СКИФ-МГУ: №22 top500 – что дальше?

27.03.08 было официально объявлено о завершении строительства самого мощного в странах Восточной Европы суперкомпьютера "СКИФ МГУ". Основные особенности проекта и перспективы дальнейшего развития суперкомпьютинга – тема публикации.

Введение

Прошло уже около 3,5 лет с тех пор, когда в ноябре 2004 г. Объединенный институт проблем информатики Национальной академии наук Беларуси, ИПС РАН, компания "Т-Платформы" и корпорация AMD объявили о создании суперкомпьютера "СКИФ К-1000", занявшего на тот момент 98 место рейтинга суперкомпьютеров top500 (www.storagenews.ru/21/SKIF-1000_21.pdf).

Новое совместное заявление МГУ им. М.В. Ломоносова, ИПС РАН, компании "Т-Платформы" и корпорации Intel, состоявшееся 27.03.08, о завершении работ по проекту создания самого мощного в странах Восточной Европы суперкомпьютера "СКИФ МГУ" позволило подняться России в мировом рейтинге суперкомпьютеров на 22 позицию – его производительность на тесте Linpack составила 47,17 TFlops (78,6% от пиковой). Проект был реализован в течение полугода (сентябрь 2007 г. – март 2008 г.) и создан по плану работ в рамках суперкомпьютерной программы "СКИФ-ГРИД" на объединенные средства МГУ им. Ломоносова и суперкомпьютерной программы "СКИФ-ГРИД", финансируемой из бюджета Союзного государства. Общая стоимость комплексного проекта составила 231 млн руб. (из них 101,0 млн. руб. (54%) из бюджета РФ первых двух лет Программы "СКИФ-ГРИД".

В настоящее время Россия в рейтинге top500 представлена семью системами и вместе со Швейцарией и Швецией занимает 9 место в списке стран, располагающих самыми высокопроизводительными компьютерами.

Одновременно была опубликована 8-я редакция списка самых мощных компьютеров СНГ – top50 (http://www.supercomputers.ru).

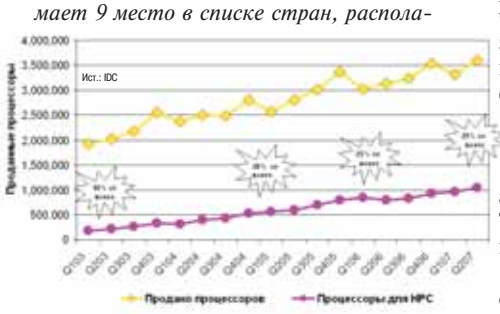

Рис. 1. По данным IDC, мировой рынок процессоров, проданных для HPC на конец 2 кв. 2007 г. составил 29% от общего числа процессоров, проданных для серверов.

Тенденции HPC-рынка

Мировой рынок кластерных HPC-систем начал развиваться примерно с 1999 г., в России – несколько позже, примерно с 2003 г. В соответствии с недавними исследованиями IDC, на конец 2 кв. 2007 г. мировой рынок процессоров, проданных для HPC-систем составил 29% от общего числа процессоров, проданных для серверов. HPC-рынок по числу продаваемых серверов показывает постоянный квартальный рост, что в среднем в год составляет 25% по сравнению с 8% общим серверным рынком. По данным Intel, их рост в секторе HPC-серверов еще больше – 37% по сравнению с 11% общего объема серверов и уже в 2008 г. HPC-рынок может достигнуть 35% рынка серверов.

HPC-рынок в России растет гораздо большими темпами, чем мировой. Так, суммарная пиковая производительность трех самых мощных суперкомпьютеров России (в TFlops) за период 2005–2008 гг. выросла с 8 до 125. И на начало 2010 г. планируется выход на уровень пета-компьютинга.

В соответствии с 8-й редакций списка top50 суммарная производительность систем на тесте Linpack за полгода выросла с 61,6 до 197,3 TFlops (на 220,3%), что является беспрецедентным ростом за все время существования списка. В целом, количество новых систем в списке (включая системы, модернизированные за последние полгода) составило 56% (28 из 50). Количество компьютеров "терафлопного диапазона" (с реальной производительностью более 1 TFlops) на территории СНГ за прошедшие полгода практически удвоилось с 13 до 25, а нижняя граница первой десятки по производительности выросла с 1,3 TFlops до 5,2 TFlops (на 300%). Для попадания в список Топ50 теперь требуется производительность на тесте Linpack не менее 432 GFlops (253,6 GFlops в предыдущей редакции).

