УДК 004.382.2 DOI: 10.14529/cmse190305

ТЕНДЕНЦИИ РАЗВИТИЯ ВЫЧИСЛИТЕЛЬНЫХ УЗЛОВ СОВРЕМЕННЫХ СУПЕРКОМПЬЮТЕРОВ*

© 2019 Е.О. Тютляева[1], И.О. Одинцов[2], А.А. Московский[1], Г.В. Мармузов[1]

[1] *ЗАО «РСК Технологии»*

(121170 Москва, пр. Кутузовский, д. 36, стр. 23),

[2] *ООО «РСК Лабс»*

(121205 Москва, ул. Большой Бульвар, д. 42, стр. 1)

E-mail: xgl@rsc-tech.ru, igor_ odintsov@rsc-tech.ru, moskov@rsc-tech.ru, gleb.marmuzov@rsc-tech.ru

Поступила в редакцию: 28.01.2019

В данной работе выполнен анализ вычислительных узлов современных суперкомпьютеров с двух точек зрения — аппаратно-компонентной и инфраструктурной. На основании проведённого анализа названы основные конструктивные элементы, которыми должен быть оснащен современный вычислительный узел. В статье приведены классификации архитектур современных универсальных и специализированных ядер с примерами; проведен обзор современных тенденций организации подсистемы памяти и внутриузлового интерконнекта; упомянуты способы использования энергонезависимых устройств хранения на узлах при организации современных высокопроизводительных систем хранения. Также разобраны основные требования к организации инфраструктуры узла современного суперкомпьютера, в частности, дана краткая классификация современных подходов к организации жидкостного охлаждения и мониторинга вычислительных узлов. Выявленные тенденции приводят к основным вариантам дизайна вычислительных узлов, состоящих из энергоэффективного универсального процессора и совокупности энергоэффективных специализированных ускорителей. В статье сделан акцент на современных технологиях, которые достигли стадии выхода в производство или, как минимум, создания рабочих прототипов. Обсуждаются современные суперкомпьютерные задачи и их отображение на архитектуру вычислительных узлов. В заключении приведено кратное обсуждение актуальных технологических проблем и основных направлений для сохранения прогресса в компьютерной отрасли.

Ключевые слова: высокопроизводительный вычислительный узел, разработка архитектуры вычислительного узла, анализ суперкомпьютерных архитектур, тенденции развития суперкомпьютеров.

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Тютляева Е.О., Одинцов И.О., Московский А.А., Мармузов Г.В. Тенденции развития вычислительных узлов современных суперкомпьютеров // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2019. Т. 8, № 3. С. 92–114. DOI: 10.14529/cmse190305.

Введение

При создании вычислительного узла для современного суперкомпьютера необходимо использование новых технологических подходов.

Простое масштабирование существующих технологий не будет эффективным решением. Задача, стоящая перед исследователями, гораздо глубже и обширнее, чем простое достижение заданной вычислительной мощности, она требует ряда инновационных решений. Можно сформулировать ряд основных задач, которые стоят перед разработчиками лидирующих суперкомпьютеров:

*Статья рекомендована к публикации программным комитетом Международной научной конференции «Параллельные вычислительные технологии (ПаВТ) 2019»

Минимизация энергопотребления. При масштабировании текущих технологий может быть получен суперкомпьютер, обладающий энергопотреблением в сотни мегаватт. Необходимо разработать новые технологии, которые позволят уменьшить планируемое энергопотребление до экономически приемлемого уровня (20-40 MW).

Обеспечение модульности и высокой вычислительной плотности в рамках одного узла. Высокоскоростные интерфейсы в рамках одного узла.

Организация энергоэффективного межузлового интерконнекта с высокой пропускной способностью и минимальными задержками.

Организация *высокопроизводительной системы хранения* с достаточно высокой пропускной способностью для избежать простоев вычислительных ресурсов.

Подбор оптимальных типов памяти для построения наиболее эффективной структуры памяти узла.

Ряд вызовов связан и с программным обеспечением (ПО), где мы выделяем следующие важные аспекты:

Сложность создания массового параллельного программного обеспечения на современных языках программирования фактически поднимает вопрос о необходимости *новой парадигмы программирования*: гипер-параллельной.

Множество различных высокопроизводительных аппаратных архитектур поднимает задачу не просто переносимости программного обеспечения, а *максимально эффективной переносимости и автоматизированного отображения на архитектуру.*

Также, во многом от программной реализации зависят *надежность и отказоустойчивость* программного обеспечения.

При разработке современного узла важно отразить все ожидаемые архитектурные особенности, такие как количество процессоров, количество процессорных ядер, соотношение объема памяти и объема вычислений в узле, количество независимых вычислительных потоков [1].

Текст статьи организован следующим образом. Раздел 1 содержит классификации современных универсальных и специализированных ядер, а также описывает подходы к отображению задач на вычислительные ядра. В разделе 2 сформулированы основные конструктивные элементы, которыми должен быть оснащен современный вычислительный узел. Далее в разделе 3 разобраны основные требования к организации инфраструктуры узла современного суперкомпьютера. В разделе 4 предложены обобщенные варианты отображения задач на архитектуру вычислительных узлов. В заключении подведены итоги проведённого обзора и приведено обсуждение актуальных технологических вызовов.

1. Задачи и вычислительные ядра

Д. Рид и Д. Донгарра в статье [2] выделяют две основные экосистемы: обработка данных и вычислительная наука. Авторы подчеркивают, что эти в настоящее время эти экосистемы имеют различия как на аппаратном уровне, так и на уровне программного обеспечения. Тем не менее, практически одинаковый аппаратный уровень может быть использован как для обработки данных, так и для вычислений, а существенная разница наблюдается лишь на уровне инструментального и промежуточного программного обеспечения для ряда задач машинного обучения и глубокого обучения. Дополнительно заметим, что если ранее научные и инженерные вычисления относили к «классическим задачам HPC» и они составляли основной объем задач, то прогнозируется, что в будущем

они будут составлять около половины задач, а оставшийся объем займут задачи обработки больших данных, искусственного интеллекта, машинного обучения и глубокого обучения.

С точки зрения аппаратного уровня более корректной будет классификация отображения задач на вычислительные ядра, а также их количество с учетом масштабируемости задач на вычислительные нити.

В предлагаемом исследовании выделим два базовых класса архитектур ядер: универсальные и ускорители, а также множество их подклассов, которые очень часто определяются конкретным производителем.

Универсальные ядра (типичная роль — хост, базовые, «классические вычислительные», «толстые») можно классифицировать следующим образом:

- традиционный набор команд (CISC):
  - x86, IA-64, x86-64 (Intel, AMD);
- упрощенный набор команд (RISC):
  - POWER (IBM);
  - ARM (много производителей);
- сверхдлинное командное слово (VLIW):
  - Эльбрус (МЦСТ).

Типов и классов специализированных ядер достаточно много. С одной стороны, обратим внимание на то, что математические сопроцессоры и сопроцессоры ввода-вывода остались историей и их функциональность реализуется в универсальных ядрах. С другой стороны, создаются новые перспективные ускорители, например, квантовые. Немаловажен тот факт, что специализированные ядра кроме их эффективности на определенном классе задач, являются также и энергоэффективными.

Специализированные ядра (типичная роль — ускоритель, сопроцессор, «тонкие») могут быть классифицированы следующим образом:

- гомогенные (на базе упрощенных универсальных), например Intel Xeon Phi [3] (Intel);
- вычислительные (графические) ускорители (GPU, GPGPU), например Tesla [4] (NVIDIA);
- узкоспециализированные (оптическое быстрое преобразование Фурье);
- тензорные ускорители (матричное умножение и свертка), например TPU [5] (Google);
- нейроморфные (самообучающиеся), например Loihi [6] (Intel);
- ускорители алгоритмов работы машинного зрения, например Movidius (Intel);
- квантовые (криптография, искусственный интеллект, молекулярное моделирование), например Tangle Lake [7] (Intel);
- перепрограммируемые (FPGA), например Intel Stratix 10 SX. FPGA [8].

Подчеркнем тот факт, что многие задачи считаются как на универсальных ядрах, так и на специализированных (в том числе и упрощенных универсальных). Конечно, важен такой критерий как эффективность, в котором интересную роль играет точность вычислений:

- двойная точность — необходима в тех сферах, где появление ошибок является недопустимым (большинство научных задач, инжиниринговых задач, ...);
- одинарная точность — допустима для задач симуляции, игровой физики;
- половинная точность — используется для глубокого обучения.

