

*Посвящается основателям вычислительного дела
в Сибирском отделении РАН
академикам Г. И. Марчуку и А. С. Алексееву*

СИБИРСКИЙ СУПЕРКОМПЬЮТЕРНЫЙ ЦЕНТР КОЛЛЕКТИВНОГО ПОЛЬЗОВАНИЯ: этапы развития, текущее состояние и перспективы

Б. М. Глинский

Институт вычислительной математики и математической геофизики
630090, Новосибирск, Россия

УДК 681.32

Описаны основные этапы развития центров коллективного пользования в СО РАН, их роль в решении больших задач математического моделирования, обучении современным суперкомпьютерным технологиям. Рассматриваются текущая структура и задачи Центра коллективного пользования "Сибирский Суперкомпьютерный Центр" (ЦКП ССКЦ), основные результаты работы центра в 2012 г. и перспективы его развития.

Ключевые слова: центр коллективного пользования, численное моделирование, супервычисления.

The main stages of development of multi-access computer centers at the Siberian Branch of the Russian Academy of Sciences, their role in solving big-size problems of mathematical modeling as well as teaching modern supercomputer technologies are considered. The current structure and tasks of the SMASCC, its basic results obtained in 2012 and prospects of its development are discussed.

Keywords: multi-access computer center, numerical modeling, supercomputer technologies.

Введение. С момента образования Сибирского отделения АН СССР академики М. А. Лаврентьев и Г. И. Марчук уделяли большое внимание развитию вычислительной техники и методов математического моделирования для различных наук [1]. Организация Вычислительного центра (ВЦ) в мае 1963 г. в составе СО АН стала ключевым моментом в создании вычислительного центра коллективного пользования (ВЦКП). Машинный парк ВЦ, оснащенный передовой отечественной вычислительной техникой, за несколько лет стал одним из самых мощных в стране.

В 1975 г. в Новосибирском Научном Центре (ННЦ) была организована первая сеть дистанционного доступа институтов к мощным по тем временам ЭВМ, расположенным в ВЦ. Тогда по телефонным каналам городской телефонной станции с помощью самодельных модемов была

реализована первая сетевая система коллективного пользования тремя машинами БЭСМ-6 с объединенной памятью.

В 1979 г. был разработан проект создания корпоративной скоростной кабельной сети, связывающей практически все институты ННЦ с базовыми вычислительными комплексами БЭСМ-6 (3 шт.) и тремя ЕС 1060, расположенными в ВЦ (проект ВЦКП). Реализация этой корпоративной сети СО АН с траншейной прокладкой кабелей между институтами сыграла важную роль при переходе к волоконно-оптическим системам связи между ЭВМ и развитию Интернета на скоростных оптических каналах [2].

Осуществление другого важного проекта в области развития вычислительных средств ВЦ началось в 1987 г. Он был направлен на создание многопроцессорного вычислительного комплекса "Сибирь" на базе средств ЕС ЭВМ: управляющего вычислителя ЕС-1068.17 и восьми сопроцессоров ЕС-2706. Пиковая производительность комплекса достигала 100 MFlops, что весьма немало для того времени. Для комплекса было разработано все необходимое системное параллельное программное обеспечение, включая расширение штатной операционной системы и систему параллельного программирования. Комплекс был ориентирован в первую очередь на обработку сейсмических данных. Примерно 30 комплексов были поставлены в промышленность. Именно на "Сибири", в частности, впервые в мировой практике у нас в Новосибирске, выполнили распараллеливание трехмерного метода частиц и провели первые численные эксперименты. В ходе работ по проекту "Сибирь" была осознана необходимость разработки нового поколения системного параллельного программного обеспечения. Уже в языке и системе параллельного программирования ИНЯ (расширение Фортрана (1989 г.)) было заложено автоматическое обеспечение в разрабатываемых прикладных программах, свойства динамической настройки на все доступные вычислительные ресурсы мультимикрокомпьютера.

На этих технических и программных ресурсах был создан Главный производственный Вычислительный Центр (ГПВЦ), предоставлявший вычислительные услуги институтам СО РАН.

Интенсивную работу по созданию вычислительной базы СО АН СССР и развитию сетевых технологий математического моделирования выполняли также Вычислительные центры в Красноярске (ВЦК) и Иркутске (ВЦИ). Эти ВЦ были первыми в Сибири институтами в области информационно-вычислительных технологий [3].

В настоящее время работу центров координирует Научный совет по супервычислениям при Президиуме СО РАН под председательством академика Б. Г. Михайленко.

В середине 1980-х успешно работали ВЦКП и Сибирский сегмент "Академсети". Эти системы обеспечивали режимы обмена информацией, совместного счета, электронной почты между институтами Новосибирска, Красноярска и Иркутска.

К сожалению, в 1990-х годах все это было демонтировано из-за быстрого роста стоимости электроэнергии и аренды международных каналов связи. Главный ГПВЦ был практически ликвидирован, и на его базе был создан Институт вычислительных технологий.

Однако потребность в высокопроизводительных вычислениях для решения больших задач математического моделирования осталась, следовательно, нужны были мощные вычислительные ресурсы.

В 2000-е годы ключевую роль в восстановлении ВЦ КП сыграл академик А. С. Алексеев. Осталась кабельная сеть ВЦ КП, и достаточно было установить в ИВМиМГ (бывший ВЦ) СО РАН компьютер Silicon Graphics с большой оперативной памятью и выходом в сеть, как в течение одного года 15 институтов СО РАН подключились к этой сети.

