СУПЕРКОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ ДЛЯ СТРАТЕГИЧЕСКИ ВАЖНЫХ ЗАДАЧ

В.Митрофанов, А.Слуцкин, Л.Эйсымонт

На рубеже нового тысячелетия разработка суперкомпьютеров (особенно кластеров) и их программного обеспечения зашла в тупик. Чрезвычайно снизилась реальная производительность (до уровня от 10% до 0,1% от пиковой), возросла потребляемая мощность и катастрофически упала продуктивность разработки программ. Проблему своевременно распознали и в США [1, 2], и в России [3]. Ее разрешение на основе новых суперкомпьютерных технологий – это начало нового научно-технического направления. Но если в США программе развития данного направления придан статус федерального закона, то в России проблеме перспективных суперкомпьютеров внимания уделяется явно недостаточно. Тем не менее, и в нашей стране уже начаты работы, которые дают шанс не слишком отстать от зарубежных проектов.

СТРАТЕГИЧЕСКИЕ ЗАДАЧИ И ТРЕБОВАНИЯ К ВЫЧИСЛИТЕЛЬНЫМ СИСТЕМАМ

Достижения микроэлектроники последних 10–15 лет позволили создавать высокопроизводительные многопроцессорные вычислительные системы (вычислительные кластеры) на базе коммерчески доступных компонентов. Кластеры позволили, как никогда ранее, расширить круг разработчиков высокопроизводительных вычислительных средств, а также их пользователей и классов решаемых задач. Такой успех и коммерческая привлекательность создали у многих иллюзию "всемогущества" кластерных технологий. Время показало, что эта точка зрения ошибочна и вредна. В ряде важнейших для государства областей (стратегические задачи) требуются либо полностью специальные технологии разработки суперкомпьютеров, либо их сочетание с доступными "кластерными" технологиями.

Стратегические задачи – это задачи обеспечения национальной безопасности и решения наиболее важных научных и научно-технических проблем. Первый класс таких задач – разработка ядерного оружия и контроль его боеготовности, радиоэлектронная разведка (криптоанализ, интеллектуальная обработка, контроль линий связи), национальные системы противоракетной обороны (обработка сигналов и изображений, автоматическое распознавание целей, принятие решений и управление), разработка систем направленной энергии, имитационное моделирование, отработка и оценка программного обеспечения военных систем. Второй класс задач – научные исследования, например управляемый термоядерный синтез, новые источники энергии, нано- и биотехнологии. Это и важнейшие научно-технические задачи – прогноз состояния окружающей среды и изменения климата; создание новых технических систем, прежде всего оружия (высокоточное оружие и роботизированные системы, оружие, основанное на новых физических принципах); создание информационных систем управления крупными военными операциями с координацией действий разных родов войск; создание информационных систем разведки, навигации и связи.

Суперкомпьютеры, предназначенные для решения стратегических задач, называют суперкомпьютерами стратегического назначения (СКСН). Но в современных условиях национальная безопасность зависит еще от одного класса задач, решаемых в перспективных интеллектуальных бортовых системах управления реального времени. Они требуют особых, встроенных суперкомпьютеров (ВСК).

Проблемы разработки СКСН и ВСК во многом едины, поскольку к ним предъявляется общее главное требование – высокая эффективность, а основной их показатель – высокий уровень реальной производительности. Причем для ВСК удовлетворить этим требованиям труднее, поскольку более сложны как класс решаемых задач (с точки зрения достижения высокой реальной производительности), так и требования по условиям эксплуатации.

Область стратегических задач и СКСН для их решения в США называют high-end computing (HEC), а СКСН – суперкомпьютеры высшего диапазона производительности (high-end computers). Это выделяет их из более широкого семейства высокопроизводительных систем и суперкомпьютеров (HPC, high performance computers).

В США государство придает СКСН весьма важное значение. Так, в 2007 году из 3,074 млрд. долл. бюджета глав-

ной программы NITRD по информационным технологиям США около половины (1,324 млрд. долл.) было запрошено на приобретение СКСН, разработку приложений и обеспечение расчетов, выполнение исследований и разработок в области СКСН. То есть чрезвычайно высоко внимании со стороны правительства США именно к классу самых мощных из HPC-компьютеров. Это удивительно, поскольку мировой рынок HPC-компьютеров составляет около 9 млрд. долл. и темп его роста в ближайшие пять лет оценивается приблизительно в 9% в год. В то же время рынок СКСН – всего лишь 2–3 млрд. долл., но именно на нем сосредоточено внимание государства. Таким образом, СКСН и технологии их создания не расцениваются в США как товар для рынка, в развитие которого надо вкладывать бюджетные деньги. Ценность СКСН и связанных с ними технологий для государства в ином – в возможности обеспечения национальной безопасности и решении главных научно-технических и социально-экономических проблем страны.

Из множества СКСН в США выделяется особый подкласс – ультракомпьютеры, или системы-лидеры (LeaderShip Systems). Существуя в единичных экземплярах, они должны превосходить по реальной производительности не менее чем на два порядка любой суперкомпьютер мира, который можно собрать из коммерчески доступных компонентов или просто купить [4, 5]. Стратегические задачи для таких систем (порядка десяти) определяются каждый год специальной комиссией и утверждаются президентом США.

Практика показала, что высокие требования к эффективности применения СКСН и ВСК недостижимы при использовании коммерчески доступных компонентов и обычных аппаратно-программных решений. Это связано с тем, что при разработке СКСН и ВСК необходимо решение "проблем преодоления пяти стен" – стены памяти, стены ограниченного параллелизма выполнения машинных команд, стены потребляемой энергии, стены тепловыделения и стены продуктивности программирования. Степень преодоления по крайней мере первых двух из перечисленных проблем отражается в достигаемых значениях шести общепринятых базовых характеристик СКСН (табл.1).

Агентство перспективных исследований МО США (DARPA) реализует программу HPCS по созданию перспективных СКСН к 2010 году [1, 2]. В ней сформулирован ряд целевых показателей, которые коррелируются с перечисленными базовыми характеристиками:

- более 2 PFlops (10^{15} Flops) реальной производительности на тесте Linpack;
- около 6,5 Пбайт/с на тесте пропускной способности памяти при регулярных обращениях (тест STREAM);
- около 3,2 Пбайт/с на тесте бисекционной пропускной способности системной коммуникационной сети (тест BISECT);
- 64000 млрд. произвольных коррекций (GUPS – Giga Random Access) в секунду на тесте нерегулярного доступа к памяти (тест RandomAccess);
- высокий полиморфизм – параллелизм типа ILP (командный), TLP (тредовый) и DLP (по данным – векторный и потоковый);
- высокая реконфигурируемость и адаптируемость к задачам;
- распределенная общая память (глобально адресуемая память) объемом до нескольких петабайт;
- увеличение продуктивности программирования по отношению к уровню 2005 года в 10 раз.

Создаваемые в рамках программы HPCS СКСН должны существенно превосходить лучшие современные СКСН, даже созданные по специальным суперкомпьютерным технологиям (табл.2), которые, в свою очередь, по показателям бисекционной пропускной способности сети и эффективности нерегулярного доступа к памяти эффективнее кластеров на один-два порядка. В эту таблицу не включен самый мощный современный суперкомпьютер IBM Roadrunner. Он действительно обладает наибольшей производительностью на тесте Linpack (более одного петафолопса), несомненно

Таблица 1. Чувствительность задач разных областей приложений к базовым характеристикам суперкомпьютеров

Базовая характеристика	Критичная область приложений
Производительность вычислений с плавающей точкой	Астрофизика, обработка радарной информации, моделирование климата, физика плазмы
Объем оперативной памяти	Разведка, материаловедение, геном, автомобильные шумы, вибрация, прочность
Пропускная способность оперативной памяти	Разведка, моделирование климата, материаловедение, астрофизика, моделирование биологических систем
Задержка выполнения операции с памятью	Разведка, ядерное моделирование, моделирование климата, астрофизика, моделирование биологических систем
Пропускная способность коммуникационной сети	Разведка, моделирование климата, материаловедение, астрофизика, моделирование биологических систем
Задержка выполнения операции с коммуникационной сетью	Разведка, ядерное моделирование, моделирование климата, астрофизика, моделирование биологических систем

ПРЕДСТАВЛЯЕМ АВТОРОВ СТАТЬИ

В.В.Митрофанов, к.т.н., генеральный директор ОАО "НИЦЭВТ".

А.И.Слуцкин, к.т.н., первый заместитель генерального директора ОАО "НИЦЭВТ", начальник управления, главный конструктор суперкомпьютера стратегического назначения (СКСН) "Ангара".

