Открытые системы

Открытые системы № 7, 2011 21.09.2011

Сети для суперкомпьютеров

Сегодня сеть все чаще оказывается узким местом, сдерживающим рост быстродействия суперкомпьютеров. В каких направлениях идут разработки современных высокоскоростных коммуникационных сетей и какие проблемы поджидают разрабочиков?

Дмитрий Макагон, Евгений Сыромятников

Вычислительная мощность суперкомпьютеров растет высокими темпами за счет увеличения числа узлов, процессоров, ядер и внедрения различных ускорителей, однако эффективность использования суперкомпьютеров и их реальная производительность на многих задачах попадает в зависимость от реализации коммуникационных сетей. Для задач с малым количеством обменов по сети, которые легко разбиваются на независимые части, не требуется специализированных решений, и такие проекты, как GPUGRID.net и SETI@Home, это наглядно показывают. Однако для эффективного решения задач с интенсивным обменом данными на распределенной вычислительной системе требуется не только использовать специальные библиотеки и языки программирования, но и привлекать алгоритмы, адаптированные на параллельное выполнение с учетом специфики оборудования.

Сегодня используются следующие модели параллельного программирования: обмен сообщениями (двусторонние коммуникации, модель Send/Receive или MPI), когда процессы обмениваются между собой данными, явно выполняя отправку и прием сообщений; глобально адресуемая память, при которой любому процессу потенциально доступна вся память системы. В последнем случае память может быть физически общей либо распределенной. В системах с общей памятью (Symmetrical Multiprocessing, SMP) доступ к памяти для всех процессоров симметричен и занимает примерно одинаковое время. В системах с распределенной общей памятью (Disributed Shared Memory, DSM) у каждого процессора имеется своя локальная память, и некоторая часть ее доступна всем остальным узлам для удаленного доступа, а совокупность всех общедоступных регионов памяти образует распределенную общую память, время обращения к которой зависит от того, насколько далеко отстоит соответствующий узел.

Работа с распределенной общей памятью может происходить посредством библиотек с поддержкой односторонних коммуникаций (one-sided communications, модель Put/Get) и с помощью языков категории PGAS (Partitioned Global Address Space), позволяющих прозрачно работать с распределенной памятью как с памятью в SMP-системе, при этом учитывая локальность данных при выполнении операций. На модели двусторонних коммуникаций основан интерфейс MPI (Message Passing Interface). Для систем с десятками тысяч узлов модели SMP и двусторонние коммуникации оказываются непрактичными: с одной стороны, технически невозможно сделать SMP-системы с очень большим количеством узлов и объемом памяти, с другой — двусторонние обмены требуют явного участия в коммуникациях и отправляющей, и принимающей сторон, что затрудняет программирование. Кроме того, для формирования сообщений и их упорядочения требуются дополнительные ресурсы, что становится особенно затруднительным в случае большого количества обменов мелкими сообщениями между многими узлами. Наиболее перспективными при использовании систем такого масштаба оказываются PGAS-языки (UPC, Co-Array Fortran, X10) и библиотеки с поддержкой односторонних коммуникаций (Shmem, ARMCI, GASNet).

1. Коммуникационные сети

Эффективность суперкомпьютера на многих прикладных задачах в значительной мере определяется профилем работы с памятью и сетью. Профиль работы с памятью обычно описывается пространственно-временной локализацией обращений — размерами обращений и разбросами их адресов, а профиль

работы с сетью описывается распределением узлов, с которыми происходит обмен сообщениями, интенсивностью обмена и размерами сообщений.

Производительность суперкомпьютера на задачах с интенсивным обменом данными между узлами (задачи моделирования, задачи на графах и нерегулярных сетках, вычисления с использованием разреженных матриц) в основном определяется производительностью сети, поэтому применение обычных коммерческих решений (например, Gigabit Ethernet) крайне неэффективно. Однако реальная сеть — это всегда компромиссное решение, при разработке которого расставляются приоритеты между ценой, производительностью, энергопотерблением и другими требованиями, во многом конфликтующими между собой: попытки улучшения одной характеристики могут приводить к ухудшению другой.

