СИСТЕМА BLUE GENE/P ФАКУЛЬТЕТА ВМК МГУ ИМЕНИ М.В. ЛОМОНОСОВА: КОНФИГУРАЦИЯ, ЭКСПЛУАТАЦИЯ И ОБЗОР ЗАДАЧ

А.В. Гуляев, Д.А. Гуляев, Е.И. Гуревич, Г.К. Митрохин, А.В. Позднеев, В.В. Ситник

С 2008 года на факультете ВМК МГУ имени М.В. Ломоносова работает суперкомпьютер IBM Blue Gene/P, который является одной из первых систем данной серии среди установленных в мире. Архитектура Blue Gene была предложена компанией IBM в рамках проекта по исследованию возможностей достижения принципиально новых рубежей в супервычислениях. Более крупные машины данной серии в настоящее время занимают лидирующие позиции в списке пятисот самых мощных компьютеров мира Top500, а машина Blue Gene/P, установленная на ВМК МГУ, была внесена в этот рейтинг в ноябре 2008 года и оказалась в нем на 128-м месте. Самой высокой позицией в списке самых высокопроизводительных компьютеров стран СНГ для системы Blue Gene/P Московского университета была 4-я строчка. На тот момент на территории СНГ это была самая мощная система по числу процессоров (2048 четырехъядерных), работающих в рамках одной установки.

Система Blue Gene/P принадлежит к новому семейству суперкомпьютеров, обладающих высокой производительностью, масштабируемостью, возможностью обрабатывать данные большего объема, потребляя при этом значительно меньше энергии и занимая меньшую площадь по сравнению с предыдущими системами. В списке Green 500 самых энергоэффективных компьютеров мира система Blue Gene/P факультета ВМК МГУ неизменно занимала лидирующие позиции, опережая по показателю производительности на ватт потребляемой энергии большинство суперкомпьютеров, основанных на конкурирующих решениях.

Суперкомпьютерные системы на факультете ВМК МГУ давно и активно используются как в научном, так и в образовательном процессе. В течение долгого времени основным инструментом была 16-процессорная вычислительная система с общей памятью IBM eServer pSeries 690 Regatta [1]. В настоящее время в качестве такового выступает система Blue Gene/P, описанию и опыту эксплуатации которой посвящено данное сообщение.

Описание вычислительного комплекса

Рис. 1. Вычислительные стойки суперкомпьютера IBM Blue Gene/P.

На факультете ВМК МГУ представлена конфигурация, состоящая из двух стоек (рис. 1), содержащих в общей сложности 2048 вычислительных узлов (рис. 2), каждый из которых включает четыре ядра PowerPC 450,

работающих на частоте 850 мегагерц, что дает пиковую производительность 27,8 триллионов операций с плавающей точкой в секунду [2, 3].

