Биатлон[1] для СКИФов: быстро и точно

С. М. Абрамов, А. И. Адамович, М. Р. Коваленко, В. А. Роганов

Институт программных систем РАН

Статья посвящена вопросу использования изделий фирмы AMD в составе перспективных моделей высокопроизводительных кластерных установок семейства СКИФ, создаваемых в рамках суперкомпьютерной Программы «СКИФ» Союзного государства. Дается анализ современных процессоров AMD Athlon MP, двухпроцессорных системных (материнских) плат для них, описывается опыт выбора конкретных изделий для выпуска ближайших серий кластеров СКИФ, результаты тестирования различных конфигураций возможных вычислительных модулей.

The paper is devoted to discussion of using AMD products in perspective models of high-efficient cluster plant of SKIF family. These models are developed in the network of supercomputer program "SKIF" of the Union State. Analysis of modern processors AMD Athlon MP and two-processor motherboards is carried out. Choice criteria for concrete parts of nearest series of SKIF clusters are described. Results of tests for differend configurations of possible computers are cited.

1. Краткое описание Программы «СКИФ»

Основной целью суперкомпьютерной Программы «СКИФ» Союзного государства является разработка и освоение в серийном производстве семейства программно-совместимых высокопроизводительных вычислительных систем и ряда прикладных систем на базе этих вычислительных систем. Государственные заказчики-координаторы Программы «СКИФ»: Национальная Академия наук Республики Беларусь и Министерство промышленности, науки и технологий Российской Федерации, в программе участвуют около 20 предприятий:

- *от Республики Беларусь*: НИО «Кибернетика» (ответственный исполнитель от Республики Беларусь), УП «Белмикросистемы», УП «НИИ ЭВМ», ИТМО НАН Беларуси и другие;
- *от Российской Федерации*: ИПС РАН (ответственный исполнитель от Российской Федерации), МГУ, НИЦЭВТ, ИВВиИС, Предприятие «Суперкомпьютерные системы» и другие.

Обычно раньше традиционными потребителями высокопроизводительных вычислений являлись оборонные предприятия и ведомства. Сегодня в таких технологиях объективно нуждаются множество крупных учреждений и организаций в различных отраслях: предприятия химической промышленности и фармацевтики, энергетики, машиностроения, структуры управления, фирмы, работающие на рынке телекоммуникационных и информационных услуг, различные научные учреждения, ВУЗы и т.д. В дальнейшем тенденция к расширению сферы применения высокопроизводительных вычислений будет только усиливаться. Именно поэтому в Программе «СКИФ» особое внимание уделяется изготовлению не одиночных рекордных систем, а разработке конструкторской и программной документации, пригодной для серийного производства суперкомпьютеров среднего класса, имеющих не только высокие технические характеристики, но и хорошие показатели «стоимость/производительность», что учитывает финансовые возможности потенциальных потребителей такой техники в Союзном государстве.

В соответствии с разработанной концепцией по Программе «СКИФ» модели семейства «СКИФ» имеют (в общем случае) двухуровневую архитектуру (Рис. 1):

[1]Биатлон — двухпроцессорная материнская плата с процессорами AMD Athlon (профессиональный юмор).

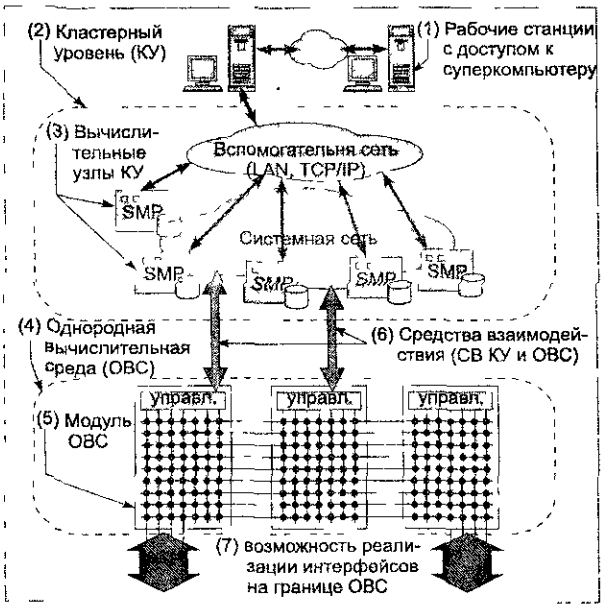


Рис 1 Структура суперкомпьютеров «СКИФ»

- *кластерный уровень* – Linux-кластер, использующий современные скоростные средства комплексирования для системной сети,
- *уровень ОВС* – отечественный спецвычислитель.

