ISBN 978-5-901795-16-3 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ. Переславль-Залесский, 2009

УДК 004.382.2

С. М. Абрамов, В. Ф. Заднепровский, А. А. Московский, А. Б. Шмелев

Суперкомпьютеры Ряда 4 семейства «СКИФ»

Аннотация. В статье описывается проект создания отечественных суперЭВМ Ряда 4 семейства «СКИФ». В рамках создания данных суперЭВМ разрабатываются самые передовые суперкомпьютерные технологии: жидкостное охлаждение печатных плат, сверхплотная упаковка вычислительной мощности, новые решения в части системной, вспомогательной и сервисных сетей. Впервые в суперЭВМ Ряда 4 семейства «СКИФ» отечественная интеллектуальная собственность будет охватывать все конструкции, все печатные платы — то есть, все, кроме микросхем. Планируемая достижимая максимальная пиковая производительность данных суперЭВМ: 0.5 Pflops к осени 2009 года, 1 Pflops к осени 2010 года, более 5 Pflops к весне 2012 года.

1. Суперкомпьютерные программы Союзного государства

1.1. Суперкомпьютерная программа «СКИФ»

Суперкомпьютерная программа «СКИФ» [1] Союзного государства выполнялась в 2000–2004 годах. Полное название суперкомпьютерной программы — «Разработка и освоение в серийном производстве семейства моделей высокопроизводительных вычислительных систем с параллельной архитектурой (суперкомпьютеров) и создание прикладных программно-аппаратных комплексов на их основе», — точно определяло содержание работ.

В программе «СКИФ» головными исполнителями являлись: со стороны Беларуси — Объединенный институт проблем информатики Национальной академии наук Беларуси, со стороны России — Институт программных систем Российской академии наук. Заказчики-координаторы программы «СКИФ»: со стороны Беларуси — Национальная академия наук Беларуси, со стороны России — Федеральное агентство на науке и инновациям.

Суперкомпьютерная программа «СКИФ» — это серьезная, комплексная программа, в рамках которой по единой концепции создавался широкий спектр моделей семейства «СКИФ» и обеспечивалась возможность подбора конфигураций, оптимальных для различных

применений. Девятнадцать мероприятий программы покрывали все слои суперкомпьютерной отрасли:

- разработка и реализация аппаратных средств;
- разработка и реализация системного программного обеспечения;
- разработка и реализация инструментальных средств и законченных (пилотных) прикладных систем;
- вспомогательные мероприятия: подготовка и переподготовка кадров, создание и эксплуатация единого информационного пространства программы.

Институт программных систем РАН работы по созданию аппаратных и программных средств для семейства суперЭВМ «СКИФ» вел в тесном сотрудничестве с исполнителями от Республики Беларусь и с основными исполнителями Программы со стороны России, среди которых были:

- ОАО «Научно-исследовательский центр электронно-вычислительной техники» (НИЦЭВТ, Москва);
- Центр научных телекоммуникаций и информационных технологий (ЦНТК РАН, Москва);
- НИИ механики МГУ имени М. В. Ломоносова (Москва);
- Институт высокопроизводительных вычислений и информационных систем (ИВВиИС, СПб.);
- Российский НИИ региональных проблем (Переславль-Залесский, РосНИИ РП);
- Компания «Суперкомпьютерные системы» (СКС, Москва);
- НИИ Космических систем (Королев).

Программа «СКИФ» была признана одной из самых успешных программ Союзного государства. Результаты, достигнутые в ходе выполнения программы, получили и высокую правительственную оценку. За работу «Разработка конструкторской и программной документации, подготовка промышленного производства и выпуск образцов высокопроизводительных вычислительных систем (суперкомпьютеров) семейства «СКИФ» Ряда I и Ряда II» была присуждена премия Правительства Российской Федерации в области науки и техники за 2006 год группе исполнителей Программы «СКИФ».

1.2. Суперкомпьютерная программа «СКИФ-ГРИД»

Суперкомпьютерная программа «СКИФ-ГРИД» Союзного государства рассчитана на выполнение в 2007–2010 годах. Полное наименование программы: «Разработка и использование программно-аппаратных средств ГРИД-технологий и перспективных высокопроизводительных (суперкомпьютерных) вычислительных систем семейства «СКИФ»».

Как и в программе «СКИФ», головными исполнителями программы «СКИФ-ГРИД» являются: со стороны Беларуси — Объединенный институт проблем информатики Национальной академии наук Беларуси, со стороны России — Институт программных систем Российской академии наук.

Заказчики-координаторы программы «СКИФ-ГРИД»: со стороны Беларуси — Национальная академия наук Беларуси, со стороны России — Федеральное агентство на науке и инновациям. Программа «СКИФ-ГРИД» включает четыре направления работ:

- GRID-технологии: развитие, исследование, внедрение высокопроизводительных вычислений на основе GRID-технологий; поддержка гетерогенных, территориально-распределенных вычислительных комплексов.
- Суперкомпьютеры семейства «СКИФ» (Ряд 3 и 4): создание суперкомпьютеров «СКИФ» нового поколения на базе новых перспективных процессоров и вычислительных узлов, новых технических средств системной сети, управления системой, спецвычислителей и гибридных узлов, разработка соответствующего программного обеспечения.
- Защита информации: реализация (аппаратных и программных) средств защиты информации в создаваемых вычислительных комплексах.
- Пилотные системы: реализация прикладных систем в перспективных областях применения создаваемых высокопроизводительных установок, решение актуальных задач на суперкомпьютерах и GRID-системах, усилия по подготовке и переподготовке кадров в области GRID- и суперкомпьютерных технологий.

