Программа создания перспективных суперкомпьютеров

Чтобы устранить нарушенный паритет с Японией в области высокопроизводительных вычислений, Министерство обороны США реализует сегодня программу создания суперкомпьютеров с перспективной архитектурой для решения стратегически важных государственных задач обеспечения национальной безопасности, а также для нужд промышленности. Программа, как это уже не раз случалось в истории ИТ, безусловно, окажет влияние на отрасль и определит дальнейшее развитие индустрии суперкомпьютеров.

Александр Фролов, Александр Семенов, Антон Корж, Леонид Эйсымонт

Программа DARPA HPCS (high productivity computing systems) определяет, что новые суперкомпьютеры должны: обладать реальной производительностью на задаче Linpack в несколько петафлопс (PFLOPS, 10^{15} операций с плавающей точкой в секунду); иметь глобально адресуемую оперативную память объемом несколько петабайт с пропускной способностью, на 4–5 порядков превосходящей современный общедоступный уровень; обладать в десять раз более простой системой программирования, чем современные системы и технологии разработки параллельных программ. Впрочем, основные задачи программы DARPA HPCS были известны еще до ее появления [1, 2], однако реальных шагов не предпринималось. Гром грянул 27 февраля 2001 года, когда компания Cray отозвала свою жалобу на японскую компанию NEC по поводу продажи последней по демпинговым ценам на американском рынке своих векторных суперкомпьютеров. Ценой отзыва стала выплата японцами компании Cray 25 млн. долл. и предоставление этой американской компании исключительного права на продажу японских суперкомпьютеров [3].

Рядовое, казалось бы, событие в мире бизнеса привело отвечающих за высокотехнологичную промышленность американских чиновников в состояние шока. Векторные суперкомпьютеры имеют ключевое значение для национальной безопасности, поскольку позволяют эффективно работать с оперативной памятью огромного объема. Это необходимо для криптоанализа, разработки и поддержания боеготовности ядерных арсеналов, создания новых вооружений, оценки эффективности ведения боевых действий, прогноза погоды и т.п. [2].

Для этого класса задач, требующих особо эффективной организации работы с памятью большого объема, характерна работа с регулярно или предсказуемо нерегулярно расположенными в памяти элементами векторов. Однако даже векторные процессоры не могли успешно осуществлять интенсивную непредсказуемо нерегулярную работу с памятью большого объема. Согласно современной терминологии, речь идет об интенсивной нерегулярной работе с памятью, имеющей плохую пространственно-временную локализацию [4]. Значение этого класса задач сегодня еще больше возросло в связи с развитием вычислительных методов, боевых и гражданских информационно-управляющих систем, необходимостью предотвращения террористических операций, актуальностью анализа данных и извлечения знаний в таких приложениях, как обработка научной и разведывательной информации.

Даже рядовое соглашение между Cray и NEC по поводу векторных процессоров фактически означало, что американская промышленность по перечисленным критическим направлениям не способна самостоятельно обеспечить национальную безопасность страны.

После анализа ситуации в США на государственном уровне было решено организовать программу DARPA HPCS, направленную на интенсивные исследования и разработки в области **суперкомпьютеров стратегического назначения (СКСН)**. Составлением программы занялись, как это обычно бывает при появлении особо перспективных разработок, военные и специалисты Агентства национальной безопасности США. Работа началась в конце апреля 2001 года [3], а к первой фазе программы приступили уже в июле 2002 года. К выполнению программы DARPA HPCS были привлечены компании IBM, Cray, SGI, HP и Sun Microsystems. Всего планировались три фазы проекта, с конкурсом среди участников при переходе с одной фазы на другую. На завершающем этапе должны были остаться две компании.

СТРАТЕГИЧЕСКИЕ И МАССОВЫЕ СУПЕРКОМПЬЮТЕРЫ: ПРОБЛЕМЫ И ЗАБЛУЖДЕНИЯ

В число участников программы DARPA HPCS не вошла компания Intel, и, насколько можно было понять из выступлений ее руководителей (например, Джастина Раттнера), это вызвало некоторую обиду. Компания организовала собственную аналогичную программу ACP (Advanced Computing Program), чтобы продемонстрировать, кто на самом деле является основным производителем компонентов суперкомпьютерных систем и суперкомпьютеров. Основания для такого утверждения имеются — так, львиная доля систем из списка Top500 используют элементную базу Intel. На этот аргумент, впрочем, стоит обратить внимание не только в связи с позицией Intel по отношению к DARPA HPCS: Top500 отражает состояние дел в суперкомпьютерной области весьма неоднозначно, часто вводя в заблуждение. Почему так происходит и о каких, собственно, суперкомпьютерах стратегического назначения идет речь в DARPA HPCS?

www.osp.ru

Top500: числом или уменьем?
http://www.osp.ru/os/2005/10/380430

Рынок вычислительных средств повышенной производительности принято делить на четыре сегмента: *высший* (technical capability) — рекордные по быстродействию стратегические вычислительные средства, установленные в крупнейших государственных лабораториях и центрах, решающих важнейшие для государства задачи; три сегмента систем, менее чувствительных к реальной производительности — *корпоративные системы* (technical enterprise), *системы уровня отделения* (technical divisional) и *системы уровня отдела* (technical department). Пиковая производительность СКСН, образующих первый сегмент, сегодня приближается к 1 PFLOPS (на момент принятия программы HPCS — к 10 TFLOPS). При этом подразумевается, что в любое время, любой ценой, усилиями мощных коллективов алгоритмистов и программистов из этих суперкомпьютеров будет извлечена высокая реальная производительность для решения конкретных задач государственного значения. Однако проблема в том, что достичь нормальной реальной производительности становится

все сложнее из-за изменений элементной базы и классов задач.

