Тенденции развития вычислительных узлов современных суперкомпьютеров

Е.О. Тютляева, И.О. Одинцов, А.А. Московский, Г.В. Мармузов (ЗАО "РСК Технологии")

Вниманию читателей предлагается анализ вычислительных узлов современных суперкомпьютеров с двух точек зрения – аппаратно-компонентной и инфраструктурной. Выявленные тенденции приводят к созданию основных вариантов дизайна вычислительных узлов, состоящих из энергоэффективного универсального процессора и совокупности энергоэффективных специализированных ускорителей. В статье рассмотрены такие компоненты, как оперативная и энергонезависимая память, внутренний интерконнект, а также организация подсистем мониторинга и охлаждения. Обсуждаются современные суперкомпьютерные задачи и их отображение на архитектуру вычислительных узлов.

1. Введение

При создании вычислительного узла для современного суперкомпьютера необходимо использовать новые технологические подходы. Простое масштабирование существующих технологий не будет эффективным решением.

Можно сформулировать ряд основных задач, которые стоят перед разработчиками лидирующих суперкомпьютеров:

- Минимизация энергопотребления – необходимо разработать новые технологии, которые позволят уменьшить планируемое энергопотребление до экономически приемлемого уровня (20÷40 *MW*). Масштабирование текущих технологий не позволяет уложиться в данный лимит.
- Обеспечение модульности и высокой вычислительной плотности в рамках одного узла. Необходимы высокоскоростные интерфейсы в рамках одного узла.
- Организация энергоэффективного межузлового интерконнекта (то есть, системы связи) с высокой пропускной способностью и минимальными задержками.
- Организация высокопроизводительной системы хранения с достаточно высокой пропускной способностью, чтобы избежать простоя вычислительных ресурсов.
- Подбор оптимальных типов памяти для построения наиболее эффективной структуры памяти узла.

Ряд вызовов связан и с программным обеспечением (ПО).

Здесь мы выделяем как минимум две существенные грани:

1. Сложность создания массового параллельного ПО на современных языках программирования фактически поднимает вопрос о необходимости новой парадигмы программирования: гиперпараллельной.
2. Наличие множества различных высокопроизводительных аппаратных архитектур выдвигает задачу обеспечить не просто переносимость ПО, а максимально эффективную переносимость и автоматизированное отображение на архитектуру.

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

2. Задачи и вычислительные ядра

Д. Рид и Д. Донгарра в статье [2] выделяют две основные экосистемы:

- обработка данных;
- вычислительная наука.

Авторы подчеркивают, что 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 ПО для ряда задач машинного обучения и глубокого обучения. Дополнительно заметим, что если ранее основной объем задач создавали научные и инженерные вычисления, которых и относили к классическим задачам *HPC* (*High-Performance Computing*), то теперь прогнозируется, что в будущем их доля будет составлять примерно половину, а оставшийся объем займут обработка больших данных и задачи искусственного интеллекта, машинного обучения и глубокого обучения.

Если смотреть с 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

Классифицировать универсальные ядра (типичная роль – хост, базовое, "классическое вычислительное", "толстое" ядро) можно по следующим критериям:

- Традиционный набор команд (*Complex Instruction Set Computing, CISC*):
  - *x86, IA-64, x86-64* (*Intel, AMD*).
- Упрощенный набор команд (*Reduced Instruction Set Computer, RISC*):
  - *POWER* (*IBM*);
  - *ARM* (много производителей).
- Сверхдлинное командное слово (*Very Long Instruction Word, VLIW*):
  - Эльбрус (МЦСТ).

✓ Специализированные ядра

Типов и классов специализированных ядер достаточно много. С одной стороны, обратим внимание на то, что

математические сопроцессоры и сопроцессоры ввода-вывода остались в истории, а их функциональность теперь реализуется в универсальных ядрах. С другой стороны, создаются новые перспективные ускорители – например, квантовые. Немаловажен тот факт, что специализированные ядра, помимо их эффективности для определенного класса задач, являются также и энергоэффективными.

