Организация эффективных вычислений в распределенной среде NumGRID

М.А. Городничев
Институт вычислительной математики и математической геофизики СО РАН, Новосибирск

Представлены результаты экспериментов по объединению вычислительных кластеров на основе программного комплекса NumGRID для решения задач численного моделирования. Показано, что существуют задачи численного моделирования, которые демонстрируют приемлемый уровень эффективности использования вычислительных ресурсов при запуске в распределенной среде NumGRID.

Введение

Программный комплекс NumGRID [1] предназначен для объединения разнородных вычислительных кластеров в единый вычислительный ресурс на основе частичной реализации стандарта MPI-2.2 [2]. NumGRID обеспечивает возможность запускать MPI-приложение так, чтобы процессы приложения были распределены по рабочим узлам нескольких кластеров. При этом процессы составляют один коммуникатор MPI и имеют возможность обмениваться между собой средствами MPI.

NumGRID позволяет:

- решать задачи, для которых недостаточно ресурсов отдельных кластеров;
- продлять жизнь устаревающего оборудования за счет объединения с новым;
- распределять части комплексных задач между специализированными кластерами в соответствии с индивидуальными требованиями частей к оборудованию и программному обеспечению;
- повысить гибкость при планировании распределения задач между кластерами в грид;
- планировать постепенное наращивание мощностей распределенной системы.

Обзор NumGRID

Схема устройства объединенного вычислительного ресурса NumGRID представлена на рис. 1.

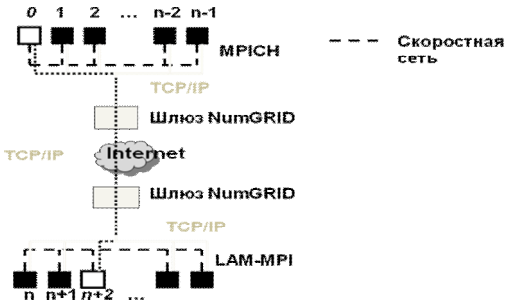

Рис. 1. Устройство NumGRID

В примере объединяются два кластера через Internet. Каждый кластер имеет рабочие узлы, объединенные высокоскоростной сетью (Myrinet, Infiniband и др.). Также узлы связаны с головным узлом сетью с меньшей пропускной способностью, поддерживающей протокол TCP. Процессы приложения MPI, находящиеся на рабочих узлах одного кластера, обмениваются между собой через высокоскоростную сеть средствами специфичной для кластера библиотеки MPI, позволяющей использовать возможности высокоскоростной сети. В примере это библиотеки MPICH и LAM-MPI. Для процессов, распределенных по рабочим узлам двух кластеров, поддерживается глобальная адресация в рамках одного коммуникатора MPI. Сообщения между процессами, расположенными на разных кластерах, проходят путь через головные узлы кластеров, где для этих целей перед стартом приложения запускаются шлюзы.

Пользователь вначале запускает шлюзы на каждом кластере. Шлюзы устанавливают между собой связь. Пользователь загружает на головные узлы кластеров исходный код своего приложения и исходные данные, собирает приложение на каждом кластере с библиотекой NumGRID-MPI, запускает необходимое число процессов на каждом кластере как локальные задачи MPI. Локальная задача получает в параметрах информацию о месте данной группы процессов в общей распределенной задаче, что позволяет обеспечить глобальную адресацию процессов. В конце работы пользователь забирает с каждого кластера результаты работы программы.

Экспериментальное исследование NumGRID

Эксперименты по запуску приложений в NumGRID проводились на объединении кластеров Сибирского суперкомпьютерного центра

(ССКЦ) [8] и Новосибирского государственного университета (НГУ). Кластеры построены на основе двухпроцессорных узлов с процессорами Intel Xeon E5540 (4 ядра) и внутренней коммуникационной сетью InfiniBand. Пропускная способность канала между головными узлами кластеров – 10 Гб/с.

Решение волнового уравнения

Ниже представлены графики, демонстрирующие характеристики исполнения программы решения волнового уравнения с помощью двухслойной явной схемы, размер задачи 10000x10000. На всех диаграммах запись «P (NxS)» означает «всего P ядер, из них N – на кластере 1 и S – на кластере 2».

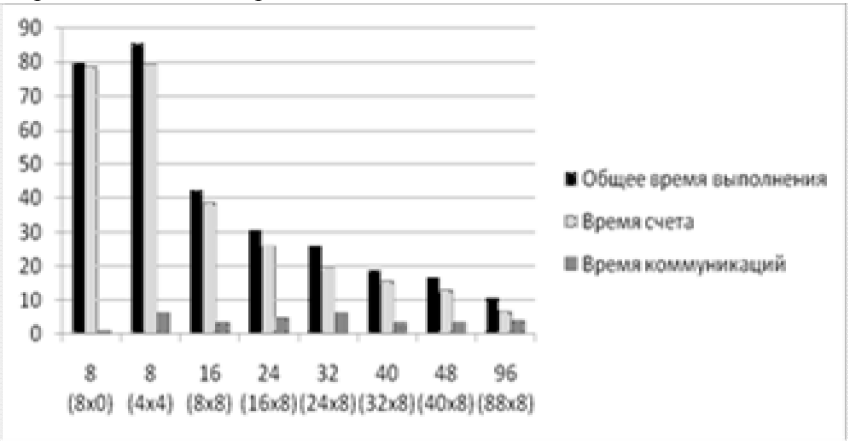

Рис. 2. Время решения волнового уравнения явным методом на NumGRID, с

Из рис. 2 видно, что при переходе к распределенным вычислениям (с 8x0 на 4x4) увеличивается общее время работы программы за счет увеличения расходов на коммуникации. Расходы на коммуникации в данном примере увеличиваются в 5 раз, что приводит к росту времени работы программы на ~6%.