Программа «СКИФ-ГРИД» примерно в два–три раза крупнее программы «СКИФ» по масштабам: по количеству привлеченных

предприятий, по объемам запланированных работ и объемам ресурсов, привлекаемым для выполнения данных работ. Так, в исполнение программы «СКИФ» было вовлечено примерно по десять предприятий со стороны Беларуси и России. В программе «СКИФ-ГРИД» только со стороны Российской Федерации сегодня участвуют уже более 20 организаций. В том числе, российских исполнителей программы — более 20, среди них: ГЦ РАН, ИКИ РАН, ИПМ им. М. В. Келдыша РАН, ИППИ РАН, ИПХФ РАН, ИХФ РАН, НИВЦ МГУ, НИИ КС, НИИФХБ МГУ, НИИЯФ МГУ, ННГУ, НПЦ «Элвис», ОИЯИ, ОАО «Т-Платформы», ООО «ЮникАйСиз», ПензГУ, СПбАЭП, СПбГПУ, ТГУ, Химический факультет МГУ, ЧелГУ, ЮУрГУ, а от белорусской стороны в программе участвуют институты Национальной академии наук Беларуси: Институт математики, ОИЭЯИ, ИБОХ; ведущие государственные университеты: БГУ, БНТУ, БГУИР, ГрГУ; отраслевые институты, конструкторские бюро и ведущие предприятия Республики Беларусь, среди которых НИИ ЭВМ, «Белмикросистемы», «КБ системного программирования», «Минский моторный завод» и др.

1.3. Создание программного обеспечения суперЭВМ семейства «СКИФ»

Как правило, когда говорят о результатах программ «СКИФ» и «СКИФ-ГРИД», внимание уделяется аппаратным средствам, мощностям разработанных суперЭВМ, а также фактам вхождения в список пятисот самых мощных суперЭВМ мира (Top500). Это, действительно, очень важно, но хотелось бы отметить, что большая часть усилий, большее время, большие трудозатраты и значительная часть финансов в обеих программах были потрачены на создание программного обеспечения (ПО). Так, чтобы оценить масштабность комплекта программного обеспечения суперкомпьютеров семейства «СКИФ», перечислим, что в него уже на момент завершения программы «СКИФ» (2004 год) входили:

- системное ПО: операционная система; базовые библиотеки поддержки параллельного счета; файловые системы; системы очередей, мониторинга и управления; стандартные системы программирования — C, C++, Fortran; и т. п.;
- средства разработки параллельных программ — программные системы, инструментальные средства и библиотеки: MIRACLE, Open TS, Grace, и др.;

- два десятка параллельных прикладных систем для различных областей.

2. Роль суперкомпьютерных технологий в государствах с экономикой, основанной на знаниях

Прежде, чем обсуждать суперкомпьютерные технологии и супер-ЭВМ, разрабатываемые в программах «СКИФ» и «СКИФ-ГРИД», обсудим роль суперкомпьютерных технологий. Сегодня критические (прорывные) технологии в государствах, строящих экономику, основанную на знаниях, исследуются и разрабатываются на базе широкого использования высокопроизводительных вычислений [2, 3]. И другого пути — нет. Без серьезной суперкомпьютерной инфраструктуры:

- невозможно создать современные изделия высокой (аэрокосмическая техника, суда, энергетические блоки электростанций различных типов) и даже средней сложности (автомобили, конкурентоспособная бытовая техника и т. п.);
- невозможно быстрее конкурентов разрабатывать новые лекарства и материалы с заданными свойствами;
- невозможно развивать перспективные технологии (биотехнологии, нанотехнологии, решения для энергетики будущего и т. п.).

Сегодня суперкомпьютерные технологии по праву считаются, пожалуй, важнейшим фактором обеспечения конкурентоспособности экономики любой страны, а единственным способом победить конкурентов объявляют возможность обогнать их в расчетах. Здесь характерны слова Президента Совета по конкурентоспособности США: *«Технологии, таланты и деньги доступны многим странам. Поэтому США стоит перед лицом непредсказуемых экономических конкурентов из-за рубежа. Страна, которая желает победить в конкуренции, должна победить в вычислениях»*.

Неважно, о конкуренции в каком секторе экономики идет речь: сказанное верно для добывающих и перерабатывающих секторов экономики, и особенно это верно при разработке новых технологий. И поэтому в развитых странах мира для перехода к экономике знаний создается новая инфраструктура государства — государственная система из мощных суперкомпьютерных центров, которые объединены сверхбыстрыми каналами связи в грид-систему. То есть, по сути, речь

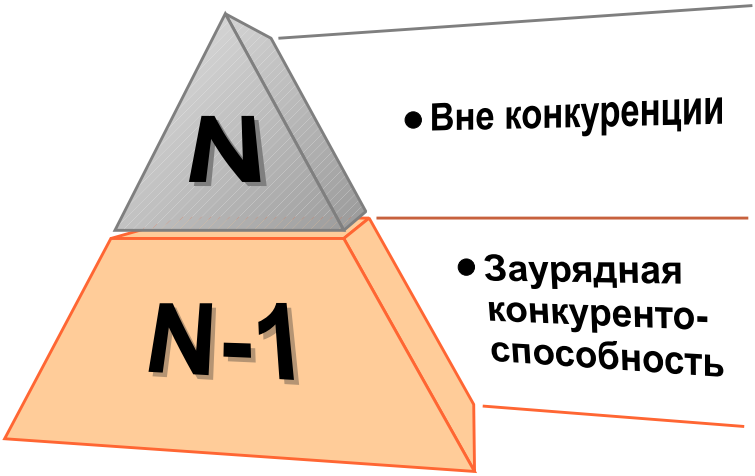


Рис. 1. Соответствие между уровнем (N или N-1) суперкомпьютерных технологий суперЭВМ и уровнем конкурентоспособности разработок, создаваемых при помощи данной суперЭВМ

идет о национальной научно-исследовательской информационно-вычислительной сети. Для такой системы часто используют термин *киберинфраструктура*. В этих странах на создание национальной киберинфраструктуры выделяются большие финансы из государственных бюджетов: в 2005–2008 гг. США тратили на эти цели 2–4 млрд. долларов в год.

Тем самым, краткое определение сегодняшней роли суперкомпьютерных технологий может быть таким: это ключевая критическая технология, единственный инструмент, дающий возможность победить в конкурентной борьбе.