Специализированные ядра (типичная роль – ускоритель, сопроцессор, "тонкое" ядро) могут быть классифицированы следующим образом:

- Гомогенные (на базе упрощенных универсальных) – например, *Intel Xeon Phi* [3].
- Вычислительные (графические) ускорители (*GPU, GPGPU*) – например, *Tesla* от *NVIDIA* [4].
- Узкоспециализированные (оптическое быстрое преобразование Фурье).
- Тензорные ускорители (матричное умножение и свертка) – например, *TPU* от *Google* [5].
- Нейроморфные (самообучающиеся) – например, *Loihi* от *Intel* [6].
- Ускорители алгоритмов работы машинного зрения – например, *Movidius* от *Intel*.
- Квантовые (криптография, искусственный интеллект, молекулярное моделирование) – например, *Tangle Lake* от *Intel* [7].
- Перепрограммируемые (*FPGA*) – например, *Intel Stratix 10 SX FPGA* [8].

✓ **Точность вычислений**

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, и здесь интересную роль играет точность вычислений:

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

Можно выделить несколько основных аспектов, общих для всех перечисленных научных и инженерных направлений:

- Очень широкий спектр временны́х и пространственных масштабов, плюс сложные, нелинейные пересечения множества биологических и физических процессов. Всё это требует качественного вычислительного моделирования, над которыми должны работать объединенные исследовательские группы специалистов из нескольких научных областей.
- Огромные объемы 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, представленные в отчете *DOE* [9]), что приблизительные вычислительные требования ряда задач к 2025 году возрастут в 100–1000 раз.

3. Компоненты вычислительного узла суперкомпьютера

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

Каждый узел оснащается следующими компонентами:

1 **Процессоры с универсальными ядрами**

Такие процессоры берут 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.

2 **Ускорители со специализированными ядрами**

Ускорители, в первую очередь, должны быть ориентированы на эффективное решение задач анализа данных; на вычислительных задачах они привлекаются для проведения специализированных вычислений. Кроме того, применение специализированных ускорителей должно способствовать достижению необходимой производительности (укладываясь при этом в разумные границы энергопотребления).

3 **Материнская плата и другие платы.**

4 **Оперативная память (ОЗУ)**

В связи с возросшими объемами данных и актуальностью задач по обработке, классификации и анализу больших данных, к памяти на узле предъявляются достаточно серьезные требования, касающиеся объема и пропускной способности. Возможно использование иерархических решений с высокопроизводительной оперативной памятью.

5 **Энергонезависимая память.**

6 **Внутриузловой (внутренний) интерконнект.**

7 **Высокопроизводительная фабрика (интерфейс передачи данных).**

8 **Контроллеры**

Контроллеры обеспечивают эффективное управление элементами инженерной инфраструктуры.

Простейшая классификация вычислительных 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.

3.1. Универсальные ядра

При выборе процессоров следует рассматривать зарекомендовавших себя производителей, выпускающих процессоры общего назначения с крупными ядрами. В настоящее время здесь можно ориентироваться на данные международного суперкомпьютерного рейтинга *Top500* и решения, разрабатываемые при создании прототипов суперкомпьютеров экзафлопного класса (экзаскейл).

Рассмотрим наиболее показательные архитектуры, опираясь на данные списка *Top500* за ноябрь 2018 года [10] и другие источники:

- Наибольшая часть попавших в рейтинг высокопроизводительных систем (46.6%) построена на базе *Intel*

Xeon E5 (*Broadwell*). Второе место занимает *Xeon Gold* (19.8%), третье– *Intel Xeon E5* (*Haswell*) с показателем 14.2%.

• Первое и второе места в мире занимают, соответственно, суперкомпьютеры *Summit* и *Sierra*, разработанные в США. Они базируются на процессорах *IBM POWER9*, которые поддерживает самые передовые технологии внутриузлового интерконнекта, включая *NVIDIA NVLink, OpenCAPI* и *PCIe Gen4*. Каждый узел суперкомпьютера *Summit* оснащен двумя процессорами *IBM POWER9* и шестью ускорителями *NVIDIA Tesla V100*.