Коммуникационная сеть состоит из узлов, в каждом из которых есть сетевой адаптер, соединенный 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 (обычно решетка, тор или толстое дерево) и набором структурных параметров: количество измерений, количество уровней дерева, размеры сторон тора, число коммутаторов на уровнях дерева, число узлов сети, портов у маршрутизаторов и т. д. На рис. 1 приведен пример топологии четырехмерный тор 3x3x3x3.

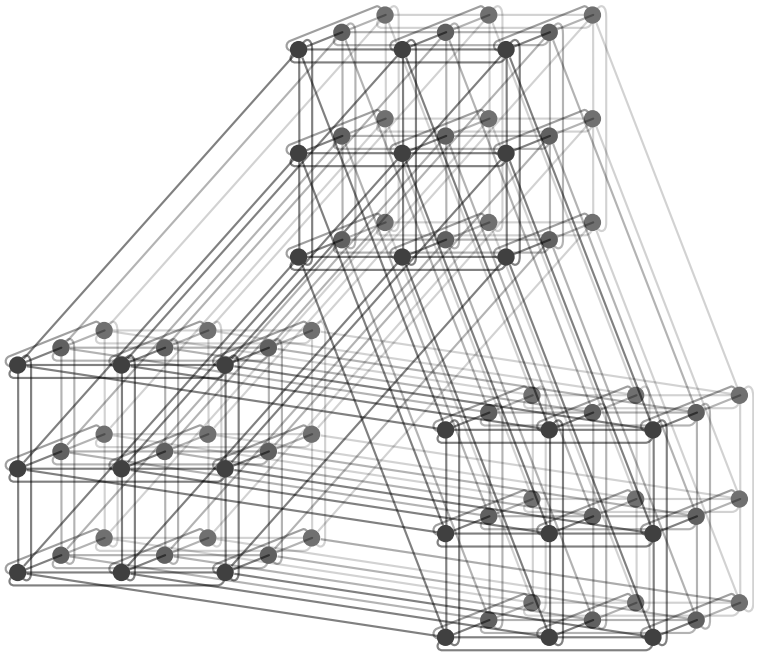

Рис. 1. Топология 4D-тор (3x3x3x3).

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 между процессором, памятью и сетью; в частности, на этом уровне осуществляется поддержка MPI-операций, RDMA (Remote Direct Memory Access — прямой доступ к памяти другого узла без участия его процессора), подтверждений получения другим узлом пакета, обработки исключительных ситуаций, агрегации пакетов.

Для оценки производительности коммуникационной сети чаще всего используются три характеристики: *пропускная способность* (количество данных, передаваемых за единицу времени); *коммуникационная задержка* (время передачи данных по сети); *темп выдачи сообщений* (обычно отдельно рассматривают темп выдачи при посылке, приеме и передаче пакетов между внутренними блоками

маршрутизатора).

Для полноты картины данные характеристики измеряются на разных видах трафика, например когда один узел рассылает данные всем остальным, либо, наоборот, все узлы шлют данные одному, либо когда все узлы посылают данные случайным адресатам. К современным сетям предъявляются требования по функциональности:

- эффективная реализация библиотеки Shmem, как варианта поддержки модели односторонних коммуникаций, и GASNet, на которой основаны реализации многих PGAS-языков;
- эффективная реализация MPI (обычно для этого требуется эффективная поддержка механизма кольцевых буферов и подтверждений для принятых пакетов);
- эффективная поддержка коллективных операций: широковещательной рассылки (посылки одинаковых данных одновремено многим узлам), редукции (применение бинарной операции, например сложения, ко множеству значений, получаемых от различных узлов), распределения элементов массива по множеству узлов (scatter), сборки массива из элементов, находящихся на разных узлах (gather);
- эффективная поддержка операций межузловой синхронизации (как минимум барьерной), эффективное взаимодействие с сетью большого количества процессов на узле, обеспечение надежной доставки пакетов.

Также важна эффективная поддержка работы адаптера с памятью узла напрямую без участия процессора.