По количеству систем, входящих в список top50, свои лидирующие позиции укрепила компания "Т-Платформы", увеличившая долю своих суперкомпьютеров в списке с 30% до 46% (с учетом суперкомпьютеров "СКИФ", поставленных компанией и разработанных с ее определяющим участием – всего 23 системы из 50 и более 20% рынка решений для высокопроизводительных вычислений, *по данным IDC*). Далее следует Hewlett-Packard (доля уменьшилась с 26% до 20%, соответственно, 13 и 10 систем) и компания IBM (доля уменьшилась с 22% до 14%, соответственно, 10 и 6 систем).

Еще одна положительная тенденция в соответствии с 8-й редакцией списка top50 в сравнении с общемировыми показателями – сокращение разрыва по отраслям по пиковой производительности (в скобках 7 ред.): исследования – в 10 раз (10), промышленность – 2,9 раз (14), финансы – 4,5 раз (11).

Однако возрастающая диспропорция, в сравнении с общемировыми показателями, в таких отраслях промышленности, как электронная, геологоразведка, нефте- и газодобыча и ряде других, не выделяемых отдельной строкой, свидетельствует о недостаточно сбалансированной стратегической политике в области суперкомпьютинга.

В области HPC-архитектуры продолжается рост числа систем в списке, построенных на процессорах Intel (с 31 до 38 – 76% списка). Также в список входит 6 систем, построенных на процессорах AMD (9 – в предыдущей редакции), 5 систем – на базе процессоров IBM (5 – полгода назад) и 1 система – на базе процессоров HP (5 – в сентябре). Вопреки ожиданиям, никакой конкуренции нет со стороны AMD с "Барселоной", Sun с "Ниагарой" и IBM с Power 6.

Заметно уменьшается число компьютеров, использующих для взаимодействия узлов лишь коммуникационную сеть Gigabit Ethernet – с 20 систем в преды-

дущей редакции до 9 в нынешней. Расширяется использование высокоскоростных коммуникационных технологий: InfiniBand (с 20 до 31 системы) и Myrinet (с 6 до 8 систем). Еще по 1 системе построено на базе коммуникационных технологий SCI и HyperPlex.

Список top50 на 100% состоит из кластеров. В нем (с 4-й редакции) отсутствуют SMP- или/и векторно-конвейерные системы. Для большинства применений – кластеры наиболее оптимальный выбор по показателям цена и производительность. И лишь для отдельных приложений, требующих общей памяти, не соизмеримой для одного сервера, предпочтение отдается SMP-системам. Пример – "Росгидромет".

"СКИФ-МГУ": самый мощный суперкомпьютер в странах Восточной Европы

За период развития суперкомпьютерной отрасли в новейшей истории России создание системы "СКИФ-МГУ" – одно из наиболее значимых событий. И не только потому, что был достигнут один из рекордных показателей по производительности, а, прежде всего, этот проект отразил переход на новый технологически зрелый уровень отечественных разработок в этой отрасли. Из наиболее значимых технологических достижений "СКИФ-МГУ" хотелось бы отметить следующее:

1. Это первый проект, реализованный на отечественных блэйд-серверах. И хотя элементная база остается покупной, конструктив – разработка компании "Т-платформы". Блэйд разработан на 45-нм Intel-процессорах, анонсированных в декабре 2007 г., и позволяет на шасси высотой 5U достичь производительности 0,96 TFlops, что обеспечивает более 7 TFlops на стойку – самая высокая плотность вычислительной мощности для blade-систем на базе процессоров Intel.

Оптимально решена проблема отвода тепла. 10 вентиляторов с "горячей заменой", расположенных с лицевой стороны каждого blade-шасси, обеспечивают эффективное охлаждение blade-модулей, причем ресурс работы вентиляторов при таком расположении в 2–3 раза больше, чем в традиционных blade-решениях. Использование "горячего коридора" и межрядных кондиционеров позволяет ограничить доступ горячего воздуха в помещение и равномерно подавать холодный воздух непосредственно к вычислительным модулям. Потенциально такое решение позволяет отводить до 60 КВт от стойки с вычислительными модулями (почти 2-кратный запас при максимальной выделяемой мощности одного лезвия – 435 Вт).

