ИССЛЕДОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ ПОДСИСТЕМЫ ВВОДА/ВЫВОДА СУПЕРКОМПЬЮТЕРА BLUEGENE/P

С.В. Коробков

На данный момент в списке 500 наиболее мощных суперкомпьютеров довольно популярны установки архитектуры BlueGene. За 5 лет вышло 3 поколения суперкомпьютеров данной архитектуры. Система BlueGene/Q, установленная в США, занимает 2 по производительности место. В суперкомпьютерной области большое внимание уделяется вычислительным ресурсам. При построении суперкомпьютерных комплексов большую роль играет размер памяти на каждом узле, количество ядер в процессоре, производительность коммуникационных и других сетей в кластере поскольку эти параметры влияют на вычислительную способность кластера. Все вычислительные комплексы, в первую очередь, оцениваются по количеству операций проводимых в секунду. Однако при решении реальных задач важную роль также играют другие параметры вычислительного комплекса. Одним из таких важных параметров является скорость обмена данным между вычислительной подсистемой комплекса и подсистемой хранения данных. Большинство задач требуют ввода или вывода больших объемов данных. Рассмотрим какие узкие места есть в подсистеме ввода/вывода на примере системы BlueGene/P, установленной в МГУ.

Общая структура подсистемы ввода/вывода в данной установке представлена схематически на рисунке 1.

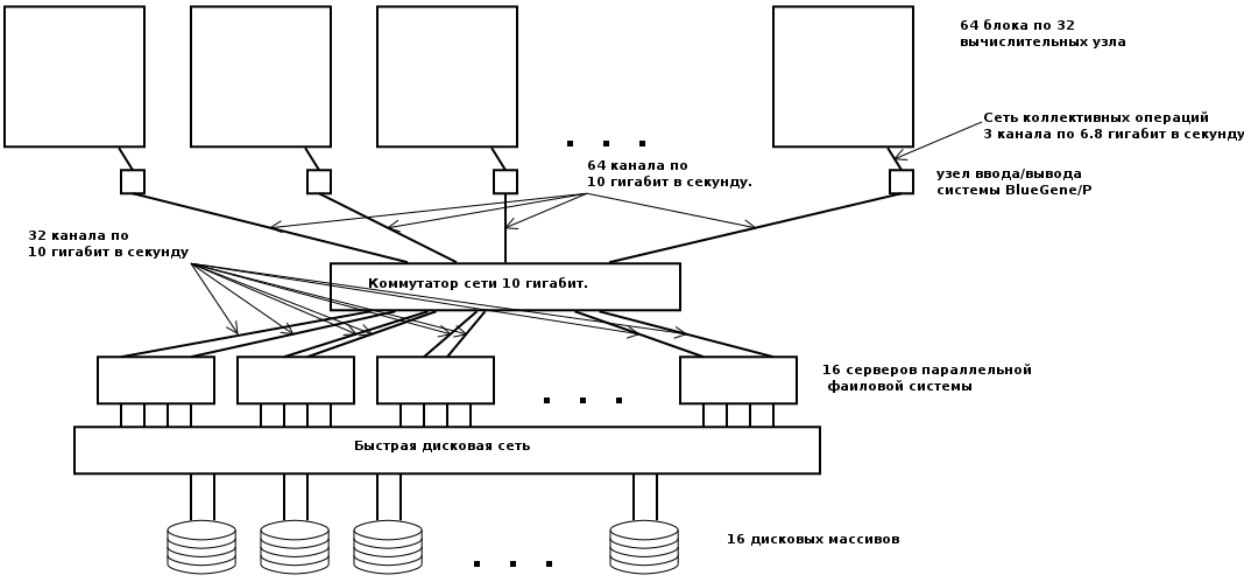

Рис. 1. Общая схема подсистемы ввода/вывода BlueGene/P установленного в МГУ

На данной схеме изображены 64 блока по 32 вычислительных узла в каждом. На каждый такой блок приходится по 1 узлу ввода/вывода, непосредственно соединенных с вычислительными узлами сетью коллективных операций. Каждый узел ввода/вывода обладает 3 соединениями с сетью коллективных операций. Скорость обмена данными по одному такому каналу 6.8 гигабит в секунду [1]. И обслуживает только свой блок вычислительных узлов. Каждый узел ввода/вывода соединен каналом 10 гигабит в секунду до коммутатора.

