

УДК 004.2

ОЦЕНОЧНОЕ ТЕСТИРОВАНИЕ КЛАСТЕРОВ НА БАЗЕ ПРОЦЕССОРОВ AMD BARCELONA И SHANGHAI С СЕТЯМИ INFINIBAND DDR И QDR

М. В. Кудрявцев¹, Д. В. Мошкин¹, М. А. Полуни², Л. К. Эйсымонт¹

Предложена методика и приведены результаты оценочного тестирования производительности кластеров на базе четырехпроцессорных и двухпроцессорных плат Supermicro с четырехъядерными микропроцессорами AMD Barcelona (тактовая частота 2,0 ГГц, пиковая производительность 32 Гфлопс) и AMD Shanghai (тактовая частота 2,6 ГГц, пиковая производительность 41,6 Гфлопс) и коммуникационных сетей Infiniband 4X DDR и 4X QDR. Используются оценочные тесты APEX-MAP, STREAM, EuroBen, NPВ и др. Проводятся сравнения с узлами и кластерами, построенными на базе четырехъядерных микропроцессоров Intel Clovertown. В перспективе предложенную методику оценочного тестирования планируется применить к кластерам на основе многоядерно-мультитредовых процессоров Intel Nehalem. Статья подготовлена по материалам доклада авторов на международной конференции “Параллельные вычислительные технологии” (ПаВТ-2009; <http://agora.guru.ru/pavt2009>).

Ключевые слова: кластеры, суперкластеры, процессоры, производительность, гигафлопс, APEX-MAP, STREAM, Infiniband, Barcelona, Shanghai.

1. Введение. Основная работа авторов данной статьи из ОАО “НИЦЭВТ” — создание суперкомпьютера стратегического назначения (СКСН) “Ангара” [1], являющегося, в некоторой степени, ответом на проект создания к 2010 г. американских СКСН с перспективной архитектурой транспетафлопсной производительности для решения задач обеспечения национальной безопасности и важнейших научно-технических задач [2, 3].

Создание СКСН такого типа в идеальном случае не предполагает использования в ней выпускавшихся до настоящего времени коммерчески доступных базовых компонентов ни в качестве процессоров, ни в качестве основной коммуникационной сети передачи данных между вычислительными узлами. Основная причина отказа в СКСН от коммерчески доступных микропроцессоров — их низкая толерантность (нечувствительность) к задержкам выполнения операций с памятью и сетью. Отказ от коммерчески доступных коммуникационных сетей (Quadrics, Myrinet, Infiniband) объясняется демонстрируемой ими низкой пропускной способностью при передаче коротких пакетов длиной до 64 байтов. Оба упомянутых свойства необходимы для реализации эффективного доступа к глобально адресуемой памяти огромного объема, от сотен терабайт до нескольких петабайт. Именно таких объемов требуют современные и перспективные приложения.

Пример современной “идеальной” (полностью заказной) СКСН такого типа — Cray BlackWidow [4, 5]. Можно привести следующие примеры СКСН как с коммерчески доступными, так и заказными компонентами: Cray XT4, XT5 (заказная сеть на маршрутизаторах Cray SeaStar [6]) и Cray XMT [7] (заказной мультитредовый микропроцессор Cray Threadstorm [8] и заказная сеть на Cray SeaStar), но с сервисными узлами на базе микропроцессоров AMD Opteron.

В настоящее время в коммерчески доступных микропроцессорах нового поколения стали появляться перспективные архитектурные свойства. Наиболее важные из этих свойств: многоядерность, мультитредовость, прямые каналы для подключения процессоров и ускорительных плат (HyperTransport — фирма AMD, QuickPath — фирма Intel), средства поддержки мелкозернистых и среднезернистых потоковых моделей вычислений.

Такие новые микропроцессоры представляют для нас интерес как средства макетирования разрабатываемых СКСН и как новые вычислительные средства для имитационного моделирования этих СКСН

¹ Научно-исследовательский центр электронной вычислительной техники, Варшавское шоссе, д. 125, 117587, Москва; М. В. Кудрявцев, нач. сектора, e-mail: atsign@yandex.ru; Д. В. Мошкин, нач. сектора, e-mail: den_m@nicvt.ru; Л. К. Эйсымонт, нач. отдела, e-mail: verger@nicvt.ru

² ЗАО ТРОНИК, Бизнес-центр “Кругозор”, ул. Обручева, д. 30/1, стр. 2, 117485, Москва; техн. директор, e-mail: mikep@tronicint.ru

(обе точки зрения на этот микропроцессор поясняются далее). Эти средства могут быть интересны также и для построения кластеров нового поколения — суперкластеров [12], в которых в какой-то степени можно уже увидеть свойства создаваемых перспективных СКСН.

Исследуемая четырехсокетная плата с четырехъядерными микропроцессорами Barcelona или Shanghai представляет интерес по следующим причинам. Во-первых, такую плату можно рассматривать в целом как прототип вычислительного узла СКСН “Ангара”, но с шестнадцатью тредовыми устройствами, которые реализуются ядрами этого микропроцессора. Суммарная производительность четырех микропроцессоров Barcelona платы — 120 Гфлопс. Во-вторых, эти микропроцессоры имеют встроенные контроллеры памяти, т.е. всего на плате имеется четыре таких контроллера с суммарной пиковой производительностью около 40 Гбайт/сек. Такие характеристики платы соответствуют параметрам разрабатываемого микропроцессора J10 в варианте J10-2 [1].