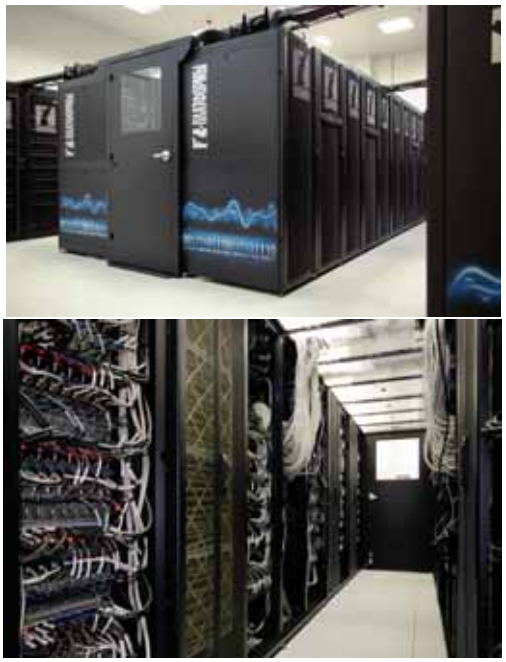

Рис. 3. Внешний вид системы СКИФ-МГУ: "снаружи" (вверху) и "изнутри" т.н "горячий коридор" (внизу).

Общее энергопотребление суперкомпьютера в стандартном режиме составляет 520 кВт и может достигать 720 кВт при теоретически возможной пиковой нагрузке. Данная мощность сосредоточена на площади менее 100 м^2, при этом температура в суперкомпьютерном центре не превышает 20^{o}C. Для отвода тепла инженерами "Т-Платформы" была спроектирована модульная система охлаждения с герметичным "горячим коридором" между стойками с вычислительными узлами. Решение гарантирует отвод до 30 кВт тепловой энергии от каждой стойки, имеет уровень резервирования всех компонентов N+1 и, в аварийном случае, обеспечивает поддержание температурного режима в помещении не менее 10 минут.

3. За счет полнофункциональной программно-аппаратной системы мониторинга разработки ИПС РАН полностью централизовано управление температурным полем стойки/шасси/лезвия и состоянием блоков питания. Автоматическое изменения скорости вращения вентиляторов позволяет экономить до 30% энергопотребления подсистемы охлаждения. Удаленное управление системой на всех уровнях осуществляется через единый иерархический веб-интерфейс.

4. "СКИФ МГУ" впервые использует российские программные средства для кластерных систем, разработанные в рамках программы "СКИФ-ГРИД" и включающие специально созданный отечественный кластерный дистрибутив ОС Linux от ALT Linux. Программные разработки ИПС РАН и НИВЦ МГУ (OpenTS и X-Com) позволят существенно упростить разработку параллельных приложений и организовать распределенные вычисления с использованием разнородных вычислительных ресурсов.

Планируется сертификация системы на "Intel Cluster Ready", что будет являться свидетельством гарантии работоспособности на СКИФ-МГУ любого кластерного ПО.

5. "СКИФ МГУ" – законченное сбалансированное решение, включающее систему хранения данных с параллельной файловой системой T-Platforms ReadyStorage ActiveScale Cluster объемом 60 Тбайт, ленточную систему резервного копирования данных. Параметры и состав всех подсистем подобраны таким образом, чтобы обеспечить максимальную эффективность

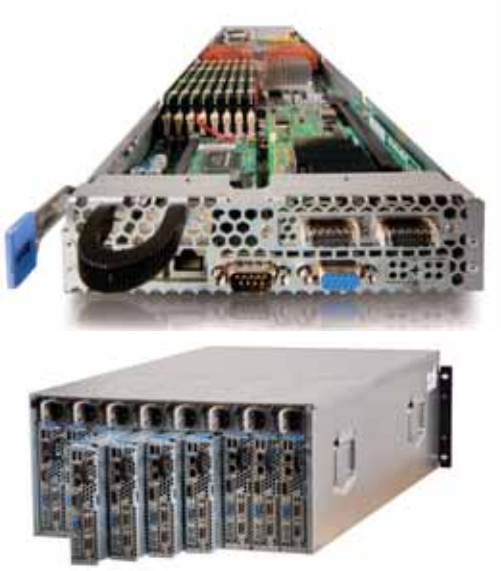

Рис. 4. Блэйд-сервер (вверху) и шасси с установленными 10 блэйд-серверами с передней стороны со снятым блоком вентиляторов (внизу), на базе которых построен суперкомпьютер "СКИФ-МГУ".

выполнения пользовательских приложений. Так, система содержит вычислительные узлы с различным количеством памяти и дискового пространства для наиболее производительной работы различных приложений с индивидуальными требованиями к ресурсам. Большая часть вычислительных узлов не содержит жестких дисков, что улучшает отказоустойчивость системы. Бездисковая загрузка ОС, в свою очередь, упрощает администрирование: при любых обновлениях достаточно изменить только единый образ ОС на управляющем узле.

