ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № ?, 2014, с. ??–??

УДК 004.7

Ю. А. Климов, А. Б. Шворин

Паутина: высокоскоростная коммуникационная сеть

Аннотация. В статье представлена разработанная в Институте программных систем им. А.К. Айламазяна РАН высокоскоростная коммуникационная сеть Паутина, основанная на активных оптических кабелях (АОК) и программируемых логических интегральных схемах (ПЛИС). Данная сеть предназначена для использования в высокопроизводительных вычислительных систем (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. Однако она не всегда способна удовлетворить все потребности, и многие фирмы

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

© Ю. А. Климов, А. Б. Шворин, 2014
© ИПМ им. М.В. Келдыша РАН, 2014
© ИПС им. А.К. Айламазяна РАН, 2014
© Программные системы: теория и приложения, 2014

(Cray, IBM, Fujitsu) разрабатывают собственные сети, которые применяются в самых крупных вычислительных системах. Например, первые шесть установок в списке Top500 июня 2014 г. [8] используют такие заказные сети. При этом, хотя доля машин с заказными сетями в списке Top500 сравнительно невелика, их суммарная производительность составляет более 50%.

	Доля в Top500	
	Количество	Производительность
InfiniBand	45%	32%
Ethernet	40%	15%
Заказные сети	15%	53%

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» [4],
- **МВС-Экспресс** — ИПМ им. М. В. Келдыша РАН и НИИ «Квант» [5],
- **СМПО-10GA-1** — РФЯЦ-ВНИИЭФ,
- **СКИФ-Аврора**, **Паутина** — ИПС им. А. К. Айламазяна РАН [1],
- **Extoll** — EXTOLL GmbH [7].

1. Сеть Паутина

В ИПС им. А.К. Айламазяна РАН в 2013–2014 гг. была разработана высокоскоростная бескоммутаторная коммуникационная сеть, основанная на программируемых логических интегральных схемах (ПЛИС) и активных оптических кабелях (АОК) со следующими 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.

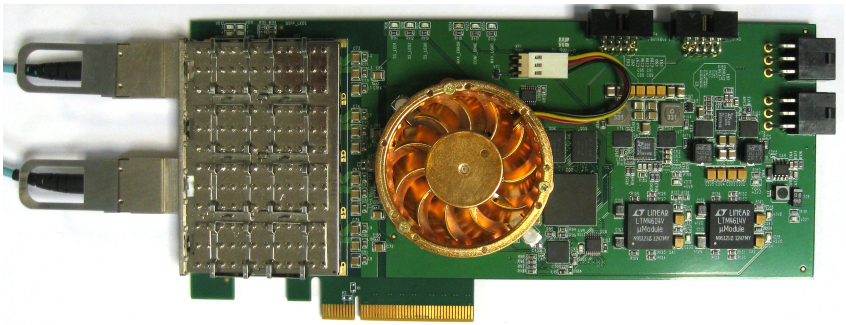

Рис. 1. Плата сетевого адаптера

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 в Паутине. Использование оптических кабелей большой длины и архитектурная гибкость, которую дает ПЛИС, позволяет на практике реализовать и другие топологии, более эффективные по 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 [6]. 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 тре-