Параметры исследуемых узлов кластера приведены в табл. 1. Процессоры AMD Shanghai способны работать с памятью DDR2-800 МГц, однако в распоряжении авторов такой памяти не было, поэтому данное тестирование было произведено с использованием памяти DDR2-667 МГц.

Таблица 1

Параметры исследуемых узлов кластера

Обозначение (кол-во, номер модели CPU)	Материнская плата Supermicro	Тактовая частота CPU, ГГц	Память
2x AMD 2382 “Shanghai”	H8DME-2	2,6	DDR2-667 SDRAM
4x AMD 8350 “Barcelona”	AS-1041M-T2	2,0	DDR2-667 SDRAM
4x AMD 8380 “Shanghai”	AS-1041M-T2	2,5	DDR2-667 SDRAM
2x Intel E5345 “Clovertown”	SBI-7125B-T1	2,3	DDR2-667 FBDRAM

2. Экспериментальная часть

2.1. Исследование подсистемы памяти на тестах STREAM и APEX-MAP. STREAM [9] — простой синтетический тест, предназначенный для измерения реальной пропускной способности подсистемы памяти на простейших вычислительных ядрах с регулярным доступом к памяти. В табл. 2 приведены их описания.

Размер каждого вектора выбирается таким образом, чтобы быть по крайней мере в четыре раза больше суммарного размера всех кэшей последнего уровня многопроцессорной системы. В данных экспериментах каждый вектор имеет размер 8000000 элементов типа double (64-бит). Доступ к элементам векторов — регулярный, с единичным шагом. Тест отличается высокой пространственной локализацией (регулярный доступ производится к участкам памяти большой длины) и низкой временной локализацией данных (каждый элемент памяти используется только один раз), очень популярен и важен на практике, поэтому и был использован в первую очередь.

Операции пересылки теста STREAM могут выполняться одним или множеством тредов. Эти треды могут принудительно размещаться для выполнения на тех или иных ядрах процессоров, находящихся в сокетах исследуемой платы. Приведенные ниже измерения производились без указаний операционной системе о “привязке” тредов к ядрам процессоров и распределении памяти программы по сокетах: все это осуществляла операционная система по алгоритмам по умолчанию. Измерения производились 25 раз, и среди них выбиралось наилучшее по пропускной способности теста Copy. Такой подход позволяет избежать случайных вмешательств процессов операционной системы в измерения. На рис. 1 приведены результаты измерений теста STREAM ядра Triad (для остальных вычислительных ядер результаты похожие), полученные с использованием компилятора Intel ICC 10.1.018 (опции -O3 -static -openmp). С

Таблица 2
Описания ядер теста STREAM

Имя	Ядро	Флоп на итерацию
Copy	$a(i) = b(i)$	0
Scale	$a(i) = q*b(i)$	1
Sum	$a(i) = b(i) + c(i)$	1
Triad	$a(i) = b(i) + q*c(i)$	2

использованием компилятора GCC результаты получились хуже, несмотря на выставленные опции оптимизации под конкретную платформу, развертку циклов, использование SSE и пр.

Пропускная способность памяти на тесте STREAM Triad при переходе от четырехsocketной платы с процессорами Barcelona к четырехsocketной плате с процессорами Shanghai практически совпадает на 1 и 2 тредах. Оказались неожиданными различия на 4 и 8 тредах: на плате с процессорами Barcelona пропускная способность выше на 0,4–1,7 Гбайт/с. Объяснение этому пока не найдено. Однако на 16 тредах виден рост пропускной способности на 2,5 Гбайт/с.

Интересен факт, что на двухsocketной плате с процессорами Shanghai достигается до 1,5 раз бóльшая пропускная способность к памяти даже на одном треде по сравнению с четырехsocketной. Это увеличение вряд ли можно объяснить на 100 МГц большей тактовой частотой процессора AMD Opteron 2382 по сравнению с AMD Opteron 8380. В качестве гипотезы объяснения такого различия рассматриваются дополнительные накладные расходы на транзакции по шине HyperTransport в четырехsocketной плате.

Ускорения пропускной способности при увеличении количества тредов для платы с микропроцессорами Clovertown практически не происходит. Очевидно, что такое преимущество платы с микропроцессорами Barcelona объясняется наличием на ней четырех двухканальных контроллеров памяти DDR2 SDRAM. На плате с микропроцессорами Clovertown имеется один чипсет только с одним контроллером памяти FBDRAM. Такое явление масштабируемости пропускной способности при увеличении тредов означает наличие определенной толерантности (нечувствительности) платы к задержкам обращений к памяти — фактическое время их выполнения начинает определяться не временами задержек их выполнения, а темпом их выдачи.

Для детального исследования работы с памятью в разных режимах использовался тест APEX-MAP [10], который строит APEX-поверхность, характеризующую эффективность выполнения операций с памятью. APEX-MAP в данном случае показывает изменение среднего количества тактов процессора, приходящегося на одно обращение к памяти (все обращения на считывание) в зависимости от изменяемой тестом по некоторой схеме пространственно-временной локализации этих обращений. На рис. 2 ось L — это изменяемая пространственная локализация, а ось A — изменяемая временная локализация. Направление стрелки показывает увеличение локализации. Такая поверхность может быть построена для одного или множества узлов исследуемой вычислительной системы.

