

С. М. Абрамов, И. А. Адамович, С. А. Блохин,  
А. В. Елистратов, Л. Я. Карачинский, Ю. А. Климов,  
И. И. Новиков, А. Ю. Орлов, А. Ю. Пономарев,  
Л. П. Прокофьева, С. С. Ранцев, И. А. Фохт, А. Ю. Хренов,  
А. Б. Шворин, Ю. В. Шевчук, В. В. Щербаков

## **Высокоскоростной интерконнект на активных оптических кабелях и программируемых логических интегральных схемах**

Аннотация. В докладе представлен разрабатываемый в Институте программных систем им. А.К. Айламазяна РАН высокоскоростной интерконнект на активных оптических кабелях (АОК) и программируемых логических интегральных схемах (ПЛИС). Данное изделие предназначено для построения коммуникационных сетей суперкомпьютеров. В рамках проекта разрабатывается плата интерконнекта, активные оптические кабели, а также аппаратное (на основе ПЛИС) и программное обеспечение. Технические характеристики интерконнекта находятся на современном уровне и будут обеспечивать скорость передачи данных до 56 Гбит/с между двумя платами по активному оптическому кабелю. Во время доклада будут продемонстрированы первые опытные образцы разрабатываемого интерконнекта.

*Ключевые слова и фразы:* интерконнект, коммуникационная сеть, активный оптический кабель, высокопроизводительная вычислительная система, суперкомпьютер.

### **Введение**

Современные высокопроизводительные вычисления невозможны без быстрых вычислительных устройств (процессоров, ускорителей вычислений) и быстрых подсистем передачи данных (как в рамках

---

Работа выполняется в рамках государственного контракта с Министерством промышленности и торговли Российской Федерации № 12411.1006899.11.105.

© С. М. Абрамов, И. А. Адамович, С. А. Блохин, А. В. Елистратов, Л. Я. Карачинский, Ю. А. Климов, И. И. Новиков, А. Ю. Орлов, А. Ю. Пономарев, Л. П. Прокофьева, С. С. Ранцев, И. А. Фохт, А. Ю. Хренов, А. Б. Шворин, Ю. В. Шевчук, В. В. Щербаков, 2013?

© ИПС им. А.К. Айламазяна РАН, 2013?

© ФТИ им. А.Ф. Иоффе, 2013?

© ЗАО «Центр ВОСПИ», 2013?

© **ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ**, 2013?

одного вычислительного узла между вычислительными устройствами и памятью, так и между вычислительными узлами). Во многих приложениях именно коммуникационные сети, связывающие узлы суперкомпьютеров, являются наиболее узким местом.

Наиболее распространенная и доступная на текущий момент коммуникационная сеть — InfiniBand FDR, обладающая высокими характеристиками: скоростью передачи данных 56 Гбит/с. Однако она не всегда способна удовлетворить все потребности, и многие фирмы (Cray, IBM, Fujitsu) разрабатывают собственные интерконнекты, которые применяются в самых крупных вычислительных системах. Так, например, первые пять установок в списке Top500 июня 2013 г. используют такие заказные сети. В то же время на многие топовые решения наложены экспортные ограничения, и, более того, весьма вероятно введение новых ограничений на последующие версии доступных в настоящее время решений. Это может повлечь недоступность коммуникационных сетей с требуемыми характеристиками для российских организаций. Указанные причины вынуждают развивать собственные технологии коммуникационных сетей.

Современные суперкомпьютеры требуют высокой скорости передачи на значительные расстояния. Поэтому применение медных кабелей, которые еще недавно были широко распространены, становится более затруднительным или вообще невозможным из-за высокой частоты передачи данных (десятки гигагерц). Для передачи данных между узлами суперкомпьютеров всё чаще используются активные оптические (оптоволоконные) кабели, способные работать на высокой скорости на значительные расстояния.

