УДК 004.272.26 DOI: 10.14529/cmse150201

ОПТИМИЗАЦИЯ ОТОБРАЖЕНИЯ НЕОДНОРОДНО ВЗАИМОДЕЙСТВУЮЩИХ MPI ПРОЦЕССОВ НА ВЫЧИСЛИТЕЛЬНУЮ АРХИТЕКТУРУ[1]

В.В. Гетманский, В.С. Чалышев, Д.И. Крыжановский, Е.И. Лексиков

Разработан метод отображения на кластерную архитектуру неоднородно взаимодействующих параллельных процессов в вычислительном приложении, использующем MPI. Метод предназначен для сокращения задержек при синхронизации за счет назначения наиболее интенсивно взаимодействующих процессов, на вычислительные ядра с наиболее быстрым интерконнектом. Метод использует представление вычислительной задачи и архитектуры кластера в виде взвешенного графа. Разработан эвристический алгоритм, дающий за приемлемое время результат отображения номеров процессов на номера вычислительных ядер кластера. На примере хорошо масштабируемого вычислительного пакета получено ускорение вычислений на 17–20 % в результате оптимизации отображения для тестов от 300 до 4800 процессов.

Ключевые слова: отображение задач, кластер, графы задачи и системы, MPI.

Введение

Проблема отображения параллельной программы на архитектуру вычислительной системы с целью уменьшения времени обмена данными рассмотрена в ряде работ отечественных [1, 2] и зарубежных [3–5] авторов. В настоящей работе рассмотрена разработка и тестирование метода отображения графа задачи на граф вычислительной системы. Вершины графа задачи соответствуют параллельным процессам, а ребра графа задачи взвешены объемами пересылаемых между процессами данных. Вершины соединены ребрами, только если соответствующие им процессы обмениваются данными. В научных вычислительных пакетах граф задачи имеет произвольную структуру. Пример такого графа для вычислительного пакета ZeusMP, запущенного на 2 узлах кластера, показан на рис. 1. Ширина линий, обозначающих ребра, для наглядности пропорциональна весу ребра. Данное обозначение будет использовано на остальных рисунках.

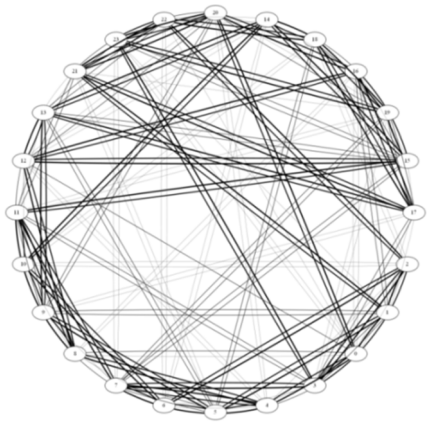


Рис. 1. Граф задачи ZeusMP

[1] Статья рекомендована к публикации программным комитетом Международной научной конференции «Параллельные вычислительные технологии-2015».

Граф системы представляет собой топологию кластера с учетом различной производительности каналов связи (интерконнекта). Ребра графа системы взвешены постоянными коэффициентами, представляющими соотношение времени передачи данных внутри узла кластера и времени передачи данных между узлами.

Современные кластеры состоят из многосокетных узлов, как правило, содержащих сопроцессоры GPGPU, FPGA или MIC. Каналы связи также работают с различной скоростью (общая память, InfiniBand, Ethernet). Таким образом, граф кластера тоже может иметь неоднородную структуру из-за сложной конфигурации узлов кластера (рис. 2).

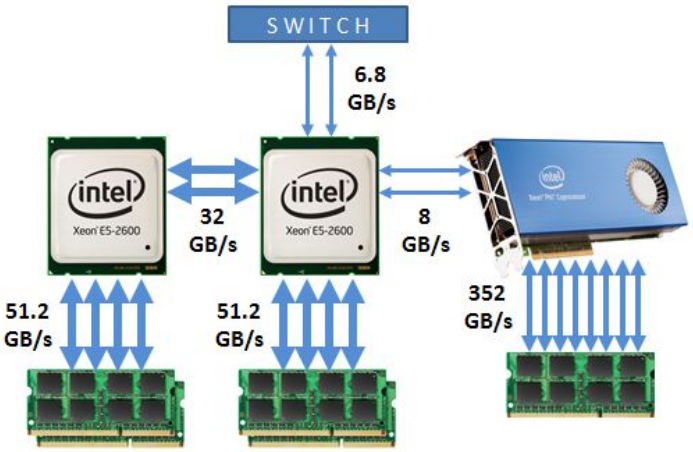


Рис. 2. Конфигурация узлов гетерогенного кластера

Основное проверяемое предположение в настоящей работе состоит в том, что обмен данными можно ускорить, если распределить процессы на кластере так, что наиболее интенсивный обмен (ребра графа задачи с максимальным весом) будет происходить по наиболее быстрому каналу связи (ребра графа системы с максимальным весом).

Статья организована следующим образом. В первом разделе рассмотрена постановка задачи отображения графа параллельной программы на граф вычислительной системы. Во втором разделе описаны два реализованных алгоритма решения этой задачи. В третьем разделе описан синтетический тест параллельной программы с сильно неоднородным обменом данными между процессами. В четвертом разделе описаны вычислительные эксперименты на кластере с использованием синтетического теста и пакета QBox, проведен анализ полученных результатов. В заключении перечислены основные результаты исследования и предложения по усовершенствованию метода.

