Информационные системы и технологии

УДК 004.052.3

Г.Д. Киселев, канд. техн. наук, В.И. Пушкар, Е.В. Оленович, О.Г. Киевський

Оценка производительности вычислительного кластера в подразделении ВУЗа

Создан учебный кластер на серверах малой производительности. Проведено его тестирование и оценка суммарной производительности. Результаты тестирования показали возможность его применения в учебном процессе для изучения технологий параллельного программирования и подготовки системных администраторов кластеров.

An educational cluster based on the low speed servers was created. Their testing and total performance evaluations were carried out. The testing results describe the possibility of cluster implementation for educational process in the meaning of parallel programming studying and clusters system administrators training.

Введение

Бурное развитие информационных технологий, рост количества обрабатываемых и передаваемых данных, рост сложности самих вычислений требуют внедрения кластерных технологий в учебный процесс ВУЗа. Все большее количество университетов становятся владельцами высокопроизводительных суперкомпьютеров, использующих кластерные принципы работы. Основной целью покупки и использования кластера вузом является выполнение научных исследований, требующих значительных вычислительных ресурсов. В то же время в учебном процессе необходимо решать задачи обучения технологиям параллелизации вычислений и системного администрирования кластеров. При огромных возможностях кластера, который обслуживает университет, загружать его достаточно мелкими задачами учебного процесса не имеет смысла. Выходом из этой ситуации является создание специальных учебных кластеров. В настоящей статье предлагается решение, в котором кластер собирается из серверов снятых с коммерческой эксплуатации. При этом, встает задача оценки производительности полученного решения, по результатам которой можно судить:

- использовать кластер только для обучения системных администраторов или на кластере можно ставить циклы достаточно сложных лабораторных работ по специальным дисциплинам, в которых изучаются технологии параллельных вычислений.
- какие из научных проектов, выполняемых на кафедре, можно «вынести» на построенный кластер и есть ли в этом смысл.

Описание и тестирование кластера

Архитектура кластера представлена на рисунке 1. Кластер состоит из 8 вычислительных узлов, сервера кластера и коммутатора Cisco 2924xl.

Вычислительный узел (нод), имеет следующие характеристики:

- CPU: Intel Celeron (601.37-МГц 686-class CPU);
- материнская плата Supermicro 370SSR/370SSE;
- SDRAM 256 Мб 100 МГц 64 бит;
- FastEthernet = 100Мбит/с.

Характеристика сервера кластера:

- CPU: Intel Celeron (601.37-МГц 686-class CPU);
- материнская плата Supermicro 370SSR/370SSE;
- SDRAM 256 Мб 100 МГц 64 бит;
- FastEthernet = 100Мбит/с;
- Жесткий диск 20Гб;
- ОС FreeBSD 7.2-STABLE.

На физическом уровне узлы кластера связаны сетью типа звезда, в которой узлы (ноды) подключены к центральному коммутатору линиями связи FastEthernet пропускной способностью 100Мбит/с. Операционная система на вычислительных узлах подгружается с сервера кластера.

Выбранная архитектура кластера имеет следующие преимущества:

- При выделении отдельного сервера, нет необходимости использовать внешние накопители информации в узлах кластера, что упрощает администрирование и обслуживание.
- Создание на сервере единой файловой структуры позволяет организовать вычисления, которые совместимы и переносимы

на существующие суперкомпьютеры (классические кластеры).

– Допускается изменение параметров узлов кластера.

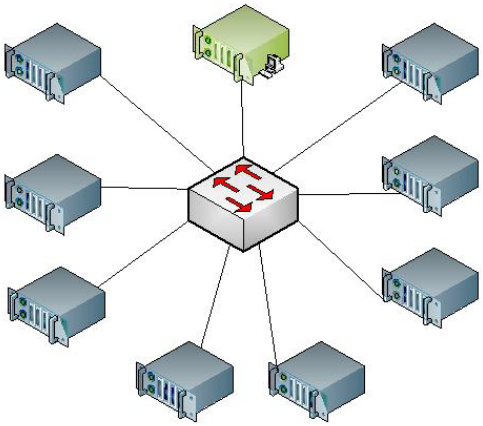

Рис. 1. Архитектура кластера

Недостатком данного решения является увеличение времени выполнения дисковых операций, таких как, считывание и запись, но это не главное для учебного кластера.

