УДК 004.725.5

«Разработка межузловой коммуникационной сети ЕС8430 «Ангара» для перспективных российских суперкомпьютеров».

«Development of ES8430 “Angara” interconnect for future Russian supercomputers».

А.И.Слуцкин, А.С.Симонов, И.Жабин, Д.Макагон, Е.Сыромятников
A.Slutskin, A.Simonov, I.Zhabin, D.Makagon, E. Syromyatnikov

Аннотация

В статье рассматриваются основные решения, заложенные в основу проекта отечественной высокоскоростной коммуникационной сети ЕС8430 «Ангара», предназначенной для объединения узлов массово-параллельных суперкомпьютеров.

Abstracts

This paper describes the basic decisions underlying the development of ES8430 “Angara” high speed interconnect for massively parallel computers.

Ключевые слова: суперкомпьютеры, многопроцессорные вычислительные системы, коммуникационная сеть.

Keywords: supercomputers, multiprocessor computer systems, interconnect.

Области применения современных суперкомпьютеров постоянно расширяются. Их используют для решения вычислительно сложных задач в физике, химии и биологии, при подготовке прогнозов изменения климата и ситуационном моделировании, при осуществлении виртуального прототипирования и при проведении виртуальных испытаний. При этом одни и те же задачи на разных суперкомпьютерах могут решаться с различной эффективностью, которая во многом определяется характеристиками коммуникационной сети, используемой для объединения узлов суперкомпьютера.

В соответствии с сетевым законом Амдала, модификацией основного закона Амдала, которая отражает потери времени на межпроцессорный обмен сообщениями, ускорение при решении задач на многопроцессорных вычислительных системах (МВС) обратно пропорционально коэффициенту сетевой деградации, значение которого является произведением технической и алгоритмической составляющих:

$$R = \frac{1}{a + \frac{1-a}{n} + c},$$

где: a − удельный вес скалярных операций;
n – число вычислительных узлов;
c – коэффициент сетевой деградации.

$$C = C_A * C_T ,$$

где: c_A − алгоритмическая составляющая;
c_T – техническая составляющая.

Значение технической составляющей зависит от соотношения реальной производительности процессора и характеристик аппаратуры сети. Значение алгоритмической составляющей обусловлено свойствами алгоритма, в том числе особенностями его реализации. Таким образом, для повышения скорости вычислений следует воздействовать на обе составляющие коэффициента деградации.

Значения алгоритмической и технической составляющих коэффициента деградации во многом определяют масштабируемость суперкомпьютера, которая определяется тем, насколько обеспечивается рост производительности при увеличении числа вычислительных узлов, задействованных для решения практической задачи. В идеальном случае производительность должна расти по логарифмическому закону обратно пропорцио-

нально удельному весу скалярных операций. В реальной жизни из-за накладных расходов на взаимодействие вычислительных узлов производительность имеет предел масштабируемости, т.е. после достижения определённого, специфичного для каждой конкретной задачи, числа узлов производительность суперкомпьютера начинает снижаться. В результате при решении многих актуальных задач, например, с использованием графов или сильно разреженных матриц, современные суперкомпьютеры показывают крайне низкую реальную производительность, часто менее 1% от пиковой.

За рубежом вопросу разработки суперкомпьютеров с использованием заказных суперкомпьютерных технологий и, в том числе, высокоэффективных коммуникационных сетей, уделяется особое внимание. В США под руководством государственных органов были приняты более десятка программ, таких как DARPA HPCS, UHPC, NSF PetaApps, NNSA ASC, DOE ASCR, которые позволяют участвующим в них национальным лабораториям, институтам, университетам и частным компаниям активно создавать новые и развивать существующие суперкомпьютерных технологии. Аналогичные программы есть в Европе (PRACE, EESI, DataGrid), Китае (Программа 863), Японии и Индии.

Результаты разработок можно проследить по рейтингам суперкомпьютеров мира. При этом наивысшие показатели производительности показывают суперкомпьютеры, созданные с использованием заказных коммуникационных сетей, таких как сеть Tofu суперкомпьютера Fujitsu K-Computer [1], сети SeaStar/Gemini суперкомпьютеров Cray серий XT/XE [2], сеть Arch суперкомпьютера Tianhe-1A [3], коммуникационные сети суперкомпьютеров IBM Blue Gene/P, /Q [4].