• Третье место в мире занимает китайский суперкомпьютер *Sunway TaihuLight*, который базируется на процессорах *Sunway SW26010*. Эти разработанные в Китае процессоры с 64-битной архитектурой *RISC* изготавливаются по технологической норме 28 нм.

• Согласно заявлению *Dan Stanzione*, директора *TACC* [11], для *Pre-Exascale* суперкомпьютера *Frontera* выбор был остановлен на будущих процессорах *Intel Xeon SP Platinum* (*Cascade Lake*).

• В рамках европейского проекта *ExaNode* [12] разрабатывается базовый интегрированный узел на общей подложке – многоядерный процессор общего назначения на основе *ARMv8 CPU*, плюс набор ускорителей *FPGA* (*Field-Programmable Gate Array* – программируемая логическая интегральная схема, ПЛИС).

• Согласно открытым источникам, японский проект построения экзафлопсного суперкомпьютера *Post-K* [13] также базируется на *ARMv8*. Процессор *Post-K* – это вариант архитектуры *ARMv8-A*, но с 512-битным векторным расширением (*SVE)* с добавленным набором математических инструкций.

Сегодня выбор представленных на рынке архитектур и моделей процессоров с универсальными ядрами очень широк, поэтому к определению ключевых критериев и необходимых характеристик целевого вычислителя следует подходить тщательно. К базовым характеристикам процессоров относят производительность и энергопотребление. Но следует также учитывать и 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 др.), количество сокетов на плате и т.п. Все перечисленные аспекты оказывают существенное влияние на итоговую производительность целевого суперкомпьютера и удобство эксплуатации.

3.2. Специализированные ядра

Специализированные ядра, такие как ускорители и сопроцессоры, чаще всего имеют узкую область применения, но за счет этого позволяют получить рекордное соотношение производительность/энергоэффективность при решении целевых задач. Математические ускорители, как правило, интегрированы в основной процессор, остальные же добавляются в вычислительный узел при помощи высокоскоростных каналов данных.

Если анализировать мировые тенденции, то видно, что лидирующее место среди ускорителей занимают *GPGPU*:

• По данным статистики [10], 12.8% высокопроизводительных систем в рейтинге *Top500* за ноябрь 2018 года опираются на ускорители *NVIDIA Pascal* – это первое место в статистике ускорителей и сопроцессоров. Второе (9.2%) и третье (2.6%) места занимают системы с *NVIDIA Volta* и *NVIDIA Kepler* соответственно. На четвертом месте появляются системы на базе *Intel Xeon Phi* (6 систем, 1.2%).

• *Pre-Exascale* суперкомпьютер *Summit*, занимающий лидирующую позицию в рейтинге, также использует *GPGPU*. Каждый его вычислительный узел оснащен шестью ускорителями *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 каналов ввода-вывода, и гибкий настраиваемый мелкозернистый параллелизм.

В настоящее время доступен широкий спектр решений на базе *FPGA* – например, *Xilinx 7-Series FPGAs* [14]. Программируемая вентильная матрица *Intel Stratix 10 SX FPGA*, выпущенная в 2018 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. К примеру, уже созданы оптические процессоры, которые позволяют выполнять быстрое преобразование Фурье практически мгновенно [15].

Использование ускорителей открывает возможность добиться необходимой производительности целевого суперкомпьютера и уложиться в разумный энергетический бюджет. Выбор ускорителей, прежде всего, зависит от специфики задач. В настоящее время при создании топовых суперкомпьютеров наблюдается тенденция к разработке многоцелевых машин, предназначенных для широкого спектра задач, включая вычислительное моделирование в различных областях науки и обработку больших объемов данных. Таким образом, при выборе ускорителей необходимо учитывать потенциальную возможность их применения для решения большинства целевых задач и наличие стека программного обеспечения, который позволит эффективно использовать полученный гетерогенный суперкомпьютер.

