С.И. Мальковский, В.В. Пересветов

Вычислительный центр Дальневосточного отделения РАН, г. Хабаровск

ГЕТЕРОГЕННЫЙ ВЫЧИСЛИТЕЛЬНЫЙ КЛАСТЕР ВЦ ДВО РАН

В докладе описывается гетерогенный вычислительный кластер ВЦ ДВО РАН, состоящий из трех разнородных групп вычислительных узлов. Рассматриваются полученные экспериментальные результаты исследования производительности сегментов этой вычислительной системы.

Гетерогенный вычислительный кластер отличается от однородного вычислительного кластера тем, что в его состав входят вычислительные узлы разных типов. Целесообразность создания гетерогенных вычислительных кластеров обусловливается приобретением в разные периоды времени различного вычислительного оборудования, которое необходимо эффективно использовать. В таком кластере можно выделять обладающие наиболее подходящими ресурсами узлы для решения определенных классов задач. При этом отсутствует необходимость поддержки нескольких специализированных вычислительных кластерных систем, так как все группы узлов в гетерогенном кластере обслуживаются одним управляющим узлом.

Для эффективного использования гетерогенной кластерной системы необходимо знать производительность всех сегментов кластера при решении основных типов задач. Также необходимо учитывать особенности работы на разнородных узлах установленного системного программного обеспечения (ПО) и различных параллельных технологий. Поэтому нами были проведены экспериментальные исследования производительности всех сегментов гетерогенного кластера, результаты которых приведены в работах [1–4]. Результаты тестирования могут быть также сопоставлены с данными [5], где экспериментально исследовалась производительность вычислительного кластера, построенного с использованием неспециализированного оборудования. В этой работе изложена также методика тестирования,

в которой проводится исследование производительности вычислительного оборудования на основе усреднения результатов многократных запусков тестовых задач с нахождением среднеквадратических отклонений.

Все узлы описываемого гетерогенного вычислительного кластера работают под управлением операционной системы Linux CentOS. Для диспетчеризации заданий используется PBS Torque с планировщиком Maui. Мониторинг осуществляется пакетом Ganglia. Установлено ПО Intel Cluster Toolkit. В качестве коммуникационной среды параллельных вычислений используется библиотека MPI в реализации Intel. Также на кластере доступна библиотека Intel MKL, компиляторы Intel C/C++/Fortran и GNU C/C++/Fortran.

Гетерогенный вычислительный кластер ВЦ ДВО РАН в настоящее время объединяет с помощью коммуникационной сети Gigabit Ethernet три типа узлов: Sun Blade X6440 Server, Sun Blade X6250 Server и HP ProLiant DL360 G5.

К первому типу относятся два сервера Sun Blade X6440 Server, каждый из которых оснащен четырьмя шестиядерными процессорами AMD Opteron Istanbul 8431 (2,4 GHz) и 96 GB оперативной памяти. Эти узлы могут использоваться по отдельности в качестве мультипроцессорных (24 вычислительных ядра) систем с общей памятью. Производительность одного такого сервера в тесте Linpack составила 182 GFlops или 79 % от пиковой. Экспериментальные исследования производительности узлов Sun Blade X6440 Server показали, что на каждом из них можно эффективно решать без взаимодействия с остальными узлами задачи широкого спектра в пределах до 24 процессов MPI или нитей OpenMP [1–2]. Хорошую масштабируемость на них показывают даже те задачи, эффективность решения которых напрямую зависит от скорости обмена данными между вычислительными ядрами.

На данных узлах компиляторы Intel icc и ifort превосходят компиляторы GNU лишь в отдельных случаях некоторых вычислительно-затратных программ с небольшим объемом меж-

процессорных взаимодействий. В остальных случаях различие в производительности программ с использованием протестированных компиляторов (Intel и GNU) несущественно. Применение технологии OpenMP в большинстве случаев приводит к получению производительности на уровне MPI. Более высокую эффективность технология OpenMP показывает лишь тогда, когда сильно загружается среда передачи данных.