Вместе с тем, по оценкам 1999 года, мировой рынок всех четырех сегментов составлял всего лишь 5,617 млрд. долл., из них рынок стратегических систем — 934 млн. долл. Рост на ближайшие пять лет стратегических систем первого сегмента оценивался тогда в 6,7% в год, а остальных систем — в 9,3%. Доля векторных суперкомпьютеров в высшем сегменте составляла лишь около 500 млн. долл. Таким образом, СКСН, о которых идет речь в программе DARPA HPCS, — явно не рыночный товар. Но насколько успешна на рынке ядерная бомба? А авианосец? СКСН — изделие как раз такого типа, к тому же он не только является видом оружия, но и средством его разработки.

Теперь вернемся к списку Top500, который до сих пор остается незаслуженно популярным. Этот список составляется по значениям реальной производительности, развиваемой на тесте Linpack с хорошей пространственно-временной локализацией. Этот тест не требует наличия эффективной подсистемы памяти, поэтому в Top500 стратегических суперкомпьютеров мало (их вообще не так много). Список Top500 отражает коммерческие успехи лишь в трех последних сегментах рынка высокопроизводительных систем. К слову, рынок даже этих сегментов мал в сравнении с рынком серверных платформ объемом 50 млрд. долл., на котором активно работает Intel. Конечно, высший сегмент заманчив и престижен, но при создании СКСН слишком высоки интеллектуальные и экономические затраты, а особой непосредственной прибыли нет, разве что создаются новые суперкомпьютерные технологии, которые потом уходят в остальные сегменты.

В России сегодня нестратегические сегменты рынка суперкомпьютеров развиваются неоправданно активно; при этом широко используются компоненты, которые предлагают зарубежные компании, работающие в сегменте серверных платформ. В сравнении с заказами в сотни миллионов рублей по созданию вычислительных кластеров преимущественно для университетов (среди них идет соревнование за место в российском аналоге Top500), не имеющих достаточного количества задач для этих систем, финансирование на создание отечественных СКСН ничтожно. Безусловно, польза от создания кластеров есть — молодежь обучается работе на суперкомпьютерах, но очевидный перекос в расходовании средств, недопонимание ограничений, присущих массовым суперкомпьютерам, и игнорирование исключительной важности СКСН для государства могут привести к печальным последствиям. Работа по исправлению ситуации сегодня ведется, но недостаточными темпами. До сих пор продолжаются дискуссии о целесообразности для России программы по созданию СКСН: «не рановато ли», «нужно ли вообще», «не лучше ли догнать производителей микропроцессоров» и т.п.

В США обсуждение научно-технической и военно-политической значимости стратегических суперкомпьютеров прекратилось после разразившегося в середине 2002 года скандала, связанного с появлением японской машины Earth Simulator. Этот многопроцессорный векторный суперкомпьютер на базе NEC SX-6 почти в десять раз превосходил лучшие американские суперкомпьютеры, причем при выполнении критически важных для государства задач. Поистине, новый Перл-Харбор! После оценки «масштабов бедствия» [9], 30 ноября 2004 года был принят закон о крупномасштабном возрождении в Министерстве энергетики США работ по стратегическим суперкомпьютерам разного типа [10], а не только вычислительным кластерам, а затем и федеральный план работ по суперкомпьютерам высшего диапазона производительности [11].

Свое лидерство — хотя бы в списке Top500 — США удалось вернуть в ноябре 2004 года. Система IBM BlueGene/L и специализированный кластер Columbia на базе вычислительных узлов SGI Altix возглавили этот список. В классе векторных мультипроцессоров сильным ответом стал векторный суперкомпьютер с глобально адресуемой памятью Cray X1. Однако надо помнить, что фора в 10–15 лет в научно-технической области отыгрывается долго и, несмотря на все принятые в США меры, ситуация с Earth Simulator может повториться. 25 октября нынешнего года было объявлено о выпуске векторной системы NEC SX-9 с глобально адресуемой памятью в 1 Тбайт для 16-процессорного узла. Всего система может содержать до 512 узлов. Сейчас США могут противопоставить ей только модифицированный BlueGene/L, систему 2007 года BlueGene/P и прототипы систем программы DARPA HPCS (Roadrunner [12] и Baker [6]), гибридные суперкомпьютеры от Cray, создаваемые в рамках переходного проекта Rainier. Смогут ли новые американские системы победить в этой гонке?

Накал страстей таков, что уже в ноябре нынешнего года компания Cray объявила о выпуске семейства суперкомпьютеров Cray XT5, в состав которых входят кластерные, векторные и реконфигурируемые узлы, объединенные новой коммуникационной сетью с топологией 3D-тор на базе новых маршрутизаторов SeaStar2+. Основная вычислительная мощность в Cray XT5 обеспечивается векторными узлами Cray X2, которые позволяют 1024 процессорам работать с глобально адресуемой памятью емкостью 16 Тбайт, но пока через коммуникационную сеть. Похоже, что Cray XT5 – переходная система, поскольку в ней заметны упрощенные подключения компонентов будущих систем BlackWidow [12] и Baker [6,8], а также имеются компоненты старых систем, о предстоящей замене которых уже известно (прежде всего это касается коммуникационной сети).

www.osp.ru

Динозавры эпохи микропроцессоров
http://www.osp.ru/os/2004/08/185070

ЦЕЛИ И ЗАДАЧИ ПРОГРАММЫ DARPA HPCS

Основные цели программы DARPA HPCS таковы.

