Суперкомпьютеры «СКИФ» ряда 4

Абрамов С.М. (*abram@botik.ru)*,
Заднепровский В.Ф. (*v.f.z_skii@mail.ru)*, Лилитко Е.П. (*gene@ks.pereslavl.ru)*
Институт программных систем имени А.К. Айламазяна РАН

Аннотация. **В статье рассматриваются суперкомпьютеры «СКИФ» ряда 4. Показаны преимущества технологических решений. Особое внимание уделено отечественному интерконнекту с топологией 3D-тор, поддержке гибридных вычислений, системе водяного охлаждения, высокой плотности упаковки вычислительной мощности.**

Ключевые слова: ***суперкомпьютеры семейства «СКИФ», интерконнект, гибридные вычисления, водяное охлаждение, энергоэффективность.***

Введение. История разработки

Суперкомпьютеры семейства «СКИФ» создавались в рамках суперкомпьютерных программ Союзного государства «СКИФ» [1] и «СКИФ-ГРИД» [2].

Это крупные, комплексные научно-технические программы Союзного государства, связанные с разработкой и исследованиями на всех уровнях суперкомпьютерных и грид-технологий: аппаратные средства, операционные системы, системы параллельного программирования и различные приложения, сервисы и т.п.

Программа «СКИФ» исполнялась в 2000–2004 годах, в ней участвовало примерно по десять организаций от России и Белоруссии, объем из бюджетов обеих стран составил около 250 млн. рублей.

По составу работ, финансированию и составу исполнителей Программа «СКИФ-ГРИД» в 2,5–3 раза крупнее программы «СКИФ». Работы в программе «СКИФ-ГРИД» велись по четырем направлениям: грид-технологии; суперкомпьютеры; информационная безопасность; приложения. В данной статье мы остановимся только на втором направлении.

Головным исполнителем от России в обеих программах являлся ИПС имени А.К. Айламазяна РАН.

Суперкомпьютерные программы «СКИФ» и «СКИФ-ГРИД» внесли серьезный вклад в развитие суперкомпьютерной отрасли и суперкомпьютерного рынка России:

- за время выполнения программ «СКИФ» и «СКИФ-ГРИД» семь систем[1] семейства «СКИФ» 18 раз вошли в престижный всемирный рейтинг суперЭВМ Top500 [3] (с максимально высокой позицией № 36 в июне 2008):
    1. **СКИФ-Аврора ЮУрГУ** (модель СКИФ 4/В), 100.4/117.0 Tflops[2] — 06'2011 № 87, 11'2011 № 121;
    2. **СКИФ-Аврора ЮУрГУ** (модель СКИФ 4/Н), 21.8/24[3] Tflops — 11'2009 № 450;

[1] — Отметим: к ноябрю 2011 года за всю историю только девять машин, разработанных в России, входили в мировой рейтинг Top500, семь из них — СКИФы.

[2] — 1 Gflops — миллиард (10^9) операций с плавающей точкой в секунду. 1 Tflops — триллион (10^{12}) операций с плавающей точкой в секунду; 1 Pflops — квинталион (10^{15}) операций с плавающей точкой в секунду.

[3] — Указана производительность на тесте Linpack и, через дробь, пиковая производительность.

3. **СКИФ МГУ «Чебышёв»**, 47.17/60 Tflops — 06’2008 № 36, 11’2008 № 54, 06'2009 № 82, 11'2009 № 103; 06'2010 № 121; 11'2010 № 179, 06'2010 № 353;
4. **СКИФ Урал**, 12.2/15.94 Tflops — 06’2008 № 283;
5. **СКИФ Cyberia**, 9.01/12 Tflops — 06’2007 № 105, 11’2007 № 200, 06’2008 № 485;
6. **СКИФ К-1000**, 2.032/2.534 Tflops — 11’2003 № 98, 06’2005 № 182, 11’2005 № 331, 06’2006 № 489;
7. **СКИФ К-500**, 0.424/0.717 Tflops — 11’2003 № 406;

- в последние годы 75–80% суперкомпьютеров **отечественной разработки** обеспечиваются суперЭВМ семейства СКИФ и установками с использованием технологических решений семейства СКИФ.

За время выполнения программ «СКИФ» и «СКИФ-ГРИД» команда исполнителей — СКИФ-кооперация,— прошла большой путь. В части разработки суперкомпьютеров семейства «СКИФ» этот путь можно поделить на некие этапы, посвященные созданию определенного *ряда* суперкомпьютеров семейства «СКИФ». Каждому ряду соответствуют (Таблица 1):

- некоторый временной отрезок разработки и выпуска моделей данного ряда;
- максимальная производительность, достижимая для изделий данного ряда;
- используемые технологические решения.

Первые установки были относительно просты, но они дали нам возможность освоить решения, которые для зарубежных коллег были стандартными и рутинными.

Затем появились и крупные проекты, такие как СКИФ К-500, СКИФ К-1000, СКИФ МГУ (названный впоследствии «Чебышёв»). Некоторые системы попадали в мировой рейтинг пятисот самых мощных машин в мире — Top500.

Таблица 1. Суперкомпьютеры семейства «СКИФ», ряды 1–4

Ряд	**Годы и пиковая производительность**	**Ядер в CPU / разрядность**	**Сетевые решения**	**Форм-фактор; *CPU/U***	**Примечания**
1	2000–2003 20–500 GFlops	1/32	Fast Ethernet, SCI (2D-тор), Myrinet	4U–1U; ***0.5–2***	Отечественная системная сеть SCI (2D-тор)
2	2003–2007 0.1–5 Tflops	1/32–64	GB Ethernet, SCI (3D-тор), InfiniBand	1U, Hyper-Blade ***2***	Сеть управления: ServNet v.1 и v.2. Ускорители: FPGA, ОВС
3	2007–2008 5–150 Tflops	2–4/64	GB Ethernet, InfiniBand DDR	1U, blades ***2–4***	Сеть управления: ServNet v.3, Три контура охлаждения: воздух–вода–фреон
4	2009–2012 ~0.5–5 Pflops	4–12/64	InfiniBand QDR, отечественная системная сеть 3D-тор	Сверхплотная упаковка, blades ***10.7***	Сеть управления: ServNet v.4 Ускорители: FPGA, GPU, МЦОС

Переход от ряда к ряду характеризовался серьёзными усовершенствованиями по тем или иным направлениям. В установках «СКИФ» всегда применялись современные процессоры, которые предоставляла индустрия. От модели к модели совершенствовались системная сеть (интерконнект, сеть, используемая для организации параллельного счета) и вспомогательная TCP/IP сеть, используемая для организации файловых обменов и управления. В части системной сети использовались как самые производительные готовые коммерчески доступные решения, так и выполнялись работы по реализации отече-

ственного интерконнекта. Системной сети вообще уделялось повышенное внимание при проектировании каждой машины.