По существу, благодаря усилиям А. С. Алексеева при ИВМиМГ началось возрождение ВЦ КП под названием "Сибирский суперкомпьютерный центр коллективного пользования СО РАН" (ЦКП ССКЦ) [4].

На начальной стадии создания ССКЦ (первые четыре года) важную организационную роль сыграл доктор физико-математических наук Г. Н. Ерохин, собравший квалифицированный коллектив инженеров и программистов-системщиков, трудом которых и был создан ССКЦ КП для ННЦ.

В 2000 г. Президиум СО РАН выделил финансирование на создание кластера МВС 1000М с 32 процессорами. Кластер создавался сотрудниками института на основе разработки Межведомственного Суперкомпьютерного центра и к концу 2003 г. достиг проектной мощности. Формально ССКЦ КП был создан в 2001 г. по Постановлению Президиума СО РАН № 100 от 6 марта 2001 г. "О создании Сибирского Суперкомпьютерного центра коллективного пользования СО РАН " на базе ИВМиМГ СО РАН и являлся штатной структурой, первоначально объединявшей отдел вычислительных систем и сетей ИВМиМГ, а также временный научно-технический коллектив "Параллель" из специалистов ИВМиМГ и других институтов СО РАН. Научным руководителем центра был назначен акад. А. С. Алексеев. В те же годы был создан сайт ССКЦ КП (www2.sssc.ru). В настоящее время существует лаборатория "Сибирский суперкомпьютерный центр" в составе ИВМиМГ СО РАН, научным руководителем ЦКП ССКЦ является академик Б. Г. Михайленко.

Важным моментом данного периода является распространение технологии параллельного программирования. Успешное решение больших задач требует объединения усилий ученых и высококвалифицированных инженеров. Эксплуатация параллельного оборудования ССКЦ обеспечивалась людьми, прошедшими школу эксплуатации комплекса БЭСМ-6 и ЕС ЭВМ (отдел вычислительных систем и сетей). Однако для решения больших и сверхбольших задач этого оказалось недостаточно, поскольку возникла необходимость также решать комплексные научные задачи в области параллельных вычислений. Работы в области параллельных вычислений ведутся в отделе математического обеспечения высокопроизводительных вычислительных систем ИВМиМГ. Отдел накопил и сохранил традиции и опыт ученых в области исследования параллельных вычислений в Новосибирском Институте математики (ИМ) и ВЦ СО АН СССР, начинавших в 1960-х годах.

Теоретические и экспериментальные исследования, выполненные в 1960-80-е годы в ИМ и ВЦ СО РАН, составили основу современных достижений в построении параллельных вычислительных систем и создании программного обеспечения для них. Так, в 1960-е годы проводились эксперименты по созданию вычислительных систем из нескольких ЭВМ, активно разрабатывалось программное обеспечение для взаимодействия ЭВМ в системе, создавались языки описания алгоритмов решения задач.

Сейчас такие системы реализуются в форме мультикомпьютеров различных типов, а их дальнейшее обобщение приводит к сетям компьютеров (GRID). Возникновение этого направления в ВЦ СО РАН связано с переходом нескольких сотрудников из ИМ в отдел математического обеспечения высокопроизводительных вычислительных систем в 1984 г. Другую часть коллектива отдела составляют ученики А. П. Ершова и В. Е. Котова, которые продолжают исследования крупноблочных вычислений и, в первую очередь, проблем параллельного программирования мультикомпьютеров [5, 6]. Такой состав отдела определил комплексный характер исследований, необходимый для создания современных параллельных вычислительных технологий. В отделе ведутся разнообразные проекты по созданию систем отладки и мониторинга исполнения параллельных программ, асинхронного программирования, программного обеспечения GRID-вычислений, сборочной технологии параллельного программирования, по разработке численных моделей большого размера и рекомендаций по эффективному параллельному программированию прикладных задач, визуализации результатов моделирования.

В настоящее время основными задачами ЦКП ССКЦ являются:

- обеспечение деятельности институтов СО РАН и университетов в области математического моделирования в фундаментальных и прикладных исследованиях в механике, физике, химии, геологии, биологии и других дисциплинах высокопроизводительными суперкомпьютерными технологиями, техническими средствами и квалифицированным обслуживанием вычислительных систем центра;

- организация обучения специалистов СО РАН, студентов и аспирантов НГУ и НГТУ методам параллельных вычислений на суперкомпьютерах;

- сетевое взаимодействие с другими Суперкомпьютерными центрами СО РАН, Москвы и других городов России, а также зарубежных стран, совместная разработка технологий распределенных вычислений.

Архитектурные особенности ЦКП ССКЦ. В настоящее время в ССКЦ имеются два кластера, используемых в режиме коллективного пользования институтами СО РАН. Один из кластеров построен на основе вычислительных узлов с Intel Xeon (архитектура MPP), пиковая производительность 30 TFlop/s, программирование с применением MPI и OpenMP, другой – с гибридным расширением на GPU NVIDIA Tesla M2090 (архитектура GPGPU), пиковая производительность 84 TFlop/s, параллельное программирование при помощи C/C++ CUDA и OpenCL. Имеется кластерная файловая система Ibrix, включающая четыре сервера и 32 Тбайта памяти. Кроме того, в состав ССКЦ входит сервер с общей памятью HP ProLiant DL980 G7 с четырьмя 10-ядерными процессорами Intel E7-4870 с тактовой частотой 2,4 ГГц, оперативной памятью 512 Гбайт и восемью SAS-дисками по 300 Гбайт. Пиковая производительность сервера в текущей конфигурации составляет 384 GFlop/s. В апреле 2012 г. сервер включен в кластер НКС-30Т как нестандартный вычислительный узел.