Таким образом, есть несколько основных аспектов, общих для всех перечисленных научных и инженерных направлений. Прежде всего, это широкий спектр временных и пространственных масштабов; и сложные, нелинейные пересечения множества

биологических и физических процессов. Это все требует качественного вычислительного моделирования, над которыми должны работать объединенные исследовательские группы из нескольких научных областей. Также важно учитывать существование огромных объемов разнообразных научных данных и беспрецедентные возможности для выявления междисциплинарных корреляций и статистического анализа. Во всех областях, от биологии до бизнеса, большие данные создают новые исследовательские возможности и предъявляют новые требования. Существуют оценки, например, в отчете DOE [9], что приблизительные вычислительные требования ряда задач к 2025 году возрастут в 100–1000 раз.

2. Компоненты узла суперкомпьютера

В данной статье мы сфокусируемся на основном конструктивном элементе — современном вычислительном узле.

Каждый узел должен быть оснащен следующими конструктивными элементами:

Процессоры с универсальными ядрами. Данные процессоры должны взять на себя основную нагрузку при решении вычислительных задач. Эти ядра также должны обеспечивать достаточную производительность для фрагментов кода, не применимых для расчетов на специальных ускорителях.

Ускорители со специализированными ядрами. Ускорители, в первую очередь, должны быть ориентированы на эффективное решение задач анализа данных; на вычислительных задачах должны использоваться для проведения специализированных вычислений. Кроме того, применение специализированных ускорителей должно способствовать достижению необходимой производительности (для решения перечисленных выше задач, стоящих на переднем крае науки) при этом уложившись в разумные границы энергопотребления.

Материнская плата и другие платы.

ОЗУ. В связи с возросшими объемами данных и актуальностью задач по обработке, классификации и анализу больших данных, к памяти на узле выдвигаются достаточно серьезные требования, касающиеся объема и пропускной способности. Возможно использование иерархических решений с высокопроизводительной оперативной памятью, дополненной Software-defined memory на базе NVMe (спецификация на протоколы доступа к твердотельным накопителям (SSD), подключенным по шине PCI Express).

Энергонезависимая память. Для организации уровня локального хранения на узле — может использоваться как кэш к системе хранения или элемент гиперконвергентной инфраструктуры.

Внутриузловой (внутренний) интерконнект. Один из ключевых элементов, обеспечивающих эффективность совместной работы всех вычислительных компонентов узла.

Высокопроизводительная фабрика, для обеспечения высокой пропускной способности при построении суперкомпьютера.

Контроллеры, обеспечивающие эффективное управление элементами инженерной инфраструктуры.

Простейшая классификация узлов может выглядеть так:

– гомогенный узел — узел, состоящий из однотипных ядер (как правило, универсальных);
– гибридный узел — узел, состоящий из универсальных ядер и ядер-ускорителей;

– гиперконвергентный узел — узел, полностью интегрирующий уровень хранения с уровнем обработки.

2.1. Универсальные ядра

При выборе процессоров с крупными ядрами следует рассматривать зарекомендовавших себя производителей, выпускающих процессоры общего назначения с крупными ядрами.

В настоящее время можно ориентироваться на данные рейтинга ТОП-500 и решения, разрабатываемые при создании прототипов суперкомпьютеров класса экзаскейл. Рассмотрим наиболее показательные архитектуры:

По данным статистики [10], 46,6% систем в рейтинге TOP-500 за ноябрь 2018-го года построены на базе *архитектур с традиционным набором команд (CISC)* — Intel Xeon E5 (Broadwell). Второе место занимают Xeon Gold (19,8%) и третье — Intel Xeon E5 (Haswell) (14,2%). Также архитектуру x86 планируется использовать для pre-exascale суперкомпьютера Frontera. Согласно заявлению директора TACC Dan Stanzione [11] выбор был остановлен на будущих процессорах Intel Xeon SP Platinum (Cascade Lake).

Процессоры с упрощенным набором команд (RISC) используются в суперкомпьютерах Summit и Sierra, которые по данным рейтинга ТОП-500 на ноябрь 2018-го года занимают первое и второе места в мире соответственно. Они базируются на процессорах IBM POWER9. IBM POWER9 поддерживает самые передовые технологии внутриузлового интерконнекта, включая NVIDIA NVLink, OpenCAPI и PCIe Gen4. Каждый узел суперкомпьютера Summit оснащен 2-мя процессорами IBM POWER9 и шестью ускорителями NVIDIA Tesla V100. Кроме этого, суперкомпьютер Sunway TaihuLight, разработанный в Китае, по данным рейтинга ТОП-500 на ноябрь 2018-го года занимает третье место в мире и базируется на процессорах Sunway SW26010. Это разработанные в Китае процессоры на базе 64-битной архитектуры RISC с технологией 28 нм.

Если говорить о универсальных ядрах с архитектурой ARM, то здесь следует упомянуть европейский проект ExaNode [12] в рамках которого проектируется многоядерный процессор общего назначения на базе ARMv8 CPU, плюс набор ускорителей FPGA. Кроме этого, согласно открытым источникам, японский проект построения экзафлопсного суперкомпьютера Post-K [13] также базируется на ARM-v8. Процессор Post-K — это вариант архитектуры ARMv8-A, но с 512-битным SVE векторным расширением с добавленным набором математических инструкций. Наконец, самый большой суперкомпьютер на базе архитектуры ARM — это Astra [14]. Производительность суперкомпьютера Astra достигает 2,322 PFLOPS с двойной точностью. Каждый узел базируется на двух процессорах ARM Cavium ThunderX2, без дополнительных ускорителей.

На данный момент выбор представленных на рынке архитектур и моделей процессоров с универсальными ядрами очень широк. Поэтому следует тщательно подходить к определению ключевых критериев и необходимых характеристик целевого вычислителя. К базовым характеристикам процессоров относят производительность и энергопотребление. Но следует также учитывать показатели надежности, совместимость с различными видами памяти, поддержку внутриузловых высокопроизводительных каналов передачи данных, совместимость со специализированными сопроцессорами, наличие прикладного программного стека (математических библиотек, оптимизированных для архитектуры прикладных библиотек, таких как BLAS, ScaLAPACK, FFTW, OpenFOAM

и т.п.), количество сокетов на плате и т.п. Все перечисленные аспекты оказывают существенное влияние на итоговую производительность целевого суперкомпьютера и удобство эксплуатации.

2.2. Специализированные ядра

Специализированные ядра, такие как ускорители и сопроцессоры чаще всего имеют узкую специфику, но за счет этого позволяют получить рекордное соотношение производительность/энергоэффективность при решении целевых задач.

Математические ускорители, как правило, интегрированы в основной процессор, остальные добавляются в вычислительный узел при помощи высокоскоростных каналов данных.

Если анализировать мировые тенденции, то лидирующее место среди ускорителей занимают GPGPU. Так по данным статистики [10], 12,8% систем в рейтинге TOP-500 за ноябрь 2018-го года используют ускорители NVIDIA Pascal — это первое место в статистике ускорителей и со-процессоров. Второе место (9,2%) и третье (2,6%) занимают системы, использующие соответственно NVIDIA Volta и NVIDIA Kepler. На четвертом месте появляются системы на базе Intel Xeon Phi (6 систем, 1,2%). Pre-Exascale суперкомпьютер Summit, занимающий лидирующую позицию в рейтинге TOP-500 за ноябрь 2018 также использует GPGPU ускорители. Каждый вычислительный узел оснащен шестью NVIDIA Tesla V100.

Другим направлением при ускорении вычислений является использование реконфигурируемой логики. В отличие от графических ускорителей, реконфигурируемые ускорители FPGA — это многоцелевые вычислительные устройства. Отличительными свойствами FPGA являются высокая пропускная способность ввода-вывода, и гибкий настраиваемый мелкозернистый параллелизм.

В настоящее время доступен широкий спектр решений на базе FPGA, например Xilinx 7-Series FPGAs [15] или программируемая вентильная матрица Intel Stratix 10 SX FPGA, выпущенная в 2018-м году, позволяет получить производительность до 10 TFlops с одинарной точностью [8].

Наконец, необходимо постоянно отслеживать прогресс в области перспективных направлений разработки ускорителей. В частности, интерес вызывают оптические процессоры и криптоакселераторы. Например, в настоящее время созданы оптические процессоры, которые позволяют выполнять быстрое преобразование Фурье практически мгновенно [16].