3.3. Память

3.3.1. Оперативная память (ОЗУ)

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

Например, для решения задач экзафлопсного масштаба потребуются и значительные объемы оперативной

памяти на узле. В статье разработчиков аппаратно-программной платформы "Эльбрус" [16] формулируются следующие требования: 5 *PB* оперативной памяти с пропускной способностью 4 *TB/s*. По оценке *Coral-2*, объем оперативной памяти должен быть не менее 8 *PB* [17]. Согласно докладу *Al Gara* (*Intel Fellow, Data Center Group*), оперативная память системы должна удовлетворять следующим требованиям: объем 6÷12 *PB*, пропускная способность 100÷200 *PB/s;* для энергонезависимой памяти – 10÷100 *TB/s* [18].

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. Это самая распространенная технология, она поддерживается большинством процессоров. В настоящий момент лидирующим стандартом является *DDR4*, а в 2020 году ожидается выпуск *DDR5* [19].
- *3D Stacked Memory* – технология трехмерного (многослойного) размещения памяти, которая позволяет интегрировать ОЗУ и логические блоки микропроцессора, тем самым существенно увеличивая пропускную способность [20]. Следует отметить, что плотность компоновки *3D Stacked Memory* требует специального подхода к охлаждению – либо специальная организация воздушных потоков, либо жидкостное охлаждение. Предлагаются следующие типы:

– *HBM* (*High Bandwidth Memory*) – память с высокой пропускной способностью. Эта непланарная (неплоская) память имеет трехмерную конструкцию в виде куба или прямоугольного параллелепипеда, где несколько микросхем памяти сложены друг на друга. Благодаря этому уменьшается площадь, занимаемая чипами памяти, что делает возможным её размещение в непосредственной близости к графическому процессору [21]. На текущий момент лидирующим является поколение *HBM2*;

– *HMC* (*Hybrid Memory Cube)* обеспечивает пропускную способность до 480 *GB/s* на устройство, но обладает лимитированным объемом – до 8 *GB*, согласно стандартам, определенным консорциумом [22].

- *SSD DIMM* – твердотельные (*SSD*) накопители с интерфейсом *DIMM;* также есть варианты накопителей *3dX Point* с интерфейсом *DIMM*. Такое решение может использоваться в качестве дополнительного уровня ОЗУ. Достоинством этого подхода является большой объем, а узким местом всё еще остается допустимое 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. Достоинством данного подхода является большой объем, а к недостаткам можно отнести небольшую пропускную способность по сравнению с *DDR*. Примером может являться технология *IMDT* [23] на базе *3DXPoint NVMe* накопителей.
- *GDDR* (*Graphics Double Data Rate*) – графическая память, предназначенная для использования в видеокартах. Представляет собой подвид энергозависимой динамической памяти с произвольным доступом (*DRAM*) и удвоенной скоростью передачи данных (*DDR*). В настоящее время доступны стандарты *GDDR5* [24] и *GDDR6* [25].

Наиболее популярным в настоящее время остается стандарт *DDR-SDRAM*, поскольку именно он поддерживается большинством процессорных архитектур. Технология *3D Stacked Memory* позволяет значительно увеличить пропускную способность, но лимитирует объем. Большой интерес представляет построение памяти с многоуровневой иерархией, что позволяет объединять технологии *3D Stacked Memory* и *DDR-SDRAM* или *DDR-SDRAM* и *SDM* для получения ОЗУ большого объема с высокой пропускной способностью.

3.3.2. Энергонезависимая память

Хотя анализ организации систем хранения выходит за рамки данной статьи, необходимо учесть, что на самóм узле должны быть 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. Сегодня 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* [26]) для работы с данными подразумевает наличие твердотельных накопителей (в том числе, устройств *SSD* с интерфейсами *DIMM* и *PCI-e* для повышения пропускной способности), что необходимо 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.

Кроме того, используется специальное ПО, которое позволяет в фоновом режиме подкачивать в *Burst Buffer* данные для обработки, и перемещать полученные результаты в постоянное хранилище. Примером может являться *Burst Buffer*, организованный на суперкомпьютере *NERSC Cori* [27].