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-

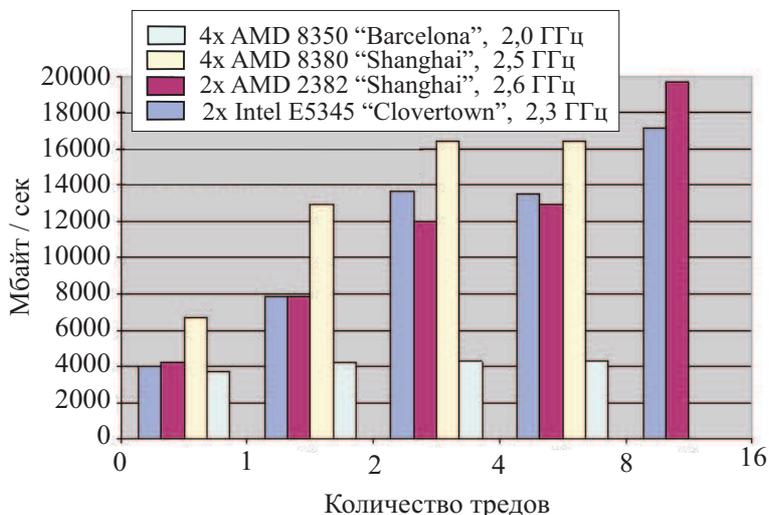


Рис. 1. Результаты теста STREAM Triad на исследуемых системах

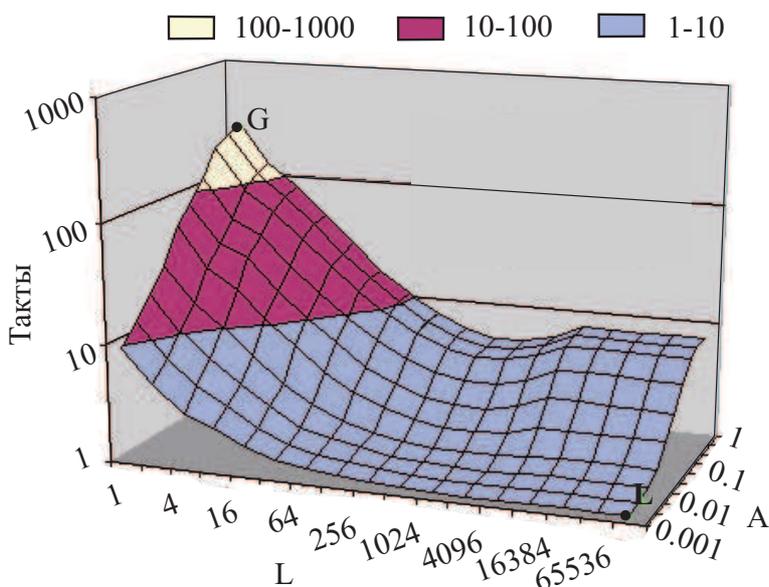


Рис. 2. APEX-поверхность для узла 2x AMD 2382 "Shanghai", полученная с использованием компилятора ICC

временной локализации.

В табл. 3 приведены данные по количеству тактов процессора, затрачиваемому на одно обращение к памяти в точках худшей (G) и лучшей (L) пространственно-временной локализации теста APEX-MAP. Это данные однотредовых запусков теста, в которых работа ведется либо с памятью из модуля одного (ближайшего) сокета, либо с памятью из модулей всех сокетов (четыре — для четырехсокетных плат и двух — для двухсокетных).

Таблица 3

Количество тактов процессора, затрачиваемое на одно обращение к памяти в точках G и L теста APEX-MAP (погрешность 10%)

Узел кластера	Компилятор	Худшая локализация (точка G)		Лучшая локализация (точка L)	
		1 сокет	все сокеты	1 сокет	все сокеты
4x AMD 8350 "Barcelona"	GCC	306	408	3,60	3,55
	ICC	325	419	1,17	1,17
4x AMD 8380 "Shanghai"	GCC	317	496	3,60	3,50
	ICC	331	517	1,14	1,17
2x AMD 2382 "Shanghai"	GCC	205	301	3,48	3,47
	ICC	231	323	1,09	1,09

На рис. 3 и 4 приведены данные по влиянию повышения количества тредов в точках худшей и лучшей пространственно-временной локализации теста APEX-MAP соответственно.

Отметим, что данные табл. 3 и диаграмм на рис. 3 и 4 приведены в тактах. Для оценки времени одного обращения в память нужно учитывать тактовую частоту процессора. Главный полученный результат этих измерений — обнаружение при увеличении параллелизма (количества тредов) резкого уменьшения задержки на одно обращение к памяти (толерантности) для разных вариантов пространственно-временного обращения к памяти. Так, в точке G задержка уменьшилась с 407 до 31 на плате с процессорами Barcelona и с 496 до 50 на четырехсокетной плате с процессорами Shanghai. В точке L задержка уменьшилась с 3,5 до 0,1 на плате с процессорами Barcelona и с 3,5 до 0,2 на четырехсокетной плате с процессорами Shanghai.

