Параллельный алгоритм отображения информационного графа MPI-программы на процессорные ядра распределенной вычислительной системы

М.Г. Курносов

Современные распределенные вычислительные системы (ВС) имеют сложноорганизованные коммуникационные среды, в большинстве случаев с неоднородными по производительности каналами связи между процессорными ядрами. Например, в системах Cray XT5 и IBM BlueGene/P время передачи сообщения между парой процессорных элементов зависит от их размещения в трехмерном торе, а в (мульти)кластерных и вычислительных GRID-системах коммуникационные среды имеют иерархическую организацию, в которых первый уровень коммуникационной среды – сеть связи между кластерами, второй уровень – сеть связи внутри кластеров, третий уровень – среда доступа процессоров вычислительного узла к общей памяти.

MPI-программы, разрабатываемые для таких систем, характеризуется информационными графами, вершинам которых соответствуют параллельные ветви (процессы) программ, а ребрам – обмены между ними.

При организации функционирования распределенных ВС важной является задача оптимального вложения параллельной MPI-программы в структуру распределенной ВС с целью минимизации накладных расходов на обмены между ветвями и дисбаланса загрузки процессорных ядер системы. Результат вложения – распределение рангов MPI-процессов по процессорным ядрам. Ниже приведены примеры запусков MPI-теста HPL (High Performance Linpack, рис. 1) на двух узловом кластере с распределением ветвей без учета структуры коммуникационной среды и с учетом производительности каналов связи. Каждый узел – два двухъядерных процессора Intel Xeon 5150, узлы объединены сетью связи стандарта Gigabit Ethernet.

Результаты выполнения теста HPL с распределением ветвей без учета структуры коммуникационной среды (рис. 2): время выполнения T = 118 с., оценка производительности R = 44 GFLOPS. Результаты запуска с учетом производительности каналов связи (рис. 3): T = 100 с., R = 53 GFLOPS.

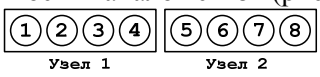

Рис. 2

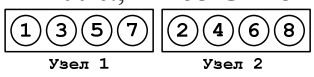

Рис. 3

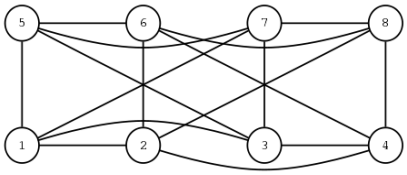

Рис. 1. Структура обменов между параллельными ветвями теста HPL (NP = 8, PMAP = 1, BCAST = 5)

Из примеров видно, что учет структуры коммуникационной среды при организации выполнения MPI-программ даже на небольшой подсистеме позволяет сократить время выполнения программы.

Задача оптимального вложения параллельной программы в структуру распределенной ВС представляет собой задачу дискретной оптимизации. Разработана группа нетрудоемких алгоритмов позволяющих отыскивать приближенные решения задачи за приемлемое время.

Создан пакет MPITaskMap оптимизации выполнения MPI-программ на (мульти)кластерных ВС. В пакет входят средства анализа протоколов выполнения MPI-программ, средства оценки производительности каналов связи коммуникационных сред и модуль вложения параллельных программ в структуру распределенной ВС с целью минимизации времени их выполнения.

Проведена серия экспериментов на вычислительных кластерах различных конфигураций с MPI-программами из пакетов NAS Parallel Benchmarks и SPEC MPI2007. Результаты показывают, что на SMP-кластерах оптимизация вложения программ в структуры систем позволяет сократить время их выполнения от 1 до 60 процентов относительно времени выполнения с распределением ветвей стандартным алгоритмом mpiexec.

Направление дальнейших исследований – разработка алгоритмического и программного инструментария распределения множества параллельных программ по процессорным ядрам ВС с учетом структуры ее коммуникационной среды.