ВЫСОКОСКОРОСТНАЯ СЕТЬ АНГАРА: АРХИТЕКТУРА И РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ

Симонов А.С.[1], Жабин И.А.[2], Куштанов Е.Р.[3], Макагон Д.В.[4], Семенов А.С.[5], Щербак А.Н.[6]

Цель статьи: *представление архитектуры, вариантов аппаратного исполнения, принципов организации системного программного обеспечения и достигнутых результатов применения высокоскоростной сети Ангара.*

Полученный результат: *В статье представлена архитектура высокоскоростной сети Ангара с топологией 4D-тор, описана ее реализация в СБИС и вариантов исполнения сетевого оборудования. Представлены результаты исследования производительности сети Ангара на двух вычислительных системах: «Ангара-К1» в АО «НИЦЭВТ» и Desmos в ОИВТ РАН. На различных тестах и приложениях сеть Ангара позволяет достичь высокого уровня производительности и масштабируемости, которые не уступают или опережают соответствующие характеристики на вычислительных системах с использованием сети Mellanox Infiniband 4xFDR.*

Ключевые слова: высокоскоростная сеть Ангара, суперкомпьютер, многомерный тор, Infiniband, библиотека MPI

DOI: *10.21681/2311-3456-2018-4-46-53*

1. Введение

Многопроцессорные вычислительные системы имеют важное значение для решения прикладных задач, направленных на повышение научно-технического потенциала экономики и укрепление обороноспособности страны. В первую очередь это задачи компьютерного моделирования, обработки больших объёмов данных, искусственного интеллекта, планирования и прогнозирования (далее целевой класс задач).

Для развития технологий компьютерного моделирования, обработки больших объёмов данных, искусственного интеллекта в промышленно развитых странах активно создаются суперкомпьютеры и вычислительные кластеры. Так, в США на сегодня сформирован пул суперкомпьютеров и вычислительных кластеров суммарной пиковой производительностью 850 Пфлопс, в Китае – 885 Пфлопс, более половины самых мощных суперкомпьютеров мира, по данным списка TOP500 (на июнь 2019 года), используется в промышленности для решения задач компьютерного моделирования. Суммарная производительность российских суперкомпьютеров составляет всего 20 Пфлопс, из них в промышленности используется лишь 8%.

Передовые зарубежные суперкомпьютеры, как правило, представляют собой заказные разработки, наиболее известными из них являются серия BlueGene компании IBM, серии XK, XC компании Cray, K Computer компании Fujitsu. В отличие от суперкомпьютеров и вычислительных кластеров, собранных из коммерчески доступных комплектующих, заказные суперкомпьютеры, во многом за счёт передовых технических решений в области создания коммуникационной среды – высокоскоростных коммуникационных сетей межпроцессорного обмена, обеспечивают значительно более высокую масштабируемость производительности при решении прикладных задач. В связи с этим крайне актуальным является вопрос разработки отечественной высокоскоростной сети, сравнимой с иностранными заказными аналогами.

В России имеется ряд проектов по созданию высокоскоростных коммуникационных сетей. Так, в ФГБУ «ИПМ им. М.В. Келдыша РАН» и ФГУП «НИИ Квант» разработана сеть МВС-Экспресс [1], представляющая собой коммуникационную сеть с отечественным стеком программного обеспечения, созданная на основе зарубежных коммутаторов PCI Express компании PLX Technology. Во ФГУП «РФЯЦ-ВНИИЭФ» разработана система межпроцессорного обмена СМПО-10G [2], основу которой составляет заказная СБИС. Информации об этой разработке в открытой печати очень мало. В ФГБУ

1 Симонов Алексей Сергеевич, кандидат технических наук, первый заместитель генерального директора АО «НИЦЭВТ», г. Москва, Россия. E-mail: simonov@nicevt.ru

2 Жабин Иван Алексеевич, заместитель начальника отдела, АО «НИЦЭВТ», г. Москва, Россия. E-mail: zhabin@nicevt.ru

