

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 эксафлопсу, 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 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 Top500

Список наиболее мощных суперкомпьютеров мира **Top500** обновляется дважды в год на международных суперкомпьютерных конференциях в июне и ноябре. Первый российский суперкомпьютер появился в списке Top500 в ноябре 2001 года. Это была система ненаучного профиля, принадлежащая Сбербанку России. В июне 2002 года в список Top500 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:

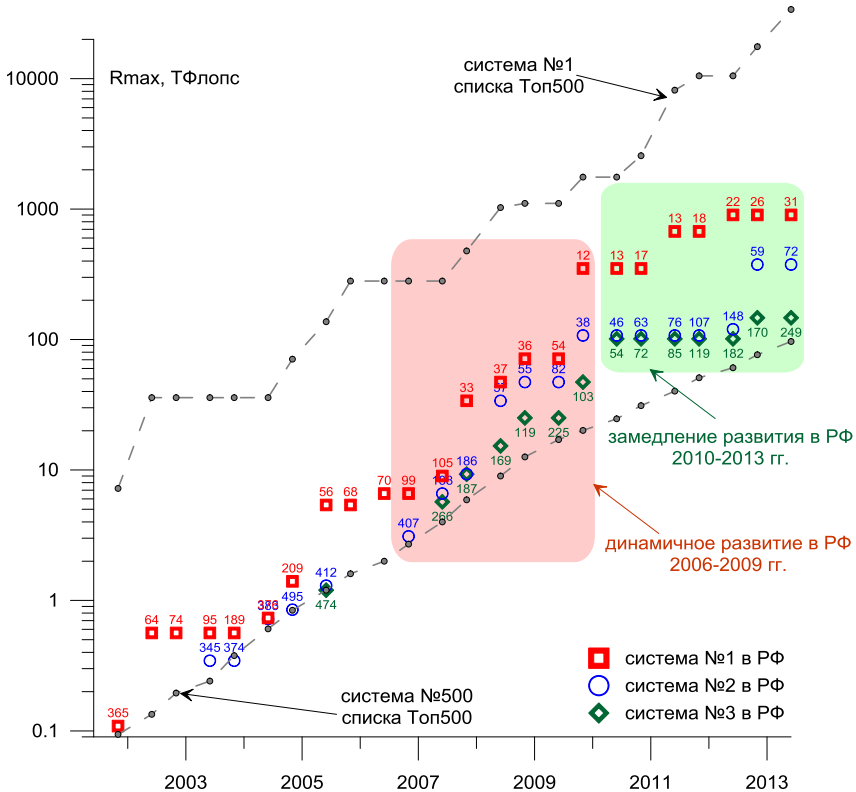


Рис. 1. Максимальная производительность лидирующих суперкомпьютеров согласно данным списка Top500 в 2001-2013 гг. Показаны верхняя и нижняя границы по производительности (системы №1 и №500) и три самые мощные системы в России, попадающие в список Top500 (с указанием их номера)

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 сот-

ню 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 накачивал 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 им. М.В.Ломоносова, занявшего 12 позицию в Top500. 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.

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 им. 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 меж-

ду обновлениями списка Топ500 (МСЦ РАН – Томский Госуниверситет – МСЦ РАН – МГУ им. М.В.Ломоносова – два раза МСЦ РАН – МГУ им. 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 эксафлопсу»

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, работаю-

щих 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)

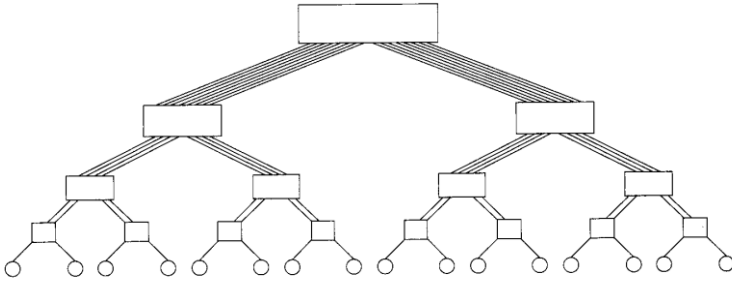


Рис. 2. Схема объединения узлов коммутаторами в топологии «толстое дерево».