Программное обеспечение кластера может создаваться с использованием программных пакетов PVM (Parallel Virtual Machine**)** или MPI (The Message Passing Interface), обеспечивающих параллельное выполнение программ на нескольких узлах кластера. Пакет PVM – предназначен больше для гетерогенных аппаратных платформ, поэтому его применение для однородной – нецелесообразно. Интерфейс обмена данными MPI не кодирует данные для гетерогенных сетей, следовательно, в однородной среде имеет большую производительность. Поэтому, выбор программного обеспечения сделан в пользу пакета MPI и рассматривались только бесплатные его реализации: MPICH2 и OpenMPI. При использовании пакета OpenMPI нужен компилятор с языкаC++, например gcc42. При этом, не все C++ компиляторы поддерживают OpenMPI. Пакет MPICH2 не нуждается во внешнем C++ компиляторе, поскольку в нем используются собственные компиляторы: mpicc, mpif77 (Fortran). В результате, на все вычислительные узлы установлена версия MPI – MPICH2, которая на сегодня является самой популярной реализацией интерфейса MPI.

Тестирование характеристик производительности кластера выполнялось тестирующими программами Flops [1], Transfer, Nettest, Mpitest [2]. Для вычисления производительности, использовали программу Flops, для определения латентности и скорости пересылок между двумя узлами – программу Transfer, для определения пропускной способности сети при сложных обменах по различным логическим топологиям – программу Nettest и для определения эффективности основных операций MPI – Mpitest.

Программа Flops вычисляет производительность кластера в Mflops для заданного количества узлов (нодов). Для этого, тестер решает задачу с известным количеством операций и подсчитывает время, за которое она была решена. Flops (**Fl**oating point **O**perations **P**er **S**econd) — величина, используемая для измерения производительности компьютеров, показывающая, сколько операций с плавающей запятой в секунду выполняет данная вычислительная система. По результатам тестирования (таб. 1, рис.2) видно, что увеличение количества узлов приводит к пропорциональному увеличению производительности кластера. Это связано с тем, что с ростом количества узлов растут и накладные расходы на организацию параллельных вычислений, в частности, пропускная способность коммуникационной среды.

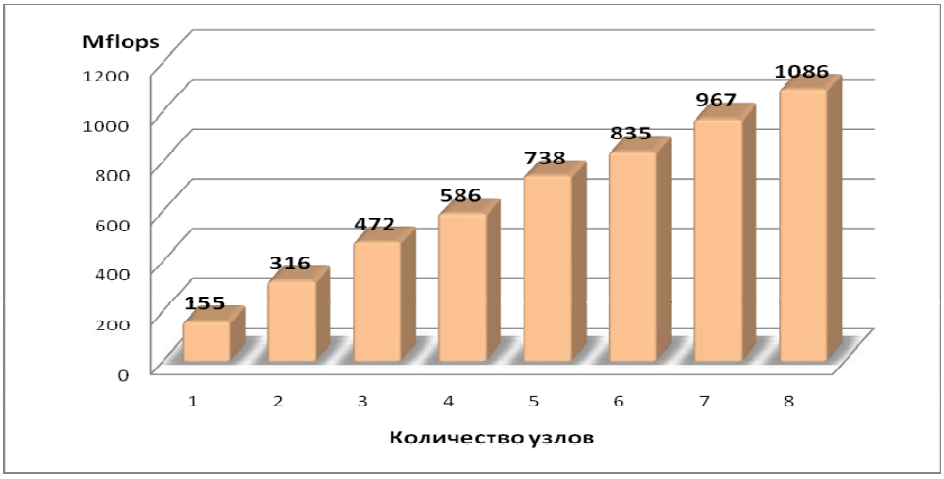

Рис. 2. Производительность кластера при разном количестве его узлов

Таблица 1. Производительность кластера для количества узлов от 1 до 8.