В Институте программных систем им. А.К. Айламазяна РАН при сотрудничестве с российскими фирмами в настоящее время ведется разработка современного интерконнекта для высокопроизводительных вычислительных систем. В рамках этого проекта ведется разработка как плат интерконнекта на основе программируемых логических интегральных схем (ПЛИС), так и активных оптических кабелей, используемых для соединения плат между собой, которые будут обеспечивать передачу данных на скоростях до 56 Гбит/с на расстояние до 50 метров. Также создается задел для разработки оптических кабелей, работающих на более высоких скоростях. В качестве аппаратной основы сетевого адаптера и коммутатора выбрана современная ПЛИС Altera Stratix V GX.

Нередко в качестве основы для интерконнекта выбирается ПЛИС как удобное, эффективное и, что важно, гибкое средство. Решения на основе ПЛИС, как правило, имеют гораздо меньший цикл разработки по сравнению аналогами на основе заказных микросхем. В дальнейшем на основе ПЛИС возможно изготовление заказных микросхем с целью удешевления изделия при массовом выпуске. В любом случае реализация на ПЛИС может выступать и как самостоятельное решение, и как прототип для будущей реализации в виде специализированной микросхемы. С другой стороны, использование ПЛИС позволяет быстро адаптировать интерконнект под различные области применения, в том числе и под различные топологии сети. ПЛИС также может использоваться в качестве ускорителя вычислений, что было успешно продемонстрировано в [1, 5].

## 1. Краткий обзор

ПЛИС используется во многих разрабатываемых интерконнектах: в проекте СКИФ-Аврора [2], в котором участвовали авторы, в проекте «Ангара», который ведется в ОАО «НИЦЭВТ» [6], в интерконнекте Extoll [8]. Общей характеристикой этих разработок также является то, что они представляют собой бескоммутаторные интерконнекты, в которых платы соединяются друг с другом без использования внешних коммутаторов, в роли которых выступают сами сетевые адаптеры.

Интерконнект Extoll разрабатывает одноименная немецкая фирма Extoll, начало которому было положено в Гейдельбергском университете. В настоящее время уже имеется две версии интерконнекта на ПЛИС, и, кроме того, фирма Extoll планирует выпустить версию на заказных микросхемах в этом году. Также Extoll развивает технологии активных оптических кабелей и производит 12-канальные кабели с пропускной способностью до 120 Гбит/с.

Отечественные разработчики также осознают важность проблемы интерконнекта и предлагают свои специализированные решения. В их число входит ОАО «НИЦЭВТ» с интерконнектом ЕС8430 «Ангара» [6]. Первоначальные прототипы разрабатывались с использованием ПЛИС, однако в настоящее время «НИЦЭВТ» готовится выпустить решение на основе заказной микросхемы. Интерконнект «Ангара» предназначен для построения сетей с топологией многомерного

тора. Заявлено наличие адаптивной маршрутизации, агрегации данных, аппаратной поддержки барьерной синхронизации и коллективных операций. В этой разработке используются 12-канальные медные кабели длиной около полутора метров и пропускной способностью 75 Гбит/с.

Нельзя не упомянуть наиболее распространенную в высокопроизводительных вычислениях сеть InfiniBand. Эта разработка, в отличие от приведенных выше, коммерчески доступна. Выпускаемая в настоящее время версия InfiniBand FDR обладает пропускной способностью 56 Гбит/с, и ведется разработка InfiniBand EDR с пропускной способностью 100 Гбит/с.

Еще один отечественный проект, в котором был успешно реализован интерконнект на основе ПЛИС, разрабатывался авторами в 2009–2010 годах. Данная сеть нашла свое применение в суперкомпьютере СКИФ-Аврора [2], установленном в Южно-Уральском государственном университете. Сеть имеет топологию 3D-тор, полная пропускная способность внешних каналов маршрутизатора составляет 60 Гбит/с. Представляемый здесь проект продолжает развитие этой технологии. Учитывая опыт предыдущих разработок, данный проект ставит своей целью достижение характеристик мирового уровня.

## 2. Разрабатываемый интерконнект

Разрабатываемая сеть состоит из плат интерконнекта, которые соединяются активными оптическими кабелями напрямую, без использования коммутаторов. Для такого рода бескоммутаторных сетей наиболее распространенная топология — многомерный тор, которая поддерживается в разрабатываемой сети. Использование оптических кабелей большой длины и архитектурная гибкость, которую дает ПЛИС, позволяет на практике реализовать и другие топологии, более эффективные или более специализированные под конкретные задачи.