Систему можно делить на независимые партиции для пользовательских приложений, гибко варьируя требования к ОП, вычислительным ресурсам, системе хранения и т.д.

Конструктивные и функциональные особенности суперкомпьютера "СКИФ-МГУ"

Суперкомпьютер "СКИФ МГУ" построен на базе 625 блейд-серверов производства "Т-Платформы" (рис. 3) с 1250 новейшими четырехъядерными 45-нм процессорами Intel® Xeon® E5472 (Harpertown) с частотой 3.0 ГГц. Его производительность на тесте Linpack была зафиксирована на уровне 47,17 TFlops (78,6% от пиковой – 60 TFlops), что является лучшим показателем эффективности среди всех систем первой сотни списка Top500 самых мощных компьютеров мира на базе четырехъядерных процессоров Intel Xeon *(www.top500.org)*.

Основу суперкомпьютера составляют *блэйд-модули* – T-Blade, позволяющие разместить 20 четырехъядерных процессоров Intel® Xeon® в шасси высотой всего 5U и обеспечивающие наибольшую вычислительную плотность среди всех представленных на рынке блэйд-решений на базе платформ Intel. Это первые блэйд-решения в отрасли с использованием нового чипсета Intel® 5400, что обеспечивает выигрыш в производительности реальных приложений до 30% и поддержку следующего поколения процессоров Intel. Модули T-Blade также совместимы с любыми

стандартными видами интерконнекта и других внешних устройств благодаря слоту расширения PCI-Express 2.0.

В качестве *системной сети* использована технология DDR InfiniBand с микросхемами компании Mellanox четвертого поколения. Всего в состав сети входят 6 144-портовых корневых коммутаторов и 54 24-портовых граничных коммутатора. Архитектура этой новейшей реализации InfiniBand не только позволяет сократить время задержки при передаче сообщений до 1,2 мкс (в реальности – 1,3–1,95 мкс в зависимости от числа коммутационных устройств при прохождении сигнала) и улучшить масштабируемость приложений, но также обеспечивает совместимость с новым, вдвое более производительным стандартом QDR InfiniBand. Скорость передачи сообщений между узлами сети составляет не менее 1450 Мбит/с.

Таким образом, архитектура кластера "СКИФ МГУ" уже сегодня ориентирована на технологии ближайшего будущего и позволяет легко и экономично модернизировать оборудование без необходимости смены блэйд-модулей. Данная архитектура и технические решения являются базовыми для 4 ряда суперкомпьютерного семейства "СКИФ".

Вспомогательная сеть построена на оборудовании Force10 (первая в России инсталляция) – лидера на рынке Ethernet-решений для HPC. В ее составе – один корневой 24-портовый коммутатор 10G Ethernet, два модульных коммутатора с 336 портами GbE и 4 портами 10G, а также два 48-портовых коммутатора GbE с 4 портами 10G Ethernet для подключения СХД.

Система хранения на основе решения компании Panasas – ReadyStorage ActiveScale Cluster – и состоит из 12 blade-шасси по 5 Тбайт с прямым параллельным обменом данными между модулями хранения и узлами кластера. Система хранения может масштабироваться без возникновения "узких мест". Производительность файловой системы – более 3 Гбайт/с. Из новых особенностей, примененных в системе хранения кластера "СКИФ МГУ", – использование технологии ActiveGuard, которая обеспечивает максимальную непрерывность работы при выходе любого компонента оборудования (модуля хранения, элемента коммутационной сети), а также за счет резервирования метаданных на модулях DirectorBlade.

Климатическая система построена по принципу герметичных "горячих коридоров", когда весь нагретый воздух попадает в полностью изолированное пространство и "забирается" вентиляторами. Холодный воздух подается 8 межрядными кондиционерами APC InRow RP с холодопроизводительностью 50кВт каждый. При этом температура в зале не превышает 20°C.

Будущее суперкомпьютинга

Своим видением перспектив развития отрасли с SN поделился Андрей Слепухин – руководитель Центра Кластерных Технологий компании "Т-Платформы".

Андрей Слепухин – руководитель Центра Кластерных Технологий "Т-Платформы".

Суперкомпьютерная отрасль как в России, так и мире бурно развивается. Реальная производительность существующих систем уже достигла порога 0,5Pflops, 7 систем превысили рубеж в 100 Tflops (4 MPP-системы и 3 кластера). Петафлопный рубеж будет преодолен уже в этом году, а уровень 10 Pflops будет "взят", как ожидается, уже в 2010–2011 гг.