	Производительность			
К-во узлов	**1**	**2**	**3**	**4**
Результат	Calculation time 11.60 seconds Cluster speed 155 MFLOPS node N00 speed 155 MFLOPS	Calculation time 5.68 seconds Cluster speed 316 MFLOPS node N00 speed 158 MFLOPS node N01 speed 162 MFLOPS	Calculation time 3.81 seconds Cluster speed 472 MFLOPS node N00 speed 157 MFLOPS node N01 speed 160 MFLOPS node N02 speed 157 MFLOPS	Calculation time 3.07 seconds Cluster speed 586 MFLOPS node N00 speed 146 MFLOPS node N01 speed 157 MFLOPS node N02 speed 146 MFLOPS node N03 speed 146 MFLOPS
К-во узлов	**5**	**6**	**7**	**8**
Результат	Calculation time 2.44 seconds Cluster speed 738 MFLOPS node N00 speed 147 MFLOPS node N01 speed 154 MFLOPS node N02 speed 147 MFLOPS node N03 speed 147 MFLOPS node N04 speed 152 MFLOPS	Calculation time 2.15 seconds Cluster speed 835 MFLOPS node N00 speed 139 MFLOPS node N01 speed 139 MFLOPS node N02 speed 144 MFLOPS node N03 speed 144 MFLOPS node N04 speed 143 MFLOPS node N05 speed 143 MFLOPS	Calculation time 1.86 seconds Cluster speed 967 MFLOPS node N00 speed 138 MFLOPS node N01 speed 146 MFLOPS node N02 speed 140 MFLOPS node N03 speed 140 MFLOPS node N04 speed 138 MFLOPS node N05 speed 144 MFLOPS node N06 speed 152 MFLOPS	Calculation time 1.66 seconds Cluster speed 1086 MFLOPS node N00 speed 135 MFLOPS node N01 speed 142 MFLOPS node N02 speed 136 MFLOPS node N03 speed 136 MFLOPS node N04 speed 136 MFLOPS node N05 speed 136 MFLOPS node N06 speed 137 MFLOPS node N07 speed 138 MFLOPS

Основными характеристиками быстродействия сети являются ее латентность (latency) и пропускная способность (bandwidth). Под пропускной способностью сети понимают количество информации, передаваемой между узлами сети в единицу времени (байт в секунду). Латентностью (задержкой) называется время, затрачиваемое программным обеспечением и коммутационными устройствами сети на подготовку к передаче информации по заданному каналу. Полная латентность складывается из программной и аппаратной составляющих.

Значения пропускной способности обозначают в мегабайтах в секунду (Мб/с), значения латентности – в микросекундах (Uc = 10^{-6} с).

Для приложений с тонкой параллельной структурой (fine-grained parallelism), какими, как правило, являются вычислительные программы, крайне важны малые величины латентности; тогда как для приложений, использующих большие объемы пересылаемых данных (а это, как правило, коммерческие приложения баз данных), более важно максимальное увеличение пропускной способности.

Для измерения пиковых характеристики латентности и пропускной способности при обмене данными между двумя узлами кластера применен тест Transfer.

При измерении пропускной способности в режиме "точка-точка" используется следующая методика. Процесс с номером 0 посылает процессу с номером 1 сообщение длины L байт. Процесс 1, приняв сообщение от процесса 0, посылает ему ответное сообщение той же дли-

ны. Для этого используются блокирующие (blocking) вызовы MPI (MPI_Send, MPI_Recv). Эти действия повторяются N раз с целью минимизировать погрешность за счет усреднения результатов. Процесс с номером 0 измеряет время T, затраченное на все эти обмены. Пропускная способность R определяется по формуле R=2NL/T. Пропускная способность двунаправленных обменов определяется по той же формуле. В этом случае используются неблокирующие (non-blocking) вызовы MPI (MPI_Isend, MPI_Irecv). При этом, производится измерение времени, затраченного процессом 0 на передачу сообщения процессу 1 и прием ответа от него, при условии, что процессы начинают передачу информации одновременно после синхронизации.