Логика разработки суперкомпьютеров «СКИФ» ряда 4

Суперкомпьютеры «СКИФ» ряда 4 создавались в рамках выполнения суперкомпьютерной программы «СКИФ-ГРИД» Союзного государства [4]. Они являются результатом работы большого коллектива специалистов из России и Белоруссии. Непосредственно в разработке суперкомпьютеров «СКИФ» ряда 4 участвовали группы из семи организаций: ИПС имени А.К. Айламазяна РАН (головной), ЮУрГУ, ЗАО «РСК СКИФ», ОАО «НИЦЭВТ», ООО «Альт Линукс Технолоджи», ОИПИ НАН Беларуси, ИПМ имени М.В.Келдыша РАН. В создании, адаптации и оптимизации системного и прикладного программного обеспечения для суперкомпьютеров «СКИФ» ряда 4 участвовали двадцать организаций: ИПС имени А.К. Айламазяна РАН, ИПМ им. М.В.Келдыша РАН, ИСА РАН, ИММ РАН, ГЦ РАН, ИКИ РАН, ИПХФ РАН, ИХФ РАН, ЮУрГУ, СПбГПУ, НИИ КС, ННГУ, ТГУ, УГАТУ, МТУСИ, ЧелГУ, ЗАО «РСК СКИФ», ООО «Альт Линукс Технолоджи», ЗАО «Каледин и Партнеры», ЗАО «Сигма технологии».

Целью разработки суперкомпьютеров ряда 4 семейства «СКИФ» было создание решения, масштабируемого до рекордных систем. То есть, надо было создать решения, позволяющие при наличии заказа и адекватного разумного финансирования реализовать систему уровня Top1–5, что в 2009–2011 годах соответствует производительности 1–5 Pflops.

Данная цель — максимальное масштабирование, максимальная производительность,— определила логику разработки и принятия решений, которая и сформировала окончательный облик суперкомпьютеров «СКИФ» ряда 4 (рис. 1):

- Максимальное масштабирование, максимальная производительность влекут необходимость использования:
    - системной сети с характеристиками, которые превышают характеристики коммерчески доступных решений. Поэтому была разработана отечественная системная сеть с топологией 3D-тор;
    - сочетания лучших доступных стандартных i86–64 процессоров и ускорителей. Были серьезные причины того, что в качестве ускорителей нами использовались FPGA (обсуждено ниже).
- Системная сеть с топологией 3D-тор и максимальное масштабирование влекут две новые проблемы:
    - в системной сети 3D-тор требуется минимизировать длины соединительных линий, по возможности как можно большую их часть реализовать без кабелей и разъемов — на печатных платах. Тем самым, требуется максимизировать плотность упаковки электроники в единицу объема, создать установку с **максимальной плотностью упаковки** вычислительной мощности;
    - рекордная установка будет иметь огромные размеры — десятки тысяч вычислительных узлов и других модулей. Потребовалось серьезное внимание уделить вопросам **надежности, мониторинга и управления**. Для повышения надежности:
        - были предприняты усилия по минимизации использования кабелей и разъемов в системе: память вычислительных узлов впаяна в материнскую плату, максимально возможное число соединительных линий реализованы на печатных платах;

- в вычислительных узлах отказались от механических жестких дисков (элемент ненадежности и источник вибрации) и заменили их на твердотельные диски;
- была разработана оригинальная отечественная система мониторинга и управления суперкомпьютером.

- Максимальная плотность упаковки электроники означает **высокое тепловыделение в единице объема**. Старые схемы охлаждения вычислителя оказываются не применимыми. В суперкомпьютерах «СКИФ» ряда 4 было принято решение и реализовано **водяное охлаждение всей электроники вычислителя**: вычислительных узлов, блоков питания, управляющих узлов. Водяное охлаждение, несомненно, усложнило разработку. Но оно привнесло и ряд технологических преимуществ:
  - отсутствие подвижных частей в вычислителе, отсутствие шума и вибрации. Это весьма позитивно сказывается на показателях надежности и эргономике установки;
  - высокая энергоэффективность системы в целом. Экономия 40–50% электроэнергии, по сравнению с системами с использованием воздуха в одном из контуров охлаждения. Экономия в оборудовании. Возможность использования простых схем рекуперации тепловой энергии.

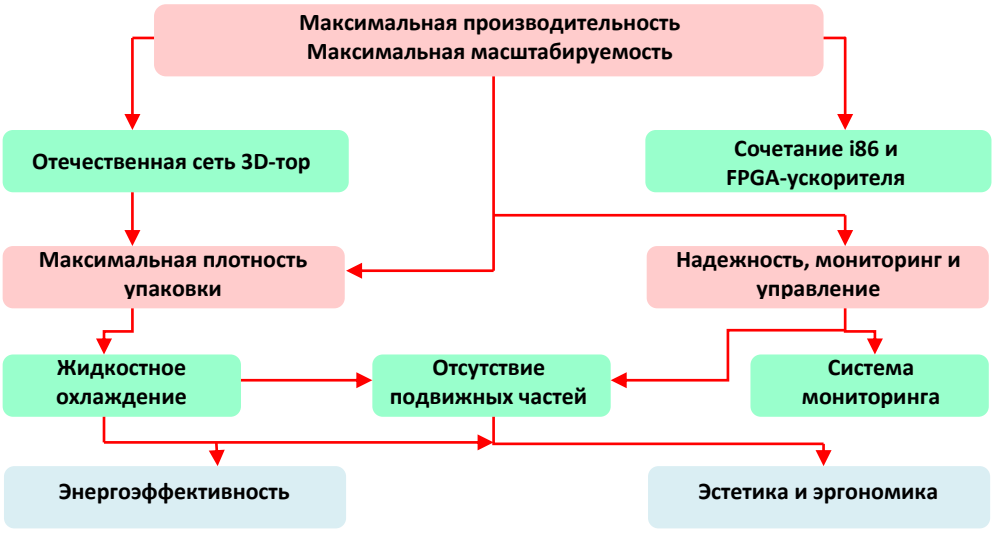

Рис. 1. Логика разработки суперкомпьютеров «СКИФ» ряд 4

Основные модули и система охлаждения суперкомпьютеров «СКИФ» ряд 4

На рис. 2 слева показан вычислительный узел суперкомпьютера «СКИФ» ряда 4. На первый, поверхностный взгляд, все выглядит достаточно стандартно: материнская плата собственной разработки, пара стандартных процессоров i86-64, память — для надежности впаяна в материнскую плату, разъем для твердотельного диска и т.д.