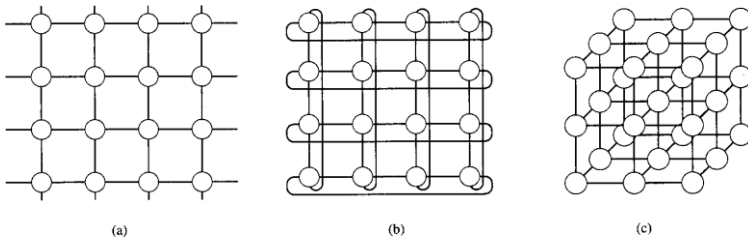


Рис. 3. Схема объединения узлов непосредственно друг с другом в топологии решетка и тор (а – решетка, б – 2-х мерный тор, с – 3-х мерная решетка).

Топология является ключевым фактором, определяющим возможность роста (масштабирования) размера суперкомпьютера.

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 коммутаторами 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 трудно-разрешимыми 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 трех 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 делом 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 MFLOPS, 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 НИ-ЦЭВТ. 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 MPP-компьютера МВС-100, в процессорных элементах которого используются микропроцессоры i860XP, а для организации коммуникаций применяются транспьютеры T805. Хотя в наличии имеются опытные образцы неко-

торых 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 C1xx, 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 Топ500 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 недорогого коммерческого коммутационного оборудования типа Ethernet, Gigabit Ethernet, на основе которого можно было создавать системы, получившие название кластеров.

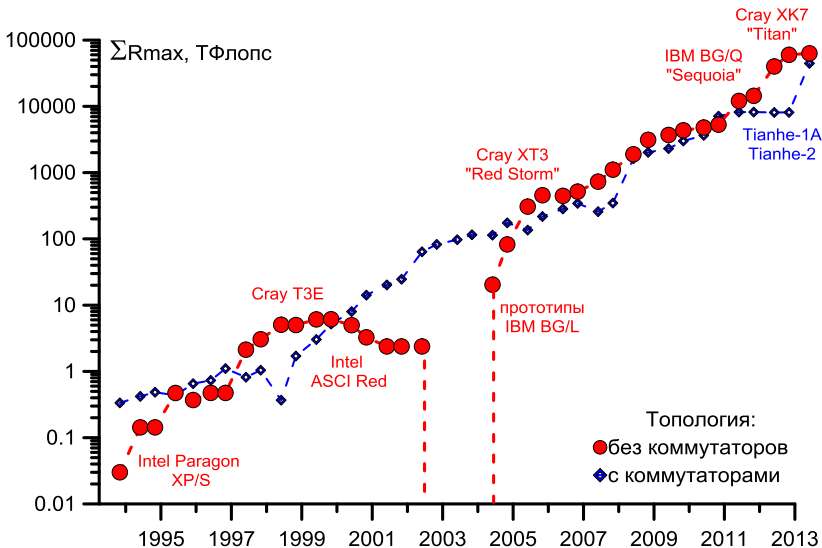


Рис. 4. Суммарная вычислительная мощность систем Rmax с коммутаторной и безкоммутаторной топологиями среди систем верхней десятки списка Top500. Отмечены названия систем, давших решающий вклад в суммарную производительность. После перерыва в 2002-2003 гг. первая система IBM BlueGene/L с топологией 3-х мерный тор попала в верхнюю десятку список Top500 в июне 2004 года

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 типа 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. Как красноречиво иллюстрирует Рис. 4: в 2002-2003 гг. «эксклюзивные» суперкомпьютеры с тороидальной топологией интерконнекта вообще исчезли из верхней десятки Top500. Правда, исчезли ненадолго.

