ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №1(23), 2015, с. 109–120

УДК 004.7

Ю. А. Климов, А. Б. Шворин, А. Ю. Хренов, И. А. Адамович,
А. Ю. Орлов, С. М. Абрамов, Ю. В. Шевчук,
А. Ю. Пономарев

Паутина: высокоскоростная коммуникационная сеть

Аннотация. В статье представлена разработанная в Институте программных систем им. А.К. Айламазяна РАН в кооперации с отечественными компаниями высокоскоростная коммуникационная сеть Паутина, основанная на активных оптических кабелях (АОК) и программируемых логических интегральных схемах (ПЛИС). Данная сеть предназначена для использования в высокопроизводительных вычислительных системах (суперкомпьютерах). В рамках проекта разработана плата сетевого адаптера, активные оптические кабели, а также аппаратное (на основе ПЛИС) и программное обеспечение. Технические характеристики сети находятся на современном уровне и обеспечивают скорость передачи данных до 56 Гбит/с между двумя платами по активному оптическому кабелю.

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

Введение

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

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

Таблица 1. Доли технологий коммуникационных сетей в Top500

Технология	Количество	Производительность
InfiniBand	45%	34%
Ethernet	38%	13%
Заказные сети	17%	53%

она не всегда способна удовлетворить все потребности, и многие фирмы (Cray, IBM, Fujitsu) разрабатывают собственные сети, которые применяются в самых крупных вычислительных системах. Например, первые шесть установок в списке Top500 ноября 2014 г. [1] используют такие заказные сети. Хотя доля машин с заказными сетями в списке Top500 сравнительно невелика (см. таблицу 1), их суммарная производительность составляет более 50%.

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.

Чтобы построить топовый суперкомпьютер, нужно обладать компетенцией во всех вопросах, в частности, в коммуникационных сетях. Интересные нам топовые решения, как правило, нельзя просто купить из-за экспортных ограничений — надо создавать самим.

Известны разработки сетей как в России, так и за рубежом:

- **Ангара** — ОАО «НИЦЭВТ» [2];
- **МВС-Экспресс** — ИПМ им. М. В. Келдыша РАН и ФГУП «НИИ Квант» [3];
- **СМПО-10G** — ФГУП «РФЯЦ-ВНИИЭФ» [4];
- **СКИФ-Аврора**, **Паутина** — ИПС им. А. К. Айламазяна РАН [5];
- **Extoll** — EXTOLL GmbH [6].

1. Сеть Паутина

В ИПС им. А.К. Айламазяна РАН совместно с отечественными компаниями ООО «Коннектор Оптикс» и ЗАО «Центр ВОСПИ» в 2012–2014 гг. была разработана высокоскоростная бескоммутаторная

коммуникационная сеть, основанная на программируемых логических интегральных схемах (ПЛИС) и активных оптических кабелях (АОК) со следующими ключевыми характеристиками:

- соединение сетевого адаптера с процессором посредством PCI Express Gen3 x8 со скоростью передачи данных 64 Гбит/с;
- межузловые соединения со скоростью 56 Гбит/с.

Сеть Паутина построена на сетевых адаптерах, которые соединяются между собой активными оптическими кабелями напрямую, без использования коммутаторов. Для такого рода бескоммутаторных сетей наиболее распространенная топология — многомерный тор, которая поддерживается и в Паутине. Использование оптических кабелей большой длины и архитектурная гибкость, которую дает ПЛИС, позволяет также реализовать другие топологии, более эффективные по общей производительности или более специализированные под конкретные задачи.

Платы сетевых адаптеров 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 (5SGXMA3K1F35C2N), 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 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+.

На рис. 2 представлена структура сетевого адаптера Паутины и соединенных с ним устройств.