1. Постановка задачи

Обозначим граф задачи как

$$G_1(P, L),\ L_i = (n_i, d_i),\ i = 1, \dots, K, \tag{1}$$

где P — набор вершин графа (процессы), L — набор связей (взаимодействие процессов), n_i — частота обмена данными, d_i — объем обмениваемых данных, K — число связей.

Обозначим граф системы как

$$G_2(V, D),\ D_j = \left(l_j, b_j\right),\ j = 1, \dots, S, \tag{2}$$

где V — множество вершин (соответствующих вычислительным ядрам центральных процессоров или сопроцессоров), D — множество ребер (каналы передачи данных), l_j — задержка, b_i — пропускная способность, S — общее число каналов передачи данных. В первом приближении в графе системы передача данных возможна между любой парой вершин, т.е. G_2 — полносвязный граф и $K < S$.

Искомое отображение можно записать как

$$G_1 \to G_2\ (P \to V)\colon L_i \to D_j,\ F(D,L) \to \min. \tag{3}$$

Минимизируемая целевая функция задает временную задержку на пересылки и имеет вид

$$F(D,L) = \sum_{k=1}^{N} T_k\ ,\ T_k\left(D_i, L_j\right) = l_i \cdot n_j + \frac{d_j}{b_i},\ F \to \min. \tag{4}$$

Для разработанного прототипа используется упрощенная формулировка. Все ребра графа задачи взвешены объемом обмениваемых данных между процессами, которым соответствуют вершины ребра. Граф системы взвешен коэффициентом задержки канала связи r_i. Для задания того, что сетевое взаимодействие заведомо медленнее взаимодействия через общую память, r_i определено как

$$r_i = \begin{cases} 1, \text{если обмен через общую память} \\ 2, \text{если обмен через } InfiniBand \end{cases} \tag{5}$$

Целевая функция вычисляется как сумма произведений весов ребер:

$$\tilde{F}\left(\widetilde{D}, L\right) = \sum_{k=1}^{N} \tilde{T}_k\ ,\ \tilde{T}_k\left(\widetilde{D}_i, L_j\right) = r_i \cdot d_j,\ \widetilde{D}_i = (r_i),\ \tilde{F} \to \min. \tag{6}$$

Таким образом, для решения задачи отображения необходимо построить граф задачи и граф системы, разработать алгоритм поиска отображения графа задачи на граф системы с целью минимизации суммы произведений весов ребер.

2. Алгоритм отображения

Разработаны два алгоритма отображения: первый использует переборную стратегию, второй — «жадную» стратегию. Для графов, содержащих более 10 вершин, полный перебор выполняется неоправданно долго, поэтому практически применим только второй алгоритм, который дает либо оптимальное решение, либо близкое к оптимальному. Экспериментально установлено, что значение целевой функции становится не хуже исходного после работы алгоритма для проведенных тестов.

2.1. Алгоритм полного перебора

Алгоритм ищет решение с помощью полного перебора с возвратом. Вычислительная сложность алгоритма $O\left(\binom{N}{M} \cdot M!\right)$, где N — число вершин в графе системы, M — число вершин в графе задачи. Таким образом, на практике перебор можно использовать только для графов системы, содержащих до 15 вершин, и для графов задачи, содержащих до 10 вершин.

2.2. Алгоритм с «жадной» стратегией

Реализация основана на похожем подходе, описанном в работе [5]. Основное отличие предлагаемой реализации в том, что отсутствует жесткое требование равенства вершин в графе задачи и графе системы.

Общий алгоритм:

1. Поиск первого приближения. В разработанном прототипе используется отображение -й вершины P_i графа кластера на j-ю вершину V_j графа системы.
2. Итеративная процедура улучшения решения:
   2.1. Перестановка отображения двух вершин. Смена отображения $P_i \to V_j,\ P_k \to V_l$, на отображение $P_i \to V_l,\ P_k \to V_j$.
   2.2. Отображение на новую вершину. Смена отображения $P_i \to P_j$ на $P_i \to P_k$.

Вычислительная сложность алгоритма $O\big(IE(M+N)\big)$, где I — число итераций, E — число ребер в графе задачи M — число вершин в графе задачи, N — число вершин в графе системы.

В общем случае оценить сходимость алгоритма к точному решению сложно. Предположим, что все ребра в графе задачи 2 типов и имеют кратные веса a и $a \cdot k$, все ребра в графе системы тоже 2 типов с весами b и $b \cdot r$. В этом случае верхняя граница сходимости $Mkr/(r-1)$, т.е. сходимость алгоритма для большинства случаев быстрее этого значения.

3. Описание синтетического теста

Синтетический тест моделирует обмен данными различного объема и с различной интенсивностью в итеративной процедуре, выполняющейся в MPI программе. Тест представляет собой наиболее оптимизируемый частный случай входных данных для алгоритма оптимизации. В тесте используется обмен данными двух типов: обмен внутри группы тесно взаимодействующих процессов с интенсивным взаимодействием и обмен между группами процессов.

Параметры синтетического теста: N_{bl} — число групп процессов, N_{perbl} — число процессов в группе, D_{bl} — объем данных, пересылаемых между группами, D — объем данных, пересылаемых внутри группы, N_{skip} — число пропускаемых итераций.

В каждой группе каждый процесс взаимодействует с каждым процессом группы (полно связанный подграф). Взаимодействие между группами включает только обмен данными между последним процессом группы i и первым процессом группы с номером $i+1$. Например, для значений параметров $N_{bl} = 2$, $N_{perbl} = 6$ взаимодействие между 11-м процессом, входящим в первую группу, и первым процессом, входящим во вторую группу, показано тонкими ребрами на графе задачи (рис. 3).