В качестве программного обеспечения на кластерном уровне используются операционная система Linux с ее штатными компиляторами и библиотеками, стандартные средства организации параллельных вычислений (MPI, PVM, и т. п.) и отечественная система автоматического динамического распараллеливания программ – Т-система. Подробное обсуждение уровня ОВС выходит за рамки данной статьи Завершая общее описание архитектуры суперкомпьютеров семейства «СКИФ», мы отметим, что

- Кластерный уровень универсален: на нем можно реализовать любую прикладную систему. Особенно он эффективен для реализации фрагментов прикладных задач со сложной логикой вычисления и с крупноблочным параллелизмом Такие фрагменты могут быть эффективно реализованы на кластерном уровне, в том числе с использованием Т-системы. Как правило, для комплексирования кластерного уровня в установках «СКИФ» сегодня используется сеть SCI[2] на базе изделий фирмы Dolphin[3].
- Уровень ОВС эффективен для реализации фрагментов прикладных задач с *простой логикой вычисления, с конвейерным или мелкозернистым явным параллелизмом*, с большими потоками информации, требующими обработки в реальном режиме времени.
- В соответствии с модульным подходом в составе установки семейства «СКИФ» может быть произвольное число модулей кластерного уровня и произвольное число модулей ОВС (в некоторых случаях ОВС может не использоваться вовсе) Для той или иной прикладной системы, после анализа целевой задачи и

[2]SCI — scalable coherent interface.

[3]Сайт фирмы Dolphin Interconnect Solutions AS: `http://www.dolphinics.com`.

дефрагментации ее на подзадачи (одни из которых окажутся «удобными» для реализации на КУ, другие - для реализации на ОВС), можно выбрать оптимальный для данного случая состав установки «СКИФ»: оптимальное количество модулей КУ и оптимальное количество модулей ОВС.

В настоящее время модули ОВС пока еще находятся в разработке. Поэтому все первые образцы изделий семейства «СКИФ» (например, см. Рис. 2) содержали только кластерный уровень

Предельная пиковая производительность	ок. 20 GFlops
Число процессоров (Intel Pentium III-600 MHz)	32 шт.
Число вычислительных узлов:	16 шт.
Оперативная память:	$16 \times 0.5 = 8$ ГБ
Дисковая память:	$16 \times 10 \approx 160$ ГБ
Системная сеть SCI	2D-тор 4×4
задержка (MPI, не хуже)	6 мкс
скорость MPI-обменов (точка-точка)	до 120 МБ/с
– физич. скорость обмена	до 800 МБ/с

Рис. 2: Первый экспериментальный образец установки семейства «СКИФ» и его технические характеристики (декабрь 2000 г.)

Дальнейшее обсуждение будет посвящено только аппаратным средствам кластерного уровня суперкомпьютеров семейства «СКИФ». То есть, по сути, все изложенное ниже можно считать справедливым для любых Linux-кластеров, построенных с использованием сети SCI на базе современных изделий фирмы Dolphin.

2. Поиск альтернативы

Концепция архитектуры изделий семейства «СКИФ» предусматривает применение в качестве вычислительных узлов кластерного уровня произвольных процессоров и произвольных материнских плат (предпочтительно, поддерживающих SMP), для которых возможно использование OS Linux. Однако по разным причинам на первом этапе исполнения Программы (2000–2001 годы) разработчики ограничивались рассмотрением процессоров фирмы Intel и соответствующих материнских плат. В этом году мы решили попробовать альтернативные подходы. Для этого мы провели ряд экспериментов с различными типами процессоров AMD AthlonMP и SMP-платами для них. Соответствующие аппаратные средства для экспериментов нам любезно предоставили Представительство[4] AMD в Москве, фирма «Традиция»[5] и ОАО «НИЦЭВТ»[6].

Целью экспериментов была оценка эффективности использования решений AMD для вычислительных узлов кластерного уровня, сравнение различных конфигураций вычислительных узлов. Для измерения различных характеристик использовались следующие тесты:

1. *Оценка производительности на вычислительных задачах*: тест linpack-odu из семейства тестов Linpack[7]. Тест состоит в решении системы линейных уравне-

[4]Сайт представительства: http://www.amd.com/ru-ru.

[5]Сайт фирмы «Традиция»: http://www.tradition.ru

[6]Сайт ОАО «НИЦЭВТ». http://www.nicevt.ru.

[7]Описание теста: http://netlib2.cs.utk.edu/benchmark/linpackc.

ний с помощью LU-факторизации. Основное время счета связано с выполнением векторных операций.

2. *Оценка эффективности MPI-обменов между вычислительными узлами по сети SCI*: тесты all2all, send-receive, ping-ping и ping-pong[8] из поставки SCALI SSP[9]. Измеряемые характеристики: скорость передачи (bandwidth) и задержка (latency).

