeng 916 Processor," in *Proc. IEEE 25th Int. Conf. on Parallel and Distributed Systems, Tianjin, China, December 4—6, 2019* (IEEE Press, New York, 2019), pp. 360–367, doi 10.1109/ICPADS47876.2019.00057.
28. Vulcan–Microarchitectures–Cavium–WikiChip. https://en.wikichip.org/wiki/cavium/microarchitectures/vulcan. Cited April 25, 2021.
29. E. Calore, A. Gabbana, S. F. Schifano, and R. Tripiccione, "ThunderX2 Performance and Energy-Efficiency for HPC Workloads," Computation **8** (1), 1–17 (2020). doi 10.3390/computation8010020.
30. Hot Chips 2020: Marvell Details ThunderX3 CPUs–Up to 60 Cores Per Die, 96 Dual-Die in 2021. https://www.anandtech.com/show/15995/hot-chips-2020-marvell-details-thunderx3. Cited April 25, 2021.
31. eMAG–Ampere–WikiChip. https://en.wikichip.org/wiki/ampere_computing/emag. Cited April 25, 2021.
32. D. Alfonso, R. Demenko, A. Kozhin, et al., "Eight-Core 'Elbrus-8C' Processor Microarchitecture," Voprosy Radioelektron., No. 3, 6–13 (2016).
33. Chinese Processor-Supercomputer Path. https://www.osp.ru/os/2017/01/13051592. Cited April 25, 2021 [in Russian].
34. SPARC64 XIfx: Fujitsu's Next Generation Processor for HPC. https://www.fujitsu.com/global/Images/20140811hotchips26_tcm100-1156766.pdf. Cited April 25, 2021.
35. MN-Core: Preferred Networks. https://projects.preferred.jp/mn-core/en. Cited April 25, 2021.
36. Cerebras Wafer Scale Engine. https://www.ixbt.com/news/2019/08/20/cerebras-wafer-scale-engine-ipad-pro-1-tdp-15.html. Cited April 25, 2021 [in Russian].
37. First GPU on CDNA Architecture. https://www.hardwareluxx.ru/index.php/news/hardware/grafikkarten/50677-pervyj-gpu-na-arkhitekture-cdna-amd-predstavila-instinct-mi100.html. Cited April 25, 2021 [in Russian].
38. K. Komatsu, R. Egawa, Y. Isobe, et al., "An Approach to the Highest Efficiency of the HPCG Benchmark on the SX-ACE Supercomputer," in *Proc. Int. Conf. on High Performance Computing, Networking, Storage, and Analysis, Austin, USA, November 15–20, 2015*, http://sc15.supercomputing.org/sites/all/themes/SC15images/tech_poster/poster_files/post277s2-file3.pdf. Cited April 15, 2021.
39. K. Komatsu, S. Momose, Y. Isobe, et al., "Performance Evaluation of a Vector Supercomputer SX-Aurora TSUBASA," in *Proc. Int. Conf. for High Performance Computing, Networking, Storage, and Analysis, Dallas, USA, November 11–16, 2018* (IEEE Press, Piscataway, 2018), pp. 685–696. doi 10.1109/SC.2018.00057.
40. J. Dongarra, *Report on the Tianhe-2A System*, Tech Report No. ICL-UT-17-04 (Univ. Tennessee, Knoxville, 2017). https://www.dropbox.com/s/0jyh5qlgok73t1f/TH-2A-report.pdf?dl=0.
41. Hardwareluxx: The Third Generation of Google TPU. https://www.hardwareluxx.ru/index.php/news/hardware/prozessoren/44682-googles-tpu.html. Cited April 25, 2021 [in Russian].