2. Внешние соединения

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, которая была выполнена компаниями ООО «Коннектор Оптикс» и ЗАО «Центр ВОСПИ». Активный оптический (оптоволоконный) кабель (АОК, Active Optical

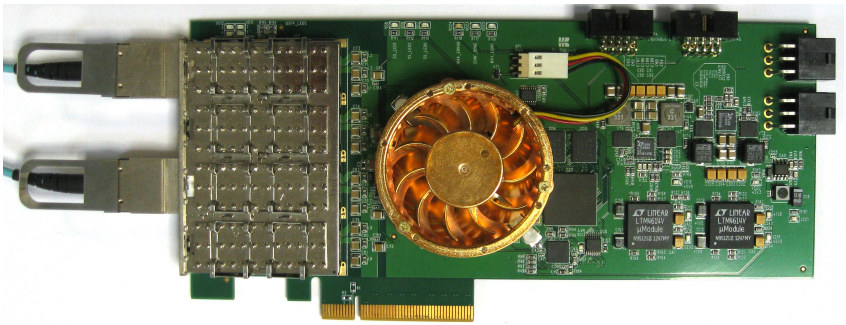


Рис. 1. Плата сетевого адаптера

Cable, AOC) 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 разработки фирмы ООО «Коннектор Оптикс», позволяющие достичь результатов мирового уровня в данной области. На микроплате в разъеме QSFP+ размещается четверка таких оптических каналов, каждый их которых имеет пропускную способность 14 Гбит/с, что дает в сумме 56 Гбит/с на кабель. Сохранена совместимость разъемов с InfiniBand FDR (4×14 Гбит/с), что позволяет использовать стороннее оборудование других фирм — как медные, так и оптические кабели.

3. Соединение с процессором

Основным интерфейсом передачи данных между сетевым адаптером и центральным процессором является высокоскоростное соединение PCI Express Gen3 x8. Реализация сетевого адаптера использует встроенные в ПЛИС модули: набор трансиверов и аппаратное ядро PCIe, которое реализует 256-битный интерфейс Avalon-ST в мультипакетном режиме.