Платы для интерконнекта разрабатываются в ИПС им. А.К. Айламазяна РАН (рис. 1). В качестве основной микросхемы, выполняющей роль маршрутизатора, используется установленная на плате ПЛИС фирмы Altera Stratix V серии GX, ориентированная на передачу значительных потоков данных [7]. Данная ПЛИС имеет большое число встроенных высокоскоростных трансиверов, рассчитанных на скорость до 14 Гбит/с, что позволяет получить скорость 56 Гбит/с



Рис. 1. Прототип платы разрабатываемого интерконнекта

на один кабель. Для подключения к узлу плата имеет разъем PCI-Express Gen3 x8, а для подключения активных высокоскоростных кабелей для межузловых соединений установлены разъемы QSFP+.

В качестве основного интерфейса передачи данных между маршрутизатором и CPU используется высокоскоростное соединение PCI-Express Gen3 x8. Обмен данными происходит по протоколу SkifCh [4], разработанному авторами. Протокол SkifCh реализован в специальной библиотеке, имеющей аппаратную поддержку в маршрутизаторе, что позволяет уменьшить накладные расходы и достичь высокой эффективности. Особенно заметный выигрыш по сравнению с InfiniBand достигается на такой характеристике как темп выдачи сообщений [3] на сообщениях короткой и средней длины.

Для унификации с имеющимся прикладным программным обеспечением реализована версия MPI, которая работает поверх интерфейса SkifCh. Поддерживаются и другие параллельные библиотеки и системы: SHMEM, Co-Array Fortran, UPC, GASNet. Для интерконнекта разрабатывается и системное программное обеспечение: драйвер ядра Linux и системные программы настройки и управления интерконнектом.

На основе программируемой логики ПЛИС реализован коммутатор, связывающий аппаратный блок PCI-Express и набор устройств, размещенных в ПЛИС. Схемы маршрутизации и арбитража также реализованы в ПЛИС, благодаря чему они могут быть сравнительно легко адаптированы под необходимую топологию сети.

Маршрутизатор занимает относительно мало основного ресурса ПЛИС — логических элементов. Поэтому одновременно с маршрутизатором в ПЛИС могут быть реализованы как специальные вычислительные устройства, тесно связанные с сетью, так и независимые ускорители вычислений [1, 5].

Функция прошивки основной ПЛИС Altera Stratix V (маршрутизатора) возложена на вспомогательную ПЛИС Altera MAX V. Эта микросхема изготовлена по технологии CPLD, благодаря чему ее собственная прошивка сохраняется в энергонезависимой памяти, что обеспечивает автоматическую инициализацию ПЛИС Altera Stratix V после включения питания. Для обновления сохраненной прошивки разработано соответствующее программное обеспечение.

Питание для всех элементов платы обеспечивает цифровой четырехтактный программно-управляемый преобразователь напряжения. Использование цифрового блока питания обеспечивает высокий КПД и полный контроль всех параметров в режиме реального времени. В микропроцессоре цифрового блока питания реализована шина VotikBus, разработанная в ИПС им. А.К. Айламазяна РАН. С помощью VotikBus можно удаленно собирать информацию с сенсоров, отвечающих за мониторинг температуры и скорости вращения вентиляторов, качество электропитания и состояние имеющегося на плате оборудования, а также управлять подсистемой питания и процессом прошивкой ПЛИС.

Существующие на рынке медные кабели, предназначенные для межузловых соединений, перестают удовлетворять современным требованиям как по скорости передачи данных, так и по необходимой длине кабеля. Поэтому существенную часть данного проекта занимает разработка активных оптических кабелей. Активный оптический (оптоволоконный) кабель (АОК, Active Optical Cable, АОС) представляет собой гибкое оптоволокно со стандартными разъемами QSFP+ на концах. В этих разъемах находятся активные оптические компоненты АОК: лазеры и фотодиоды, преобразующие электрический сигнал в оптический и обратно. Такая компоновка позволяет применять активные оптические кабели вместо традиционных медных кабелей без какого-либо изменения оборудования.