В каждый момент времени, если посмотреть уровень развития суперкомпьютерной отрасли (например, список Top500), то можно выделить два слоя (см. Рис. 1):

- *Технологии уровня «N»*. Инновационные, совершенно новые суперкомпьютеры, которые сильно вырываются вперед. Они сделаны по технологиям будущего, которые еще не вполне освоены, а только-только разрабатываются в мире. Такие

машины соответствуют первым 10–20 местам списка Top500. Эти суперЭВМ обладают мощностью, которая радикально отличает их от всех других машин. И на платформе таких суперЭВМ можно выполнить вычисления — разработать новые материалы, новые технологические решения, — которые позволят обладающей ими стране быть вне конкуренции и существенно оторваться от других производителей материалов, лекарств, механизмов и тому подобного.

- *Технологии уровня «N-1»*. Технологии более низкого уровня, отработанные решения, воспроизводить их способны многие страны. Соответственно, расчеты, выполняемые на таких машинах, позволяют достичь нормальной (обычной, заурядной) конкурентоспособности. То есть, позволяют создавать материалы, механизмы, решения такие же, как и у многих других стран. В данном случае мы получаем рядовую конкурентоспособность: с разработками можно выходить на мировой рынок, на котором нам придется вести изнурительную конкурентную борьбу с десятком подобных разработок.

Надо честно отметить, что разработанные в предыдущие годы суперЭВМ Рядов 1–3 относились к технологическому уровню N-1. А вот суперЭВМ Ряда 4 планируется разрабатывать на технологическом уровне N.

2.1. СуперЭВМ семейства «СКИФ» Рядов 1, 2 и 3

Суперкомпьютеры семейства «СКИФ» [2,4] выпускались отдельными группами, называемыми «рядами» (Таблица 1, Рис. 2). На сегодняшний день:

- разработаны и выпущены опытные и серийные образцы суперЭВМ Рядов 1, 2 и 3;
- проработаны технические решения, подготовлена эскизная конструкторская документация суперЭВМ Ряда 4 семейства «СКИФ».

2.2. Ряд 1 суперЭВМ семейства «СКИФ»

Конструкторская документация для суперЭВМ Ряда 1 (семейства «СКИФ»), а также их опытные образцы разрабатывались и выпускались в 2000–2003 годах. Решения, которые были при этом выработаны, были способны обеспечивать мощность суперЭВМ 20–500 GFlops[1].

Для этих суперЭВМ были характерны следующие технические решения:

- использовались 32-х-разрядные одноядерные CPU;
- для системной сети использовался SCI (2D-тор) и Myrinet;
- как вспомогательная сеть использовался FastEthernet;
- форм-фактор для вычислительных узлов в данных суперЭВМ — монтируемые в стойки корпуса — от 4U до 1U.

В эти же годы была разработана и освоена в производство отечественная системная сеть — плата SCI (2D-тор). Работа была выполнена исполнителем программы «СКИФ» ОАО НИЦЭВТ.

2.3. Ряд 2 суперЭВМ семейства «СКИФ»

Конструкторская документация и опытные образцы Ряда 2 суперЭВМ семейства «СКИФ» разрабатывались и выпускались в 2003–2007 годах. Полученные здесь решения позволяли выпускать суперЭВМ мощностью 0,1–5 Tflops.

Для этих суперЭВМ были характерны следующие технические решения:

- использовались одноядерные CPU как 32-х-разрядные, так и 64-х-разрядные (для старших моделей Ряда 2);
- в качестве системной сети использовались сети SCI (3D-тор) и Infiniband;
- в качестве вспомогательной сети использовался GbEthernet;
- существенно повысилась плотность упаковки вычислительной мощности — использовались серверы с форм-фактором 1U и даже так называемые решения Hyper-Blade.

[1]Единицы производительности суперЭВМ: 1 Gflops — миллиард операций с плавающей точкой в секунду, 1 TFlops = 1 000 GFlops — триллион операций с плавающей точкой в секунду, 1 Pflops = 1 000 TFlops — тысяча триллионов операций с плавающей точкой в секунду.

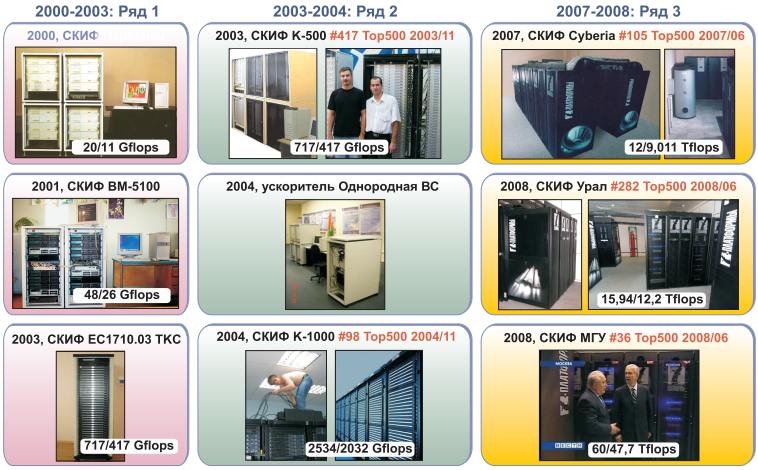


Рис. 2. Ряды 1, 2, и 3 семейства суперЭВМ «СКИФ»

Отметим, что в эти же годы были разработаны системы управления и мониторинга суперкомпьютеров ServNet v.1 и ServNet v.2 (разработка ИПС РАН). Также начались работы по изучению и применению ускорителей, как построенных на FPGA, так и ускорителей, выполненных полностью на отечественной элементной базе (так называемые однородные вычислительные системы, ОВС).

2.4. Ряд 3 суперЭВМ семейства «СКИФ»

Конструкторская документация и опытные образцы Ряда 3 [4] суперЭВМ семейства «СКИФ» разрабатывались в 2007–2008 гг. Полученные здесь технические решения позволяют строить суперкомпьютеры с производительностью 5–150 Tflops. Для этих суперЭВМ были характерны следующие технические решения:

- использовались 2–4-х-ядерные 64-х-разрядные CPU;
- в качестве системной сети использовалась сеть Infiniband DDR;
- в качестве вспомогательной сети использовался GbEthernet;

- в младших моделях использовались монтируемые в 19"монтажный шкаф серверы с форм-фактором 1U, в старших моделях использовались отечественные blade-решения, позволяющие в 5U упаковывать 10 вычислительных узлов.