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

42. System Architecture: Cloud TPU. https://cloud.google.com/tpu/docs/system-architecture. Cited April 25, 2021.
43. Google Cloud Blog. https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu. Cited April 25, 2021.
44. What Is a Data Processing Unit (DPU). https://www.forbes.com/sites/janakirammsv/2020/10/11/what-is-a-data-processing-unit-dpu-and-why-is-nvidia-betting-on-it. Cited April 25, 2021.
45. What Is a DPU? — NVIDIA Blog. https://blogs.nvidia.com/blog/2020/05/20/whats-a-dpu-data-processing-unit. Cited April 30, 2021.
46. Hardwareluxx: NVIDIA + Mellanox. https://www.hardwareluxx.ru/index.php/news/hardware/grafikkarten/50439-nvidia-mellanox-konvergentsiya-tekhnologij-v-budushchikh-dpu.html. Cited April 25, 2021 [in Russian].
47. NVIDIA BLUEFIELD-2 DPU. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-2-dpu.pdf. Cited April 25, 2021.
48. INTUIT: Lecture. https://intuit.ru/studies/courses/1156/190/lecture/4942?page=4. Cited April 25, 2021 [in Russian].
49. NUMA Deep Dive. Part 1. https://frankdenneman.nl/2016/07/07/numa-deep-dive-part-1-uma-numa. Cited April 25, 2021.
50. D. Kanter, "The Common System Interface: Intel's Future Interconnect," Real World Technologies (2007). https://www.realworldtech.com/common-system-interface/. Cited April 25, 2021.
51. T. Wang, Z. Su, Y. Xia, et al., "NovaCube: A Low Latency Torus-Based Network Architecture for Data Centers," in *Proc. Global Communications Conf., Austin, USA, December 8—12, 2014* (IEEE Press, New York, 2014), pp. 2252–2257, doi 10.1109/GLOCOM.2014.7037143.
52. Fat Tree — Википедия. https://ru.wikipedia.org/wiki/Fat_Tree. Cited April 25, 2021 [in Russian].
53. X. Yuan, "On Nonblocking Folded-Clos Networks in Computer Communication Environments," in *IEEE International Parallel & Distributed Processing Symposium.* (IEEE Press, Anchorage, USA, 2011), pp. 188–196, doi 10.1109/IPDPS.2011.27.
54. Butterfly Network: Wikipedia. https://ru.qaz.wiki/wiki/Butterfly_network. Cited April 25, 2021 [in Russian].
55. J. Kim, W. J. Dally, and D. Abts, "Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks," ACM SIGARCH Computer Architecture News **35** (2) (2007). doi 10.1145/1273440.1250679.
56. Intel® Omni-Path Architecture Performance Tested for HPC. https://www.intel.ru/content/www/ru/ru/high-performance-computing-fabrics/omni-path-architecture-performance-overview.html. Cited April 25, 2021.