К совершенно непривычным техническим решениям следует отнести способ охлаждения электроники. На вычислительный узел накладывается так называемая охлаждающая пластина (рис. 2, справа), имеющая такой профиль, что все электронные компоненты материнской платы надежно прижимаются к охлаждающей платине. Пластина имеет два разъемных водяных соединителя — входной и выходной quick-коннектор,— для подключения к системе водяного охлаждения установки. Естественно, в пластине имеется фи-

гурная полость для протока воды от входного к выходному водяному разъему. Проходя по пластине, вода отнимает тепло, выделяемое электроникой, и уносит его в систему охлаждения. Заметим, что путь охлаждающей жидкости по плате вычислительного узла от впускного водяного разъема до выпускного показан на рис. 3 весьма условно; проектирование оптимальной формы движения жидкости для наиболее эффективного охлаждения электроники — не простая задача.

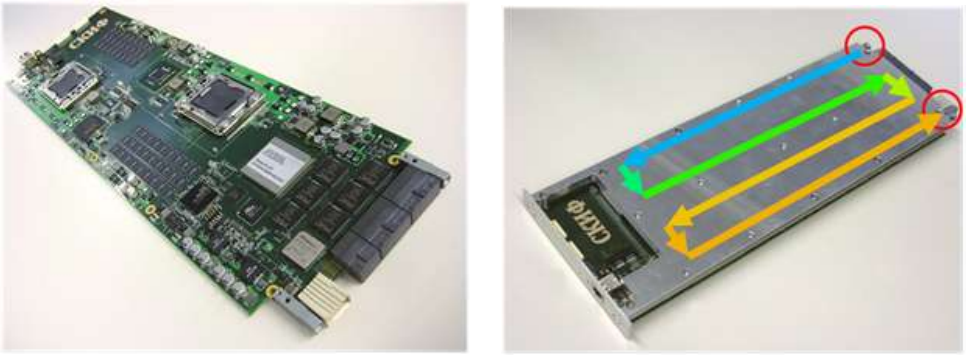

Рис. 2. Слева: вычислительный узел суперкомпьютера «СКИФ» ряда 4; справа: он же, с охлаждающей пластиной. Красным цветом обведены водяные разъемы — quick-коннекторы. Цветными стрелками показан путь охлаждающей жидкости.

На рис. 3 показаны основные модули и узлы вычислителя суперкомпьютеров «СКИФ» ряда 4: вычислительный узел, блок питания (48 V DC — 48 V DC, 6 КВатт), управляющий (корневой) узел. Один блок питания, один управляющий узел и 16 вычислительных узлов заполняют полушасси — основной модуль суперкомпьютера «СКИФ» ряда 4. Термин полушасси подчеркивает тот факт, что его глубина соответствует половинной глубине стандартной монтажной стойки.

Верхний ряд, слева направо: вычислительный узел, он же — с охлаждающей пластиной; блок питания и он же — с охлаждающей пластиной.
Средний ряд, слева направо: управляющий (корневой) узел, он же — с охлаждающей пластиной; полушасси со смонтированным блоком питания (снизу в полушасси) и управляющим узлом (сверху в полушасси); это же полушасси с установленными всеми шестнадцатью вычислительными модулями.
Нижний ряд: в центре — это же полушасси в изометрии; слева — оно же, но с сенсорным экраном в качестве крышки, справа — оно же, вид сзади, видны общие гидравлические разъемы полушасси.

Рис. 3. Основные модули и узлы вычислителя суперкомпьютеров «СКИФ» ряда 4

Полушасси является минимальным «строительным блоком», из которого собирается установка. Вся электроника в полушасси охлаждается водой. Не требуется сохранять проход для воздуха через полушасси — спереди полушасси закрывается как крышкой сенсорным LCD экраном. Экран доступен для работы с корневого узла и в штатном режиме используется как устройство ввода/вывода системой управления и мониторинга суперкомпьютера. Полушасси потребляет до 6 КВатт электроэнергии (48 V DC) и имеет пиковую производительность от 1.5 Tflops (модель СКИФ 4/N, 16×2×4=128 ядер CPU в полушасси) до 2.6 Tflops (модель СКИФ 4/W, 16×2×6=192 ядра CPU в полушасси).

Полушасси имеет высоту 6U и рассчитано для монтажа в стандартной монтажной стойке 48U. Так как глубина полушасси соответствует половинной глубине монтажной стойки и не требуется обеспечивать проход воздуха через монтажную стойку, то штатно в монтажной стойке 48U предусмотрена установка шестнадцати полушасси — восемь с одной стороны стойки и восемь — с другой. Каждое из шестнадцати полушасси содержит шестнадцать вычислительных модулей. В результате мы получаем монтажную стойку, в которой смонтированы 256 вычислительных модулей, содержащих 512 процессоров.

Такая монтажная стойка потребляет 100 Кватт электроэнергии (48 V DC) и обеспечивает пиковую производительность от 24 Tflops (модель СКИФ 4/N, 16×16×2×4 = 2048 ядер CPU в стойке) до 41 Tflops (модель СКИФ 4/W, 16×16×2×6 = 3072 ядра CPU в стойке). Вся электроника охлаждается водой.

Первый опытный образец суперкомпьютера «СКИФ» ряда 4 в масштабе одной монтажной стойки был установлен в суперкомпьютерном центре Южно-Уральского государственного университета — ЮУрГУ. Данная система была названа «СКИФ-Аврора ЮУрГУ». При пиковой производительности в 24 TFlops система показала 21.8 TFlops на тесте Linpack, что обеспечило ее вхождение в список Top500 на 450 место в ноябре 2009 года. На рис. 4 показан машинный зал установки площадью всего 30 кв. м, в центре которой расположена стойка вычислителя «СКИФ-Аврора ЮУрГУ», слева от которого расположена монтажная стойка с системой хранения данных и коммутаторами InfiniBand QDR.

Рис. 4. суперкомпьютер «СКИФ-Аврора ЮУрГУ», состояние на 2009–2010 годы

По левой и правой стене машинного зала стоят распределительные шкафы системы электропитания и другое вспомогательное оборудование. Фронтальная и задняя сторона стойки с вычислителем является сенсорным мультиэкраном из 8 LCD-панелей с отображением информации о текущем состоянии системы и с управлением «на кончиках пальцев». Система абсолютно беззвучная, не содержит ни одной механически подвижной части.

Монтажная стойка суперкомпьютера «СКИФ» ряда 4 сама по себе является модулем более высокого уровня. Если требуется вычислитель с производительностью больше, чем у одной стойки, то можно несколько стоек установить стена к стене, вдоль непрерывной линии и объединить в единую систему — обсуждается ниже. На рис. 5 представлен проект суперкомпьютера производительностью в 1 PFlops из 25 стоек вычислителя (модель СКИФ 4/W, 41 Tflops в стойке).