Для данных суперЭВМ использовалась новая версия управляющей сети — ServNet v.3 (разработка ИПС РАН). Повысилась плотность упаковки процессоров до уровня 4 CPU на 1 U. Соответственно повысилась и плотность выделения тепловой энергии на единицу объема. И если до этого в суперЭВМ семейства «СКИФ» использовалось воздушное охлаждение, то в машинах Ряда 3 уже использовалось трехконтурное охлаждение «воздух–вода–фреон».

3. Что есть отечественного в суперЭВМ семейства «СКИФ»?

Когда обсуждаются суперЭВМ семейства «СКИФ», то всегда задается вопрос: «А что отечественного есть в этих суперЭВМ, ведь в них же используются импортные комплектующие?» Это правда, пока еще в странах-участниках Союзного договора не развито производство необходимых для суперЭВМ отечественных микропроцессоров и сопутствующих комплектующих. В результате приходится использовать импортную элементную базу. Впрочем, такая ситуация не является исключением. Суперкомпьютеры (впрочем, как и компьютеры) — технически сложные устройства. Как правило, такого рода изделия создаются с широким использованием мирового распределения труда. Общая практика, когда в суперЭВМ, разрабатываемой одной компанией некоторой страны, широко используются компоненты, разработанные и производящиеся в самых различных компаниях в разных странах мира. В настоящее время ни одна страна мира (за исключением разве что США), не производит все без исключения компоненты компьютерной техники и суперкомпьютеров в частности.

В полном соответствии с данной тенденцией суперкомпьютеры семейства «СКИФ» основываются на использовании передовой зарубежной элементной базы, что позволяет обеспечить конкурентоспособность по такому важнейшему параметру как производительность. Суперкомпьютеры семейства «СКИФ» разрабатываются, собираются, налаживаются и тестируются нашими специалистами. При этом

Таблица 1. Суперкомпьютеры семейства «СКИФ» Ряд 1, 2, 3 и 4

Ряд	Годы, пиковая производительность (расчетный диапазон)	Ядер в CPU/ разрядность	Сетевые решения вспомогательной/ системной сети	Форм-фактор **(CPUs/U)**	Примечание
1	2000–2003, 0.020–0.5 TFlops	1/32	FastEthernet/ SCI (2D-тор), Myrinet	4U–1U **(0.5–2)**	Отечественный SCI (2D-тор). Охлаждение: воздух
2	2003–2007, 0.1–5 Tflops	1/32–64	GbEthernet/ SCI (3D-тор), Infiniband	1U, HyperBlade **(2)**	ServNet v.1, v.2. Ускорители: FPGA, ОВС. Охлаждение: воздух
3	2007–2008, 5–150 Tflops	2–4/64	GbEthernet/ Infiniband DDR	1U, blades 20 CPU в 5U **(2–4)**	ServNet v.3. Охлаждение: воздух-вода-фреон
4	2009–2011, 500–5 000 Tflops	4–8/64	Infiniband QDR/ отечественная системная сеть (3D-тор)	blades 64 CPU в 6U **(10.667)**	Новые подходы к охлаждению. Ускорители: FPGA, GPU, МЦОС...

Союзное государство является собственником конструкторской документации на узлы суперЭВМ семейства «СКИФ» и на изделия целиком. На часть разработок имеются патенты. Это еще одно документальное подтверждение оригинальности отечественных разработок.

Независимая экспертиза страны происхождения суперЭВМ выполняется и при включении суперЭВМ в рейтинг Top500. Поданные заявителем сведения о стране происхождения и о производителе проверяются составителями списка и, если нужно, исправляются — такие случаи известны. Во всех случаях вхождения всех суперЭВМ

семейства «СКИФ» данная проверка страны происхождения проходила успешно — составители списка оставляли без изменения сведения об отечественном происхождении суперЭВМ семейства «СКИФ»: «СКИФ К-500», «СКИФ К-1000», «СКИФ Cyberia», «СКИФ МГУ» и «СКИФ Урал»[2].

В целом, за всю историю Top500 отечественное происхождение [2] признавалось только у этих пяти суперЭВМ семейства «СКИФ» и еще у «МВС-1000М» (НИИ «Квант», редакции рейтинга 06/2002–06/2004). Все остальные установленные в России системы, попавшие в Top500, являются импортными — производства: IBM, Sun Microsystems и Hewlett-Packard. Еще одно объективное доказательство отечественного происхождения суперЭВМ семейства «СКИФ» — превышение зарубежных аналогов по показателям. Если некоторая суперЭВМ обладает характеристиками, которые превышают достижения отрасли, то это является неоспоримым доказательством уникальности, оригинальности установки. СуперЭВМ семейства «СКИФ» часто показывали лучшие в отрасли результаты. Например:

- «СКИФ К-500», «СКИФ Cyberia», «СКИФ МГУ», «СКИФ Урал» продемонстрировали лучший показатель КПД для суперЭВМ на процессорах Intel. Да, в суперЭВМ семейства «СКИФ» используются импортные процессоры, но отечественным разработчикам удается их использовать лучше, чем кому бы то ни было!
- В ноябре 2004 «СКИФ К-1000» занял первое место в мире на тесте «столкновение 3 автомобилей» в рейтинге TopCrunch (`www.topcrunch.org`, поддержан DARPA).
- В феврале 2007 «СКИФ Cyberia» выдает показатели лучшие, чем у современных суперЭВМ (Cray, HP, IBM, SUN): лучший (на 8..13%) КПД, лучшую (в 2–1.5 раза) масштабируемость на прикладном инженерном пакете STAR-CD.