На рис. 1 показана структурная схема гетерогенного кластера НКС 30Т+GPU. В состав кластера входят 576 процессоров (2688 ядер) Intel Xeon E5450/E5540/X5670; 120 процессоров GPU Tesla M 2090 (61440 ядер); SMP сервер с общей памятью hp DL980 G7 (4 процессора, 40 ядер Intel E7-4870, оперативная память 512 Гбайт); кластерная файловая система IBRIX (4 сервера, 32 Тбайта памяти).

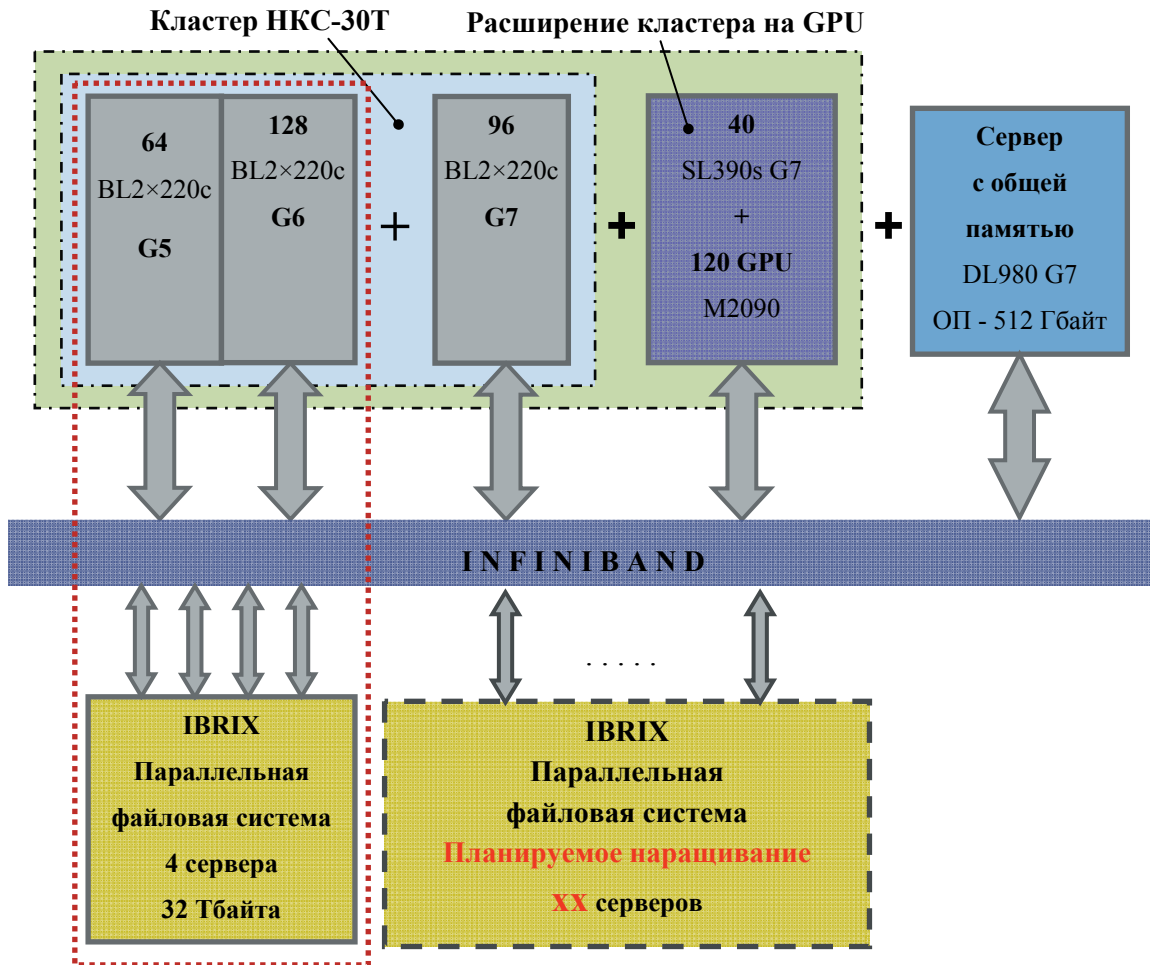


Рис. 1. Структурная схема соединений основных узлов кластера

Таким образом, в состав гетерогенного кластера входят вычислительные блоки с MPP-архитектурой, гибридной архитектурой с использованием карт NVIDIA Tesla M2090 (40 узлов, на каждый узел по три карты) и SMP-архитектурой (планируется увеличение количества вычислительных ядер до 80, оперативной памяти – до 1 Тбайта). Все узлы кластера связаны между собой через Infiniband QDR.

Такая структура кластера отвечает требованиям центров коллективного пользования, поскольку приходится решать самые разнообразные задачи из различных областей знаний, и наличие нескольких архитектур в центре дает возможность выбрать оптимальную, исходя из специфики решения задачи. Например, для плохо распараллеливаемых задач часто используют SMP-архитектуру.