2. Зарубежные высокоскоростные сети

Все коммуникационные сети можно разделить на два класса: коммерческие и заказные, разрабатываемые в составе вычислительных систем и доступные только вместе с ними. Среди коммерческих сетей рынок поделен между InfiniBand и Ethernet — в списке Top500 (июнь 2011 года) 42% систем используют InfiniBand и 45% — Gigabit Ethernet. При этом, если InfiniBand ориентирована на сегмент высокопроизводительных систем, рассчитанных на сложные вычислительные задачи с большим количеством коммуникаций, то Ethernet традиционно занимает нишу, где обмен данными между узлами некритичен. В суперкомпьютерах сеть Ethernet, благодаря своей дешевизне и доступности, зачастую используется в качестве вспомогательной сервисной сети с целью снижения интерференции управляющего трафика и трафика задач.

Сеть Inifiniband изначально была ориентирована на конфигурации с топологией Fat tree, но последние версии коммутаторов и маршрутизаторов (в первую очередь, производства QLogic) поддерживают топологию «многомерный тор» (с помощью Torus-2QoS Routing Engine), а также гибридную топологию из 3D-тора и Fat tree. Суперкомпьютер Sandia RedSky, собранный в начале 2010 года и находящийся сейчас на 16-м месте в Top500, является одним из первых масштабных проектов с сетью InfiniBand и топологией 3D-тор (6x6x8). Также много внимания сейчас уделяется эффективной поддержке RDMA-операций и библиотеки Shmem (в частности, Qlogic Shmem).

Популярность InfiniBand обусловлена ее относительно низкой стоимостью, развитой экосистемой программного обеспечения и эффективной поддержкой MPI. Однако InfiniBand имеет и свои недостатки: невысокий темп выдачи сообщений (40 млн сообщений в секунду в последних решениях от Mellanox), низкая эффективность передачи коротких пакетов, относительно большая задержка (более 1,5 мкс на передач узел-узел и дополнительно 0,1-0,5 мкс на каждый транзитный узел), слабая поддержка тороидальной топологии. В целом можно утверждать, что InfiniBand — это продукт для массового пользователя и при его разработке был сделан компромисс между эффективностью и универсальностью.

Можно также отметить готовящуюся к выводу на рынок сеть Extoll — разработку Гейдельбергского университета под руководством профессора Ульриха Брюнинга. Основной упор при разработке этой сети сделан на минимизацию задержек и повышение темпа выдачи при односторонних коммуникациях. Планируется, что Extoll будет иметь топологию 3D-тор и использовать оптические линки с пропускной способностью 10 Гбит/с на лейн (последовательный канал передачи данных в рамках линка) и шириной 12 лейнов на линк. Сейчас имеются прототипы сети Extoll на FPGA: R1 — на базе Virtex4,

R2 Ventoux — двухузловой макет на базе Virtex6. Односторонняя пропускная способность на один линк составляет 600 Мбайт/с (для R1). Также будут поддерживаться два интерфейса (HyperTransport 3.0 и PCI Express gen3) с процессором, что позволит интегрировать данную сеть в платформы Intel и AMD. В Extoll поддержаны несколько способов организации односторонних записей, собственный MMU (Memory Management Unit, блок трансляции виртуальных адресов в физические) и атомарные операции.

В отличие от коммерческих сетей, заказные занимают гораздо меньшую долю рынка, однако именно они используются в наиболее мощных суперкомпьютерах от Cray, IBM, SGI, Fujitsu, NEC и Bull. При проектировании заказных сетей разработчики имеют больше свободы и стараются применять более прогрессивные подходы в силу меньшей значимости рыночной привлекательности конечного продукта, решая в первую очередь задачу получения максимальной производительности на конкретном классе задач.

В суперкомпьютере K Computer используется коммуникационная сеть Tofu (TOrus FUsion) собственной разработки, представляющая собой масштабируемый 3D-тор, в узлах которого содержатся группы по 12 узлов (группы узлов соединены 12 сетями 3D-тор, а каждый узел из этой группы имеет выход в свою сеть 3D-тор). Узлы внутри каждой группы соединены между собой 3D-тором со сторонами 2x3x2 без дублирующих линков, что эквивалентно 2D-тору со сторонами 3x4 (итого получается 5D-тор с фиксированными двумя измерениями). Таким образом, узел сети Tofu имеет 10 линков с односторонне пропускной способностью в 40 Гбит/с каждый. Аппаратно поддерживаются барьерная синхронизация узлов и редукция (целочисленная и с плавающей запятой).