Л.К. Эйсымонт, к.ф.-м.н., начальник отдела ОАО "НИЦЭВТ", заместитель главного конструктора СКСН "Ангара".

verger@nicevt.ru, (495) 319-19-36

Таблица 2. Сравнение характеристик создаваемых к 2010 году СКСН программы DARPA HPCS с характеристиками лучших современных СКСН

Характеристика	Существующие в 2007–2008 году		Перспективные
	IBM Blue Gene	Cray XT3	
Производительность на тесте Linpack, TFlops	260	101	>2000
Пропускная способность памяти при регулярных обращениях, Тбайт/с	128	196	>6500
Бисекционная пропускная способность коммутационной сети, Тбайт/с	0,36	11,7	>3200
Скорость нерегулярного доступа к памяти, GUPS	35	29	>64000

высока и пропускная способность его системы памяти при регулярных обращениях (этот показатель пока не опубликован). Однако, зная структуру этого суперкомпьютера, можно утверждать, что у него плохие результаты на тестах BISECT и RandomAccess, которые наиболее принципиальны для перспективных СКСН.

В последние год-два появилось дополнительное требование к СКСН и ВСК – возможность эффективной обработки огромных внешних потоков данных от всевозможных источников, прежде всего – при работе с сетями сбора и обработки данных на фоне интенсивного нерегулярного обмена с памятью (класс приложений типа вычислений с интенсивной работой с данными – Data Intensive Computing, DIC [6]). Оно еще не отражено в четко формулируемых характеристиках, аналогичных приведенным в табл.2, но это – вопрос ближайшего времени.

Внедрение новых СКСН будет происходить в 2010–2015 годы. Они настолько необычны, что готовить к ним пользователей начали уже сейчас. По эффективности эти СКСН должны превосходить современные кластерные системы на 4–5 порядков [1, 6]. Предполагается также, что создаваемая перспективная архитектура СКСН и языки программирования класса PGAS нового поколения (Chapel, X10, Fortress) на порядок повысят и продуктивность программирования СКСН [1]. Это связано с тем, что резко возросший в перспективных СКСН объем эффективно доступной памяти позволит перейти к моделям вычислений с общей памятью и односторонним взаимодействиям параллельных процессов вместо неэффективных двусторонних взаимодействий, типичных для используемых на современных суперкомпьютерах параллельных MPI-программ.

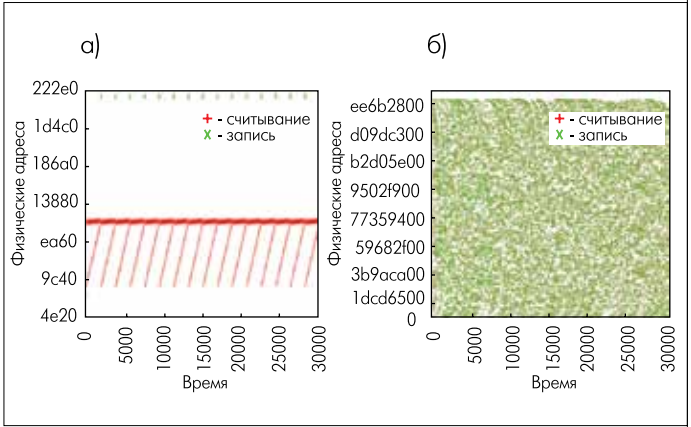

Рис.1. Профиль обращений: с регулярным доступом на задаче MMX (а) и предельно нерегулярные обращения на тесте RandomAccess (б)

Прокомментируем подробнее целевые характеристики работы с памятью, поскольку для СКСН это основное направление оптимизации, влияющее как на развиваемую на разных типах задач реальную производительность, так и на продуктивность программирования. Ранее негативное влияние на реальную производительность высоких задержек обращений к памяти (несколько сотен тактов процессора при обращении к локальной памяти вычислительного узла и несколько тысяч тактов при обращении к памяти удаленного узла) обычно сглаживалось быстрой кэш-памятью и блоками спекулятивной преднакачки данных в нее. Однако для современных задач (особенно стратегических) такие приемы явно недостаточны. Различные типы задач могут быть дружественными к использованию кэш-памяти и блоков преднакачки (CF-задачи, cache friendly) и недружественными (DIS-задачи, data intensive systems). При выполнении CF-задач реальная производительность составляет 60–90% от пиковой производительности системы, а DIS-задач – лишь 5–10% и менее. Для задач обоих типов важно понятие пространственно-временной локализации обращений к памяти.

Пространственная локализация показывает, как часто в заданных интервалах последовательности обращений к памяти встречаются обращения по близким и/или предсказуемым адресам. Для задач с хорошей пространственной локализацией эффективно используются смежные данные в кэш-строке, а также оказывается точным угадывание данных, подкачиваемых в кэш-память блоками спекулятивной преднакачки.

Временная локализация показывает, как часто в заданных интервалах обращений происходят обращения по одним и тем же адресам. Этот показатель также связан с эффективностью использования кэш-памяти. Задачи CF-класса обладают высокой пространственно-временной локализацией. Напротив, у задач DIS-класса пространственно-временная локализация (или хотя бы один из этих параметров) низка.

Пример профиля работы с памятью для задачи CF-класса показан на рис.1 – это задача MMX умножения плотно заполненных матриц. Для нее характерна высокая вероятность повторного использования данных (высокая временная локализация, горизонтальные линии на профиле обращений), а также высокая пространственная локализация и предсказуемость обращений к данным (наклонные участки регулярного вида на профиле).

Пример профиля обращений к памяти экстремальной задачи DIS-класса также дан на рис.1. Это профиль для специ-

ального теста RandomAccess (его другое используемое название – GUPS [6]). Хоть это и специальный тест, но полученные на нем характеристики настолько важны, что являются главным улучшаемым параметром разрабатываемых в настоящее время перспективных СКСН (см. табл.2).

Чрезвычайно высокая скорость произвольного доступа к оперативной памяти – 64 тыс. GUPS (Giga Updates Per Second) (см. табл.2) – главная особенность перспективных СКСН. Этот параметр измеряется посредством теста RandomAccess [6] нерегулярного доступа к памяти. Тест RandomAccess заключается в выполнении коррекций (чтение-коррекция-запись) ячеек памяти по псевдослучайным адресам практически во всей физически доступной памяти (т.е. адреса корректируемых ячеек имеют самую плохую пространственно-временную локализацию). 1 GUPS соответствует одному миллиарду коррекций в секунду.

Задачи со степенью нерегулярности, аналогичной тесту RandomAccess, – не такая уж абстракция, как утверждают некоторые отечественные специалисты. Для примера на рис.2 показаны профили задач вычисления одномерного быстрого преобразования Фурье (FFT) и задачи поиска вширь на графе (BFS). Первая задача – основной инструмент из области обработки сигналов, а вторая – важнейшее средство для анализа разведывательных данных, построения систем борьбы с терроризмом, построения сложных систем управления, решения важнейших задач биологии, химии, медицины и научно-технических расчетов.

Из анализа профилей обращений к памяти DIS-задач видно, что использование кэш-памяти и блоков спекулятивной преднакачки здесь практически бесполезно. Одно из направлений решения проблемы – обеспечение толерантности микропроцессора к задержкам обращений к памяти за счет аппаратно поддерживаемой мультитредовой архитектуры, реализующей много одновременно выполняемых обращений к памяти без задержки счета.

Насколько в действительности серьезна проблема "стены памяти", т.е. как влияет изменение пространственно-временной локализации на эффективность работы с ней, наиболее наглядно характеризует тест APEX-MAP [2, 3], предложенный недавно специалистами Ливерморской лаборатории Министерства энергетики США. Тест APEX-MAP обращается на считывание к элементам массива длины M, сопоставимого по объему со всей физической памятью тестируемой системы. Тест периодически производит регулярный доступ с единичным шагом в пределах области длины L массива M. Параметр L определяет пространственную локализацию. Области длины L после прохода по ним меняются случайным образом (это уже имитация нерегулярного доступа), начальный адрес очередной области определяется как $Adr = Base + L\cdot[rand^{1/A}\cdot(M/L - 1)]$, где Base – начальный адрес, rand – случайная равномерно распределенная величина от 0 до 1; A – параметр временной локализации. При A = 1 временная локализация наихудшая, доступ к памяти абсолютно случаен. Когда $A \to 0$, временная локализация увеличивается до предельно возможной.