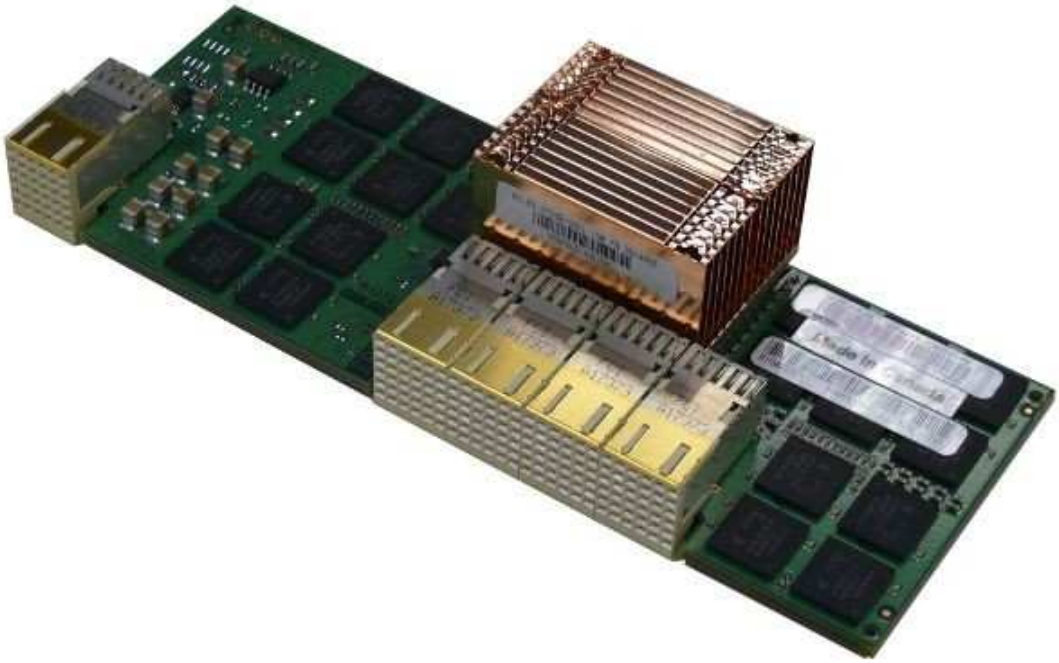

Рис. 2. Вычислительный узел суперкомпьютера IBM Blue Gene/P. (Фото: А.П. Григорьев.)

Архитектура Blue Gene спроектирована для вычислительных кодов, которые хорошо масштабируются до сотен и тысяч процессоров. В рамках этой архитектуры для межпроцессорных обменов и глобальных операций существенно наличие отдельных коммуникационных сетей, причем предназначенная для операций «точка-точка» сеть общего назначения, объединяющая все вычислительные узлы, представляет из себя трехмерный тор. Несмотря на то, что индивидуальные процессорные ядра системы Blue Gene работают на относительно низкой тактовой частоте, для приложений, способных эффективно использовать большое число процессорных элементов, удается достигнуть значительно более высокой производительности по сравнению с традиционными суперкомпьютерами.

Текущая конфигурация установленного на ВМК МГУ Blue Gene/P включает 16 карт ввода-вывода, таким образом минимальный доступный пользователю раздел состоит из 128 процессорных узлов. Вообще же, пользователи могут запускать задания на 128, 256, 512, 1024 и 2048 вычислительных узлов, причем, начиная с 512 процессоров коммуникационная сеть представляет собой уже не трехмерную решетку, а трехмерный тор.