К сожалению, заказные коммуникационные сети недоступны на рынке. В этой связи цель работы состоит в создании коммуникационной сети, обеспечивающей, во-первых, низкое значение технической составляющей коэффициента деградации за счёт существенно лучших основных характеристик, латентности и пропускной способности, по сравнению с коммерчески доступными сетями Ethernet, Infiniband, и приближенными к значениям, характерным для заказных разработок коммуникационных сетей зарубежных суперкомпьютеров, и во-вторых, снижение значения алгоритмической составляющей за счёт обеспечения возможности использования комбинированной модели программирования SHMEM+MPI, а также расширенной аппаратной поддержки основных операций библиотек MPI и SHMEM.

При разработке концепции коммуникационной сети был принят ряд решений относительно модели исполнения прикладных задач на вычислительной системе. Одно из наиболее важных состоит в том, что на вычислительном узле исполняются процессы только одной прикладной задачи. Это позволяет, во-первых, сократить накладные расходы на переключение между процессами, а во-вторых, необходимо поддерживать всего два виртуальных адресных пространства — решаемой задачи и операционной системы. Последняя необходима, помимо своих основных функций, для реализации параллельной файловой системы, функционирующей поверх коммуникационной сети и обеспечивающей возможность выполнения не только стандартных операций ввода/вывода, но и реализации контрольных точек.

Для объединения узлов МВС используется топология «4D-тор» (рис.1). При этом в каждый вычислительный узел или узел ввода/вывода МВС устанавливается маршрутизатор и специализированное программное обеспечение, обеспечивающие поддержку распределённой глобально адресуемой памяти.

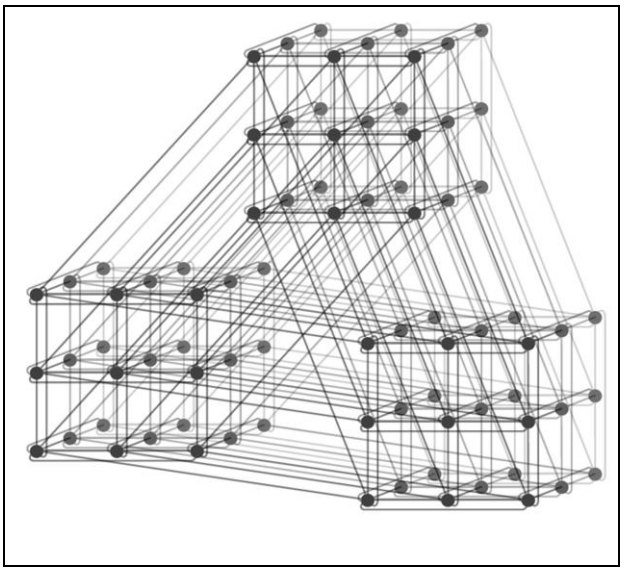

Рис. 1. Топология 4D-тор (3x3x3x3)

Выбор топологии «многомерный тор» сделан исходя из особенностей шаблонов обмена сообщениями по сети на интересующем классе задач. По сравнению с топологией Fat tree, используемой в сети Infiniband, данная топология значительно более толерантна к неравномерной загрузке, что подтверждается не только результатами имитационного моделирования, но и тестовыми расчётами на существующих суперкомпьютерах, например IBM Blue Gene/P, который имеет основную сеть с топологией 3D-тор. Кроме того, обеспечивается хорошая масштабируемость производительности МВС при решении сильно связанных задач.

Работы по разработке данной коммуникационной сети ведутся в ОАО «НИЦЭВТ» с 2006 года [5,6]. На подготовительном этапе было проведено тщательное изучение результатов зарубежных исследований и разработок, в том числе работ Уилльяма Дэйли [7] и Хосе Дуато [8], а также архитектур IBM Blue Gene и Cray SeaStar.

За это время было создано три поколения макетных образцов маршрутизаторов на FPGA (рис.2) и вычислительные кластеры на их основе (рис.3), отработаны решения по передаче сообщений и взаимодействию с коммерчески доступными процессорами вычислительных узлов с использованием современного интерфейса PCI Express. В настоящее время выполняется разработка заказной СБИС маршрутизатора на технологических нормах 65 нм, прототипы планируется изготовить на фабрике TSMC в первой половине 2012 г.