Для различных значений L и A измеряются различные параметры доступа к памяти и в результате строятся трехмерные графики измеренного параметра эффективности памяти от пространственной и временной локализации обращений к ней (APEX-поверхности) (рис.3). Такую поверхность можно построить для одного или множества узлов исследуемой вычислительной системы. Для одного узла, как правило, изме-

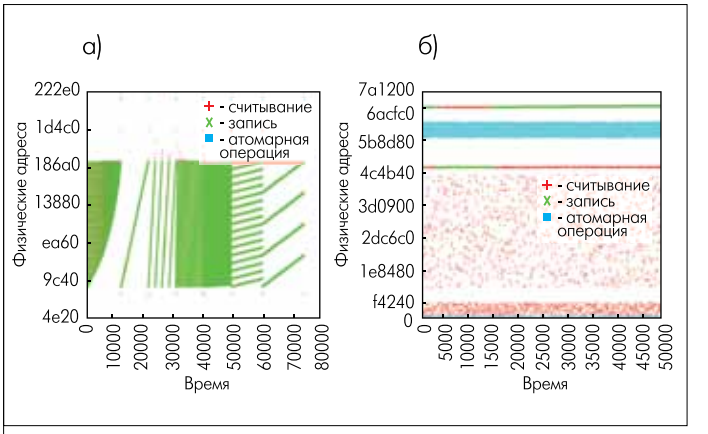

Рис.2. Профили нерегулярного обращения к памяти реальных задач: а – FFT, б – BFS

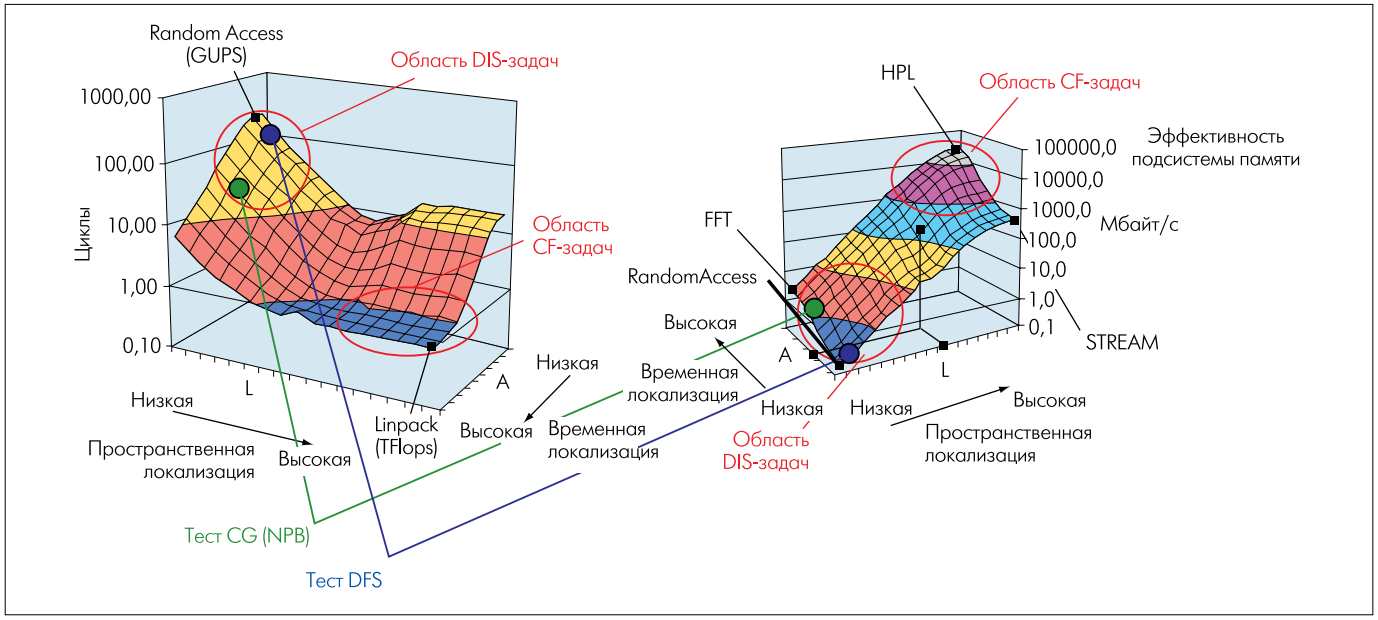

Рис.3. Интегральная оценка эффективности подсистемы памяти посредством APEX-поверхностей (тест APEX-MAP). Показаны точки, соответствующие динамике работы с памятью тестов CG, DFS (поиск в глубину на графе), Linpack (HPL), быстрое преобразование Фурье (FFT), тест Мак-Калпина эффективности пересылок в памяти с регулярным доступом к ней (STREAM), тест Random Access (GUPS)

ряется среднее число тактов процессора, за которое происходит одно обращение к памяти на считывание. Для множества узлов обычно показывают приведенную к одному вычислительному узлу пропускную способность памяти.

Все известные тесты и прикладные программы могут быть поставлены в соответствие определенным точкам APEX-поверхности. Например, до недавнего времени основным тестом вычислительных систем считался тест Linpack. Ему соответствует точка APEX-поверхности с наилучшей пространственно-временной локализацией. В современном наборе оценочных тестов HPC Challenge, который вытесняет применение только Linpack, предусмотрено еще несколько тестов, каждому из которых на APEX-поверхностях соответствуют различные предельные точки.

На рис.3 видно, что скорости обращения к локальной памяти при наихудшей и наилучшей пространственно-временной локализации могут отличаться на два порядка, а время доступа к распределенной памяти через сеть – на 4–5 порядков даже при наличии всего лишь 256 процессоров в системе. Данная ситуация типична для всех существующих СКСН (включая IBM Roadrunner), причем для кластеров она еще хуже. В этом и проблема.

Разработчики программы DARPA HPCS полагали, что перспективные СКСН должны эффективно решать задачи с любой пространственно-временной локализацией обращений к памяти. По этой причине APEX-поверхности для таких систем должны иметь вид горизонтальной плоскости, а не "горки", как сейчас. С этим и связано выбранное в качестве целевого параметра значение на тесте Random Access в 64 тыс. GUPS.

Первые модели (опытные образцы) суперкомпьютеров программы DARPA HPCS должны появиться в конце 2009 года (например, Cray Baker). Переходные модели появились уже в 2008 году (суперкомпьютеры Cray XT5, Cray XT5h и Cray XMT). В этих системах значительно улучшены характеристики работы с памятью, в них предпринята попытка активного внедрения для широкого круга пользователей высокопродуктивных моделей вычислений с общей памятью – интерфейс OpenMP, языки UPC и CAF.

ВОЗМОЖНОСТИ КОММЕРЧЕСКИХ ПРОЦЕССОРОВ И КЛАСТЕРНЫЕ ТЕХНОЛОГИИ

Основной элемент кластерных технологий – коммерчески доступный микропроцессор. Современные коммерческие микропроцессоры резко теряют эффективность при ухудшении пространственно-временной локализации работы с памятью. В частности, это показывают результаты тестов APEX-MAP новых 4-ядерных микропроцессоров Clovertown корпорации Intel и Barcelona компании AMD (рис.4). Видно, что при наихудшей локализации (L = 1, A = 1) задержки обращений к памяти в однотредовом варианте теста, загружающем одно ядро, составляют для Clovertown 290 тактов процессора, а для Barcelona – 350 тактов. В точке с наилучшей пространственно-временной локализацией (L = 65536, A = 0,001) задержка обращения к памяти составляет 1,1 такта для Clovertown и 1,2 – для Barcelona. Достижение таких показателей обеспечивается за счет эффективной работы кэш-памяти. В точке с хорошей пространственной, но плохой временной локализацией (L = 65536, A = 1) задержки для Clovertown и Barcelona составляют 5,7 и 14,3 такта, соответственно.

Если с общей памятью вычислительного узла работают все ядра микропроцессора и программа выполняется в многопотоковом (мультитредовом) режиме, то появляется неко-