1. Разработка нового поколения высокопродуктивных вычислительных

Таблица 1. Базовые характеристики СКСН, создаваемых по программе DARPA HPCS

1	Более 2 PFLOPS реальной производительности на тесте Linpack
2	~ 6,5 Пбайт/с на тесте пропускной способности памяти при регулярных обращениях (тест Stream)
3	~ 3,2 Пбайт/с на тесте бисекционной пропускной способности системной коммуникационной сети (тест BISECT)
4	64000 GUPS на тесте нерегулярного доступа к памяти (тест RandomAccess)
5	Высокий полиморфизм — параллелизм типа ILP, TLP и DLP (векторная и потоковая обработка)
6	Высокая реконфигурируемость и адаптируемость к задачам
7	Распределенная общая память
8	Увеличение продуктивности программирования по отношению к уровню 2005 года в десять раз

Примечание. ILP (instruction level parallelism) — параллелизм выполнения машинных команд программы; TLP (thread level parallelism) — параллелизм выполнения легковесных процессов программы; DLP (data level parallelism) — параллелизм выполнения операций над множеством данных одновременно.

систем и программного обеспечения, призванных преодолеть проблемы низкой реальной производительности, плохой масштабируемости, высокой сложности программирования, а также физических ограничений.

2. Ликвидация разрыва в суперкомпьютерных технологиях. Современные технологии базируются на достижениях конца 80-х годов, а создаваемые технологии типа квантовых вычислений имеют слишком отдаленную перспективу внедрения.

3. Разработка к 2010 году рентабельных высокопродуктивных вычислительных систем для решения стратегических задач национальной безопасности и промышленности, отвечающих следующим требованиям:

- реальная производительность должна быть в 10–40 раз выше, чем у существующих систем;
- время на программирование и общие затраты на разработку, наладку и сопровождение прикладных программ должны быть ниже нынешних СКСН в десять раз;
- для достижения приемлемой переносимости программная реализация должна быть отделена от архитектуры вычислительной системы;
- должна быть обеспечена повышенная отказоустойчивость и надежная защита информации.

4. Привлечение к работам широких слоев академической общественности с целью восстановления в США потерянной инфраструктуры фундаментальных исследований и разработок. Воспитание нового поколения специалистов для работы в области перспективных СКСН.

Основными областями использования создаваемых систем должны стать обработка разведывательной информации, средства наблюдения и слежения, криптоанализ, разработка вооружений, моделирование распространения примесей в атмосфере, биотехнологии.

www.osp.ru

Будущее высокопроизводительных вычислительных систем
http://www.osp.ru/os/2003/05/183009

УТОЧНЕННЫЕ ХАРАКТЕРИСТИКИ СОЗДАВАЕМЫХ СУПЕРКОМПЬЮТЕРОВ

После выполнения первой фазы программы появились уточнения. Они приведены в таблице 1, а в таблице 2 дано сравнение характеристик создаваемых СКСН с лучшими системами 2007 года.

Что же означает создание систем с указанными характеристиками? Чтобы ответить на этот вопрос, сравним их с вычислительными кластерами, которые легко собираются из доступных на рынке компонентов и которых так много в Top500 и в российском Top50.

Один из базовых компонентов вычислительных кластеров — это коммерчески доступный микропроцессор, высокий показатель реальной производительности которого достижим лишь в случае, когда удается эффективно использовать его небольшую быструю кэш-память. Это достижимо на задачах с хорошей пространственно-временной локализацией, например, на Linpack. Однако на задачах с плохой пространственно-временной локализацией кластерные системы развивают крайне низкую реальную производительность — 0,1–1% от пиковой даже на одном вычислительном узле, а на задачах со средней пространственно-временной локализацией — 5–10% от пиковой. Чтобы убедиться в этом, достаточно запустить на таком кластере задачи из области аэрогидродинамики пакета NPB 3.1. Каждый тест после своего выполнения выдает данные о достигнутой реальной производительности, которая определяется как количество необходимых для решения задачи содержательных операций (например, операций над числами с плавающей точкой), деленное на время решения этой задачи.

Обычно выделяют четыре главных причины такой низкой реальной производительности («четыре всадника Апокалипсиса», как их назвал специалист по архитектуре суперкомпьютеров Томас Стерлинг [5]):

- *latency* — задержки выполнения операций (по большей части с памятью и коммуникационной сетью, от сотен до тысяч тактов процессора);
- *overhead* — накладные расходы, связанные с организацией выполнения содержательных для задачи операций;

Таблица 2. Сравнение характеристик создаваемых к 2010 году СКСН и характеристик лучших современных систем

	Характеристика	Существующие системы		Перспективные СКСН (США, 2010 год)
		IBM BlueGene/L	Cray XT3	
	Производительность на тесте Linpack, TFLOPS	260	101	> 2000
	Пропускная способность памяти при регулярных обращениях, Тбайт/с	128	196	> 6500
Тесты ОВ	Бисекционная пропускная способность системной коммуникационной сети, Тбайт/с	0,36	11,7	> 3200
Тесты ОВ	Тест нерегулярного доступа к памяти, GUPS	35	29	> 64000

Примечание: ОВ — тесты особой важности, их показатели отражают уровень реальной производительности.