Часто разработанные нами решения превышают зарубежные аналоги и по техническим возможностям:

- blade-решение по суперЭВМ Ряда 3 «СКИФ МГУ», «СКИФ Урал» имело (на момент выпуска): плотность упаковки вычислительной мощности процессоров Intel — на 20% лучше всех аналогичных изделий в мире; стандартный разъем PCI

[2]Редакции рейтинга 11/2003, 11/2004–06/2006, 06/2007, 11/2007, 06/2008, 11/2008.

Express; «N+1» резервирование и «горячую замену» как блоков питания, так и вентиляторов. Такое сочетание этих важных эксплуатационных свойств встречалось только в данной blade-системе;
- система управления СКИФ ServNet версии 1, 2, 3 (разработана в ИПС РАН) поддерживает ряд уникальных возможностей. Например, функцию «черного ящика» — сохранение последних записей о событиях в отказавшем блоке.

СуперЭВМ семейства «СКИФ» являются отечественными системами, разработанными на базе импортных комплектующих, с постепенно нарастающей долей импортозамещения. В суперкомпьютерах «СКИФ» Ряда 1 отечественными были:

- схемотехнические решения;
- конструкторская документация (КД) корпусов и стоек, при этом стойки и корпуса выпускались в Минске;
- программное обеспечение (ПО) кластерного уровня семейства «СКИФ» — ПО КУ СКИФ.

При этом набор отечественного базового программного обеспечения (ПО КУ СКИФ) создавался и на основе оригинальных разработок, и на основе доработок и адаптации программного обеспечения с открытыми исходными текстами. СуперЭВМ «СКИФ» Ряда 2 также разработаны по оригинальному проекту. И здесь отечественными являлись:

- схемотехнические решения;
- конструкторская документация (КД) корпусов и стоек;
- разработка и программное обеспечение — ПО КУ СКИФ.

Кроме того:

- отдельные компоненты узлов были доработаны по документации наших разработчиков — например, материнские платы для «СКИФ К-500» и «СКИФ К-1000»;
- суперЭВМ «СКИФ» Ряда 2 оснащались сетью управления и мониторинга отечественной разработки — ServNet версии 1 и 2, разработка ИПС РАН;
- в суперкомпьютере «СКИФ ЕС1710.03» использовался интерконнект отечественного производства (НИЦЭВТ, интерконнект SCI 2D-тор).

Суперкомпьютеры «СКИФ» Ряда 3 «СКИФ МГУ» и «СКИФ Урал» созданы на основе blade-серверов отечественной разработки, имеющих уникальные показатели. Таким образом, здесь отечественными были:

- схемотехнические решения;
- конструкторская документация (КД) на сами blade-серверы и шасси;
- программное обеспечение — ПО КУ СКИФ;
- конструкторская и программная документация на сервисную сеть ServNet версии 3 (платы ServNet T-60 и ServNet CMB).

Тем самым, суперЭВМ Рядов 1–3 по праву называют отечественными. Правда, надо заметить, что в них использовались целые блоки, на которые отсутствовала и отечественная конструкторская документация, и интеллектуальная собственность (включая право на производство и право на модификацию). И к таким блокам относились не только элементная база (не только микросхемы), но и, например, практически все печатные платы. За исключением ServNet (разработанного ИПС РАН) все печатные платы (материнские, соединительные и т. п.) суперЭВМ Рядов 1–3 были импортными. В рамках реализации суперЭВМ Ряда 4 семейства «СКИФ» планируется серьезно изменить данное положение вещей — подробнее ниже, в разделе 4.8.

4. СуперЭВМ семейства «СКИФ» Ряда 4

Конструкторская документация и опытные образцы Ряда 4 суперЭВМ семейства «СКИФ» запланированы к разработке в 2008–2012 гг. Данные суперЭВМ будут иметь пиковую производительность 200–5 000 Tflops (0.2–5 Pflops) и выше (см. Рис. 3).

В суперкомпьютерах Ряда 4 семейства «СКИФ» предусмотрены самые современные решения[3]:

- в вычислительных узлах использованы многоядерные (4–8 ядер и выше) 64-х-битные процессоры стандартной архитектуры x86. В дополнение к ним в узле предусмотрена ПЛИС, ресурсы которой могут быть использованы для ускорения специализированных алгоритмов;

[3]По сути, речь идет о разработке технологий уровня N.

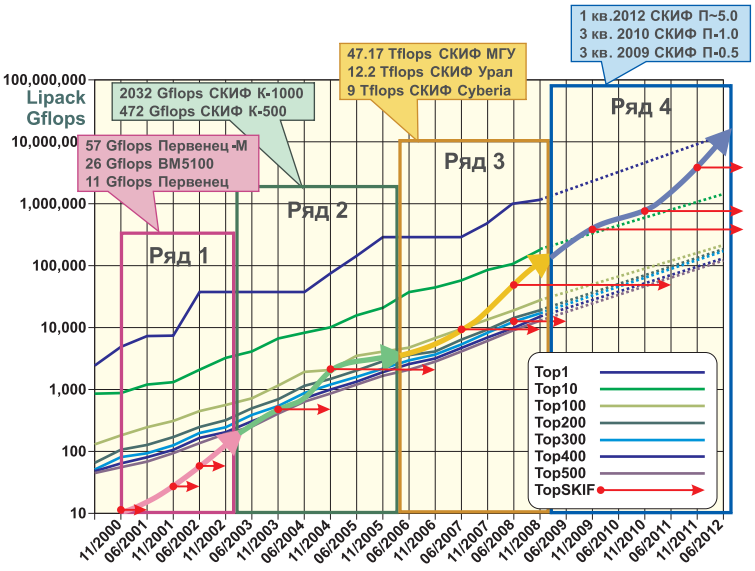


Рис. 3. Семейство суперЭВМ «СКИФ»: Ряд 1, 2, 3 и 4

- достигнута более высокая плотность упаковки вычислительной мощности. Разработаны оригинальные blade-системы, позволяющие упаковать более 10 процессоров в 1U стоечного пространства;
- такая высокая плотность упаковки требует новых подходов к охлаждению вычислительной установки. В СуперЭВМ семейства «СКИФ» Ряда 4 применяется система охлаждения вычислительных узлов с использованием воды либо этиленгликоля;
- в качестве системной сети в суперЭВМ используется отечественная системная сеть (3D-тор на базе FPGA), а в качестве вспомогательной сети — Infiniband QDR или 10GbEthernet.