57. Low-Latency Ethernet Solutions for High-Performance Computing. https://www.cisco.com/c/dam/en_us/solutions/industries/docs/education/ethernet-solutions-high-performance-computing-education.pdf. Cited April 25, 2021.
58. Y. Ajima, T. Kawashima, T. Okamoto, et al., "The Tofu Interconnect D," in *Proc. IEEE Int. Conf. on Cluster Computing, Belfast, UK, September 10–13, 2018* (IEEE Press, New York, 2018), pp. 646–654, doi 10.1109/CLUSTER.2018.00090.
59. Y. Ajima, T. Inoue, S. Hiramoto, and T. Shimizu, "Tofu: Interconnect for the K Computer," Fujitsu Sci. Tech. J. **48** (3), 280—285 (2012).
60. Aries Network on Theta — Argonne Leadership Computing Facility. https://www.alcf.anl.gov/support-center/theta/aries-network-theta. Cited April 25, 2021.
61. S. Parker, S. Chunduri, K. Harms, and K. Kandalla, "Performance Evaluation of MPI on Cray XC40 Xeon Phi Systems," https://cug.org/proceedings/cug2018_proceedings/includes/files/pap131s2-file1.pdf. Cited April 25, 2021.
62. K. Harms, T. Leggett, B. Allen, et al., "Theta: Rapid Installation and Acceptance of an XC40 KNL System," Concurr. Comput. **30** (1) (2018). doi 10.1002/cpe.4336.
63. HPE Cray. https://buy.hpe.com/ru/ru/servers/cray-systems/cray-supercomputer/cray-supercomputer/hpe-cray-supercomputers/p/1012927320. Cited April 25, 2021 [in Russian].
64. D. Sensi, S. Girolamo, K. H. McMahon, et al., "An In-Depth Analysis of the Slingshot Interconnect," in *Proc. Int. Conf. for High Performance Computing, Networking, Storage and Analysis, Atlanta, USA, November 9–19, 2020* (IEEE Press, New York, 2020), pp. 481–494, doi 10.1109/SC41405.2020.00039.
65. T. F. Ismagilov, A. S. Semyonov, and A. S. Simonov, "Results of Evaluation Testing of the Angara Domestic High-Speed Communication Network," in *Russian Supercomputing Days* (Mosk. Gos. Univ., Moscow, 2016), pp. 626–639.
66. A. Simonov, I. Zhabin, E. Kushtanov, et al., "Angara Interconnect: Architecture and Performance Results," Voprosy Kiberbezopasn., No. 4, 46–53 (2019).
67. Xilinx's Project Everest Looks Like Bad News for Intel. https://www.fool.com/investing/2018/04/17/xilinxs-project-everest-looks-like-bad-news-for-in.aspx. Cited April 25, 2021.
68. Xilinx vs. Intel High-End FPGA Series Comparison. https://hardwarebee.com/xilinx-vs-intel-high-end-fpga-series-comparison. Cited April 25, 2021.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