Рис. 5. 3D-модель проекта системы «СКИФ-Аврора» 1 Pflops

Еще раз подчеркнем, что подобный суперкомпьютер «СКИФ» ряда 4 бесшумен и эстетически красив, обладает высокими эргономичными свойствами: управление суперкомпьютером — на кончиках пальцев, через любой сенсорный экран в любой точке установки.

Преимущества водяного охлаждения

Вычислители крупнейших суперкомпьютеров потребляют очень много электроэнергии — от единиц до десятков мегаватт. Вся эта энергия в вычислителе преобразуется в тепло, которое выводится из вычислителя системой охлаждения. Сегодня в большинстве систем охлаждения вычислителей используется несколько контуров[4] с теплообменниками, причем контур, непосредственно контактирующий с электроникой — воздушный.

[4] — Классической является трехконтурная схема: воздух — вода — фреон.

Системы охлаждения, содержащие воздушный контур, имеют несколько слабых черт:

- Такие системы приводят к значительным затратам электроэнергии на охлаждение. В современных суперкомпьютерах, чтобы охладить вычислитель 100 КВатт, требуется затратить около 60 КВатт на систему охлаждения. Таким образом, на подобные системы охлаждения тратится от 45% до 50% подведенной к суперкомпьютерному центру электрической мощности;
- Кроме подведенной мощности на систему охлаждения тратятся и другие ресурсы суперкомпьютерного центра. Например, в крупных установках сегодня более половины[5] всей установочной площади используется не для размещения аппаратуры, а для её охлаждения. Так, по результатам анализа документации одного из крупнейшего Российского суперкомпьютерного центра было определено, что:
    - в машинном зале электроника вычислителя занимает 39 кв.м. и весит 50 500 кг;
    - в машинном зале компоненты подсистемы охлаждения (горячий коридор, внутрирядные кондиционеры) занимают 40 кв.м. и весят 14 800 кг;
    - вне стен машинного зала компоненты подсистемы охлаждения (чилеры, резерв холодной воды и т.п.) занимают еще 800 кв.м. и весят 180 000 кг;
- Низкая теплоемкость воздуха сильно снижает надежность охлаждения — в случае даже кратковременной остановки потока воздуха электроника быстро перегревается и может выйти из строя;
- В воздушном контуре для организации движения воздуха используются вентиляторы и мощные потоки воздуха. Это приводит в больших установках к низкой эргономике (шум, часто опасный для органов слуха человека), вибрации и ветровой нагрузке на печатные платы, контакты и кабели;
- Трудно организовать движение воздуха (как и любого газа) по траекториям, оптимальным для охлаждения электроники;
- Воздухом, из-за его низкой теплоемкости, сложно (а часто просто невозможно) охладить современные установки, с высокой плотностью тепловыделения[6].

С учетом сказанного, большинство разработчиков перспективных суперкомпьютеров исследуют новые схемы охлаждения. И одним из самых популярных направлений здесь является разработка систем водяного (жидкостного) охлаждения.

Большинство преимуществ водяного охлаждения вытекают из того, что теплоемкость кубометра воды 4 000 раз выше, чем теплоемкость кубометра воздуха и вода является жидкостью, а не газом. Поэтому в системе с жидкостным (водяным) охлаждением мы сразу получаем:

повышение надежности охлаждения — даже остановившаяся жидкость некоторое время успешно отводит тепло от электроники за счет своей высокой теплоемкости.

- Несложно организовать движение жидкости таким образом, чтобы обеспечить оптимальный отвод тепла от электроники.
- Не требуется для организации системы охлаждения тратить площадь машинного зала (на горячие коридоры, на внутрирядные кондиционеры).

---

[5] — Отметим, что с ростом вычислительной мощности установки, все затраты на её охлаждение будут расти нелинейно — чем мощнее установка, тем большая доля энергии и площади будет уходить на её охлаждение. Всё это приводит к выводу, что мощные вычислительные установки должны создаваться не с воздушным, а с водяным охлаждением.

[6] — Плотность тепловыделения в суперкомпьютерах «СКИФ» около 10 КВатт на 200 литров объема вычислителя. Для интуитивной оценки: 10 КВатт — это средняя мощность каменки в классической русской бане, а размер 200-литровой бочки представить себе не трудно.

- Можно использовать схемы с одним или максимум с двумя контурами охлаждения. В каждом контуре достаточно иметь одну[7] помповую станцию с дублированием. Помповую станцию легко вынести из вычислителя и, тем самым, полностью убрать шум и вибрацию.
- Охлаждение может быть обеспечено горячей водой и это проверено в суперкомпьютерах «СКИФ» ряда 4. При этом на входе в систему используется вода с температурой 50°C, а на выходе — 55°C. В этом случае все расходы на охлаждение можно свести практически к нулю[8], и для охлаждения нагретой воды (с 55°C до 50°C) могут использоваться как схемы с рекуперацией тепла, так и схемы «free cooling» — пассивные радиаторы, сухие градирни и другие устройства, позволяющие отдать тепло от воды с температурой 55°C в окружающую среду.

Сравнение с другими системами водяного охлаждения

Отметим, что «СКИФ-Аврора» — первая крупная вычислительная установка на стандартных процессорах, полностью охлаждаемая водой. Однако, преимущества водяного охлаждения столь очевидны, что разработки подобных систем в настоящее время ведутся практически во всех странах мира.

Например, всего лишь на несколько дней позже «СКИФ-Авроры» в США была запущена установка с охлаждением горячей водой «Aquasar» (IBM) [5].

Представляет интерес сравнить подходы к реализации охлаждения в «Aquasar» и в «СКИФ-Авроре» (см. рис. 6).

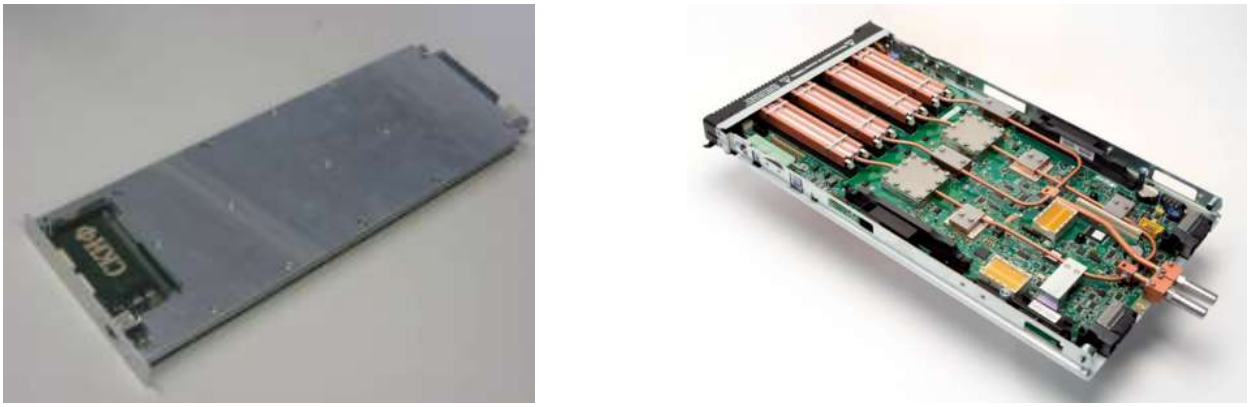

Рис. 6. Слева — плата «СКИФ-Аврора», справа — плата «Aquasar»

В «Aquasar» к охлаждаемым элементам жидкость подводится системой трубок различной толщины, проходящим вдоль платы. В «СКИФ-Авроре» имеется специальная, прилегающая к печатной плате, пластина охлаждения и канал для движения воды расположен внутри нее.