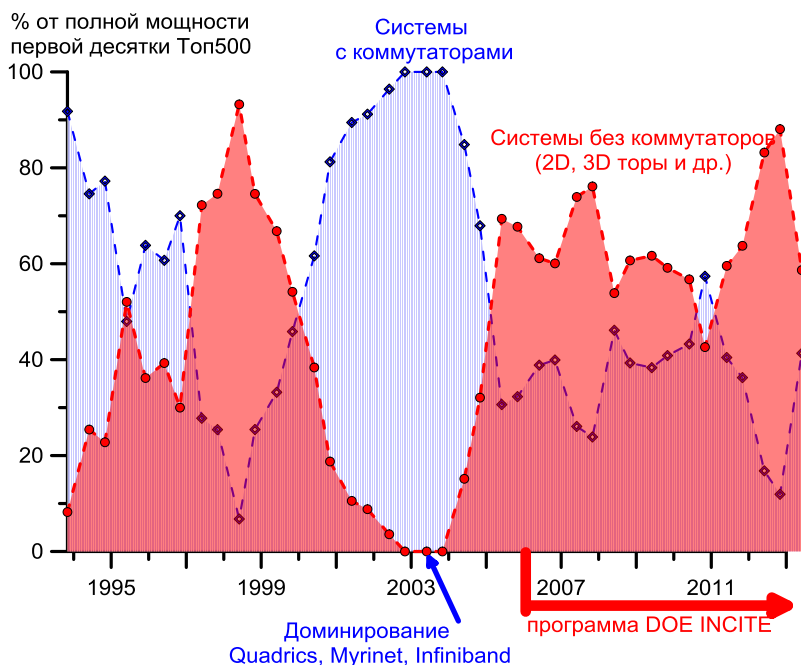


Рис. 5. Процент суммарной вычислительной мощности систем с коммутаторной и безкоммутаторной топологиями среди систем верхней десятки списка Top500

Однако именно на это время пришлось зарождение современной суперкомпьютерной отрасли в России (ознаменованное созданием в 2002 г. кластера МВС-1000М в МСЦ РАН). 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.

ТАБЛИЦА 2. Число и суммарные параметры систем с тороидальной топологией по странам мира в текущем списке Топ500 июня 2013 г.: количество подобных систем, суммарное число ядер (в скобках – прирост по отношению к предыдущему списку), суммарная производительность, вклад в суммарную производительность текущего списка Топ500 (224 Пфлопс), распределение систем по производителям. (Китай пока не представлен в Топ500 такими суперкомпьютерами)

Регион/страна	Кол-во	Число ядер	$\Sigma R_{\max}$, ТФлопс	% от $\Sigma R_{\max}$	Название
Россия	1	8192	23.9	0.01	IBM ² – 1.
Европа	17	1 406 240 (+45 712)	14 371.6	6.4	IBM – 8, Cray – 9.
Австралия, Бразилия, Канада, Сауд.Аравия	4	194 944 (- 8 192)	1 465.8	0.65	IBM – 3, Cray – 1.
США	35	4 810 064 (+367 416)	59 812.9	26.7	IBM – 18, Cray – 16, Sun – 1.
Япония и Южная Корея	10	1 042 688 (+23 808)	14 202.9	6.3	Fujitsu -3, IBM – 2, Cray – 5.
Итого:	67	7 462 128 (+428 744)	89 877.1	40.1	Fujitsu -3, IBM – 32, Cray – 31, Sun – 1.

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

¹ Существуют лишь две небольших системы такого типа: IBM BlueGene/P ВМК МГУ им. М. В. Ломоносова и экспериментальный тороидальный интерконнект части кластера «Аврора» ЮУрГУ.

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

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 топологии «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).

Анализ распределения в мире систем с тороидальной топологией показывает, что в версию списка Топ500 ноября 2013 г. входят 67 систем. При этом в США располагаются 35 из них. Из 224 ПФлопс суммарной вычислительной мощности суперкомпьютеров, входящих в список Топ500 июня 2013 г. на системы с тороидальной топологией приходится 89.9 ПФлопс, т.е. 40%, включая шесть систем из первой десятки лидеров (см. Рис. 5).