Полная интеграция уровня хранения с уровнем обработки (гиперконвергентность) – такой подход практикуется в больших, гипермасштабируемых центрах обработки данных (ЦОД), таких как *Google*, *Facebook* или *Amazon Web Services*. Многие современные *HPC*-платформы видят будущее именно в создании гиперконвергентных решений [28].

Группа компаний РСК успешно разработала и продемонстрировала на суперкомпьютерной выставке ***ISC'18*** гиперконвергентный вычислительный узел "РСК Торнадо" с прямым жидкостным охлаждением и твердотельными дисками *Intel* [29].

3.4. Внутриузловой интерконнект

Для обеспечения эффективного взаимодействия все процессоры, ускорители и память должны быть архитектурно интегрированы. В настоящее время внутриузловой интерконнект становится главным узким местом серверных узлов. Для создания современного высокопроизводительного узла необходимо рассмотреть перспективные варианты организации внутриузлового интерконнекта, различные топологии и технологии кластеризации (*die-stacking*).

Возможные варианты объединения можно классифицировать по применяемому подходу:

1. Объединение высокоскоростными интерфейсами.
2. 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 – это позволяет разработчику серверного решения интегрировать топовые решения от ведущих мировых производителей. Доминирует в этой области стандарт *PCIe*.

✓ **Стандарт *PCIe:***

- *PCIe 3.0* – пропускная способность порядка 1 *GB/s* на одиночную линию, интегрированная пропускная способность на 16 линий может достигать 32 *GB/s* в двух направлениях. Этот стандарт поддерживается большинством производителей процессоров, сопроцессоров, а также устройств ввода-вывода;
- *PCIe 4.0* – обновленный стандарт, обеспечивающий двукратное увеличение скорости передачи данных (до 2 *GB/s* на одиночную линию и до 64 *GB/s* на 16 линий). Поддерживается только некоторыми современными моделями процессоров, но список устройств с поддержкой *PCIe 4.0* расширяется с каждым годом.

Помимо *PCIe* существуют и альтернативные решения: в первую очередь, это новые открытые стандарты *OpenCAPI* и *Gen-Z*, а также частные решения – *NVIDIA NVLink, IBM Infinity Fabric, Intel UltraPath* и т.п. Рассмотрим подробнее некоторые альтернативные решения, используемые наиболее часто.

✓ **Альтернативные открытые стандарты:**

- *Gen-Z* [30] – открытая технология (стандарт опубликован в 2018 году), которая позволяет получить пропускную способность 32 *GB/s* на один канал, а на несколько каналов – вплоть до 400 *GB/s*.
- *OpenCAPI* [31] – открытый стандарт, разработанный широким консорциумом участников. *OpenCAPI 3.0* обеспечивает пропускную способность до 25 *GB/s* на один канал, а число каналов может достигать восьми. Этот стандарт поддерживает процессор *IBM POWER9*.

✓ **Частные разработки**

Более широкий спектр вариантов предоставляют частные разработки компаний. К их недостаткам можно отнести ограниченную совместимость с устройствами других производителей.

Наибольшей популярностью пользуются:

- *NVIDIA NVLink* [32] – интерфейс, изначально разработанный для интеграции *NVIDIA GPU* и предоставления общей памяти, сейчас находит более широкий спектр применения, в том числе для организации высокоскоростных каналов *CPU – GPU*. К примеру, в топовом суперкомпьютере списка *Top500* за июнь 2018 года используется соединение *NVlink* между процессорами *IBM POWER9* и ускорителями *NVIDIA*. Пропускная способность (двунаправленная) одного линка *NVLink 2.0* достигает 50 *GB/s*, агрегированная пропускная способность шести линков (двунаправленная) – 300 *GB/s*.
- *Intel Ultra Path Interconnect* (*Intel UPI*) [33] – проприетарный канал для объединения двух процессоров *Intel Xeon Scalable*.
- *Infinity Fabric* [34] – интерконнект от *AMD*, который можно использовать 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 42.667 *GB*/s, общая двунаправленная пропускная способность – 170.667 *GB/s*.
- Российские микропроцессоры "Эльбрус-8С" поддерживают по три дуплексных канала с двунаправленной производительностью 16 *GB/s* [35].
- *Cavium Coherent Processor Interconnect CCPI2* – соединяет два процессора *Cavium*, обеспечивая производительность 600 *GB/s*.

