Кристалл для «Ангары»

Текст И. Жабин,
Д. Макагон,
А. Симонов,
Е. Сыромятников,
А. Фролов,
А. Щербак

В октябре компания «НИЦЭВТ» представила СБИС EC8430 маршрутизатора отечественной высокоскоростной сети «Ангара» для кластеров и суперкомпьютерных комплексов. Что представляет собой сеть «Ангара» и разработанная СБИС? На этот вопрос мы ответим в данной статье.

Сотни тысяч вычислительных ядер в суперкомпьютерах – это уже реальность. Очевидно, что от того, насколько эффективно обеспечивается обмен данными между ядрами, в конечном итоге и зависит, насколько эффективно их можно использовать одновременно для решения одной задачи. В этом смысле коммуникационная сеть является ключевым компонентом суперкомпьютера. Медленная сеть, не способная эффективно подстраиваться под возникающие отказы оборудования, которые при наличии миллионов компонентов перестают быть маловероятными, становится «узким местом», что для многих задач с большой долей сетевых обменов может быть весьма критичным. Фактически мы говорим о потенциале масштабируемости, который напрямую определяется именно характеристиками используемой коммуникационной сети.

Современные высокоскоростные сети

Прежде чем перейти к предметному обсуждению, необходимо определиться с терминологией. Коммуникационная сеть (интерконнект) состоит из узлов, в каждом из которых есть сетевой адаптер, соединённый с одним или несколькими маршрутизаторами, которые, в свою очередь, соединяются между собой высокоскоростными каналами связи (линками). Структура сети, определяющая, как именно связаны между собой узлы системы, задается топологией сети. В настоящее время наиболее распространены топологии «многомерный тор», fat tree, dragonfly. Архитектура маршрутизатора определяет структуру и функциональность блоков, отвечающих за передачу данных между узлами сети, а также необходимые свойства протоколов канального, сетевого и транспортного уровней, включая алгоритмы маршрутизации, арбитража и управления потоком данных. Архитектура сетевого адаптера определяет структуру и функциональность блоков, отвечающих за взаимодействие с хост-системой: процессором и памятью. На этом уровне, в частности, может осуществляться поддержка операций библиотеки MPI, обработка исключительных ситуаций, агрегация пакетов, поддержка механизма RDMA (Remote Direct Memory Access), обеспечивающего прямой доступ к памяти другого узла без участия его процессора.
Если посмотреть на статистику списка TOP500 (top500.org), то можно заметить, что большинство представленных в нем систем используют коммерчески доступные сети InfiniBand и Gigabit Ethernet. Однако суперкомпьютеры из первой десятки списка – китайские системы

Tianhe-2 и Tianhe-1A, японский K Computer, американские Cray Titan, IBM Blue Gene/Q – используют собственные уникальные («заказные») коммуникационные сети, разрабатываемые в составе этих вычислительных систем и доступные только совместно с ними. То есть, хотя в отличие от коммерчески доступных сетей, «заказные» сети занимают гораздо меньшую долю рынка, именно они используются в наиболее мощных суперкомпьютерах. Это, конечно же, неслучайно. Основная причина здесь в том, что для получения высокой производительности, сеть должна быть максимально интегрирована с вычислительной подсистемой и программным обеспечением (ОС на вычислительных узлах, библиотеки параллельного программирования, средства мониторинга и управления ресурсами системы). Более того, часто «заказные» системы (и их сети) подстраиваются под определённые классы целевых задач и, соответственно, предполагают необходимость тесного сотрудничества с организациями, решающими задачи в интересах государства.
Приобретение подобных машин в России в ряде случаев затруднено, а часто является просто невозможным. В то же время коммерчески доступные сети InfiniBand и Ethernet далеко не всегда подходят для эффективной реализации систем со столь высокими требованиями по масштабируемости, надежности и производительности. Можно также заметить, что в списке TOP500 нет сетей, использующих ПЛИС (FPGA). В связи с этим крайне актуальным является вопрос разработки отечественной высокоскоростной сети, сравнимой с западными «заказными» аналогами.
В таблице приведены основные характеристики сетей, используемых в наиболее мощных суперкомпьютерах, а также для сравнения – характеристики сети «Ангара» на базе ПЛИС и СБИС.
Необходимо отметить, что ряд отечественных организаций также достиг определенных успехов в разработке коммуникационных сетей для суперкомпьютеров, в том числе РФЯЦ ВНИИЭФ, ИПС РАН, ИПМ РАН и НИИ «Квант».

