Распараллеливание задачи умножения разреженной матрицы на вектор на вычислительных кластерах с минимальной аппаратной поддержкой PGAS

А.А. Корж

В данной работе предлагается исследовать способ распараллеливания задачи умножения разреженной матрицы на вектор, который не реализуется эффективно в парадигме передачи сообщений MPI, но может быть эффективно реализован в парадигме, основанной на использовании общей памяти.

Рассматривается задача умножения разреженной матрицы на вектор. Известным бенчмарком, основанным на данной задаче, является метод сопряженных градиентов CG из пакета NAS Parallel Benchmarks. Стандартные схемы распараллеливания делятся на две группы: разбивающие матрицу по узлам горизонтально или вертикально. Каждый узел локально выполняет обычное умножение разреженной матрицы на вектор. При этом, получаемая реальная производительность оказывается чрезвычайно низкой (не более 5 процентов от пиковой). Основной причиной этого является нерегулярность работы с памятью, преобладание коммуникаций над вычислениями и разбалансировку нагрузки в случае несимметричной матрицы. Задача содержит большой уровень внутреннего параллелизма, но из-за его мелкозернистости масштабируемость оказывается крайне низкой.

Предлагаемая идея заключается в том, чтобы не подкачивать плотный вектор как он есть, а подкачивать те элементы вектора и в том порядке в каком они будут умножаться на данном узле на ненулевые элементы матрицы. Целью является организовать коммуникации таким образом, чтобы после их завершения у каждого узла в локальной памяти получался массив такой же длины как и первый массив части матрицы в формате CRS, то есть содержащий ненулевые элементы построчно. Получится, что все обращения к локальной памяти имеют плотный последовательный характер «удобный» для суперскалярных процессоров, а вся нерегулярность задачи окажется «спрятана» в нерегулярности удаленных записей.

Для эффективной реализации описанной выше схемы нам потребуется поддержка записей в удаленную память со следующими свойствами: 1) возможность выдачи каждым десятков миллионов в секунду операций записи 64х-битного слова в память другого узла 2) возможность определить, что все нужные нам записи получены узлом и мы можем начинать вычисления.

Вслед за [1] предлагается реализовать аппаратную поддержку распределенной общей памяти за счет внесения небольших изменений в аппаратуру адаптера сетевого интерфейса коммуникационной сети. Для выполнения свойства предлагается дальнейшая модификация адаптера, позволяющая считать количество полученных пакетов.

Стандартные схемы распараллеливания умножения разреженной матрицы на вектор в парадигме MPI, основными недостатками которых, являются растущие объемы коммуникаций и нерегулярность доступа к локальным данным. Предложенный метод исправляет оба этих недостатка, «расплатой» же за это является переход к пересылкам небольших сообщений, что снизит пропускную способность сети.

Автором получена теоретическая оценка, что при $N_{PE} > 4s$ (s - число ненулевых элементов в строке) предложенный метод окажется эффективней классических схем.

Литература

1. Underwood, K. D., Levenhagen, M. J., and Brightwell, R. 2007. Evaluating NIC hardware requirements to achieve high message rate PGAS support on multi-core processors. In Proceedings of the 2007 ACM/IEEE Conference on Supercomputing (Reno, Nevada, November 10 - 16, 2007). SC '07. ACM, New York, NY, 1-10.