3. Используемые аппаратные средства

В тестах использовались два типа процессоров: AMD Athlon MP 1800+, AMD Athlon MP 2000+; три типа системных (материнских) плат: ASUS A7M266-D, TYAN Thunder K7X, TYAN Tiger MPX; адаптеры SCI: Dolphin SCI PCI-64/66 / D330.

4. Оценка производительности платформ AMD на вычислительных задачах

В результате измерения реальной производительности на тесте Linpack было обнаружено, что производительность определяется в основном типом процессора: она не зависит (с точностью около 1%) от типа используемых системных плат.

На рисунке 3 результаты теста Linpack для предоставленных нам процессоров AMD сравниваются с результатами, полученными для имеющихся у нас процессо-

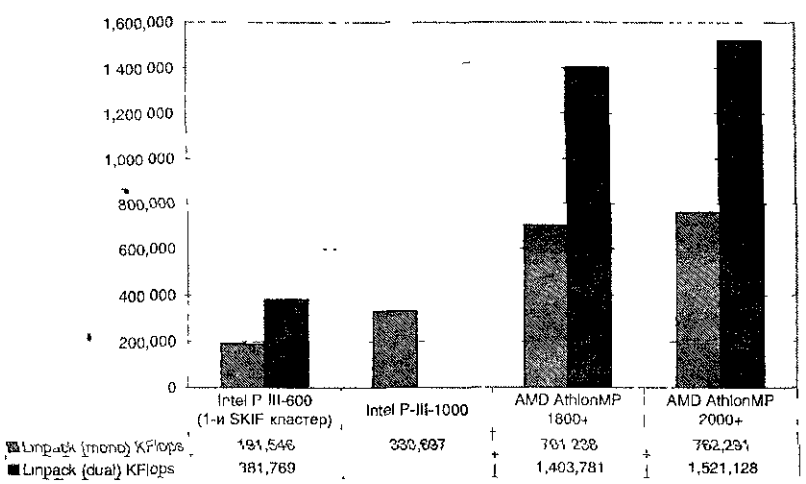


Рис. 3 Результаты теста Linpack на различных платформах

Таблица 1

Эффективность реализации процессоров: Linpack Flops/Hz

Тип процессора	частота (MHz)	Linpack Flops/Hz
Intel P-III-600 (1-й SKIF-кластер)	600	0.318
Intel P-III-1000	1,007	0.328
AMD AthlonMP 1800+	1,533	0.458
AMD AthlonMP 2000+	1,666	0.457

[8]Из за ограничений места в данной статье приводятся результаты только для теста ping-pong.
[9]Описание пакета SSP: `http://scali.com/products/ssp.html`.

5. Оценка эффективности MPI-обменов между вычислительными узлами по SCI-сети

К сожалению, мы не могли сравнить между собою по эффективности MPI-обменов по SCI-сети конфигурации на основе Intel и AMD: у нас не было вычислительных узлов с процессорами Intel, пригодных для установки адаптеров Dolphin SCI PCI-64/66 D330. Однако, по словам специалистов ОАО «НИЦЭВТ» (обладающих обширной базой данных по тестированию SCI-сетей), показатели, полученные для рассматриваемых здесь вычислительных узлов с процессорами AMD, являются на сегодняшний день самыми высокими:

- максимальная пропускная способность, показанная на тестах, составила: ping-ping — 276 МБ/с, ping-pong — 234 МБ/с (Рис. 4), send-receive — 258 МБ/с.
- задержка при передаче коротких пакетов (от 0 до 32 байтов) составила: ping-ping — 3.6-4.6 мкс, ping-pong — 3.5–4.6 мкс (Рис. 5), send-receive — 2.5–3.3 мкс.