При такой схеме построения центра имеется возможность использовать все ресурсы гетерогенного кластера при решении одной задачи.

Подробнее о составе технических и программных средств, пакетах прикладных программ можно посмотреть на сайте ССКЦ (www2.ssc.ru).

Инфраструктура машинного зала ССКЦ. Центр обработки данных (ЦОД) ССКЦ занимает четыре помещения общей площадью 205 м²: машинный зал № 1 площадью 66,7 м²; машин-

ный зал № 2 площадью 59,9 м²; узел электропитания площадью 58,5 м²; помещение гидромодуля площадью 20 м².

ЦОД оборудован системой газового пожаротушения, пожарной и охранной сигнализацией, источниками бесперебойного питания и прецизионными кондиционерами, системой мониторинга температуры и влажности (дополнительную информацию см. www2.sssc.ru/Information/Infrastr/2012/Infrastr-2012.htm). Общая мощность двух источников бесперебойного электропитания составляет 240 кВт, общая мощность прецизионных кондиционеров по холоду составляет 276 кВт. Вычислительная техника работает в круглосуточном режиме. ЦКП ССКЦ подключен по выделенному каналу 1 Гбит/с к сети ННЦ и дополнительно по скоростному каналу 10 Гбит/с к его суперкомпьютерной сети.

Программное обеспечение и инструментальные средства разработки. На кластере НКС-30Т установлен Intel MPI 4, компиляторы Intel C++ и Intel Fortran Composer XE for Linux Version 2011 Update 5, включающие библиотеки Intel MKL, Intel IPP и Intel TBB. На кластере также установлены параллельные версии Gromacs, Quantum Espresso и Bioscope. Для программирования на GPU Nvidia установлен CUDA Toolkit 5.0. В 2012 г. закуплены компиляторы PGI Accelerator (годовая лицензия) для работы на GPU, коммерческие пакеты ANSYS CFD (годовая поддержка) и Gaussian 09. В декабре 2012 г. ANSYS CFD обновлен до версии 14.5, в январе 2013 PGI Accelerator обновлен до версии 13.1 (release 13.1, updated January 28, 2013).

Поскольку сервер с общей памятью HP ProLiant DL980 G7 включен в НКС-30Т, то на нем доступно программное обеспечение кластера. На многопроцессорном сервере с общей памятью ProLiant DL580 G5 также установлены компиляторы Intel C++ и Intel Fortran Composer XE for Linux. Одинаковый комплект базового программного обеспечения на кластерах и серверах упрощает работу пользователей.

Особенностью программирования задач на кластере с MPP-архитектурой, ориентированной на решение больших задач, прежде всего 3D, является применение параллельных языков MPI и OpenMP, поскольку это обусловлено архитектурой кластера, построенного с использованием многопроцессорных серверов с общей памятью (SMP). При таком подходе внутри каждого вычислительного модуля формируются несколько потоков с помощью OpenMP. Таким образом поддерживаются две современных парадигмы параллельных вычислений – MPI для систем с распределенной памятью (кластеров) и OpenMP для систем с общей памятью. Схема вычислений предусматривает запуск одного MPI-процесса на каждом вычислительном узле кластера, который запускает внутри каждого вычислительного модуля несколько потоков с помощью OpenMP.

Другая технология высокопроизводительных вычислений связана с реализацией параллельного алгоритма на гибридной архитектуре следующего вида: суперкомпьютер состоит из набора соединенных между собой узлов, для обмена данными используется MPI; каждый узел состоит из одного CPU и трех GPU; на каждом узле запускается один процесс MPI, управляющий вычислениями (процесс выполняется на CPU); из MPI процесса запускаются нити (threads) CUDA, каждая из которых предназначена для выполнения на своем ядре (CUDA core); для управления тремя GPU из одного CPU используется технология Multi-GPU.

Перспективы развития ЦКП ССКЦ. Необходимость наращивания мощности ЦКП ССКЦ обусловлена моральным старением оборудования и возрастающими потребностями институтов СО РАН в использовании современных вычислительных средств для решения задач в следующих исследуемых областях: математического моделирования свойств наноматериалов; развития нанотехнологий в физике, химии, биологии, геологии, наноэлектронике; создания трехмерных моделей Земли; трехмерных моделей атмосферы; синтеза веществ в фармакологии; моделирования физики атмосферы и океана; аэродинамики летательных аппаратов; лазерной оптики (трехмерные связанные резонаторы) и др. В настоящее время ЦКП ССКЦ обслуживает 19 институтов СО РАН и 3 университета, более 160 пользователей решают разнообразные научно-технические задачи. Однако по вычислительным мощностям ССКЦ (115 TFlop/s) находится на одном из последних мест среди организаций занимающихся высокопроизводительными вычислениями. Достаточно упомянуть МГУ, кластер "Ломоносов" (1700 TFlop/s), МСЦ (227 TFlop/s); кластер на Intel Xeon Phi (523 TFlop/s), который должен служить прототипом суперкомпьютера производительностью 10 PFlop/s (10000 TFlop/s), также в ЮрГУ запущен кластер на Intel Xeon Phi (236 TFlop/s).