Основными целями при разработке суперкомпьютера Tianhe-1A были достижение высокой энергоэффективности, разработка собственного процессора и сети, превосходящей InfiniBand QDR. Суперкомпьютер состоит из 7168 вычислительных узлов, объединенных сетью Arch собственной разработки с топологией «толстое дерево». Сеть строится из 16-портовых маршрутизаторов, односторонняя пропускная способность линка — 8 Гбайт/с, задержка — 1,57 мкс. Поддержаны операции RDMA и оптимизированы коллективные операции.

Классическими представителями систем, использующих тороидальную топологию для объединения вычислительных узлов, являются системы для серии IBM Blue Gene, в первых двух поколениях которых — Blue Gene/L (2004) и Blue Gene/P (2007) — использовалась топология 3D-тор. Сеть в Blue Gene/P имеет относительно слабые линки с односторонней пропускной способностью 0,425 Гбайт/с, что на порядок меньше пропускной способности линка ее современника InfiniBand QDR, однако аппаратная поддержка барьерной синхронизации и коллективных операций (по отдельным древовидным сетям) позволяет достигать хорошей масштабируемости на реальных приложениях. Кроме того, все интерфейсы и блоки маршрутизации встроены в микропроцессор BPC (Blue Gene/P Chip), что значительно снижает задержки при передаче сообщений. Коммуникационная сеть следующего поколения Blue Gene/Q имеет топологию 5D-тор, и в отличие от предшественников в ней нет отдельных сетей для барьерной синхронизации и коллективных операций. Чип Blue Gene/Q впервые стал многоядерно-мультитредовым — четыре аппаратных треда на одно ядро при числе ядер 16, что позволяет ослабить требования к сети и обеспечить толерантность к задержкам. Пропускная способность линка увеличена до 2 Гбайт/с, но все равно остается небольшой по сравнению с Cray Gemini или Extoll. Низкая пропускная способность в этих системах нивелируется большой размерностью тора (большим количеством линков) и, как следствие, малым диаметром сети (значительно меньшим, чем у сетей с топологией 3D-тор с тем же числом узлов). В доступных источниках сообщается о создании двух транспетафлопсных суперкомпьютеров Blue Gene/Q: Sequoia с производительностью 20 PFLOPS и Mira — 10 PFLOPS. Можно сделать вывод, что Blue Gene/Q ориентирован на задачи, которые будут использовать десятки и сотни тысяч вычислительных узлов с сетевым трафиком типа «все — всем».

Другим приверженцем подхода к построению коммуникационных сетей с тороидальной топологией является компания Cray, которая продолжает использовать топологию 3D-тор, при этом наращивая пропускную способность и количество линков, соединяющих соседние узлы. Современным поколением тороидальной сети Cray является сеть Cray Gemini. Один маршрутизатор Gemini соответствует двум маршрутизаторам предыдущего поколения SeaStar2+, то есть фактически двум узлам сети, поэтому в Gemini вместо 6 линков для соединения с соседними узлами используется 10 (2 служат для соединения двух адаптеров между собой).

Компоненты (сетевые адаптеры, коммутаторы, маршрутизаторы) сети для суперкомпьютера, в отличие от процессоров, часто дороже, а доступ к ним более ограничен. Например, сейчас коммутаторы для сети InfiniBand, являющейся основной коммерческой сетью для суперкомпьютеров, производят всего две компании, при этом обе контролируются США. Это означает, что при отсутствии собственных разработок в области высокоскоростных сетей создание современных суперкомпьютеров в любой стране, кроме США, Китая или Японии, может легко контролироваться.

3. Отечественные сети

Разработка коммуникационных сетей для использования в суперкомпьютерах ведется рядом отечественных организаций: РФЯЦ ВНИИЭФ (о данных разработках имеется очень мало информации в открытых источниках); Институтом программных систем РАН и РСК «СКИФ»; ИПМ РАН и НИИ «Квант» (сеть «МВС-Экспресс»).

Коммуникационная сеть 3D-тор для российско-итальянского суперкомпьютера «СКИФ-Аврора» полностью построена с использованием FPGA Altera Stratix IV, что объясняет довольно небольшую пропускную способность на один линк — 1,25 Гбайт/с (ресурсы FPGA сильно ограничены).