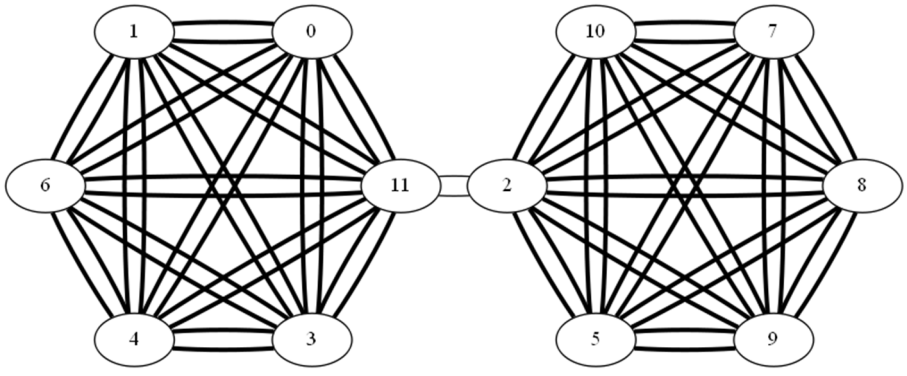


Рис. 3. Пример графа для 2 блоков с 6 процессами в каждом

Взаимодействие между группами происходит не на каждой итерации, поэтому задается пропуск итераций. Номера процессов перемешаны случайным образом (с постоянным параметром генератора случайных чисел для воспроизводимости коммуникационного взаимодействия).

Смещения и длины в линейном массиве с данными строятся на каждой итерации для операций MPI_Alltoallv коллективного обмена MPI-библиотеки. Эти массивы дополняются на каждой $1 + N_{skip}$ итерации данными для обмена между группами. Например, в случае $N_{skip} = 2$ массивы для графа задачи, показанном на рис. 3, имеют вид, представленный в табл. 1.

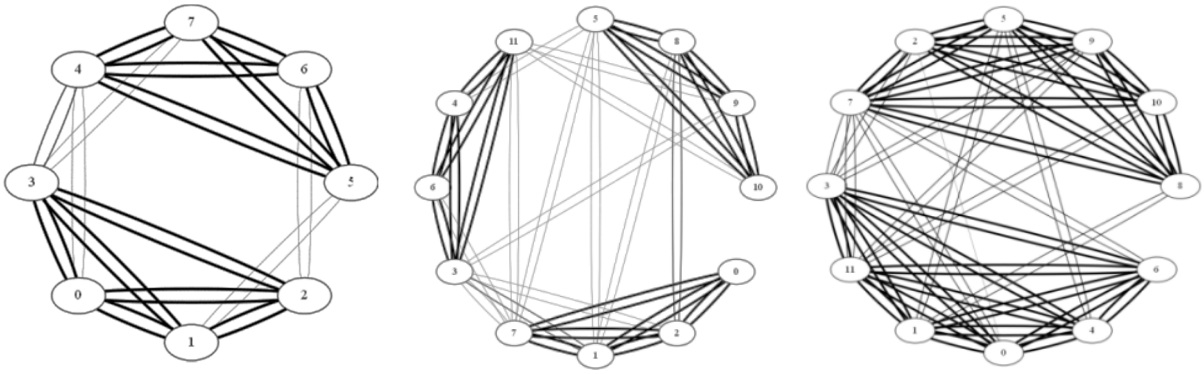
Таблица 1

Смещения и длины для блоков обмениваемых данных MPI-процессов

Процесс назначения		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
итерации $3k$, $1+3k$, процесс 2	смещение	0	0	0	0	0	0	D	D	$2D$	$3D$	$4D$	$5D$
	длина	0	0	0	0	0	D	0	D	D	D	D	0
итерации $3k$, $1+3k$, процесс 11	смещение	0	D	$2D$	$2D$	$3D$	$4D$	$4D$	$5D$	$5D$	$5D$	$5D$	$5D$
	длина	D	D	0	D	D	0	D	0	0	0	0	0
итерация $2+3k$, процесс 2	смещение	0	0	0	0	0	0	D	D	$2D$	$3D$	$4D$	$5D$
	длина	0	0	0	0	0	D	0	D	D	D	D	D_{bl}
итерация $2+3k$, процесс 11	смещение	0	D	$2D$	$2D + D_{bl}$	$3D + D_{bl}$	$4D + D_{bl}$	$4D + D_{bl}$	$5D + D_{bl}$	$5D + D_{bl}$	$5D + D_{bl}$	$5D + D_{bl}$	$5D + D_{bl}$
	длина	D	D	D_{bl}	D	D	0	D	0	0	0	0	0

В табл. 1 показано, что массив смещений и длин обмениваемых блоков для процессов, участвующих в обмене между группами (процессы 2 и 11), на каждой итерации с номером $0 + 3k$ и $1 + 3k$ содержит 5 обмениваемых блоков длиной D, а на каждой итерации с номером $2 + 3k$ содержит дополнительный блок обмена между группами длиной D_{bl}. У остальных процессов массив длин и смещений не меняется.

Пример графов задач, полученных с помощью запусков синтетического теста с различными параметрами, показан на рис. 4. Тонкими линиями показаны системные взаимодействия процессов при синхронизации, полученные при сборе статистики.



а б в

Рис. 4. Пример графов задачи, построенных по синтетическим тестам: а) 2 блока с 4 процессами на блок, б) 3 блока с 4 процессами на блок, в) 2 блока с 6 процессами на блок

Таким образом, с помощью синтетического теста моделируется неоднородное взаимодействие заданного количества групп параллельных процессов с возможностью менять объем данных для обмена и интенсивность обмена данными между группами. При задании числа процессов, соответствующего числу ядер на один узел кластера и перенумерации процессов случайным образом, собирается статистика обмена данными, которая подается на вход алгоритму оптимизации.