Приведенная статистика (Таблица 2) иллюстрирует очевидный факт лидерства США в мировом развитии суперкомпьютерной отрасли. В тоже 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.

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, насколько нам известно, нет ясных продекларированных планов создания или закупки новых подобных систем.

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 интерконнекта «на пути к экзафлопсу». В данный момент по заказу оборонного агентства 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 [2][3], но пока что имеется мало информации об успешных примерах его использования [4]. 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 эксафлопса».

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 Top500 июня 2004 для BlueGene/L с 4096 ядрами указана производительность 8.7ТФлопс, 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.

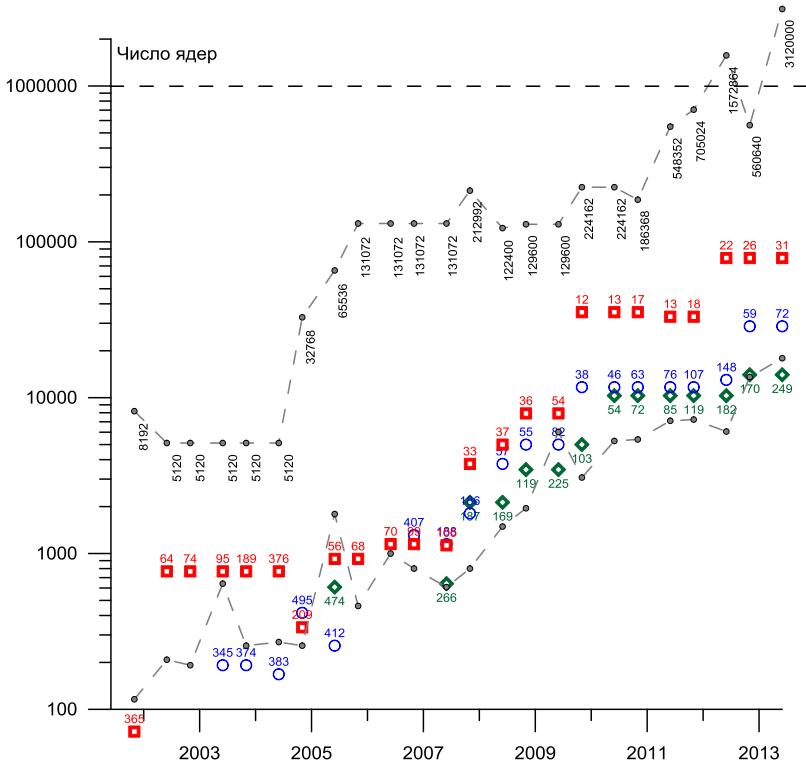


Рис. 6. Число вычислительных элементов в лидирующих суперкомпьютерах согласно данным списка Top500 в 2001-2013 гг. Символами показаны три самые мощные системы в России, попадающие в список Top500 (с указанием их номера). Серым пунктиром показаны данные для систем №1 и №500 списка Top500. Для системы №1 над точками указано число вычислительных элементов. Пунктиром выделен уровень 1 млн. ядер

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. Суперкомпьютер IBM BlueGene/Q объединил в одно целое уже более 1.5 миллионов ядер (см. РИС. 6)!

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

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).

Следующая таблица показывает сравнение реальной производительности суперкомпьютеров с суммарной теоретической производительностью их вычислительных элементов.

