ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ 5(19), 2013, с. 75–116

УДК 519.6

В. В. Стегайлов, Г. Э. Норман

Проблемы развития суперкомпьютерной отрасли в России: взгляд пользователя высокопроизводительных систем

Аннотация. За прошедшее десятилетие активная господдержка ускорила развитие суперкомпьютерной отрасли в России. Сегодня в стране работают несколько суперкомпьютеров большой производительности, на которых решается все большее число научно-технических задач. Набирает силу суперкомпьютерное образование в ВУЗах. В то же время можно констатировать «однобокое» развития отрасли в отношении представленных в стране суперкомпьютерных архитектур и недостаточное развитие работ в области развития массового параллелизма при решении прикладных задач.

В статье кратко рассмотрены основные тенденции того, как развивалась архитектура и коммутационные сети лучших суперкомпьютеров мира, начиная с 1990х гг. Выделено главное направление, которое побеждает в условиях жесткой конкуренции возрастающего спроса на вычисления, использующие все большее число процессоров (вычислительных ядер) в одной задаче. Сформулированы предложения того, что нужно предпринять, чтобы суперкомпьютерная отрасль в России не отстала бы от мирового уровня, а встала бы на это же главное направление.

Критический настрой статьи ни в коей мере не нацелен на преуменьшение достигнутых успехов в развитии суперкомпьютерной отрасли нашей страны. Задача авторов состоит в попытке акцентирования внимания суперкомпьютерного сообщества России на вызовах, которые чувствует сегодня российский ученый, использующий высокопроизводительные вычисления, при проведении исследований, которые могли бы стать конкурентными в международном контексте.

Ключевые слова и фразы: топология интерконнекта, путь к экзафлопсу, масштабируемость параллельных алгоритмов, перспективные архитектуры.

Введение

В начале развития советского атомного проекта академик Ландау привлек будущего академика Тихонова к численному решению задач, которые теоретическая физики могла сформулировать для описания ядерного взрыва. С тех пор началось интенсивное развитие вычислительной математики и численных методов в Советском Союзе. Важнейшую роль в успехах этой области играло сопутствующее развитие вычислительной техники.

К сожалению, научно-технический потенциал советской науки в области аппаратно-технического обеспечения высокопроизводительных вычислений потерял былой темп роста в эпоху 1990-х годов в России. За эти тяжелые годы накопились кадровые и инфраструктурные проблемы. Выход из кризиса в 2000-х годах принес много существенных позитивных изменений. Тем не менее ситуация в нашей стране продолжает вызывать сильное беспокойство в контексте бурного развития суперкомпьютерной отрасли в мире. Успехи, достигнутые в России за последнее десятилетие, скрывают нерешенные проблемы и слабые места в инфраструктуре и научно-организационном обеспечении российской суперкомпьютерной отрасли. Описанию данных проблем и возможных путей их решения посвящена эта статья.

1. Россия в списке Топ500

Список наиболее мощных суперкомпьютеров мира Топ500 обновляется дважды в год на международных суперкомпьютерных конференциях в июне и ноябре. Первый российский суперкомпьютер появился в списке Топ500 в ноябре 2001 года. Это была система ненаучного профиля, принадлежащая Сбербанку России. В июне 2002 года в список Топ500 на 64 место попал первый российский суперкомпьютер научного профиля (кластер МВС-1000М в МСЦ РАН). Таким образом, развитие суперкомпьютерной отрасли в России вышло на режим устойчивого роста около 10 лет назад.

Информационный портал INauka.ru писал 30 июня 2005 года в связи с вводом в строй суперкомпьютера МВС-15000 в МСЦ РАН:

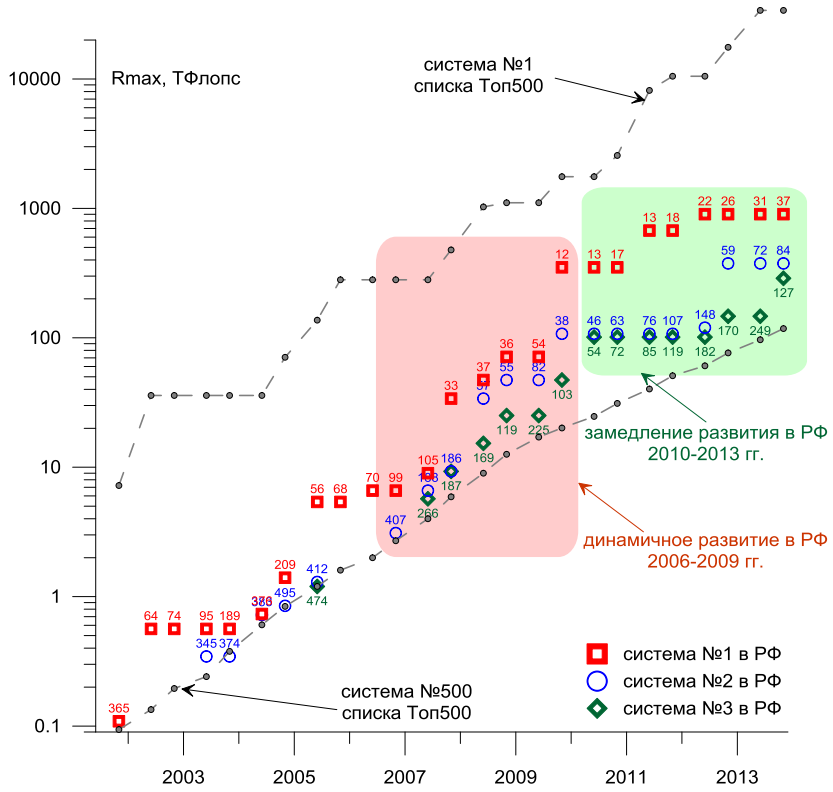

Рис. 1. Максимальная производительность лидирующих суперкомпьютеров согласно данным списка Топ500 в 2001–2013 гг.

В 1999 году был открыт Межведомственный суперкомпьютерный центр РАН, где был установлен первый российский суперкомпьютер. Это одно из немногих научных мероприятий, которое посетил тогдашний премьер-министр В. Путин. На тот момент, что подтверждено сертификатами, кластер вошел в сотню самых мощных компьютеров мира. Производить такие машины могут лишь три страны - США, Япония и Россия. С годами наш суперкомпьютер, несмотря на то, что каждый год он нака-

чивал мышцы, спускался в мировой иерархии на нижние строки. В 2005 году его обогнал суперкомпьютер СКИФ К-1000, сделанный совместно Россией и Белоруссией и установленный в Минске.

Как же выглядело развитие суперкомпьютерной отрасли у нас в стране за прошедшее десятилетие? На Рис. 1 приведены данные по глобальному росту производительности суперкомпьютеров в мире и в России. Видно, что после 2002 года на стадии первоначального формирования отрасли и накопления опыта рост российских суперкомпьютеров в общемировом рейтинге шел достаточно медленно. Повышение экспертизы суперкомпьютерного сообщества России и повышение государственного финансирования привело к динамичному развитию в 2006–2009 гг., итогом которого можно считать создание кластера «Ломоносов» МГУ имени М.В. Ломоносова, занявшего 12 позицию в Топ500. Это до сих пор непревзойденный результат. Начиная с 2010 года, рост суперкомпьютерной отрасли замедлился, и стали проявляться тенденции несбалансированного развития.

Для анализа на Рис. 1 приведены данные по трем лидирующим системам нашей страны. На рисунке показаны верхняя и нижняя границы по производительности (системы № 1 и № 500) и три самые мощные системы в России, попадающие в список Топ500 (с указанием их номера)

О динамичном развитии отрасли можно говорить, если растет как производительность суперкомпьютера-лидера, так и ближайших к нему по мощности систем. Мы видим, что период интенсивного роста, на который вышла наша суперкомпьютерная отрасль в 2006–2009 гг., сменился периодом с чертами стагнации в 2010–2013 гг. Большую часть указанного срока тройка лидеров оставалась неизменной: № 1 - кластер «Ломоносов» МГУ имени М.В. Ломоносова, № 2 - кластер МВС-100К МСЦ РАН и № 3 - кластер РНЦ КИ, аналогичный МВС-100К. Рост мощности кластера «Ломоносов» обеспечивался подключением к нему дополнительных модулей.

Ранее в 2006–2009 гг. позиция суперкомпьютера-лидера в стране «мигрировала» по различным организациям в промежутке между обновлениями списка Топ500 (МСЦ РАН — Томский Госуниверситет — МСЦ РАН — МГУ имени М.В. Ломоносова — два раза МСЦ РАН — МГУ имени М.В. Ломоносова). Менялись и системы № 2 и № 3 в России, представляя не только Москву, но и региональные центры (Уфу и Красноярск). Подобная ротация — существенное организационное условие динамичного развития отрасли — прекратилась с конца 2008 г. Только в конце 2012 г. системой № 3 среди суперкомпьютеров научных организаций стал кластер «Торнадо» в ЮУрГУ.

2. Мировые тенденции развития суперкомпьютеров: «на пути к экзафлопсу»

2.1. Развитие суперкомпьютерных архитектур

Основную задачу суперкомпьютерной отрасли можно сформулировать, как необходимость объединения большого числа вычислительных элементов для синхронизированной работы с общими данными — т.е. для решения определенной задачи.

В 1990-е годы обсуждение развития суперкомпьютерных технологий велось в терминах выбора между двумя типами систем. SMP (Symmetric Multiprocessing) — архитектура многопроцессорных компьютеров, в которой два или более одинаковых процессоров подключаются к общей памяти и, следовательно, имеют доступ к общим данным. Альтернативой была массивно-параллельная архитектура MPP (Massive Parallel Processing) — класс архитектур параллельных вычислительных систем, которые состоят из отдельных узлов, память которых физически разделена, и поэтому в процессе решения задачи необходим обмен данными между узлами.

В результате развития обе технологии перестали конкурировать и заняли каждая свое место в архитектуре современных суперкомпьютеров.

Технологические возможности наращивания производительности SMP-систем ограничены проблемой доступа к общей памяти.