4. Результаты тестирования

4.1. Запуск тестов

Запуск тестов проводился с использованием программы запуска Intel® MPI Library [6]. Передача набора параметров, необходимых для назначения процессов на вычислительные ядра, реализована с помощью конфигурационного файла, передаваемого по ключу командной строки «--configfile». Строка запуска имеет вид: «mpirun --configfile config.txt».

Разработан инструментарий для генерации файла конфигурации. В качестве входного файла используется список хостов кластера. Первый шаг — запуск непродолжительного теста для сбора статистики коммуникаций MPI-процессов. Входные данные для генерации конфигурационного файла (config.txt) — это список хостов с именами узлов машин на кластере и числом ядер на каждом узле (hostlist.txt).

Например, для кластера из 4 узлов с 4-ядерными процессорами содержимое файла hostlist.txt имеет вид в листинге на рис. 5:

```
line 01: node1 4
line 02: node2 4
line 03: node3 4
line 04: node4 4
```

Рис. 5. Пример файла hostlist.txt для 4 узлов

Здесь и в остальных листингах начало строки обозначено как line xx, xx — номер строки. Генерируемый файл конфигурации config.txt для сбора статистики имеет вид, приведенный в листинге на рис. 6.

```
line 01: -env I_MPI_STATS 4 -env I_MPI_STATS_FILE
/path/to/stat.txt -n 4 -host node1 /path/to/application
line 02: -env I_MPI_STATS 4 -env I_MPI_STATS_FILE
/path/to/stat.txt -n 4 -host node2 /path/to/application
line 03: -env I_MPI_STATS 4 -env I_MPI_STATS_FILE
/path/to/stat.txt -n 4 -host node3 /path/to/application
line 04: -env I_MPI_STATS 4 -env I_MPI_STATS_FILE
/path/to/stat.txt -n 4 -host node 4 /path/to/application
```

Рис. 6. Пример файла config.txt для сбора статистики на четырех узлах

Запуск MPI программы с использованием файла конфигурации config.txt указанного формата приводит к формированию средствами библиотеки Intel® MPI Library файла статистики обменов stats.txt. Программе оптимизации подаются на вход файлы stats.txt и hostlist.txt, на выходе генерируется файл config.txt с оптимизированным отображением процессов на узлы кластера. Для рассматриваемого примера содержимое файла config.txt приведено в листинге на рис. 7.

```
line 01:-n 2 -host node1 /path/to/application
line 02:-n 2 -host node4 /path/to/application
line 03:-n 1 -host node2 /path/to/application
line 04:-n 2 -host node1 /path/to/application
line 05:-n 3 -host node2 /path/to/application
line 06:-n 4 -host node3 /path/to/application
line 07:-n 2 -host node4 /path/to/application
```

Рис. 7. Пример файла config.txt для четырех узлов после оптимизации

Оптимизированный файл конфигурации используется для запуска длительного теста. Методология тестирования состоит в замере времени выполнения длительного теста до и после оптимизации размещения процессов на кластере. Показатель эффективности — ускорение, вычисляемое как $S = T/T_{opt}$, где T — время работы теста до оптимизации, T_{opt} — после оптимизации.

Для запуска тестов использован суперкомпьютер «Торнадо ЮУрГУ» [7]. Конфигурация выделенных узлов: двухсокетные узлы с процессорами Intel® Xeon® X5680 (2 процессора по 6 ядер), Intel® Xeon Phi™ SE10X (61 ядро по 1,1 ГГц), 2 Гб RAM, InfiniBand QDR, топология сети — толстое дерево. При тестировании были поставлены задачи исследования синтетического теста до и после оптимизации отображения процессов на ядра двух сопроцессоров и центральных процессоров кластера при различной размерности блоков. На центральных процессорах кластера также было необходимо провести запуск масштабируемого на большое число узлов вычислительного приложения с неоднородным взаимодействием процессов, в качестве которого был выбран CORAL QBox [8].

4.2. Синтетический тест для двух узлов кластера с сопроцессорами

Конфигурация для тестирования состоит из двух узлов с установленными сопроцессорами Intel® Xeon Phi™, содержащими по 60 ядер, доступных для вычислений. Такая конфигурация была выбрана для оценки влияния работы MPI тестов, характеризующихся слабой связанностью по данным групп параллельных процессов с интенсивным

обменом, что имеет место, например, при моделировании связанных физических задач. Данный тест был необходим для оценки влияния сетевого взаимодействия в случае интенсивного обмена групп с большим числом параллельных процессов, поэтому в графе системы было только 2 типа весов ребер: связи между ядрами и сетевой интерконнект, взаимодействие ядер сопроцессора принималось приближенно равнозначным. Результаты тестирования приведены в табл. 2. Параметры теста: $D = 60$ Мб, $D_{bl} = 60$ Кб, $N_{skip} = 0$.