Использование ускорителей может позволить добиться необходимой производительности целевого суперкомпьютера и уложиться в разумный энергетический бюджет. Выбор ускорителей прежде всего зависит от специфики задач, которые будут решаться на целевом суперкомпьютере. В настоящее время при создании топовых суперкомпьютеров наблюдается тенденция к разработке многоцелевых машин, которые используются для широкого спектра задач, включая задачи вычислительного моделирования в различных областях науки и задачи обработки больших объемов данных. Таким образом, при выборе ускорителей необходимо учитывать потенциальную возможность применения ускорителей выбранного класса для решения большинства целевых задач и наличие стека программного обеспечения, который позволит эффективно использовать полученный гетерогенный суперкомпьютер.

2.3. Память

2.3.1. Оперативная память (ОЗУ)

К оперативной памяти предъявляются очень высокие требования. Практически все консорциумы заявляют о необходимости использования высокоскоростных интерфейсов и, дополнительно к ним, энергонезависимой памяти.

Например, для решения задач экзафлопсного масштаба также потребуются значительные объемы оперативной памяти на узле. В статье разработчиков аппаратно–программной платформы «Эльбрус» [17] формулируются следующие требования: 5 ПБ оперативной памяти с пропускной способностью 4 ТВ/s. Согласно оценке Coral-2, размер оперативной памяти должен быть не менее 8 ПБ; только проведение стандартных тестов потребует 5 ПБ) [18].

Согласно докладу Al Gara (Intel Fellow, Data Center Group), для того, чтобы суперкомпьютер класса экзафлопс смог успешно решать как задачи вычислительного моделирования, так и задачи искусственного интеллекта, оперативная память системы должна удовлетворять следующим требованиям: Объем оперативной памяти равен 6–12 ПБ, пропускная способность оперативной памяти — 100–200 ПБ/с, а энергонезависимой памяти I/O — 10–100 ТБ/с [19].

В настоящее время можно выделить следующие подходы к организации оперативной памяти на узле:

DDR-SDRAM (Double Data Rate Synchronous Dynamic Random Access Memory — синхронная динамическая память с произвольным доступом и удвоенной скоростью передачи данных). Синхронная динамическая память с произвольным доступом и удвоенной скоростью передачи данных. Самая распространенная технология, поддерживается большинством процессоров. В настоящий момент лидирующий стандарт: DDR4. В следующем году ожидается выпуск стандарта DDR5 [20].

3D Stacked Memory — это технология трехмерного размещения памяти, которая позволяет интегрировать ОЗУ и логические блоки микропроцессора, тем самым существенно увеличивая пропускную способность [21]. Следует отметить, что плотность организации 3D Stacked Memory требует специального подхода к охлаждению — требуется либо специальная организация воздушных потоков, либо жидкостное охлаждение. Можно выделить следующие виды:*HBM (High Bandwidth Memory)* — память с высокой пропускной способностью). По конструкции HBM — это непланарная память с трехмерной конструкцией в виде куба или прямоугольный параллелепипед. В HBM несколько микросхем памяти сложены друг над другом, чтобы сформировать кубическую структуру. Благодаря этому снижается площадь, занимаемая чипами памяти, что делает возможным размещение ее в непосредственной близости к графическому процессору [22]. На текущий момент лидирующим является поколение HBM2. *HMC (Hybrid Memory Cube).* Предоставляет пропускную способность до 480 ГБ/с на устройство, но обладает лимитированным объемом — до 8 ГБ, согласно стандартам, определенным консорциумом [23].

SSD DIMM — SSD накопители с интерфейсом DIMM, также есть варианты 3dX Point накопителей с интерфейсом DIMM. Такое решение может использоваться в качестве дополнительного уровня ОЗУ. Достоинствами данного подхода являются большой объем. Узким местом все еще остается доступное количество циклов перезаписи.

SDM (Software Defined Memory) — программно определяемая память. При наличии соответствующих программно-аппаратных решений может быть организован дополнительный уровень ОЗУ на базе энергонезависимой памяти. Достоинствами данного подхода является большой объем, а к недостаткам можно отнести небольшую пропускную способность по сравнению с DDR. Примером может являться технология IMDT [24] на базе 3DXPoint NVMe накопителей.

Графическая память. Память предназначенная для использования в графических картах. GDDR (Graphics Double Data Rate) — память предназначенная для использования в графических картах (видеокартах). Подвид энергозависимой динамической памяти с произвольным доступом (DRAM) и удвоенной скоростью передачи данных (DDR). В настоящее время доступны стандарты GDDR5 [25] и GDDR6 [26].

Наиболее популярным в настоящее время остается стандарт DDR-SDRAM, т.к. именно этот стандарт поддерживается большинством процессорных архитектур. Технология 3D Stacked Memory позволяет значительно увеличить пропускную способность, но обладает лимитированным объемом. Большой интерес представляют многоуровневые иерархии памяти позволяющие объединять технологии 3D Stacked Memory и DDR-SDRAM или DDR-SDRAM и SDM для получения больших объемов ОЗУ с высокой пропускной способностью.

2.3.2. Энергонезависимая память

Хотя анализ организации систем хранения выходит за рамки данной статьи, необходимо учесть, что на самом узле также должны присутствовать устройства для энергонезависимого хранения информации.

В настоящее время популярным решением становится интегрирование мощностей хранения в вычислительные узлы и объединение высокопроизводительным интерконнектом. В зависимости от выбранной конфигурации это могут быть твердотельные накопители (в том числе более дорогостоящие NVMe). Отдельное хранилище в большинстве случаев не может обеспечить пропускную способность, необходимую для задач обработки данных. В настоящее время можно выделить два основных пути для обеспечения надлежащих характеристик ввода-вывода:

– создание промежуточного буфера для работы с данными;
– полная интеграция уровня хранения с уровнем обработки (гиперконвергентность).

Создание промежуточного буфера (Burst Buffer [27]) для работы с данными подразумевает наличие твердотельных накопителей (в том числе возможно использование устройств SSD с интерфейсами DIMM и PCI-e для увеличения пропускной способности) для увеличения производительности ввода-вывода. Эти накопители могут быть установлены как непосредственно на вычислительных узлах, так и на специальных выделенных узлах, которые объединены высокопроизводительным интерконнектом наравне с вычислительными узлами.

Также используется специальное ПО, которое позволяет в фоновом режиме подкачивать на Burst Buffer данные для обработки, и перемещать в постоянное хранилище полученные результаты. Таким образом, приложение пользователя большую часть времени работает с локальными накопителями для быстрого ввода-вывода информации, а трансфер данных между Burst Buffer и основным хранилищем выполняется в фоновом режиме специальным ПО.

Например, на суперкомпьютере NERSC Cori используется BurstBuffer [28], организованный с использованием технологии DataWarp [29] от Cray.

Полная интеграция уровня хранения с уровнем обработки (гиперконвергентность) — это решение, которое используется в больших, гипер-масштабируемых ЦОД, таких как Google, Facebook или Amazon Web Services. Многие современные HPC платформы видят будущее именно в создании гиперконвергентных решений [30].

Группа компаний РСК также успешно разработала и продемонстрировала на европейской суперкомпьютерной выставке ISC'18 Гиперконвергентный вычислительный узел «РСК Торнадо» с прямым жидкостным охлаждением с использованием твердотельных дисков Intel SSD DC P4511 (NVMe, M.2) и Intel Optane SSD DC P4800X M.2 Series с Intel Memory Drive Technology (IMDT) [31].

2.4. Внутриузловой интерконнект

Процессоры, ускорители и память должны быть архитектурно интегрированы для эффективного взаимодействия.

В настоящее время внутриузловой интерконнект становится главным узким местом серверных узлов. Для создания современного высокопроизводительного узла необходимо рассмотреть лидирующие и перспективные варианты организации внутриузлового интерконнекта, различные топологии и технологии кластеризации (die-stacking).

Возможные варианты объединения можно классифицировать на два подхода:

- объединение высокоскоростными интерфейсами;
- интеграция на одном чипе:
  - интеграция универсальных и специализированных ядер на одном чипе;
  - интеграция вычислительных ядер и памяти на одном чипе (Memory-driven computing).

Объединение основного процессора и ускорителей-сопроцессоров высокоскоростными интерфейсами является классическим подходом, который позволяет разработчику серверного решения интегрировать топовые решения от ведущих мировых производителей.