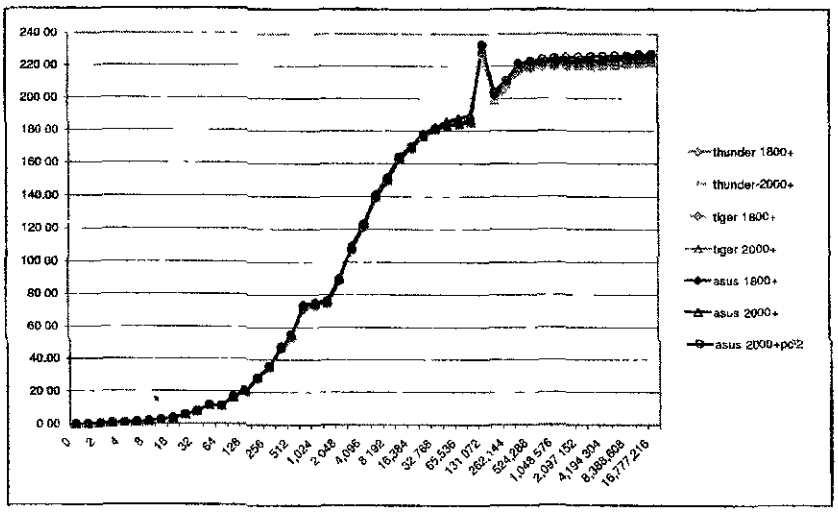
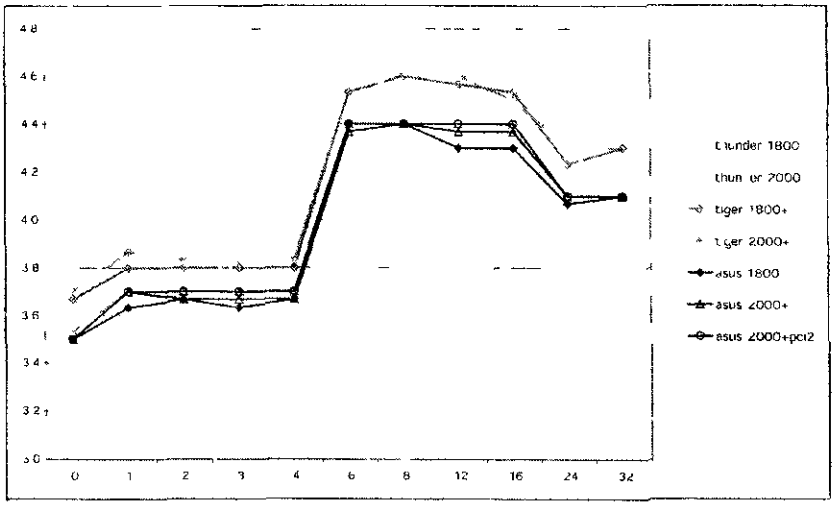


Рис. 4: Тест ping-pong, пропускная способность SCI-сети при различных размерах пакета

6. Выводы

В результате проведенных экспериментов мы пришли к следующим выводам:

1. Процессоры AMD AthlonMP следует признать перспективной и эффективной платформой для построения кластерных систем. Привлекательными являются и высокие технические характеристики такого решения, и отношение стоимости к производительности.
2. Из шести исследованных нами конфигураций по совокупности всех тестов наиболее удачной является конфигурация с процессорами AMD Athlon MP 1800+ и с системной платой ASUS A7M266-D. Нами принято решение в ближайшем будущем использовать данную конфигурацию, как базовую для изготовления моделей семейства «СКИФ».
3. Системные платы TYAN Thunder K7X и TYAN Tiger MPX несколько уступают плате ASUS по исследуемым показателям.



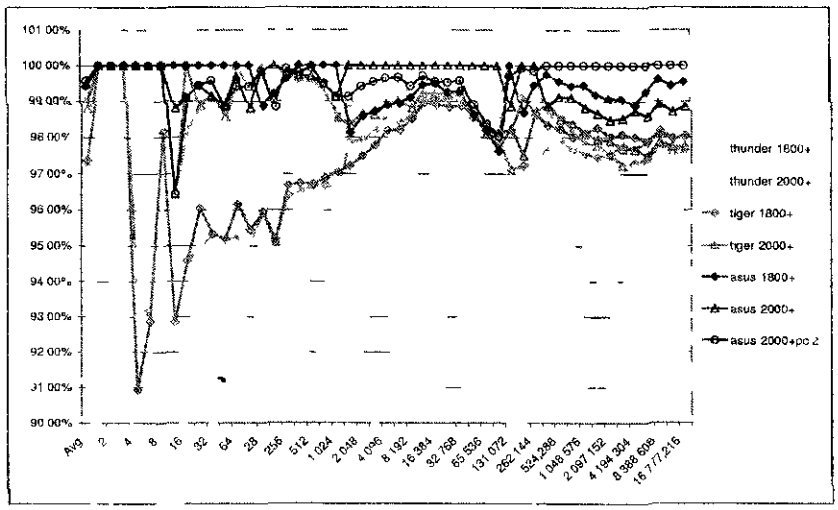


Рис. 6. Выбор конфигурации, оптимальной по пропускной способности SCI-сети на тесте ping pong

Благодарности. Авторы благодарят Представительство AMD в Москве, фирму «Традиция» и ОАО «НИЦЭВТ», которые предоставили все необходимые для исследования аппаратные средства и тем самым, оказали огромную помощь в подготовке данной статьи. Работа частично поддержана грантом РФФИ 02-01-81024-Бел2002а