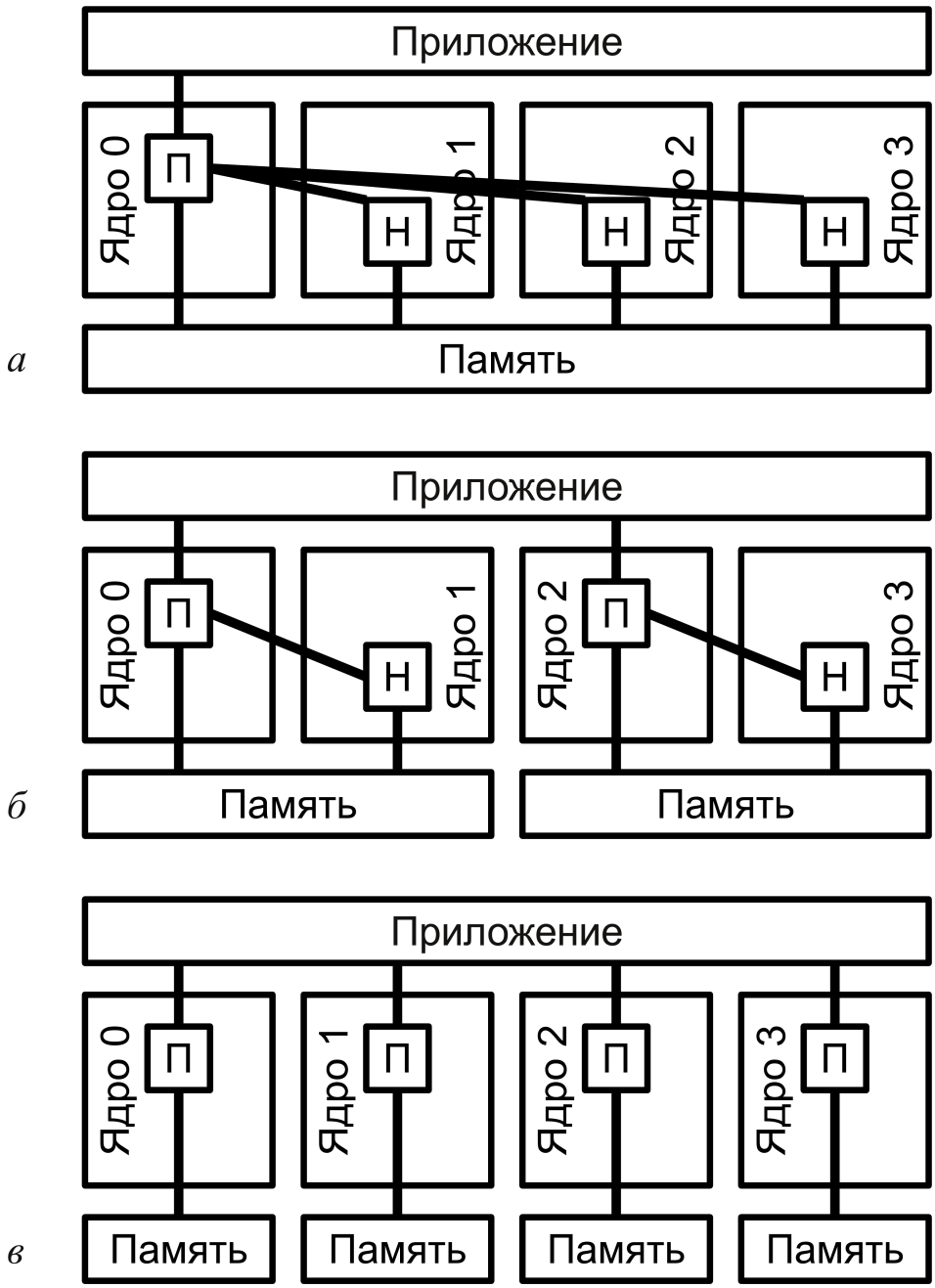

Рис. 3. Режимы исполнения процессов на вычислительных узлах системы Blue Gene/P: режим симметричного мультипроцессора SMP (а), режим двухъядерных вычислительных узлов DUAL (б), режим виртуальных вычислительных узлов VN (в). Процессы MPI обозначены квадратами с буквой «П», легковесные нити OpenMP или pthreads — квадратами с буквой «Н».

На вычислительных узлах системы Blue Gene/P возможны три режима исполнения процессов: режим симметричного мультипроцессора (symmetrical multiprocessing, SMP) — рис. 3, *а*; режим двухъядерных вычислительных узлов (dual node mode, DUAL) — рис. 3, *б*; режим виртуальных вычислительных узлов (virtual node mode, VN) — рис. 3, *в*. В режиме SMP на каждом вычислительном узле выполняется один MPI-процесс, который может породить до трех дополнительных нитей с помощью механизмов OpenMP или pthreads. В режиме DUAL на вычислительном узле запущено два процесса MPI, которые могут породить еще по одной OpenMP- или pthreads-нити. В режиме VN на каждом вычислительном узле запущено четыре MPI-процесса, и порождение дополнительных нитей невозможно. Таким образом, в любом из режимов на каждом из четырех ядер процессорного узла запущено не более одной нити. Выбор режима исполнения определяет пользователь, исходя из специфики задачи, требований по памяти и использованного механизма распараллеливания. В таблице 1 сведены данные о количестве процессорных ядер, которые будут задействованы процессами, в зависимости от числа заказанных вычислительных узлов и выбранного режима выполнения задачи при запуске

программы, распараллеленной с использованием механизма MPI, на системе Blue Gene/P, установленной на факультете ВМК МГУ.