• *contention* — конкуренция за совместно используемые ресурсы;
• *starvation* — «голодание», простаивание из-за недостаточного параллелизма и асинхронности, отсутствия сбалансированной загрузки.

ПРОСТРАНСТВЕННО-ВРЕМЕННАЯ ЛОКАЛИЗАЦИЯ И РЕАЛЬНАЯ ПРОИЗВОДИТЕЛЬНОСТЬ

Падение эффективности подсистемы памяти при ухудшении пространственно-временной локализации хорошо заметно на специальном тесте APEX-MAP [4], который строит некоторую APEX-поверхность, характеризующую эффективность выполнения операций с памятью. Такая поверхность может быть построена для одного или множества узлов. Для одного узла строится зависимость от пространственно-временной локализации среднего количества тактов процессора, за которое выполняется одно обращение к памяти на считывание. Для множества узлов обычно строится приведенная к одному вычислительному узлу пропускная способность памяти, также в зависимости от пространственно-временной локализации. Примеры APEX-поверхностей приведены на рис. 1, на нем также показана область CF-задач (дружественные к использованию кэш-памяти задачи, реальная производительность при их выполнении составляет 60–90%) и DIS-задач (недружественные к использованию кэш-памяти задачи, реальная производительность на которых составляет 5–10% от пиковой и ниже). Также показаны точки, которые коррелируются по профилю обращений к памяти с известными тестами: Linpack (HPL), быстрое преобразование Фурье (FFT), тест Мак-Калпина по эффективности пересылок в памяти с регулярным доступом к ней (STREAM), тест интенсивного нерегулярного доступа к памяти (RandomAccess). Вообще говоря, любая задача может быть поставлена в соответствие определенной точке APEX-поверхности по некоторой методике нахождения наибольшей корреляции ее профиля обращений к памяти с профилем обращений APEX-теста в этой точке поверхности (точка T на рис. 1). Это полезное свойство используется на практике при оптимизации программ.

Из рис. 1 видно, что обращения к локальной памяти при худшей и лучшей локализации могут различаться на два порядка, а обращения к распределенной памяти через сеть могут при худшей и лучшей локализации различаться на 4–5 порядков. Это типичная ситуация для всех существующих СКСН, а для кластеров эти соотношения еще хуже. В программе DARPA HPCS ставится задача повысить реальную производительность на DIS-задачах на 3-4 порядка. APEX-поверхности в таком случае должны иметь вид горизонтальной плоскости, а не «горки», как сейчас.

Об этом также свидетельствует требование достижения показателя в 64000 GUPS (Giga Updates Per Second, миллиард коррекций в секунду) на тесте Random Access (см. табл. 1). Тест RandomAccess заключается в выполнении коррекций (чтение-запись) ячеек памяти по псевдослучайным адресам,

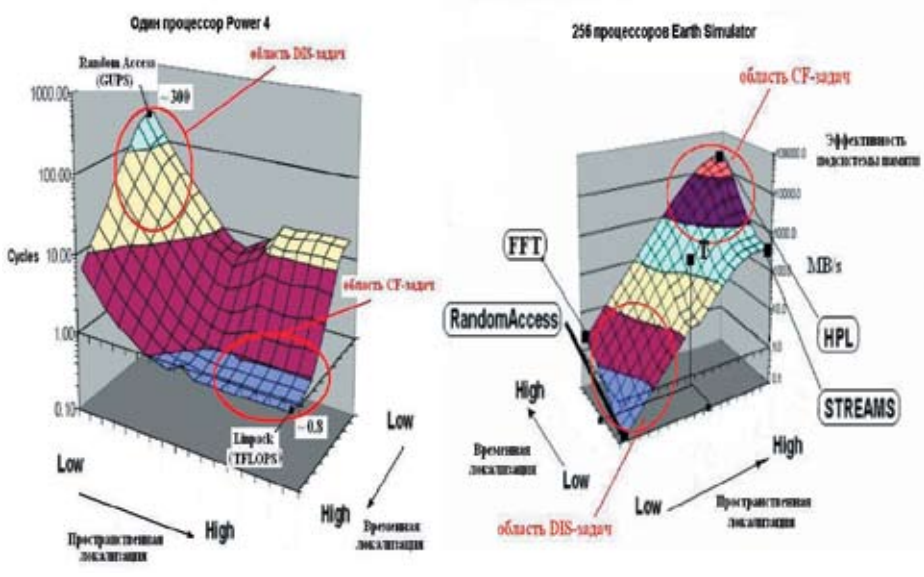


Рис. 1. Интегральная оценка эффективности подсистемы памяти посредством APEX-поверхностей (тест APEX-MAP)

Объективное оценочное тестирование

Новшество DARPA HPCS — хорошо поставленная система оценочного тестирования, которое ведется на всех этапах создания суперкомпьютеров. По сути это уточнение технического задания, которое может контролироваться. Ответственность за проведение тестирования возложена на независимую организацию — Линкольновскую лабораторию Массачусетского технологического института (лаборатория известна, в частности, своим участием в разработке программы СОИ).

Применяется иерархическая система оценочного тестирования. На самом верхнем уровне находятся тесты пакета HPC Challenge, оценивающие границы реальной производительности. В пакете имеется восемь тестов, в том числе четыре теста вычислительного узла и четыре теста системы. Все это — тесты границ APEX-поверхности.