69. S. Craven and P. Athanas, "Examining the Viability of FPGA Supercomputing," EURASIP J. Embed. Syst. (2007), doi 10.1155/2007/93652.
70. Specifications: Supercomputer Fugaku. https://www.fujitsu.com/global/about/innovation/fugaku/specifications. Cited April 25, 2021.
71. Y. Ajima, S. Sumimoto, and T. Shimizu, "Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers," Computer **42** (11), 36–40 (2009). doi: 10.1109/MC.2009.370.
72. FEFS: Scalable Cluster File System. https://www.fujitsu.com/downloads/TC/sc11/fefs-sc11.pdf. Cited April 25, 2021.
73. Next Generation File System Design. http://oss-tsukuba.org/wp-content/uploads/2018/09/2018-GFarmWS-Fujitsu.pdf. Cited April 25, 2021.
74. About Fugaku: RIKEN Center for Computational Science. https://www.r-ccs.riken.jp/en/fugaku/about. Cited April 25, 2021.
75. Fact Sheet: Collaboration of Oak Ridge, Argonne, and Livermore (CORAL). https://www.energy.gov/downloads/fact-sheet-collaboration-oak-ridge-argonne-and-livermore-coral. Cited April 25, 2021.
76. Summit: Oak Ridge Leadership Computing Facility. https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit. Cited April 25, 2021.
77. Sierra: High Performance Computing. https://hpc.llnl.gov/hardware/platforms/sierra. Cited April 25, 2021.
78. H. Fu, J. Liao, J. Yang, et al., "The Sunway TaihuLight Supercomputer: System and Applications," Sci. China Inf. Sci. **59** (2016). doi 10.1007/s11432-016-5588-7.
79. AI of the Storm: How We Built the Most Powerful Industrial Computer in the U.S. in Three Weeks During a Pandemic. https://blogs.nvidia.com/blog/2020/08/14/making-selene-pandemic-ai. Cited April 25, 2021.
80. Role of the New Machine: Amid Shutdown, NVIDIA's Selene Supercomputer Busier Than Ever. https://blogs.nvidia.com/blog/2020/12/18/nvidia-selene-busy. Cited April 25, 2021.
81. Forschungszentrum Jülich: JUWELS. https://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUWELS/JUWELS_node.html. Cited April 25, 2021.
82. HPC5: the Supercomputer Working for Energy. https://www.eni.com/en-IT/operations/green-data-center-hpc5.html. Cited April 25, 2021.
83. FRONTERA. https://frontera-portal.tacc.utexas.edu. Cited April 25, 2021.
84. Texas Advanced Computing Center: TACC LAUNCHES EXPANDED FRONTERA SUPERCOMPUTER TO SUPPORT URGENT COMPUTING. https://www.tacc.utexas.edu/-/tacc-launches-expanded-frontera-supercomputer-to-support-urgent-computing. Cited April 25, 2021.
85. Aramco and STC Unveil Dammam 7 Supercomputer. https://www.aramco.com/en/news-media/news/2021/aramco-and-stc-unveil-dammam-7-supercomputer. Cited April 25, 2021.
86. NVIDIA DGX SuperPOD for Enterprise. https://www.nvidia.com/en-us/data-center/dgx-superpod. Cited April 25, 2021.
87. #1 Green500 Supercomputer Delivers the World's Best Performance-Per-Watt. https://www.supermicro.com/en/success-story/green500-pfn-number1. Cited April 25, 2021.
88. Preferred Networks' MN-3 Supercomputer. https://www.preferred.jp/en/news/pr20201117. Cited April 25, 2021.
89. SberCloud: Christofari. https://sbercloud.ru/ru/christofari. Cited April 25, 2021 [in Russian].
90. The Christofari Supercomputer. https://tass.ru/ekonomika/8121173. Cited April 25, 2021 [in Russian].
91. V. Voevodin, A. Antonov, D. Nikitenko, et al., "Lomonosov-2: Petascale Supercomputing at Lomonosov Moscow State University," in *Contemporary High Performance Computing: from Petascale toward Exascale* (CRC Press, Boca Raton, 2019), Vol. 3, pp. 305–330.
92. V. V. Voevodin, A. S. Antonov, D. A. Nikitenko, et al., "Supercomputer Lomonosov-2: Large Scale, Deep Monitoring and Fine Analytics for the User Community," Supercomput. Front. Innov. **6** (2), 4–11 (2019). doi 10.14529/jsfi190201.
93. PARALLEL.RU: Supercomputer Lomonosov-2. https://parallel.ru/cluster/lomonosov2.html. Cited April 25, 2021 [in Russian].
94. Blue Waters User Portal. https://bluewaters.ncsa.illinois.edu. Cited April 25, 2021.
95. D. E. Shaw, J. P. Grossman, J. A. Bank, et al., "Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer," in *Proc. Int. Conf. for High Performance Computing, Networking, Storage, and Analysis, New Orleans, USA, November 16–21, 2014* (IEEE Press, Piscataway, 2014), pp. 41–53, doi 10.1109/SC.2014.9.
96. ARM's Secret Recipe for Power Efficient Processing. https://www.androidauthority.com/arms-secret-recipe-for-power-efficient-processing-409850. Cited April 25, 2021.
97. EPYC: A Study in Energy Efficient CPU Design. https://www.amd.com/system/files/documents/The-Energy-Efficient-AMD-EPYC-Design.pdf. Cited April 25, 2021.