В сети «МВС-Экспресс» для объединения вычислительных узлов используется PCI Express 2.0, при этом узлы объединяются через 24-портовые коммутаторы. Сеть имеет топологию, близкую к Fat tree. Сетевой адаптер в вычислительном узле имеет один порт шириной 4 лейна, вследствие чего односторонняя пиковая пропускная способность на линк составляет 20 Гбит/с без учета накладных расходов на кодирование. Преимуществом применения PCI Express в «МВС-Экспресс» является эффективная поддержка общей памяти с возможностью односторонних коммуникаций. Как следствие, сеть удобна для реализации библиотеки Shmem и PGAS-языков (UPC, CAF).

В ОАО «НИЦЭВТ» при поддержке Минпромторга РФ ведутся работы по разработке коммуникационной сети «Ангара» с топологией 4D-тор, которая может стать основой для создания отечественных технологий разработки суперкомпьютеров.

4. Сеть «Ангара»

Основные цели разработки сети «Ангара»:

- эффективная поддержка односторонних коммуникаций (put/get) и PGAS-языков (как основных средств параллельного программирования);
- эффективная поддержка MPI;
- выпуск собственного кристалла (для достижения высоких скоростей передачи данных и низких задержек);
- адаптивная отказоустойчивая передача пакетов;
- эффективная работа с современными процессорами и чипсетами.

На первом этапе разработки данной сети (2006 год) было проведено имитационное моделирование различных вариантов сети и приняты основные решения по топологии, архитектуре маршрутизатора, алгоритмам маршрутизации и арбитражу. Помимо тороидальной топологии рассматривались сети Кэли и «толстое дерево». Четырехмерный тор был выбран в силу более простой маршрутизации, хорошей масштабируемости, высокой связности по сравнению с торами меньшей размерности. Моделирование сети позволило детально изучить влияние различных параметров архитектуры сети на основные характеристики производительности, понять закономерности для трафика задач с интенсивным нерегулярным доступом к памяти. В результате были подобраны оптимальные размеры буферов, число виртуальных каналов и проанализированы потенциальные узкие места.

В 2008 году появился первый прототип маршрутизатора на FPGA — макет сети из шести узлов на Virtex4 [1], соединенных в тор 2x3, на котором была отлажена базовая функциональность маршрутизатора, отработана отказоустойчивая передача данных, были написаны и отлажены драйвер и библиотека нижнего уровня, портированы библиотеки Shmem и MPI. Сейчас запущен макет третьего поколения, состоящий из девяти узлов, соединенных в двухмерный тор 3x3. Собран стенд с двумя узлами для тестирования новых разъемов и каналов передачи данных, предполагаемых к использованию

с будущими кристаллами маршрутизатора ВКС. При разработке принципов работы сети ряд деталей был позаимствован из работ [2] и [3], а также в том или ином виде из архитектур IBM Blue Gene и Cray SeaStar.

Сеть «Ангара» имеет топологию 4D-тор. Поддерживается детерминированная маршрутизация, сохраняющая порядок передачи пакетов и предотвращающая появление дедлоков (взаимных блокировок), а также адаптивная маршрутизация, позволяющая одновременно использовать множество путей между узлами и обходить перегруженные и вышедшие из строя участки сети. Особое внимание было уделено поддержке коллективных операций (широковещательной рассылки и редукции), реализованных с помощью виртуальной подсети, имеющей топологию дерева, наложенного на многомерный тор. В сети на аппаратном уровне поддерживаются удаленные записи, чтения и атомарные операции двух типов (сложение и исключающее ИЛИ). Схема выполнения удаленного чтения (посылка запроса и прием ответа) приведена на рис. 2 (удаленная запись и атомарные операции выполняются аналогично). В отдельном блоке реализована логика по агрегации пришедших из сети сообщений с целью повышения доли полезных данных на транзакцию при передаче через интерфейс с хостом (хостом называется связка процессор-память-мосты).