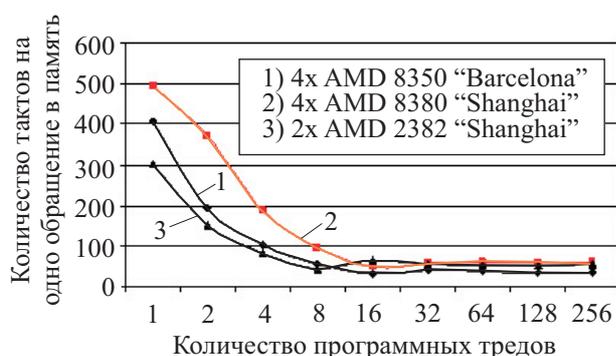


Рис. 3. Влияние мультитредовости в точке G теста APEX-MAP. Комппилятор GCC

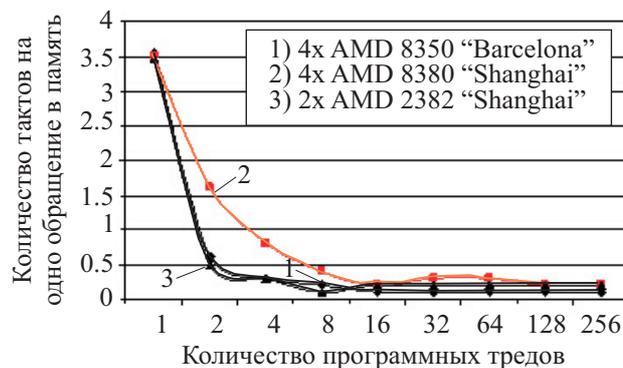


Рис. 4. Влияние мультитредовости в точке L теста APEX-MAP. Комппилятор GCC

2.2. Исследование производительности одного ядра на пакете EuroBen. Выше было показано, что исследуемые системы на базе процессоров AMD обладают свойствами толерантности при регулярной и нерегулярной работе с памятью. Отметим, что это явление уникально, для других плат оно не наблюдалось.

Возникает вопрос, как отразится такое свойство на развиваемой реальной производительности не на специальных, а на прикладных оценочных тестах и программах. При этом сравнение будет вестись с платами, которые, как выяснилось, толерантностью не обладают. В данном исследовании — это платы с микропроцессором Intel Clovertown E5345 с частотой 2,3 ГГц.

Понятно, что развиваемая реальная производительность зависит не только от работы с памятью, но и от качества реализации собственно ядер процессора, что включает функциональные устройства, кэши данных разного уровня и устройство конвейера команд. Эти сравнения исследуемого оборудования

проводились на пакете тестов EuroBen 4.2, модуле mod1ac.

При тестировании выделены группы тестов (всего 10 групп), составленные таким образом, чтобы показать ту или иную особенность тестируемого оборудования. Тесты внутри группы в некоторых случаях выбирались по определенной логике изменения нагрузки с целью выявления дополнительных особенностей объектов тестирования.

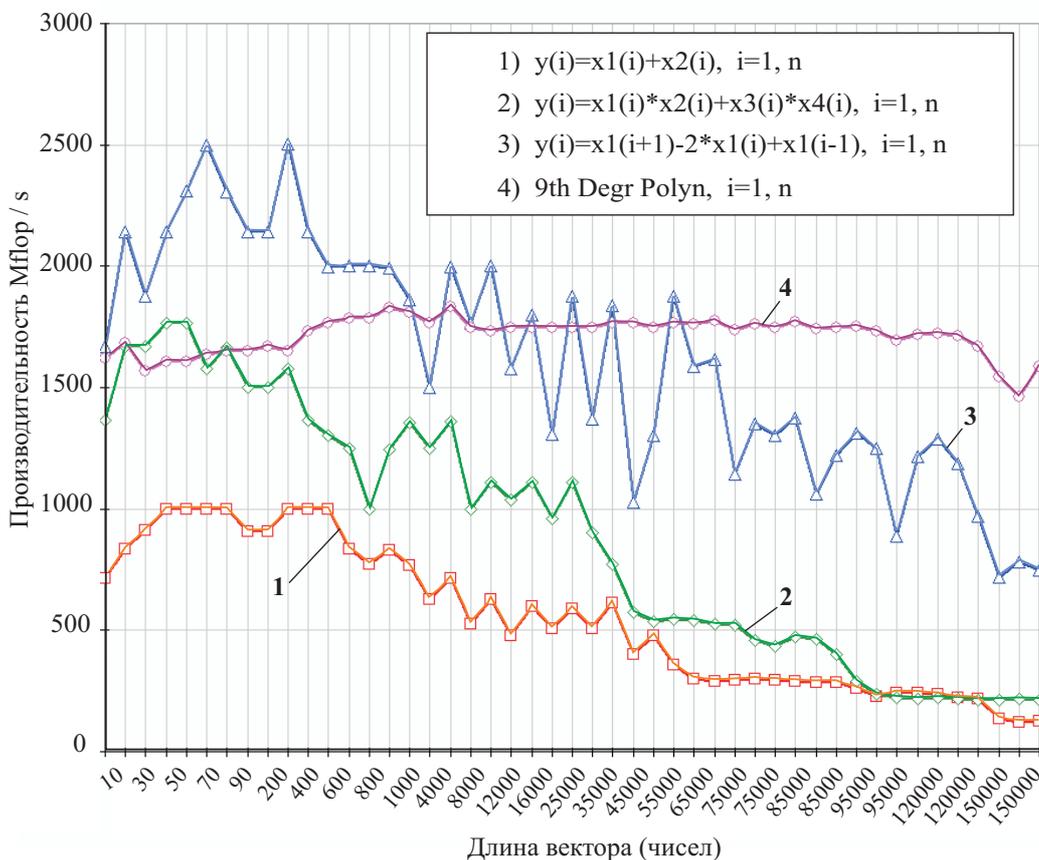


Рис. 5. Зависимость реальной производительности одного ядра процессора Barcelona (на четырехсокетной плате) от количества арифметических операций, приходящихся на одно обращение к памяти