Доминирующим решением в этой области является стандарт PCIe. В настоящее время актуальны спецификации версий 3.0 и 4.0.

PCIe 3.0. Пропускная способность у которого достигает порядка 1 ГБ/с на одиночную линию; интегрированная пропускная способность на 16 линий может достигать 32 ГБ/с в двух направлениях. Стандарт PCIe 3.0 поддерживается большинством производителей процессоров, сопроцессоров, а также устройств ввода-вывода.

PCIe 4.0. Обновленный стандарт PCIe 4.0 поддерживает двукратное увеличение скорости передачи данных (до 2 ГБ/с на одиночную линию и до 64 ГБ/с на 16 линий). Поддерживается некоторыми современными моделями процессоров, список устройств, поддерживающих PCIe 4.0, расширяются с каждым годом.

Кроме PCIe существуют и альтернативные решения — в первую очередь это новый открытый стандарт OpenCAPI, и Gen-Z а также частные решения NVIDIA NVLink, IBM Infinity Fabric, Intel UltraPath и т.п.

Рассмотрим подробнее некоторые (наиболее часто используемые) альтернативные решения. Среди открытых стандартов следует отметить ниже перечисленные технологии.

Gen-Z [32] — открытая технология, которая позволяет получить пропускную способность в 32 ГБ/с на канал и вплоть до 400 ГБ/с на несколько каналов. Открытый стандарт опубликован в 2018-м году.

OpenCAPI [33] — также открытый стандарт, разработанный широким консорциумом. Процессор IBM POWER9 поддерживает OpenCAPI. OpenCAPI 3.0 поддерживает пропускную способность до 25 Гбит/с на один канал, и вплоть до восьми каналов.

Более широкий спектр вариантов представляют частные разработки компаний. К недостаткам можно отнести ограниченную совместимость с устройствами других производителей. Ниже перечислен ряд проприетарных технологий внутриузлового интерконнекта.

NVIDIA NVLink [34]. Изначально разработанный для интеграции NVIDIA GPU и предоставления общей памяти, сейчас может быть применен для более широкого спектра решений, в том числе для организации высокоскоростных каналов «CPU – GPU». В частности, в топовом суперкомпьютере по данным рейтинга июнь 2018 используется соединение NVlink между процессорами IBM POWER9 и ускорителями NVIDIA. Пропускная способность (двунаправленная) одного линка NVLink 2.0 — 50 ГБ/с, аггрегированая пропускная способность шести линков (двунаправленная) – 300 ГБ/с.

Intel Ultra Path Interconnect (Intel® UPI) [35] — проприетарный канал для объединения двух процессоров Intel Xeon Scalable.

Infinity Fabric [36] — интерконнект от AMD, может быть использован для объединения процессоров AMD (семейства Zen) и графических ускорителей (например, Vega). В мультисокетной конфигурации с процессорами EPYC и оперативной памятью DDR4-2666 каждый линк может достигать производительности в 42,667 ГБ/с, общая двунаправленная пропускная способность – 170,667 ГБ/с.

Отечественные микропроцессоры «Эльбрус-8С» поддерживают по *3 дуплексных канала* с двунаправленной производительностью 16 ГБ/с [37].

Cavium Coherent Processor Interconnect CCPI2 — соединяет 2 процессора Cavium с производительностью 600 Гбит/с.

Кроме того, есть ряд решений, которые являются расширениями PCIe — например CCIX [38] и IBM CAPI [39].

Внутриузловой интерконнект оказывает существенное влияние на производительность вычислительного узла. При прочих равных (пиковая производительность, энергопотребление, наличие ПО для выбранной архитектуры) – предпочтение следует отдавать тем технологиям, которые поддерживают более производительные каналы передачи данных. С точки зрения каналов передачи данных наиболее перспективным в настоящее время является процессор IBM POWER9, который поддерживает NVLink и OpenCAPI.

Также следует следить за развитием перспективных технологий, которые могу позволить увеличить пропускную способность на несколько порядков. Особенный интерес в данном контексте представляет передача информации при помощи полупроводниковых лазеров. Потенциально данная технология может быть использована как для быстрой передачи данных внутри узла, так и для сверх-быстрого межузлового интерконнекта. В качестве примера успешной реализации приведем новый суперкомпьютер «Жорес», разработанный учеными Сколковского института науки и технологий, в котором

для передачи информации между узлами используются оптоволоконные каналы и полупроводниковые лазеры, основанные на полупроводниковых гетероструктурах [40].

3. Инфраструктура узла современного суперкомпьютера

3.1. Охлаждение

С целью увеличения вычислительной плотности, для охлаждения современных суперкомпьютеров следует использовать жидкостное охлаждение, а наиболее перспективным направлением являются охлаждающие пластины (coldplates) — плотно прилегающие пластины с хладагентом. В настоящее время существует огромное количество компаний, которые занимаются жидкостным охлаждением. В списке ниже рассмотрены наиболее распространенные подходы [41] к организации жидкостного охлаждения. *Охлаждающие пластины (coldplates)* — использование охлаждаемых жидкостью пластин, которые полностью покрывают всю элементосодержащую поверхность вычислительного узла. Примеры компаний, использующих данную технологию: Aquila [42], Dell. В России лидером по разработке вычислительных кластеров, охлаждаемых колдплейтами является группа компаний РСК [43].

Индивидуальные теплообменники (бобышки) — специальные элементы, позволяющие подводить охлаждающую жидкость напрямую к индивидуальным компонентам вычислительного узла. Примеры компаний: Asetek [44], Ebullient [45].

Жидкостное охлаждение на уровне шкафа. Примеры компаний: CoolIT, Inspur, HPE's Apollo.

Погружные (иммерсионные) системы — разработки, подразумевающие полное погружение вычислительных узлов в специальную диэлектрическую жидкость. Примеры компаний: ExaScaler Inc [46], 3M'S HPC [47]. Представителями в России являются IMMERS [48], а также погружные системы охлаждения реконфигурируемых вычислительных систем [49].

Другие разработки. Например, компания Liquid MIPS [50] представляет интересное направление — геотермальный кулинг.

В настоящее время можно говорить, что жидкостное охлаждение является однозначно лидирующим методом для охлаждения суперкомпьютеров больших масштабов. По сравнению с воздушным охлаждением методы жидкостного охлаждения обеспечивают более высокую надежность и низкое энергопотребление.

3.2. Датчики и сенсоры

Датчики и сенсоры, отвечающие за мониторинг состояния узла становятся критически важным элементом инфраструктуры. Наличие датчиков не только позволяет отслеживать критичные сбои в режиме реального времени, но и поможет прогнозировать потенциальные отказы оборудования, а также использовать данные о температуре и энергопотреблении для «умной» энергоэффективной балансировки вычислительной нагрузки.

Установленные датчики и сенсоры должны обеспечивать [51]:

- высокую точность и надежность измерений;
- измерение состояния отдельных компонентов вычислительного узла — таких как процессоры, оперативная память, ускорители, а также интерконнект и система охлаждения;
- корректное измерение энергопотребления;

– высокую частоту снятия данных.

Наиболее распространенным способом отслеживания состояния вычислительного узла являются встроенные датчики, информацию с которых можно получить через Intelligent Platform Management Interface (IPMI), который опрашивает board management controller (BMC). Другим способом является использование средств мониторинга и программных моделей, предоставляемых производителями, таких как Intel RAPL или IBM Amester. Существует и ряд других решений, таких как использование внешних устройств для измерения напряжения.

Независимо от выбранного способа организации аппаратного уровня мониторинга, необходимо предусмотреть наличие удобного пользовательского интерфейса для мониторинга присутствующих датчиков. Крайне желательно наличие унифицированного API, который позволит в едином стиле получать информацию с системных датчиков и сенсоров [52].

При наличии соответствующей аппаратной и программной инфраструктуры, позволяющей отслеживать основные характеристики вычислительного кластера на уровне базовых вычислительных элементов возможно «умное» управление кластером, в том числе отслеживание потенциальных сбоев по изменению характеристик температуры и энергопотребления, использование адаптивных алгоритмов балансировки нагрузки [53]. Данная инфраструктура должна существенно повысить время функциональной работы вычислительной системы.

4. Отображение вычислительных задач на архитектуру

Некоторая часть суперкомпьютеров создается для специфических задач, но большая часть должна быть готова к задачам из разных прикладных областей. Определим несколько основных классов задач и предложим обобщенные варианты архитектур узлов для них.