Латентность измеряется как время, необходимое на передачу сигнала или сообщения нулевой длины. При этом, для снижения влияния погрешности и низкого разрешения системного таймера, важно повторить операцию посылки сигнала и получения ответа N раз. Таким образом, если время для N итераций пересылки сообщений нулевой длины туда и обратно составило Тмкс., то латентность измеряется как S=T/(2N). Результаты тестирования латентности кластера при передаче сообщений от нулевого узла к оставшимся 7 приведены на рисунке 3.

Результаты тестирования пропускной способности кластера при однонаправленной передаче сообщений от 0 узла, показаны на рис.4. Результаты тестирования пропускной способности кластера при двунаправленной передаче сообщений от 0 узла, показаны на рис.5.

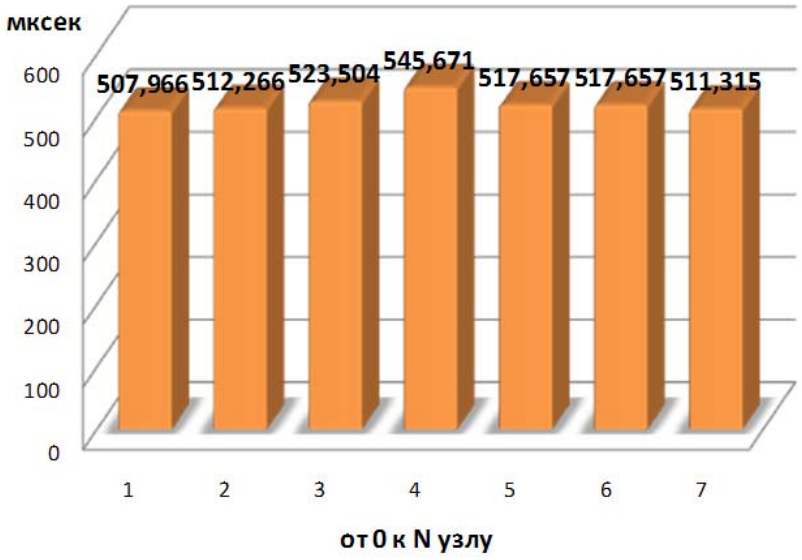

Рис. 3. Задержки передачи сообщений от нулевого узла кластера к остальным

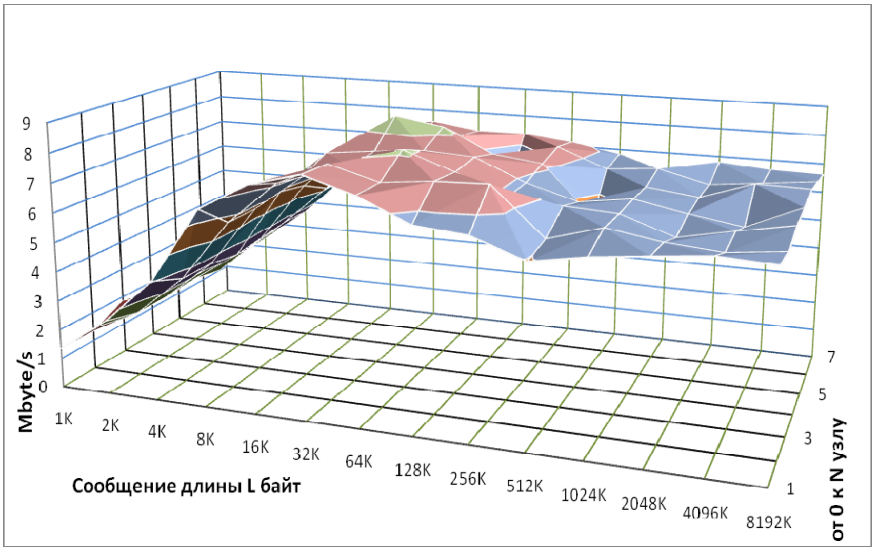

Рис. 4. Пропускная способность при однонаправленной передаче сообщений

Тестирование показало, что разброс латентности при передаче сообщений между узлами кластера меньше 7%. В тесте на пропускную способность, максимальная эффективность 8,2Мб/с достигается при двунаправленной передаче данных и длине сообщений 32Кб. При длине сообщений выше 32Кб наблюдается уменьшение пропускной способности, что связано с насыщением каналов передачи данных, переполнением буферов, коллизиями и т.д.