К преимуществам «Aquasar» следует отнести то, что конструкция получается более лёгкая по весу, однако системой водяного охлаждения охвачены не все электронные компоненты на плате.

Преимущества «СКИФ-Авроры», несомненно, в том, что цельнометаллическая охлаждающая пластина, присоединенная к печатной плате, придаёт последней большую жёсткость и способствует, таким образом, повышению надёжности системы в целом. Кроме

[7] — Вместо сотен и тысяч вентиляторов, используемых в воздушном контуре.
[8] — А именно, свести все затраты к электропитанию одной–двух помповых станций.

того, охлаждающая пластина примыкает ко всем электронным компанентам печатной платы.

Система электропитания суперкомпьютеров «СКИФ» ряда 4

Другим серьёзным техническим прорывом в архитектуре «СКИФ-Авроры» стало применение электропитания постоянного тока (48 V DC) на входе в машинный зал. Это решение позволяет (рис. 7) избежать лишнего преобразования электроэнергии в системе, отказаться от части оборудования в системе электропитания и, таким образом, удешевить установку, уменьшить её вес и габариты, повысить КПД подсистемы электропитания.

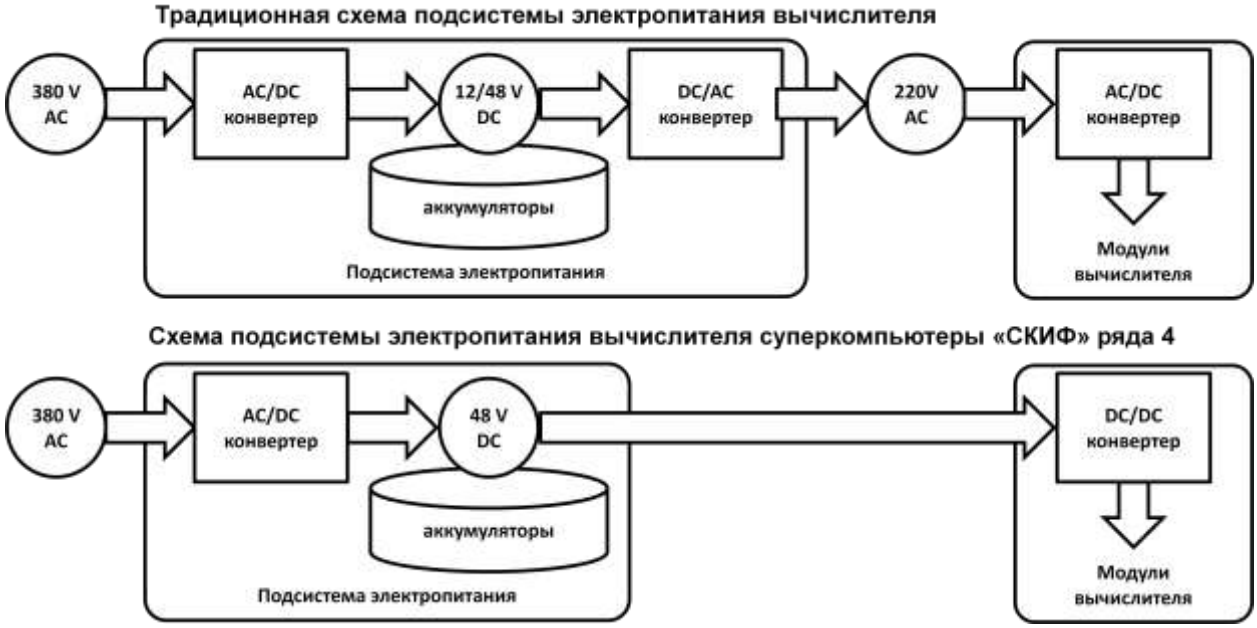

Рис. 7. Сравнение схем организации подсистемы электропитания:
предыдущие решения и суперкомпьютеры «СКИФ» ряда 4

Вычислительный модуль суперкомпьютера «СКИФ» ряда 4

Вычислительный модуль суперкомпьютера «СКИФ» ряда 4 содержит — рис. 8:

- два современных процессора Intel® Xeon® — Nehalem с 4 ядрами в моделях СКИФ 4/N; Westmare с 6 ядрами СКИФ 4/W;
- микросхему FPGA Altera Stratix IV, которая используется как FPGA-ускоритель и для реализации маршрутизатора отечественной системной сети «SKIF 3D-torus»;
- шесть трансиверов системной сети «SKIF 3D-torus», каждый из которых обеспечивает пропускную способность 10 Gbps — то есть суммарно 60 Gbps в каждом узле системной сети «SKIF 3D-torus»;
- микросхема сетевого адаптера InfiniBand QDR.

Отметим несколько обстоятельств:

- Вспомогательная сеть InfiniBand QDR используется для реализации TCP/IP сети для обмена файлами, управления задачами и т.п. Сетевой адаптер InfiniBand QDR имеет пропускную способность 40 Gbps и подключен к вычислительному узлу одинарной связью PCI Express Gen2, имеющей пропускную способность 40 Gbps;
- FPGA подключен к вычислительному узлу двойной связью PCI Express Gen2 с пропускной способность 80 Gbps. В этой связи только 60 Gbps в пределе может потребоваться

на поддержку обменов с системной сетью «SKIF 3D-torus». И, как минимум, 20 Gbps остаются доступными для организации взаимодействия с FPGA-ускорителем.