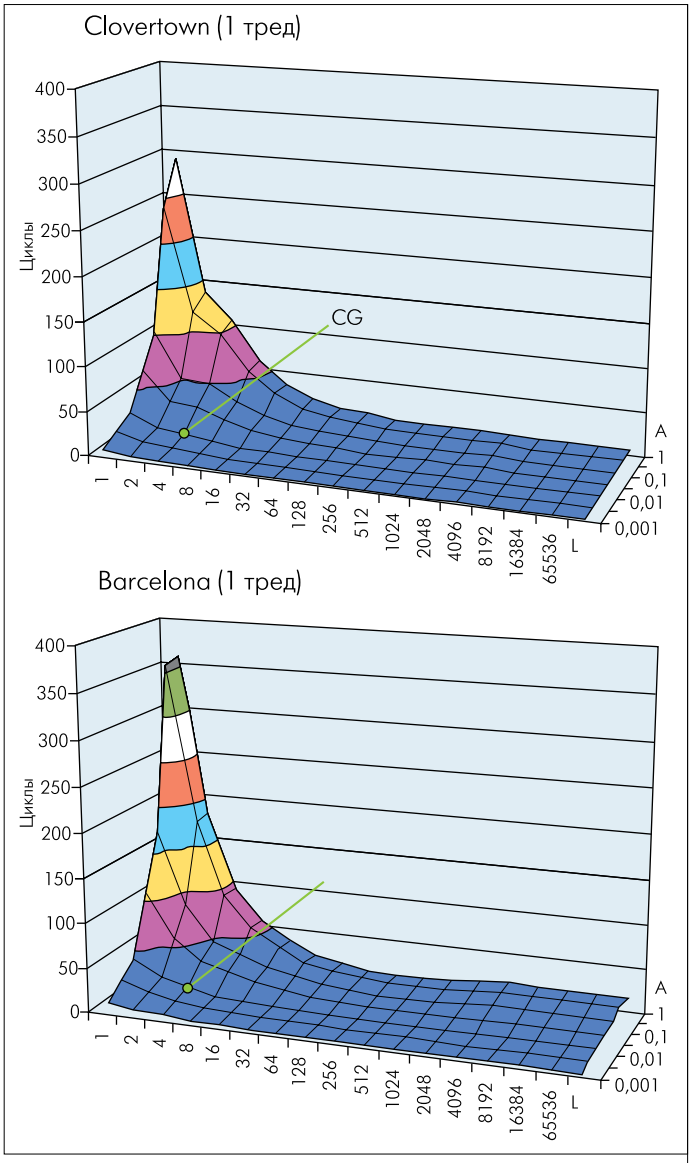

Рис.4. APEX-поверхности для одного процессорного ядра процессоров Clovertown и Barcelona

торая толерантность всего многоядерного процессора к задержкам, поскольку возрастает общее число обращений к памяти и удается работать со скоростью, определяемой не задержками, а темпом выдачи обращений к памяти. Количественная оценка этого эффекта была определена экспериментально на мультитредовом варианте теста APEX-MAP, в котором используются POSIX-треды*. Исследовалась плата с двумя 4-ядерными микропроцессорами Clovertown и плата с четырьмя 4-ядерными микропроцессорами Barcelona (рис.5). На 16-тредовом тесте в наихудшей точке удалось уменьшить видимую тестом задержку выполнения операции с памятью обоими процессорами примерно до 70 тактов. На 64-тредовом тесте данный показатель снизился до 47 тактов на плате с Barcelona (см. рис.4) и остался неизменным для Clovertown. Это говорит о резерве толерантности Barcelona, когда тредов становится больше, чем ядер процессора.

Для 64-тредового варианта APEX-поверхности (рис.6) в точке с лучшей пространственно-временной локализацией (L = 65536, A = 0,001) видимая задержка обращения к памяти составляет 0,2 для Clovertown и 0,1 для Barcelona. Значение видимой задержки памяти в точке с хорошей пространственной, но наихудшей временной локализацией (L = 65536, A = 1) для Clovertown равно 4,3; а для Barcelona – 1,8. Для сравнения – микропроцессор Cell показал в точке с наихудшей пространственно-временной локализацией задержку обращений к памяти в 1000 тактов. Это говорит об ограниченных возможностях этого уникального изделия, необходима особая осмотрительность при оценке его возможностей.

* POSIX-треды – легкие параллельные процессы (потоки) выполняемой программы в соответствии со стандартами POSIX (Portable Operating System Interface for UNIX – переносимый интерфейс операционной системы для UNIX), описывающими интерфейс между операционной системой и прикладными программами.

HT – блок интерфейса HyperTransport;
MC – двухканальный контроллер памяти (2×64 разряда)

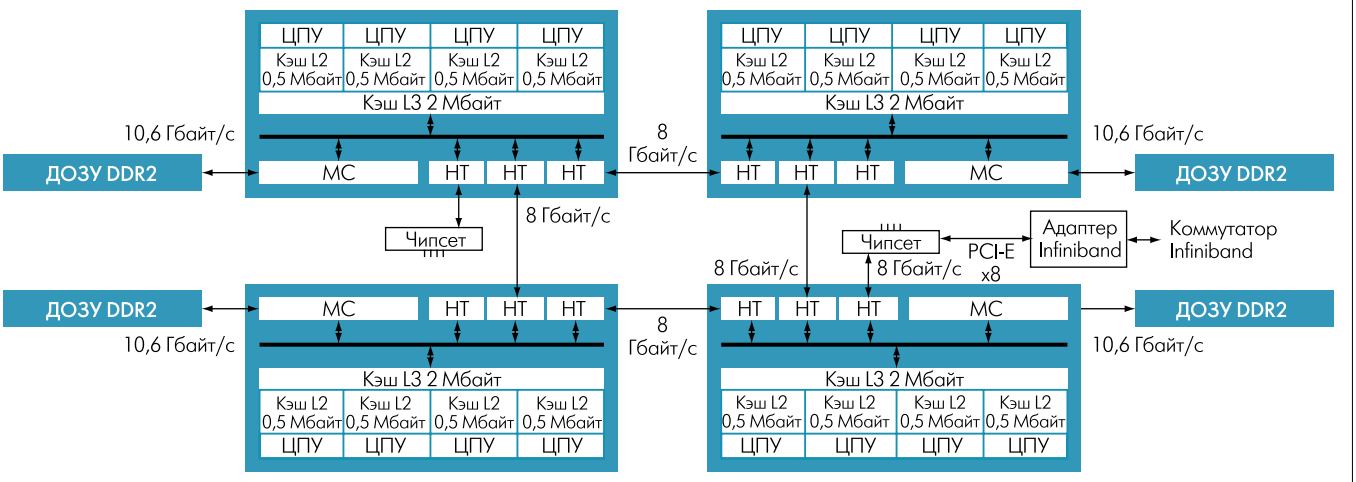

Рис.5. Структура платы Supermicro AS-1041M-T2 с четырьмя 4-ядерными микропроцессорами Barcelona (ЦПУ) и адаптером сети Infiniband

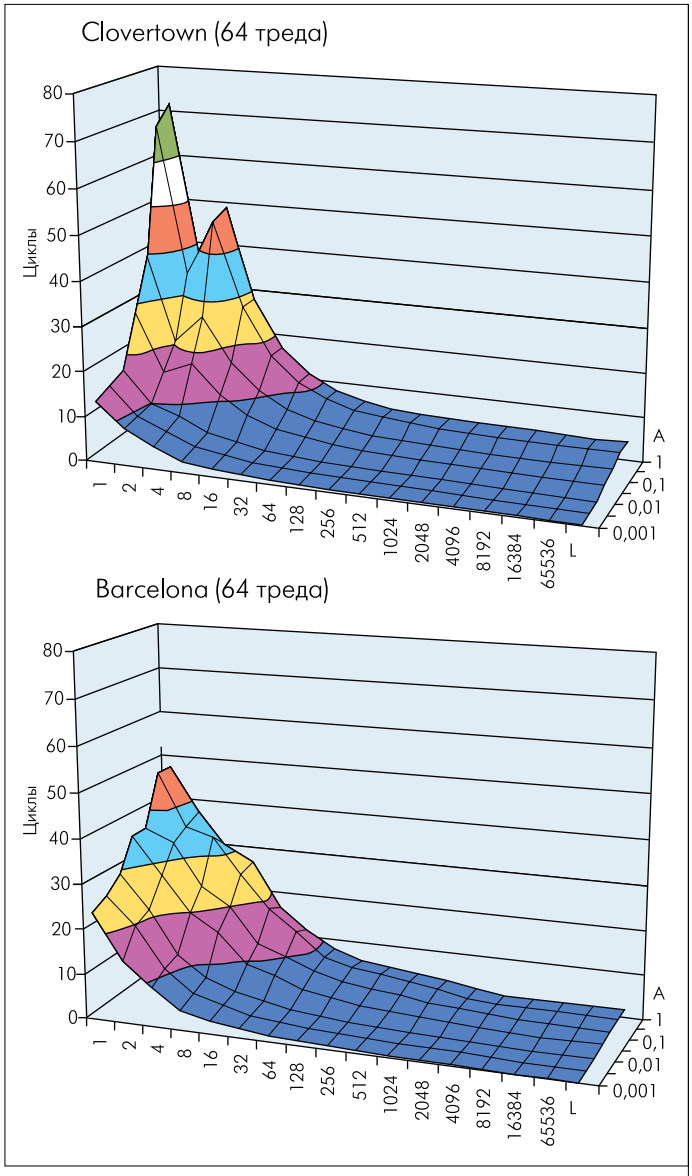

Рис.6. APEX-поверхности при полной загрузке ядер 64 тредами

Чем выше эффективность работы с памятью, тем лучше показатели выполнения многих прикладных программ, особенно – DIS-класса. Для примера рассмотрим MPI-версию теста CG, входящего в пакет NPB 3.1 аэродинамических оценочных программ YFCF. На APEX-поверхности этому тесту соответствует точка (L=3, A=0,01) (см. рис.3). При выполнении этого теста процессоры Clovertown и Barcelona обращаются к памяти за 15–20 тактов (см. рис.4). Как это сказывается на реальной производительности?

