СТРУКТУРНО-ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ КЛАСТЕРА СПИИРАН

М.Ю. Петров, Е.Л. Евневич, Е.В. Белаш, О.М. Селянин

Санкт-Петербургский институт информатики и автоматизации РАН

СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178

<eva@iias.spb.su>

УДК 681.3

Петров М.Ю., Евневич Е.Л., Белаш Е.В., Селянин О.М. **Структурно-функциональный анализ кластера СПИИРАН** // Труды СПИИРАН. Вып. 6. — СПб.: Наука, 2008.

Аннотация. *В данной статье представлен анализ структурного состава и функциональных возможностей кластера СПИИРАН. Приведено описание тестирования производительности кластера и проведено сравнение по мощности с другими кластерными системами. Сформулированы рекомендации по дальнейшему использованию кластера.*— Библ. 6 назв.

UDC 681.3

Petrov M.Yu., Evnevich E.L., Belash E.V., Selyanin O.M. **Analysis of SPIIRAS cluster structure and functional aspects.** // SPIIRAS Proceedings. Issue 6. — SPb.: Nauka, 2008.

Abstract. *Structure and function analysis of SPIIRAS cluster is presented in the paper. Cluster performance testing is described. Comparison in performance is carried out to some other cluster systems. Suggestions are made for further cluster use.* — Bibl. 6 items.

1. Введение

Анализ характеристик и возможностей кластера может осуществляться на разных этапах его жизненного цикла [1]. Для формализации процесса анализа вводится ряд категорий свойств системы, например: категория структурных свойств, функциональных свойств, эксплуатационных свойств, свойств безопасности, экономических свойств. К категории структурных свойств относятся следующие количественные характеристики: число узлов в системе, число процессоров, пропускная способность каналов связи, объем памяти для хранения данных и объем оперативной памяти [1]. С точки зрения этой классификации кластер СПИИРАН [2] представляет собой гетерогенную вычислительную систему, состоящую из 12 модулей. На данный момент в его состав входят 38 процессоров, 7 двухпроцессорных счетных модулей (узлов) с двухъядерными PIV (Pentium IV) (четыре процессорных ядра на один модуль) с сетевой загрузкой, 4 двухпроцессорных счетных модуля (узла) PIV с сетевой загрузкой, управляющая двухпроцессорная ЭВМ (PIV), дисковый массив (SCSI, RAID-0, RAID-1). Суммарный объём оперативной памяти составляет 19 Гб. Используется коммуникационная транспортная сеть Gigabit Ethernet и операционная система: Scientific Linux 4.4 [3,4]. Для параллельных приложений используются библиотеки и пакеты LAM MPI версии 7.0.6 и MPICH версии 1.2.7. Производительность на тестовом пакете прикладных программ линейной алгебры linpack достигает 18 Gflop/s [5].

2. Основные характеристики модулей кластера

Основные характеристики модулей кластера приведены в таб. 1. В одну вычислительную систему объединены узлы с процессорами разных поколений и с памятью, различающейся по быстродействию. Очевидно, что включение в

вычислительный процесс модулей различных поколений может существенно снизить производительность системы в целом, например, для матричных вычислений. Для всех модулей используется одна операционная система одной архитектуры (i386), что в некоторой степени снижает вычислительные возможности модулей Xeon 5130, так как новыми процессорами поддерживаются 64-битные вычисления. Ещё одно отличие процессоров Xeon 5130 – это их архитектура. На более низких частотах процессор обладает более высокой надежностью.

Функционирование кластера осуществляется следующим образом: управляющей является машина с дисковым массивом. Все остальные модули являются вычислительными (ВМ) и загружаются по сети через управляющую машину и коммутатор. Головная машина также распределяет вычислительные задачи по модулям.

Таблица 1

Характеристики модулей кластера

№	Количество модулей	Тип	Тип процессора	Количество процессоров	HT*	RAM	HDD
1	1	Упр.	Intel Pentium 4 Xeon 2.4 Ghz	2	да	1 Гб	~ 360 Гб
2	2	Выч.	Intel Pentium 4 Xeon 2.8 Ghz	2	да	1 Гб	Нет
3	2	Выч.	Intel Pentium 4 Xeon 3.0 Ghz	2	да	1 Гб	Нет
4	7	Выч.	Intel Xeon 5130 2.0 Ghz	4	нет	2 Гб	Нет

*HT – поддержка технологии Hyper Threading

Дисковый массив управляющей машины включает следующие компоненты:

1) системный массив – два диска по 70 Гб, объединённые в RAID-1;

2) массив для данных – четыре диска по 160 Гб, объединённые в RAID-10.