В маршрутизаторе реализованы следующие функциональные возможности:

- Протокол надёжной доставки пакетов.
- Детерминированная и адаптивная передача данных.
- Аппаратная поддержка многопоточности.
- Аппаратная реализация операций: запись в память удалённого узла, запись в память удалённого узла с сохранением концептуальности, атомарные операции в памяти удалённого узла, чтение из памяти удалённого узла, неблокирующая запись больших массивов в память удалённого узла, чтение больших массивов из памяти удалённого узла.
- Аппаратная поддержка операций барьерной синхронизации.
- Реализация сборки массивов в памяти маршрутизатора с последующим копированием в память узла.
- Сеть коллективных операций (операции broadcast, reduce).
- Система обеспечения отказоустойчивости.

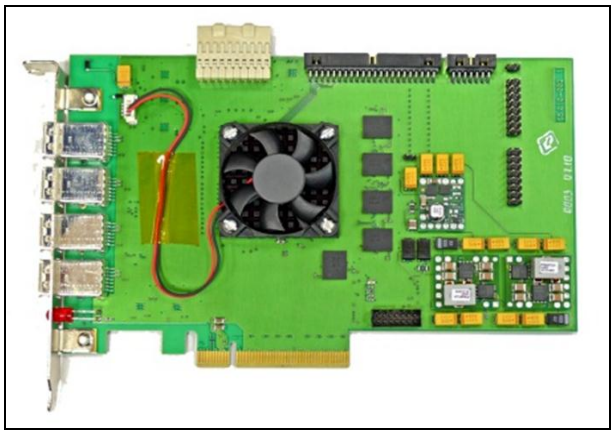

Рис. 2. Макетный образец маршрутизатора ЕС8430 «Ангара»

Рис. 3. Высислительный кластер ЕС1720.05, оснащенный макетными образцами машрутизаторов межузловой коммуникационной сети ЕС8430 «Ангара», представленный на ChipExpo-2010 (Москва)

Основным рекомендуемым режимом работы разрабатываемой коммуникационной сети, на котором обеспечиваются наилучшие показатели производительности МВС, является режим с использованием библиотеки SHMEM. При этом общий объём глобально адресуемой маршрутизатором памяти МВС составляет до 128 ПБайт, а объём памяти, адресуемой на каждом узле, — до 2 ТБайт.