Однако развитие подобных устройств — современных многоядерных процессоров и графических процессоров (GPU) — проходит достаточно интенсивно. Ранее речь шла о десятках ядер, работающих с общими данными. Сегодня (с появлением GPU) речь идет уже о тысячах. В результате подобная многоядерная комбинация процессора и специального ускорителя образуют высокопроизводительный вычислительный элемент.

Дальнейший путь наращивания производительности — а сегодня это «путь к экзафлопсу» — лежит на следующем масштабном уровне объединения сотен тысяч отдельных вычислительных элементов (=узлов) в системы, содержащие миллионы вычислительных ядер. Для решения этой задачи решающим является использование адекватной коммутационной сети (интерконнекта), объединяющей систему в единое целое.

2.2. Коммутационная сеть (интерконнект)

«Сердце» современного суперкомпьютера — коммутационная сеть (или интерконнект) проявляется на 3 уровнях:

- оборудование и топология сети, т.е. принцип физического объединения узлов каналами обмена данных;
- системное программное обеспечение, реализующее стандартные процедуры обмена данными (один-одному, один-всем, все-всем и т.п.);
- алгоритмы параллельного решения математической задачи, основанные на указанном системном программном обеспечении.

Возможные варианты аппаратного оборудования сети можно с некоторой долей огрубления разделить на два класса:

- объединение узлов коммутатором или единой шиной данных;
- объединение узлов непосредственно друг с другом, коммутация осуществляется самими узлами.

По возможной топологии эти классы представлены соответственно следующими наиболее распространенными вариантами:

- топология «толстое дерево» (Рис. 2);
- топологии типа «решетка» и «многомерный тор» (Рис. 3).

Важнейшими характеристиками сети являются пропускная способность, латентность и темп выдачи сообщений. Оптимальная производительность суперкомпьютера достигается при сбалансированном сочетании указанных характеристик и производительности узлов.

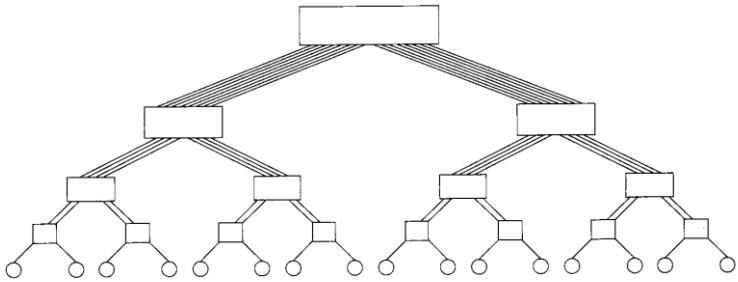

Рис. 2. Схема объединения узлов коммутаторами в топологии «толстое дерево»

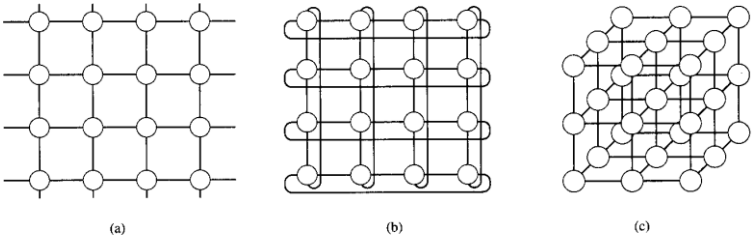

Рис. 3. Схема объединения узлов в топологии решетка и тор (a — решетка, b — 2-х мерный тор, c — 3-х мерная решетка)

Топология интерконнекта является ключевым фактором, определяющим возможность увеличения размеров суперкомпьютеров.

Древовидная топология связана с использованием коммутирующих элементов (switches). Эти устройства могут объединять десятки или даже многие сотни узлов. Однако увеличение их емкости технологически весьма сложно. Поэтому для дальнейшего «роста дерева» сами коммутаторы нужно объединять коммутато-

рами следующего уровня. При этом возникает проблема необходимости существенного роста пропускной способности каналов обмена данными. В результате применение топологии «толстое дерево» для систем более десятков тысяч элементов сталкивается с трудноразрешимыми проблемами.

Системы с топологией решетка или многомерный тор представляет собой альтернативную концепцию построения сетей (3-х, 5-ти, 6-ти мерные торы сегодня уже используются в лидирующих суперкомпьютерах списка Топ500). Основа подобной топологии — соединение узлов с ближайшими соседями по принципу топологической близости. Очевидно, что наращивание размера подобных систем не имеет серьезных технологических барьеров. Однако существенно усложняется организация обмена сообщениями — сообщения передаются от узла к узлу через цепочку узлов-посредников. С одной стороны — это замедляет передачу данных, с другой стороны — в подобных топологиях имеется множество альтернативных путей передачи (т.н. адаптивная маршрутизация). Это может компенсировать потерю в скорости и, что крайне существенно для сверхбольших систем, позволяет компенсировать возможные технические неисправности на узлах в процессе работы суперкомпьютера.

В настоящее время в стадии активного развития находятся сети со сложной топологией, наследующие принципы древовидной и тороидальной топологий, такие как «стрекоза» (dragonfly) компании Cray и PERCS (Productive Easy-to-use Reliable Computing System) компании IBM. Выбор типа коммутационной сети или его смена характеризуется большой инерционностью, т.е. архитектура сети определяет выбор системного программного обеспечения и оптимальные алгоритмы прикладных программ. Цикл полной адаптации под данную сеть можно оценить 10 годами, что определяет потребность долгосрочного планирования. Возможность «маневра» в выборе аппаратного обеспечения сети ограничивается и требованием обратной совместимости, т.е. необходимостью иметь

возможность использовать все созданное прикладное программное обеспечение на новой архитектуре.

3. «Особый путь» России в развитии суперкомпьютеров

В начале 1990-х годов на заре развития суперкомпьютерной отрасли требовалось объединять в единое вычислительное поле порядка 10–100 узлов. Создание суперкомпьютеров было уделом крупных компаний (Cray, Hitachi, IBM, Intel, NEC и др.). Создавались суперкомпьютеры с разнообразными вариантами топологии сети. В значительной степени это была эпоха экспериментирования, поиска оптимальных вариантов в условиях бурного развития мощности микропроцессоров.

К сожалению, в связи с разрушением Советского Союза в России в это время объем аналогичных работ существенно упал, что привело к технологическому отставанию. Российский журнал «Открытые системы» в 1995 году писал [1]:

Положение с разработками суперкомпьютеров в России, очевидно, оставляет сегодня желать лучшего. Работы над отечественными суперЭВМ в последние годы велись сразу в нескольких организациях. Под управлением академика В.А.Мельникова была разработана векторная суперЭВМ "Электроника СС-100" с архитектурой, напоминающей Cray-1. В ИТМиВТ РАН проводятся работы по созданию суперкомпьютеров "Эльбрус-3". Этот компьютер может иметь до 16 процессоров с тактовой частотой 10 нс. По оценкам разработчиков, на тестах LINPACK при $N = 100$ быстродействие процессора составит 200 MFL0PS, при $N = 1000$ — 370 MFLOPS. Другая разработка, выполненная в этом институте, — Модульный Конвейерный Процессор (МКП), в котором используется оригинальная векторная архитектура, однако по быстродействию он, вероятно, должен уступать "Эльбрус-3".

Другим центром работ над отечественными суперкомпьютерами является известный своими работами по ЕС ЭВМ НИЦЭВТ. Там был выполнен ряд интересных разработок — различные модели векторных суперЭВМ ЕС 1191 на ECL-технологии и

идут работы над новым суперкомпьютером "АМУР", в котором используется КМОП-технология. Ряд организаций во главе с ИПМ РАН ведут работы по созданию МРР-компьютера МВС-100, в процессорных элементах которого используются микропроцессоры i860XP, а для организации коммуникаций применяются транспьютеры T805. Хотя в наличии имеются опытные образцы некоторых из вышеупомянутых отечественных компьютеров, ни один из них промышленно не производится.

В итоге, в сложившейся в России в 1990-е годы ситуации необходимость в оборудовании для проведения высокопроизводительных расчетов приходилось удовлетворять за счет импортного оборудования [1]:

В большинстве инсталляций суперкомпьютеров используется, вероятно, продукция фирмы Convex. В нескольких организациях эксплуатируются старые модели минисуперкомпьютеров серий Clxx, C2xx, которые по производительности уже уступают современным рабочим станциям. В Санкт-Петербурге в системе Госкомвуза инсталлирована минисуперЭВМ Convex серии C3800, в Москве в ИПМ РАН недавно установлена суперкомпьютерная система SPP 1000/CD. Имеются планы инсталляции и других суперкомпьютеров (например, SGI POWER CHALLENGE) в ряде институтов РАН.

Экономические трудности не позволили российским ученым быстро и в полном объеме отреагировать на следующий шаг развития суперкомпьютерной области, ознаменованный появлением в 1995 году суперкомпьютера Cray T3E, основанном на топологии 3-х мерного тора. Эта система оказалась одной из наиболее коммерчески успешных разработок суперкомпьютеров. В 1998 году на суперкомпьютере Cray T3E впервые был преодолен барьер производительности 1ТФлопс при расчете прикладной программы. Значение, которая данная система имела для общего развития суперкомпьютерной отрасли, показывает Таблица 1.

Таблица 1. Число суперкомпьютеров Cray T3E в верхней десятке списков Top500 в 1996–1999 гг.

11/1996	06/1997	11/1997	06/1998	11/1998	06/1999	11/1999
1	6	5	8	6	5	4

Доминирование одного типа суперкомпьютера, естественно, отразилось и на тенденциях разработки системного и прикладного программного обеспечения для высокопроизводительных суперкомпьютерных расчетов.

К сожалению, в России подобная система не появилась.

Прогресс в суперкомпьютерной отрасли у нас в стране наметился в связи с появлением к концу 1990-х годов широкого спектра недорого коммерческого коммутационного оборудования типа Ethernet, Gigabit Ethernet, на основе которого можно было создавать системы, получившие название кластеров.