Таблица 2. Число процессов MPI, создаваемых на системе Blue Gene/P для различного числа процессорных узлов в зависимости от режима исполнения.

Режим исполнения	**Число процессорных узлов**				
	128	**256**	**512**	**1024**	**2048**
SMP	128	256	512	1024	2048
DUAL	256	512	1024	2048	4096
VN	512	1024	2048	4096	8192

С точки зрения конечного пользователя файловая система вычислительного комплекса состоит по сути из двух частей: /home и /gpfs. /home — домашние директории, предназначенные для хранения пользовательских данных, разработки и компиляции программ, постобработки результатов расчетов; узлы ввода-вывода не имеют доступа к этой части файловой системы. /gpfs — высокопроизводительная часть файловой системы, к которой имеют непосредственный доступ узлы ввода-вывода; служит для хранения временных файлов, необходимых для счета (исполняемых файлов и данных).

Окружение Blue Gene/P включает фронтэнд, сервисный узел и систему управления распределенной файловой системой IBM General Parallel File System (GPFS). Фронтэнд — система, открытая для доступа по протоколу SSH; служит для доступа пользователей на вычислительный комплекс; вся связь с комплексом осуществляется только через эту машину; предназначена для разработки пользователями программ, компилирования проектов и постановки задач в очередь; работа с ней осуществляется в интерактивном режиме. Сервисный узел обеспечивает контроль над системой Blue Gene/P; к этой машине пользовательского доступа. Фронтэнд и сервисный узел — это серверы IBM pSeries 55A, за управление GPFS отвечают два сервера IBM pSeries 510. Все серверы построены на базе процессоров POWER5+ и работают под 64-битной версией операционной системы SUSE Linux Enterprise Server 10.

Для коммутации оптических линий служит высокопроизводительный 10-гигабитный свитч Cisco Catalyst 6509 с 24 портами: 16 портов используются для подключения узлов ввода-вывода вычислительной системы, к шести портам подключены фронтэнды, сервисные узлы и GPFS-сервера, два порта зарезервированы для будущего использования. Гигабитный Ethernet скоммутирован на 48-портовый свитч Cisco. Высокопроизводительная часть файловой системы, к которой имеют непосредственный доступ узлы ввода-вывода, находится на хранилище IBM DS4700.

Системное программное обеспечение

В состав системного программного обеспечения входят компиляторы семейства IBM XL версии «Advanced Edition for Blue Gene/P». Также доступны компиляторы GCC. На системе установлены математические библиотеки Engineering and Scientific Subroutine Library (ESSL) и Mathematical Acceleration Subsystem (MASS).

Так как фронтэнды построены на базе процессоров POWER5+, а вычислительные узлы — на базе процессоров PowerPC 450, то для сборки программ используется механизм кросс-компиляции. Для сборки программ, использующих OpenMP- или pthreads-нити, доступны потокобезопасные версии компиляторов. Для отлаженных кодов можно применять быструю версию MPI, в которой полностью отключена проверка ошибок.

В реализации MPI на Blue Gene/P доступны три дополнительные функции, расширяющие функциональность механизма передачи сообщений с учетом особенностей аппаратуры системы. Они используются в рамках коммуникаций по сети трехмерного тора и упрощают отображение процессов на наборы процессоров. Имеется функция создания четырехмерного декартового коммуникатора и функции объединения процессов, принадлежащих одному и разным наборам процессоров. Под набором процессоров подразумевается множество процессоров, совместно использующих один и тот же узел ввода-вывода.