Таблица 2

Результаты синтетического теста для 2 узлов с сопроцессорами

N_{bl}	N_{perbl}	T, с	T_{opt}, с	S
2	60	660,847	84,125	7,855
3	40	440,512	194,514	2,264
4	30	320,938	153,668	2,088
5	24	261,291	86,774	3,011
6	20	221,371	35,210	6,287
8	15	165,581	34,700	4,771
15	8	93,015	39,100	2,378
20	6	68,347	30,007	2,277

Каждый тестовый запуск проводился в Native режиме (код выполняется только на сопроцессорах) и использовал все вычислительные ядра сопроцессора: $N_{bl} \cdot N_{perbl} = 120$. Содержимое сгенерированного конфигурационного файла для сбора статистики приведено в листинге на рис. 8.

```
line 01:-env I_MPI_STATS 4 -env I_MPI_STATS_FILE
/path/to/stat.txt -n 60 -host mic0 /path/to/application
line 02:-env I_MPI_STATS 4 -env I_MPI_STATS_FILE
/path/to/stat.txt -n 60 -host mic1 /path/to/application
```

Рис. 8. Пример файла config.txt для сбора статистики на 2 сопроцессорах

Фрагмент файла конфигурации config.txt после оптимизации на основе собранной статистики приведен в листинге на рис. 9.

```
line 01:-n 5 -host mic0 /home/test/MPICommunication 20 6
line 02:-n 1 -host mic1 /home/test/MPICommunication 20 6
line 03:-n 4 -host mic0 /home/test/MPICommunication 20 6
line 04:-n 2 -host mic1 /home/test/MPICommunication 20 6
line 05:-n 1 -host mic0 /home/test/MPICommunication 20 6
line 06:-n 3 -host mic1 /home/test/MPICommunication 20 6
. . .
```

Рис. 9. Фрагмент файла config.txt для 2 сопроцессоров после оптимизации

Файл для сбора статистики состоит из 2 строк (по числу узлов), а сгенерированный после оптимизации конфигурационный файл для запуска содержит 55 строк, так как номера процессов изначально перемешаны случайным образом.

В результате отображения параллельных процессов на ядра получено ускорение выполнения тестов в несколько раз, что свидетельствует о существенной зависимости расположения параллельных процессов на вычислительных ядрах и целесообразности использования алгоритма отображения.

4.3. Синтетический тест для 10 узлов с InfiniBand интерконнектом

В качестве тестового стенда использовался кластер с 10 двухсокетными узлами и 6-ядерными процессорами Intel® Xeon®. Тестирование проводилось для фиксированного объема данных, но с разным числом итераций (N_{iter}), в соответствии с параметрами, приведенными в табл. 3.

Таблица 3

Результаты синтетического теста для 10 узлов кластера

N_{iter}	D_{bl}, Кб	D, Мб	T, с	T_{opt}, с	S
100	8	8	1755	880	1,99
1000	0,8	0,8	23,21	9,56	2,42
10000	0,08	0,08	5,96	1,01	5,9

Для обмена внутри групп общий объем передаваемых данных между парой процессов составляет примерно 800 Мб, для обмена между группами — 800 Кб. В тесте сгенерированы 10 групп по 12 процессов в каждой (по числу ядер каждого узла кластера). Фрагмент графа задачи приведен на рис. 10.

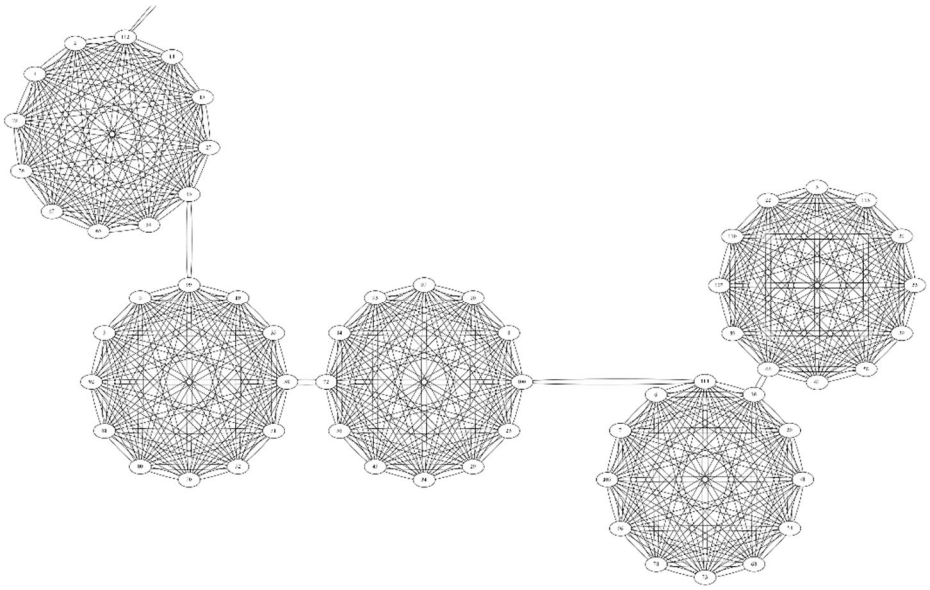


Рис. 10. Фрагмент графа задачи для синтетического теста (5 групп по 12 процессов)

Результаты свидетельствуют о том, что более частые обмены небольших порций данных с помощью наиболее эффективных в данном случае операций коллективного обмена библиотеки MPI оптимизируются лучше, чем менее интенсивные обмены больших порций данных. Частые обмены небольших порций характерны для вычислительных приложений, использующих, например, декомпозицию расчетной области или параллельную реализацию явных методов интегрирования систем дифференциальных уравнений. Для некоторых задач при неполной загрузке узлов (тест с менее чем 12 процессами на узел) выполнение параллельной программы может оказаться быстрее при использовании InfiniBand, чем при использовании общей памяти, например, для описанного в [9] случая. Для рассматриваемой задачи эксперименты показали проявление такого же эффекта: в случае неполной загрузки узлов (по 4–8 ядер из 12) тест срабатывал

медленнее при переназначении наиболее интенсивного обмена на общую память, чем при переназначении интенсивного обмена на InfiniBand. Разница времени выполнении составляла не более 5 %.