В рамках проекта впервые в России осуществляется разработка высокоскоростных многоканальных активных оптических кабелей. В ключевых компонентах активных оптических кабелей — линейных массивах вертикально-излучающих лазеров и p-i-n фотодиодов

— использованы уникальные разработки фирмы ООО «Коннектор Оптикс», позволяющие достичь результатов мирового уровня в данной области. На микроплате в разьеме QSFP+ размещается четверка таких оптических каналов, каждый из которых имеет пропускную способностью 14 Гбит/с, что дает в сумме 56 Гбит/с на кабель. Сохранена совместимость разъемов с InfiniBand FDR (4x14 Гбит/с), что позволяет использовать стороннее оборудование других фирм: как медные, так и оптические кабели. Маршрутизатор, как было отмечено выше, реализован в ПЛИС на основе программируемой логики. Его четыре внешних дуплексных канала выходят на встроенные в ПЛИС высокочастотные трансиверы, которые, в свою очередь, по плате выходят на разъемы QSFP+.

## Заключение

В статье представлен обзор проекта, выполняемого Институтом программных систем им. А.К. Айламазяна РАН в кооперации с отечественными компаниями, по разработке высокоскоростного интерконнекта на основе активных оптических кабелях и программируемых логических интегральных схемах (ПЛИС). Данная работа ведется на основе опыта, полученного коллективом при разработке интерконнекта для суперкомпьютера СКИФ-Аврора. Ключевым направлением данной разработки является освоение отечественными разработчиками скорости 56 Гбит/с во всех компонентах интерконнекта: ПЛИС, платы, активные оптические кабели.

## Список литературы

- [1] 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* // Суперкомпьютерные технологии: разработка, программирование, применение (СКТ-2010) : Материалы международной научно-технической конференции (27 сентября – 2 октября 2010 г., с. Дивноморское) — Таганрог : Изд-во ТТИ ЮФУ, 2010. Т. 1, с. 11–21. ↑**1**, **2**
- [2] Абрамов С. М., Заднепровский В. Ф., Лилитко Е. П. *Суперкомпьютеры «СКИФ» ряда 4* // Информационные технологии и вычислительные системы, 2012, № 1, с. 3–16. ↑**1**
- [3] Климов Ю. А., Орлов А. Ю., Шворин А. Б. *Темп выдачи сообщений как мера качества коммуникационной сети* // Научный сервис в сети Интернет: суперкомпьютерные центры и задачи : Труды Международной суперкомпьютерной конференции (20–25 сентября 2010 г., г. Новороссийск) — Москва : Изд-во МГУ, 2010, с. 414–417. ↑**2**

- [4] Климов Ю. А., Орлов А. Ю., Шворин А. Б. *SkifCh: эффективный коммуникационный интерфейс* // Вестник Южно-Уральского государственного университета. Серия «Математическое моделирование и программирование», 2011, № 25 (242), с. 98–106. ↑<sup>2</sup>
- [5] Лацис А. О., Дбар С. А., Плоткина Е. А., Андреев С. С. *Система программирования Автокод HDL и опыт ее применения для схемной реализации численных методов в FPGA* // Научный сервис в сети Интернет: масштабируемость, параллельность, эффективность : Труды Международной суперкомпьютерной конференции (21–26 сентября 2009 г., г. Новороссийск) — Москва : Изд-во МГУ, 2009, с. 237. ↑<sup>1</sup>, <sup>2</sup>
- [6] Слущкин А. И., Симонов А. С., Жабин И. А., Макагон Д. В., Сыромятников Е. Л. *Разработка межузловой коммуникационной сети EC8430 «Ангара» для перспективных российских суперкомпьютеров* // Успехи современной радиоэлектроники, 2012, № 1, с. 6–10. ↑<sup>1</sup>
- [7] ПЛИС Altera Stratix V GX, <http://www.altera.com/devices/fpga/stratix-fpgas/stratix-v/stxv-index.jsp>. ↑<sup>2</sup>
- [8] Extoll, <http://www.extoll.de/>. ↑<sup>1</sup>