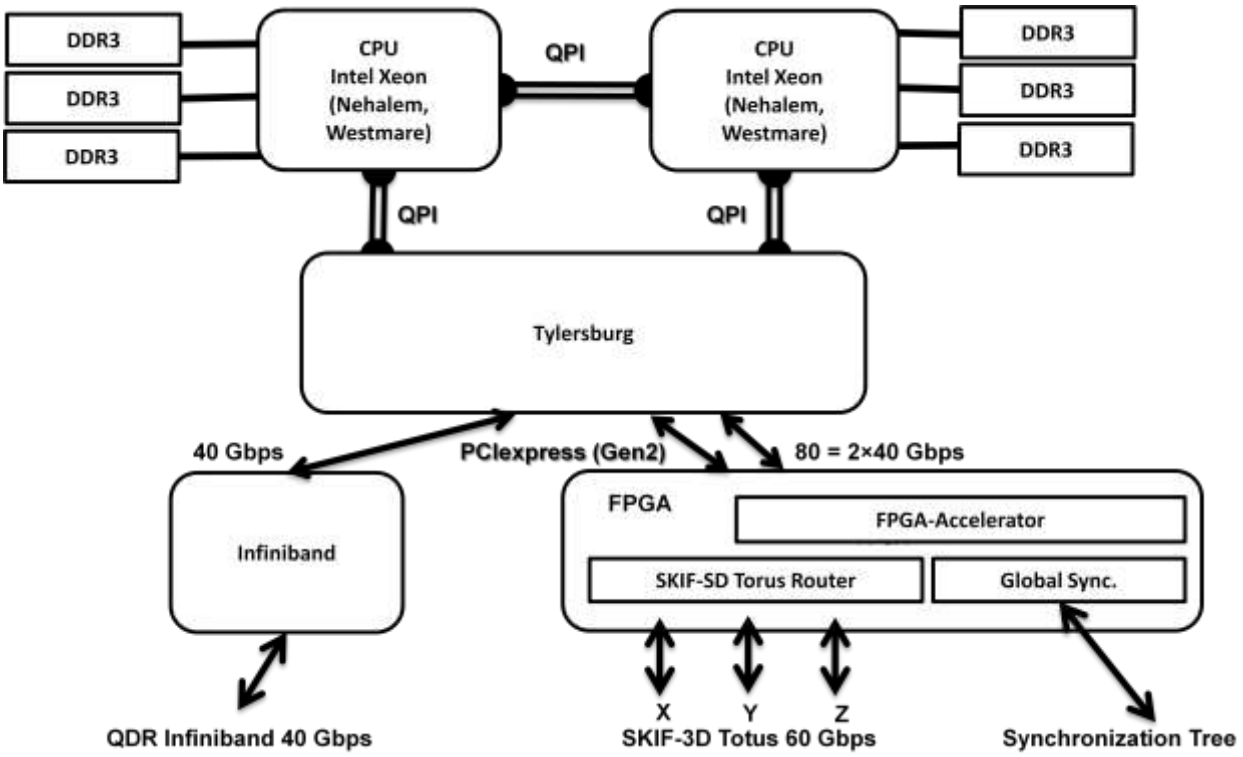

Рис. 8. Схема вычислительного узла суперкомпьютеров «СКИФ» ряда 4

Отечественная системная сеть «SKIF 3D-torus»

Максимальное масштабирование и максимальная производительность, поставленные как цели разработки суперкомпьютеров «СКИФ» ряда 4, повлекли за собою необходимость разработки собственной системной сети с характеристиками — пропускная способность, темп выдачи сообщений, задержка,— которые превышают характеристики коммерчески доступных решений. В результате этого была создана отечественная системная сеть «SKIF 3D-torus» с топологией трехмерного тора.

Трехмерный тор является хорошо масштабируемой топологией системной сети. Кроме того, для задач, связанных с моделированием процессов нашего реального трехмерного мира такая топология позволяет естественным образом отобразить задачу на системную сеть суперкомпьютера.

Серьезным преимуществом отечественной системной сети «SKIF 3D-torus» является ее гибкость, что включает в себя несколько обстоятельств:

- Сеть поддерживает разбиение ее на несвязанные трехмерные подторы меньшего размера и предоставление каждой решаемой задачи ее собственной системной подсети. Это повышает эффективность работы сети и исключает взаимное влияние одной задачи на другую.
- Маршрутизаторы сети реализованы в FPGA и, тем самым, поддерживают:
    - гибкую перенастройку. Это обеспечивает возможность одновременной поддержки разных алгоритмов маршрутизации и даже создания алгоритмов маршрутизации, учитывающих ту или иную специфику задачи;
    - гибкую аппаратную поддержку перспективных технологий организации параллельного счета и аппаратную поддержку операций, отличных от чистых опера-

ций передачи данных — например, поддержку вычислений во время передачи, поддержку примитивов класса all–reduce в сети, без использования процессоров вычислительных узлов.

Организация физических связей в отечественной системной сети «SKIF 3D-torus» очень хорошо отображена на различные уровни модульной конструкции суперкомпьютеров «СКИФ» ряда 4:

- первое измерение 3D-тора — ось X,— реализовано без кабелей — на соединительной панели полушасси и представляет из себя кольцо, включающее в себя шестнадцать вычислительных узлов в полушасси. Тем самым в каждом полушасси сформирован одномерный тор (кольцо) размером 16;
- второе измерение 3D-тора — ось Y,— реализовано кабельным соединением в кольцо шестнадцати корзин, смонтированных в стойке. Тем самым в каждой стойке сформирован двумерный тор размером 16×16;
- третье измерение 3D-тора — ось Z,— реализовано кабельным соединением нескольких стоек, расположенных вдоль некоторой непрерывной линии. Тем самым в системе формируется трехмерный тор размером 16×16×N, где N — количество стоек в системе.

Данное решение позволило унифицировать кабельное хозяйство — во всей системе используются кабели одной и той же длины.

Стек программного обеспечения системной сети «SKIF 3D-torus»

Схемная реализация (на языке VHDL) маршрутизатора на базе FPGA и весь стек программного обеспечения системной сети «SKIF 3D-torus» реализован в Институте программных систем имени А.К. Айламазяна РАН и (в части реализации библиотек SKIF-ARMCI, SKIF-GASNET) в ОАО «НИЦЭВТ».

На нижнем уровне стека находится драйвер SKIF-Driver для ОС «ALT Linux SKIF Cluster» и коммуникационная библиотека SkifCh, поддерживающие примитивы работы с маршрутизатором. Над этим уровнем надстроена реализация:

- SKIF-MPI — реализация стандартной библиотеки MPI версии 2;
- SKIF-SHMEM — реализация библиотеки, подобной библиотеке SHMEM компании CRAY [6];
- SKIF-ARMCI, SKIF-GASNET — реализация стандартных библиотек ARMCI [7], GASNET [8].

Подчеркнем, что отдавая дань поставленной цели — максимальное масштабирование и максимальная производительность,— были потрачены усилия на эффективную поддержку как прежней технологии параллельных вычислений (MPI), так и перспективных — SHMEM, ARMCI, GASNET и PGAS [9], что является базисом для перспективных языков и систем программирования: Unified Parallel C, Co-array Fortran, Titanium, Fortress, Chapel и др.