4.4. Результаты тестирования на удаленном кластере пакета для моделирования молекулярной динамики CORAL QBox

CORAL QBox — это пакет моделирования молекулярной динамики для определения свойств материала, использующий теорию функционала плотности. Метод имеет вычислительную сложность $O(N^3)$, где N — общее число валентных электронов в системе.

Тесты собраны и запущены для различного числа узлов удаленного кластера вплоть до 400. В тестах задействовано разное число узлов (N_{nodes}), при максимальной загрузке узлов число параллельных процессов составляет $12N_{nodes}$.

Основной параметр, влияющий на производительность QBOX, — это параметр максимального числа строк в $2D$ сетке параллельных процессов ($nrowmax$). Количество строк и столбцов в этой сетке подбирается так, что $nrowmax$ уменьшается, пока число параллельных процессов (N) не будет делиться на $nrowmax$, после чего число столбцов принимается равным $N/nrowmax$. Эксперименты показали, что максимальная эффективность имеет место при $nrowmax = N_{nodes}$, соответственно число столбцов составляет при этом 12. Число атомов при генерации входных данных было выбрано как максимальное число атомов, при котором задача помещается в оперативную память, значение которой ограничено 2 Гб на узел.

Граф задачи для пакета QBox имеет общий вид из-за неоднородного обмена данными. По статистике, собранной при запуске на двух узлах с $nrowmax = 2$, построен граф задачи, на котором можно выделить 2 группы по 12 процессов с тесным взаимодействием (рис. 11), причем порядок соответствующих вершин графа соответствует номерам ядер процессоров.

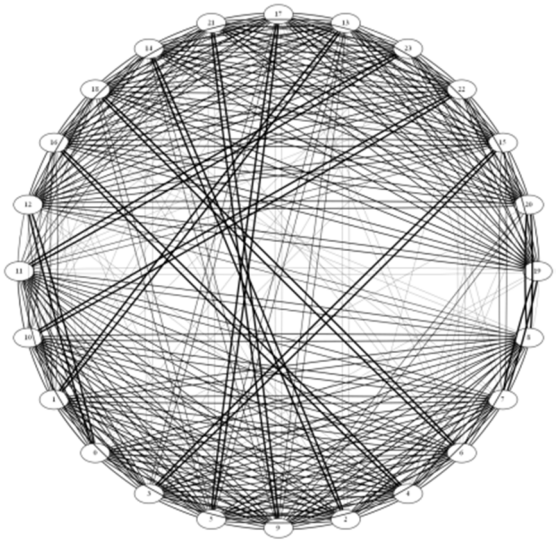


Рис. 11. Две группы из 12 процессов в графе задачи для пакета QBox

Каждой подгруппе соответствует полносвязный подграф графа задачи, в то время как между подграфами набор связей неполный и их суммарный вес, соответствующий объему переданных данных сопоставим по величине с суммарным весом ребер, инци-

дентных ребрам одной вершины в подгруппе (для каждой вершины средний объем переданных данных составил 185 Мб, а общий объем переданных данных между подграфами — 421 Мб).

Тесты в пакете QBox были подобраны для наилучшего времени выполнения при размерностях сеток 25×12, 50×12, 100×12, 150×12, 200×12, 400×12 для 25–400 узлов соответственно (число процессов и ядер 300–4800). Соотношение весов ребер графа системы составило 2:1 (общая память по отношению к InfiniBand). Результаты тестирования показаны в табл. 4.

Таблица 4

Результаты тестирования QBox

Число узлов	**Размерность сетки**	T, с	T_{opt}, с	S
25	25×12	773,5	660,5	1,171083
50	50×12	398,3	336,4	1,184007
100	100×12	216,8	180,2	1,203108
150	150×12	180,5	155,1	1,163765
200	200×12	172,4	148,3	1,162508
400	400×12	170	145,2	1,170799

На основании собранных данных видно, что оптимизация обменов данными дает практически постоянное улучшение производительности на 17–20 %. Максимум приходится на 100 узлов (1200 параллельных процессов). С этого же числа узлов падает масштабируемость задачи, так как размерность задачи ограничена оперативной памятью и задачу большей размерности для сохранения масштабируемости рассчитать при такой конфигурации кластера нельзя. Время выполнения рассматриваемой задачи в пакете QBox после 100 узлов сходится к постоянному значению (рис. 12). Тесты на большем числе процессов (до 4800 процессов на 400 узлах) показывают примерно одинаковый результат как по времени выполнения, так и по ускорению за счет оптимизации отображения процессов.