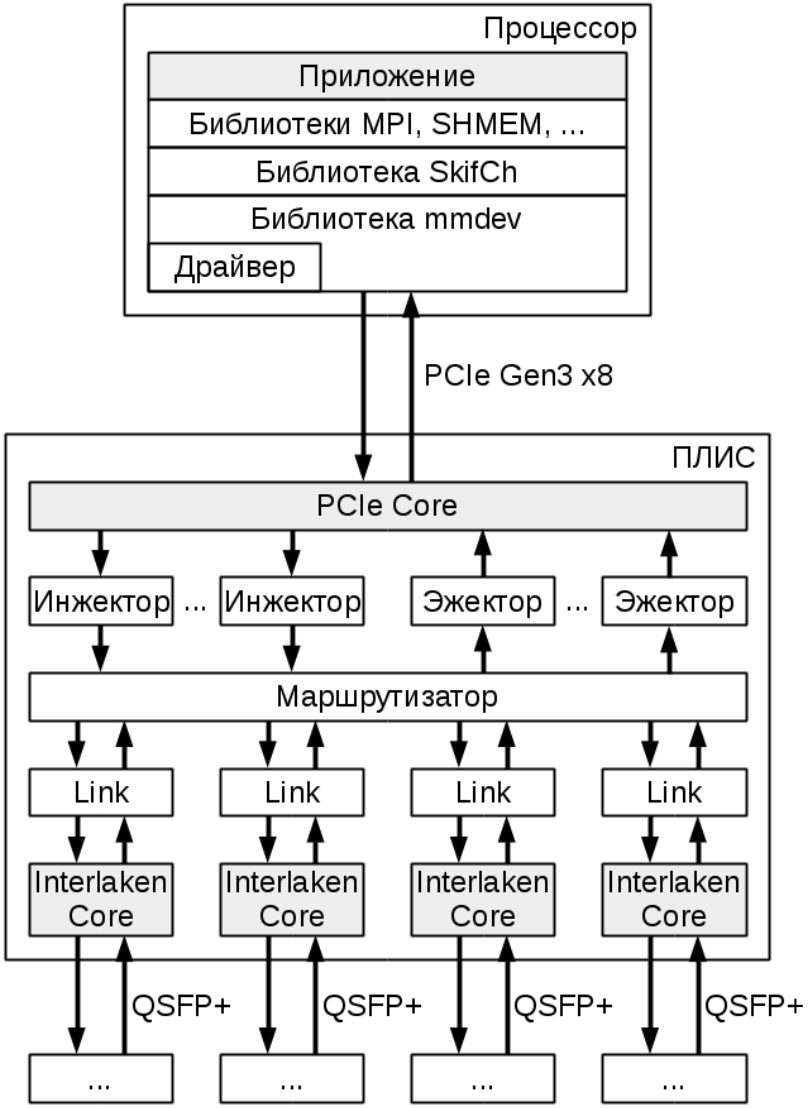

Рис. 2. Структура адаптера Паутины

бованиям 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, 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 способностью 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.

Обмен данными происходит по протоколу SkifCh, разработанному авторами. Протокол SkifCh основан на кольцевых буферах и реализован в виде 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 [2] на сообщениях короткой и средней длины. В работе [3] 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.

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 в 64 байта.
- Запись данных производится блоками по 64 байта с дополнением сообщения мусором.
- Аппаратный учет приходящих данных дает толерантность к нарушению порядка приходящих 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 в расположен в системной памяти и имеет размер ячейки в 256 байт. Размер ячейки выбран равным макси-

мальному 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 протокола SkifCh и его реализации в аппаратуре, 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.

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) — 0.4 мкс,
- скорость передачи данных между узлами — 40.6 Гбит/с,
- темп выдачи сообщений — 50–75 млн сообщений в секунду,
- задержка при передаче между узлами — 1.2 мкс,
- задержка при передаче между процессором и ПЛИС — 0.8 мкс.

В статье представлен обзор проекта, выполненного 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).

Список литературы

[1] С. М. Абрамов, В. Ф. Заднепровский, Е. П. Лилитко. *Суперкомпьютеры «СКИФ» ряда 4* // Информационные технологии и вычислительные системы, 2012. Т. **1**, с. 3–16. ↑2

[2] 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). — Москва: Изд-во МГУ, 2010, с. 414–417. ↑5

[3] Ю. А. Климов, А. Ю. Орлов, А. Б. Шворин. *SkifCh: эффективный коммуникационный интерфейс* // Вестник Южно-Уральского государственного университета. Серия «Математическое моделирование и программирование», 2011. Т. **25 (242)**, с. 98–106. ↑5

[4] 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. Т. **1**, с. 6–10. ↑2

[5] MVS Express. http://paco2012.ipu.ru/procdngs/P202.pdf. ↑2

[6] ПЛИС Altera Stratix V GX. http://www.altera.com/devices/fpga/stratix-fpgas/stratix-v/stxv-index.jsp. ↑3

[7] Extoll. http://www.extoll.de/. ↑2

[8] Рейтинг производительности суперкомпьютеров Top500. http://www.top500.org/lists/2014/06/. ↑2

Об авторах:

Юрий Андреевич Климов

Старший научный сотрудник ИПМ им. М.В. Келдыша РАН, ведущий инженер-программист ИПС им. А.К. Айламазяна РАН, к.ф.-м.н. Разработчик метода специализации на основе частичных вычислений для программ на объектно-ориентированных языках, принимал активное участие в разработке коммуникационного программного обеспечения для сетей SCI, 3D-тор суперкомпьютера СКИФ-Аврора и «МВС-Экспресс» суперкомпьютера К-100. Область научных интересов: суперкомпьютеры, оптимизация и преобразование программ, функциональное программирование.

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

Образец ссылки на эту публикацию:

Ю. А. Климов, А. Б. Шворин. *Паутина: высокоскоростная коммуникационная сеть* // Программные системы: теория и приложения: электрон. научн. журн. 2014. Т. ??, № ?, с. ??–??.

URL: http://psta.psiras.ru/read/

Yu. Klimov, A. Shvorin. *Pautina: the High Performance Interconnect.*

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