В дальнейших разделах подробно обсуждаются различные характеристики суперЭВМ Ряда 4 семейства «СКИФ».

4.1. Производительность, компактность, надежность

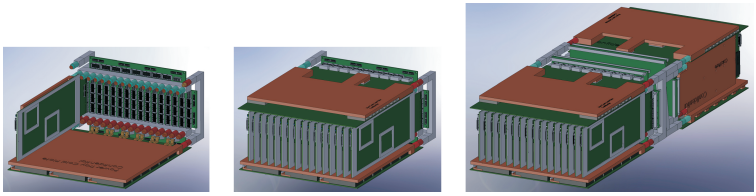
СуперЭВМ высокой производительности по необходимости содержит большое количество узлов. При росте числа вычислительных узлов критическими становятся такие параметры как надежность и размер установки (с ростом физических размеров растет задержка при передаче данных в системной сети, что снижает характеристики суперЭВМ). К счастью, и повысить надежность, и уменьшить размер установки удается одним и тем же приемом: повышение плотности упаковки вычислительных узлов. По мере того как все большее количество вычислительных узлов упаковывается в рамки монтажного шасси, мы достигаем следующих эффектов:

- уменьшаются физические размеры установки, уменьшаются длины соединительных линий между вычислительными узлами, уменьшаются задержки;
- большое количество соединений выполняется в рамках монтажного шасси. Такие соединения выполнены либо в виде контактных дорожек на печатных платах, либо в виде соединений через разъемы соединительной печатной платы (backplane). Таким образом, происходит существенное снижение количества соединительных кабелей и кабельных разъемов в системе, за счет чего серьезно улучшается надежность.

В суперкомпьютерах Ряда 4 семейства «СКИФ» в шасси с размером 6U входят (см. Рис. 4) две соединительные панели (backplane), к которым подключены две группы печатных плат, каждая из которых включает:

- плату поддержки электропитания (Рис. 5);
- корневую плату, содержащую средства управления и мониторинга аппаратурой шасси и коммутатор Infiniband QDR (Рис. 5);
- 16 плат-лезвий вычислительных узлов (Рис. 7, раздел 4.5).

Существенная часть соединений вычислительных узлов в системной сети и во вспомогательной сети (Infiniband QDR) выполнены в рамках шасси за счет соединительной панели (не при помощи кабельных соединений). В моделях СКИФ 4/Н и СКИФ 4/В суперкомпьютеров семейства «СКИФ» Ряда 4 шасси содержит 32 двухпроцессорных вычислительных узла.



(1) *Начало сборки:* снизу плата блоков питания с охлаждающей пластиной, слева — вычислительный узел, сзади — соединительная плата (backplane);
(2) *Собрана половина шасси:* снизу плата блоков питания с охлаждающей пластиной, в середине — 16 вычислительных узлов, сверху — корневая плата с охлаждающей пластиной, сзади — соединительная плата (backplane);
(3) *Две половины объединены в одно шасси 6U:* для модели СКИФ 4/Н — 32 вычислительных узла, 64 процессора Intel Nehalem, 256 ядер.

Рис. 4. Схемы компоновки шасси суперЭВМ Ряда 4 семейства «СКИФ»

При работе передняя и задняя стороны шасси закрываются сенсорными ЖКД-мониторами, при помощи которых поддерживаются отображение состояния и управление аппаратурой шасси. Шасси не содержит подвижных частей (вентиляторов, механических дисков), является законченной (стационарной или возимой) суперЭВМ (пиковая производительность 3 Tflops для модели СКИФ 4/Н) и блоком для более крупных систем.

4.2. Охлаждение: передовые решения

Такая высокая плотность упаковки требует новых подходов к охлаждению вычислительной установки. В суперЭВМ Ряда 4 семейства «СКИФ» применена система непосредственного водяного охлаждения вычислительных узлов.

Решения подобного класса сегодня, несомненно, относятся к технологиям уровня N. Лидеры в области суперкомпьютерных технологий переходят от уже освоенных схем охлаждения «вода на уровне

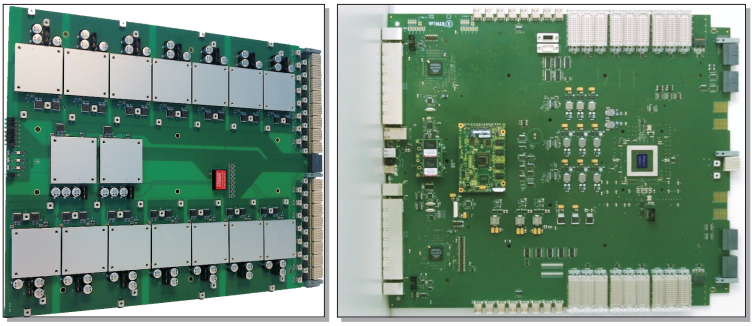


Рис. 5. Плата электропитания (слева) и корневая плата (справа) шасси для суперЭВМ Ряда 4 моделей СКИФ 4/Н и СКИФ 4/В

шкафа», «горячий коридор», «воздух–вода–фреон» к новым подходам к охлаждению вычислительной установки. Примером могут послужить разработки SGI (система охлаждения Kelvin), водяное охлаждение процессоров у фирм IBM и Fujitsu и разработки компаний Cray и IBM по использованию фазового перехода (испарения) как способа охлаждения микросхем.

Заметим, что, по сравнению со схемами охлаждения, где в качестве теплоносителя используется воздух, у водяного охлаждения имеется ряд серьезных преимуществ:

- данная схема охлаждения требует как минимум в 2 раза меньше энергозатрат;
- при остановке циркуляции теплоносителя за счет большей теплоемкости вода в течение некоторого времени сохраняет способность охлаждать микросхемы;
- система охлаждения в вычислителе не содержит ни одной механической подвижной части. Это повышает надежность установки и ее эргономические качества (бесшумность).