ССКЦ совместно с корпорацией HP и ООО "Нонолет-ИТ" разработал проект гетерогенного кластера с использованием ускорителей Intel Phi. Последнее обстоятельство очень важно для программистов, поскольку используется единая система команд, в отличие от имеющегося в ССКЦ кластера с GPU NVidia Tesla M2090. В последнем для программирования графических карт необходимо применять специализированный язык CUDA, что является существенным препятствием для широкого применения пользователями гибридного кластера. Создание кластера нового поколения позволит существенно повысить эффективность решения проблем по вышеперечисленным направлениям, проектам и грантам. В основе предлагаемого решения лежит модульное серверное шасси HP Proliant SL6500, занимающее в стойке 4U пространства и вмещающее четыре 2-юнитовых сервера половинной ширины HP Proliant SL250 Gen8. Каждое шасси обеспечивает централизованное электропитание и охлаждение установленных в него серверов, для этого в шасси установлены четыре высокоэффективных блока питания мощностью 1200 Вт каждый и восемь отказоустойчивых вентиляторов охлаждения. Высокую производительность серверов HP SL250Gen8 обеспечивают два центральных восьмиядерных процессора последнего поколения Intel® Xeon® Processor E5-2670 (20M Cache, 2.60 GHz, 8.00 GT/s Intel® QPI), 64 Гбайта оперативной памяти и два вычислительных модуля Intel Xeon Phi 5110P с 8 ГБ локальной памяти. Для установки операционной системы предусмотрен жесткий диск SATA на 500 Гбайт.

Сопроцессоры Intel Xeon Phi 5110P обеспечивают более высокую производительность при пониженном энергопотреблении. Они демонстрируют производительность с удвоенной точностью на уровне 1,011 GFlop/s (1,01 TFlop/s) и поддерживают 8 Гбайт памяти GDDR5 с пропускной способностью 320 Гбит/с. Сопроцессоры Intel Xeon Phi 5110P с показателем TDP на уровне 225 Вт, оснащенные пассивным охлаждением, обеспечивают энергоэффективность, оптимально соответствующую средам с высокой плотностью размещения вычислительного оборудования, и предназначены для рабочих задач с ограничениями по скорости вычислений, включая разработку цифрового контента и исследования в области энергетики.

Применение 102 таких серверов обеспечит производительность вычислений с двойной точностью на уровне $R_{peak} = 1011 \text{ GFlop} \times 204 = 206, 244 \text{ TFlop/s}$.

Для высокоскоростного обмена данными между узлами кластера в каждом сервере имеется адаптер FDR Infiniband (QSFP), подключенный к коммутатору Voltaire IB, который обеспечивает интеграцию с параллельной кластерной системой хранения данных.

Следует отметить, что по такому же принципу построен лидер последнего списка Топ-500 китайский кластер Tianhe-2.

Создание данного кластера и его пиковая производительность полностью зависят от финансирования, выделяемого Президиумом СО РАН.

Подготовка кадров. Высокая квалификация сотрудников института и их большой опыт позволяют не только обеспечивать научную составляющую работ ССКЦ, но успешно готовить новые кадры. Несмотря на большую востребованность на рынке труда специалистов в параллельном программировании и их постоянный отток из института, столь же постоянно кадровый состав восстанавливается за счет успешной работы базирующихся в институте кафедр: Математических методов геофизики НГУ; Вычислительной математики НГУ; Параллельных вычислений НГУ; Вычислительных систем НГУ; Параллельных вычислительных технологий НГТУ. С 2002 г. отдел высокопроизводительных вычислений института регулярно проводит весенние и осенние школы для пользователей ССКЦ и студентов по параллельным алгоритмам и программам.

При институте создан Учебный научный центр (УНЦ) на 12 компьютеризированных рабочих мест, который по оптоволоконной линии связан с ССКЦ. В этом центре проводятся занятия со студентами, на его базе проходят летние и зимние школы по параллельному программированию, школы Intel по высокопроизводительным вычислительным технологиям. В 2012 г. на базе ИВМиМГ проведена Международная конференция "Параллельные и вычислительные технологии-2012", в работе которой приняли участие 242 участника из России, Казахстана, Украины, Германии, Франции, США (agoga.gugu.ru/display.php?conf=pavt2012). В апреле 2012 г. при поддержке специалистов NVIDIA на вычислительных ресурсах кластера организована трехдневная школа по технологии NVIDIA CUDA, в которой прошли обучение 118 слушателей из институтов СО РАН, вузов и фирм (программу и учебные материалы см. www2.sccc.ru/Seminars/Nvidia%20Cuda-1.htm). В декабре 2012 г. проведена школа по параллельному программированию гибридных кластеров (www2.sccc.ru/Seminars/Shool-2012.htm). Организован регулярный семинар "Архитектура, системное и прикладное программное обеспечение кластерных супер-ЭВМ" на базе ССКЦ ИВМиМГ СО РАН, кафедры Вычислительных систем НГУ и Центра Компетенции по высокопроизводительным вычислениям СО РАН – Intel (презентации семинаров см. www2.sccc.ru/Seminars/NEW/Seminars.htm).

Решение прикладных задач. ЦКП ССКЦ СО РАН предоставляет вычислительные и консалтинговые услуги девятнадцати академическим институтам СО РАН РАН и трем университетам, более 160 пользователей используют ресурсы центра для решения своих задач. Решается большое количество задач из различных областей знаний, в том числе, определенных приоритетными направлениями развития науки и техники: индустрия наносистем (ИВМиМГ, ИК, ИТПМ, ИХКиГ, ИФП, ИЯФ, ИХиХТ (Красноярск)), ОНЦ (Омск), ИКЗ (Тюмень); информационно-



Рис. 2. Диаграмма распределения процессорного времени за 2012 г.