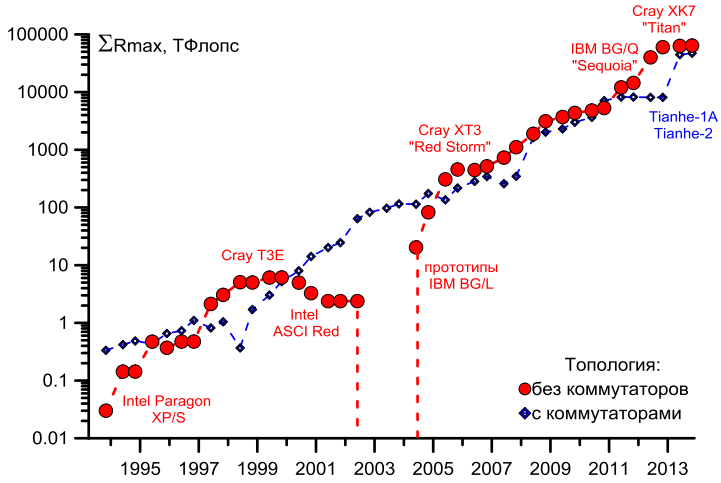

Рис. 4. Суммарная вычислительная мощность систем Rmax с коммутаторной и безкоммутаторной топологиями среди систем верхней десятки списка Top500

Создание кластеров из комплектующих стало доступным большому числу организаций и научных коллективов, как в мире, так и в России. В результате появления в начале 2000-х более производительного коммерческого коммутационного оборудования ти-

па Myrinet, Quadrics и Infiniband эпоха доминирования дорогих и «эксклюзивных» суперкомпьютеров прервалась. Существенный момент заключается в том, что все указанные коммерческие продукты были ориентированы на создание кластеров с коммутаторной топологией «толстое дерево». Именно такие суперкомпьютеры вышли в лидеры списка Top500 в 2000-х годах. На Рис. 4 отмечены названия систем, давших решающий вклад в суммарную производительность. После перерыва в 2002–2003 гг. первая система IBM BlueGene/L с топологией 3-х мерный тор попала в верхнюю десятку список Top500 в июне 2004 года. Как красноречиво иллюстрирует Рис. 4, в 2002–2003 гг. «эксклюзивные» суперкомпьютеры с тороидальной топологией интерконнекта вообще исчезли из верхней десятки Top500. Правда, исчезли ненадолго.

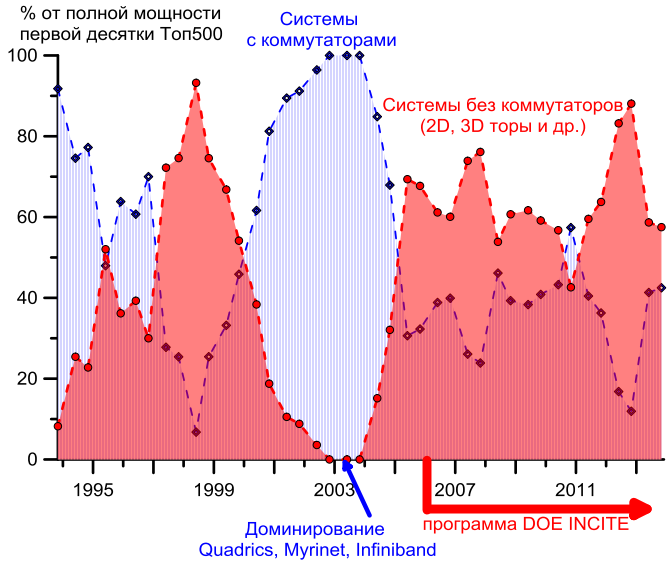

Рис. 5. Процент суммарной вычислительной мощности систем с коммутаторной и безкоммутаторной топологиями среди систем верхней десятки списка Top500

Однако именно на это время пришелся этап развития суперкомпьютерной отрасли в России, когда стали активно создаваться системы большой мощности. По-видимому, этим объясняется тот факт, что все российские суперкомпьютеры на протяжении всех десяти лет развития отрасли ориентируются на сборку из импортных комплектующих и имеют топологию «толстое дерево»[1]. Так именно в рамках топологии «толстое дерево» созданы текущие системы №№1–3 в России.

Проблема заключается в том, что процесс поиска оптимальной топологии в 2000-х годах отнюдь не закончился. И не остановился на топологии «толстое дерево» как единственном эффективном варианте. Как видно из Рис. 5, уже в 2006 году в верхней десятке списка Топ500 суммарная производительность систем с тороидальной топологией опять превысила суммарную производительность систем с топологией «толстое дерево». 2012 год показал стремительное падение удельного вклада систем топологии «толстое дерево». В списке Топ500 июня 2013 г. вклад систем с топологие «толстое дерево» возрос из-за введения в строй китайской системы Tianhe-2 с производительностью 33.86 ПФлопс. Ее детальное обсуждение выходит за рамки данной статьи, однако отметим, что такая впечатляющая производительность достигнута в большей степени за счет использования ускорителей, а не развития интерконнекта (система содержит всего 16 000 узлов).

Нельзя не отметить с глубочайшим сожалением, что начав возрождаться на волне роста производительности коммерчески доступных комплектующих, суперкомпьютерная отрасль России не сохранила темп развития собственной элементной базы. А этот вектор развития, шедший из 1980-х годов, состоял именно в развитии массивно-параллельных систем с топологией, которую впоследствии стали называть тороидальной.

Концепция массового параллелизма, вероятно, впервые проявила свой потенциал в системах на основе *транспьютеров* —

[1] Существуют лишь две небольших системы иного типа: IBM BlueGene/P ВМК МГУ имени М.В. Ломоносова и кластер СКИФ-Аврора ЮУрГУ с экспериментальным тороидальным интерконнектом.

микросхем, сочетавших на одном кристалле процессор, оперативную память и несколько коммуникационных модулей (транспьютеры могли использоваться и в связке с более производительными процессорами SPARC, Intel i860 и др.). С небольшим запаздыванием по отношению к таким зарубежным производителям как INMOS, Transtech и Meiko, в Советском Союзе были начаты работы по развитию транспьютерной элементной базы [2], [3]. Большие надежды на параллельные вычисления были связаны с созданием «компьютеров пятого поколения», от которых ожидалось возникновения систем искусственного интеллекта [4]. Оригинальной отечественной системой стала мультипроцессорная транспьютерная система МАРС на базе процессоров Кронос [5], [6], [7]. Проект дошел до стадии мелкосерийного производства, но, к сожалению, не получил дальнейшего развития.

Этапной серией отечественных суперкомпьютеров начала 1990х гг. стало (упоминавшееся в цитате выше) семейство МВС-100. Коммуникационная система этих машин строилась на основе транспьютеров, пригодных для построения топологии 2-х мерного тора. Реально использовались «квазиматрицы» — образования на основе двумерных решеток. Предполагалось, что производительность данного типа систем может достичь 100 ГФлопс, однако реальные образцы дошли до уровня 10 ГФлопс. Проект систем следующего поколения МВС-1000 первоначально разрабатывался также с ориентацией на транспьютероподобную коммуникационную систему (в ней использовались цифровые сигнальные процессоры DSP, похожие на транспьютеры и также пригодные для построения двумерных торов, решеток и т.п.) [8], [9]. Несколько таких машин (с топологией неполного двумерного тора) было построено, но вскоре стало ясно, что по основным показателям эффективности коммуникационной системы они уступают, кластерам основанным на только что появившемся коммерчески доступном интерконнекте Myrinet с топологией «толстое дерево» (прототип системы МВС-1000М с производительностью порядка 200 ГФлопс была установлен в 1999 г. в МСЦ РАН).

Необходимо отметить, что производительность транспьютерных систем МВС-100 и МВС-1000, вообще говоря, допускала их попадание в список Топ500 в середине 1990-х гг. (хотя и не на лидирующие позиции). Препятствием послужили технические сложности в запуске стандартных тестов LINPACK. В результате, только МВС-1000М стал первой российской системой в этом списке и ознаменовал эпоху доминирования в России систем с коммерческим интерконнектом типа «толстое дерево».

Развитие систем тороидальной топологии продолжалось в рамках совместной российско-белорусской программы «СКИФ». В 2003 г. была создана система СКИФ К-500 на основе SCI-интерконнекта Dolphin с топологией 3-х мерного тора [10], попавшая на 385 место в списке Топ500 ноября 2003 г. (отнесена к Беларуси). Сеть SCI, к сожалению, не долго продержалась на рынке, и последующих подобных тороидальных систем большего размера создано не было. Использование коммерческого интерконнекта (Myrinet, Infiniband и др.) было, очевидно, выгодно в то время по соображениям текущего момента: как с точки зрения итоговой производительности (определяющей место в списке Топ500), так и стоимости создаваемых систем. Доступный тороидальный интерконнект проигрывал по этим приоритетным параметрам, а специального обоснованного спроса на системы определенной топологии — не было. В результате этого процесса тороидальные сети в суперкомпьютерах по всему миру сохранились, в основном, в классе систем особо большого размера, не представленном в России.

В тоже время в рамках программы СКИФ все-таки велись работы по созданию оригинальной коммунационной сети с топологией 3-х мерного тора [11]. Прототип этого интерконнекта был установлен на системе СКИФ-Аврора в ЮУрГУ (с основной коммутационной сетью типа Infiniband).

Вероятно, первой отечественной разработкой интерконнекта для MIMD систем, доведенной до реального применения, стала сеть МВС-Экспресс [12], которая, однако, была воплощена только в древовидных топологиях. Другим текущим российским проектом по развитию интерконнекта является сеть Ангара с топологией 4-х

мерного тора, разрабатываемая в НИЦЭВТ [13], [14]. На Национальном суперкомпьютерном форуме 2013 г. (НСКФ-2013) были представлены экземпляры данных устройств, готовые к серийному производству. Работы по созданию интерконнекта «Система межпроцессорных обменов» СМПО-10G, подразумевающего создание и тороидальных систем [15], ведутся РФЯЦ-ВНИИЭФ совместно с НИИИС им. Ю. Е. Седакова.

Таблица 2. Число и суммарные параметры систем с тороидальной топологией по странам мира в списке Топ500 ноября 2013 г.