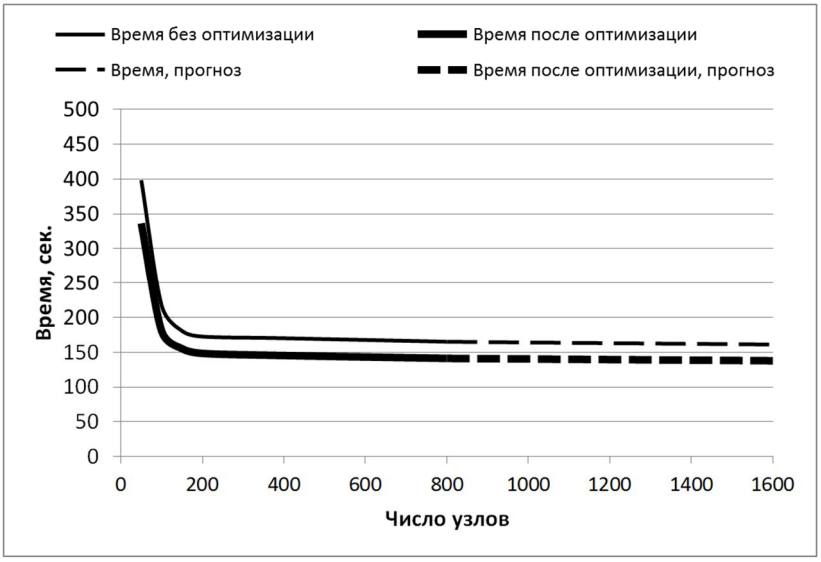


Рис. 12. Время выполнения QBox до и после оптимизации

В рассматриваемом случае сильного масштабирования время пересылок на несколько порядков меньше по сравнению со временем расчета одной итерации, поэтому, экстраполируя данные на большее число процессов, можно сделать предположение о том, что время выполнения MPI-программы будет сохраняться постоянным, так же, как и эффект от оптимизации, для достаточно большого числа процессов.

При существенном увеличении числа процессов время расчета одной итерации будет приближаться ко времени выполнения пересылки данных, что приведет к увеличению влияния пересылок. При этом время выполнения параллельной программы будет расти, а эффект от оптимизации, предположительно, будет усиливаться из-за более существенного влияния времени пересылок на общее время выполнения. Последний сценарий не используется на практике, так как предполагает неэффективное использование ресурсов.

Рассмотренный вычислительный пакет CORAL QBox хорошо масштабируется до 100 узлов. На большем числе узлов эффективность параллельного расчета снижается, так как размерность задачи недостаточно большая и время обмена данными сопоставимо с расчетным временем. Для задач с большей размерностью пакет масштабируется лучше (анализ проведен в [8]), возможно получение более хороших результатов по ускорению за счет оптимизации обмена данными.

Заключение

Предложенный подход обеспечивает улучшение производительности параллельного расчета в случае, если граф задачи MPI-приложения имеет неоднородное распределение весов ребер и выраженные подграфы, вершины в которых сильно связаны между собой, а количество этих вершин близко к количеству ядер на одном узле кластера.

Положительный эффект ускорения в 2–7 раз получен для синтетического теста при тестировании на узлах с сопроцессорами Intel® Xeon Phi™ и процессорами Intel® Xeon®. Основные факторы, влияющие на эффективность метода: неоднородность взаимодействия процессов, частота обмена данными и объемы пересылаемых данных.

Положительный эффект ускорения открытого пакета CORAL QBox на 17–20 % получен предложенным методом для числа процессов до 4800.

Предложенный метод не дает ощутимого эффекта в случае плохо масштабируемого приложения, задач с равномерным обменом между параллельными процессами и задач с малой интенсивностью обменов.

Экспериментально установлено, что использование подхода при неполной загрузке узлов кластера дает отрицательный эффект из-за сопоставимой в этом случае скорости передачи данных через интерконнект между узлами кластера и через общую память на каждом узле.

Метод отображения задачи можно усовершенствовать, используя физические параметры интерконнекта (пропускную способность и задержки канала связи), а также учитывая полную топологию кластера.

Литература

1. Копысов, С.П. Методы привязки параллельных процессов и потоков к многоядерным узлам вычислительных систем / С.П. Копысов, А.К. Новиков, Л.Е. Тонков

и др. // Вестн. Удмуртск. ун-та. Матем. Мех. Компьют. науки. — 2010. — Вып. 1. — С. 123–132.

2. Курносов, М.Г. Назначение ветвей параллельной программы на процессорные ядра распределенной вычислительной системы / М.Г. Курносов // Материалы Межд. научно-технической конференции «Многопроцессорные вычислительные и управляющие системы» (пос. Дивноморское, Геленджик, 2007). — Таганрог: ТТИ ЮФУ, 2007. — Т. 1. — С. 227–231.
3. Karlsson, C. Optimizing Process-to-Core Mappings for Application Level Multi-dimensional MPI Communications / C. Karlsson, T. Davies, Z. Chen // Cluster Computing (CLUSTER), 2012 IEEE International Conf. Proceedings (Beijing, China, September, 24–28, 2012). — Beijing, 2012. — P. 486–494. DOI: 10.1109/cluster.2012.47
4. Zhang, J. Process Mapping for MPI Collective Communications / J. Zhang, J. Zhai, W. Chen, et al. // Lecture Notes in Computer Science. — 2009. — Vol. 5704. — P. 81–92. DOI: 10.1007/978-3-642-03869-3_11
5. Chen, H. MPIPP: an Automatic Profile-Guided Parallel Process Placement Toolset for SMP Clusters and Multiclusters / H. Chen, W. Chen, J. Huang, et al. // ICS'06 Proceedings of the 20th annual international conference on Supercomputing (Queensland, Australia, June, 28 – July, 01, 2006). — Queensland, 2006. — P. 353–360. DOI: 10.1145/1183401.1183451
6. Intel® MPI Library Reference Manual. URL: http://software.intel.com/sites/products/documentation/hpc/ics/impi/41/lin/Reference_Manual/index.htm (дата обращения: 20.12.2014).
7. Larsson, P. Shared Memory Communication vs. Infiniband / P. Larsson. URL: http://www.nsc.liu.se/~pla/blog/2013/09/12/smp-vs-infiniband (дата обращения: 20.12.2014).
8. Gygi, F. Large-Scale First-Principles Molecular Dynamics Simulations on the BlueGene/L Platform using the Qbox Code / F. Gygi, R.K. Yates, J. Lorenz, et al. // Proceedings of the ACM/IEEE SC 2005 Conference (Seattle, WA, USA, November, 12–18, 2005). — Seattle, 2005. — 24 p. DOI: 10.1109/sc.2005.40
9. Суперкомпьютер «Торнадо ЮУрГУ». URL: http://supercomputer.susu.ac.ru/computers/tornado (дата обращения: 20.12.2014).