Уровнем ниже находятся тесты получения индикаторов производительности на базовых алгоритмах конкретных прикладных областей. Всего здесь около 40 тестов, разделенных по шести подобластям: дискретная математика; обработка графов; решение систем линейных уравнений, представленных плотно заполненными и разреженными матрицами; моделирование физических процессов; обработка сигналов; ввод/вывод.

Следующий уровень — десять масштабируемых компактных приложений. На этом уровне получают индикаторы реальной производительности и продуктивности разработки программного обеспечения — программы оптимального сопоставления образцов, анализа графов, моделирования физических процессов и технических систем, обработки сигналов с получением знаний.

Наконец, последний уровень — системы, состоящие из девяти приложений трех областей: разведка, моделирование, наблюдение.

охватывающим практически всю физически доступную память. Сегодня максимально достижимый результат составляет 35 GUPS для суперкомпьютера BlueGene/L и 29 GUPS для Cray XT3 (см. табл. 2).

Увеличение производительности систем на тесте RandomAccess на 3–4 порядка позволит поднять нижнюю точку APEX-поверхности для многопроцессорных систем вверх до горизонтального уровня. Для этого используются новые архитектурно-программные принципы построения процессоров, коммуникационной сети и памяти, применяются новые модели организации вычислений в программах.

Известны три стратегических направления повышения эффективности задач DIS-класса, которые используются сегодня в программе DARPA HPCS:

• обеспечение толерантности процессора к задержкам обращений к памяти за счет применения высокопроизводительной подсистемы памяти, способной обслуживать одновременно много запросов, а также процессора со специальной организацией, способного выдать и одновременно выполнять большое количество обращений к памяти без задержки счета по программе, коммуникационной системы, способной пропустить такое большое количество запросов;

• локализация не только данных при процессоре, но и вычисления при данных;

• использование моделей вычислений в виде графов задач с передачей данных между такими задачами через быстрые ресурсы, минуя обращения к памяти.

ПРОДУКТИВНОСТЬ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Рассматривая цели и задачи программы DARPA HPCS, следует упомянуть о продуктивности вычислительных систем, которая должна быть повышена на порядок. Продуктивность — сложное понятие, связанное как с используемыми аппаратными средствами, так и со средствами и технологиями программирования. Предполагается, что создаваемые системы будут программироваться на языках более высокого уровня (класс языков PGAS), чем применяемые сегодня Фортран и Си. Фундаментальным свойством этих языков, которое приведет к более простому и эффективному программированию задач, будет работа с глобально адресуемой памятью, в которой возможно выделение эффективно доступных подобластей.

Программная модель PGAS (Partitioned Global Address Space — «разделенное глобальное адресное пространство») предусматривает поддержку на уровне синтаксических конструкций языка программирования глобального адресного пространства и выделения в нем подобластей, отображаемых на локальную физическую память. Другими словами, в модели PGAS обеспечивается прозрачный доступ к памяти всех вычислительных узлов с учетом неравномерной задержки выполнения команд обращений к памяти. Обязательной составляющей являются средства управления локализацией данных и вычислений за счет распределения данных и удаленного вызова процедур.

ВЫПОЛНЕНИЕ ПРОГРАММЫ DARPA HPCS

Выполнение второй фазы программы DARPA HPCS было начато в 2003 году силами компаний IBM, Cray и Sun (проекты PERCS, Cascade и HERO соответственно). Третья — заключительная — фаза должна завершиться в 2010 году созданием опытных образцов. К ее выполнению 21 ноября 2006 года были допущены IBM и Cray, которые получили по 250 млн. долл. каждая. Еще столько же ими получено летом 2006 года от Министерства энергетики США для построения двух петафлопсных систем. Судя по [6], речь идет о прототипах будущих систем, но с использованием частично коммерчески доступных компонентов и уже готовых вариантов компонентов новых суперкомпьютеров. Cray выполняет проект Baker для Окриджской лаборатории [6], применяя в качестве базовых элементов микропроцессор Opteron, коммуникационный сопроцессор Gemini [7] и маршрутизатор YARC