Регион/страна	Кол-во	Число ядер	Rmax, ТФлопс	% от ΣRmax	Название
Россия	1	8192	23.9	0.01	IBM[2] – 1.
Европа	23	1 678 384 (+317 856)	23 148.5	9	IBM – 9, Cray – 14.
Австралия, Бразилия, Канада, Саудовская Аравия	5	204 384 (+ 1 248)	1 670.6	0.7	IBM – 3, Cray – 2.
США	28	4 780 928 (+338 280)	59 589.8	24	IBM – 14, Cray – 13, Sun – 1.
Япония и Южная Корея	9	1 036 928 (+18 048)	14 098.0	6	Fujitsu -3, IBM – 2, Cray – 4.
Итого по Топ500 ноября 2013 г.:	65	7 700 624 (+428 744)	98 316.0	39	Fujitsu -3, IBM – 28, Cray – 33, Sun – 1.

Анализ распределения в мире систем с тороидальной топологией показывает, что в версию списка Топ500 ноября 2013 г. входят 65 систем. При этом в США располагаются 28 из них. Из 250 ПФлопс суммарной вычислительной мощности суперкомпьютеров, входящих в список Топ500 ноября 2013 г. на системы с тороидаль-

[2] Российская система IBM BlueGene/P была № 381 в списке за ноябрь 2009 года. В текущий список не входит.

ной топологией приходится 98.3 ПФлопс, т.е. 39%, *включая семь систем из первой десятки лидеров* (см. Рис. 5).

Таблица 2 иллюстрирует очевидный факт лидерства США в мировом развитии суперкомпьютерной отрасли. В таблице показано количество систем с тороидальной топологией, суммарное число ядер (в скобках — прирост по отношению списку ноября 2012 г.), суммарная производительность, вклад в суммарную производительность текущего списка Top500 (250 ПФлопс), распределение систем по производителям (Китай пока не представлен в Top500 такими суперкомпьютерами). В то же время европейские страны не допускают образования качественного разрыва в техническом оснащении, и в силу того, что собственных производителей подобной техники в Европе (пока) нет, для этого активно закупаются американские суперкомпьютеры IBM и Cray. За прошедший год было установлено 9 новых Cray XC30 (Австралия — 1, Германия — 4, Швейцария — 1, Великобритания — 1, Япония — 2) и одна новая система IBM BlueGene/Q в Великобритании.

Япония, затратив в 2006–2012 гг. беспрецедентный объем госинвестиций (более $1.2 млрд.), разработала собственный интерконнект с топологией 6-ти мерного тора и уже построила 4 суперкомпьютера на его основе. К сожалению, Россия в этом списке — явный аутсайдер. У нас имеется только одна небольшая система IBM BlueGene/P и нет ясных продекларированных планов создания или закупки новых подобных систем[3].

В то же время приведенные данные однозначно свидетельствуют, что суперкомпьютеры с тороидальной топологией являются кандидатами на роль систем экзафлопсного класса.

Различные архитектуры суперкомпьютеров имеют свои сильные и слабые стороны. Нельзя утверждать, что текущий тренд развития в сторону тороидальной топологии указывает на окончательное решение проблемы выбора оптимальной топологии интер-

[3] Новые системы должны представлять собой не минимальные примеры, пригодные только для тестовых исследований (1 стойка IBM BlueGene/Q — 1024 узла), а быть машинами для практического использования (минимум 10–20 тыс. узлов).

коннекта «на пути к экзафлопсу». В данный момент по заказу оборонного агенства США DARPA компания IBM разрабатывает интерконнект PERCS, который можно рассматривать как развитие топологии «толстое дерево». Одновременно компания Cray разрабатывает топологию «стрекоза» (dragonfly), имеющую черты как тороидальной, так и древовидной структуры.

Очевидно, что суперкомпьютерная архитектура неразрывно связана с системным и прикладным программным обеспечением, а разработка и совершенствование программного обеспечения требует наличия соответствующего оборудования. Поэтому «однобокое» развитие аппаратной базы суперкомпьютерной отрасли в России приводит к тому, что отечественные разработчики не имеют возможности разрабатывать программное обеспечение для перспективного типа систем будущей экзафлопсной эры.

Опять, как и 20 лет назад, доминирующие позиции заняли суперкомпьютеры, поставляемые компаниями IBM и Cray — готовые решения «под ключ». Коммерчески доступное оборудование для создания сетей с топологией тор от компании Mellanox появилось на рынке [15]–[16], но пока что имеется мало информации об успешных примерах его использования [17]. При этом существенным является эффект запаздывания: указанный коммерческий продукт компании Mellanox появился через 7–9 лет после начала интенсивного развития тороидальных систем IBM BlueGene и Cray. Можно с уверенностью предположить, что такие технологические новинки, как сети с топологией 5-ти и 6-ти мерного тора и сеть PERCS, если и появятся на рынке, то с таким же запаздыванием.

Очевидный вывод из приведенной ретроспективы развития отрасли состоит в том, что и государство, и российские компании-интеграторы должны инициировать и поддерживать развитие отечественных разработок интерконнекта для обеспечения независимого и своевременного вхождения России «в эру экзафлопса».

4. Отдельные проблемы развития суперкомпьютерной отрасли в России

Отмеченное «однобокое» развитие суперкомпьютерной области у нас в стране за последние 10 лет с абсолютным доминированием систем одной топологии, на наш взгляд, уже привело к определенному инфраструктурному, технологическому и методическому отставанию. Выделим несколько отдельных проблем.

4.1. Массовый параллелизм: необходимость или главная цель?

Глобальный путь развития суперкомпьютеров лежит в направлении увеличения числа вычислительных элементов (размера вычислительного поля), объединенных интерконнектом. Текущий уровень и скорость развития производительности узлов превосходит уровень и скорость развития коммутационной сети. Поэтому не удивительно, что основной фокус внимания сосредоточен на разработках в области интерконнекта.

Ярким примером является серия IBM BlueGene, которая началась с системы BlueGene/L, основанной на процессоре PowerPC 440 с тактовой частотой 700 МГц. «Слабый» процессор был выбран для снижения энерговыделения и облегчения плотной компоновки. Разработчики поставили цель добиться высокой производительности за счет совершенствования интерконнекта. Система BlueGene/L получила 3 сети обмена данными, которые могут использоваться одновременно. В результате в списке Топ500 июня 2004 г. для BlueGene/L с 4096 ядрами указана производительность 8.7 ТФлопс, а для BlueGene/L с 8192 ядрами — 11.7 ТФлопс. В том же списке для системы с интерконнектом Myrinet и 2500 одноядерными процессорами Intel Xeon 3.06 ГГц указана производительность 9.8 ТФлопс.

Таким образом, развитие технологии — естественно, требовавшее существенного финансирования — шло по пути массового параллелизма. Хотя с точки зрения производительности системы меньшего размера на существующем интерконнекте обеспечивали на тот момент такую же производительность.

Сегодня мы видим результат этого пути развития. Как показано на Рис. 6, суперкомпьютер IBM BlueGene/Q объединил в одно целое уже более 1.5 миллионов ядер! На рисунке символами показаны три самые мощные системы в России, попадающие в список Топ500 (с указанием их номера). Серым пунктиром показаны данные для систем № 1 и № 500 списка Топ500. Для системы № 1 под точками указано число вычислительных элементов. Пунктиром выделен уровень 1 млн. ядер.

Примечательно, что системы IBM BlueGene, объединяющие короткими связями большое число относительно простых вычислительных элементов, могут рассматриваться как новый виток развития идей создателей транспьютерных систем [18].

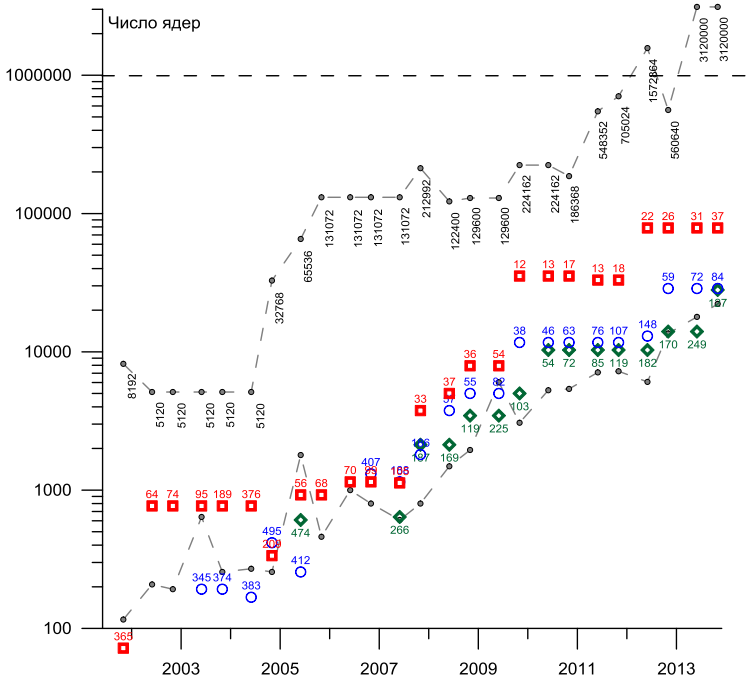

Рис. 6. Число вычислительных элементов в лидирующих суперкомпьютерах согласно данным списка Топ500 в 2001–2013 гг.

На Рис. 6 следует отметить еще одну особенность, а именно — тот факт, что формат представления данных о числе ядер, принятый в Топ500, отчасти скрывает реальное устройство этих систем. Таблица 3 показывает устройство узлов трех лидирующих суперкомпьютеров текущего списка. Видно, что производительность систем № 1 и № 2 достигается за счет увеличения производительности узлов. Причем в случае Tianhe-2 рост производительности узлов явно противоречит эффективности их использования из-за сложности организации совместного доступа к данным для двух ЦПУ и трех ускорителей. Cray Titan в этом отношении более сбалансированная система. С точки зрения массового параллелизма очевидным лидером является IBM Sequoia, каждый узел которой обладает лишь одним 18 ядерным процессором Power PC A2 с теоретической производительностью 205 ГФлопс (2 ядра — служебные).

Таблица 3. Особенности устройства суперкомпьютеров №№ 1–3 текущего списка Топ500 июня 2013 г.