Диски «разбиты» на разделы с использованием технологии LVM2 (Logical Volume Manager). При разбиении остальной части жёсткого диска использовалась та же технология LVM, которая увеличивает гибкость файловой системы, но, являясь просто промежуточным слоем, не отменяет ограничения и использования других слоёв и усложняет работу. То есть по-прежнему нужно создавать и изменять разделы и осуществлять форматирование. Изменение размера должно поддерживаться также и самой файловой системой.

Системный массив разбит следующим образом: /boot — данные загрузчика и ядро ОС, и занимает 100 Мб.

Создана группа логических томов SysVolGroup, занимающая всё оставшееся пространство на системном массиве и включающая следующие логические тома:

1) MainRoot — / — Корневая файловая система объемом 10 Гб,

2) SysSwap — файл подкачки, равный 2*RAM, в данном случае – 2 Гб,

3) DisklessRoot — /diskless/root — корневая файловая система – 10 Гб,

4) SpecProgs — /usr/local — спецпрограммы, общие для всех узлов системы и занимающие 760 Мб,

5) DisklessSnap — /diskless/snapshot — необходимые индивидуальные файлы для ВМ – 5 Гб,

6) DisklessSwap — /diskless/swap — раздел для swap-файлов ВМ – 25 Гб. В резерве остается приблизительно 15 Гб.

Массив данных разбит следующим образом: создана группа логических томов DataVolGroup, занимающая всё пространство. Она включает единственный том UserData — /home для данных пользователей объемом приблизительно 217 Гб. В резерве остается около 53 Гб неразмеченного пространства.

3. Тестирование кластера

Тестирование проводилось в следующем порядке: тестирование «новых» модулей на параллельной версии linpack, отдельное тестирование модулей с предыдущим поколением процессоров, этап измерения производительности системы из «новых» и «старых» модулей. Полученные изменения в производительности отражены на графиках (см. рис. 1,2,3,4).

При тестировании модулей с процессорами Xeon 5130 размерность матрицы для тестовых запусков варьируется от 500 до 15000 и верхняя граница определяется объёмом доступной оперативной памяти. Несмотря на то, что на 28 процессорах (процессорные ядра) есть возможность использовать тестовый набор с квадратной матрицей размерностью 20000, на одном процессоре запуск задачи с матрицей 16000 оказался невозможным из-за недостатка памяти. График тестов приведен на рис. 1.

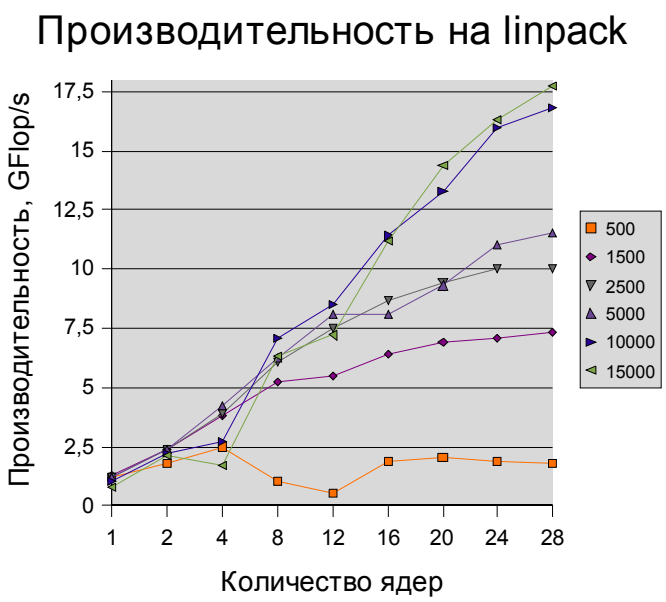

Рис. 1. Результаты теста 1.

В следующем тесте (рис.2) вместо последних восьми ядер Xeon 5130 использовались более ранние модели (P4 Xeon с 3,0 и 2,8 Ghz соответственно): На графике хорошо видно, как падает производительность системы на матрицах больших размерностей, а на матрице малого размера присутствует небольшой прирост производительности на 24 ядрах по сравнению с запуском теста на 8, 16 и 20 ядрах.