За операции с жесткими дисками отвечают 16 серверов параллельной файловой системы GPFS. Каждый сервер соединен 2 каналами 10 гигабит в секунду с коммутатором. Каждый сервер общается с дисковыми системами по высокопроизводительной сети.

Как видно из этой схемы, есть 3 узких места:

1. Узел ввода/вывода BlueGene/P.
2. Связь серверов GPFS с коммутатором. Её производительность ограничена 320 гигабитами в секунду по сравнению с 640 гигабитами в секунду от узлов ввода/вывода.
3. Скорость взаимодействия с жесткими дисками.

В данной статье рассматривается запись данных, как более медленная операция, чем чтение данных. Также это связано с тем, что параллельная запись организуется сложнее, чем параллельное чтение. Рассмотрим первое узкое место и оценим производительность узла ввода/вывода. На данном узле установлен клиент параллельной файловой системы GPFS. Тестирование осуществлялось записью 10 миллиардов байт данных, созданных случайным образом, и замером времени потраченного на запись этих данных. Скорость записи составила 695 мегабит в секунду. Тест передачи данных по сети производился без записи данных в файл для

исключения влияния скорости работы клиента параллельной файловой системы . В результате была получена скорость 880 мегабит в секунду. При этом, на узле ввода/вывода одно из четырех вычислительных ядер было занято на 100 процентов процессом ciodrd, остальные ядра простаивали. Этот процесс обеспечивает ввод/вывод в системе.

Эти результаты говорят о том, что ввод/вывод с одного узла не позволяет достичь большой производительности. Для проверки этого необходимо запустить ввод/вывод на нескольких вычислительных узлах из блока относящегося к одному узлу ввода/вывода. Тестирование производилось на одном блоке из 32-х узлов. Для тестирования использовалось 3 различных метода ввода/вывода данных:

1. Стандартный POSIX IO в несколько различных файлов.
2. Встроенные в MPI параллельные файловые операции(MPI IO).
3. Использование сетевых сокетов и запись в файлы уже на серверах параллельной файловой системы.

Самый простой метод вывода данных с вычислительных узлов это простая запись файлы. На каждом из узлов открывается отдельный файл для вывода данных. Результаты для различного количества узлов, выводящих данные, показаны на рисунке 2. Каждый из узлов выводит 10 миллиардов байт параллельно с остальными узлами. Из графика видно, что в случае более 3 выводящих узлов скорость вывода растет незначительно.