ТАБЛИЦА 4. Реальная производительность ($R_{\max}$) суперкомпьютеров Cray XK7 «Titan» и IBM BlueGene/Q «Sequoia» по данным Top500 (июнь 2013 г.), суммарная теоретическая производительность (R_{peak}) их вычислительных элементов, их отношение и потребляемая мощность. Для сравнения приведены данные для кластера «Ломоносов»

	$R_{\max}$, ПФлопс	R_{peak} , ПФлопс	$R_{\max}/R_{\text{peak}}$, %	МВт
Cray XK7 «Titan»	17.6	27.1	65%	8.21
IBM BlueGene/Q «Sequoia»	17.2	20.1	85%	7.89
Кластер «Ломоносов»	0.9	1.7	53%	2.8

По этим данным видно, что система BlueGene/Q более эффективно использует объединенную вычислительную мощность узлов. При этом потребляемая мощность электропитания практически одинакова. Еще один немаловажный плюс системы BlueGene/Q – это использование вычислительных элементов, не требующих специальных методик программирования. Вычислительные узлы Cray Titan 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 порядка 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.

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 [5].

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 [6]. 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 [7].

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.

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 эксафлопса» 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 [8] и [9]. На работу [8] 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 [9] 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 двух вариантов российского интерконнекта представлены всего в нескольких статьях. По-видимому, основными статьями в рецензируемых журналах являются работы [10] и [11].

Появление технологии параллельного программирования MPI (совпавшее с появлением рейтинга Top500) в значительной степени скрыло архитектурные различия высокопроизводительных систем.

Однако технологические трудности продвижения «по пути к эксафлопсу» показывают, что оптимизация приложений на сверхбольших 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: нарастающее отставание!

Еще один аспект качественного отставания российской суперкомпьютерной отрасли связан с 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. Па-

раллельный 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 нужно 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) [12]. Так в 2012 году на 60 проектов было выделено 1672 миллиона процессор-часов на суперкомпьютерах IBM BlueGene/P и Cray XT5.

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

ТАБЛИЦА 5. Число проектов, получивших компьютерное время на системах IBM BlueGene L/P/Q и Cray XT3/4/5/XK7 по программе INCITE в 2006-2013 гг.

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

Тематика проектов INCITE охватывает практически все области науки и техники. Принимается большое число проектов с международным участием, даже без партнера из США.

Обязательным условием получения расчетного времени 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 [12]. 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 году – 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 будет накоплен как опыт разработки программ для суперкомпьютеров экзафлопсного класса тороидальной архитектуры, так 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!

атомистическое моделирование, расчеты электронной структуры и квантовая химия [14][15]. Существенный 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, [16]).

ТАБЛИЦА 6. Демонстрация сохранения эффективности распараллеливания решения одной задачи на системах BlueGene/P и Cray XT5 до 126 тыс. и 186 тыс. ядер соответственно (из работы [17])

Число ядер	Время загрузки ЦПУ, с	Эффективность
BlueGene/P (4 ядра на узел)		
28 672	3205.58	
61 440	1399.12	1.07
126 976	665.79	1.02
Cray XT5 (12 ядер на узел)		
17 280 (Kraken)	2194	
25 920 (Kraken)	1177	1.24
34 560 (Jaguar)	806	1.10
93 312 (Jaguar)	280	1.07
186 624 (Jaguar)	206	0.68

Прикладную направленность проектов INCITE можно проиллюстрировать работой 2011 г., посвященной моделированию кровотока в артериях головного мозга [17]. Единая многомасштабная модель объединяет гидродинамическое описание течения крови в рамках модели Навье-Стокса и детальное описание холестеринových частиц методом молекулярной динамики (см. Рис. 7).

Для решения одной задачи в работе используется от 20 до 200 тыс. ядер. Рассматривается структура участка системы артерий. Структура подразделяется на области, каждая из которых рассчитывается на вычислительном поле порядка 5-50 тыс. ядер. Таблица 6 показывает возможность высокой параллельной эффективности

расчетов данной модели на размерах вычислительного поля до 150-200 тыс. ядер.

Исследование, выполненное в приведенной работе, 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 данной работы 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.

Заключение: пора менять профессию?