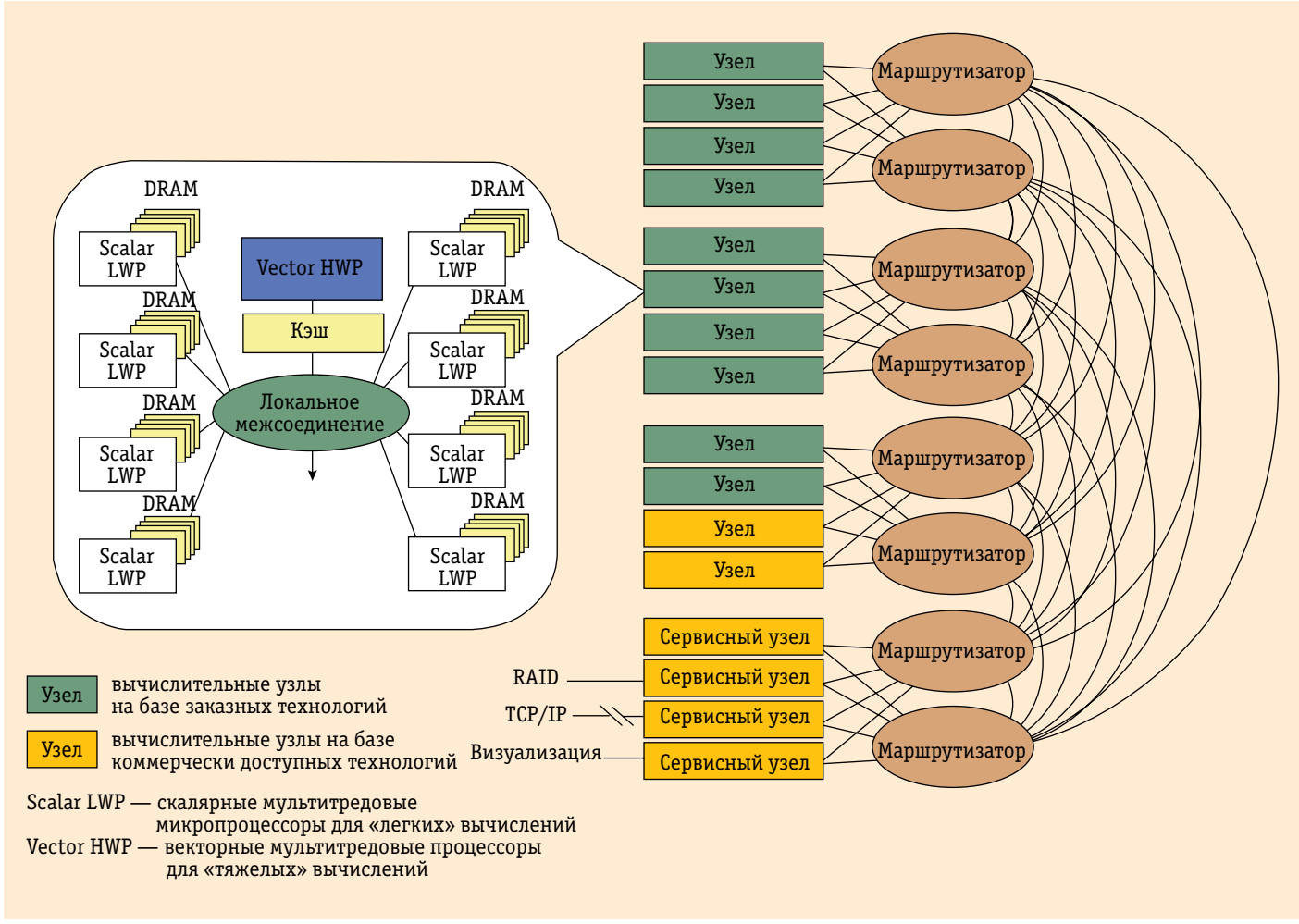


Рис. 2. Концептуальная архитектура системы Cray Cascade

[8], реализующие 14-мерный гиперкуб. IBM ведет проект Roadrunner [12] для Лос-Аламосской лаборатории на микропроцессорах Cell и Opteron и коммуникационной сети Infiniband.

Больше информации в открытых источниках имеется о ходе реализации программы в компании Cray.

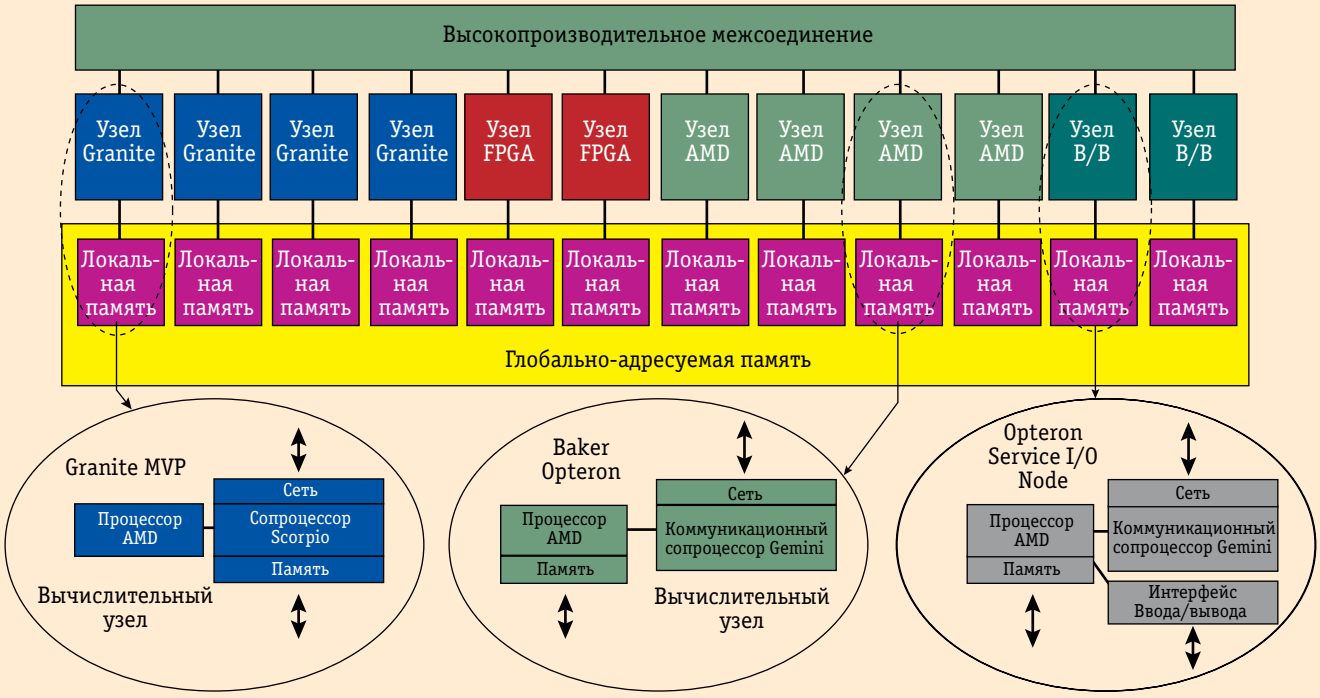
Архитектура Cascade основана на следующих принципах:

- глобально адресуемая память с унифицированной для всех типов узлов архитектурой;
- конфигурируемые сеть, память, процессоры и ввод/вывод;
  - гетерогенная обработка на множестве узлов разного типа и внутри потоково-векторных (MVP) узлов;
  - возможность адаптации при конфигурировании, компиляции, а также в процессе выполнения.