Кроме того, есть ряд решений, которые являются расширениями *PCIe* – например, *CCIX* [36] и *IBM CAPI* [37].

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*.

Следует принимать во внимание и развитие перспективных технологий, обещающих увеличить пропускную способность на несколько порядков. Особенный интерес в данном контексте представляет передача информации при помощи полупроводниковых лазеров. Потенциально эта технология может найти применение как для быстрой передачи данных внутри узла, так и для организации сверхбыстрого 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 [38].

4. Инфраструктура узла современного суперкомпьютера

4.1. Охлаждение

С целью увеличения вычислительной плотности, для охлаждения современных суперкомпьютеров следует использовать жидкостное охлаждение, а наиболее перспективным направлением считается применение плотно прилегающих охлаждающих пластин с хладагентом. В настоящее время жидкостным охлаждением занимается огромное количество компаний.

Рассмотрим наиболее распространенные подходы [39] к организации жидкостного охлаждения:

- Использование охлаждаемых жидкостью пластин (*Coldplates*), которые полностью покрывают всю элементосодержащую поверхность вычислительного узла. Такую технологию взяли на вооружение компании *Aquila* [40] и *Dell*. В России лидером по разработке вычислительных кластеров, охлаждаемых колдплейтами, является группа компаний РСК [41].
- Индивидуальные теплообменники (бобышки) – специальные элементы, позволяющие подводить охлаждающую жидкость напрямую к индивидуальным компонентам вычислительного узла. Примеры компаний: *Asetek* [42], *Ebullient* [43].
- Жидкостное охлаждение на уровне шкафа. Примеры компаний: *CoolIT, Inspur.*
- Погружные (иммерсионные) системы – разработки, подразумевающие полное погружение вычислительных узлов в специальную диэлектрическую жидкость. Примеры компаний: *LiquidCool Solution* [44], *ExaScaler Inc.* [45]. Представителями этого подхода в России являются система *IMMERS* [46], а также погружные системы охлаждения реконфигурируемых вычислительных систем на основе ПЛИС [47].

Имеются и другие разработки. Например, компания *Liquid MIPS* [48] 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, этот метод обеспечивает более высокую надежность и низкое энергопотребление.

4.2. Датчики и сенсоры

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

Установленные датчики и сенсоры должны обеспечивать [49]:

- 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 является применение встроенных датчиков, информацию с которых можно получить через интерфейс *IPMI* (*Intelligent Platform Management Interface*), который опрашивает *BMC* (*Board Management Controller*). Другой способ – использование средств мониторинга и программных моделей, предоставляемых производителями (к примеру, *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 удобный пользовательский интерфейс для мониторинга датчиков. Крайне желательно наличие унифицированного программного интерфейса (*API*), который позволит отображать информацию с системных датчиков и сенсоров в едином стиле [50].

При наличии соответствующей аппаратной и программной инфраструктуры, возможно обеспечить "умное" управление кластером – в том числе, отслеживать потенциальные сбои по изменению температурных характеристик и энергопотребления, использовать адаптивные алгоритмы балансировки нагрузки.

5. Заключение

Некоторая часть суперкомпьютеров создается для специфических задач, но большинство из них должно быть готово к задачам из разных прикладных областей. Основным фактором для быстрого решения конкретных задач на суперкомпьютерах является наличие не только соответствующей аппаратной части, но и эффективного прикладного программного обеспечения, функционирующего на данной аппаратуре.

Определим несколько основных классов задач и предложим для них обобщенные варианты архитектур узлов:

1 Тяжелые вычислительные задачи

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

2 Обработка больших данных

Данный класс задач для эффективной работы требует применения гиперконвергентных узлов или высокопроизводительной системы хранения данных, построенной с использованием специальных технологий организации промежуточных буферов между системой хранения и вычислительными узлами.