Об авторах:

- Сергей Михайлович Абрамов**  
e-mail: [abram@botik.ru](mailto:abram@botik.ru)
- Игорь Алексеевич Адамович**  
e-mail: [iaadamovich@gmail.com](mailto:iaadamovich@gmail.com)
- Сергей Анатольевич Блохин**  
e-mail: [blokh@mail.ioffe.ru](mailto:blokh@mail.ioffe.ru)
- Алексей Викторович Елистратов**  
e-mail: [concept@pereslavl.ru](mailto:concept@pereslavl.ru)
- Леонид Яковлевич Карачинский**  
e-mail: [karach@switch.ioffe.ru](mailto:karach@switch.ioffe.ru)
- Юрий Андреевич Климов**  
e-mail: [yuri@klimov.net](mailto:yuri@klimov.net)
- Иннокентий Игоревич Новиков**  
e-mail: [novikov@switch.ioffe.ru](mailto:novikov@switch.ioffe.ru)
- Антон Юрьевич Орлов**  
e-mail: [orlov@mccme.ru](mailto:orlov@mccme.ru)
- Александр Юрьевич Пономарев**  
e-mail: [harry@opus.botik.ru](mailto:harry@opus.botik.ru)
- Лариса Петровна Прокофьева**  
e-mail: [info@centervospi.ru](mailto:info@centervospi.ru)
- Сергей Сергеевич Ранцев**

*e-mail:* [diklaker@gmail.com](mailto:diklaker@gmail.com)  
**Игорь Анатольевич Фохт**

*e-mail:* [vogt@vgt.botik.ru](mailto:vogt@vgt.botik.ru)  
**Андрей Юрьевич Хренов**

*e-mail:* [hau@hau.botik.ru](mailto:hau@hau.botik.ru)  
**Артем Борисович Шворин**

*e-mail:* [shvorin@gmail.com](mailto:shvorin@gmail.com)  
**Юрий Владимирович Шевчук**

*e-mail:* [shevchuk@botik.ru](mailto:shevchuk@botik.ru)  
**Владислав Васильевич Щербаков**

*e-mail:* [info@centervospi.ru](mailto:info@centervospi.ru)

*Образец ссылки на эту публикацию:*

С. М. Абрамов, И. А. Адамович, С. А. Блохин, А. В. Елистратов, Л. Я. Карачинский, Ю. А. Климов, И. И. Новиков, А. Ю. Орлов, А. Ю. Пonomarev, Л. П. Прокофьева, С. С. Ранцев, И. А. Фохт, А. Ю. Хренов, А. Б. Шворин, Ю. В. Шевчук, В. В. Щербаков. *Высокоскоростной интерконнект на активных оптических кабелях и программируемых логических интегральных схемах* // Программные системы: теория и приложения : электрон. научн. журн. 2013?. Т. 4?, № 4(17)?, с.??-??.

*URL:* <http://psta.psiras.ru/read/>

S. M. Abramov, I. A. Adamovich, S. A. Blokhin, A. V. Elistratov, L.Ya. Karachinsky, Yu. A. Klimov, I. I. Novikov, A.Yu. Orlov, A.Yu. Ponomarev, L. P. Prokofyeva, S.S. Rantsev, I. A. Fohkt, A.Yu. Khrenov, A. B. Shvorin, Yu. V. Shevchuk, V. V. Shcherbakov. *High performance interconnect based on active optic cables and field-programmable gate arrays.*

ABSTRACT. This article presents new high performance interconnect based on active optic cables (AOC) and field-programmable gate arrays (FPGA). A prototype of the interconnect is now in the process of development in Program systems institute of RAS. The interconnect is aimed to be used as the main communication network in supercomputers. Performance of the interconnect is about 56 Gbit/s of bandwidth per cable. During the presentation the sample interconnect boards and software are to be demonstrated. (*in Russian*).

*Key Words and Phrases:* interconnect, network, active optic cable, HPC, supercomputer.