Технологические решения семейства «СКИФ» широко использовались во многих российских разработках за рамками программ «СКИФ» и «СКИФ-ГРИД». Новый яркий пример этого: в суперкомпьютере К-100 (41.1/107.9 Tflops, 2011 год, ИПМ имени М.В.Келдыша РАН, НИИ «Квант») для оригинальной отечественной системной сети «МВС-экспресс» был осуществлен перенос (с минимальной адаптацией) всего стека ПО «SKIF 3D-torus».

Технические характеристики системной сети «SKIF-3D-torus»

Разработанная в рамках программы «СКИФ-ГРИД» отечественная системная сеть «SKIF-3D-torus» показала следующие результаты — в сравнении с лучшим на сегодня коммерчески доступным решением InfiniBand QDR:

- пропускная способность 60 Gbps — в 1,5 раза выше, чем у InfiniBand QDR;
- темп выдачи сообщений 14 MT/s — примерно в 4–5 раз лучше, чем у InfiniBand QDR;
- задержка около 1–1.5 μs — сравнимо с InfiniBand QDR.

Отметим, что у современных приложений и перспективных подходов к организации параллельного счета высокие требования именно к темпу выдачи сообщений все чаще выходят на первый план.

Перспективы развития системной сети «SKIF-3D-torus»

Следующие направления работ являются перспективным развитием системной сети «SKIF-3D-torus»:

- реализация аппаратной поддержки коллективных операций — например, эффективная реализация традиционно «тяжёлых» операций MPI «all-reduce»;
- более эффективная аппаратная поддержка реализации парадигм SHMEM, GASNET и PGAS;
- специализация маршрутизатора под различные задачи — система со специализируемым под нужды конкретной задачи маршрутизатором могла бы сочетать в себе положительные стороны универсальной и специализированной машин, не унаследовав недостатков этих подходов.

Использование ускорителей в суперкомпьютерах «СКИФ» ряда 4

Сегодня многие разработчики суперкомпьютеров используют ускорители с фиксированной архитектурой, такие как GPGPU, ClearSpeed, Cell и др. Такие ускорители, в силу специализации (то есть, оптимизации) своей архитектуры под специфику целевого узкого класса задач:

- **в своей области специализации** — на задачах из узкого целевого класса,— являются существенно более эффективными, чем универсальные вычислители;
- **вне своей области специализации** — на задачах, не входящих в их узкий целевой класс,— являются существенно менее эффективными, чем универсальные вычислители.

В суперкомпьютерах «СКИФ» ряда 4 в качестве ускорителя используется FPGA. Для такого решения имеется ряд предпосылок.

Во-первых, FPGA можно рассматривать как реконфигурируемый ускоритель, у которого нет никакой архитектуры до той поры, пока его не запрограммировали. Таким образом, можно создавать свою собственную архитектуру ускорителя буквально для каждой задачи. В результате можно получить ускоритель, наиболее адекватно отражающий специфику решаемой задачи.

Во-вторых, FPGA уже имеется в составе каждого вычислительного модуля суперкомпьютера «СКИФ» ряда 4 и для реализации маршрутизатора системной сети «SKIF-3D-Torus» использовано весьма немного ресурсов (памяти, логических элементов) FPGA: 5–10%. Большая часть ресурсов не использована и может быть задействована для реализации ускорителей, специализированных под ту или иную задачу.

Суперкомпьютеры «СКИФ» ряда 4 не первые системы, где в качестве ускорителей используются FPGA. Примерами таких решений являются, например, суперкомпьютеры

CRAY XD1 (и другие изделия семейства Cray XD*), SGI Origin 2000/3000 + SGI Tensor Processing Unit, SGI Altix+FPGA RASC (Reconfigurable Application Specific Computing) и т.п. В силу своей исключительной эффективности подобные решения подпадают под ограничения на поставки в Россию.

Однако во всех подобных системах FPGA-ускорители организованы следующим образом:

- в каждом вычислительном узле имеется свой FPGA-ускоритель;
- FPGA-ускоритель может взаимодействовать только с соответствующей стандартной частью (со стандартным процессором) вычислительного узла;
- два FPGA-ускорителя не могут обмениваться данными между собою напрямую, только посредством нескольких косвенных передач: FPGA — стандартный процессор — системная сеть — стандартный процессор — FPGA.

Особенностью суперкомпьютеров «СКИФ» ряда 4 является то, что именно FPGA связаны между собою в системную сеть «SKIF 3D-torus». Тем самым, FPGA-ускорители могут обмениваться данными между собою напрямую, не вмешивая в этот процесс стандартные части своих вычислительных модулей. Это новая архитектура, новые возможности, которые еще предстоит до конца осмыслить и использовать в интересах решения конкретных задач [10].

Однако уже первые эксперименты с FPGA-ускорителями суперкомпьютеров «СКИФ» ряда 4 показали их высокую эффективность: задачи численных расчетов с плавающей точкой с двойной точностью выполнялись на FPGA-ускорителе в 4 раза быстрее, чем на двух процессорах (то есть на 8 ядрах — пиковая производительность 94 Gflops) Intel Nehalem. Если положить, что это дает нам основание оценить пиковую производительность одного FPGA как 0.375 Tflops, то пиковая производительность всех FPGA-ускорителей из одной стойки суперкомпьютера «СКИФ» ряда 4 составит 96 Tflops. Неплохая добавка к производительности стандартной части этой же стойки — 41 Tflops.

Подсистема мониторинга и управления суперкомпьютеров «СКИФ» ряда 4

Для поддержки возможности создания рекордных установок в рамках разработки суперкомпьютеров «СКИФ» ряда 4 пришлось создать уникальную подсистему мониторинга и управления суперкомпьютером. Данная подсистема состоит из трех взаимодействующих, но независимых (способны работать самостоятельно, в некоторой части дублируя друг друга) уровней:

- уровень 1 — охватывает вычислительные модули, использует сеть TCP/IP поверх InfiniBand QDR, реализует все возможности стандарта IPMI;
- уровень 2 — охватывает вычислительные модули и блоки питания, использует сеть TCP/IP поверх InfiniBand QDR, реализует селективную сериальную консоль к любому вычислительному узлу, возможность включить или выключить (непосредственно на блоке питания) электропитание на любом вычислительном узле, на группе из 8 вычислительных узлов, на корневом узле;
- уровень 3 — независимая сенсорная сеть СКИФ ServNet v.4, охватывающая все блоки электропитания, все корневые и вычислительные узлы, использует выделенную сеть передачи данных с резервными каналами, содержит независимые сенсоры температуры, протечек и влажности, уровней напряжения в разных точках системы, механизмы включения/выключения (непосредственно на блоке питания) электропитания на любой группе из 8 вычислительных узлов и на корневом узле.