Зависимости приведенной к одному ядру производительности от числа используемых ядер при выполнении программы для различных вычислительных кластеров 2002–2007 годов показывают (рис.7), что эффективность использования аппаратных средств на задаче CG составляет лишь несколько процентов от пиковой производительности для одного ядра, и порядка 1% и менее – для 64 ядер. Поскольку с ростом числа ядер реальная производительность падает, то ускорение при распараллеливании слабое или вообще отсутствует.

Если загружать работой лишь часть ядер процессора, снижая нагрузку на процессор, то эффективность использования загруженного ядра значительно увеличивается, например – для кластера на основе процессоров Clovertown с сетью Infiniband (рис.8). Это говорит о том, что ядрам явно не хватает эффективности работы с памятью для решения данных задач.

Тем не менее, были замечены и определенные положительные явления. Из рис.8 видно, что при увеличении загрузки ядер платы с Barcelona 8350 деградация долго не происходит, чего нельзя сказать о плате с Clovertown 5345. Это явный успех разработчиков AMD, и таким свойством непременно следует воспользоваться при разработке кластеров. Но все равно реальная производительность одного ядра остается низкой, хотя уже и стабильной при увеличении числа задействованных процессорных ядер. В целом же даже для задачи из нижней части "пика" APEX-поверхности (см. рис.4) реальные производительности малы и имеют тенденцию к деградации.

Кроме "стены памяти" для кластерных технологий возникают и серьезные проблемы "стены потребляемой мощности". Так, в проекте национальной лаборатории им. Лоуренса в Беркли по созданию комплексной модели атмосферы Земли, что необходимо для качественного прогноза погоды и предсказания климата, требуется СКСН с пиковой производительностью в 200 PFlops. Причем на этой задаче должна развиваться реальная производительность хотя бы 10 PFlops. По оценкам, строительство такого СКСН на коммерческих компонентах обойдется в 1 млрд долл., а мощность его энергетической установки составит 200 МВт, что в 1,5 раза превышает мощность ядерных реакторов современного ударного авианосца! Применение специальных решений позволит построить аналогичный СКСН за 75 млн. долл. с потреблением не более 4 МВт.

Потребляемая энергия – одна из сторон эксплуатации. Еще важно, насколько кластеры просты для разработки программ, т.е. как высока продуктивность программирования. К сожалению, полученный даже в России опыт говорит о низкой продуктивности, что определяется высокой сложностью разработки MPI-программ. Необходимы новые методы высокопродуктивного программирования даже в этом классе массовых и достаточно простых суперкомпьютеров.

Все это свидетельствует о том, что старые кластерные технологии неэффективны, а для некоторых задач вообще не применимы. Появление многоядерных процессоров (это было вынужденной мерой, поскольку примерно с 2003 года прекратилось влияния закона Мура на производительность микропроцессоров) пока лишь ухудшило ситуацию даже для нестратегических задач. Таким образом, эру механической сборки и бесхитростного применения кластерных технологий можно считать завершенной, во всяком случае для крупных систем. Специалисты сейчас озадачены проблемами как эффективного использования многоядерных

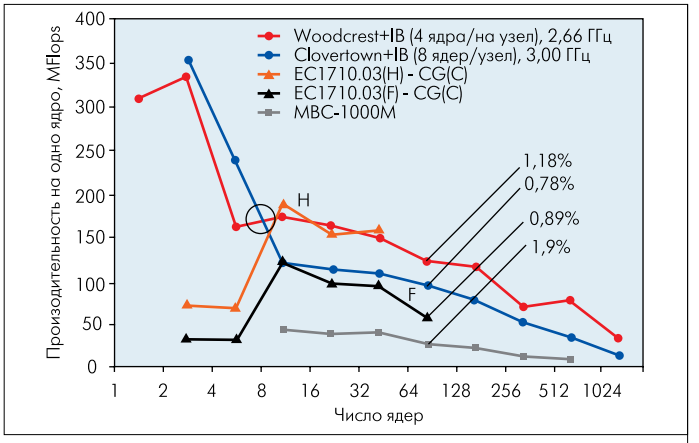

Рис.7. Эффективность распараллеливания и использования аппаратных средств на задаче CG (класс сложности C). Данные получены для случая, когда на каждое ядро загружается MPI-процесс

микропроцессоров, так и оптимизации их микроархитектуры. Это означает, что кластерные технологии сами вынуждены модифицироваться, вбирая в себя элементы новых решений, в полной мере реализуемых в проектах перспективных СКСН. Более того, такая "мутация" кластерных технологий делает их частичное применение в некоторых СКСН не таким уж бесполезным занятием, что показывает опыт разработки первых образцов перспективных СКСН.

СОВРЕМЕННЫЕ СУПЕРКОМПЬЮТЕРЫ И ПСЕВДОКОММЕРЧЕСКИЕ СУПЕРКЛАСТЕРЫ

Программа DARPA HPCS предусматривает поэтапную реализацию (сейчас реализуется фаза III, включающая три этапа). Поэтому она подразумевает применение не только специальных, но и новых кластерных суперкомпьютерных технологий. Особенно важны новые кластерные технологии, включающие такие архитектурные свойства, как многоядерность, мультитредовость, прямые каналы для подключения процессоров и ускорительных плат (Hyper Transport фирмы AMD и Quick Path компании Intel), средства поддержки мелкозернистых и среднезернистых потоковых моделей вычислений (реконфигури-

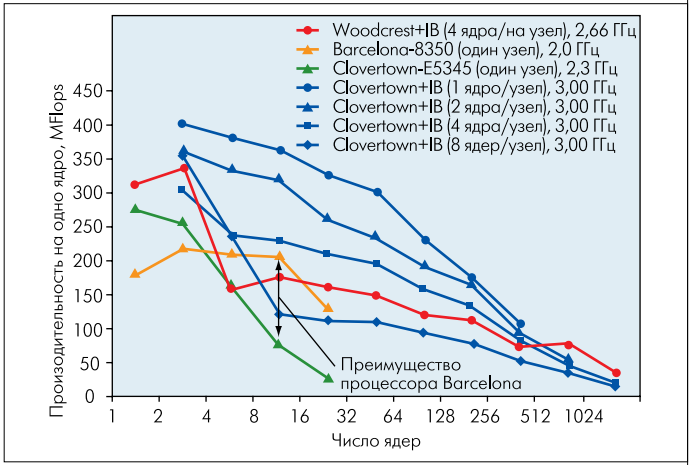

Рис.8. Зависимость реальной производительности одного ядра от числа применяемых ядер при вычислении на тесте CG (C)

руемые решающие поля вычислительных элементов, встраиваемые в микропроцессор или реализуемые на ПЛИС, а также графические сопроцессоры).

Например, для повышения эффективности работы с глобально адресуемой памятью в СКСН Baker компании Cray (создается в рамках программы DARPA HPCS, первый этап фазы III) применяется неоднородная мультитредовая организация вычислительного процессора. Это обеспечивает толерантность СКСН к задержкам обращений к памяти и дополнительно реализует принцип разделения программ на "тяжелые" вычислительные треды и "легкие" треды работы с данными. В результате повышается как параллелизм, так и асинхронность. Для реализации такого подхода в вычислительном узле СКСН используется стандартный многоядерный микропроцессор фирмы AMD (например, 4-ядерный Barcelona). К этому микропроцессору через интерфейс Hyper Transport подключен специальный ускорительный блок (мультитредовый сопроцессор Gemini), связанный со специальной коммуникационной сетью с топологией трехмерного тора или сети Клоса [2, 7]. Функции сопроцессора – трансляция виртуальных адресов глобально адресуемой памяти и реализация легких тредовых вычислений работы с данными. Тяжелые тредовые вычисления при этом выполняют ядра микропроцессора AMD.

Дальнейшее повышение толерантности к задержкам возможно за счет увеличения числа вычислительных тредов и введения векторных операций, что уже реализуется в проекте Granite компании Cray (второй этап фазы III DARPA HPCS). Для этого вводится векторно-тредовый сопроцессор Scorpio, который также подключается к микропроцессору AMD. Вероятнее всего, на третьем этапе фазы III такие промежуточные решения будут заменены специальным суперкомпьютерным многоядерно-мультитредовым микропроцессором. Он значительно эффективнее, хотя дороже и его создание сопряжено с большим риском. Но у компании Cray уже есть пример похожего решения – векторный мультипроцессорный СКСН с глобально адресуемой памятью BlackWidow [7, 8].