На рис. 2 приведена концептуальная архитектура системы Cascade — гетерогенного стратегического суперкомпьютера, содержащего вычислительные узлы, непосредственно разрабатываемые по проекту (Custom Compute Locale), а также вычислительные узлы на коммерческих процессорах (COTS Compute Locale) и сервисные узлы на коммерческих процессорах (COTS Service Locale).

Необходимая вычислительная мощность достигается благодаря заказным узлам, использующим архитектурные принципы мультитредовости, разделения вычислений, доступа к памяти по разным процессам, принцип размещения обработки вблизи модулей памяти. Vector HWP — это векторный мультитредовый мультипроцессор, способный эффективно выполнять вычисления с подготовленными ему данными в быстродействующей памяти. Предварительную «накачку» данных для этого процессора и простые вычисления (например, адресные) осуществляют скалярные мультитредовые микропроцессоры Scalar LWP, которые находятся вблизи микросхем памяти DRAM и хорошо справляются с задачами, отличающимися плохой пространственно-временной локализацией.

Вычислительная сеть (Router) связывает вычислительные узлы всех типов с модулями распределенной памяти, доступной через единое глобальное адресное пространство. Основное требование к этой сети — высокая пропускная способность на коротких пакетах, что и отражено в требованиях по развиваемой бисекционной пропускной способности сети (см. табл. 1). В этом случае толерантные (за счет мультитредовости) к задержкам обращений к памяти процессоры могут использовать ее высокую пропускную способность и работать на темпе выполнения обращений, а не на их задержках.



Узел Granite — вычислительные узлы Granite на базе суперскалярных микропроцессоров AMD Opteron и векторно-мультитредовых сопроцессоров Scorpio.
Узел FPGA — вычислительные узлы с реконфигурируемыми ускорителями на ПЛИС.
Узел AMD — вычислительные узлы с процессорами AMD Opteron.
Узел B/B — сервисные узлы ввода/вывода.

Рис.3. Структура проекта Cascade на этапах Granite и Baker

Память с высокой пропускной способностью — еще одна проблема, особенно при обращениях с непредсказуемой нерегулярностью (RandomAccess). Именно поэтому в табл. 1 указаны такие высокие требования к ней.

В качестве языка программирования разрабатывался Chapel, который должен еще быть принят программистским сообществом.

Работа по проекту Cascade задумывалась грандиозная, но жизнь, как это бывает, внесла свои коррективы, сделав первые реализации менее амбициозными. Это не всем понравилось, и некоторые «идейные отцы» разработки, прежде всего Бартон Смит и Томас Стерлинг, прекратили участие в проекте и переключились на близкие изначальной идее исследования в других организациях.

Разработчики Cray выбрали прагматичный путь выполнения проекта (см. рис. 3), добавив новые вычислительные средства, основанные на перспективных архитектурных концепциях, к уже хорошо зарекомендовавшим себя микропроцессорам Opteron, имеющим высокопроизводительный интерфейс HyperTransport для подключения внешних устройств. При этом задачи разработки микропроцессорных СБИС мультитредового типа не снимаются с повестки дня, просто они разрабатываются пока в упрощенном виде, что снижает технические риски: наличие Opteron в узле позволяет подстраховаться в случае неудачных аппаратных решений при разработке компонентов с новой архитектурой.

Заключительная фаза разработки системы Cray Cascade в рамках DARPA HPCS состоит из трех этапов: Baker, Granite и Marble (см. рис. 3).

На первом этапе создается коммуникационный сопроцессор Gemini, имеющий интерфейс с разными системами от Cray, включая новый векторный процессор BlackWidow [12], который будет выпущен в 2008 году. Для построения сети используется кристалл YARC сети Клоса, созданный совместно со специалистами Стэнфордского университета. Коммуникационный сопроцессор Gemini оптимизирован под особенности MPI и эффективно выполняет обращения к удаленным узлам. В нем будет реализована трансляция виртуальных адресов глобально адресуемой памяти и выполнение легких тредов для фрагментов задач с плохой пространственно-временной локализацией. Предусмотрены две версии этого коммуникационного сопроцессора. На втором этапе создается мощный векторно-потоковый сопроцессор Scorpio, который должен резко повысить возможности вычислительного узла, усилив его толерантность к задержкам обращений к памяти за счет векторной и потоковой организации.

Надо полагать, что на данный момент Vector HWP воплотился в конкретном сопроцессоре Scorpio, а Scalar LWP — в коммуникационном микропроцессоре Gemini. Отметим также, что

в системе появились и узлы с реконфигурируемой архитектурой на базе программируемых логических матриц (FPGA Compute Nodes, рис. 3).