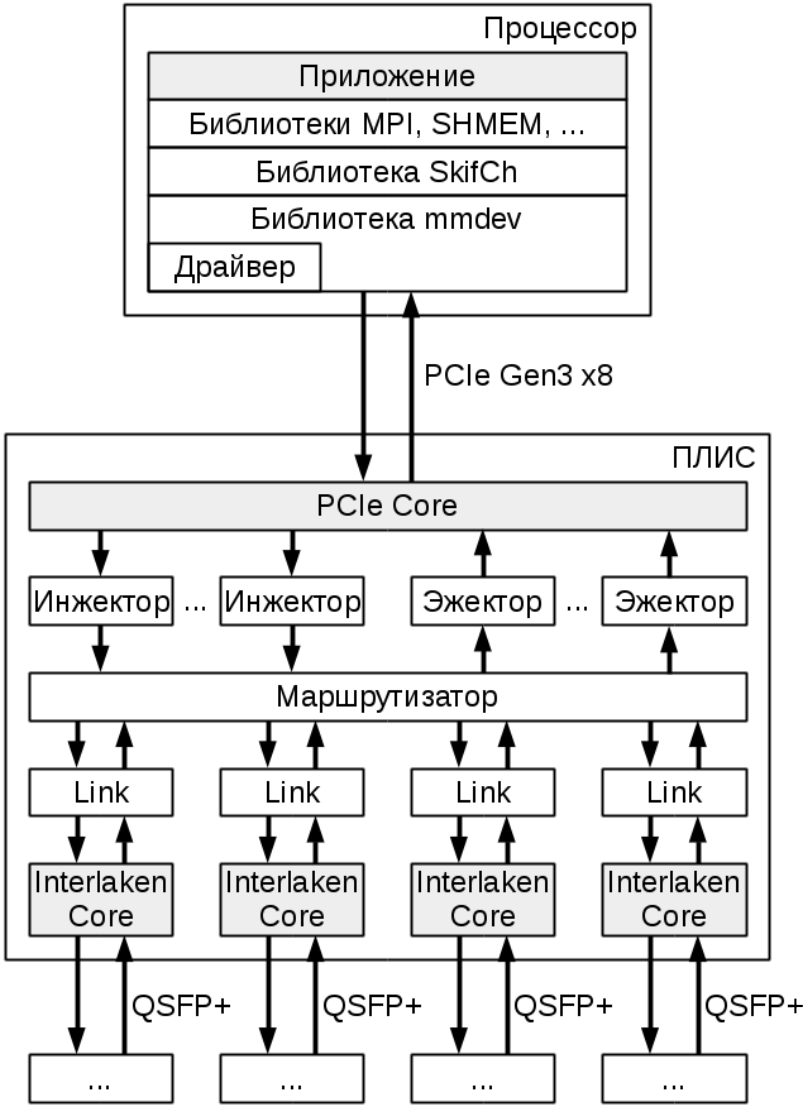


Рис. 2. Структура адаптера Паутины

Обмен данными происходит по протоколу SkifCh, разработанному авторами. Протокол SkifCh основан на кольцевых буферах и реализован в виде специальной библиотеки, 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 [8] на сообщениях короткой и средней длины. В работе [9] 2011 года приводится описание первой версии протокола SkifCh, которая использовалась в проекте СКИФ-Аврора. К настоящему времени протокол был существенно переработан для достижения большей эффективности.

Для унификации с имеющимся прикладным программным обеспечением реализована версия MPI, работающая поверх интерфейса SkifCh. Поддерживаются и другие параллельные библиотеки и системы: SHMEM, Co-Array Fortran, UPC, GASNet. Для поддержки сетевого оборудования разработано также системное программное обеспечение: драйвер ядра Linux и системные программы настройки и управления сетью.

Протокол SkifCh, используемый в Паутине, имеет ряд особенностей, способствующих достижению высокой производительности:

- Используются только операции записи PCIe. Интерфейс PCIe предоставляет также операции чтения, но, поскольку они требуют существенно бóльших накладных расходов, от них было решено отказаться.
- Все PCIe-пакеты выровнены на 64 байта.
- Используются PCIe-пакеты максимально допустимой длины.
- Отсутствует дополнительная синхронизация, связанная с уведомлением получателя о факте доставки сообщения.
- Заголовок сообщения размером 8 байт (включает в себя размер сообщения, адрес получателя, другие метаданные).

При этом протокол существенно асимметричен — передача от процессора к адаптеру и в обратном направлении организована поразному. Для «нисходящего» потока (от процессора в ПЛИС) характерны следующие особенности:

- Кольцевой буфер расположен в памяти ПЛИС и имеет размер ячейки 64 байта.
- Запись данных производится блоками по 64 байта с дополнением сообщения мусором.

- Аппаратный учет приходящих данных дает толерантность к нарушению порядка приходящих в ПЛИС PCIe-пакетов и «бесплатную» синхронизацию.
- В самом начале сообщения располагается заголовок размером 8 байт. Заголовок включает в себя размер сообщения, адрес получателя, бит четности, обозначающий четность номера прохода по кольцевому буферу, и другие метаданные.

«Восходящий» поток (из ПЛИС к процессору) организован со следующими особенностями:

- Кольцевой буфер расположен в системной памяти и имеет размер ячейки в 256 байт. Размер ячейки выбран равным максимальному размеру PCIe-пакета, который поддерживается аппаратурой.
- Для передачи сообщения используется минимально возможное количество PCIe-пакетов.
- Заголовок сообщения размещается в самом конце ячейки и содержит бит четности, что обеспечивает простую синхронизацию.

Инженерные решения, принятые при разработке и реализации в аппаратуре протокола SkifCh, опираются на детальное исследование особенностей взаимодействия современных процессоров с PCI Express. В частности, выравнивание PCIe-пакетов на 64 байта и дополнение полезных данных мусором необходимо для активации механизма аппаратной агрегации Write Combining, применение которого в несколько раз повышает реальную пропускную способность.

Для упрощения процесса разработки и аппаратной реализации протокола SkifCh был создан инструмент моделирования — эмулятор PCI Express [**10**], который позже выделился из Паутины в независимый проект [**11**].

4. Маршрутизация и пакетная передача данных

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.

В адаптере используется пузырьковая маршрутизация, которая гарантирует наличие буферного пространства на следующем узле

до начала передачи пакета, что обеспечивает отсутствие deadlock'ов. Минимальность маршрута гарантирует отсутствие livelock'ов.