Проект разработки высокоскоростной сети «Ангара»

В сети «Ангара» маршрутизатор и адаптер находятся в одном кристалле (в отличие, например, от сети InfiniBand). Упрощенная блок-схема показана на рис. 1.
Топология сети – «многомерный тор» (до 4 измерений). Поддерживается надёжная передача пакетов по линку, детерминированная и адаптивная маршрутизация. Для предотвращения взаимных блокировок (deadlocks) детерминированной маршрутизации используется комбинация двух методов: метода «порядка направлений» (direction order) и «правило пузырька» (bubble-rule). Поддерживаются три RDMA-операции: асинхронные записи в память удаленного узла, асинхронные чтения и атомарные операции с удаленной памятью. Отдельный виртуальный канал используется для доставки ответов на чтения, чтобы предотвратить возникновение логических взаимных блокировок, обусловленных взаимозависимостью запросов и ответов. Эффективная работа с сетевым адаптером многоядерных процессоров поддерживается с помощью нескольких инжекционных конвейеров.
Взаимодействие вычислительного узла (т. е. кода, исполняемого на центральном процессоре) с маршрутизатором осуществляется путем записи данных по адресам памяти,

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ СОВРЕМЕННЫХ «ЗАКАЗНЫХ» КОММУНИКАЦИОННЫХ СЕТЕЙ И СЕТИ «АНГАРА»

СЕТЬ (СУПЕРКОМПЬЮТЕР)	TH Express-2 (Tianhe-2)	Cray Gemini (Titan)	IBM BlueGene/Q (Sequoia)	Tofu (K Computer)	InfiniBand FDR (Stampede)	Cray Aries (Cray XC30)	Ангара (ПЛИС)	Ангара (СБИС)
ГОД СОЗДАНИЯ СЕТИ	2013	2010	2011	2011	2011	2012	2010	2013
ТОПОЛОГИЯ	fat tree	3D-тор	5D-тор	6D-тор	fat tree	dragonfly	2D-тор	4D-тор
ПС ИНТЕРФЕЙСА С ХОСТ-СИСТЕМОЙ, ГБ/С	8 PCIe 2.0 x16	9,6 HyperTransport 3	~20 Custom	6,25 Custom	16 PCIe 3.0 x16	16 PCIe 3.0 x16	2 PCIe 1.0 x8	8 PCIe 2.0 x16
ПС ЛИНКА, ГБ/С	~4,55	9,375	2	5	6,8	5,25	0,625	7,5
ЗАДЕРЖКА МЕЖДУ СОСЕДНИМИ УЗЛАМИ, МКС	н/д	1,4	< 1,0	< 1,0	1,0	< 1,0	2,5	1,0
ТЕХПРОЦЕСС КРИСТАЛЛОВ СЕТ. АДАПТЕРОВ	90 nm	90 nm	45 nm Integrated into Compute chip	65 nm	65 nm	40 nm	(65 nm) FPGA	65 nm