Мощные вычислительные задачи. Для этого класса задач мы рекомендуем гомогенные узлы с универсальными вычислителями, а также при наличии ограничений по энергопотреблению — узлы на базе универсальных процессоров RISC или CISC и дополнительных ускорителей GPGPU.

Задачи обработки больших данных. Данный класс задач для эффективной работы требует гиперконвергентные узлы или высокопроизводительную систему хранения данных, построенную с использованием специальных технологий организации промежуточных буферов между системой хранения и вычислительными узлами.

Задачи машинного обучения и глубокого обучения. Для этого класса задач мы рекомендуем гиперконвергентные узлы на базе универсальных процессоров RISC или CISC и дополнительных ускорителей GPGPU. Особое внимание следует уделить характеристикам внутриузлового интерконнекта, доступного для интеграции выбранных моделей процессоров и ускорителей.

Задачи из некоторых специфических областей. Для этого класса задач мы рекомендуем узлы на базе универсальных процессоров RISC или CISC и дополнительных ускорителей FPGA, реализующих алгоритмы для данных задач.

Тем не менее, основным критерием для быстрого решения конкретных задач на суперкомпьютерах является не только наличие соответствующей аппаратной части, но и эффективного прикладного программного обеспечения, функционирующего на данной

аппаратуре. Таким образом, важнейшими для узла характеристиками, на основе которых должен производиться детальный выбор узла являются:

- наличие необходимого прикладного программного обеспечения для данных архитектур;
- теоретическая производительность узла;
- максимальное энергопотребление узла;
- стандартизация и модульность компонент узла (как с точки зрения замены вышедших из строя, так и обновленных узлов).

Заключение

В данной статье рассмотрены основные конструктивные элементы современных вычислительных узлов. Даны классификации и наиболее популярные примеры универсальных и специализированных ядер, проведён анализ тенденций на основе лидирующий суперкомпьютеров рейтинга ТОП-500.

Основные итоги проведённого обзора заключаются в следующем:

Выбор архитектуры универсальных ядер и ускорителей во многом зависит от задач, которые планируется решать на проектируемом суперкомпьютере. В работе определены основные классы задач и предложены наиболее перспективные варианты архитектур для них.

При организации ОЗУ наиболее популярным в настоящее время остается стандарт DDR-SDRAM. Альтернативными решениями являются иерархические решения с использованием дополнительного уровня ОЗУ на базе энергонезависимой памяти, а также использование 3D Stacked Memory, которая позволяет увеличить пропускную способность, но обладает лимитированным объемом.

На узле также должны присутствовать энергонезависимые устройства хранения, которые могут быть использованы для создания промежуточного буфера для работы с данными или для полной интеграции уровня хранения с уровнем обработки (гиперконвергентность) в зависимости от выбранного типа организации системы хранения данных.

В связи с возрастающими требованиями к вычислительной плотности, для охлаждения современных суперкомпьютеров следует использовать жидкостное охлаждение. В статье рассмотрены популярные подходы к организации жидкостного охлаждения — погружное, с использованием индивидуальных теплообменников и охлаждающих пластин (coldplates) и приведены примеры успешных решений на базе перечисленных подходов.

Наконец, современный суперкомпьютер должен быть оснащён системой мониторинга. Соответственно, на уровне вычислительного узла должны присутствовать датчики и сенсоры, обеспечивающие высокую точность и надежность измерений; высокую частоту снятия данные, а также корректное измерение состояния отдельных компонентов вычислительного узла.

В заключение обсудим некоторые технологические проблемы. Прежде всего, это исчерпание потенциала закона Мура (точнее бизнес-прогноза Мура), вызванное тем, что проектная норма технологических процессов подходит к физически допустимому пределу. Согласно докладу [54] для сохранения прогресса в компьютерной отрасли есть три основных направления:

- изобретение новых устройств;

– изобретение новых архитектур;
– новые парадигмы вычислений.

Некоторые актуальные разработки, относящиеся к первым двум направлениям, упомянуты в статье. Прежде всего, это узкоспециализированные ускорители (аналоговые, квантовые, тензорные и нейроморфные). Существенный прогресс в настоящее время наблюдается в использовании различных наборов инструкций (ISA) — кроме традиционного CISC доступны процессоры и соответствующий стек ПО для различных RISC и VLIW архитектур. Также следует отметить прогресс в области организации иерархии памяти, особенно 3D Stacked Memory.

В статье мы старались сделать акцент на тех технологиях, которые достигли стадии выхода в производство или, как минимум, создания рабочих прототипов. За пределами обзора оказались направления работ, которые требуют преодоления существенных технологических барьеров. Тем не менее, в контексте потенциально перспективных технологий нельзя не отметить идеи использования сверхпроводников для создания цифровых и квантовых компьютеров, идеи использования других материалов и структур для создания вычислителей (например, углеродных нанотрубок или графена). Если существующие на текущий момент технологические барьеры в любом из этих направлений будут преодолены, это может положить начало новой эпохе энергоэффективных вычислений.

Очень важно, чтобы основные усилия были направлены на разработку соответствующего программного стека, который будет поддерживать новые программные парадигмы и аппаратные технологии. Это — гибридные вычисления, параллелизм на миллионы и более потоков, использование нестандартных архитектур и т.п. Необходимо решить ряд противоречий, например между необходимостью глубокой аппаратно-зависимой оптимизации ПО и необходимости портирования этого ПО на широкий диапазон разноархитектурных вычислителей. Еще одно противоречие заключается в необходимости разработки новой сверх-масштабируемой парадигмы параллельных вычислений, которая позволит эффективно использовать вычислительные суперкомпьютеры будущего, и необходимости портирования на эти же суперкомпьютеры существующих основных вычислительных пакетов.

Таким образом, в ближайшем будущем прогресс в области повышения производительности машинных вычислений будет во многом зависеть от эффективности интеграции новых технологий (ускорителей, внутриузлового интерконнекта, межузлового интерконнекта, новых наборов инструкций, иерархий памяти) на уровне стека программного обеспечения. Прогресс может быть неравномерным, зависеть от предметной области, алгоритмической специфики и соответствующих архитектурных требований и степени адаптации (или разработки с нуля) специализированных программных пакетов к современным суперкомпьютерам.

Литература

1. Проект Российской Академии Наук: «Создание вычислительной системы для моделирования суперкомпьютера с производительностью экзафлопсного уровня». Институт прикладной математики им. М.В. Келдыша Российской академии наук. URL: `http://www.keldysh.ru/projects/exaflops.pdf` (дата обращения: 23.01.2019).