Гетманский Виктор Викторович, к.т.н., младший научный сотрудник кафедры высшей математики, Волгоградский государственный технический университет (Волгоград, Российская Федерация), victor.getmanski@gmail.com.

Чалышев Владимир Сергеевич, ООО «Сингулярис Лаб» (Волгоград, Российская Федерация), cmdsoft@gmail.com.

Крыжановский Дмитрий Иванович, к.т.н., ООО «Сингулярис Лаб» (Волгоград, Российская Федерация), dmitry.kryzhanovsky@singularis-lab.com.

Лексиков Евгений Иванович, Intel (Нижний Новгород, Российская Федерация), evgeny.leksikov@intel.com.

Поступила в редакцию 19 марта 2015 г.

Bulletin of the South Ural State University
Series "Computational Mathematics and Software Engineering"
2015, vol. 4, no. 2, pp. 5–19

DOI: 10.14529/cmse150201

OPTIMIZING PROCESSES MAPPING FOR TASKS WITH NON-UNIFORM DATA EXCHANGE RUN ON CLUSTER WITH DIFFERENT INTERCONNECTS

V.V. Getmanskiy, Volgograd State Technical University (Volgograd, Russian Federation) victor.getmanski@gmail.com,
V.S. Chalyshev, Singularis Lab Ltd. (Volgograd, Russian Federation) cmdsoft@gmail.com,
D.I. Kryzhanovskiy, Singularis Lab Ltd. (Volgograd, Russian Federation) dmitry.kryzhanovsky@singularis-lab.com,
E.I. Leksikov, Intel (Nizhniy Novgorod, Russian Federation) evgeny.leksikov@intel.com

The problem of mapping the parallel task to the nodes of computing cluster is considered. MPI software with non-uniform communication and heterogeneous interconnect of computing cluster require to appropriate parallel processes mapping for optimization of data exchange. The graph mapping algorithm is developed. It uses parallel program representation as a task graph and cluster topology representation as system graph. The proposed optimization technique is tested on synthetic benchmark and on real QBox software to study its efficiency on large number of computing cores. The positive results of optimization are achieved and the summary is presented in the paper. Speedup of 17–20 % is obtained on scalable benchmarks using 300–4800 parallel processes.

Keywords: task mapping, cluster, communication graph, MPI.

References

1. Kopysov S.P., Novikov A.K., Tonkov L.E., et al. Metody privyazki parallelnyx processov i potokov k mnogoyadernym uzlam vychislitelnyx sistem [Methods of Parallel Processes and Threads Binding to Multicore Cluster Nodes]. Vestnik Udmurtskogo universiteta. Matematika. Mehanika. Kompyuternye nauki [Bulletin of Udmurt University: Mathematics, Mechanics and Computing Science]. 2010. Issue 1. P. 123–132.
2. Kurnosov M.G. Naznachenie vetvej parallelnoj programmy na processornye yadra raspredelennoj vychislitelnoj sistemy [Mapping of parallel program branches to computing cores in distributed system]. Materialy Mezhd. nauchno-texnicheskoj konferencii "Mnogoprocessornye vychislitelnye i upravlyayushhie sistemy" (pos. Divnomorskoe, Gelendzhik, 2007) [Multiprocessor computing and control systems, International Conference Proceedings (Divnomorskoe, Gelendzhik, Russia, 2007)]. P. 227–231.
3. Karlsson C., Davies T., Chen Z. Optimizing Process-to-Core Mappings for Application Level Multi-dimensional MPI Communications. Cluster Computing (CLUSTER), 2012 IEEE International Conf. Proceedings (Beijing, China, September, 24–28, 2012). P. 486–494. DOI: 10.1109/cluster.2012.47

4. Zhang J., Zhai J., Chen W., et al. Process Mapping for MPI Collective Communications. Lecture Notes in Computer Science. 2009. Vol. 5704. P. 81–92. DOI: 10.1007/978-3-642-03869-3_11
5. Chen H., Chen W., Huang J., et al. MPIPP: an Automatic Profile-Guided Parallel Process Placement Toolset for SMP Clusters and Multiclusters. ICS'06 Proceedings of the 20th annual international conference on Supercomputing (Queensland, Australia, June, 28 – July, 01, 2006). P. 353–360. DOI: 10.1145/1183401.1183451
6. Intel® MPI Library Reference Manual. URL: http://software.intel.com/sites/products/documentation/hpc/ics/impi/41/lin/Reference_Manual/index.htm (accessed: 20.12.2014).
7. Larsson P. Shared Memory Communication vs. Infiniband. URL: http://www.nsc.liu.se/~pla/blog/2013/09/12/smp-vs-infiniband (accessed: 20.12.2014).
8. Gygi F., Yates R.K., Lorenz J., et al. Large-Scale First-Principles Molecular Dynamics Simulations on the BlueGene/L Platform using the Qbox Code. Proceedings of the ACM/IEEE SC 2005 Conference (Seattle, WA, USA, November, 12–18, 2005). 24 p. DOI: 10.1109/sc.2005.40
9. Tornado SUSU Supercomputer. URL: http://supercomputer.susu.ac.ru/en/computers/tornado/ (accessed: 20.12.2014).

Received March 19, 2015.