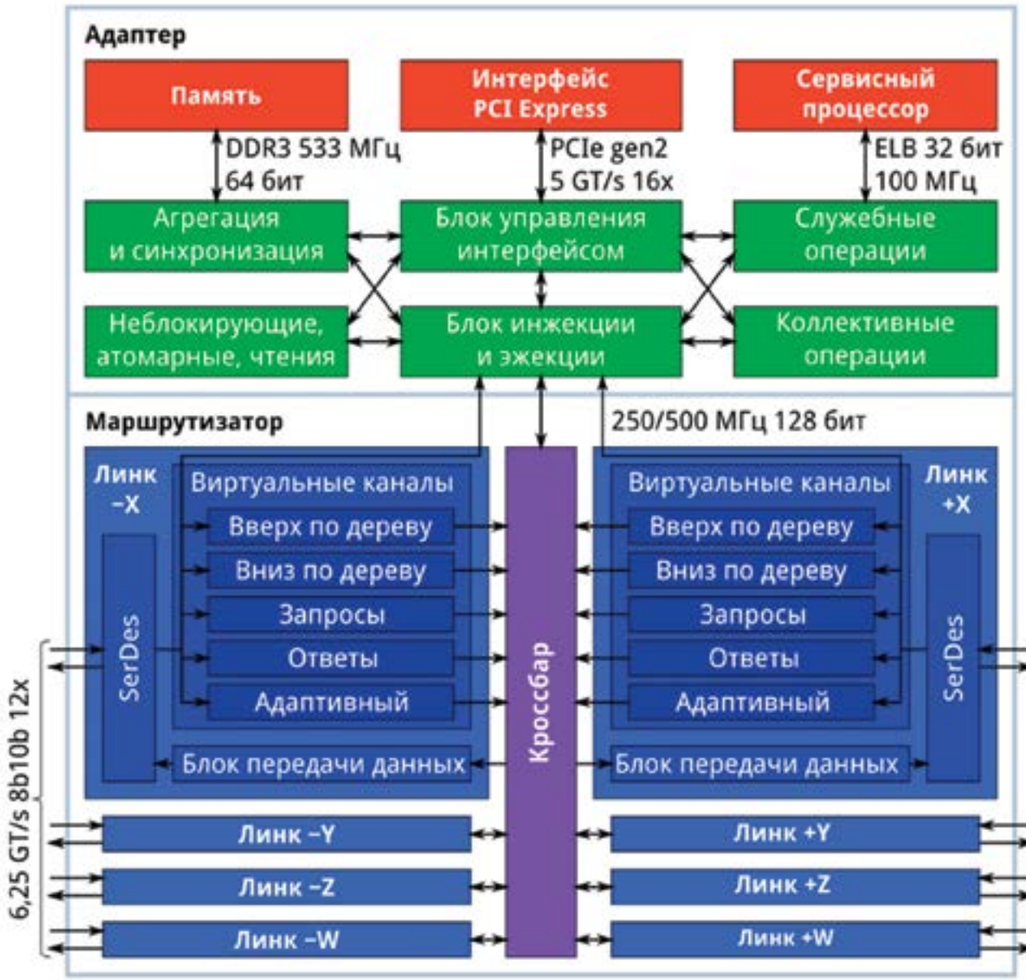

Рис. 1. Упрощенная блок-схема СБИС EC8430

которые отображены на адреса ресурсных регионов маршрутизатора (memory-mapped input/output). Это позволяет приложению взаимодействовать с маршрутизатором без участия ядра ОС, что снижает накладные расходы при отправке пакетов, поскольку переключение в контекст ядра и обратно занимает существенное время, в сравнении со временем отправки пакета. Аппаратная поддержка коллективных операций (например, broadcast – один узел рассылает данные группе узлов) реализуется на базе основной сети с топологией «многомерный тор», при этом используются отдельные виртуальные каналы, образующие виртуальную подсеть с древовидной топологией. В дереве задается корень, относительно которого вводятся два возможных направления движения по дереву: от корня и к корню. Каждому из направлений соответствует свой виртуальный канал. Чтобы предотвратить появление взаимных блокировок, дерево строится с учетом порядка измерений (dimension order). Для достижения большей эффективности было принято решение исключить из рассмотрения случаи, когда две разные задачи используют пересекающиеся группы узлов. С учетом этого каждый узел может относиться только к одной вычислительной задаче. Это позволяет исключить накладные расходы, связанные с использованием виртуальной памяти, избежать интерференции задач, упростить архитектуру маршрутизатора за счет отсутствия необходимости в полноценном MMU и избежать всех связанных с его работой коммуникационных задержек, упростить модель безопасности сети, исключив из нее обеспечение безопасности процессов различных задач на одном узле. Принятое решение не повлияло на функциональность сети, поскольку она предназначена в первую очередь для задач большого размера. Аналогичное решение было принято в IBM Blue Gene, с той разницей, что там ограничение на единственность задачи вводится для раздела. Основным режимом программирования для сети «Ангара» является совместное использование MPI, OpenMP и Shmem. Также поддерживаются библиотеки и языки параллельного программирования GASNet, UPC, ARMCI, Charm++.