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Следующий тест (рис.3) проводился на 20 ядрах Xeon 5130 и на двух процессорах Xeon 3.0 Ghz с использованием технологии Hyper Threading (HT):

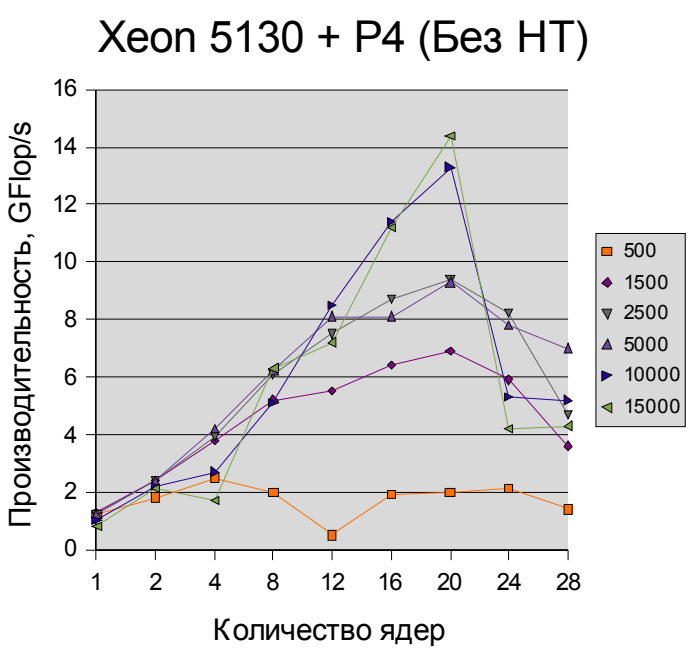

Рис.2. Результаты теста 2.

По графику видно, как сильно падает суммарная производительность системы. Несмотря на то, что для некоторых задач технология Hyper Threading дает существенный прирост производительности, на данном тесте ситуация ухудшилась, что связано, по-видимому, с ограничением на размер памяти.

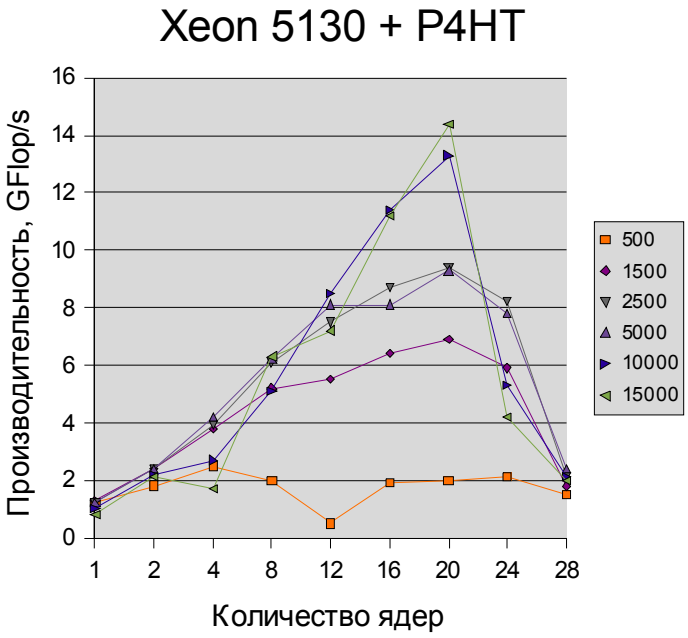

Рис. 3. Результаты теста 3.

Следующий график – результаты отдельного тестирования производительности модулей на базе Intel Xeon P4 (рис. 4). Очень хорошо видно, что мо-

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

дули на базе Intel Xeon P4 едва пересекают отметку в 3 Gflop/s, в этом случае, вероятно, играет некоторую роль разброс в тактовых частотах модулей.