	Кол-во узлов	Кол-во вычисл. ядер ЦПУ	Узел		
			ЦПУ (кол-во ядер)	Ускоритель (кол-во ядер)	Суммарная теор. произв. узла, ТФлопс
Tianhe-2	16 000	384 000	2 x Intel Xeon (2 x 12)	3 x Intel Phi (3 x 57)	3.43
Cray Titan	18 688	299 008	AMD Opteron (16)	Nvidia Kepler (14 потоковых мультипроцессора)	1.45
IBM Sequoia	98 304	1 572 864	IBM PowerPC (16 + 2)		0.205

Таблица 4 показывает сравнение реальной производительности суперкомпьютеров с суммарной теоретической производительностью их вычислительных элементов. В таблице указаны суммарная теоретическая производительность (Rpeak) их вычислительных элементов, их отношение, потребляемая мощность и удельная производительность на Вт по данным списка Green500. Для сравнения приведены данные для кластера «Ломоносов».

По этим данным видно, что система IBM BlueGene/Q более эффективно использует объединенную вычислительную мощность узлов, чем Cray XK7. При этом потребляемая мощность электропитания практически одинакова. Еще один немаловажный плюс системы IBM BlueGene/Q — это использование вычислительных элементов, не требующих специальных методик программирования. Вычислительные узлы Cray XK7 требуют специального переноса (переписывания) прикладных программ с использованием технологии Nvidia CUDA. Отметим, что для систем типа Tianhe-2 характерна аналогичная ситуация: несмотря на использование ядер с архитектурой x86 в ускорителях Intel Phi, эффективный параллельный алгоритм должен учитывать особенности внутренней организации узлов. Для многих прикладных программ подобная адаптация и оптимизация алгоритмов требует не менее 1 года. При наличии соответствующих специалистов.

Поэтому можно утверждать, что уровень технологии в суперкомпьютерной области в большей степени определяется не производительностью в петафлопсах, а коммуникационной сетью и количеством вычислительных элементов, которые могут быть эффективно ею объединены. Анализ отечественных систем по размеру вычислительного поля (Рис. 6) демонстрирует достаточно медленное сокращение отрыва от мировых лидеров в этом отношении.

Таблица 4. Реальная производительность (Rmax) суперкомпьютеров Cray «Titan» и IBM «Sequoia» по Топ500 ноября 2013 г.

	Rmax, ПФлопс	Rpeak, ПФлопс	Rmax/Rpeak, %	МВт	ГФлопс /Вт
Tianhe-2	33.7	54.9	61%	17.8	0.733
Cray XK7 «Titan»	17.6	27.1	65%	8.21	2.143
IBM BlueGene/Q «Sequoia»	17.2	20.1	81% $\rightarrow$ 85%	7.89	2.177
Кластер «Ломоносов»	0.9	1.7	53%	2.8	0.322

4.2. Необходимы новые тесты производительности: LINPACK устарел!

Стандартным способом сравнения суперкомпьютеров является измерение их производительности по тесту LINPACK. Соответствующие результаты, измеренные в единицах Флопс, ранжируют системы в списке Top500. Однако необходимо учитывать, что тест LINPACK представляет собой достаточно специфический набор задач линейной алгебры. Этот тест был предложен для тестирования производительности суперкомпьютеров в 1979 году. В то время задачи линейной алгебры составляли значительную часть приложений из набора потенциальных задач для высокопроизводительных вычислений. К настоящему времени ситуация сильно изменилась, и к результатам теста LINPACK не следует относиться как к главному критерию эффективности суперкомпьютера, делать из него «священную корову».

Зарубежный опыт демонстрирует примеры передовых систем, создатели которых не планировали попадание в Top500. Японский суперкомпьютер «Protein Explorer», созданный на основе чипов MDGRAPE-3 для задач молекулярной динамики, достиг мощности 1 ПФлопс еще в 2003 году. Создаваемая в данный момент система IBM Blue Waters на основе гибридного интерконнекта PERCS — возможный будущий лидер в суперкомпьютерной области — скорее всего не будет проходить тесты для попадания в список Top500 [19].

В настоящее время активно развиваются альтернативные наборы тестов, которые нацелены на имитирование нескольких типов современных алгоритмов высокопроизводительных вычислений. Сложность выработки новых критериев состоит в том, что в области высокопроизводительных вычислений сформировалось несколько подобластей с различной специализацией. Решение научно-технических задач требует физико-математического моделирования на основе широкого класса вычислительных методов. Попыткой их репрезентативного представления, является, например, набор тестов Mantevo [20]. С другой стороны алгоритмы обработки больших объемов данных и веб-поиска сформировали другой класс

вычислительных методов (т.н. Data Intensive Supercomputer — DIS), представленный набором тестов Graph500 [22].

К сожалению, в России при планировании нового суперкомпьютера попадание в список Топ500 на основе теста LINPACK зачастую рассматривается как самоцель. Эта точка зрения, на наш взгляд, приводит к перекосам при развитии суперкомпьютерной отрасли. Зачастую, для повышения параметров суперкомпьютера по тесту LINPACK достаточно максимально повысить мощность вычислительных элементов, жертвуя тем самым принципами массового параллелизма и препятствуя, тем самым, развитию новых алгоритмов и программного обеспечения экзафлопсного уровня.

4.3. Ввод и вывод данных: проблема без решения

Как для задач крупномасштабного суперкомпьютерного физико-математического моделирования, так и для задач обработки больших объемов данных отдельной важнейшей технической задачей на пути в «эру экзафлопса» является организация параллельного ввода-вывода с сотен тысяч и миллионов ядер. Сохранение «точек останова» счета и перезапуск задач при использовании сверх больших размеров вычислительного поля даже с применением современных технологий может занимать больше времени, чем сам процесс расчета. Требуется разработка принципиально новых решений в этой области. Речь идет об организации специального уровня интерконнекта для задач ввода-вывода, дополняющего сеть обмена данными. Родственной является проблема обработки и визуализации данных расчетов «на лету». При этом решение подобных задач чаще всего привязано к конкретной проблеме. Внимание к этим задачам в мире уделяется, но не акцентируется (например, нет специальных рейтингов, сопоставляющих суперкомпьютеры в этом отношении). В России эта тематика пока еще практически не развивается, по-видимому, из-за отсутствия выраженного спроса со стороны пользователей.

4.4. Системное программное обеспечение: новые тенденции не замечены?

Отставание в производительности суперкомпьютерных систем может рассматриваться как количественное отставание (т.е. мы решаем задачи на мировом уровне, но в меньшем количестве и дольше). К сожалению, отсутствие суперкомпьютеров с перспективной топологией интерконнекта и отставание в размере вычислительного поля является уже в большей степени качественным отставанием. Это значит, что российские исследователи вообще лишены возможности разрабатывать алгоритмы мирового уровня.

С начала 2000-х годов активное развитие систем с топологией типа тор стимулирует целую область исследований на стыке computer science и вычислительной математики, а именно — исследование алгоритмов коллективных коммуникаций в сетях с подобной топологией. В связи с отсутствием подобных суперкомпьютерных систем в России данная область развивается крайне медленно.

В качестве примеров зарубежных работ этой тематики можно привести две публикации ученых из США [23] и [24]. На работу [23] за период немногим больше 3 лет была дана 31 ссылка в научной периодике (авторы цитирующих статей из Австрии, Германии, Индии, Италии, Испании, Китая, США и Турции). На работу [24] за период немногим больше 1 года было дана 21 ссылка в научной периодике (авторы цитирующих статей из Бельгии, Германии, Индии, Испании, Китая и США).

Пример данных статей показывает высочайшую востребованность в мире исследований в данной области. К сожалению, косвенно подобный анализ литературы выявляет скромный вклад российских ученых в данном направлении. Результаты исследований по разработке вариантов российского интерконнекта представлены всего в нескольких статьях [11], [12], [13], [14], [15].

Появление технологии параллельного программирования MPI (совпавшее с появлением рейтинга Топ500) в значительной степени скрыло архитектурные различия высокопроизводительных систем. Однако технологические трудности продвижения «по пути к экзафлопсу» показывают, что оптимизация приложений на сверх-

больших вычислительных полях не может не учитывать деталей их аппаратного устройства. Эффективное использование системы с тороидальным интерконнектом основано на использовании топологически определенных сегментов системы (например, «замкнутых» сегментов из 512 ядер) и специальной организации расположения данных на узлах с учетом топологии для минимизации обменов. Проблемы такого рода для систем с топологией «толстое дерево» выражены слабее и, главное, требуют принципиально иных подходов для решения. На уровне узлов стандарт MPI тоже не может «автоматически» обеспечить высокую эффективность. Большинство современных гибридных GPU-систем основано на связке MPI+CUDA. На первый взгляд архитектурно однородные системы с ускорителями Intel Phi также требуют решения проблемы MPI-обменов как внутри одного ускорителя, так и внутри одного узла.

4.5. Параллельные алгоритмы и прикладное программное обеспечение: почти катастрофическое отставание!

Еще один аспект качественного отставания российской суперкомпьютерной отрасли обусловлен тем, что создание алгоритмов решения конкретных задач неразрывно связано с топологией интерконнекта и системным программным обеспечением. Достижение максимальной параллельной эффективности — это тонкий процесс, основанный на учете специфики решаемой задачи и специфики имеющейся суперкомпьютерной техники. Можно констатировать, что текущая ситуация определяет исключение российских ученых из важнейшего мирового тренда развития вычислительных методов для высокопроизводительных расчетов.

Эффективность алгоритма распараллеливания определяется уменьшением времени расчета при росте числа задействованных вычислительных элементов. В идеальном случае при росте вычислительного поля с 10 до 100 ядер решение задачи должно пройти в 10 раз быстрее. В действительности передача данных между узлами приводит к отклонению от идеальной масштабируемости. Параллельный алгоритм тем лучше, чем большее вычислительное

поле можно эффективно использовать. Например, реальный случай может выглядеть так: при переходе с 10 на 100 ядер время расчета падает в 7 раз, при переходе с 10 на 1000 ядер — падает в 20 раз, при переходе с 10 на 2000 ядер — возрастает в 2 раза. В последнем случае распараллеливание на такое большое вычислительное поле становится неэффективным. Основная задача создателя параллельного алгоритма — повысить подобную границу потери параллельной эффективности.