Генерация случайных чисел и расчет статистик

Ниже представлены результаты тестирования программы, которая параллельно генерирует случайные величины и вычисляет статистики. Программа используется для моделирования физических процессов методами Монте-Карло. Программа осуществляет редкие коллектив-

ные коммуникации для сбора статистик, в остальное время процессоры выполняют существенные по объему вычисления независимо. Это позволяет ожидать, что относительно плохая пропускная способность сети между кластерами незначительно скажется на общей производительности программы.

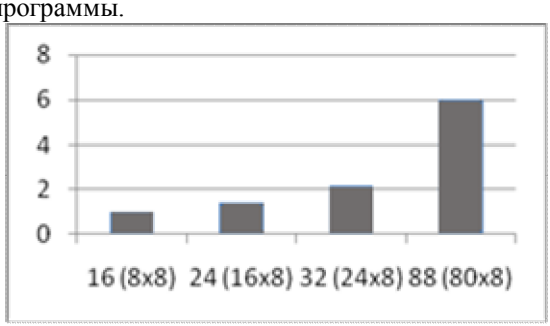

Рис. 3. Время генерации 2000 значений случайной величины и расчета статистик на NumGRID, с. При этом время генерации 2000 значений случайной величины и расчета статистик на одном ядре кластера 1: 0,0109 с

Время исполнения программы при переходе от решения на одном ядре к решению на 16 ядрах, по 8 ядер на каждом кластере, уменьшается в 21 раз. Объяснить подобное сверхлинейное ускорение можно более эффективным использованием кэшей процессоров при увеличении количества процессоров и низкими коммуникционными расходами, свойственными данной задаче. При дальнейшем увеличении количества вовлеченных ресурсов так же отмечается сверхлинейное ускорение. Рисунок 4 дает более ясное представление об ускорении времени решения при увеличении количества процессоров.

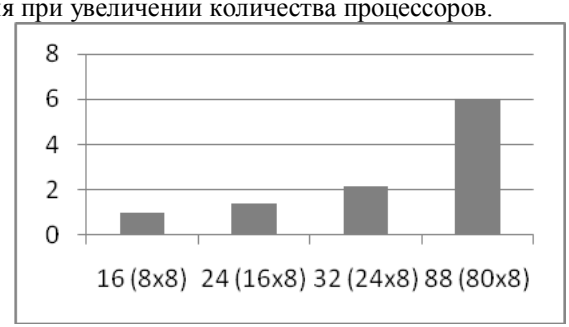

Рис. 4. Ускорение генерации 2000 значений случайной величины и расчета статистик на NumGRID относительно времени работы программы на двух узлах: один узел кластера НГУ (8 ядер) и один узел кластера ССКЦ (8 ядер)

В ходе экспериментов показано, что использование дополнительных процессоров из другого кластера позволяет уменьшать время выполнения приложения при определенных конфигурациях запуска. В частности, на задаче решения волнового уравнения достигается эффективность 90% на 40 процессорных ядрах относительно производительности программы на 8 ядрах, и 65% – на 96 ядрах. На задаче генерации случайных чисел достигается эффективность 92% на 24 ядрах и 109% на 88 ядрах.

Несмотря на существенную разницу в производительности систем коммуникации внутри кластеров и между кластерами, время решения задач увеличивается умеренно для решения волнового уравнения на том же количестве процессоров и уменьшается для решения задачи поиска статистик. Можно заключить, что существует класс задач, которые могут быть решены на объединении кластеров за приемлемое для пользователей время. Вместе с тем объединение кластеров может позволить решать задачи большего объема, чем был бы способен решить один кластер за то же время.

Заключение

Проведены испытания системы NumGRID по объединению кластеров Сибирского суперкомпьютерного центра и Новосибирского государственного университета. Эксперименты демонстрируют, что существуют классы задач численного моделирования, для решения которых распределенные вычислительные системы на основе NumGRID применимы.

Дальнейшая работа заключается в расширении поддержки стандарта MPI-2.2, адаптации существующих приложений для работы в среде NumGRID, создании методов и средств разработки программ численного моделирования для неоднородных вычислительных сред.

Литература

1. **Fougere D., Gorodnichev M., Malyshkin N. et al.** NumGrid Middleware: MPI Support for Computational Grids // Parallel Computing Technologies: 8th International Conference, PaCT 2005, Krasnoyarsk, Russia, September 5–9, 2005. Proceedings. – N.Y. : Springer, 2005. – LNCS Vol. 3606. – P. 313–320.

2. MPI Documents [Электронный ресурс]. – Режим доступа: http://www.mpi-forum.org/docs/docs.html (дата обращения: 15.12.2011).