Оптимизация работы MPI-программ для кластеров, использующих интерконнект Ангара

М.Р.Халилов, А.В.Тимофеев

МИЭМ ВШЭ

В рамках оптимизации запуска и работы параллельных MPI-программ на вычислительных кластерах, использующих коммуникационную сеть Ангара, рассмотрен метод, основанный на подборе топологии вычислительной системы и исследована эффективность предложенной оптимизации.

На основе анализа работы параллельной программы составляется информационный граф программы, который используется эвристическим алгоритмом для эффективного распределения её процессов по процессорным ядрам с целью минимизации суммарного времени выполнения обменов между ветвями MPI-программы. Для оптимизации отображения на первом этапе выполняется анализ информационного графа обмена сообщений между процессами данной MPI-программы. Идея метода оптимизации отображения параллельной программы заключается в разбиении информационного графа программы на непересекающиеся подмножества интенсивно обменивающихся процессов и привязке этих подмножеств к узлам/процессорам, соединённым наиболее быстрыми каналами связи. Разбиение выполняется для минимизации суммы рёбер, соединяющих разные подмножества разбиения. Разбиение рекурсивно выполняется сначала для уровня описывающего обмены между вычислительными узлами, а затем для уровня описывающего обмены внутри каждого из узлов в случае, если внутри узла установлено несколько процессоров. Целью такого разбиения является минимизация времени выполнения программы. Задача оптимального отображения процессов MPI-программы является NP-полной, так как её можно свести к задаче разбиения графов. Для её решения целесообразно использовать эвристические алгоритмы дающие решения, близкие к оптимальным.

Показано уменьшение времени выполнения MPI-программ при использовании алгоритма оптимального отображения. Рассмотрены результаты анализа зависимости времени выполнения MPI-программ от параметров кластера.