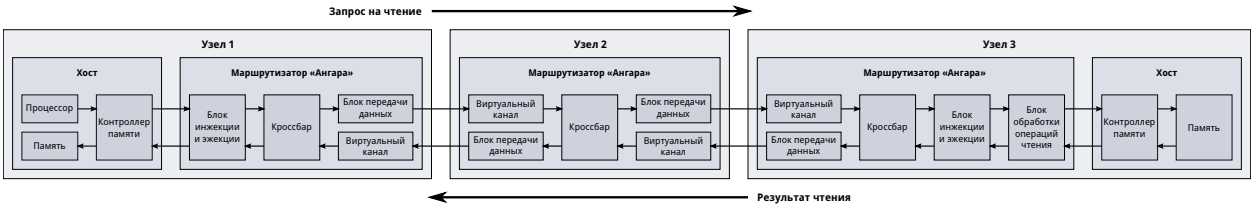

Рис. 2. Схема выполнения удаленного чтения в сети «Ангара».

На канальном уровне поддерживается отказоустойчивая передача пакетов. Существует также механизм обхода отказавших каналов связи и узлов с помощью перестройки таблиц маршрутизации. Для выполнения различных сервисных операций (в частности, настройки/перестройки таблиц маршрутизации) и выполнения некоторых расчетов используется сервисный процессор. В качестве интерфейса с хостом применяется PCI Express.

Основные блоки маршрутизатора (рис. 3):

- интерфейс с хост-системой, отвечающий за прием и отправку пакетов по хост-интерфейсу;
- блок инжекции и эжекции, формирующий пакеты на посылку в сеть и разбирающий заголовки пакетов, пришедших из сети;
- блок обработки запросов, обрабатывающий пакеты, требующие информации из памяти хост-системы (например, чтения или атомарные операции);
- блок сети коллективных операций, обрабатывающий пакеты, связанные с коллективными операциями, в частности, с выполнением редукционных операций, порождением пакетов широковещательных запросов;
- блок служебных операций, обрабатывающий пакеты, идущие в служебный сопроцессор и из него;
- коммутатор, соединяющий входы с различных виртуальных каналов и входы с инжекторов с выходами на различные направления и эжекторы;
- каналы связи для передачи и приема данных по определенному направлению;
- блок передачи данных для отправки пакетов по данному направлению и блок приема и маршрутизации для приема пакетов и принятия решения о дальнейшей их судьбе.

Взаимодействие хоста (код, исполняемый на центральном процессоре) с маршрутизатором осуществляется посредством записи по адресам памяти, отображенным на адреса ресурсных регионов маршрутизатора (memory-mapped input/output). Это позволяет приложению взаимодействовать с маршрутизатором без участия ядра, что снижает накладные расходы при отправке пакетов, поскольку переключение в ядерный контекст и обратно отнимает не одну сотню тактов. Для отправки пакетов используется один из регионов памяти, рассматриваемый как кольцевой буфер. Также имеется отдельный регион для выполнения операций без копирований память-память (данные читаются из памяти