Применение смешанных – новых кластерных и суперкомпьютерных – технологий сейчас популярно. Появился даже удачный термин – псевдокоммерческий суперкластер, промежуточное изделие между обычным кластером и настоящей СКСН. В области суперкластеров можно достаточно быстро создавать системы с высокими характеристиками (хотя и не совсем отвечающие требованиям стратегических задач, но близкие к ним) при относительно малых затратах и с меньшими рисками в сравнении с полномасштабной разработкой СКСН. По этому пути, помимо фирмы Cray, пошла и компания IBM (СКСН Roadrunner для Лос-Аламосской лаборатории), а также НАСА и корпорация SGI (проект Pleiades, 1 PFlops к 2009 году, 10 PFlops к 2012 году, наиболее вероятно, что это развитие системы Columbia). Аналогичный подход применяется во Франции в проекте создания к 2009 году системы NovaScale с пиковой производительностью 300 TFlops (с переходом к PFlops), а также в Японии в проекте системы KEISOKU с пиковой производительностью 10 PFlops (ожидается к 2010 году). По имеющейся информации, такой подход использован и в трех проектах создания петафлопсных систем в КНР. Но там работают и над собственными микропроцессорами, как со стандартной, так и нестандартной многоядерно-мультитредовой архитектурой.

Суперкластеры [9] – это вынужденный шаг, обусловленный различными причинами: техническими, политическими, вопросами престижа. Эти изделия действительно необходимо создавать, но их нельзя рассматривать как полноценный вариант СКСН для решения стратегических задач России. Суперкластеры совсем не гарантируют сохранение стратегического паритета. Необходим адекватный конечным целям программы DARPA HPCS российский проект.

РОССИЙСКИЙ ПРОЕКТ СКСН "АНГАРА"

Единственный в России проект, похожий на DARPA HPCS, – это проект СКСН "Ангара" [3]. Он предполагает создание нового российского многоядерного мультитредово-потокового микропроцессора, коммуникационных СБИС и СБИС управления модулем памяти, а далее на их основе – СКСН с реальной производительностью свыше 1 PFlops, развиваемой на широком классе задач. Образцы такого СКСН реально изготовить после 2011 года, причем их характеристики будут близки к показателям перспективных американских СКСН.

В СКСН "Ангара" применяются такие перспективные аппаратно-программные суперкомпьютерные технологии, как:

- гетерогенная мультитредовая организация процессора и выполняемых программ с целью обеспечения толерантности к задержкам операций с памятью и коммуникационной сетью;
- управление вычислениями потоком данных с использованием статических и динамических графовых моделей;
- активное применение механизмов удаленного вызова процедур и удаленного выполнения команд;
- организация работы процессора и программ с выделением вычислительной и обслуживающей ее невычислительной частей для повышения параллелизма и асинхронности выполнения вычислений и работы с данными в программе;
- распределенная общая (глобально адресуемая) память;
- многоядерные однокристальные технологии;
- модули памяти со встроенными процессорами обработки данных;
- обеспечение повышенной отказоустойчивости на аппаратном, системном и прикладном уровнях;
- языки программирования для работы в глобальном адресном пространстве с неоднородным доступом (PGAS), обладающие повышенным уровнем абстракции, ориентированные на описание сверхпараллельных и асинхронных вычислений, с возможностью выполнения процедур над данными на удаленных узлах (локализации вычислений на данных);

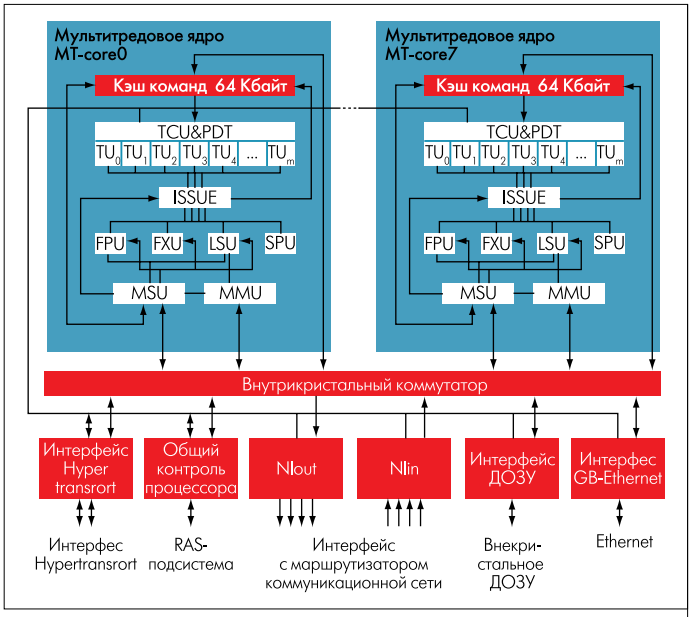

Рис.9. Многоядерный мультитредово-потоковый микропроцессор J10. TU_0–TU_m – тредовые устройства; TCU&PDT – блок управления и контроля тредовыми устройствами и таблицами регистров доменов защиты; FPU / FXU – функциональные устройства выполнения операций над числами с плавающей / фиксированной точкой; LSU – устройства операций с памятью; SPU – устройства специальных операций; MMU – блок трансляции виртуальных адресов; MSU – блок приема-выдачи пакетов сообщений, выдачи запросов запуска тредовых устройств и команд в функциональные устройства; ISSUE – блок выборки и проверки готовности команд для выполнения в функциональных устройствах и завершения выполненных команд

Таблица 3. Конфигурации микропроцессоров J7 и J10

Характеристика	J7-2	J7-3	J10-1	J10-2	J10-4
Частота процессора, ГГц	0,5	0,5	1	1	2
Число ядер/тредовых устройств в ядре	2/64	4/64	8/64	8/64	8/128
Число арифметических устройств с фиксированной/плавающей точкой	2/2	4/4	2/2	4/4	4/4
Число команд в ядре, выдаваемых за такт (ILP)	4	8	4	8	8
Пиковая производительность, GFlops	4	16	32	64	128
Объем кэш-памяти – расслоение, Мбайт/блоки	1/4	1/4	2/8	2/8	4/16
Пропускная способность кэш-памяти, Гбайт/с	64	64	128	128	256
Пропускная способность ДОЗУ, Гбайт/с	25,6	25,6	51,2	102,4	204.8
Дуплексная пропускная способность интерфейса сети, Гбайт/с	12	12	25	25	64

- адаптивный суперкомпьютинг, наличие аппаратных средств выполнения разнотипных вычислений, возможность их выбора и настройки даже в процессе счета.

Перечисленные принципы известны. Суть современных работ по СКСН – поиск их оптимального сочетания, что в равной мере относится и к американским, и к российским перспективным СКСН.

В проекте "Ангара" предусмотрены два варианта реализации микропроцессора – J7 (простой) и J10 (сложный) (рис.9, табл.3).

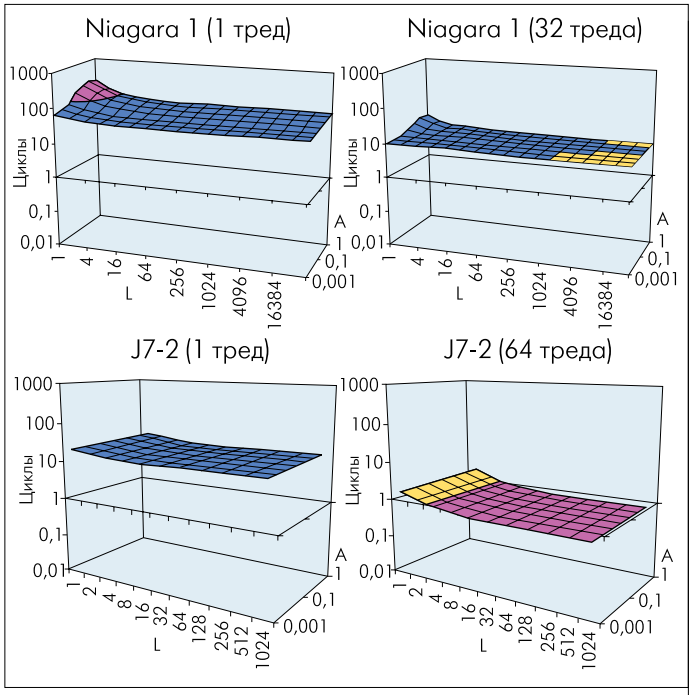

Рис.10. APEX-поверхности для многоядерных микропроцессоров с аппаратной поддержкой мультитредовости