2. Reed D.A., Dongarra J. Exascale Computing and Big Data // Communications of the ACM.

2015. Vol. 58, No. 7. P. 56-68. DOI: 10.1145/2699414.

3. Chrysos G. Intel® Xeon Phi coprocessor (codename Knights Corner) // Proceedings of the 2012 IEEE Hot Chips 24 Symposium, HCS, August 27–29, 2012, Cupertino, CA. P. 1–31. DOI: 10.1109/HOTCHIPS.2012.7476487.
4. Lindholm E.,Nickolls J., Oberman S., Montrym J. NVIDIA Tesla: A Unified Graphics and Computing Architecture // IEEE Micro. 2008. Vol. 28, No. 2. P. 39–55. DOI: 10.1109/MM.2008.31.
5. Jouppi N., Young C., Patil N., Patterson D. Motivation for and Evaluation of the First Tensor Processing Unit // IEEE Micro. 2018. Vol. 38, No. 3. P. 10–19. DOI: 10.1109/MM.2018.032271057.
6. Davies M. et al. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning // IEEE Micro. 2018. Vol. 38, No. 1. P. 82–99. DOI: 10.1109/MM.2018.112130359.
7. Hsu J. CES 2018: Intel's 49-Qubit Chip Shoots for Quantum Supremacy. IEEE Spectrum Tech Talks. 2018. URL: https://spectrum.ieee.org/tech-talk/computing/hardware/intels-49qubit-chip-aims-for-quantum-supremacy (дата обращения: 23.11.2018).
8. Intel® Stratix® 10 SoC FPGAs. URL: https://www.intel.com/content/www/us/en/products/programmable/soc/stratix-10.html (дата обращения: 23.11.2018).
9. Exascale Requirements Review. An Office of Science review sponsored jointly by Advanced Scientific Computing Research and High Energy Physics. June 10–12, 2015, Bethesda, Maryland. URL: http://hepcce.org/files/2016/11/DOE-ExascaleReport-HEP-Final.pdf (дата обращения: 13.11.2018).
10. Top500 List Statistics. Release November 2018. URL: https://www.top500.org/statistics/list/ (дата обращения: 16.11.2018).
11. Hemsoth N. Cascade Lake at Heart of 2019 TACC Supercomputer. Онлайн ресурс технологических новостей Next Platform поддерживаемый Stackhouse Publishing Inc в партнерстве с The Register. 2018. URL: https://www.nextplatform.com/2018/08/29/cascade-lake-heart-of-2019-tacc-supercomputer/ (дата обращения: 13.11.2018).
12. Bartsch V. D6.3 Initial Project Press Release // ExaNoDe Consortium Public deliverable. Пресс-релиз по проекту ExaNode. 2016. URL: http://exanode.eu/wp-content/uploads/2017/04/D6.3.pdf (дата обращения: 16.11.2018).
13. ARMv8 — A Scalable Vector Extension for Post-K. FUJITSU LIMITED. 2016. URL: http://www.fujitsu.com/global/Images/armv8-a-scalable-vector-extension-for-post-k.pdf (дата обращения: 22.01.2019).
14. Astra. Top500 The List. URL: https://www.top500.org/system/179565 (дата обращения: 16.11.2018).
15. Xilinx. High Performance Computing and Data Storage. URL: https://www.xilinx.com/applications/high-performance-computing.html (дата обращения: 23.11.2018).
16. Timmel A.N., Daly J.T. Multiplication with Fourier Optics Simulating 16-bit Modular Multiplication. URL: https://arxiv.org/pdf/1801.01121.pdf (дата обращения: 23.11.2018).
17. Ким А.К., Перекатов В.И., Фельдман В.М. На пути к российской экзасистеме: планы разработчиков аппаратно-программной платформы «Эльбрус» по созданию

суперкомпьютера экзафлопсной производительности // Вопросы радиоэлектроники. Вычислительные системы на базе многоядерных микропроцессоров. 2018. № 2. С. 6-–13.

18. CORAL Collaboration: Briefing on CORAL-2 RFP and Draft Technical Requirements // Vendor Webinar Meeting. December 6, 2017 URL: `https://procurement.ornl.gov/rfp/CORAL2/Brief-of-Draft-SOW-20171206-SA.PDF` (дата обращения: 23.11.2018).

19. Farber R. HPC and AI — Two Communities Same Future. HPCwire: Global News and Information on High Performance Computing. 2018. URL: `https://www.hpcwire.com/2018/01/25/hpc-ai-two-communities-future/` (дата обращения: 23.11.2018).

20. JEDEC DDR5 & NVDIMM-P Standards Under Development. Global Standards for the Microelectronics Industry. 2017. URL: `https://www.jedec.org/news/pressreleases/jedec-ddr5-nvdimm-p-standards-under-development` (дата обращения: 23.11.2018).

21. Hadidi R. et al. Demystifying the Characteristics of 3D-Stacked Memories: A Case Study for Hybrid Memory Cube // Proceedings of the IEEE International Symposium on Workload Characterization, IISWC 2017, October 1–3, 2017, Seattle, WA, USA, P. 66-75. DOI: 10.1109/IISWC.2017.8167757.

22. High Bandwidth Memory (HBM) DRAM. JESD235A. Global Standards for the Microelectronics Industry. 2015. URL: `https://www.jedec.org/standards-documents/docs/jesd235a` (дата обращения: 23.11.2018).

23. Hybrid Memory Cube (HMC). Hybrid Memory Cube Consortium Page. URL: `http://hybridmemorycube.org/` (дата обращения: 16.11.2018).

24. Intel® Memory Drive Technology Application Note. URL: `https://www.intel.com/content/dam/support/us/en/documents/memory-and-storage/intel-mem-drive-tech-appnote.pdf` (дата обращения: 23.11.2018).

25. Graphics Double Data Rate (GDDR5) SGRAM standard. JESD212C. Global Standards for the Microelectronics Industry. 2016. URL: `https://www.jedec.org/standards-documents/docs/jesd212c` (дата обращения: 23.11.2018).

26. Graphics Double Data Rate 6 (GDDR6) SGRAM standard. JESD250A. Global Standards for the Microelectronics Industry. 2017. URL: `https://www.jedec.org/standards-documents/docs/jesd250a` (дата обращения: 23.11.2018).

27. Ferreira da Silva R., Callaghan S., Deelman E. On the use of burst buffers for accelerating data-intensive scientific workflows // Proceedings of the 12th Workshop on Workflows in Support of Large-Scale Science, WORKS '17. ACM, 2017. P. 2:1–2:9. DOI: 10.1145/3150994.3151000.

28. Bhimji W., Bard D., Romanus M., Paul, D., Ovsyannikov A., Friesen B., et al. Accelerating Science with the NERSC Burst Buffer Early User Program. Lawrence Berkeley National Laboratory. 2016. URL: `https://escholarship.org/uc/item/9wv6k14t` (дата обращения: 23.11.2018).

29. Cray® DataWarp™ Applications I/O Accelerator. URL: `https://www.cray.com/products/storage/datawarp` (дата обращения: 23.11.2018).

30. Morgan T.P. For many hyperconverged is the next platform. Онлайн ресурс технологических новостей Next Platform поддерживаемый Stackhouse Publishing Inc в

партнерстве с The Register. 2018. URL: https://www.nextplatform.com/2018/01/29/hyperconverged-next-platform-many-jobs/ (дата обращения: 23.11.2018).

31. Бахур В. РСК представила гиперконвергентное HPC-решение на новейших компонентах. Интернет-издание о высоких технологиях cnews. 2018. URL: http://www.cnews.ru/news/line/2018-06-27_rsk_predstavila_giperkonvergentnoe_hpcreshenie (дата обращения: 23.11.2018).
32. The GEN-Z Consortium. URL: https://genzconsortium.org/ (дата обращения: 23.11.2018).
33. The OpenCAPI Consortium. URL: https://opencapi.org/ (дата обращения: 23.11.2018).
34. NVLink Fabric. URL: https://www.nvidia.com/ru-ru/data-center/nvlink/ (дата обращения: 23.11.2018).
35. Краткое описание продукции: платформа масштабируемых процессоров Intel® Xeon®. URL: https://www.intel.ru/content/www/ru/ru/processors/xeon/scalable/xeon-scalable-platform-brief.html (дата обращения: 23.11.2018).
36. Infinity Fabric (IF) — AMD. URL: https://en.wikichip.org/wiki/amd/infinity_fabric (дата обращения: 23.11.2018).
37. Российские микропроцессоры серии «Эльбрус» и МЦСТ R и системные платы на их основе. Каталог продукции 2017. URL: http://mcst.ru/files/59db45/cf0cd8/50a21b/000000/katalog_produktsii_mtsst_hq.pdf (дата обращения: 23.11.2018).
38. CCIX. URL: https://www.ccixconsortium.com/(дата обращения: 23.11.2018).
39. Coherent Accelerator Processor Interface (CAPI). URL: https://developer.ibm.com/linuxonpower/capi/ (дата обращения: 23.11.2018).
40. Шустиков В. Ученые Сколтеха создали суперкомпьютер «Жорес». Фонд «Сколково», пресс-релизы. 2019. URL: https://sk.ru/news/b/pressreleases/archive/2019/01/18/uchenye-skolteha-sozdali-superkompyuter-zhores.aspx (дата обращения: 25.01.2019).
41. Is Liquid Cooling Ready to Go Mainstream? HPCwire: Global News and Information on High Performance Computing. URL: https://www.hpcwire.com/2017/02/13/liquid-cooling-ready-go-mainstream/ (дата обращения: 16.11.2018).
42. Aquila. URL: https://www.aquilagroup.com/cooling/ (дата обращения: 26.11.2018).
43. Группа компаний РСК. URL: http://www.rscgroup.ru/ru (дата обращения: 25.01.2019).
44. Asetek. URL: https://www.asetek.com/ (дата обращения: 26.11.2018).
45. Ebullient. URL: http://ebullientcooling.com/ (дата обращения: 26.11.2018).
46. ExaScaler Inc. Overview. URL: http://www.exascaler.co.jp/en/company (дата обращения: 26.11.2018).
47. 3M Server Solutions for Data Centers. URL: https://www.3m.com/3M/en_US/data-center-us/solutions/data-center-servers/ (дата обращения: 26.11.2018).
48. Абрамов С.М., Амелькин С.А., Романенко А.Ю., Симонов А.С., Чичковский А.А. Опыт реализации высокопроизводительных вычислительных систем с погружной жидкостной системой охлаждения // Труды 3-й Всероссийской научно-технической конференции

«Суперкомпьютерные технологии» (СКТ-2014) (Дивноморское, Геленджик, 29 сентября – 4 октября 2014 г.). С. 9–15.