Тестирование пропускной способности сети или тестирование коммуникационной производительности выполнялось с помощью программы Nettest. Тест используется для проверки "выживаемости" сетевого оборудования при пиковых нагрузках и для определения пиковой пропускной способности коммутатора. При тестировании рассматривались три логические топологии взаимодействия процессов:

- Звезда (Star), т.е. взаимодействие основного процесса с подчиненными;
- Кольцо (Ring), т.е. взаимодействие каждого процесса со следующим;
- Полный граф (Chaos), т.е. взаимодействие каждого процесса с каждым.

Логическая топология определяется используемыми логическими каналами. Под логическим каналом понимается неупорядоченная пара узлов кластера (A,B), которые в данной топологии обмениваются сообщениями. Для каждой топологии рассматриваются однонаправленные и двунаправленные обмены сообщениями. В ходе теста, по каждому логическому каналу от узла A к узлу B и обратно передается по L байт информации. Однако, в случае однонаправленных обменов, один из узлов, например B, ждет получения сообщения от A, и только тогда может передавать A ответное сообщение. В случае двунаправленных обменов, информация может передаваться в обе стороны параллельно.

Конкретные способы организации пересылок средствами MPI (блокирующие, не блокирующие, и т.д.) не регламентируются. Предполагается, что из всех, соответствующих данной топологии и способов обменов, будет выбран вариант с наименьшими накладными расходами. Результаты тестирования приведены на рисунке 6. Результаты тестирования однонаправленных сообщений для соответствующих топологий помечены метками Star, Ring, Chaos. Метки Star2, Ring2, Chaos2 выделяют графики двунаправленных режимов передачи сообщений.