телекоммуникационные системы (ИВТ, ИВМиМГ, ИГ, ИНХ, НГУ, НГТУ, ИЦиГ, ИЯФ); энергоэффективность, энергосбережение, ядерная энергетика (ИВТ, ИВМиМГ, ИГ, ИК, ИТ, ИХКиГ, ИЯФ, НГУ, НГТУ, ИНГиГ); науки о жизни (ИХБиФМ, ИЦиГ, ИВМиМГ, НГУ, ИГ, ИКЗ (Тюмень), ОФ ИМ (Омск)); рациональное природопользование (ИВМиМГ, ИНГиГ, ИТ, НГУ, ИХиХТ (Красноярск), ИКЗ (Тюмень)); транспортные и космические системы (ИТПМ, НГТУ).

На рис. 2 приведена диаграмма распределения процессорного времени ССКЦ. Видно, что основными пользователями центра являются научные сотрудники СО РАН.

Из большого количества задач, решаемых в центре, приведем только некоторые из комплекса задач ИВМиМГ.

Задачи обработки данных в физике высоких энергий. На части кластера НКС-30Т развернута основанная на KVM виртуализованная вычислительная среда, используемая для обработки данных физических экспериментов в области физики высоких энергий, проводимых в ИЯФ СО РАН. Обмен данными между ИЯФ СО РАН и ССКЦ осуществляется через суперкомпьютерную сеть ННЦ (10 Гбит/с) [7].

Эксперимент КЕДР: работа проводится на электронно-позитронном коллайдере ВЭПП-4М с детектором КЕДР. Эксперименты в области рождения ψ -резонансов (J/ψ , $\psi(2S)$, $\psi(3770)$) и τ -лептона.

Эксперимент ATLAS: работа проводится на Большом адронном коллайдере (БАК) (ЦЕРН, Швейцария). Анализ данных эксперимента ATLAS в рамках ATLAS Exotics Working Group.

Эксперимент СНД: работа проводится на коллайдере ВЭПП-2000 со сферическим нейтральным детектором (СНД). Изучение процессов электрон-позитронной аннигиляции в области энергии до 2 ГэВ в системе центра масс.

Имитационное моделирование алгоритмов для экзафлопсных суперкомпьютеров. В настоящее время в ИВМиМГ развивается новое направление исследований "Развитие суперкомпьютерных технологий и методов моделирования архитектур и алгоритмов для пета- и экзафлопсных супер-ЭВМ" (руководители проекта – д-р техн. наук Глинский Б. М., д-р техн. наук Родионов А. С.). Направление связано с исследованием свойств масштабируемости параллельных алгоритмов при их реализации на будущих супер-ЭВМ экзафлопсной производительности.

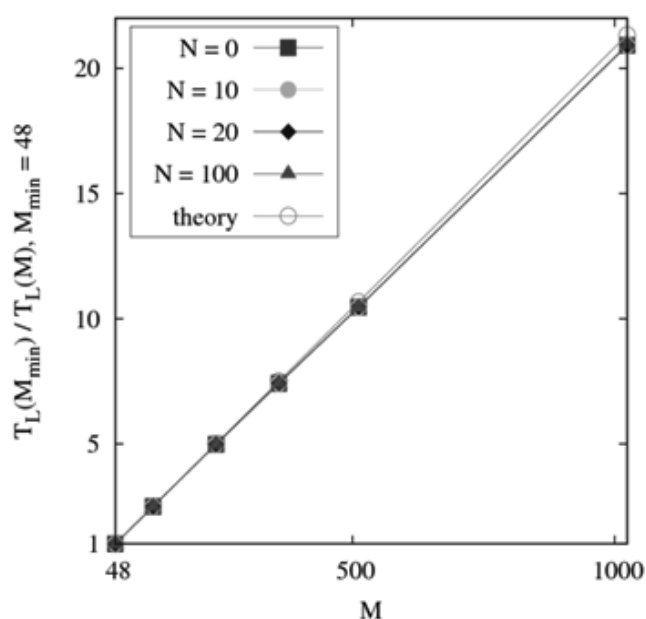


Рис. 3. Сравнение ускорения до $M=1000$. Результаты ускорения для модели совпадают с ускорением при расчетах с использованием PARMONC

Уже сейчас можно оценить поведение алгоритмов, разработать модифицированные схемы вычислений путем реализации их на имитационной модели, отображающей тысячи и миллионы вычислительных ядер. Имитационная модель позволяет выявить узкие места в алгоритмах, понять, как нужно модифицировать алгоритм, какие параметры необходимо настраивать при его масштабировании на большое количество ядер [8, 9]. Для моделирования используется система моделирования AGNES (AGent NETwork Simulator), разработанная в ИВМиМГ и установленная на кластере ССКЦ (см. www2.sccc.ru/PPP/Mat-Libr/agnes.htm). Все расчеты проводились на кластере НКС-30Т+GPU.

Приведем примеры моделирования двух задач: 1) имитация распределенного статистического моделирования (задача динамики разреженного газа по методу ПСМ, канд. физ.-мат. наук М. А. Марченко); 2) численное моделирование 3D сейсмических полей (канд. физ.-мат. наук Д. А. Караваев).