3 Машинное обучение и глубокое обучение

Для таких 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.

4 Задачи из некоторых специфических областей

Здесь мы можем рекомендовать узлы на базе универсальных процессоров *RISC* или *CISC* и дополнительных ускорителей *FPGA*, реализующих алгоритмы, ориентированные на данные задачи.

Важнейшими для узла характеристиками, на основе которых должен производиться детальный выбор узла, являются:

- наличие необходимого прикладного ПО для данных архитектур;
- теоретическая производительность узла;
- максимальное энергопотребление узла;
- стандартизация и модульность компонентов узла (как с точки зрения замены вышедших из строя компонентов, так и с позиции обновления узлов).

В заключение обсудим некоторые технологические проблемы. Прежде всего, это исчерпание потенциала "закона Мура" (точнее, бизнес-прогноза Гордона Мура), вызванное тем, что проектная норма технологических процессов подходит к физически допустимому пределу. Согласно докладу [51], прогресс в компьютерной отрасли могут обеспечить 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:* сегодня доступны процессоры и соответствующий стек ПО для различных *RISC*- и *VLIW*-архитектур. Следует также отметить прогресс в области иерархической компоновки 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 применения новых материалов и структур (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. Требуется разрешить ряд противоречий – например, между необходимостью глубокой аппаратно-зависимой оптимизации ПО и необходимостью портирования этого ПО на широкий диапазон разноархитектурных вычислителей. Еще одно противоречие заключается в том, что необходимо разработать новую сверхмасштабируемую парадигму параллельных вычислений, которая позволит эффективно использовать вычислительные суперкомпьютеры будущего, но при этом обеспечить и портирование на эти же суперкомпьютеры основных существующих вычислительных пакетов.

Прогресс в области повышения производительности машинных вычислений в ближайшем будущем 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 этот может быть неравномерным, зависеть от предметной области, алгоритмической специфики и соответствующих архитектурных требований, а также от степени успешности адаптации (или разработки с нуля) специализированных программных пакетов для современных суперкомпьютеров.

Литература

1. Описание проекта РАН: Создание вычислительной системы для моделирования суперкомпьютера с производительностью экзафлопсного уровня // *www.keldysh.ru/projects/exaflops.pdf*
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 (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 // *https://spectrum.ieee.org/tech-talk/computing/hardware/intels-49qubit-chip-aims-for-quantum-supremacy*
8. Intel Stratix 10 SoC FPGAs // *www.intel.com/content/www/us/en/products/programmable/soc/stratix-10.html*
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) // *https://press3.mcs.anl.gov//hepfce/files/2016/11/DOE-ExascaleReport-HEP-Final.pdf*

10. Top500 List Statistics. Release November 2018 // www.top500.org/statistics/list

11. Hemsoth N. Cascade Lake at Heart of 2019 TACC Supercomputer. 2018 // https://www.nextplatform.com/2018/08/29/cascade-lake-heart-of-2019-tacc-supercomputer

12. Bartsch V. D6.3 Initial Project Press Release // ExaNoDe Consortium Public deliverable 2016 // exanode.eu/wp-content/uploads/2017/04/D6.3.pdf

13. ARMv8 – A Scalable Vector Extension for Post-K. Fujitsu Limited, 2016 // https://www.fujitsu.com/global/Images/armv8-a-scalable-vector-extension-for-post-k.pdf

14. Xilinx. High Performance Computing and Data Storage // www.xilinx.com/applications/high-performance-computing.html.

15. Timmel A.N., Daly J.T. Multiplication with Fourier Optics Simulating 16-bit Modular Multiplication // arxiv.org/pdf/1801.01121.pdf

16. 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.