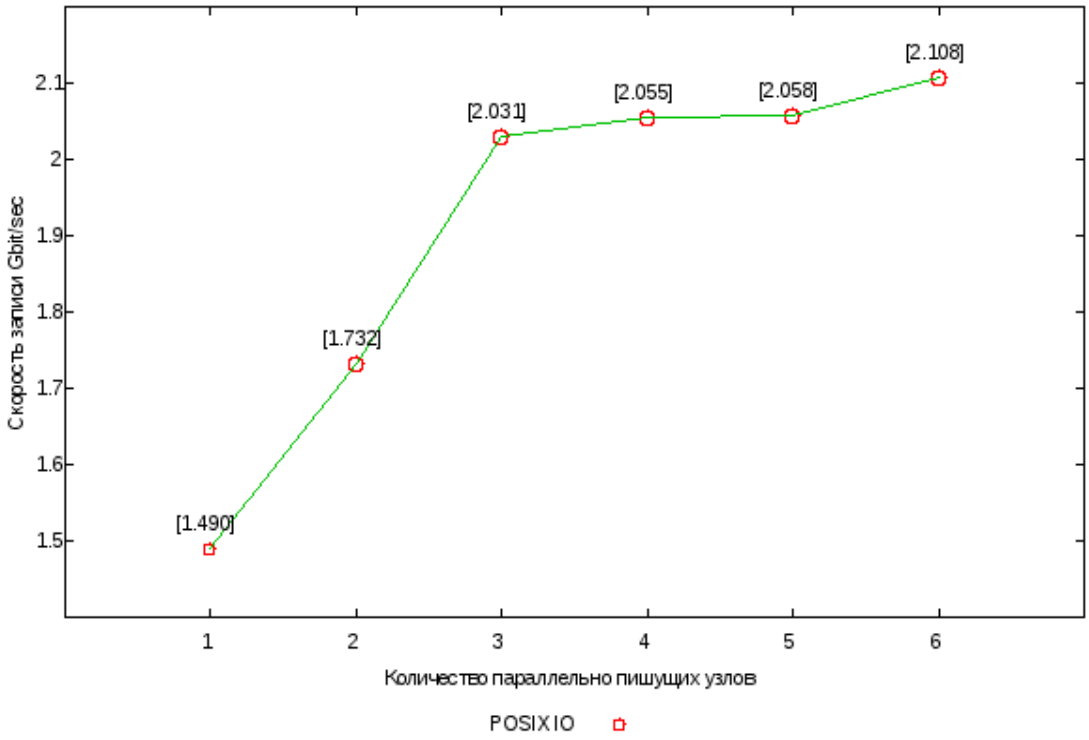

Рис. 2. Скорость записи в зависимости от количества пишущих узлов для POSIX IO

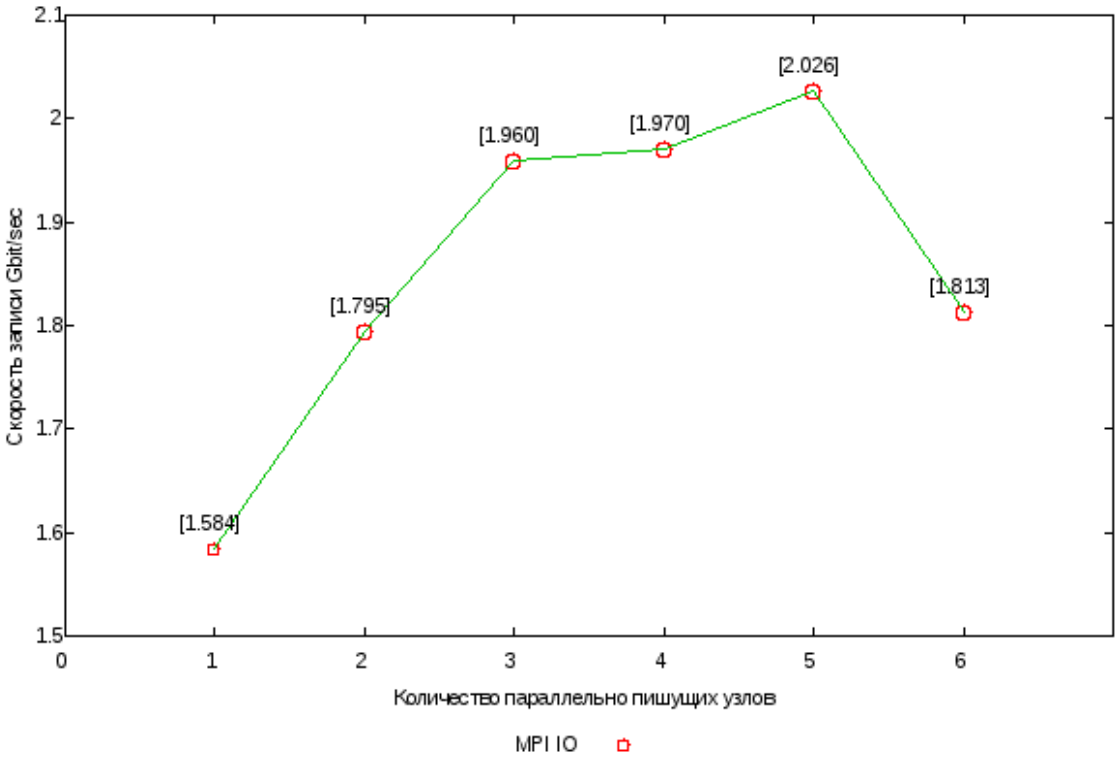

Рис. 3. Скорость записи в зависимости от количества пишущих узлов для MPI IO

Более правильно использовать для параллельного вывода данных функции MPI IO. Каждый из узлов выводил по 10 миллиардов байт параллельно в отведенную область одного файла. Результаты для различного числа выводящих узлов представлены на рисунке 3. Как видно здесь, аналогично предыдущему случаю, производительность вывода данных практически перестает расти при использовании 4-х и более узлов и начинает падать при 6 узлах.

При использовании POSIX IO или MPI IO загрузка узла ввода/вывода практически полная. При этом одно ядро занимает процесс управления ввода/вывода, клиент параллельной файловой системы использует ещё одно ядро и еще половину производительности третьего. Остальная производительность делится поровну между процессами ввода/вывода по одному на каждый вычислительный узел осуществляющий вывод данных.