2021, **22** (2), 138–181. DOI: 10.26089/NumMet.v22r210

98. Arm Supercomputer Captures The Energy Efficiency Crown. https://www.nextplatform.com/2019/11/22/arm-supercomputer-captures-the-energy-efficiency-crown. Cited April 25, 2021.
99. K. J. Barker, K. Davis, A. Hoisie, et al., "Entering the Petaflop Era: The Architecture and Performance of Roadrunner," in *Proc. 2008 ACM/IEEE Conference on Supercomputing, Austin, USA, November 15–21, 2008* (IEEE Press, Austin, 2008), pp. 1–11, doi 10.1109/SC.2008.5217926.
100. P. Thibodeau, "Scientists, IT Community Await Exascale Computers," (2009). https://www.computerworld.com/article/2550451/scientists--it-community-await-exascale-computers.html. Cited April 25, 2021.
101. Folding@Home Network Breaks the ExaFLOP Barrier in Fight Against Coronavirus. https://www.tomshardware.com/news/folding-at-home-breaks-exaflop-barrier-fight-coronavirus-covid-19. Cited April 25, 2021.
102. P. Thibodeau, "Obama Sets $126M for Next-Gen Supercomputing" (2011). https://www.computerworld.com/article/2513219/obama-sets--126m-for-next-gen-supercomputing.html. Cited April 25, 2021.
103. Frontier. https://www.olcf.ornl.gov/frontier. Cited April 25, 2021.
104. Aurora: Argonne Leadership Computing Facility. https://alcf.anl.gov/aurora. Cited April 25, 2021.
105. B. D. Zarley, "America's First Exascale Supercomputer to be Built by 2021," (2019). https://www.theverge.com/2019/3/18/18271328/supercomputer-build-date-exascale-intel-argonne-national-laboratory-energy. Cited April 25, 2021.
106. HPE Slingshot Interconnect: High Performance Network for HPE Cray Supercomputers. https://www.hpe.com/us/en/compute/hpc/slingshot-interconnect.html. Cited April 25, 2021.
107. El Capitan Supercomputer at Lawrence Livermore National Lab. https://www.hpe.com/us/en/com\protect\discretionary{\char\hyphenchar\font}{}{}pute/hpc/cray/doe-el-capitan-press-release.html. Cited April 25, 2021.
108. China Fleshes Out Exascale Design for Tianhe-3 Supercomputer. https://www.nextplatform.com/2019/05/02/china-fleshes-out-exascale-design-for-tianhe-3. Cited April 25, 2021.
109. China Launches Exascale Supercomputer Prototype. http://www.xinhuanet.com/english/2018-08/06/c_137369865.htm. Cited April 25, 2021.
110. China Launches Third Prototype Exascale Computer. http://www.xinhuanet.com/english/2018-10/22/c_137550589.htm. Cited April 25, 2021.
111. HPC AI500: A Benchmark Suite for HPC AI Systems. https://www.benchcouncil.org/HPCAI500. Cited April 25, 2021.
112. Japan's Fugaku Supercomputer. https://www.japantimes.co.jp/news/2021/01/07/business/tech/japans-fugaku-supercomputer. Cited April 25, 2021.
113. CRESTA: Developing Techniques and Solutions. http://www.cresta-project.eu. Cited April 25, 2021.
114. DEEP Projects. https://www.deep-projects.eu. Cited April 25, 2021.
115. Home: Mont-Blanc. https://www.montblanc-project.eu. Cited April 25, 2021.
116. Home: European High Performance Computer Joint Undertaking. https://eurohpc-ju.europa.eu. Cited April 25, 2021.
117. LUMI Front Page. https://www.lumi-supercomputer.eu. Cited April 25, 2021.
118. MareNostrum. https://www.bsc.es/marenostrum. Cited April 25, 2021.
119. Leonardo Pre-Exascale Supercomputer. https://www.cineca.it/en/hot-topics/Leonardo. Cited April 25, 2021.
120. NVIDIA Extends Data Center Infrastructure Processing Roadmap with BlueField-3 — NVIDIA Newsroom. https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3. Cited April 25, 2021.

Received
April 6, 2021

Accepted for publication
June 2, 2021

Information about the authors

Alexander S. Antonov — PhD., Leading Scientist, Lomonosov Moscow State University, Research Computing Center, Leninskie Gory, 1, building 4, 119991, Moscow, Russia.

Ilya V. Afanasiev — PhD., technician, Lomonosov Moscow State University, Research Computing Center, Leninskie Gory, 1, building 4, 119991, Moscow, Russia.

Vladimir V. Voevodin — Dr. Sci., Professor, Corresponding Member of Russian Academy of Sciences, Director, Lomonosov Moscow State University, Research Computing Center, Leninskie Gory, 1, building 4, 119991, Moscow, Russia.