Что будет реализовано на этапе Marble — пока неизвестно; возможно, произойдет возвращение к первоначальному проекту. По поводу подготавливаемых в Cray технологий программирования можно сказать, что на этапе Baker явно лидируют MPI в сочетании с языками Фортран и Си, средством мультитредового программирования OpenMP и прошедшими практическую проверку языками класса PGAS — UPC и Co-Array Fortran (CAF). Любопытно, что активные попытки внедрить UPC и CAF заметны в уже упоминавшемся семействе Cray XT5.

О работе IBM по программе DARPA HPCS информации мало. Известно лишь, что в создаваемом стратегическом суперкомпьюетере будет применен микропроцессор Power 7 и разрабатывается язык нового поколения X10. Некоторые специалисты утверждают, что создается 128-потоковая версия микропроцессора Power 7. Какая из множества имеющихся разработок будет выбрана (Cyclops64, TRIPS, HPC Cell), пока неясно. Однако в создании системы Roadrunner также заметен прагматичный подход — мощный, но капризный при работе с данными микропроцессор Cell объединяется с Opteron, который может взять на себя подкачку части данных для реализации задач с плохой пространственно-временной локализацией.

ЗАКЛЮЧЕНИЕ

Важный результат, уже достигнутый в рамках программы DARPA HPCS, состоит в том, что удалось сформулировать принципиальные проблемы развития СКСН, убедить чиновников в необходимости решать эти проблемы, восстановить инфраструктуру исследований и разработок в данной области. Выполнение программы приведет к появлению новых технологий создания процессоров, коммуникационных сетей, памяти и языков программирования. Применения их следует ожидать не только при разработке стационарных наземных комплексов сверхвысокой мощности и эффективности, но и в бортовых системах воздушного и космического базирования.

Успешное завершение программы может привести к технологическому отрыву США от остальных стран и дать абсолютное превосходство, которое можно сравнить с единоличным обладанием ядерным оружием в 40-х годах. Не следует питать иллюзий по поводу якобы аналогичных СКСН возможностей вычислительных кластеров на базе коммерчески доступных технологий. Созданные в ходе DARPA HPCS суперкомпьютерные технологии не появятся в обозримом будущем на открытом рынке. Напомним, что инициаторами проекта были структуры, отвечающие за национальную безопасность. Отсутствие собственных аналогичных работ по СКСН у других стран грозит серьезным технологическим отставанием со всеми вытекающими отсюда последствиями.

www.osp.ru

Направления развития отечественных высокопроизводительных систем
http://www.osp.ru/os/2003/05/183021

Между тем, DARPA HPCS закрывает технологическую брешь лишь на ближайшие пять-десять лет и направлена на создание СКСН до 2015 года. Дальнейшее развитие требует более радикальных мер, что и рассматривается в федеральной программе США [11]. Важно отметить, что базовые архитектурные и программные решения DARPA HPCS могут быть применены и при переходе на новые электронные технологии, в частности, на одноквантовую логику сверхпроводимости. Однако даже на основе кремниевых технологий возможно появление к 2020 году экзафлопсных систем (EFLOPS, 10^{18} операций с плавающей точкой в секунду), использующих основные архитектурные решения программы DARPA HPCS [13]. ■

ЛИТЕРАТУРА

1. R.F. Nesbit (chairman). Report of the Defense Science Board Task Force on DoD Supercomputing Needs. 11 October 2000, Office of the Under Secretary of Defense For Acquisition and Technology, Pentagon, Washington.

2. R. Games. Survey and Analysis of the National Security High Performance Computing Architectural Requirements. The MITRE Corporation, 4 June 2001.

3. C.J. Holland. DoD Research and Development Agenda for High Productivity Computing Systems. White Paper, 11 June 2001, Secretary of Defense for Science and Technology.

4. E. Strohmaier, H. Shan, Apex-Map: A Global Data Access Benchmark to Analyze HPC Systems and Parallel Programming Paradigms. Proceeding of the 2005 ZCM/IEEE SC|05 Conference, 2005.

5. T. Sterling. Critical Factors and Directions for Petaflops-scale Supercomputers. California Institute of Technology, NASA Jet Propulson Laboratory, Presentation to IFIP WG10.3 e-Seminar Series. 4 January 2005.

6. S. Scott. Thinking Ahead: Future Architectures from Cray. Presentation. 26 Feb 2007.

7. N. Wichmann. Network changes and its effect on applications. Presentation, 2007.

8. S. Scott, D. Abts, J. Kim, W. Dally. The BlackWidow High-Radix Clos Network. Stanford Univ., 2006.

9. D.A. Reed. Workshop on The Roadmap for Revitalization of High-End Computing. June 16-18, 2003.

10. Public Law 108-423-NON.30, 2004. Department of Energy High-end Computing Veritalization Act of 2004.

11. Federal Plan for High-End Computing. Report of the High-End Computing Revitalization Task Force. May, 2004, National Science and Technology Council Committee on Technology. Executive Office of the President of the United States.

12. K. Koch, P. Henning. Beyond a Single Cell. Los Alamos National Laboratory. Cell Workshop University of Tennessee, October 25, 2006. Presentation.

13. T. Sterling, Architecture Paths to Exaflops Computing. Is Multicore the next Moor's Law? What about Memory? Invited Presentation to the DOE E3SGS Town Hall Meeting. April 18, 2007.

Александр Фролов, Александр Семенов, Антон Корж, Леонид Эйсымонт (frolov/semenov/anton/verger@nicevt.ru) — сотрудники ОАО НИЦЭВТ (Москва).