Для уменьшения задержки используется «червячная» передача данных, которая реализует принцип VCT (virtual cut-through). Смысл VCT в том, что сообщение не накапливается в промежуточном узле, а начинает передаваться сразу, как только возможно. Таким образом, части одного сообщения в некоторый момент времени могут быть распределены по нескольким промежуточным узлам. При этом протокол передачи данных гарантирует, что в промежуточном узле, начавшем прием сообщения, достаточно места в буферной памяти, чтобы сохранить всё сообщение целиком.

Поскольку на межузловых соединениях возможны ошибки, приводящие к порче передаваемых данных, в Паутине используется механизм защиты от ошибок, основанный на буфере переповтора.

Паутина поддерживает следующие топологии сети:

- полносвязанная топология (до 5 узлов),
- 1D/2D-тор и решетка,
- сети Кэли.

Заключение

Проведенные измерения показывают следующие результаты:

- скорость передачи по линку — 53.4 Гбит/с;
- задержка в кабеле (при длине 50 м) — 0.45 мкс;
- скорость передачи данных между узлами — 45.8 Гбит/с;
- темп выдачи сообщений — 60–80 млн сообщений в секунду;
- задержка при передаче между узлами — 1.2 мкс;
- задержка при передаче между процессором и ПЛИС — 0.75 мкс.

В статье представлен обзор проекта по разработке высокоскоростного интерконнекта, базирующейся на активных оптических кабелях и программируемых логических интегральных схемах (ПЛИС), выполненного Институтом программных систем им. А.К. Айламазяна РАН в кооперации с отечественными компаниями ООО «Коннектор Оптикс» и ЗАО «Центр ВОСПИ».

Список литературы

[1] *Рейтинг производительности суперкомпьютеров Top500*, URL http://www.top500.org/lists/2014/11/ ↑ 110.

[2] А. И. Слуцкин, А. С. Симонов, И. А. Жабин, Д. В. Макагон, Е. Л. Сыромятников. «Разработка межузловой коммуникационной сети ЕС8430 «Ангара» для перспективных российских суперкомпьютеров», *Успехи современной радиоэлектроники*, 2012, №1, с. 6–10 ↑ 110.

[3] В. К. Левин, Б. Н. Четверушкин, Г. С. Елизаров, В. С. Горбунов, А. О. Лацис, В. В. Корнеев, А. А. Соколов, Д. В. Андрюшин, Ю. А. Климов. «Коммуникационная сеть МВС-Экспресс», *Информационные технологии и вычислительные системы*, 2014, №1, с. 10–24 ↑ 110.

[4] В. Г. Басалов, В. М. Вялухин. «Адаптивная система маршрутизации для отечественной системы межпроцессорных обменов СМПО-10G», *Вопросы атомной науки и техники. Серия: Математическое моделирование физических процессов*, 2012, №3, с. 64–70 ↑ 110.

[5] С. М. Абрамов, В. Ф. Заднепровский, Е. П. Лилитко. «Суперкомпьютеры «СКИФ» ряда 4», *Информационные технологии и вычислительные системы*, 2012, №1, с. 3–16 ↑ 110.

[6] *Extoll*, URL http://www.extoll.de/ ↑ 110.

[7] *ПЛИС Altera Stratix V GX*, URL http://www.altera.com/devices/fpga/stratix-fpgas/stratix-v/stxv-index.jsp ↑ 111.

[8] Ю. А. Климов, А. Ю. Орлов, А. Б. Шворин, «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), Изд-во МГУ, М., 2010, с. 414–417 ↑ 114.

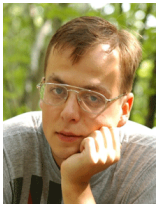
[9] Ю. А. Климов, А. Ю. Орлов, А. Б. Шворин. «SkifCh: эффективный коммуникационный интерфейс», *Вестник Южно-Уральского государственного университета. Серия «Математическое моделирование и программирование»*, **25(242)** (2011), с. 98–106 ↑ 114.

