Московский Государственный Университет

имени М.В. Ломоносова

Курс: «Параллельное программирование и суперкомпьютеры»

Параллельная сортировка Бэтчера

**Выполнил:**

**Сактаганов Нуржан (аспирант)**

**Дата подачи: 21.10.2017**

**Москва 2017**

## Постановка задачи

Дана структура (или класс)

Point{

float coord[2];

int index;

}

P[n1\*n2];

Данная структура будет использоваться для работы с регулярной сеткой. Точки данной сетки имеют координаты P[i\*n2+j].coord[0] = x(i, j), P[i\*n2+j].coord[1] = y(i,j), где i = 0, …, n1-1, j = 0, …, n2-1. Индекс определяется соотношением P[i\*n2+j].index = i\*n2+j.

На каждом процессе одинаковое количество элементов структуры Point. (Если на некоторых процессах элементов структуры Point меньше чем во всех остальных, тогда необходимо ввести фиктивные элементы, например, с отрицательным значением индекса).

Надо реализовать параллельную сортировку Бэтчера для структур Point вдоль одной из координат x (или y). То есть с начала необходимо реализовать сортировку на каждом отдельном процессе, а потом реализовать сеть слияния Бэтчера.

На выходе должны получить на каждом процессе одинаковое количество элементов структуры Point. Каждый элемент структуры Point одного процесса находиться левее по координате x ( или y) по сравнению с элементом структуры Point любого процесса с большим рангом, за исключением фиктивных элементов.

Программа должна удовлетворять следующим требованиям:

1. программа может быть гибридной: одновременно использовать технологию MPI, для обеспечения взаимодействия вычислительных узлов, и одну из двух технологий posix threads или OpenMP, для взаимодействия процессов, запущенных на ядрах процессоров.
2. программа должна демонстрировать эффективность не менее 50% от максимально возможной, на числе вычислительных ядер, не менее 128.

## Метод решения

На входе имеется распределенный массив структур из length элементов на n узлах. Далее считаем, что на каждом узле запущен только один MPI процесс. Без ограничения общности можно предположить, что на каждом узле находится одинаковое количество элементов (если это не так, то нужно дополнить глобальный массив фиктивными элементами).

Для начала, на каждом узле надо построить расписание сети сортировки Бэтчера для n элементов (с целью экономии памяти, в расписание можно добавлять только компараторы вида (x, r) и (r, x), где r-ранг текущего процесса).

На самом первом такте (параллельном) сортировки, сортируем каждый локальный массив. После чего будем производить итерации comp\_exchange, то есть обмениваться локальными массивами с удаленным узлом согласно расписанию, а потом делать слияние локального массива и полученного массива. Новым локальным массивом становится половина слитого массива. Причем, если ранг удаленного узла был выше, то берем первую половину, иначе вторую половину массива.

Когда закончится выполнение расписания, на выходе получим отсортированный распределенный массив.

В качестве последовательного алгоритма сортировки берем dh\_sort. Это гибрид между пирамидальной сортировкой и сортировки слиянием [1]. Этот алгоритм можно распараллелить по аналогии с паралельной сортировкой слиянием.

## Описание вычислительной системы

Вычислительная система представляет собой комплекс IBM Blue Gene/P. В составе имеется две стойки по 1024 четырехъядерных процессоров в каждой стойке. Пиковая производтельность — 27.9 терафлопс. Размер оперативной памяти — 2 Гб на каждый процессор. Каждое ядро модели PowerPC 450 с рабочей частотой 850 МГц, с 32 битной адресацией. Производительность ядра — 3.4 Гфлопс.

Описание коммуникационной сети:

* сеть общего назначения, объединяющие все вычислительные узлы; предназначена для операций типа «точка-точка»
* вычислительный узел имеет двунаправленные связи с шестью соседями
* пропускная способность каждого соединения — 425 MB/s (5,1 GB/s для всех 12 каналов)
* латентность (ближайший сосед):
  + 32-байтный пакет: 0,1 μs
  + 256-байтный пакет: 0,8 μs

## Оценка времени выполнения параллельного алгоритма

Для упрощения модели будем считать, что:

1. все обмены не влияют друг на друга;
2. везде одинаковая латентность;
3. везде одинаквая пропускная способность;
4. между узлами есть двухканальная связь.

Тогда итоговое время выполнения будет состоять из трех основных составляющих T = T1 + T2 + T3:

* T1 — время начальной сортировки локальных массивов;
* T2 — время, затраченное на коммуникации (обмен локальным массивами);
* T3 — время, затраченное на слияние после каждого обмена.

Время начальной сортировки T1 можно обозначить как С\*ls\*log(ls), где C — некая константа, ls — размер локального массива. Константа C вычисляется эмпирически, по сути это время на одну операцию: C = T\_sort / NlogN, T\_sort — время сортировки N элементов последовательным алгоритмом.

Коммуникация происходит после каждого такта кроме последнего. Поэтому, T2 можно оценить как takts \* (T\_latency + T\_bandwith \* ls \* sizeof(Point)).

Время слияния T3 можно оценить как C \* takts \* ls, т. к. можно сразу слить только нужную половину.

## Анализ полученных результатов

Ниже приведена таблица времен (в секундах) запусков на различном наборе сеток и процессоров.

|  |  |  |  |
| --- | --- | --- | --- |
| Сетка | 1 процессор | 128 процессоров (1 поток или ядро) | 128 процессоров (4 потока или ядра) |
| 7500 х 10000 | 43.0559 | 1.6916 | 1.6051 |
| 80000 х 60000 | Не доступно | 119.895 | 107.794 |

На 128 процессорах сортировка Бэтчера занимает 28 параллельных тактов. Сруктруа Point занимает 12 байтов (2 float и 1 int). В первой сетке 75 млн. элементов. Во второй сетке 4.8 млрд. элементов. Соответственно по 585938 и 37.5 млн элементов на каждом узле. Тогда размеры будут соответственно 7031256 байт (7 Мб) и 450 Мб на узел.

Вычислим константу C = 43.0559 / (75 млн. \* log(75 млн.)). Получим, что C приблизительно 2.1945 на десять в минус восьмой степени.

Рассмотрим варианты с 1 потоком на процессор. T1 будет 0.2464с и 20.7055с соответственно. А T2 будет, если за латентность взять 0.8мкс, 0.4632c и 39.5304с. И, наконец, Т3 будет 0.36с и 23.043с. В итоге получим T 1.07с и 83.2789с. На самом деле, программы выполнялись 1.6916с и 119.895с. Это 63.25 и 69.46 процентов от теоретически максимальной эффективности.

Следует отметить, что на меньшей сетке index структуры Point имел тип int (32бита), а на большей — int64\_t. Следовательно, sizeof(Point) — 12 и 16 байтов соответственно. Оценка времени сделана с учетом этого.

## Список литературы

1. «Параллельные алгоритмы сортировки больших объемов данных», М.В. Якобовский, 2004г.