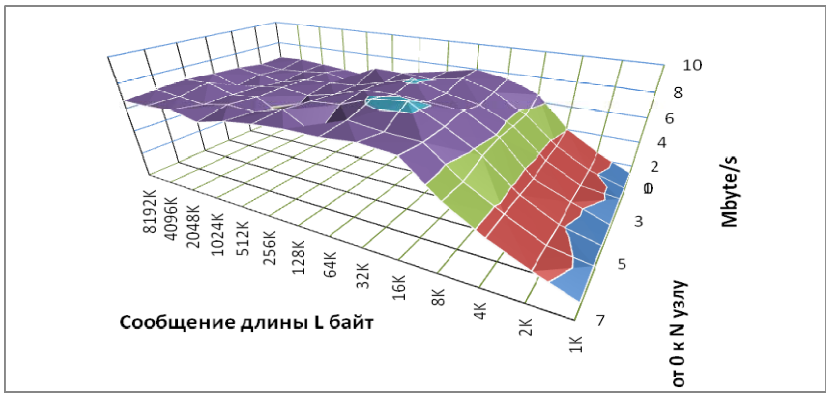

Рис. 5. Пропускная способность кластера при двунаправленной передаче сообщений

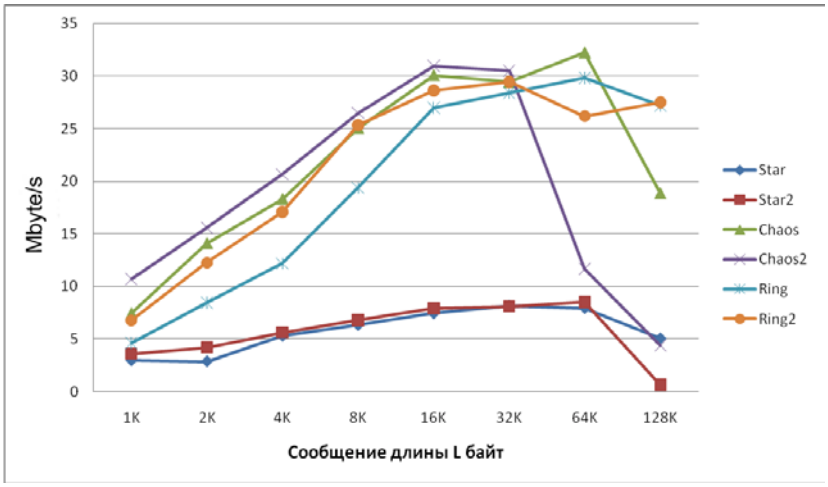

Рис. 6.Пропускная способность сети в логических топологиях "звезда", "полный граф" и "кольцо"

Из рис.6 видно, что наилучшая пропускная способность сети из 8 узлов достигается на топологии "хаос" при двунаправленных обменах сообщениями и составляет 30.97 Мбайт/сек. У логической топологии «звезда» самая низкая пропускная способность, что объясняется загрузкой канала к главному узлу, в остальных топологиях данной проблемы нету. Самой стабильной топологией, как видно, является «кольцо».

MPI - это библиотека функций передачи сообщений, облегчающих обмен данными и синхронизацию задач между процессами параллельной программы с распределенной памятью. Определение эффективности выполнения основных операций MPI выполнено с помощью теста Mpitest. В методике измерений данного теста использованны понятия «быстрых» и «медленных» операций. Под «быстрыми» понимаются операции синхронизации процессов в узлах кластера, определение латентности сети и замер текущего времени. К «медленным» относят вычислительные операции, операции передачи данных и рассылки сообщений между узлами кластера. Эффективность выполнения основных операций MPI отображается сумой времени выполнения отдельных операций и времени операций синхронизаций.

При тестировании, программа mpitest была запущена на 8 процессорах. Тест оперировал массивами данных размером в 1, 10 и 100 целых чисел. Для "продолжительных" операций тестовая процедура повторилась 100 раз, для "коротких" - 100 тыс. раз. Все измеренные времена выполнения операций не превышали 10^{-6} с, что достаточно для учебного кластера.

Выводы

С целью определения основных параметров собранного учебного кластера выполнена серия тестов. По результатам тестирования видно, что кластер пригоден к эксплуатации и способен решать достаточно сложные прикладные задачи. Для сравнения производительности созданного кластера, возьмем процессор C2D E4600 2.40ГГц с памятью ОЗУ 2048Мб, его производительность для той же задачи составляет: для 1 ядра – 578 мегафлоп/с и для 2 ядер - 1118 мегафлоп/с. Таким образом, собранный кластер по производительности сравним с современным двуядерным процессором среднего класса. В частности, на кластере эффективно решаются задачи параллельного программирования в среде MPI, кроме того, студенты изучают основы системного администрирования кластеров.

Литература

1. *http://parallel.ru/computers/benchmarks/perf.html*-Тесты производительности процессора
2. *http://www.intuit.ru/department/supercomputing/tbucs/4/4.html*-Оценка производительности кластерных систем
3. *http://www.ccas.ru/mmes/educat/lab04k/01/basics.html#Sec1.1* – Основы программирования в Message Passing Interface (MPI). Вычислительный центр им. А.А. Дородницына РАН
4. *Богачёв К.Ю*. Основы параллельного программирования. - М.: БИНОМ. Лаборатория знаний, 2003. - 342 с.