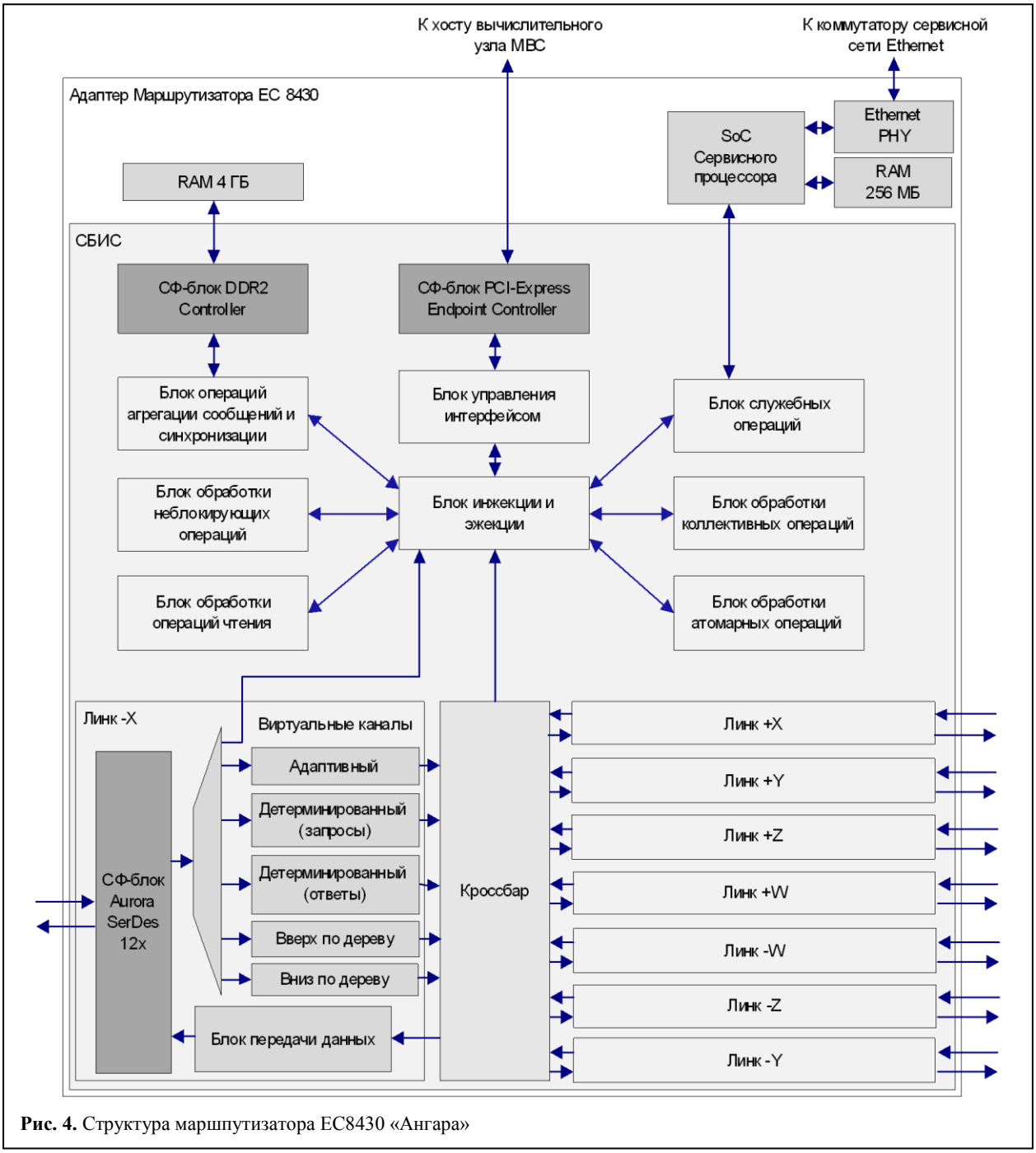

Рис. 4. Структура маршпутизатора EC8430 «Ангара»

Поддержка глобально адресуемой памяти работает следующим образом. На каждом вычислительном узле выделяется непрерывный регион физической памяти, адрес и размер которого сообщается маршрутизатору. Общие данные задача должна хранить в выделенном регионе, обращение к нему должно осуществляться с помощью функций библиотеки SHMEM. Получив данные, маршрутизатор записывает их в выделенную физическую память с использованием механизма трансляции. Таким образом, полученные данные становятся сразу видны решаемой задаче.

Структура маршрутизатора приведена на рис.4. Маршрутизатор поддерживает многопоточность в рамках одной задачи. При этом каждый поток использует для передачи сообщений собственную область кольцевого буфера передачи сообщений в маршрутизаторе, что позволило обеспечить гарантированную целостность транзакций и отсутствие дополнительных накладных расходов на синхронизацию потоков.

Также поддержан механизм прямого формирования пакетов маршрутизатором на основе команд обращения к памяти вычислительного уз-

ла. Этот механизм позволяет существенно снизить накладные расходы со стороны процессора, предпочтителен при выполнении операций пересылки единичных слов в память удалённого узла, однако обладает ограничением на количество одновременно выданных операций чтения из памяти удалённого узла из-за небольшого размера буфера отложенных операций процессора.

В каждом маршрутизаторе имеется встроенная оперативная память объёмом 4 ГБайта, которая используется для агрегирования пакетов, поступивших от других узлов, и пересылки полученных таким образом массивов данных в память вычислительного узла с использованием механизма прямого доступа в память. Таким образом, удаётся существенно снизить накладные расходы при пересылках по интерфейсу PCI Express.

Маршрутизатор имеет восемь линков, используемых для соединения с соседними узлами МВС. Линки работают в дуплексном режиме, при этом информация в каждую сторону одновременно передаётся по 12 дифференциальным парам с использованием встроенных в СБИС блоков Aurora SerDes с пропускной способностью 6,25 Гбит/с каждый. Таким образом, односторонняя пропускная способность каждого линка составляет 75 Гбит/с, а агрегатная пропускная способность всех линков маршрутизатора — 1,2 Тбит/с. В качестве кабеля для соединения узлов на расстояниях до 1,5 м может использоваться витая пара, а на расстояния до 300 м — ВОЛС.