На рис. 5 для примера показаны характеристики, полученные для одной из этих групп. Группа подобрана так, что от теста к тесту меняется количество арифметических операций, приходящихся на одно обращение к памяти. Видно, что реальная производительность заметно зависит от этого параметра задачи-теста.

Исследование на тестах EuroBen 4.2 позволило сделать вывод, что качество реализации ядра процессоров Clovertown несколько выше, чем процессора Barcelona, да и частота у них выше. Тем не менее, результаты на прикладных бенчмарках показали, что при таком отставании по качеству реализации ядра процессор Barcelona за счет своей уникальной организации подсистемы памяти и толерантности к ее задержкам позволяет получить близкую, а иногда и большую реальную производительность, особенно при высокой степени загрузки ядер. Ядра процессора Shanghai уже не уступают ядрам процессора Clovertown на тестах EuroBen 4.2.

2.3. Исследование производительности нескольких ядер на пакете NPВ. На рис. 6 и 7 для двух задач, LU и CG (пакет задач гидрогазодинамики NPВ 3.1), представлены зависимости реальной производительности, приходящейся на одно ядро микропроцессора, от степени распараллеливания задачи, а главное — от степени загрузки ядер платы. Часть результатов получена на кластере МВС-100К Межведомственного суперкомпьютерного центра РАН. В каждом его узле установлено по два четырехъядерных процессора Clovertown/3.0 ГГц. В многоузловых запусках использовалась сеть Infiniband DDR 4X с однопортовым подключением к маршрутизатору. Числа над линиями графиков в кружках показывают, сколько ядер на плате загружено: 1 — одно ядро из всех имеющихся, 2 — два ядра и т.д. Видно, что

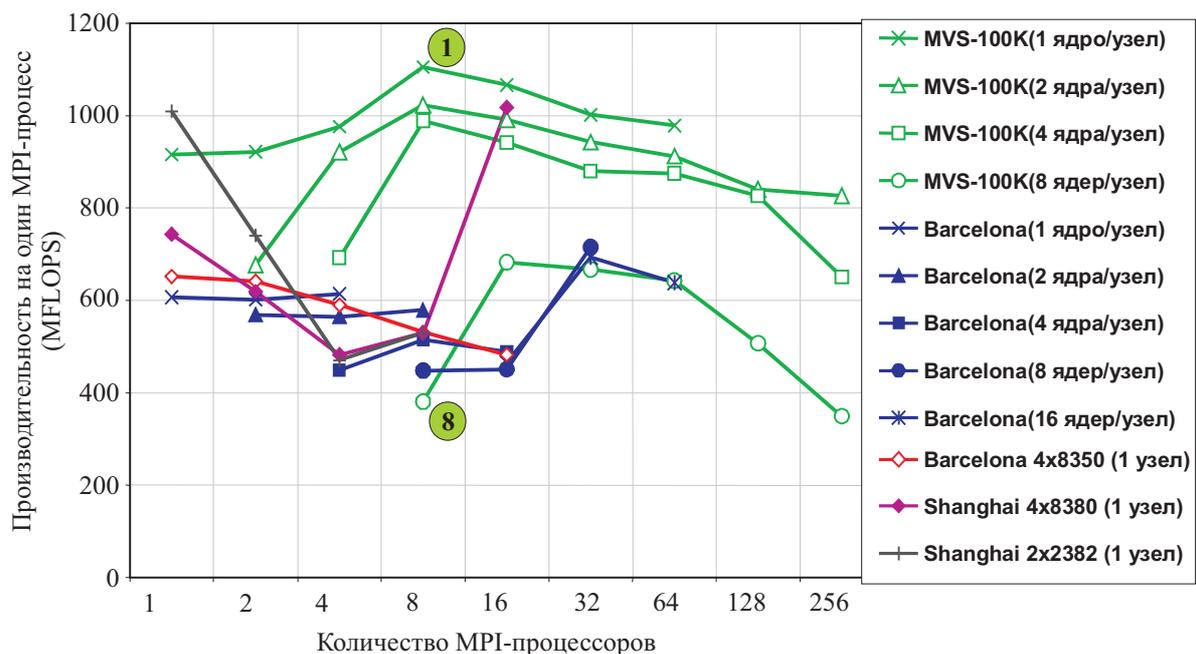


Рис. 6. Реальная производительность на один процессор для задачи LU класса C в зависимости от количества MPI-процессов и степени загрузки процессорных ядер