В первой задаче исходные данные для имитационного моделирования получены с использованием библиотеки PARMONC, предназначенной для использования на современных суперкомпьютерах тера- и петафлопсного уровня [10]. Библиотека также установлена на кластере ССКЦ (см. www2.sccc.ru/SORAN-INTEL/paper/2011/parmonc.pdf).

Схема вычислений для этой задачи требует наличия "ядер-сборщиков", которые периодически собирают статистику с "ядер-вычислителей" [8]. Проведенное имитационное моделирование показало, что при большом числе используемых вычислительных ядер (более 10000) реальное ускорение от распараллеливания существенно отличается от теоретического, что связано с большой загрузкой выделенных "ядер-сборщиков", обрабатывающих пакеты данных, поступающие с "ядер-вычислителей". При этом, при использовании до 1000 ядер ускорение в модели совпадает с ускорением в реальных расчетах (рис. 3).

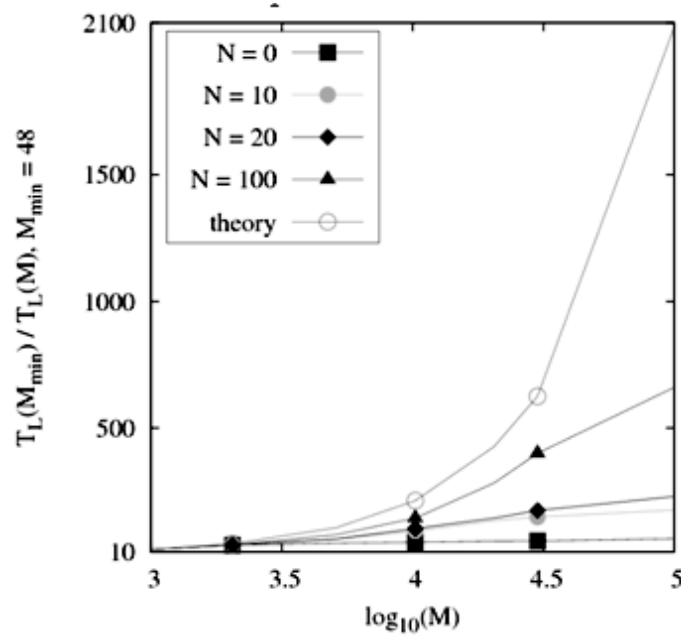


Рис. 4. Сравнение ускорения распределенного статистического моделирования для разных вариантов организации обмена данными для числа ядер M до 100 000 (горизонтальная ось – в логарифмическом масштабе)

Ускорение от распараллеливания при расчетах на M ядрах определим так: $S_L(M) = T_L(M_{\min})/T_L(M)$, где $T_L(M)$ – машинное время на центральном "ядре-сборщике", затраченное на моделирование и сохранение выборочных средних для L реализаций случайной оценки; M_{\min} – наименьшее число ядер, использованных при расчетах. В двухуровневом варианте число "ядер-вычислителей" было поделено на N равных частей ($N = 10, 20, 100$), для каждой из которых данные с "ядер-вычислителей" сначала отправлялись на специально выделенное промежуточное "ядро-сборщик". В свою очередь, N промежуточных "ядер-сборщиков" отправляли данные на главное "ядро-сборщик". В одноуровневом варианте (будем считать, что число промежуточных "ядер-сборщиков" равно нулю ($N = 0$)) данные с "ядер-вычислителей" непосредственно отправлялись на главное "ядро-сборщик".

Следующий результат имитационного моделирования приведен на рис. 4. В данном случае была использована логарифмическая шкала для количества ядер. Видно сильное отклонение от теоретической кривой. Еще большее расхождение получилось при дальнейшем увеличении количества вычислительных ядер до 500 000 и более.

На основе анализа полученных результатов сделан вывод о том, что схему вычислений необходимо изменять. В частности, были предложены многоуровневые схемы для "ядер-сборщиков".

Для имитации сеточных методов при численном моделировании 3D сейсмических полей реализован класс функциональных агентов Grid — узел-вычислитель, имитирующий расчет сеточных методов на одном вычислителе. Моделируются вычисления, при которых область исследования режется вдоль одной оси, и полученные области загружаются на вычислители. Таким образом, получается, что у каждого вычислителя есть пересечение по данным максимум с двумя вычислителями ("крайние" вычислители обмениваются только с одним соседом). Каж-

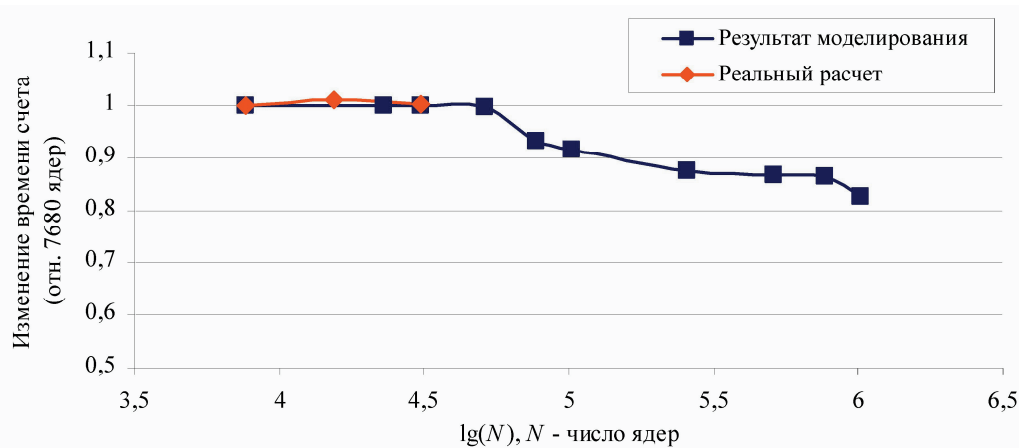


Рис. 5. Изменение времени расчета алгоритма численного моделирования в зависимости от числа вычислительных ядер (горизонтальная ось отображена в логарифмическом масштабе)