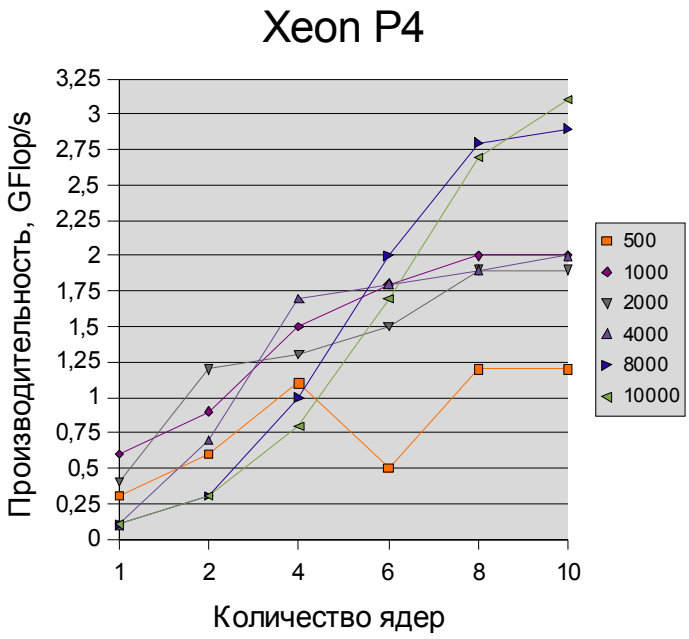

Рис. 4. Результаты теста 4.

4. Сервисы пользователей

В настоящий момент подавляющее большинство пользователей, имеющих задачи с большим объемом вычислений, не занимаются написанием собственного программного обеспечения, а адаптируют и используют готовые пакеты прикладных программ. Поэтому была выбрана такая конфигурация кластера, которая позволяет гибкое реконфигурирование и использование для решения прикладных задач как всей системы в целом, так и отдельных ее частей. В частности, предоставляется возможность использовать различные реализации параллельных библиотек (MPICH, LAM MPI). Пользователи имеют возможность работать с уже установленными пакетами и добавлять новые.

В 2005 году СПИИРАН на основе кластера выделил вычислительные и технологические ресурсы и включился в проект EGEE-RDIG. В рамках этого проекта проведен эксперимент по созданию сегмента сети EGEE с участием ПИЯФ (г. Гатчина), СПбГУ, ФТИ, ИВВиБД.

В настоящий момент ведутся работы по созданию веб-сервиса на базе портала СПбНЦ РАН для доступа к ресурсам кластера.

5. Причины низкой производительности кластера

Можно сформулировать следующие причины низкой производительности кластера СПИИРАН:

1) использование коммуникационной среды кластера для сервисных нужд;

2) неоптимальный учет поведения кластерных систем с гетерогенными вычислительными модулями;

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

3) отсутствие высокопроизводительных сетевых интерфейсов (SCI, Myrinet и т.д.);

4) использование версии библиотеки BLASC, не оптимизированной для данных моделей процессоров;

5) использование ОС с архитектурой, не полностью реализующей преимущества новых процессоров.

6. Заключение

Для рационального использования кластера СПИИРАН необходимо:

1) определить класс задач и области использования кластера для повышения производительности на данном классе;

2) разработать стратегию реконструкции и модернизации кластера с учетом результатов тестирования;

3) минимизировать размеры предполагаемых затрат на реконструкцию с учетом «узких мест» в структуре кластера;

4) в процессе дальнейшего развития кластера целесообразно учитывать требования, связанные с установкой и эксплуатацией программного обеспечения GRID-систем;

5) для ряда приложений можно рассматривать вопрос создания специальных кластерных систем под конкретные задачи.

Следует отметить, что в настоящее время работы по модернизации кластера уже проводятся. Планируется к середине 2008 года увеличить его производительность примерно в два раза.

Литература

1. *Шелестов А.Ю.* Структурно-функциональный анализ компонентов GRID- систем.// Проблемы управления и информатики. 2007. № 5.

2. *Петров М.Ю*. Автоматизация программирования на кластерных и ГРИД-системах // Материалы конференции «Региональная информатика – 2006» (СПб, 24 – 26 октября 2006 г.) – СПб.: 2006.

3. Руководство по системному администрированию Red Hat Enterprise Linux 4 (перевод) [Электронный ресурс] // <http://www.rhd.ru/docs/manuals/enterprise/RHEL-4-Manual/sysadmin-guide/> (по состоянию на 13.01.2008).

4. Руководство по системному администрированию Red Hat Enterprise Linux 4 (оригинал) [Электронный ресурс] // <http://www.redhat.com//docs/manuals/enterprise/RHEL-4-Manual/sysadmin-guide/> (по состоянию на 13.01.2008).

5. Практическое руководство по параллельным вычислениям. [Электронный ресурс] // <http://linux-cluster.org.ru> (по состоянию на 13.01.2008).

6. Кластерные установки России и СНГ. [Электронный ресурс] // <http://parallel.ru> (по состоянию на 13.01.2008).