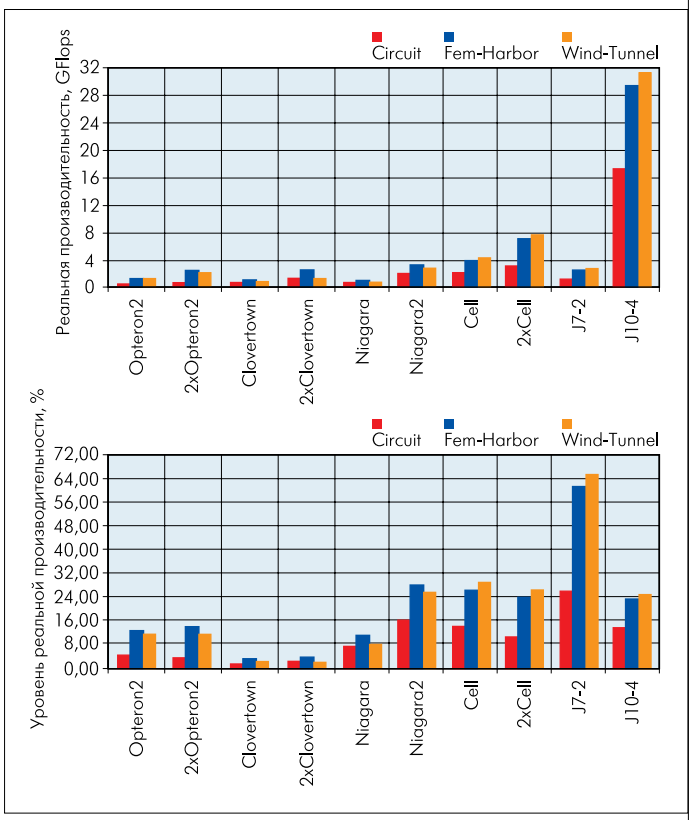

Рис.11. Реальная производительность и уровень этой производительности по отношению к пиковой на одном узле СКСН "Ангара" в разном исполнении в сравнении с узлами других систем

Первый поддерживает только гомогенную мультитредовую модель организации программ, работу с глобально адресуемой памятью и аппаратную обработку пакетов сообщений. Микропроцессор J10 дополнительно поддерживает гетерогенную мультитредовую модель, виртуализацию тредов и тредовых устройств, модели представления вычислений в виде статических и динамических графов потоков данных. Преложенные микропроцессоры зарубежных аналогов не имеют, их предполагается реализовать в виде заказных СБИС.

Аппаратная мультитредовость позволяет в процессорах J7 и J10 добиться толерантности при обращении к памяти (рис.10). Отметим, что микропроцессор Niagara 1 компании Sun также аппаратно поддерживает мультитредовость, но лишь до четырех тредов на ядро. Результаты тестов этих процессоров можно сравнить с аналогичными тестами для Clovertown и Barcelona (рис.4 и 6) – разница существенная.

Микропроцессоры J7 и J10 демонстрируют более высокие результаты и при сравнении эффективности на задаче CG, в которой преобладает выполнение операции умножения разреженной матрицы на вектор (рис.11). В эксперименте использовались различные разреженные матрицы (Circuit, Fem-Harbor и Wind-Tunnel) [10]. Похожий рекордный результат был получен на тесте поиска вширь BFS. Рассмотрены и другие тестовые задачи из области обработки сигналов и сеточных вычислительных методов.

Заслуживает особого внимания, что для СКСН "Ангара" при распараллеливании реальная производительность масштабируется гораздо лучше, чем у обычных систем (рис.12 и 13). Она близка к масштабированию на новейшем мультитредовом СКСН Cray XMT (Eldorado) [11], но на гораздо меньшем числе процессоров.

Авторам неизвестно о полученных где-либо в мире результатах, близких к приведенным на рис. 11–13. Обращаем внимание, что уровень реальной производительности J7-2 в десятки раз выше, чем у современных коммерческих микропроцессоров, а по абсолютному уровню J10-4 в несколько раз превосходит микропроцессор Cell с невероятно сложно оптимизированной программой [10]. Общий результат – на задачах DIS-класса наблюдается сильнейшее преимущество по отношению к известным системам, а на задачах CF-класса по крайней мере нет проигрыша.

ПРОЦЕССОР ДЛЯ ВСТРОЕННЫХ СУПЕРКОМПЬЮТЕРОВ

В настоящее время по проекту Минобрнауки РФ для встроенных суперкомпьютеров прорабатывается мультитредовый микропроцессор J10-M. В нем поддерживаются мелкозернистые статические графы на реконфигурируемом поле функциональных устройств. Аналогичный подход используется в микропроцессоре MONARCH [12], первом изделии программы DARPA PCA создания полиморфных микропроцессоров для перспективных ВСК (программа, родственная с DARPA HPCS).

В J10-M1 интегрированы четыре мультитредовых ядра (рис.14). Треды этих ядер обеспечивают не только толерантность к задержкам по памяти, но и быструю обработку внешних событий без прерывания вычислений. Внешние события реализуются активными сообщениями, поступающими в блок

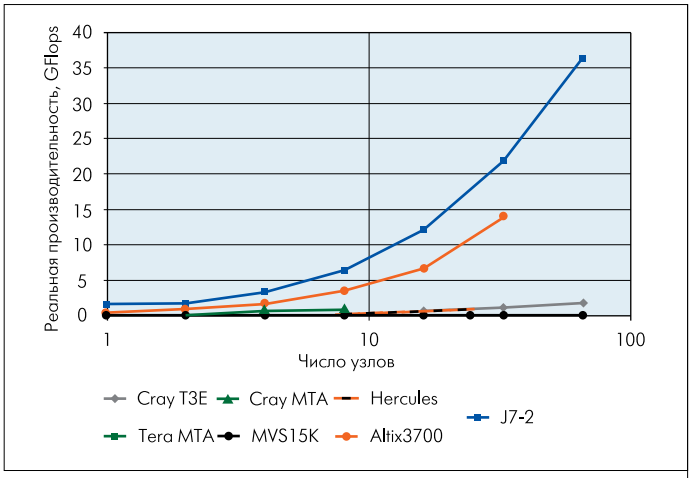

Рис. 12. Реальная производительность, получаемая при распараллеливании умножения разреженной матрицы на вектор на СКСН "Ангара" с микропроцессором Л7-2 в сравнении с другими системами

приема-выдачи пакетов сообщений в функциональные устройства (MSU) мультитредовых ядер. Эти сообщения могут автоматически запустить тредовое вычисление без приостановки остальных. Ядра через сетевые интерфейсы (NI) напрямую подключены к высокоскоростным четырехбитовым линкам. Сообщения с линков могут поступать в MSU-блок каждого ядра либо во встроенное динамическое ОЗУ ядра (ED). Все ядра связаны с реконфигурируемым полем арифметических кластеров (A) и кластеров памяти (M). Кластер A содержит восемь 32-разрядных арифметических устройств с плавающей и фиксированной точкой. Кластер M содержит блоки памяти, адресные сумматоры, контроллеры выполнения цепочек команд. Кроме того, реализована внутрикристальная сеть (PIRX) с соединениями "точка-точка" и с маршрутизаторами (P), а также четыре встроенных контроллера памяти с блоками выполнения атомарных операций (L2&AMO&FE) и два интерфейса RapidIO.

Сочетая мультитредовые и потоковые архитектурные средства (средней и мелкой зернистости), в будущем возможно создать суперкомпьютер даже экзафлопсного класса

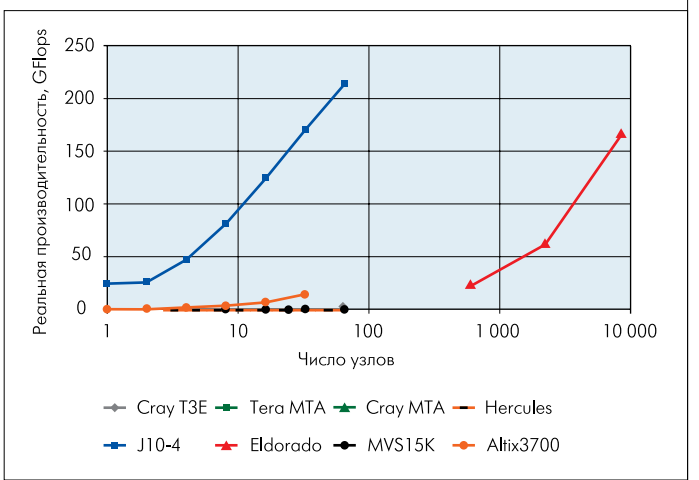

Рис. 13. Реальная производительность, получаемая при распараллеливании умножения разреженной матрицы на вектор на СКСН "Ангара" с микропроцессором Л10-4 в сравнении с другими системами