Для тонкого измерения эффективности программ пользователи могут обращаться к внутреннему счетчику производительности (Universal Performance Counter (UPC) Unit) вычислительного узла суперкомпьютера через системные программные интерфейсы (system programming interfaces, SPIs).

За управление задачами отвечает планировщик IBM Tivoli Workload Scheduler LoadLeveler. Помимо стандартных команд планировщика LoadLeveler, разработаны или адаптированы скрипты, информирующие о состоянии очереди и размещении задач по разделам системы. Для постановки задач на счет можно использовать как командные файлы, так и применять специально разработанную утилиту, позволяющую указать все параметры запуска в качестве аргументов командной строки.

Вычислительная система активно используется в образовательном процессе. Она применяется как в рамках спецкурсов факультета ВМК и кафедральных практикумов, так и в межфакультетских учебных курсах. На время аудиторных занятий и для подготовки домашних занятий средствами планировщика LoadLeveler для

студентов создаются так называемые «бронирования» (reservations) — разделы, выделенные эксклюзивно для студентов практикума.

Эксплуатация системы и поддержка пользователей

Для обеспечения функционирования вычислительного комплекса на факультете создана группа поддержки пользователей высокопроизводительных систем факультета ВМК МГУ под руководством заместителя декана по информационным технологиям доцента А.В. Гуляева. Техническая поддержка пользователей осуществляется посредством электронной почты и через веб-сайт. Адаптированная для русскоязычной аудитории оригинальная документация компании IBM, учитывающая специфичные для установленной на ВМК МГУ системы Blue Gene/P настройки, представлена на веб-сайте «hpc@cmc» [2]. Процесс получения новым пользователем учетной записи на вычислительном комплексе включает заполнение электронной заявки на веб-сайте и предоставление письменного заявления.

Мониторинг системы, диагностика и подготовка ее к замене оборудования осуществляется преимущественно через веб-интерфейс «Blue Gene Navigator» [4]. В его рамках доступна текущая загрузка системы, полная история задач, история замены комплектующих, текущая температура на каждом из вычислительных узлов. Диагностика проводится на разделах, включающих 512 процессоров; при этом не затрагиваются пользовательские задачи, запущенные на других разделах, и не нарушается нормальная работа планировщика. В рамках сервисных действий можно отключить, подготовить к замене, а затем включить любой элемент оборудования.

Обзор задач пользователей

Сделаем краткий обзор некоторых задач прикладного и образовательного характера, которые были решены пользователями системы за время, прошедшее с момента ее установки.

Структурная стабильность ультрананокристаллов алмаза. Работа выполнена Д.В. Чачковым, Д.С. Тарасовым и Р. Фрейтасом-мл. с использованием программного пакета CPMD. В ее рамках с беспрецедентной точностью проанализирована стабильность ультрананокристаллов алмазов, содержащих более 1500 атомов. Ожидается, что моделируемый материал найдет свое применение в микроэлектромеханических устройствах и наномеханических системах, в том числе механических вычислительных устройствах для наноробототехники.

Численное моделирование фарлей-бунемановской неустойчивости. В работе, выполненной Д.В. Ковалёвым под руководством А.П. Смирнова, исследуются спектральные характеристики и процесс развития плазменных неустойчивостей в Е-слое ионосферы Земли, которые фиксируются радарами и приводят к помехам в радиоэфире; по результатам этой работы Д.В. Ковалёвым была защищена диссертация на соискание ученой степени кандидата физико-математических наук.

Структурно-функциональное сравнение белков на основе анализа их энергетических ландшафтов. Под руководством П.С. Иванова студенты кафедры биофизики физического факультета МГУ имени М.В. Ломоносова И.В. Оферкин и М.Г. Годзи выполнили работу, в которой был предложен первый универсальный алгоритм функционального сравнения белковых структур из банка данных и выполнена его программная реализация. Работа важна с точки зрения развития методов генетической диагностики наследственных и приобретенных заболеваний, ускорения и удешевления процесса синтеза новых лекарственных форм.