49. Левин И.И., Дордопуло А.И., Доронченко Ю.И., Раскладкин М.К., Федоров А.М. Погружная система охлаждения реконфигурируемых вычислительных систем на основе ПЛИС // Программные системы: теория и приложения. 2016. №4 (31). URL: https://cyberleninka.ru/article/n/pogruzhnaya-sistema-ohlazhdeniya-rekonfiguriruemyh-vychislitelnyh-sistem-na-osnove-plis (дата обращения: 28.11.2018).
50. Liquid MIPS. URL: http://www.liquidmips.com/cms/en-us/howitworks.aspx (дата обращения: 26.11.2018).
51. Libri A., Bartolini A., Benini L. Dwarf in a Giant: Enabling Scalable, High-Resolution HPC Energy Monitoring for Real-Time Profiling and Analytics. URL: https://arxiv.org/pdf/1806.02698.pdf (дата обращения: 19.11.2018).
52. Grant R.E., Levenhagen M., Olivier S.L., DeBonis D., Pedretti K.T., Laros III J.H. Standardizing Power Monitoring and Control at Exascale // Computer. 2016. Vol. 49, No. 10. P. 38–46. DOI: 10.1109/MC.2016.308.
53. Georgiou Y., Glesser D., Trystram D. Adaptive Resource and Job Management for Limited Power Consumption // IEEE International Parallel and Distributed Processing Symposium Workshop, 2015, Hyderabad. P. 863–870. DOI: 10.1109/IPDPSW.2015.118.
54. Shalf J.M., Leland R. Computing beyond Moore's Law // Computer. 2015. Vol. 48, No. 12. P. 14–23. DOI: 10.1109/mc.2015.37.

Тютляева Екатерина Олеговна, инженер НИР ЗАО «РСК Технологии», Переславль-Залесский, Россия.

Одинцов Игорь Олегович, руководитель отдела НИР, ООО «РСК Лабс», Санкт-Петербург, Россия.

Московский Александр Александрович, к.х.н., генеральный директор ЗАО «РСК Технологии», Москва, Россия.

Мармузов Глеб Владимирович, к.х.н., директор по техническому маркетингу, ЗАО «РСК Технологии», Москва, Россия.

DOI: 10.14529/cmse190305

DEVELOPMENT TRENDS OF MODERN SUPERCOMPUTERS

© 2019 E.O. Tyutlyaeva[1], I.O. Odintsov[2], A.A. Moskovsky[1], G.V. Marmuzov[1]

[1] *ZAO RSC Technologies*

(121170, Moscow, Kutuzovskiy av., 36, building 23),

[2] *OOO RSC Labs*

(121205, Moscow, Bolshoi Bulvar str., 42, building 1)

E-mail: xgl@rsc-tech.ru, igor_ odintsov@rsc-tech.ru, moskov@rsc-tech.ru, gleb.marmuzov@rsc-tech.ru

Received: 28.01.2019

This work includes analysis of the computation nodes of modern supercomputers from two perspectives; first the hardware components focus and secondly discuss the infrastructure. Identified trends leads to basic options of computation node design. The paper classifies the modern architectures of universal processing cores and specialized hardware accelerators cores; studies the recent trends in memory hierarchy design and intra-node interconnect; the paper includes ways of using the non-volatile memory in modern memory hierarchy. Furthermore, the paper analyses the recent trends in HPC infrastructure, in particular in modern liquid cooling approaches and monitoring. The basic variants of HPC computing nodes design are based on energy efficient universal processor and set of energy-efficient specialized hardware accelerators cores, according to the observed trends. The paper focuses on recent technologies that are currently at various stages of production or at the functional prototype stage. The study also discusses state-of-the-art computational challenges and algorithms-to-architecture mapping issues. Lastly, the paper discusses the current technological problems and main areas to maintain the progress in HPC area.

Keywords: high-performance compute node, architecture development of a high-performance compute node, analysis of supercomputer architectures, development trends of supercomputers.

FOR CITATION

Tyutlyaeva E.O., Odintsov I.O., Moskovsky A.A., Marmuzov G.V. Development Trends of Modern Supercomputers. *Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering.* 2019. vol. 8, no. 3. pp. 92–114. (in Russian) DOI: 10.14529/cmse190305.

References

1. The project of the Russian Academy of Science: «Development of the compute system for simulation of the exascale supercomputer». Available at: `http://www.keldysh.ru/projects/exaflops.pdf` (accessed: 23.01.2019).

2. Reed D.A., Dongarra J. Exascale Computing and Big Data. *Communications of the ACM.* 2015. vol. 58, no. 7. pp. 56-68. DOI: 10.1145/2699414.

3. Chrysos G. Intel® Xeon Phi coprocessor (codename Knights Corner). Proceedings of the 2012 IEEE Hot Chips 24 Symposium, HCS, August 27–29, 2012, Cupertino, CA. pp. 1–31. DOI: 10.1109/HOTCHIPS.2012.7476487.

4. Lindholm E.,Nickolls J., Oberman S., Montrym J. NVIDIA Tesla: A Unified Graphics and Computing Architecture. *IEEE Micro*. 2008. vol. 28, no. 2. pp. 39–55. DOI: 10.1109/MM.2008.31.

5. Jouppi N., Young C., Patil N., Patterson D. Motivation for and Evaluation of the First Tensor Processing Unit. *IEEE Micro*. 2018. vol. 38, no. 3. pp. 10–19. DOI: 10.1109/MM.2018.032271057.

6. Davies M. et al. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro*. 2018. vol. 38, no. 1. pp. 82–99. DOI: 10.1109/MM.2018.112130359.

7. Hsu J. CES 2018: Intel's 49-Qubit Chip Shoots for Quantum Supremacy. IEEE Spectrum Tech Talks. 2018. Available at: `https://spectrum.ieee.org/tech-talk/computing/hardware/intels-49qubit-chip-aims-for-quantum-supremacy` (accessed: 23.11.2018).

8. Intel® Stratix® 10 SoC FPGAs. Available at: `https://www.intel.com/content/www/us/en/products/programmable/soc/stratix-10.html` (accessed: 23.11.2018).

9. Exascale Requirements Review. An Office of Science review sponsored jointly by Advanced Scientific Computing Research and High Energy Physics. June 10–12. 2015 BETHESDA, MARYLAND. Available at: `http://hepcce.org/files/2016/11/DOE-ExascaleReport-HEP-Final.pdf` (accessed: 13.11.2018).

10. Top500 List Statistics. Release November 2018. Available at: `https://www.top500.org/statistics/list/` (accessed: 16.11.2018).

11. Hemsoth N. Cascade Lake at Heart of 2019 TACC Supercomputer. Technology publication resource The Next Platform is published by Stackhouse Publishing Inc in partnership with The Register. Available at: `https://www.nextplatform.com/2018/08/29/cascade-lake-heart-of-2019-tacc-supercomputer/` (accessed: 13.11.2018).

12. Bartsch V. D6.3 Initial Project Press Release. ExaNoDe Consortium Public deliverable. 2016. Available at: `http://exanode.eu/wp-content/uploads/2017/04/D6.3.pdf` (accessed: 16.11.2018).

13. ARMv8 — A Scalable Vector Extension for Post-K. FUJITSU LIMITED. 2016 Available at: `http://www.fujitsu.com/global/Images/armv8-a-scalable-vector-extension-for-post-k.pdf` (accessed: 22.01.2019).

14. Astra. Top500 The List. Available at: `https://www.top500.org/system/179565` (accessed: 16.11.2018).

15. Xilinx. High Performance Computing and Data Storage. Available at: `https://www.xilinx.com/applications/high-performance-computing.html` (accessed: 23.11.2018).

16. Timmel A.N., Daly J.T. Multiplication with Fourier Optics Simulating 16-bit Modular Multiplication. Available at: `https://arxiv.org/pdf/1801.01121.pdf` (accessed: 23.11.2018).