По нашим субъективным наблюдениям (статистика не известна), большинство задач, решаемых на отечественных суперкомпьютерах, используют порядка 10–100 вычислительных ядер (без учета GPU ускорителей). Большее количество ядер на практике не востребовано из-за потери эффективности распараллеливания используемых кодов. Малое число задач использует порядка 1000 ядер. Задачи, которые эффективно решаются с использованием порядка 10000 ядер, в российской практике редки.

В то же время, развитие суперкомпьютерных технологий в США (и в Европе) встало на путь радикального наращивания числа вычислительных элементов. Как было отмечено выше, серия систем IBM BlueGene была с самого начала задумана для развития технологии массового параллелизма. В 2006 году, когда системы IBM BlueGene прошли начальную апробацию и появились суперкомпьютеры Cray XT3/4 (Рис. 5), для развития алгоритмов параллельного решения математических задач на этих новых системах Департамент энергетики США (DOE) расширил уже существовавшую ранее программу Innovative and Novel Computational Impact on Theory and Experiment (INCITE) [25]. Так в 2012 году на 60 проектов (Таблица 5) было выделено 1672 миллиона процессорчасов на суперкомпьютерах IBM BlueGene/P и Cray XT5.

Таблица 5. Число проектов, получивших компьютерное время на системах IBM BlueGene L/P/Q и Cray XT3/4/5/XK7 по программе INCITE в 2006–2014 гг.

Год	2006	2007	2008	2009	2010	2011	2012	2013	2014
Число проектов	15	45	58	66	70	57	60	61	59

Показательным является тот факт, что в программе INCITE участвуют системы только с тороидальной топологией.

Как показано на Рис. 7, тематика проектов INCITE охватывает практически все области науки и техники. На рисунке показано (а) число проектов по каждой тематике и (б) выделенное на каждую тематику вычислительное время в миллионах процессорочасов. Исходя из доступного краткого описания, все проекты с некоторой долей условности разделены на следующие тематики: атомистические модели (Ab initio — расчеты из первых принципов и MD – молекулярная динамика), модели в рамках механики сплошных сред (CFD), астрофизические модели (Astrophys), физика плазмы (Plasma), квантовая хромодинамика (QCD), физика ядра (Nuclear), computer science (CS). В программу принимаются проекты с международным участием, даже без партнера из США.

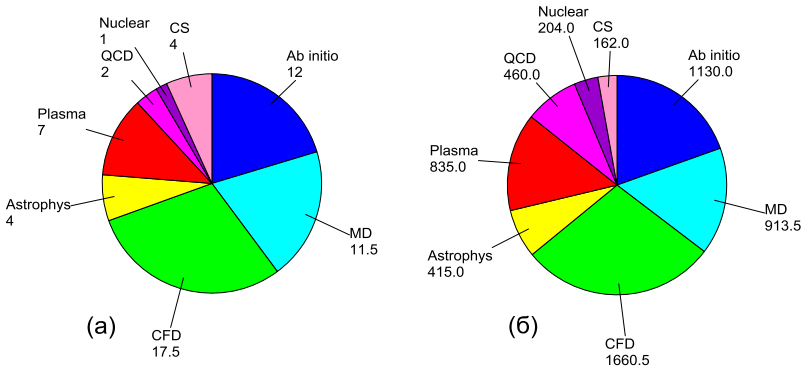

Рис. 7. Деление проектов INCITE 2014 г. по тематикам.

В рамках программы INCITE традиционно распределяется 60% вычислительного времени на одной системе IBM и одной системе Cray Департамента энергетики США. Интересные данные представляет анализ распределения проектов между ними по выделенному вычислительному времени. Таблица 6 показывает суммарные квоты времени, выделенные на каждой из двух систем в 2010 и

2014 гг. для проектов INCITE. Сравнение показывает, что доля выделяемого вычислительного времени определяется числом процессорных ядер, а не вычислительной мощностью систем.

Таблица 6. Распределение вычислительных ресурсов по программе INCITE в 2010 и 2014 гг.

Система	Число ядер	Rpeak, ПФлопс	Rmax, ПФлопс	Время, млн. проц.-часов
2010				
Cray XT5 «Jaguar»	224 162	2.331	1.759	940
IBM BlueGene/P «Intrepid»	163 840	0.557	0.459	646
Соотношение	1.4	4.2	3.8	1.5
2014				
Cray XK7 «Titan»	299 008 /560 640	27.11	17.59	2250
IBM BlueGene /Q «Mira»	786 432	10.07	8.59	3530
Соотношение	0.4/0.7	2.7	2	0.64

Обязательным условием получения расчетного времени по программе INCITE является использование массового параллелизма. При этом задачи, связанные с одновременным запуском большого числа однотипных задач (тривиальный параллелизм для набора статистики), рассматриваются, однако не являются приоритетными [25]. От алгоритма решения прикладной задачи требуется демонстрировать параллельную эффективность на вычислительном поле порядка 20% используемой машины. В 2011 году это соответствовало размеру вычислительного поля порядка 20 тысяч ядер, а в 2013–14 гг. — 200 тысяч ядер!

Эти цифры демонстрируют переход к новой эре использования вычислительных методов в науке и технике.

Очевидно, что в результате работ в рамках программы INCITE и аналогичных ей (ALCC и др.) будет накоплен опыт разработки программ для суперкомпьютеров экзафлопсного класса тороидальной архитектуры и созданы коды для предсказательного физико-математического моделирования во всех областях науки и техники.

Естественно, аналогичное развитие имеет место и в области технологий обработки больших объемов данных (DIS).

К сожалению, сегодня российские исследователи практически лишены возможности участвовать в подобных международных проектах из-за отсутствия необходимой инфраструктуры для предварительной отладки задач. Даже № 1 в России кластер «Ломоносов» может предоставить только 50 тыс. процессорных ядер!

Можно констатировать, что к настоящему времени с точки зрения приложений суперкомпьютерная отрасль в России прошла этап накопления критического объема опыта [26], [27]. Во многих областях исследователи активно используют параллельные вычисления в различных областях науки и техники. Имеется ярко выраженный дефицит вычислительного времени на ресурсах коллективного доступа. Для решения большого числа задач пользователи активно привлекают как программные решения с открытым кодом, так и коммерческие продукты. Мировая практика показывает, что разработка параллельного программного обеспечения перестает быть частным делом узкого коллектива специалистов и становится предметом больших коллаборативных (зачастую международных) проектов. Примером одной из областей, демонстрирующей активный спрос на параллельные вычисления, является молекулярное и атомистическое моделирование, расчеты электронной структуры и квантовая химия [28], [29]. Значительный прогресс в достижении параллельной эффективности на системах с тороидальной архитектурой достигается с помощью распределения данных по узлам с учетом топологии интерконнекта. Таким образом удается существенно минимизировать объемы межузловых обменов данными. С помощью подобных приемов удается добиться относительно высокой параллельной эффективности даже для такого чрезвычайно требовательного к обменам класса задач, как задачи расчета электронной структуры (см., например, [30]). Различные примеры хорошо масштабируемых атомистических моделей были представлены нашим коллективом на НСКФ-2013 [31].

Прикладную направленность проектов INCITE можно проиллюстрировать работой 2011-2012 гг., посвященной моделированию кровотока в артериях головного мозга [32]. Единая многомасштабная модель объединяет гидродинамическое описание течения крови в рамках модели Навье-Стокса и детальное описание холестериновых частиц методом молекулярной динамики (см. Рис. 8).

Для решения одной задачи в работе используется от 20 до 300 тыс. ядер. Рассматривается структура участка системы артерий. Структура подразделяется на области, каждая из которых рассчитывается на вычислительном поле порядка 5–70 тыс. ядер. Таблица 7 показывает возможность высокой параллельной эффективности расчетов данной модели на размерах вычислительного поля до 200–300 тыс. ядер.

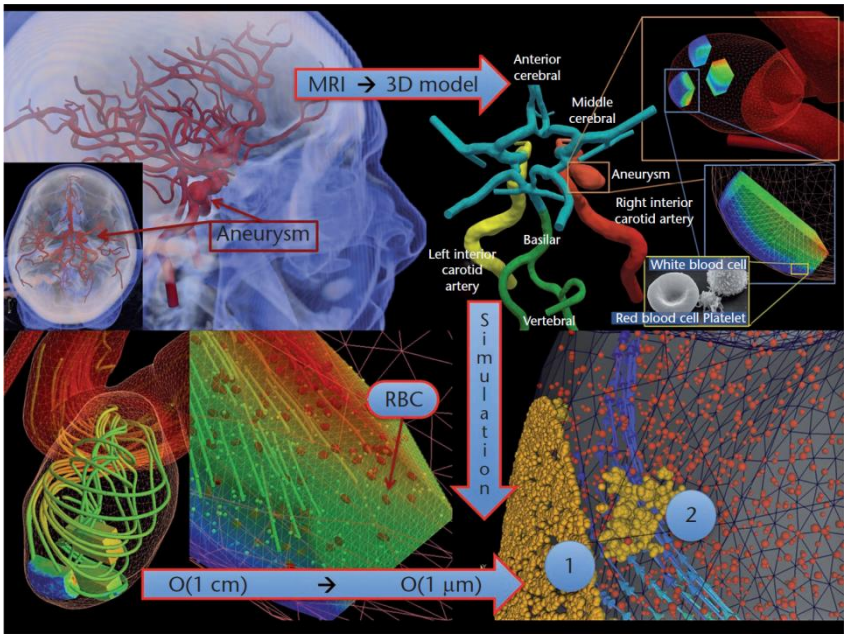

Рис. 8. Многомасштабная модель кровотока с эффектами отложения холестериновых бляшек, сочетающей атомистическое моделирование с моделью жидкости Навье-Стокса [32]

Исследование, выполненное в приведенной работе [32], состоит из следующих частей:

(1) связь континуального и дискретного алгоритмов с разными временными шагами;
(2) обмен данными в процессе расчета между частями системы и между континуальным и дискретным уровнями;
(3) оптимизация производительности с учетом специфики систем IBM BlueGene/P и Cray XT5;
(4) анализ параллельной эффективности на обеих системах.