Однако дальнейшее развитие суперкомпьютинга сдерживается 3 барьерами:

- *Power wall*: несмотря на то, что закон Мура остается в силе, дальнейшее увеличение производительности единичного процессора ведет к увеличению количества рассеиваемого тепла, которое невозможно отвести;
- *Memory wall*: увеличение производительности в терминах арифметических операций сдерживается низкой производительностью памяти; типичное время доступа к памяти может составлять порядка 200 тактов;
- *ILP wall*: потенциал увеличения производительности за счет улучшения архитектуры процессоров и внутреннего параллелизма практически исчерпан.

Преодоление этих барьеров возможно двумя путями: развитие архитектуры HPC-систем в сторону ее большей специализации и разработки принципиально новых технологий для элементной базы HPC-систем. Реализация первого направления лежит в следующих областях:

- переходе от архитектуры multicore к manycore: новые процессорные архитектуры: IBM Cell, Sun Niagara/Rock, Intel ???...;
- более массовом использовании различных аппаратных ускорителей: FPGA (программируемые матрицы), GPU (графические ускорители), специализированные процессоры (ClearSpeed, MD-GRAPE и др.);
- использование новых высокоскоростных средств коммуникаций;
- возможном возрождении векторных суперкомпьютеров.

Основные применения аппаратных ускорителей: молекулярная динамика, обработка сейсмических данных, обработка сигналов, финансовый анализ. В этой области уже достаточно много разработок, среди наиболее используемых FPGA: Nallatech, DRC, Celoxica, Alpha Data, GiDEL, SRC, Pico. Способы их подключения самые разные: процессорный сокет (AMD/Intel); слоты расширения PCI-X/PCI-E; слоты оперативной памяти.

Среди существующих графических ускорителей можно выделить три решения:

- *NVidia Tesla C870* (500+ Gflops – одинарная точность; 1.5GB памяти GDDR3; пропускная способность памяти 76.8GB/sec);
- *сервер Tesla S870* (4 платы C870 в корпусе 1U);
- *AMD (ATI) FireStream 9170* (500+ Gflops – одинарная точность; поддержка арифметики двойной точности; 2GB памяти GDDR3; интерфейс PCI-E 2.0).

На рынке специализированных процессоров можно отметить разработку CLEARSPEED, позволяющую достичь 1Tflops в корпусе 1U (12 плат) и 12Tflops в стандартной стойке. Это примерно в 4-6 раз лучшее соотношение цена/производительность на пакетах молекулярной динамики (Amber, MolPro).

Среди анонсированных продуктов в области интерконнекта интересны решения:

- *Mellanox ConnectX* (унифицированная архитектура, поддерживающая в одном чипе SDR/DDR/QDR InfiniBand или 10Gbit Ethernet; задержки до 1.2usec; интерфейс PCI-E 2.0);
- *Mellanox InfiniScale IV* (36 портов QDR InfiniBand на одном чипе; коммутирующая способность 2.88Tb/s; задержка коммутации 60ns; поддержка механизмов адаптивной маршрутизации);
- *QLogic QLE7240/QLE7280* (дальнейшее развитие архитектуры InfiniPath; поддержка DDR InfiniBand; задержка 1.25usec; пропускная способность 1900MB/sec; до 13 млн сообщений в секунду при 4-х процессорах).

Среди принципиально новых технологий для использования в HPC, которые уже существуют в опытных образцах или должны появиться в ближайшие 2-3 года, можно выделить:

- *Luxtera:* 40Gbps оптический трансивер, интегрированный на кристалл;
- *IBM:* оптический интерконнект на печатной плате с пропускной способностью 200Gbps;
- *Nantero:* память на углеродных нанотрубках (Non-volatile; время переключения порядка 2ns; CMOS 180nm тех. процесс; ожидаемое время выхода на массовый рынок – через 2–3 года).

Развивая направление перехода на новые процессорные архитектуры, компания "Т-Платформы" уже в конце 2007 г. объявила о начале разработки собственных решений на базе процессора Cell нового поколения и продвижения этой процессорной архитектуры на российский рынок. Первые результаты уже ожидаются к концу 2008 г.:

- семейство вычислительных модулей (сервер, рабочая станция);
- адаптация отечественного дистрибутива Linux;
- адаптация и оптимизация 3–5 прикладных программных пакетов;
- создание и поддержка сообщества разработчиков ПО для архитектуры Cell.