17. Kim A.K., Perekatov V.I., Feldman V.M. On the way to russian exasistemes: plans of the Elbrus hardware-software platform developers on creation of an exaflops performance supercomputer. *Voprosy radioelektroniki*, 2018, no. 2, pp. 6–13. (in Russian)

18. CORAL Collaboration: Briefing on CORAL-2 RFP and Draft Technical Requirements. Vendor Webinar Meeting. 2017. Available at: `https://procurement.ornl.gov/rfp/CORAL2/Brief-of-Draft-SOW-20171206-SA.PDF` (accessed: 23.11.2018).

19. Farber R. HPC and AI — Two Communities Same Future. HPCwire: Global News and Information on High Performance Computing. 2018. Available at: https://www.hpcwire.com/2018/01/25/hpc-ai-two-communities-future/ (accessed: 23.11.2018).

20. JEDEC DDR5 & NVDIMM-P Standards Under Development. Global Standards for the Microelectronics Industry. 2017. Available at: https://www.jedec.org/news/pressreleases/jedec-ddr5-nvdimm-p-standards-under-development (accessed: 23.11.2018).

21. Hadidi R. et al. Demystifying the Characteristics of 3D-Stacked Memories: A Case Study for Hybrid Memory Cube. Proceedings of the IEEE International Symposium on Workload Characterization, IISWC 2017, October 1–3, 2017, Seattle, WA, USA. pp. 66-75. DOI: 10.1109/IISWC.2017.8167757.

22. High Bandwidth Memory (HBM) DRAM. JESD235A. Global Standards for the Microelectronics Industry. 2015. Available at: https://www.jedec.org/standards-documents/docs/jesd235a (accessed: 23.11.2018).

23. Hybrid Memory Cube (HMC). Hybrid Memory Cube Consortium Page. Available at: http://hybridmemorycube.org/ (accessed: 16.11.2018).

24. Intel® Memory Drive Technology Application Note. Available at: https://www.intel.com/content/dam/support/us/en/documents/memory-and-storage/intel-mem-drive-tech-appnote.pdf (accessed: 23.11.2018).

25. Graphics Double Data Rate (GDDR5) SGRAM standard. JESD212C. Global Standards for the Microelectronics Industry. 2016. Available at: https://www.jedec.org/standards-documents/docs/jesd212c (accessed: 23.11.2018).

26. Graphics Double Data Rate 6 (GDDR6) SGRAM standard. JESD250A. Global Standards for the Microelectronics Industry. 2017. Available at: https://www.jedec.org/standards-documents/docs/jesd250a (accessed: 23.11.2018).

27. Ferreira da Silva R., Callaghan S., Deelman E. On the use of burst buffers for accelerating data-intensive scientific workflows. Proceedings of the 12th Workshop on Workflows in Support of Large-Scale Science, WORKS '17. ACM, 2017. pp. 2:1–2:9. DOI: 10.1145/3150994.3151000.

28. Bhimji W., Bard D., Romanus M., Paul, D., Ovsyannikov A., Friesen B., et al. Accelerating Science with the NERSC Burst Buffer Early User Program. Lawrence Berkeley National Laboratory. 2016. Available at: https://escholarship.org/uc/item/9wv6k14t (accessed: 23.11.2018).

29. Cray® DataWarp™ Applications I/O Accelerator. Available at: https://www.cray.com/products/storage/datawarp (accessed: 23.11.2018).

30. Morgan T.P. For many hyperconverged is the next platform. Technology publication resource The Next Platform is published by Stackhouse Publishing Inc in partnership with The Register. 2018. Available at: https://www.nextplatform.com/2018/01/29/hyperconverged-next-platform-many-jobs/ (accessed: 23.11.2018).

31. Bahur V. The RSC Technologies Company introduces hyperconverged HPC-solution based on state-of-the-are components. 2018. Available at: http://www.cnews.ru/news/line/2018-06-27_rsk_predstavila_giperkonvergentnoe_hpcreshenie (accessed: 23.11.2018). (in Russian)

32. The GEN-Z Consortium. Available at: https://genzconsortium.org/ (accessed: 23.11.2018).

33. The OpenCAPI Consortium. Available at: https://opencapi.org/ (accessed: 23.11.2018).

34. NVLink Fabric. Available at: https://www.nvidia.com/ru-ru/data-center/nvlink/ (accessed: 23.11.2018).

35. Product Brief: Intel® Xeon® Scalable Platform. Available at: https://www.intel.sg/content/www/xa/en/processors/xeon/scalable/xeon-scalable-platform-brief.html (accessed: 23.11.2018).

36. Infinity Fabric (IF) — AMD. Available at: https://en.wikichip.org/wiki/amd/infinity_fabric (accessed: 23.11.2018).

37. Russian microprocessors «Elbrus» and «MCST R» series and boards based thereon. Production catalogue 2017. Available at: http://mcst.ru/files/59db45/cf0cd8/50a21b/000000/katalog_produktsii_mtsst_hq.pdf (accessed: 23.11.2018). (in Russian)

38. CCIX. Available at: https://www.ccixconsortium.com/ (accessed: 23.11.2018).

39. Coherent Accelerator Processor Interface (CAPI). Available at: https://developer.ibm.com/linuxonpower/capi/ (accessed: 23.11.2018).

40. Shustikov V. Skoltex researches developed supercomputer «Zhores». The Skolkovo Foundation Press-Release. 2019. Available at: https://sk.ru/news/b/pressreleases/archive/2019/01/18/uchenye-skolteha-sozdali-superkompyuter-zhores.aspx (accessed: 25.01.2019). (in Russian).

41. Is Liquid Cooling Ready to Go Mainstream? HPCwire: Global News and Information on High Performance Computing. Available at: https://www.hpcwire.com/2017/02/13/liquid-cooling-ready-go-mainstream/ (accessed: 16.11.2018).

42. Aquila. Available at: https://www.aquilagroup.com/cooling/ (accessed: 26.11.2018).

43. RSC Group. Available at: http://www.rscgroup.ru/en (accessed: 25.01.2019).

44. Asetek. Available at: https://www.asetek.com/ (accessed: 26.11.2018).

45. Ebullient. Available at: http://ebullientcooling.com/ (accessed: 26.11.2018).

46. ExaScaler Inc. Overview. Available at: http://www.exascaler.co.jp/en/company (accessed: 26.11.2018).

47. 3M Server Solutions for Data Centers. Available at: https://www.3m.com/3M/en_US/data-center-us/solutions/data-center-servers/ (accessed: 26.11.2018).

48. Abramov S.M., Amelkin S.A., Romanenko A.Y., Simonov A.S., Chichkovsky A.A. The experience of implementing the high-performance computing systems with immersion cooling. Proceedings of the All-Russian Scientific and Technical Conference «Supercomputer Technologies» (Divnomorskoe, Russia, September, 29 – October, 4, 2014). pp. 9–15. (in Russian).

49. Levin I., Dordopulo A., Doronchenko Y., Raskladkin M., Fedorov A. Immersion cooling system for FPGA-based reconfigurable computer systems. *Program systems: theory and applications*. 2016. vol. 7:4(31), pp. 65–81. Available at: http://psta.psiras.ru/read/psta2016_4_65-81.pdf (accessed: 24.01.2019). (in Russian).

50. Liquid MIPS. Available at: http://www.liquidmips.com/cms/en-us/howitworks.aspx (accessed: 26.11.2018).

51. Libri A., Bartolini A., Benini L. Dwarf in a Giant: Enabling Scalable, High-Resolution HPC Energy Monitoring for Real-Time Profiling and Analytics. Available at: https://arxiv.org/pdf/1806.02698.pdf (accessed: 19.11.2018).

52. Grant R.E., Levenhagen M., Olivier S.L., DeBonis D., Pedretti K.T., Laros III J.H. Standardizing Power Monitoring and Control at Exascale. *Computer*, 2016. vol. 49, no. 10. pp. 38–46. DOI: 10.1109/MC.2016.308.

53. Georgiou Y., Glesser D., Trystram D. Adaptive Resource and Job Management for Limited Power Consumption. IEEE International Parallel and Distributed Processing Symposium Workshop, 2015, Hyderabad. pp. 863–870. DOI: 10.1109/IPDPSW.2015.118.

54. Shalf J.M., Leland R. Computing beyond Moore's Law. *Computer*. 2015. vol. 48, no. 12. pp. 14–23. DOI: 10.1109/mc.2015.37.