Пример работы [32] показывает нетривиальность самого процесса использования новых суперкомпьютерных возможностей, начиная с процесса задания начальных условий и заканчивая обработкой больших массивов выходных данных. Решение подобных проблем, очевидно, имеет высокую научно-техническую ценность.

Таблица 7. Демонстрация сохранения эффективности распараллеливания решения одной задачи на системах BlueGene/P и Cray XT5 до 295 тыс. и 191 тыс. ядер соответственно (из работы [32])

Число ядер	Время загрузки ЦПУ, с	Эффективность
BlueGene/P (4 ядра на узел)		
32 768	3580.34	1.00
131 072	861.11	1.04
262 144	403.92	1.07
294 912	389.85	0.92
Cray XT5 (12 ядер на узел)		
21 396 (Kraken)	2194	
30 036 (Kraken)	1177	1.24
38 676 (Jaguar)	806	1.10
97 428 (Jaguar)	280	1.07
190 740 (Jaguar)	206	0.68

Обратившись к истории развития связей аппаратной и программной составляющих суперкомпьютерной отрасли в мире с точки зрения физического моделирования можно выделить одну важную «развилку». В 1992 г. в одном из наиболее престижных физи-

ческих журналов Physical Review Letters одновременно вышло две статьи, описывающие результаты расчетов на основе квантово-механических первопринципных (ab initio) моделей с использованием массивно-параллельных алгоритмов [33], [34]. В целях демонстрации новых возможностей в этих статьях расчитаны свойства поверхности кристаллического кремния, имеющие прямое отношение к решению фундаментальных проблем микроэлектроники. Для расчетов использовались две различные массивно-параллельные системы: тороидальная система Meiko Computing Surface (использовалось вычислительное поле из 64 транспьютерных узлов с процессорами Intel i860) и система типа гиперкуб Thinking Machines CM-2 (использовалось 16384 однобитовых процессора совместно с 512 арифметическими ускорителями Weitek). Указанное аппаратное обеспечение стало доступно в 1986–1987 годах и, таким образом, на разработку программного обеспечения ушло, по-видимому, всего около 3–4 лет. Можно с уверенностью утверждать, что созданный в то время методический задел определил сегодняшние лидирующее позиции зарубежных коллективов в области программного обеспечения для компьютерного материаловедения, вычислительной химии и других смежных областей, связанных с атомистическим моделированием. Для задач других типов (газовая динамика, квантовая хромодинамика и др.) подобный задел начал формироваться в то же время (причем и в Советском Союзе — см., например, [35]).

Эпоха потрясений начала 1990-х годов помешала российским ученым принять активное участие в развитии программного обеспечения для массивно-параллельных систем в момент их появления. За прошедшее двадцатилетие отставание удалось сократить. Однако уже наступивший новый этап развития суперкомпьютерных массивно-параллельных вычислительных технологий XXI века еще не вызывал должного отклика в России и требует незамедлительных и решительных действий.

Выводы: нам пора менять профессию или уезжать?

Критическое отношение, выражаемое нами в данной работе к суперкомпьютерам с топологией «толстое дерево», не означает, что такие системы не нужны, устарели и от них необходимо отказаться. На данном этапе развития суперкомпьютеры с древовидной и тороидальной топологиями можно условно уподобить различным видам транспорта — автомобилям и самолетам. И тот, и другой вид техники необходим для решения задач своего класса. При этом странно развивать транспорт, принимая во внимание только автомобили и игнорируя авиацию. Однако в суперкомпьютерной области России в данный момент сложилась подобная ситуация.

Способность быстро обрабатывать большие объемы информации и проводить предсказательное моделирование в различных областях науки и техники — это ключевые факторы, определяющие конкурентоспособность в современном мире. Создание суперкомпьютерных систем экзафлопсного класса — очередной рубеж на пути развития соответствующих технологий. При этом развитие технологий коммуникационных сетей совместно с соответствующим промежуточным и прикладным программным обеспечением — важнейший инструмент для преодоления данного барьера. А решение передовых задач сверхбольших масштабов (grand challenges) является движущей силой инновационного развития всей суперкомпьютерной отрасли.

Можно выделить следующие особенности развития суперкомпьютерной отрасли в России за последние десять лет.

(1) В начале интенсивного развития, начавшегося в 2002 г., объем появлявшихся суперкомпьютерных ресурсов опережал спрос со стороны научно-технического сообщества. Именно в результате десятилетнего развития такой спрос постепенно вырос. И сегодня можно говорить о возможности формулировать конкретные потребности (таким примером является и данная статья);

(2) На протяжении первых 5 лет (ориентировочно до 2007 г.) наблюдалась недозагрузка многих введенных в эксплуатацию в России суперкомпьютеров. В этот переходный период форми-

ровались как коллективы, обеспечивающие работу суперкомпьютеров, так и коллективы их пользователей, осваивавшие соответствующие вычислительные методики. К сожалению, за этот же период у многих руководителей сформировалось мнение о невозможности полноценного использования уже созданных суперкомпьютеров. Вероятно, этот факт отчасти объясняет замедление развития отрасли в 2010–2013 гг. В настоящее время ситуация уже стала прямо противоположной — имеет место большой дефицит вычислительного времени и вычислительных ресурсов;

(3) Развитие аппаратно-технического оснащения суперкомпьютерной отрасли проходит в условиях отсутствия явно продекларированных пожеланий со стороны пользователей. При подобной неопределенности важнейшим критерием успешности является производительность суперкомпьютера, формальное определение которой постепенно утрачивает связь с реальными задачами. Это приводит к «однобокому» развитию, в значительной степени противоречащему магистральному направлению прогресса суперкомпьютерной отрасли в мире.

Необходимо принятие срочных мер по ликвидации аппаратно-технического отставания российской суперкомпьютерной отрасли. Для этих целей,

во-первых, необходимо оснастить научные центры страны оборудованием мирового уровня зарубежных производителей;

во-вторых, необходимо существенно ускорить развитие отечественного аппаратного обеспечения (в первую очередь — интерконнекта) и довести его уровень в кратчайшие сроки до уровня установки на больших системах и тестирования практических задач;

в-третьих, необходимо продекларировать требования создания/адаптации программного обеспечения на технологиях массового параллелизма мирового масштабного уровня.

Для обсуждения плана действий по указанным направлениям необходимо коллегиальное обсуждение с привлечением всех заинтересованных организаций и ведомств, например, в рамках Национальной суперкомпьютерной технологической платформы.

При отсутствии изменений подобного рода российские специалисты, связанные с использованием высокопроизводительных вычислений в разных областях науки и техники, через 1–2 года окажутся перед дилеммой: менять профессию или уезжать в страны с суперкомпьютерной инфраструктурой высшего уровня.

Предложения по развитию суперкомпьютерной инфраструктуры

Национальный суперкомпьютерный форум 2013 г. в Переславле-Залесском (НСКФ-2013), на наш взгляд, однозначно продемонстрировал, что в России существуют «точки роста», активное развитие которых способно изменить текущую безрадостную ситуацию. Считаем возможным сформулировать следующие предложения организационного характера.

(1) Для ликвидации технологического отставания необходимо организовать закупку новейших суперкомпьютеров для создания новых центров коллективного пользования (ЦКП) и модернизации существующих. Исходя из личного общения с представителями фирм-производителей на НСКФ-2013, сложилось впечатление, что существовавшие ранее запреты и ограничения на поставку в Россию вычислительной техники экстра-класса существенно ослаблены, и фирмы-производители охотно рассмотрели бы соответствующие заявки;

(2) Необходима ротация лидирующих по мощности суперкомпьютеров в России по разным организациям (3–5 крупнейшим ЦКП), чтобы каждая следующая лидирующая по мощности система вводилась в строй другим коллективом специалистов;

(3) В целях стимулирования перехода научных коллективов на вычислительные методики крупномасштабного параллелизма необходимо ввести конкурсные процедуры, аналогичные INCITE и PRACE, для решения задач на вычислительных полях особенно большого размера. Попытки организации такого рода специальных вычислительных сессий уже были ранее предприняты МСЦ РАН, а в этом году и НИВЦ МГУ имени М.В. Ломоносова на суперкомпьютере «Ломоносов».

Предложения научно-технического характера.

(4) Необходимо всемерное усиление собственных исследований в области перспективного интерконнекта и соответствующего системного и программного обеспечения. Уже имеются действующие реализации интерконнекта МВС-Экспресс [12] (ИПМ им. М.В. Келдыша РАН и ФГУП «НИИ Квант») и 3-х мерного тороидального интерконнекта СКИФ-Аврора [11] (ИПС им. А.К. Айламазяна РАН). На НСКФ-2013 были представлены работающие образцы 4-х мерного тороидального интерконнекта Ангара [14] (НИЦЭВТ). Интерконнект СМПО-10G [15] (РФЯЦ-ВНИИЭФ) также находится на стадии внедрения;

(5) Основными критериями для анализа эффективности использования суперкомпьютера должны быть: 1) размер вычислительного поля (число узлов, процессоров и ядер), используемого при решении каждой отдельной задачи, и 2) параллельная эффективность используемых кодов, т.е. высокая степень ускорения расчетов на больших размерах вычислительного поля. Необходима классификация систем по размеру вычислительного поля и по классам соответствующих им задач (аналогично делению Tier0, Tier1, Tier2, принятому в программе PRACE);

(6) Для продвижения «по пути к экзафлопсу» наряду с интерконнектом особое внимание необходимо уделить проблеме параллельного ввода-вывода данных при размерах вычислительного поля порядка сотен тысяч и миллионов элементов.

Благодарности. *Авторы признательны С. М. Абрамову, Е. Р. Куштанову, А. О. Лацису, А. С. Семенову, А. И. Слуцкину и С. А. Степаненко за полезные советы и обсуждения на НСКФ-2013. Работа выполнена при финансовой поддержке РФФИ (проект 13-08-12070-офи_м).*

Список литературы

[1] Волков Д., Кузьминский М. Современные суперкомпьютеры: состояние и перспективы // Открытые системы. 1995. № 06. URL: http://www.osp.ru/os/1995/06/178750/

[2] Бахтияров С. Д., Дудников Е. Е., Евсеев М. Ю. Транспьютерные технологии. М. Радиосвязь. 1992. 250 с.