4.3. Модели Ряда 4 и повторное использование разработок

СуперЭВМ Ряда 4 запланированы к разработке и производству в течение 2008–2012 гг. За это время произойдет выпуск как минимум

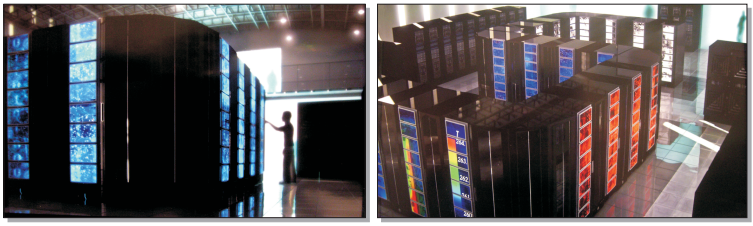


Рис. 6. Компьютерная модель суперЭВМ СКИФ П-0.5 Ряда 4 семейства «СКИФ» с производительностью 500 Tflops

трех различных семейств микропроцессоров. Основываясь на прогнозах и планах ведущих компаний, мы предусматриваем выпуск четырех последовательностей моделей в рамках Ряда 4: «СКИФ 4/Н», «СКИФ 4/В», «СКИФ 4/С», «СКИФ 4/П». Ожидается, что к концу жизненного цикла Ряда 4 плотность упаковки вырастет более чем в 8 раз, а энерго-эффективность (производительность на Ватт) более чем в 5 раз по сравнению с сегодняшним днем.

При этом предусмотрено широкое повторное использование конструкторской документации различных блоков и модулей. Так, для всех моделей одинаковыми будут являться все конструкции и соединительная инфраструктура шкафа и шасси, а также большинство печатных плат: соединительные, корневые и подсистемы электропитания. Изменяться будут (и то лишь частично) только печатные платы вычислительных узлов.

4.4. Не просто рекордные установки, а широкий ряд изделий

Каждая последовательность моделей охватывает широкий спектр производительности от нескольких Tflops до нескольких тысяч Tflops и предусматривает доступность для потребителя трех видов изделий:

- *Персональная или мобильная суперЭВМ* представляет собой одно шасси, которое можно расположить на рабочем месте сотрудника, тем более что это изделие бесшумное и имеет вполне приемлемое (для рабочего места) электропотребление. Пиковая производительность такого вычислителя может быть до трех Tflops. Заметим, что вся коммутация

вычислительных узлов системной и вспомогательной сети уже выполнена в рамках шасси. Шасси является первым уровнем законченного изделия и строительным блоком для более крупных систем (шкаф, система из нескольких шкафов).

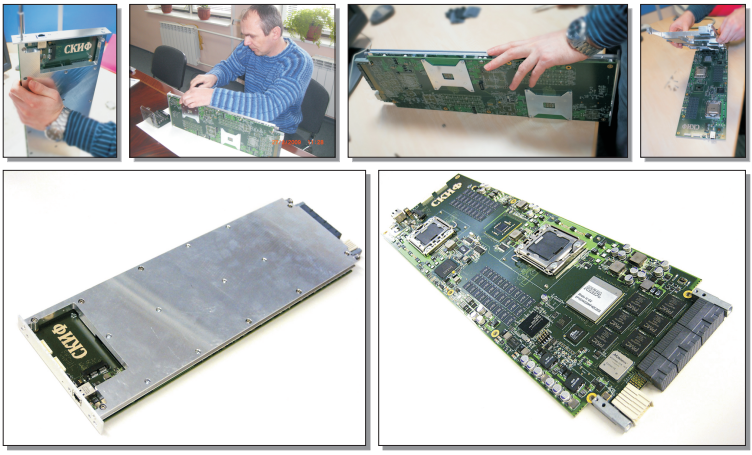
- *СуперЭВМ для лабораторий (конструкторских отделов и т. п.)* и региональных ВУЗов представляет собой один или несколько шкафов, каждый из которых содержит от двух до восьми шасси и всю необходимую соединительную инфраструктуру для них — соединения системной сети, вспомогательной сети, сервисной сети, подсистем электропитания и охлаждения. Пиковая производительность такого вычислителя может быть от шести до 350 Tflops.
- *Суперкомпьютерная система* для крупных суперкомпьютерных национальных центров представляет собой несколько (15 и более) шкафов, объединенных общей инфраструктурой — соединения системной сети, вспомогательной сети, сервисной сети, подсистем электропитания и охлаждения. Пиковая производительность такого вычислителя может быть от 350 Tflops до 10 Pflops (см. Рис. 6).

Таким образом, суперкомпьютеры Ряда 4 семейства «СКИФ» охватывают большое разнообразие областей применения и широкий диапазон производительности.

4.5. Вычислительный узел моделей

В состав вычислительного узла суперкомпьютер Ряда 4 семейства «СКИФ» входит:

- два современных стандартных (x86) многоядерных (четыре ядра и больше) 64-х-разрядных микропроцессора;
- память (RAM) объемом 6, 12 или 24 Гбайт;
- микросхема адаптера (NIC) Infiniband QDR;
- твердотельный жесткий диск (SSD) для хранения образа операционной системы, вспомогательных файлов, записей контрольных точек и раздела для организации виртуальной памяти;
- микросхема FPGA, которая используется, с одной стороны, для организации системной сети, а, с другой стороны, оставшиеся свободными ресурсы FPGA могут быть использованы для ускорения некоторых вычислений.



- *Верхний ряд:* первичный осмотр и разборка вычислительного узла, февраль 2009;
- *Нижний ряд слева:* вычислительный узел с охлаждающей пластиной;
- *Нижний ряд справа:* вычислительный узел без охлаждающей пластины.

Рис. 7. Вычислительный узел суперЭВМ Ряда 4 семейства «СКИФ» моделей СКИФ 4/Н и СКИФ 4/В

Все компоненты вычислительного модуля размещаются на одной печатной плате. К этой печатной плате прижимается (вплотную ко всем микросхемам) так называемая охлаждаемая пластина, через которую организован поток охлаждающей жидкости (см. Рис. 7).