Группа профессора В.А. Крюкова выполнила настройку и адаптацию для суперкомпьютера Blue Gene/P системы программирования DVM. На широком классе тестов ими было проведено исследование эффективности распараллеливания DVM-программ и осуществлено сравнение с результатами, полученными на суперкомпьютере МВС-100к, установленном в МСЦ РАН.

В рамках работы над кандидатской диссертацией на тему *«Алгоритмы ветвей и границ в распределенных вычислительных системах»* В.С. Махнычёв с помощью метода ретроанализа и алгоритма асинхронной обработки данных занимается расчетом эндшпильных таблиц при идеальной игре сторон в шахматах.

Под руководством академика Ю.Г. Евтушенко на системе Blue Gene/P проводится экспериментальная оценка трудоемкости задач криптоанализа с целью исследования потенциальной применимости новых подходов к решению задач поиска секретного ключа.

Два исследования на Blue Gene/P под руководством Н.Н. Поповой проводит В.Ю. Воронов. Первое из них посвящено разработке программного обеспечения для моделирования электрических цепей большой размерности, второе — созданию программной системы для автоматического построения параллельных программ на основе методов машинного обучения. Результаты этих исследований вошли в защищенную им диссертацию, представленную на соискание ученой степени кандидата физико-математических наук.

Заключение

Система Blue Gene/P, установленная на факультете ВМК МГУ имени М.В. Ломоносова, уже в течение двух лет обеспечивает сотрудников и студентов МГУ вычислительными мощностями мирового уровня. За это

время пользователями системы решено значительное число задач прикладного и образовательного характера. Более чем у 150 студентов прошли практикумы по параллельным вычислениям: как потоковые и кафедральные, так и межфакультетские. В рамках суперкомпьютерных практикумов студенты занимались разработкой и модификацией параллельных вычислительных методов линейной алгебры, решали задачи математической и вычислительной физики, исследовали эффективность параллельных алгоритмов. По результатам проведения расчетов на Blue Gene/P, пользователями системы защищено пять кандидатских диссертаций. Всего с начала эксплуатации системой воспользовалось более 270 пользователей, запустив на счет более 77 тысяч задач. Среднее число процессоров, применявшихся в одной задаче, составило 175 вычислительных узлов (700 ядер).

Из характерных особенностей системы, позволивших пользователям добиться высокой масштабируемости приложений, можно выделить следующие: значительное число процессорных узлов, их многоядерность, коммуникационную сеть трехмерного тора, сеть глобальных коллективных операций, совокупно большую вычислительную мощность системы Blue Gene/P. Отметим наличие нескольких режимов исполнения процессов, что позволяет разрабатывать гибридные программы. Система проста в обслуживании и администрировании.

Ближайшие планы развития состоят в следующем. Планируется увеличить объем дискового пространства высокопроизводительной части файловой системы, что важно для возможности создания контрольных точек работающих задач и сохранения большого объема промежуточных данных между расчетами. Также планируется активировать еще 16 узлов ввода-вывода (что удвоит их общее число). Ожидается, что это позволит повысить производительность ввода-вывода и даст возможность запускать большее число задач меньшего размера, что имеет принципиальное значение при организации проведения студенческих практикумов.

ЛИТЕРАТУРА:

1. «Вычислительная система Regatta». — http://www.regatta.cmc.msu.ru.
2. «hpc@cmc — высокопроизводительные вычисления на ВМК МГУ». — http://hpc.cmc.msu.ru.
3. C. Sosa, B. Knudson «IBM System Blue Gene Solution: Blue Gene/P Application Development». — 4th ed. — ibm.com/redbooks, 2009. — SG24-7287-03.
4. G. Lakner «IBM System Blue Gene Solution: Blue Gene/P System Administration». — 4th ed. — ibm.com/redbooks, 2009. — SG24-7417-03.