[10] А. Б. Шворин, «Эмулятор PCI Express для HDL-моделирования», *Научный сервис в сети Интернет: многообразие суперкомпьютерных миров*, Труды Международной суперкомпьютерной конференции (22–27 сентября 2014 г., г. Новороссийск), Изд-во МГУ, М., 2014, с. 395–400 ↑ 115.

[11] *Исходные коды эмулятора PCIe*, URL https://github.com/shvorin/pcie-emu ↑ 115.

Рекомендовал к публикации *Программный комитет*
Третьего национального суперкомпьютерного форума ***НСКФ-2014***

Об авторах:



Юрий Андреевич Климов

Ведущий инженер-программист ИПС им. А.К. Айламазяна РАН, с. н. с. ИПМ им. М.В. Келдыша РАН, к.ф.-м.н. Разработчик метода специализации на основе частичных вычислений, принимал активное участие в разработке коммуникационного программного обеспечения для сетей SCI, 3D-тор суперкомпьютера СКИФ-Аврора и «МВС-Экспресс» суперкомпьютера К-100.

e-mail: yuri@klimov.net

Артем Борисович Шворин

Инженер-программист ИПС имени А.К. Айламазяна РАН. Принимал активное участие в разработке коммуникационной сети 3D-тор суперкомпьютера СКИФ-Аврора. Область научных интересов: метавычисления, моделирование, функциональное программирование.

e-mail: shvorin@gmail.com

Андрей Юрьевич Хренов

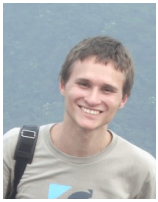
Инженер-электронщик Института программных систем имени А.К. Айламазяна РАН. Область интересов: проектирование цифровых высокоскоростных электронных устройств в системах автоматизированного проектирования.

e-mail: hau@hau.botik.ru

Игорь Алексеевич Адамович

Инженер-исследователь Института программных систем имени А.К. Айламазяна РАН. Область интересов: программирование ПЛИС.

e-mail: iaadamovich@gmail.com



Антон Юрьевич Орлов

Инженер-исследователь ИПС им. А.К. Айламазяна РАН. Разработчик новой версии языка Рефал+ и интегрированной среды разработки для него, принимал активное участие в разработке коммуникационной сети 3D-тор суперкомпьютера СКИФ-Аврора.

e-mail: orlov@mccme.ru

Сергей Михайлович Абрамов

Доктор физико-математических наук, член-корреспондент РАН. Директор Института программных систем имени А.К. Айламазяна РАН, ректор УГП имени А.К. Айламазяна РАН. Научный руководитель от России суперкомпьютерных программ «СКИФ» и «СКИФ-ГРИД» Союзного государства.

e-mail: abram@botik.ru

Юрий Владимирович Шевчук

Зав. лабораторией телекоммуникаций Института программных систем имени А.К. Айламазяна РАН, к.т.н. Область интересов: системное программирование, цифровая электроника, сети компьютеров, сенсорные сети, мониторинг и управление территориально распределенными объектами.

e-mail: shevchuk@botik.ru

Александр Юрьевич Пономарев

Ведущий инженер Института программных систем имени А.К. Айламазяна РАН. Область интересов: цифровая и аналоговая схемотехника, импульсные преобразователи напряжения.

e-mail: harry@opus.botik.ru

Пример ссылки на эту публикацию:

Ю. А. Климов, А. Б. Шворин и др. «Паутина: высокоскоростная коммуникационная сеть», *Программные системы: теория и приложения*, 2015, **6**:1(23), с. 109–120.

URL http://psta.psiras.ru/read/psta2015_1_109-120.pdf

Yuriy Klimov, Artem Shvorin, Andrey Khrenov, Igor Adamovich, Anton Orlov, Sergey Abramov, Yuriy Shevchuk, Aleksandr Ponomarev. *Pautina: the High Performance Interconnect.*

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 currently under development in the 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. (*in Russian*).

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

Sample citation of this publication

Yu. A. Klimov, A. B. Shvorin et al. "Pautina: the High Performance Interconnect", *Program systems: theory and applications*, 2015, **6**:1(23), pp. 109–120. (*In Russian.*)

URL http://psta.psiras.ru/read/psta2015_1_109-120.pdf