Сенсорная сеть СКИФ ServNet v.4 имеет свою собственную систему гарантированного электропитания. Электропотребление СКИФ ServNet v.4 — около 3 Ватт на стойку. СКИФ ServNet v.4 является интеллектуальной сенсорной сетью: в каждом узле этой сенсорной сети — а они расположены в каждом блоке электропитания, в каждом корневом или вычислительном узле,— имеется микроконтроллер или микропроцессор, поддерживающий возможности:

- дистанционного перепрограммирования;
- передачи в центральный сервер подсистемы мониторинга и управления информации и исполнения команд от центрального сервера;
- локального автономного принятия решения и реализации действий по предотвращению нежелательных последствий нештатных ситуаций — в случаях отсутствия связи с центральным сервером.

Центральный сервер подсистемы мониторинга и управления реализует сбор, хранение и обработку информации от всех уровней подсистемы мониторинга, визуализацию, обнаружение нештатных ситуаций, прогноз нештатных ситуаций, оповещение персонала, автономное принятие решения и реализации действий по предотвращению нежелательных последствий нештатных ситуаций.

В целом в суперкомпьютерах «СКИФ» ряда 4 реализована уникальная отечественная система мониторинга и управления с тройным резервированием и очень высокой степенью надежности.

Выводы: современное состояние и перспективы проекта

Первый опытный образец «СКИФ-Аврора ЮУрГУ» претерпел модернизацию (от модели СКИФ 4/N до СКИФ 4/W) и расширение. Это позволило системе занять 87-ое место в рейтинге Top500 в июне 2011 года с производительностью 100.40/117.00 Tflops и 121-ое место в ноябрьском рейтинге 2011 года [3]. Продолжается разработка прикладного программного обеспечения. Здесь следует отметить, что всё программное обеспечение системы, начиная с ОС, драйверов, интерконнекта и заканчивая прикладными системами,— отечественные разработки, в том числе и на базе программного обеспечения с открытыми кодами.

Обсуждая перспективы развития семейства суперкомпьютеров «СКИФ», конечно, надо думать о применимости накопленного опыта для движения в сторону к эксафлопсному рубежу. Общепринято, что с этим связывают следующий круг проблем:

- высокая плотность компоновки вычислителя;
    - сокращение физической длины соединений;
- снижение удельного потребления электроэнергии;
- эффективный и надежный отвод тепла;
- система обменов между вычислительными узлами с низкой задержкой, высокой пропускной способностью, поддержкой интеграции очень большой системы ($N \times 10^6$);
- мониторинг и управление большой системой;
- устойчивость системы к отказу части оборудования, парирование ошибок;
- новые подходы к организации параллельного выполнения программ;
- поддержка использования неоднородных ядер и ускорителей в составе процессоров и вычислительных узлов.

Было бы рано и излишне самонадеянно говорить, что решения всех перечисленных проблем уже найдены в наших проектах. Однако, проект создания суперкомпьютеров «СКИФ» ряда 4 даёт хорошую стартовую площадку для этой работы. По сути, в выполнен-

ном проекте исполнители прикоснулись ко всем проблемам и продвинулись в правильном направлении к их решению. Это верно для всех перечисленных проблем, за исключением, пожалуй, последней.

Авторы благодарны всем, с кем им посчастливилось работать вместе в рамках создания суперкомпьютеров семейства «СКИФ», а также руководству Союзного государства и Министерства образования и науки Российской Федерации за предоставленную возможность выполнения научно-технических программ «СКИФ» и «СКИФ-ГРИД».

Литература

1. Абламейко С.В., Абрамов С.М., Анищенко В.В., Парамонов Н.Н., Чиж О.П. Суперкомпьютерные конфигурации СКИФ. Минск: ОИПИ НАН Беларуси, 2005, цв. ил. — 170 с. — ISBN 985-6744-19-9.
2. Абрамов С.М. 2009. Суперкомпьютеры «СКИФ» — эффективная платформа для вычислений в интересах развития наукоемких технологий и наноиндустрии. Уфимский государственный авиационный технический университет, 2009. Материалы научно-технического совещания «Высокопроизводительные вычислительные ресурсы России для создания наукоемких технологий и развития инфраструктуры наноиндустрии», 14-17 октября 2008 г., Уфа. Изд-во: Уфимский государственный авиационный технический университет, 2009 г. стр. 35-51 ISBN 978-5-86911-919-3.
3. Top500 Supercomputer Sites — мировой рейтинг пятисот самых мощных компьютеров мира. — Информационный ресурс в сети Интернет, http://www.top500.org/.
4. С.М. Абрамов, В.Ф. Заднепровский, А.Б. Шмелев, А.А. Московский. 2009. Супер ЭВМ ряда 4 семейства СКИФ: штурм вершины суперкомпьютерных технологий. // Труды Международной научной конференции «Параллельные вычислительные технологии (ПаВТ'2009)», Нижний Новгород, 30 марта–3 апреля 2009 г., изд. Нижегородского государственного университета имени Н.И. Лобачевского, с. 5–16. ISBN 978-5-696-03854-4.
5. «Суперкомпьютер IBM Aquasar с системой водяного охлаждения» — HARDWAREPORTAL.RU, http://www.hwp.ru/news/Superkompyuter_IBM_Aquasar_s_sistemoy_vodyanogo_ohlazhdeniya_80768/
6. Электронный ресурс «Cray T3ETM Fortran Optimization Guide - 004-2518-002», Chapter 3. SHMEM, // Cray Research, Inc, 1999, доступен в Интернет как http://docs.cray.com/books/004-2518-002/html-004-2518-002/z826920364dep.html
7. Электронный ресурс «ARMCI — Aggregate Remote Memory Copy Interface» // Pacific Northwest National Laboratory, США, 2010, доступен в Интернет как http://www.emsl.pnl.gov/docs/parsoft/armci/
8. Электронный ресурс «GASNet» // Berkeley Lab, США, 2010, доступен в Интернет как http://gasnet.cs.berkeley.edu/
9. Электронный ресурс «PGAS — Partitioned global address space» // George Washington University, HPC Lab., http://hpcl2.hpcl.gwu.edu/
10. Абрамов С.М., Дбар С.А., Климов А.В., Климов Ю.А., Лацис А.О., Московский А.А., Орлов А.Ю., Шворин А.Б. Возможности суперкомпьютеров «СКИФ» ряда 4 по аппаратной поддержке в ПЛИС различных моделей параллельных вычислений // Материалы международной научно-технической конференции «Суперкомпьютерные технологии: разработка, программирование, применение» (СКТ–2010), 27 сентября–2 октября 2010, Дивноморское, Россия. — Таганрог: Издательство Технологического института Южного федерального университета, том 1, стр. 11–21. ISBN 978-5-8327-0383-1.