Для обеспечения эффективного ввода-вывода используется параллельная файловая система Lustre. Во время подготовки к выпуску СБИС были добавлены функции для поддержки мониторинга, отладки и профилирования. В числе прочего был проработан механизм генерации прерываний для оповещения хост-системы о возникновении той или иной нештатной ситуации, например, об отказе какого-либо блока адаптера или о получении некорректного пакета из сети. Также были добавлены несколько сотен счетчиков производительности и подсистема конфигурирования отдельных блоков.

СБИС EC8430

Сеть «Ангара» — первый в России проект высокоскоростной сети с маршрутизаторами на основе СБИС отечественной разработки.

Микросхема EC8430 стала итогом семилетней работы подразделения ОАО «НИЦЭВТ» – разработчика высокоскоростной сети «Ангара». СБИС выпущена на фабрике TSMC с использованием технологии 65 нм. Размер кристалла – 13,0×10,5 мм, количество транзисторов – 180 миллионов. Кристалл размещен в корпусе flip-chip BGA, имеет 1521 вывод, размер подложки – 40×40 мм. СБИС работает на частоте 250/500 МГц и потребляет 36 Вт. Поддерживается топология сети «четырёхмерный тор», каждый сетевой узел может иметь до 8 соединений с соседними узлами, пропускная способность каждого соединения – 75 Гбит/с (12 линий по 6.25 Гбит/с, кодирование 8b10b). Взаимодействие с вычислительным узлом осуществляется через интерфейс PCI Express 2.0×16.

Размещение функциональных блоков в СБИС показано на рис. 2. Выделены следующие блоки:

- **PCIe** — *блок интерфейса PCI Express 2.0 (16x5 бит/с);*
- **BUI** — *блок приема и передачи данных через PCIe;*
- **PE** *и* **NI** — *блоки инжекции и эжекции сетевых пакетов;*
- **CROSSBAR** — *блок коммутации для передачи пакетов между линками и блоками инжекции и эжекции;*
- **LINK[0-7]** *и* **LINK[0-7] SerDes** — *блоки приема и передачи данных через линки в соседние узлы (8 линков по 12x6,25 Гбит/с);*
- **DDR3** — *интерфейс к DRAM- памяти (ширина 64+8 ECC = 72 бита, частота 533 Мгц (1066 MT/с));*
- **MIDP** — *блок приема и передачи данных через DDR3;*
- **GPIO** — *служебные интерфейсы конфигурирации (SPI Flash), отладки (JTAG), подключения сервисного процессора (SBUS);*
- **PLL** — *блок фазовой автоподстройки частоты.*

СБИС EC8430 используется в сетевых адаптерах «Ангара» в формфакторе плат расширения PCI Express для кластерных систем с коммерчески доступными процессорами (рис. 3). В настоящее время ведутся работы по интеграции СБИС в разрабатываемую в ОАО «НИЦЭВТ» вычислительную платформу для отечественного суперкомпьютерного комплекса «Ангара». Работы по созданию суперкомпьютера «Ангара», включая разработку СБИС EC8430, выполняются при финансовой поддержке Министерства промышленности и торговли РФ.

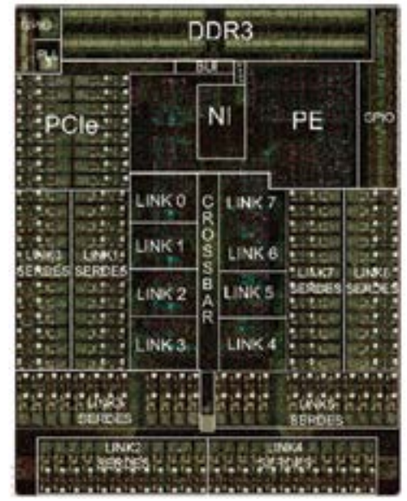

Рис. 2. Размещение функциональных блоков в СБИС EC8430

Рис. 3. Сетевой адаптер сети «Ангара» на базе СБИС EC8430