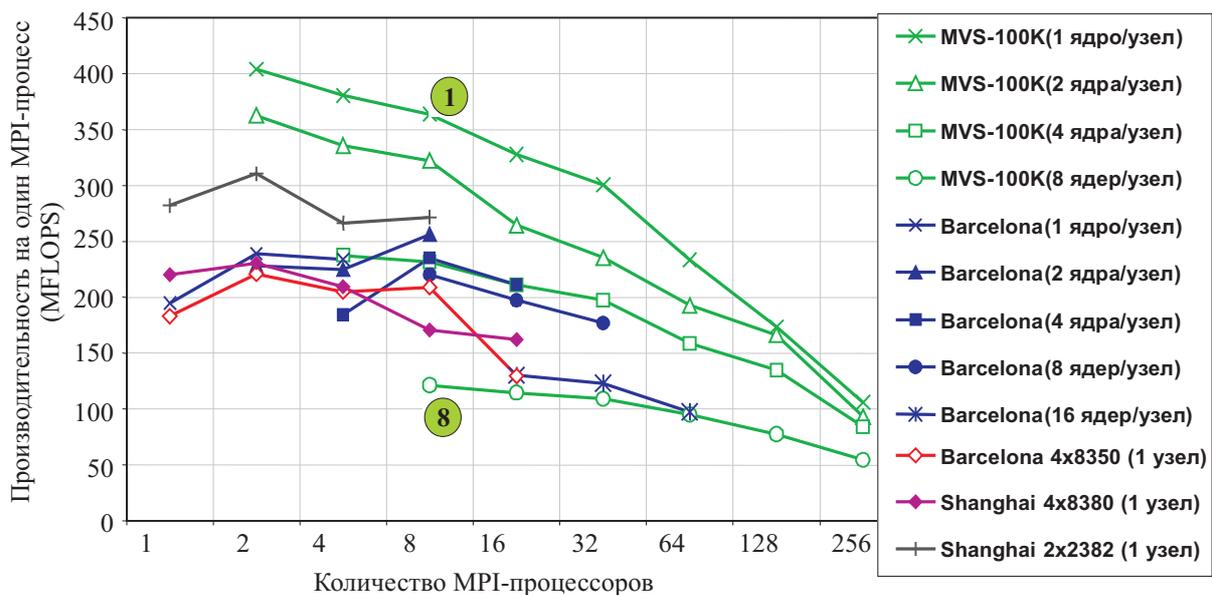


Рис. 7. Реальная производительность на один процессор для задачи CG класса C в зависимости от количества MPI-процессов и степени загрузки процессорных ядер

процессор Clovertown лучше всего работает, когда на восьмиядерной плате загружено только одно ядро, а остальные ядра не используются. Если же загрузить все 8 ядер, то производительность может упасть в три раза и более, т.е. общая производительность за счет многоядерности тогда повышается приблизительно в два, два с половиной раза.

Рассмотрение зависимостей рис. 6 и 7 показывает, что разброс реальной производительности при использовании большого количества ядер для платы с микропроцессорами Barcelona значительно меньше. В связи с этим общую реальную производительность в целом с платы удастся получить ненамного меньше, а иногда и больше, чем на процессоре Clovertown с большей в 1,5 раза частотой и лучшими по реализации ядрами.

При оценке результатов этих исследований важно помнить, что необходимость достижения требуемой при распараллеливании эффективности за счет отказа от использования ядер на плате приводит к тому,

что используется большее количество узлов. Это уже другие затраты денег и энергии — в сравнении со случаем, когда удастся эффективно загрузить все ядра платы.

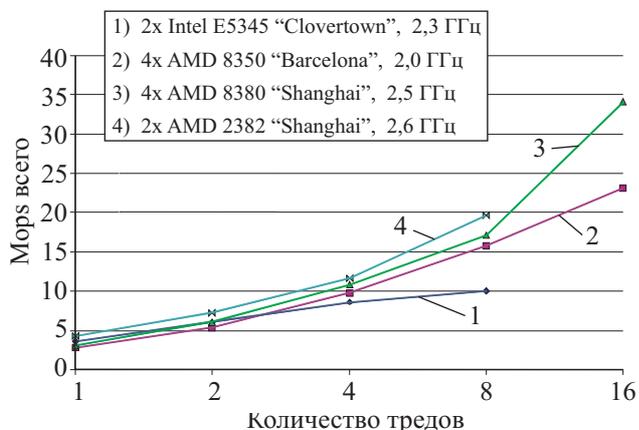


Рис. 8. Реальная производительность на один процессор для задачи UA класса C в зависимости от количества OpenMP-тредов

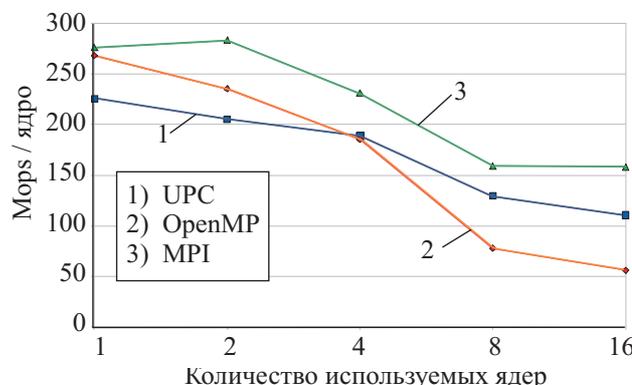


Рис. 9. Реальная производительность на одно ядро задачи CG класса B в зависимости от количества используемых ядер 4x AMD 8380 "Shanghai"

При оценочном тестировании рассматривался также тест UA из пакета NPB 3.3. Это расчет на неоднородной, перестраиваемой в процессе счета сетке (рис. 8). Процессор Barcelona/2 ГГц на этой задаче при загрузке 8 ядер одного узла показывает приблизительно в 1,5 раза большую реальную производительность в сравнении с 8 ядрами процессора Clovertown/2,3 ГГц. Эта производительность при использовании 16 ядер платы с процессорами Barcelona превосходит 8 ядер Clovertown уже в 2,4 раза (и в 3,5 раза при использовании процессоров Shanghai/2,6 ГГц).