Каждый линк содержит пять виртуальных каналов, образующих соответствующие подсети:

- подсеть для адаптивной передачи пакетов. Это основная подсеть для организации обмена данными между узлами;
- подсетей запросов и ответов с детерминированной передачей пакетов, т. е. с гарантированным сохранением порядка следования пакетов;
- подсеть коллективных операций для операций broadcast и reduce.

Передача пакетов на следующий узел осуществляется с учетом кредитной информации о наличии свободного места в буферах соответствующего виртуального канала на приёмной стороне. Блоки маршрутизатора, осуществляющие передачу пакетов на следующий узел, содержат специальный буфер повторной передачи. Флиты передаваемого пакета передаются и одновременно записываются в данный буфер. В случае, если на приёмной стороне будет зафиксирована ошибка, пакет будет передан повторно. В качестве интерфейса с процессором узла МВС в маршрутизаторе используется PCI Express 2.0.

Системное ПО маршрутизатора включает реализацию низкоуровневого API и поверх него полнофункциональные реализации MPI 2.0 и Cray SHMEM. Кроме того, разрабатывается адаптированная для работы с данной сетью версия параллельной файловой системы Lustre, обеспечивающая возможность включения узлов ввода/вывода в единую структуру многомерного тора вместе с вычислительными узлами для выполнения загрузки в вычислительные узлы исходных данных для расчётов, выгрузки результатов и реализации механизма контрольной точки. Также для разработки прикладных программ поддерживаются стандартные математические библиотеки, такие как BLAS, LAPACK, SCALAPACK , FFTW и др.

В планах разработка специализированной операционной системы для вычислительных узлов на базе облегчённой версии ядра Linux, необходимой для минимизации накладных расходов при решении вычислительных задач.

Результат данной работы, СБИС маршрутизатора коммуникационной сети с топологией «многомерный тор», планируется использовать в составе разрабатываемой в рамках проекта «Ангара» вычислительной платформы, что позволит объединить до 64 тысяч узлов в составе суперкомпьютера транспетафлопсного уровня производительности.

Кроме того, планируется производство на её основе линейки коммерческих продуктов — коммуникационных адаптеров в форм-факторах PCI Express для использования в составе вычислительных кластеров, и PC/104-Express для применения в составе бортовых комплексов и средств промышленной автоматизации.

Литература

1. Yuichiro Ajima, Shinji Sumimoto, Toshiyuki Shimizu, *Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers*, Computer, vol. 42, no. 11, pp. 36-40, Nov. 2009
2. Robert Alverson, Duncan Roweth, and Larry Kaplan, *The Gemini System Interconnect*, In Proceedings of the 2010 18th IEEE Symposium on High Performance Interconnects (HOTI '10), IEEE Computer Society, Washington, DC, USA, 83-87.
3. Xue-Jun Yang et al. *The TianHe-1A Supercomputer: Its Hardware and Software*, Journal of Computer Science and Technology, 2011,V26(3): 344-351.
4. N. R. Adiga, M. A. Blumrich, D. Chen, P. Coteus, A. Gara, M. E. Giampapa, P. Heidelberger, S. Singh, B. D. Steinmacher-Burow, T. Takken, M. Tsao, P. Vranas. *Blue Gene/L torus interconnection network*, IBM J. RES. & DEV. VOL. 49 NO. 2/3 MARCH/MAY 2005.
5. А.И.Слуцкин, А.С.Симонов. *«Развитие суперкомпьютерных технологий в ОАО «НИЦЭВТ»*, Научно-техническая конференция «Перспективные направления развития средств вычислительной техники»: Сборник тезисов докладов (г.Москва, 28 июня 2011 г.).
6. А.С. Симонов, И. Жабин, Д.Макагон. «Разработка межузловой коммуникационной сети с топологией «многомерный тор» и поддержкой глобально адресуемой памяти для перспективных отечественных суперкомпьютеров», Научно-техническая конференция «Перспективные направления развития средств вычислительной техники»: Сборник тезисов докладов (г.Москва, 28 июня 2011 г.).
7. William Dally and Brian Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.
8. Jose Duato, Sudhakar Yalamanchili, and Ni Lionel. *Interconnection Networks: An Engineering Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.