дый вычислитель на первом шаге рассчитывает свои граничные области, затем асинхронно передает насчитанные результаты соседям. Расчет внутренних областей идет на втором шаге: получив данные от соседей и просчитав изменение своей области, агент переходит к шагу один.

Общие результаты изменения времени счета в зависимости от количества доступных ядер GPU (при пропорциональном увеличении размера 3D модели) в логарифмическом масштабе приведены на рис. 5. Показано хорошее соответствие экспериментальных и модельных результатов на начальном участке кривой (до 30720 ядер). При значительном увеличении количества вычислительных узлов с пропорциональным увеличением размера 3D модели время счета увеличивается, но незначительно (при росте числа узлов от 7680 до 1 024 000 время увеличилось на 17,5 %) [9].

Исследование масштабируемости алгоритма на большое количество ядер проводилось с использованием агентно-ориентированной системы имитационного моделирования (AGNES). Исследование показало, что даже при явном распараллеливании алгоритма прямого статистического моделирования на большое количество ядер не происходит ожидаемого ускорения, близкого к линейному закону. Это связано с тем, что при числе ядер, достигающем порядка сотен тысяч или нескольких миллионов, возникают проблемы с большой загрузкой "ядер-сборщиков", которые периодически собирают статистику с "ядер-вычислителей". Следовательно, при масштабировании необходима модификация параллельной вычислительной программы, например, увеличение количества "ядер-сборщиков".

Аналогичные эксперименты проведены с численным моделированием сейсмических полей в 3D неоднородных упругих средах. В качестве метода решения используется сеточный разностный метод, а область моделирования представляется изотропной, 3D неоднородной сложно построенной упругой средой. Моделирование показывает, что при решении этой задачи можно использовать 1 млн. и более вычислительных ядер, следовательно, можно значительно ускорить время счета прямых задач, необходимых для интерпретации данных вибросейсмического зондирования [9].

Таким образом, проведенные исследования показывают эффективность имитационного моделирования при настройке параметров масштабируемых алгоритмов и изучении их поведения при реализации на большом количестве вычислительных ядер.

Список литературы

1. АЛЕКСЕЕВ А. С. Информационные и вычислительные технологии / А. С. Алексеев, М. А. Лаврентьев. Наука в Сибири. Новосибирск. 2000. № 44–45.
2. О ПРОГРАММЕ работ по созданию сети информационно вычислительных систем (Центров) в Сибирском отделении АН СССР/ Алексеев А. С. и др. Новосибирск, 1987. (Препринт / РАН. Сиб. отд-ние. ВЦ; 467).
3. АКАДЕМИЧЕСКАЯ региональная сеть Сибири / Алексеев А. С. [и др.]. Новосибирск, 1983. (Препринт / РАН. Сиб. отд-ние. ВЦ; 467).
4. ИСТОРИЯ развития Сибирского суперкомпьютерного центра, его текущее состояние и перспективы развития / Алексеев А. С. и др. // Сиб. журн. вычисл. математики РАН. 2005. Т. 8. № 3. С. 179–187.
5. КОТОВ В. Е., НАРИНЬЯНИ А. С. Асинхронные вычислительные процессы над памятью // Кибернетика. 1966. №3. С. 64–71.
6. КОТОВ V. E., NARINYANI A. S. On transformation of sequential programs into asynchronous parallel programs // Proc. of the IFIP Congress 68, Edinburg, 1968; Amsterdam, 1969. P. 121 –127.
7. ИСПОЛЬЗОВАНИЕ виртуализованной суперкомпьютерной инфраструктуры Новосибирского научного центра для обработки данных экспериментов физики высоких энергий / Белов С. Д. и др. // Выч. технологии. 2012. Т. 17. № 6. С. 36–46.
8. АГЕНТНО-ОРИЕНТИРОВАННЫЙ подход к имитационному моделированию супер-ЭВМ экзафлопсной производительности в приложении к распределенному статистическому моделированию / Глинский Б. М. и др. // Вестн. ЮУрГУ. 2012. № 18 (277), вып. 12. С. 93–106.
9. SCALING the Distributed Stochastic Simulation to Exaflop Supercomputers / B. Glinsky, et al // Proc. of the 2012 IEEE 14th Intern. conf. on high performance computing and communications. P. 1131–1136.
10. MARCHENKO M. A. PARMONC. A software library for massively parallel stochastic simulation // LNCS. 2011. V. 6873. P. 302–315.

Глинский Борис Михайлович – д-р техн. наук, проф., исполнит. дир. ЦКП "Сибирский Суперкомпьютерный центр" ИВМиМГ СО РАН; тел. (383) 330-62-79; e-mail: gbm@sscc.ru

Дата поступления – 29.05.13