[3] Дудников Е. Е. Мир ПК. 1993. Транспьютеры — новое средство построения параллельных архитектур // № 6. С. 15–22. Транспьютерные ускорители для ПК // № 8. С. 30–35. Программное обеспечение транспьютерных усилителей // № 9. С. 56–59.

[4] Широков Ф. В. На пути к пятому поколению компьютеров / Международный научно-исследовательский институт проблем управления. М., 1985. 170 с.

[5] Масалович А. Советский суперкомпьютер — возможен ли он сегодня? // Russian Dr. Dobb's Journal. 1991. №. 3. С. 12–15. URL: http://iam.ru/world/pic/super1991.doc

[6] http://www.kronos.ru/

[7] Богатырев Р. Язык как основа архитектуры. Проект «Кронос» и путь к технологиям XDS // ComputerWeek-Moscow. 1998. № 20.
URL: http://www.computer-museum.ru/histussr/kronos.htm

[8] Zabrodin A. V., Levin V. K., Korneev V. V. The massively parallel computer system MBC-100 // Lecture Notes in Computer Science, No. 964. Parallel Computing Technologies. Third International Conference, PaCT-95, St. Petersburg, Russia, Sept. 1995, Springer. P. 341–355.

[9] Забродин А. В., Левин В. К., Корнеев В. В. Массово параллельные системы МВС-100 и МВС-1000 // Научная сессия МИФИ-2000. Т. 2 Информатика и процессы управления. Информационные технологии. Сетевые технологии и параллельные вычисления. С. 194–195. URL: http://library.mephi.ru/data/scientific-sessions/2000/2/778.html

[10] http://www.cnews.ru/news/top/index.shtml?2003/12/16/153053#

[11] Адамович И. А., Климов А. В., Климов Ю. А., Орлов А. Ю., Шворин А. Б. Опыт разработки коммуникационной сети суперкомпьютера «СКИФ-Аврора» // Программные системы: теория и приложения. 2010. Т. 1. Вып. 3. С. 107–123.

[12] Елизаров Г. С., Горбунов В. С., Левин В. К., Лацис А. О., Корнеев В. В., Соколов А. А., Андрюшин Д. В., Климов Ю. А. Коммуникационная сеть МВС-Экспресс // Вычислительные методы и программирование. 2012. Т. 13. С. 103–109.

[13] Корж А. А., Макагон Д. В., Бородин А. А., Жабин И. А., Куштанов Е. Р., Сыромятников Е. Л., Черемушкина Е. В. Отечественная коммуникационная сеть 3D-тор с поддержкой глобально адресуемой памяти // Вест. ЮУрГУ — серия «Мат. модел. и программирование». 2010. Т. 211. № 35. С. 41–53.

[14] Слуцкин А. И. Симонов А. С., Жабин И. А., Макагон Д. В., Сыромятников Е. Л. Разработка межузловой коммуникационной сети ЕС8430 «Ангара» для перспективных российских суперкомпьютеров // Успехи совр. радиоэлектроники. 2012. №1.

[15] Басалов В. Г., Вялухин В. М. Адаптивная система маршрутизации для отечественной системы межпроцессорных обменов СМПО-10G // Вопросы атомной науки и техники. Серия: Математическое моделирование физических процессов. 2012. № 3. С. 64–70.

[16] Epperson M., Naegle J., Schutt J., Bohnsack M., Monk S., Rajan M., Doerfler D. HPC Top 10 InfiniBand Machine — A 3D Torus IB Interconnect on Red Sky // Sonoma Workshop. 2010. URL: http://www.openfabrics.org/archives/sonoma2010.htm

[17] Wilde T. 3D Torus for InfiniBand // HPC Advisory Council Switzerland Conference. 2012. URL: http://www.hpcadvisorycouncil.com/events/2012/Switzerland-Workshop/

[18] 3D Torus Topology with InfiniBand at San Diego Supercomputing Center // HPCwire. 2012. URL: http://www.hpcwire.com/hpcwire/2012-01-30/3d_torus_Topology_with_infiniband_at_san_diego_supercomputing_center.html

[19] Mielke U. Transputer Architecture and Occam — the Fascination of early, true Parallel Computing Anno 1983. URL: http://os.inf.tu-dresden.de/EZAG/abstracts/abstract_20130712.xml

[20] Blue Waters Opts Out of TOP500 // HPCwire. 2012. URL: http://www.hpcwire.com/hpcwire/2012-11-16/blue_waters_opts_out_of _top500.html

[21] Mantevo Project Home Page. URL: http://www.mantevo.org

[22] The Graph 500 List. URL: http://www.graph500.org

[23] Faraj A., Kumar S., Smith B., Mamidala A., Gunnels J., MPI collective communications on the Blue Gene/P supercomputer: algorithms and optimizations // 17th IEEE Symposium on High Performance Interconnects. 2009. P. 63–72.

[24] Chen D., Eisley N. A., Heidelberger P., Senger R. M., Sugawara Y., Kumar S., Salapura V., Satterfield D. L., Steinmacher-Burow B., Parker J. J. The IBM BlueGene/Q interconnection network and message unit // Proc. Int. Conf. for High Perf. Computing, Networking, Storage and Analysis (SC), 2011.

[25] INCITE program. URL: http://www.doeleadershipcomputing.org/

[26] Суперкомпьютерные технологии в науке, образовании и промышленности / Под ред.: академика В. А. Садовничего, академика Г. И. Савина, чл.-корр. РАН Вл. В. Воеводина. — М.: Издательство МГУ, 2012. С. 42–49.

[27] Sadovnichy V., Tikhonravov A., Voevodin V., Opanasenko V. "Lomonosov": supercomputing at Moscow State University // In Contemporary High Performance Computing: From Petascale toward Exascale; Vetter J. S. Ed., CRC Press: Boca Raton, FL, 2013. P. 283–307.

[28] Янилкин А. В., Жиляев П. А., Куксин А. Ю., Норман Г. Э., Писарев В. В., Стегайлов В. В. Применение суперкомпьютеров для молекулярно-динамического моделирования процессов в конденсированных средах // Вычислительные методы и программирование. 2010. Т. 11. С. 111–116.

[29] Жиляев П. А., Стегайлов В. В. Ab initio молекулярная динамика: перспективы использования многопроцессорных и гибридных суперЭВМ // Вычислительные методы и программирование. 2012. Т. 13. С. 37–45.

[30] Hutter J., Curioni A. Car-Parrinello molecular dynamics on massively parallel computers // Chemical Physics 2005. V. 6. P. 1788–1793.

[31] Куксин А. Ю., Ланкин А. В., Морозов И. В., Норман Г. Э., Орехов Н. Д., Писарев В. В., Смирнов Г. С., Стариков С. В., Стегайлов В. В., Тимофеев А. В. Предсказательное моделирование свойств и многомасштабных процессов в материаловедении. Для каких задач нужны суперкомпьютеры эксафлопсного класса? // Труды Национального суперкомпьютерного форума (НСКФ-2013), ИПС им. А.К. Айламазяна РАН, г. Переславль-Залесский.

[32] Grinberg L., Insley J. A., Fedosov D., Morozov V., Papka M. E., Karniadakis G. E. Tightly coupled atomistic-continuum simula-

tions of brain blood flow on petaflop supercomputers // Computing in Science and Engineering. 2012. V. 14. N. 6. P. 58-67.

[33] Štich I., Payne M. C., King-Smith R. D., Lin J.-S. Ab initio total-energy calculations for extremely large systems: application to the Takayanagi reconstruction of Si(111) // Phys. Rev. Lett. 1992. V. 68. P. 1351–1354.

[34] Brommer K. D., Needels M., Larson B. E., Joannopoulos J. D. Ab initio theory of the Si(111)-(7x7) surface reconstruction: a challenge for massively parallel computation // Phys. Rev. Lett. 1992. V. 68. P. 1355–1358.

[35] Дуйсекулов А. Е., Елизарова Т. Г. Использование многопроцессорных вычислительных систем для реализации кинетически-согласованных разностных схем газовой динамики // Математическое моделирование. 1990. Т. 2. № 7. С. 139–147.

Рекомендовал к публикации Программный комитет
Второго национального суперкомпьютерного форума НСКФ-2013

Об авторах:

Владимир Владимирович Стегайлов

Доктор физико-математических наук, заведующий отделом Объединенного института высоких температур РАН, доцент кафедры молекулярной физики Московского физико-технического института (Государственного университета).

e-mail: stegailov@gmail.com

Генри Эдгарович Норман

Доктор физико-математических наук, главный научный сотрудник Объединенного института высоких температур РАН, профессор кафедры молекулярной физики Московского физико-технического института (Государственного университета).

e-mail: genri.norman@gmail.com

Образец ссылки на публикацию:

В. В. Стегайлов, Г. Э. Норман. *Проблемы развития суперкомпьютерной отрасли в России: взгляд пользователя высокопроизводительных систем* // Программные системы: теория и приложения: электрон. научн. журн. 2013. Т. 4, № 5(19), с. 75–116.

URL: http://psta.psiras.ru/read/psta2013_5_75-116.pdf

V. V. Stegailov, G. E. Norman. *Challenges to the supercomputer development in Russia: a HPC user perspective.*

Abstract. Over the past decade, active government support accelerated development of supercomputer industry in Russia. Today, there are several large supercomputers of large performance, which solve a growing number of problems. Supercomputer education in Russia is maturing. At the same time, one can observe an unbalanced development with respect to the supercomputer architectures, and a lack of massive parallelism in the solution of applied problems.

The article briefly describes the main trends of how the supercomputer architecture and interconnect evolved since the 1990s. The main trends highlighted, which is winning in a competitive environment of increasing demand for high performance computing using a growing number of processors (cores) for a given problem. We discuss what could be done to put Russia into the same main trend.

Critical attitude of this article seeks in no way to underestimate the progress in the development of Russian supercomputer industry. This is an attempt to focus the attention of the community on the challenges we feel today using HPC in research that could be competitive in the international context. (in Russian).

Key Words and Phrases: Interconnect topology, way to exaflops era, scalability of parallel algorithms, perspective architectures.