Во время переходной эпохи 1990-х гг. многие новейшие тенденции в суперкомпьютерной отрасли не были замечены. В результате в начале 2000-х имела место нехватка кадров, знающих технические и системные аспекты создания и работы с суперкомпьютерами. Создание и использование новых суперкомпьютеров стало уделом оставшихся профессионалов-энтузиастов. Так же мало было научно-технических коллективов, имеющих адекватные задачи и готовых их решать на появившихся высокопроизводительных системах. В результате, на протяжении первых 5 лет (ориентировочно до 2007 г.) в среднем по России наблюдалась недозагрузка введенных в эксплуатацию суперкомпьютеров. В этот переходный период формировались 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 супер-

компьютеров. Вероятно, этот факт отчасти объясняет замедление развития отрасли в 2010-2013 гг. Уместно выделить две особенности этого процесса:

- (1) В начале 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 (примером такой попытки является и данная статья).
- (2) Развитие аппаратно-технической области суперкомпьютерной отрасли проходило практически в условиях отсутствия явно продекларированных пожеланий со стороны потенциальных пользователей. При подобной неопределенности важнейшим критерием успешности развития стала формальная производительность суперкомпьютера, постепенно утрачивающая связь с реальными задачами. Это привело к «однобокому» развитию, в значительной степени отклоняющемуся от магистрального направления прогресса суперкомпьютерной отрасли в мире.

Необходимо принятие срочных мер по ликвидации аппаратно-технического отставания российской суперкомпьютерной отрасли. Считаем возможным сформулировать следующие меры:

Во-первых, необходимо дополнительно оснастить научные центры страны (центры коллективного пользования) суперкомпьютерами с современными вариантами топологии интерконнекта.

Во-вторых, одновременно необходимо существенно ускорить развитие отечественного аппаратного обеспечения (в первую очередь — интерконнекта) и довести его уровень в кратчайшие сроки до установки на кластерах и тестирования практических задач.

В-третьих, необходима программа создания/адаптации программного обеспечения с учетом требований массового параллелизма аналогичная программам INCITE и PRACE.

Для выработки плана действий по указанным направлениям необходимо коллегиальное обсуждение с привлечением всех заинтересованных организаций и ведомств.

При отсутствии изменений подобного рода российские специалисты, связанные с использованием высокопроизводительных вы-

числений в разных областях науки и техники, через несколько лет, по-видимому, должны будут менять профессию.

***Благодарности.** 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. № 06. 1995. URL: <http://www.osp.ru/os/1995/06/178750/>
- [2] 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>
- [3] Wilde T. **3D Torus for InfiniBand** // HPC Advisory Council Switzerland Conference. 2012. URL: <http://www.hpcadvisorycouncil.com/events/2012/Switzerland-Workshop/>
- [4] **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
- [5] **Blue Waters Opts Out of TOP500** // HPCwire. 2012. URL: http://www.hpcwire.com/hpcwire/2012-11-16/blue_waters_opts_out_of_top500.html
- [6] Mantevo Project Home Page. URL: <http://www.mantevo.org>
- [7] The Graph 500 List. URL: <http://www.graph500.org>
- [8] 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.
- [9] 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.

- [10] 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.
- [11] 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.
- [12] INCITE program. URL: <http://www.doeleadershipcomputing.org/>
- [13] Суперкомпьютерные технологии в науке, образовании и промышленности / Под ред.: академика В. А. Садовниченко, академика Г. И. Савина, чл.-корр. РАН Вл. В. Воеводина. - М.: Издательство МГУ, 2012. С. 42-49.
- [14] Янилкин А. В., Жилияев 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.
- [15] Жилияев 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.
- [16] Hutter J., Curioni A. Car-Parrinello molecular dynamics on massively parallel computers // Chemical Physics 2005. V. 6. P. 1788–1793.
- [17] Grinberg L., Morozov V., Fedosov D., Insley J. A., Papka M. E., Kumaran K., Karniadakis G. E. A new computational paradigm in multiscale simulations: Application to brain blood flow // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2011.

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

S. 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.