В связи с тем, что клиент параллельной файловой системы использует около половины производительности узла вводы/вывода, целесообразно использовать параллельную файловую систему на отдельных серверах. Для подтверждения были проведены тесты с использованием простых сетевых сокетов для передачи данных и их последующей записи. Передача данных производилась на узлы параллельной файловой системы и запись производилась там. Результаты тестов приведены на рисунке 4.

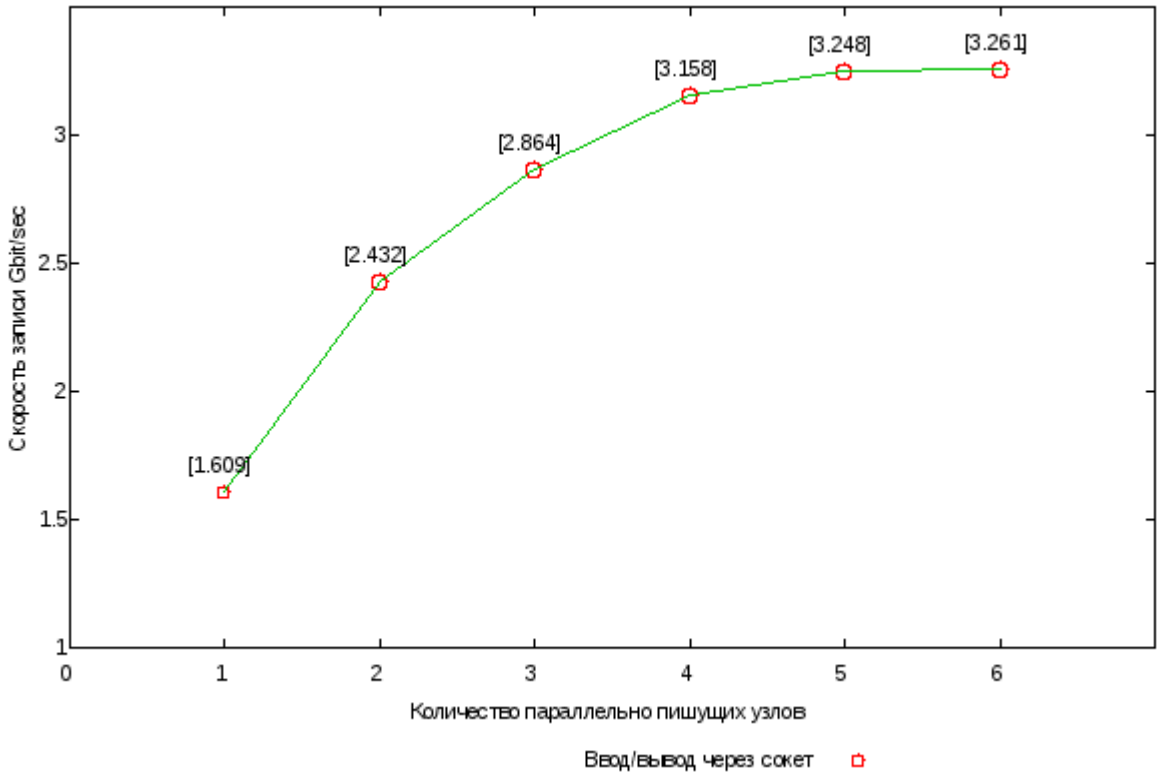

Рис. 4. Скорость записи в зависимости от количества пишущих узлов для ввода/вывода через сокет

Как видно при использовании сокетов можно достичь большей производительности ввода/вывода. Основным узким местом при вводе/выводе является узел ввода вывода BlueGene/P. Но недостаток в производительности этого узла ввода/вывода можно компенсировать распределением нагрузки по обработке потоков ввода/вывода на вычислительные узлы и узлы параллельной файловой системы. В результате, при оптимальном использовании ресурсов, можно достичь ускорения ввода/вывода в полтора раза. Оптимальной является скорость в 3.2 гигабита в секунду на один узел ввода вывода BlueGene/P.

ЛИТЕРАТУРА:

1. Гуляев А.В., Гуляев Д.А., Гуревич Е.И. и др. Система Blue Gene/P факультета ВМК МГУ имени М.В. Ломоносова: конфигурация, эксплуатация и обзор задач // Сборник трудов международной суперкомпьютерной конференции «Научный сервис в сети Интернет: суперкомпьютерные центры и задачи». Новороссийск, Россия: М.: Изд-во МГУ, 20–25 сентября 2010. С. 373–378.