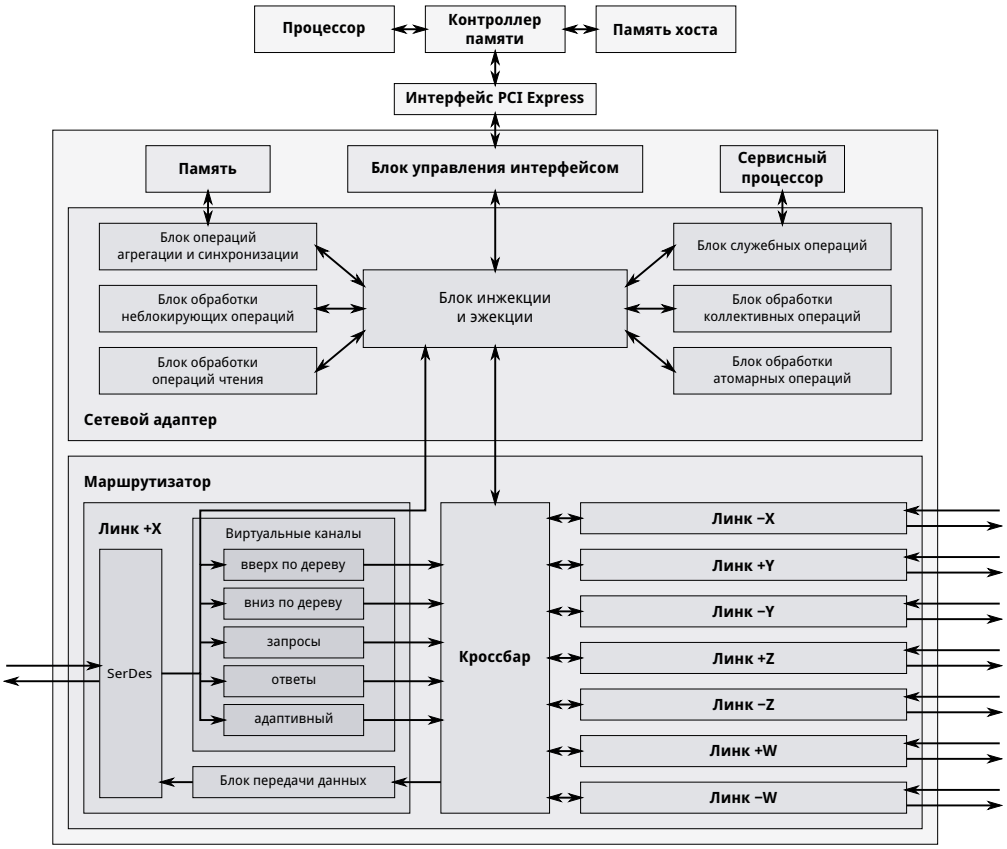

Рис. 3. Структура вычислительного узла с сетевым адаптером/маршрутизатором «Ангара».

и записываются адаптером коммуникационной сети посредством операций DMA) и регион с управляющими регистрами. Доступ к тем или иным ресурсам маршрутизатора контролируется ядерным модулем.

Для достижения большей эффективности было принято решение, что на одном узле должна выполняться только одна вычислительная задача, это позволило 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, а также упростить модель безопасности сети, исключив из нее обеспечение безопасности процессов различных задач на одном узле. Данное решение не повлияло на функциональность сети, как предназначающуюся в первую очередь для задач большого размера (в противовес InfiniBand, универсальной сети для задач различного размера). Аналогичное решение было принято в IBM Blue Gene, где ограничение на единственность задачи вводится для раздела.

На аппаратном уровне поддерживается одновременная работа с маршрутизатором многих потоков/процессов одной задачи — она реализована в виде нескольких инжекционных каналов, доступных для использования процессам посредством нескольких кольцевых буферов для записи пакетов. Количество и размер этих буферов могут изменяться динамически.

Основной режим программирования для сети «Ангара» — совместное использование MPI, OpenMP и Shmem, а также GASNet и UPC.

После завершения верификации и макетирования сети планируется выпустить кристалл СБИС. Прототипная партия СБИС будет предназначена для отладки основных технологических решений, технологического процесса и экспериментальной проверки результатов моделирования. Прототип будет содержать всю базовую функциональность, работать с интерфейсом PCI Express gen2 x16 и линками с пропускной способностью 75 Гбит/с.

Продвижение сети «Ангара» на рынок планируется осуществлять в двух вариантах: как отдельную коммерческую сеть в виде плат PCI Express для кластерных систем со стандартными процессорами и чипсетам, и в составе разрабатываемой в НИЦЭВТ четырехсокетной лезвийной системы на базе процессоров AMD.

Список литературы

[1] А. А. Корж, Д. В. Макагон, И. А. Жабин, Е. Л. Сыромятников и др. *«Отечественная коммуникационная сеть 3D-тор с поддержкой глобально адресуемой памяти для суперкомпьютеров транспетафлопсного уровня производительности»*, Параллельные вычислительные технологии (ПаВТ'2010): Труды международной научной конференции (Уфа, 29 марта — 2 апреля 2010 г.), `http://omega.sp.susu.ac.ru/books/conference/PaVT2010/full/134.pdf` — Челябинск: Издательский центр ЮУрГУ, 2010. — с. 227—237.

[2] William Dally and Brian Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[3] Jose Duato, Sudhakar Yalamanchili, and Ni Lionel. *Interconnection Networks: An Engineering Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

Дмитрий Макагон, Евгений Сыромятников ({makagond, syromiatnikov}@nicevt.ru) — сотрудники ОАО «НИЦЭВТ» (Москва).

Постоянный URL статьи: `http://www.osp.ru/os/2011/07/13010500/`
PDF-версия статьи: `http://dislab.org/docs/os-interconnects.pdf`