4.6. Больше, чем просто системная сеть

Системная сеть в суперЭВМ Ряда 4 организована с использованием FPGA. В качестве топологии для системной сети используется трехмерный тор. За счет прошивки FPGA и его подключения к различным компонентам системы реализуется:

- быстрый обмен между FPGA и системной шиной вычислительного модуля, например, PCI Express;
- шесть двусторонних каналов, позволяющих объединять вычислительные узлы по топологии трехмерный тор;

- аппаратный маршрутизатор сообщений в системной сети топологии 3D тор;
- аппаратная поддержка коллективных операций библиотеки MPI, например, all_reduce.

Тем самым, разрабатывается масштабируемая в широких пределах системная сеть с явными чертами технологического уровня N.

Упомянем также, что в суперЭВМ Ряда 4 семейства «СКИФ» будут использованы еще две независимые сети, аналоги которых встречаются только в топовых моделях суперкомпьютеров (уровня N):

- отдельная сеть для реализации операций барьерной синхронизации;
- отдельная подсистема синхронизации системных часов всех микропроцессоров в вычислителе.

Среди прочего это позволяет на уровне операционной системы реализовать поддержку контрольных точек.

4.7. Мониторинг и управление

Для обеспечения высокой надежности в суперЭВМ Ряда 4 запланировано использовать расширенный состав сенсоров, располагаемых на различных печатных платах вычислителя, и три независимые сенсорные сети — сети мониторинга и управления. Опуская подробности, упомянем, что третья из этих сетей является новой версией сети ServNet. Она использует собственную подсистему электропитания и сетевую инфраструктуру для передачи данных.

4.8. Интеллектуальная собственность Союзного государства и перспективы массового производства

В реализации суперЭВМ Ряда 4 семейства «СКИФ» впервые интеллектуальная собственность на изделия принадлежит Союзному государству. В частности, в нашем распоряжении находится полный комплект конструкторской и производственной документации. Это дает право и возможность разместить изготовление всех блоков и узлов на любых предприятиях, в том числе и отечественных. А также возможность вносить изменения в конструкторскую документацию, создавать новые модификации суперЭВМ Ряда 4 семейства «СКИФ», в том числе на различной микропроцессорной базе (включая отечественную, если такая будет доступна).

Тем самым мы будем в максимальной готовности к восприятию отечественной элементной базы по мере ее появления.

Заключение

В рамках программы «СКИФ-ГРИД» обеспечивается разработка конструкторской документации и выпуск опытных образцов вычислительных узлов, шасси, шкафов суперЭВМ Ряда 4. В настоящий момент завершена разработка эскизной конструкторской документации суперЭВМ Ряда 4 семейства «СКИФ». Ведется выпуск опытных образцов вычислительных узлов (февраль 2009), шасси (март–апрель 2009) и шкафов (апрель–май 2009) этих суперкомпьютеров. В мае 2009 года организуется серийный выпуск и поставка потребителям изделий последовательности «СКИФ 4/Н». Суперкомпьютеры семейства «СКИФ» Ряда 4 при организации их массового производства становятся основой оснащения отечественной суперкомпьютерной техникой учреждений образования и науки, исследовательских и конструкторских бюро, предприятий промышленности и государственных структур Союзного государства.

Список литературы

[1] Абрамов С. М. *Итоги суперкомпьютерной программы «СКИФ» Союзного государства и перспективы ее развития* // «Пути ученого. Е. П. Велихов» ред. Академик РАН В. П. Смирнов. — М.: РНЦ «Курчатовский институт», 2007. — ISBN 978–5–9900996–1–6, с. 325–333. ↑1.1

[2] Абрамов С. М., Заднепровский В. Ф., Московский А. А. *Отечественные суперЭВМ и грид-системы. Проблемы развития национальной киберинфраструктуры в России* // «Российские суперкомпьютеры: Наука. Технологии. Производство»: Сборник статей. — Т. **2**. — М.: Библиотека ЦСПП, 2008. — ISBN 5–8027–0061–0, с. 36–54. ↑2, 2.1, 3

[3] Абрамов С. М., Заднепровский В. Ф., Московский А. А. *Опыт использования СуперЭВМ для эффективного развития «прорывных технологий» (на примере нанотехнологий)* // XII научно-практическая конференция Университета города Переславля «Программные системы: теория и приложения». — Т. **1**. — Переславль-Залесский: Изд-во «Университет города Переславля», 2008. — ISBN 978–5–901795–11–8, с. 37–50. ↑2

[4] Абрамов С. М., Анищенко В. В., Заднепровский В. Ф., Московский А. А., Криштофик А. М., Опанасенко В. Ю., Парамонов Н. Н. *Развитие семейства отечественных суперкомпьютеров «СКИФ» в рамках программы Союзного государства «СКИФ-ГРИД»* // Научный сервис в сети Интернет: решение больших задач. Труды Всероссийской научной конференции. 22–27 сентября 2008 г. г. Новороссийск. — М.: Изд-во МГУ имени М. В. Ломоносова, 2008. — ISBN (CD)978–5–211–05616–9, с. 286–291. ↑2.1, 2.4

Исследовательский центр мультипроцессорных систем ИПС имени А. К. Айламазяна РАН

ИПС имени А. К. Айламазяна РАН

S. M. Abramov, V. F. Zadneprovskiy, A. A. Moskovskiy, A. B. Shmelev. *Supercomputers SKIF series 4* // Proceedings of Program Systems institute scientific conference "Program systems: Theory and applications". — Pereslavl-Zalesskij, v. **1**, 2009. — p. 193–216. — ISBN 978-5-901795-16-3 (*in Russian*).

Abstract. The paper outlines "SKIF" series 4 supercomputers, which feature many technology advances: ultra-dense packaging of computational power, liquid cooling on node level, new solutions for system, supplementary and service networks. For the first time in "SKIF" projects, Russian organizations will have intellectual property rights for all components down to, but excluding semiconductor chips. Planned achievable maximum peak performance for "SKIF" series 4 is 500 TFlops by fall 2009, 1 PFLops by fall 2010 and more than 5 PFlops by spring 2012.