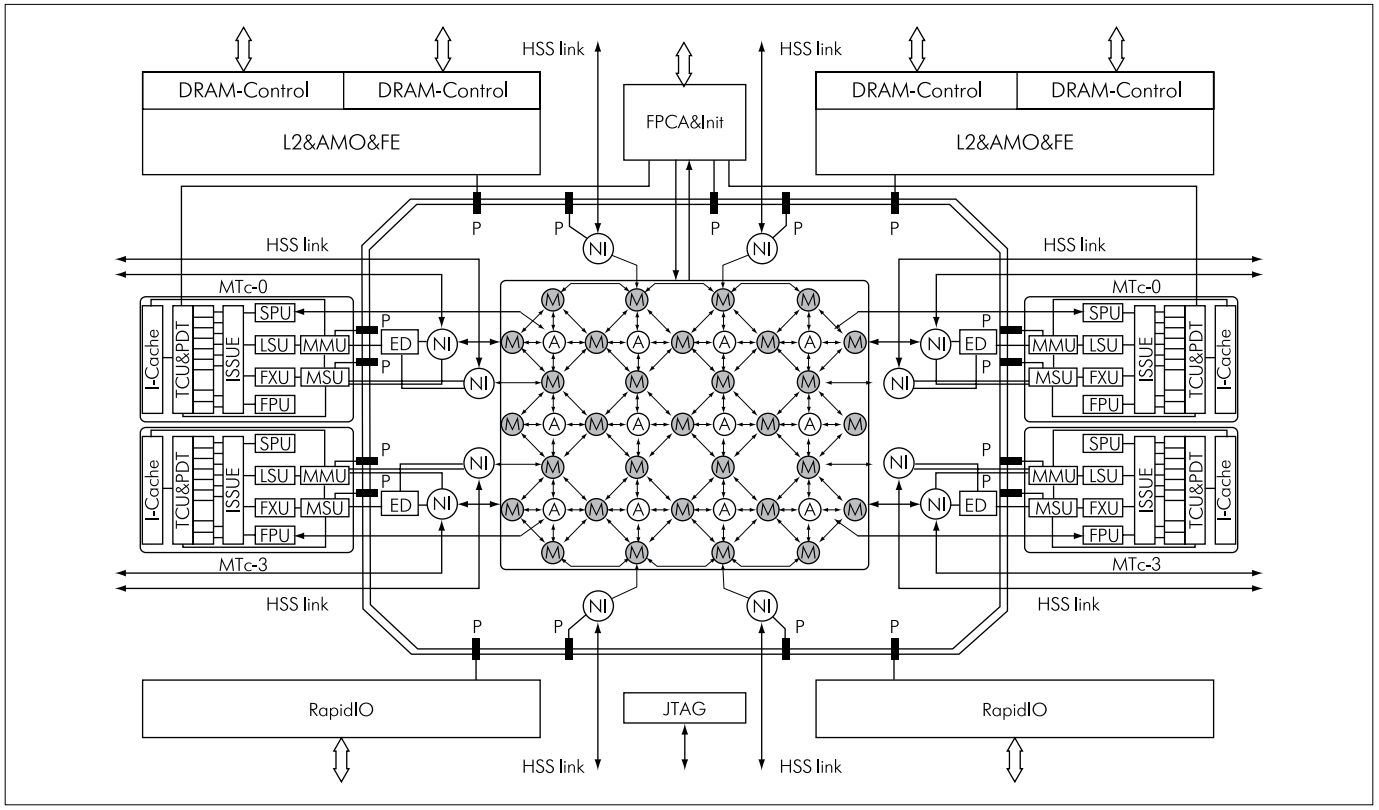

Рис.14. Структура модернизированного многоядерного мультитредово-потокового микропроцессора J10-M1 с решающим полем для реализации мелкозернистых потоковых моделей вычислений

(10^{18} операций в секунду). Это подтверждают данные по перспективным американским проектам.

НАПРАВЛЕНИЯ РАБОТ В ОБЛАСТИ СУПЕРКОМПЬЮТЕРНЫХ ВЫЧИСЛЕНИЙ

На начальном этапе проекта СКСН "Ангара" были сомнения, можно ли на основе относительно простых новых архитектурно-программных решений получить высокие результаты по реальной производительности и продуктивности. Но все сомнения развеяли оценочные программы и проведенные расчеты на имитационных моделях [3]. Это – главный фактор успеха. Остальное – хоть и очень сложная, но обычная инженерная работа. Но допустим, что проект СКСН "Ангара" решит поставленные задачи. Что делать дальше?

В США суперкомпьютерные технологии для новых СКСН будут создаваться по федеральному плану [5], аналога которому в России пока нет. Исследовательские работы по суперкомпьютерным технологиям будущего можно было бы организовать непосредственно в рамках проекта СКСН "Ангара", что обеспечило бы его развитие до уровня, близкого к экзафлопсному. Актуальны такие направления, как PGAS-языки и транзакционная память, реконфигурируемые решающие поля, сети с высокой связностью. Однако такой подход может оказаться еще одной причиной задержки развертывания проекта СКСН "Ангара".

Правильнее как можно быстрее запустить проект СКСН "Ангара", а вслед за ним подготовить основательную федеральную целевую программу по суперкомпьютерам типа федерального плана США [5]. Она могла бы включать разные направления (даже суперкластеры), но обязательно:

- технологии СКСН нового поколения;
- реверсивную логику и реверсивные вычисления;
- многозначную логику;
- RSFQ-технологии быстрой одноквантовой логики, квантовые клеточные автоматы и многое другое, что позволит расширить область применимости СКСН до задач, решение которых прежде считалось фантастикой [11].

По отношению к этим исследованиям проект "Ангара" и аналогичные ему могли бы играть роль флагманских, что повысило бы целенаправленность и качество фундаментальных исследовательских работ. При выполнении таких проектов важен и вопрос восстановления инфраструктуры отрасли создания суперкомпьютеров. Сегодня реально воссоздать существовавшую еще в СССР кооперацию организаций – ОАО "НИЦЭВТ", ИТМиВТ им. С.А. Лебедева РАН, ИПМ им. М.В.Келдыша РАН. К этой структуре реально подключение еще не менее 15 ведущих федеральных центров и предприятий России.

Резюмируя, отметим, что необходима разработка высокопроизводительных систем и суперкомпьютеров разного типа и мощности. Однако не следует забывать о главном. Область стратегических расчетов – это область заказных суперкомпьютерных технологий, специальных СКСН и ВСК. Лидеры в области суперкомпьютеров (СКСН и ВСК) обязаны быть впе-

реди в освоении новых идей и технологических возможностей. Допустимы компромиссные варианты в их разработке, но они не должны сбивать с основной идеи их создания в виде предельно эффективных уникальных аппаратных средств и программных систем, разработка аналогов которых вызовет трудности у соперников.

ЛИТЕРАТУРА

1. **J.Dongarra** et al. DARPA's HPCS Program: History, Models, Tools, Languages. – 2008.
2. **Фролов А., Семенов А., Корж А., Эйсымонт. Л.** Программа создания перспективных суперкомпьютеров. – Открытые системы, 2007, №9, с.20–29.
3. **Слуцкин А., Эйсымонт Л.** Российский суперкомпьютер с глобально адресуемой памятью. – Открытые системы, 2007, №9, с.42–51.
4. PUBLIC LAW 108-423-NON.30, 2004. DEPARTMENT OF ENERGY HIGH-END COMPUTING REVITALIZATION ACT OF 2004.
5. Federal Plan for High-End Computing. Report of the High-End Computing Revitalization Task Force. – EXECUTIVE OFFICE of the PRESIDENT of the UNITED STATES, National Science and Technology Council Committee on Technology, May, 2004.
6. **Волков Д., Фролов А.** Оценка быстродействия нерегулярного доступа к памяти. – Открытые системы, 2008, №1, с.15–19.
7. **S. Scott, D. Abts, J. Kim, W. Dally**. The BlackWidow High-Radix Close Network. – Stanford Univ., 2006.
8. **D.Abts, A.Batanieh, S.Scott, G.Faanes, J.Scwarzmeier, E.Lundberg, T.Jonson, M.Bye, G.Schwoerer**. The Cray BlackWidow: A Highly Scalable Vector Multiprocessor. SC07, November 10–16, 2007.
9. **Кудрявцев М., Мошкин Д., Полунин М., Эйсымонт Л.** Суперкластеры – между прошлым и будущим. – Открытые системы, 2008, №8, с.40–47.
10. **Williams, L.Oliker, R.Vuduc, J.Shalf, K.Yelick, J.Demmel**. Optimization of Sparse Matrix-Vector Multiplication on Emerging Multicore Platforms. – SC'07, November 10–16, 2007.
11. DOE/DOD Workshop on Emerging High Performance Architectures and Applications. – Washington, November 29–30, 2007.
12. **M.Vahey** et al. MONARCH: A First Generation Polymorphic Computing Processor. 2007.