17. CORAL Collaboration: Briefing on CORAL-2 RFP and Draft Technical Requirements // Vendor Webinar Meeting, December 6, 2017 //procurement.ornl.gov/rfp/CORAL2/Brief-of-Draft-SOW-20171206-SA.pdf

18. HPC and AI – Two Communities Same Future // www.hpcwire.com/2018/01/25/hpc-ai-two-communities-future

19. JEDEC DDR5 & NVDIMM-P Standards Under Development (March 30, 2017) // www.jedec.org/news/pressreleases/jedec-ddr5-nvdimm-p-standards-underdevelopment

20. 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 (October 1–3, 2017, Seattle, WA, USA), p. 66–75 // *DOI: 10.1109/IISWC.2017.8167757*

21. High Bandwidth Memory (HBM) DRAM. JESD235A. Nov 2018 // www.jedec.org/standards-documents/docs/jesd235a

22. Hybrid Memory Cube (HMC) Consortium // https://en.wikipedia.org/wiki/Hybrid_Memory_Cube

23. Intel Memory Drive Technology Application Note. Document Number: 33 5925-001 US. May 2017, Revision 001 // https://www.intel.com/content/dam/support/us/en/documents/memory-and-storage/intel-mem-drive-tech-appnote.pdf

24. Graphics double data rate (GDDR5) SGRAM standard. JESD212C. February 2016 // www.jedec.org/standards-documents/docs/jesd212c

25. Graphics Double Data Rate 6 (GDDR6) SGRAM Standard. JESD250A. July 2017 // www.jedec.org/standards-documents/docs/jesd212c

26. 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 (November 12–17, 2017, Denver, CO, USA). New York, USA: ACM, article 2, 9 p. // *DOI: 10.1145/3150994.3151000*.

27. 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 // escholarship.org/uc/item/9qv6k14t

28. Morgan T.P. For Many Hyperconverged is the Next Platform, 2018 // https://www.nextplatform.com/2018/01/29/hyperconverged-next-platform-many-jobs

29. РСК представила гиперконвергентное HPC-решение на новейших компонентах // https://www.cnews.ru/news/line/2018-06-27_rsk_predstavila_giperkonvergentnoe_hpcreshenie

30. The GEN-Z Consortium // https://genzconsortium.org

31. The OpenCAPI Consortium // https://opencapi.org

32. NVLink Fabric // www.nvidia.com/ru-ru/data-center/nvlink

33. Краткое описание продукции: платформа масштабируемых процессоров Intel Xeon // www.intel.ru.

34. Infinity Fabric (IF) – AMD // en.wikichip.org/wiki/amd/infinity_fabric

35. Микропроцессоры серии "Эльбрус", каталог продукции // http://mcst.ru/files/59db45/cf0cd8/50a21b/000000/katalog_produktsii_mtsst_hq.pdf

36. CCIX Consortium // www.ccixconsortium.com

37. Coherent Accelerator Processor Interface (CAPI) // https://developer.ibm.com/linuxonpower/capi

38. Шустиков В. Ученые Сколтеха создали суперкомпьютер "Жорес" // Фонд "Сколково", 18 января 2019 // https://sk.ru/news/b/pressreleases/archive/2019/01/18/uchenye-skolteha-sozdali-superkompyuter-zhores.aspx

39. Is Liquid Cooling Ready to Go Mainstream? // www.hpcwire.com/2017/02/13/liquid-cooling-ready-go-mainstream

40. Aquila // www.aquilagroup.com/cooling

41. Группа компаний РСК // www.rscgroup.ru/ru

42. Asetek // www.asetek.com

43. Ebullient // http://ebullientcooling.com

44. Liquid Cooled Servers // www.liquidcoolsolutions.com

45. ExaScaler Inc. Overview // www.exascaler.co.jp/en/company

46. 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" (Дивноморское, Геленджик, 29 сентября – 4 октября, 2014 г.), с. 9–15.

47. 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 // https://cyberleninka.ru/article/n/pogruzhnaya-sistema-ohlazhdeniya-rekonfiguriruemyh-vychislitelnyh-sistem-na-osnove-plis.

48. Liquid MIPS // www.liquidmips.com/cms/en-us/howit-works.aspx

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

50. 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*

51. Shalf J.M., Leland R. Computing beyond Moore's Law // Computer, 2015, vol. 48, No. 12, p. 14–23 // *DOI: 10.1109/MC.2015.374*