Для выявления возможностей эффективного использования вычислительных моделей OpenMP и UPC на исследуемых вычислительных узлах с процессорами AMD были проведены эксперименты над задачей NPB CG. На рис. 9 представлена производительность на одно ядро, развиваемая на задаче CG класса B, запрограммированной с использованием различных средств программирования — MPI, OpenMP, UPC. Заметна интересная тенденция — на большом количестве ядер (8–16) задача CG, написанная на языке UPC, превосходит по производительности OpenMP-реализацию. Возможно, это связано с тем, что в интерфейсе OpenMP отсутствуют средства локализации данных, что не позволяет разрабатывать эффективные программы для систем с неоднородным доступом к памяти.

2.4. Исследование бисекционной пропускной способности сетей Infiniband 4X DDR и 4X QDR. Тест оценки бисекционной пропускной способности входит в пакет тестов MPPTEST, доступный на сайте [11]. Бисекционная пропускная способность — пропускная способность между равными частями системы. В тесте все процессы разбиваются на пары, внутри каждой пары осуществляются пересылки.

В данном оценочном тестировании определялась бисекционная пропускная способность между двумя узлами: первый процесс из каждой пары расположен на первом узле, второй — на втором. Смысл исследования заключается в том, чтобы выяснить, повысится ли бисекционная пропускная способность при увеличении количества взаимодействующих пар. Если повысится, то это будет означать, что платы обладают толерантностью к задержкам сети. Для тестирования бисекционной пропускной способности программе mppctest необходимо указать опцию “-bisect”. В тесте можно использовать несколько типов обмена сообщений: “Round Trip”, “Round Trip with nonblocking sends/receives” и “Head to head with nonblocking sends/receives”. Был выбран тип “Round Trip”, в котором обмен сообщений между двумя процессами происходит так же, как и в стандартном тесте Ping-Pong. На рис. 10 представлено ядро теста bisect с таким типом обмена.

Процесс 1 (PING-процесс)	Процесс 2 (PONG-процесс)
<pre>for (i=0; i<reps; i++) { { MPI_Send(...); MPI_Recv(...); } }</pre>	<pre>for (i=0; i<reps; i++) { MPI_Recv(...); MPI_Send(...); }</pre>

Рис. 10. Ядро теста Bisect

Приведенные на рис. 11 характеристики для сетей DDR и QDR показывают, что есть масштабируемость бисекционной пропускной способности, а значит, есть и толерантность. Сеть QDR ведет себя