В дополнение к стандартным тестам для узлов Sun Blade X6440 было проведено в [2] исследование зависимости производительности системы от привязки процессов к различным процессорным ядрам. Оно показало, что в случае когда процессы совместно работают с большими объемами данных, не помещающимися в кэш-память, наиболее верным будет разнесение их на разные физические процессоры. В случае передачи сообщений небольшого объема максимальная производительность будет достигаться при их запуске на одном физическом процессоре. В коллективных операциях MPI наблюдается большая зависимость производительности системы передачи данных от способа привязки процессов к различным процессорным ядрам, чем в двухточечных. Подробно результаты исследования производительности этого типа узлов описаны в [1–2].

Ко второму типу вычислительных узлов относятся пять серверов Sun Blade X6250 Server. Каждый из них оснащен двумя четырехъядерными процессорами Intel Xeon E5450 (3 GHz) и 16 GB оперативной памяти. Пиковая производительность одного такого узла составляет 80 GFlops. В тесте Linpack для пяти узлов достигнут уровень производительности 305 GFlops.

При тестировании этой группы узлов в качестве сети передачи данных использовалась сеть, основанная на технологии Gigabit Ethernet. Такие исследования позволили оценить производительность данной коммуникационной среды в современных вычислительных кластерах. Они показали, что для достигнутого к настоящему времени уровня производительности процессоров, эффективности кэш-пямяти и скорости обменов с оперативной

памятью коммуникационная сеть Gigabit Ethernet должна быть признана устаревшей для использования в таких системах [3]. Только в ограниченном числе задач она может быть использована без существенной потери производительности. В случае использования технологии Gigabit Ethernet при программировании нужно избегать коллективных операций, осуществлять пересылки данных равномернее, отдавать предпочтение асинхронным двухточечным операциям. Использование компиляторов Intel icc и ifort на узлах этого типа приводит иногда к очень большому отрыву от других, и они в целом являются более эффективными. Компиляторы GNU иногда незначительно превосходят другие компиляторы. Компиляторы LLVM для C/C++ и Fortran показали эффективность на уровне компиляторов GNU. В пределах одного узла технология OpenMP показывает, как правило, производительность выше, чем MPI. Подробно результаты исследования производительности второго типа узлов описаны в [3].

Третий тип узлов представлен восемью серверами HP ProLiant DL360 G5, построенными на базе двух двухъядерных процессоров Intel Xeon 5060 (3,2 GHz). Каждый из таких узлов оснащен 4 GB оперативной памяти и имеет пиковую производительность, равную 25,6 GFlops. Подробно результаты исследования производительности этого типа узлов описаны в [4].

В табл. 1 представлены результаты теста IMB для трех групп узлов гетерогенного вычислительного кластера с различным числом процессов n. Размер пересылаемых сообщений 8 байт. Эти результаты показывают, что задержки при передаче данных минимальны при запуске взаимодействующих процессов в пределах одного узла вычислительного кластера. Если же в качестве коммуникационной среды используется сеть Gigabit Ethernet – задержки сильно возрастают. Эти и нижеприведенные результаты были получены на наборе тестов NPB и IMB, собранных при помощи компилятора Intel версии 10.1. В качестве библиотеки MPI использовалась Intel MPI.

Таблица 1

Результаты теста IMB

Тип сообщений	Sun Blade X6440		Sun Blade X6250		HP ProLiant DL360	
	$n = 4$	$n = 24$	$n = 4$	$n = 40$	$n = 4$	$n = 16$
Sendrecv	1,5	2,0	1,1	14	1,6	63
Bcast	1,9	2,9	1,4	33	2,1	149
Allreduce	2,8	3,8	2,6	183	3,7	145

В табл. 2 представлены усредненные результаты теста NPB EP (класс C) для трех групп узлов, которые позволяют оценить максимальную производительность (в MFlops) в отсутствие заметных межпроцессорных взаимодействий. Значения производительности в расчете на одно вычислительное ядро (результаты делятся на число задействованных ядер процессоров) практически одинаковы для любого числа ядер процессоров.

Таблица 2

Производительность в тесте EP

Sun Blade X6440	Sun Blade X6250	HP ProLiant DL360 G5
31	51	23

В табл. 3 представлены аналогичные результаты в расчете на одно процессорное ядро для тестов NPB LU (класс C), в котором интенсивно используется коммуникационная среда для передачи сообщений. Взаимодействие параллельных процессов осуществляется большим числом синхронных передач сообщений MPI_Send небольшой длины. Среднеквадратические отклонения результатов лежат в диапазоне 0,3–4 %. Показатели разброса результатов возрастают с уменьшением класса сложности.

В табл. 4 представлены результаты теста NPB IS (класс C). В нем осуществляется параллельная сортировка большого массива целых чисел. Передача сообщений между процессами реализуется с помощью операций MPI_Alltoall и MPI_Allreduce. Из

всех тестов, входящих в состав NPB, он является самым требовательным к производительности среды передачи данных. Из результатов видно, что этот тест эффективно работает только тогда, когда все вычислительные процессы запущены в пределах одного узла кластера (Sun Blade X6440). В остальных случаях, когда для передачи данных используется Gigabit Ethernet, производительность падает до неприемлемо низкого уровня. Среднеквадратические отклонения экспериментальных результатов находились в пределах 1 % для класса сложности C.

Таблица 3

Производительность в тесте LU

Число процессов	Sun Blade X6440	Sun Blade X6250	HP ProLiant DL360 G5
4	1180	511	470
16	904	765	650

Полученные экспериментальные результаты показывают, что в вычислительных кластерах производительность процессоров, на которых построены узлы, является важным, но не единственным фактором, определяющим возможности вычислительной системы. Другим важным фактором является производительность коммуникационной среды. Приемлемым вариантом построения гетерогенных вычислительных кластеров при использовании устаревшей технологии Gigabit Ethernet является применение вычислительных узлов с большим числом ядер. На каждом из таких узлов можно эффективно решать без взаимодействия с остальными узлами задачи широкого спектра в пределах до 24 процессов MPI или нитей OpenMP, как в случае описанных здесь узлов Sun Blade X6440.

Таблица 4

Производительность в тесте IS

Число процессов	Sun Blade X6440	Sun Blade X6250	HP ProLiant DL360 G5
4	54	28	21
16	27	5,2	4,5

Во многих случаях технология MPI показала производительность на уровне технологии OpenMP. Применение компиляторов GNU обычно позволяет получить скорость вычислений на уровне компиляторов Intel, поэтому их можно рекомендовать для использования в научной и образовательной сферах деятельности.

Список литературы

1. Мальковский С.И., Пересветов В.В. Исследование производительности четырехпроцессорных узлов в составе вычислительного кластера // Суперкомпьютеры: вычислительные и информационные технологии: материалы междунар. науч.-практ. конф., Хабаровск, 30 июня – 2 июля 2010 г. – Хабаровск: Изд-во Тихоокеанского гос. ун-та, 2010. – С. 77–84.

2. Мальковский С.И., Пересветов В.В. Зависимость производительности передачи сообщений MPI от порядка распределения процессов в четырехпроцессорном узле вычислительного кластера // Суперкомпьютеры: вычислительные и информационные технологии: материалы междунар. науч.-практ. конф., Хабаровск, 30 июня – 2 июля 2010 г. – Хабаровск: Изд-во Тихоокеанского гос. ун-та, 2010. – С. 85–92.

3. Мальковский С.И., Пересветов В.В. Оценка производительности вычислительного кластера на четырехъядерных процессорах // Информационные и коммуникационные технологии в образовании и научной деятельности: материалы межрегион. науч.-практ. конф., Хабаровск, 21–23 сентября 2009 г. – Хабаровск: Изд-во Тихоокеанского гос. ун-та, 2009. – С. 261–268.

4. Щерба С.И., Пересветов В.В. Сравнительный анализ эффективности программного обеспечения для вычислительных кластеров // Информационные и коммуникационные технологии в образовании и научной деятельности: материалы межрегион. науч.-практ. конф. (г. Хабаровск, 21–23 мая 2008 г.). – Хабаровск: Изд-во Тихоокеанского гос. ун-та, 2008. – С. 363–369.

5. Пересветов В.В., Сапронов А.Ю., Тарасов А.Г. Вычислительный кластер бездисковых рабочих станций. Препринт № 83. Вычислительный центр ДВО РАН. – Хабаровск, 2005. – 50 с.