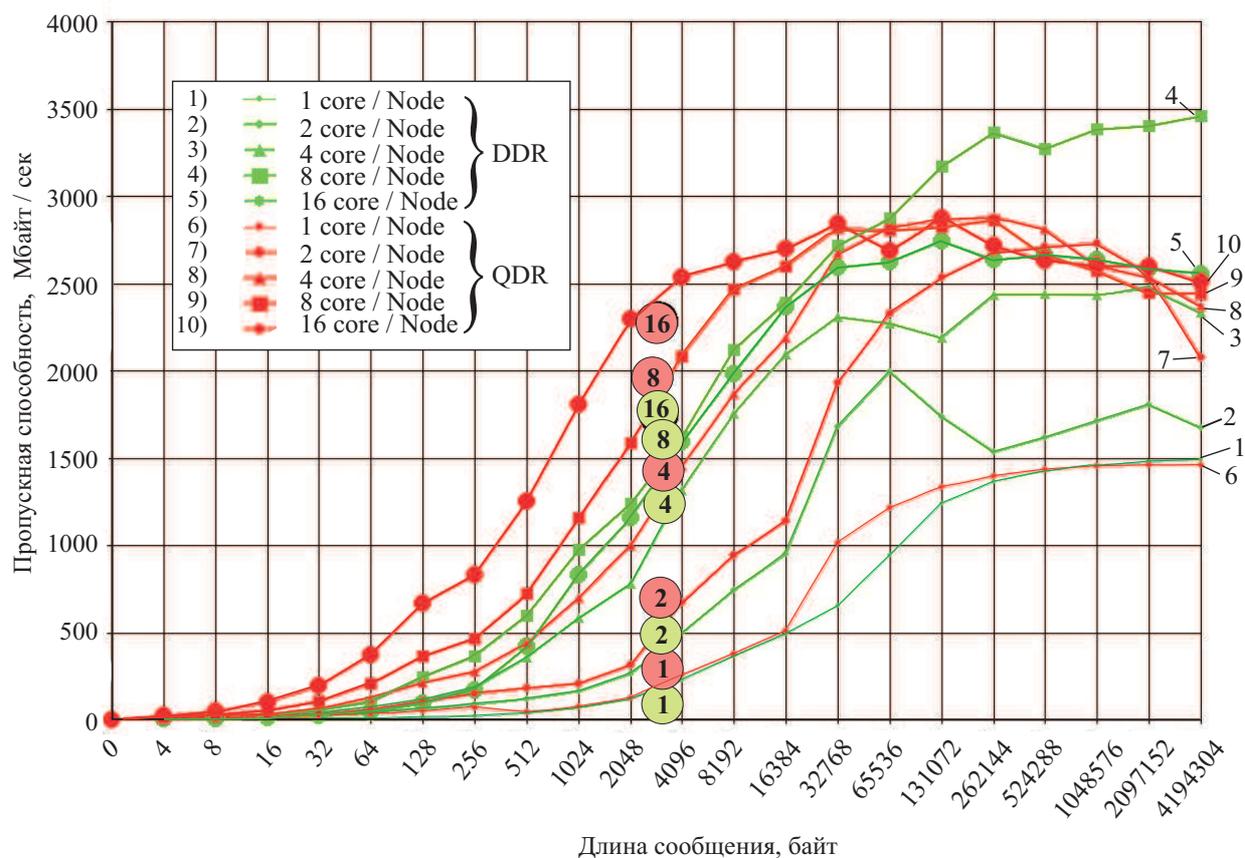


Рис. 11. Сравнение бисекционной пропускной способности для узлов кластера Barcelona, сетей 4X DDR и 4X QDR с двухпортовым подключением узла к коммутатору

значительно лучше на пакетах средней и малой длины — важное качество для приложений со средней и малой зернистостью пересылок.

3. Заключение. Рассмотренная плата с четырьмя четырехъядерными микропроцессорами Barcelona и Shanghai представляет, на наш взгляд, уникальное явление как возможный компонент усовершенствованных кластеров или суперкластеров [12]. Впервые в коммерчески доступной плате удалось добиться стабильной эффективности ядер вместо резкой их деградации при увеличении загрузки, что было раньше и что явно видно на платах с микропроцессором Clovertown. Исследования такого типа будут проводиться и впредь. Например, по методике данной работы планируется в марте 2009 г. исследование узлов кластера на базе процессора Intel Nehalem.

Правильное использование таких плат с применением перспективных вычислительных моделей и системного программного обеспечения нового поколения позволит выйти на классы приложений, где раньше эффективно применялись дорогостоящие вычислительные системы фирм SGI и IBM с большими многопроцессорными узлами с общей памятью (SMP-узлами).

Большое количество ядер и наличие эффективных прямых каналов прямого подключения сопроцессоров и прямого соединения кристаллов (HyperTransport и QuickPath) открывает значительные возможности создания кластеров нового типа (нам встречалось название “псевдокоммерческие суперкластеры”), в которых были бы, пусть частично, реализованы перспективные архитектурные свойства, аналогичные тем, которые в полном объеме будут доступны в СКЧН “Ангара”.

СПИСОК ЛИТЕРАТУРЫ

1. Слущкин А., Эйсымонт Л. Российский суперкомпьютер с глобально адресуемой памятью // Открытые системы. 2007. № 9. 42–51.
2. Фролов А., Семенов А., Корж А., Эйсымонт Л. Программа создания перспективных суперкомпьютеров // Открытые системы. 2007. № 9. 20–29.
3. Митрофанов В., Слущкин А., Эйсымонт Л. Современное состояние и перспективы развития суперкомпьютер-

- ных технологий для стратегически важных задач // Материалы конференции “Перспективы развития высокопроизводительных вычислительных архитектур”. № 1. М.: ИТМиВТ им. С. А. Лебедева РАН, 2008.
4. *Abts D. et al.* The Cray BlackWidow: a highly scalable vector multiprocessor // Proceedings of the 2007 ACM/IEEE Conference on Supercomputing (SC’07). New York, 2007.
 5. *Scott S., Abts D., Kim J., Dally W.J.* The BlackWidow high-radix clos network // International Symposium on Computer Architecture (ISCA). Boston, MA, USA. 2006. 16–28.
 6. *Brightwell R., Pedretti K.T., Underwood K.D., Hudson T.* SeaStar Interconnect: Balanced Bandwidth for Scalable Performance // IEEE MICRO. **26**, Issue 3. 2006. 41–57.
 7. *Konecny P.* Introducing the Cray XMT // Proc. Cray User Group Meeting (CUG 2007). Seattle, WA. 2007.
 8. CRAY/MTA Principles of Operation // Cray Inc. 2005.
 9. <http://www.cs.virginia.edu/stream/FTP/Code/stream.c>
 10. *Strohmaier E., Shan H.* Apex-Map: A Global Data Access Benchmark to Analyze HPC Systems and Parallel Programming Paradigms // Proceedings of the 2005 ACM/IEEE Conference on Supercomputing (SC’05). Washington, DC, 2005.
 11. <http://www-unix.mcs.anl.gov/mpi/mpptest/>
 12. *Кудрявцев М., Мошкин Д., Полунин М., Эйсымонт Л.* Суперкластеры — между прошлым и будущим // Открытые системы. 2008. № 8. 40–47.

Поступила в редакцию
28.02.2009