3 Куштанов Евгений Рустамович, начальник управления, АО «НИЦЭВТ», г. Москва, Россия. E-mail: evgeny.kushtanov@nicevt.ru

4 Макагон Дмитрий Викторович, начальник отдела, АО «НИЦЭВТ», г. Москва, Россия. E-mail: makagond@nicevt.ru

5 Семенов Александр Сергеевич, начальник отдела, АО «НИЦЭВТ», г. Москва, Россия. E-mail: semenov@nicevt.ru

6 Щербак Андрей Николаевич, начальник отдела, АО «НИЦЭВТ», г. Москва, Россия. E-mail: andrey.shcherbak@nicevt.ru

«ИПС им. А. К. Айламазяна РАН» разработаны основанные на ПЛИС коммуникационные сети СКИФ-Аврора [3] и Паутина [4].

Среди заказных коммуникационных сетей зарубежных суперкомпьютеров наибольший интерес представляют Cray Gemini [5] с топологией 3D-тор и Cray Aries [6] с топологией Dragonfly, применяемые в составе линеек суперкомпьютеров Cray, сети суперкомпьютеров IBM BlueGene/Q с топологией 5D-тор [7], сети Tofu [8], Tofu2 [9], TofuD [10] суперкомпьютеров компании Fujitsu с топологией 6D-тор, европейская сеть Extoll с топологией 3D-тор [11], китайская сеть Sugon с топологией 3D-тор [12]. Перечисленные сети во многом и обеспечивают более высокую масштабируемость производительности зарубежных заказных суперкомпьютеров по сравнению с суперкомпьютерами, построенными на основе коммерчески доступных коммуникационных сетей.

В данной статье представлены результаты многолетней работы коллектива АО «НИЦЭВТ» по разработке российской высокоскоростной коммуникационной сети Ангара. В статье описана архитектура сети Ангара, программное обеспечение, а также основные результаты достигнутой производительности.

2. Архитектура сети Ангара

Высокоскоростная сеть Ангара поддерживает топологию 4D-тор. Топология 4D-тор не подразумевает использование коммутаторов, в каждом узле имеется сетевой адаптер, соединённый с одним маршрутизатором, который, в свою очередь, соединяется высокоскоростными каналами связи с соседними маршрутизаторами. В сети Ангара адаптер и маршрутизатор реализованы в рамках одной СБИС.

В маршрутизаторе реализована бездедлоковая детерминированная и адаптивная маршрутизации, основанные на правилах «пузырька» (англ. Bubble flow control, [13]) и порядка направлений +X, +Y, +Z, +W, −X, −Y, −Z, −W (англ. Direction ordered routing, DOR, [14; 15]) с использованием битов направлений [14]. Метод маршрутизации First Step/Last Step [14] «нестандартного первого и последнего шага» позволяет ослабить требования правила порядка направлений, благодаря этому методу усилены возможности по обходу отказавших узлов и линков.

В каждом направлении имеется пять виртуальных каналов: два канала для детерминированной маршрутизации − канал запросов и канал ответов; отдельный виртуальный канал используется для адаптивной маршрутизации с возможностью перехода в детерминированный канал в случае потенциального дедлока; ещё два виртуальных канала используются для передачи сообщений по виртуальной подсети для коллективных операций.

Детерминированная маршрутизация сохраняет порядок передачи пакетов и предотвращает появление дедлоков (англ. deadlock); адаптивная маршрутизация использует для доставки пакетов один из возможных минимальных маршрутов, игнорируя порядок направлений, что позволяет обходить перегруженные и вышедшие из строя участки сети. Поддержка коллективных операций — широковещательной рассылки и редукции — реализована с помощью виртуальной подсети, имеющей топологию дерева, наложенного на многомерный тор [16].

Адаптер сети на аппаратном уровне поддерживает удаленные операции (RDMA) записи, чтения и атомарные операции. Доступны атомарные операции двух типов — сложение и исключающее ИЛИ.

На канальном уровне поддерживается отказоустойчивая передача пакетов с помощью нумерации пакетов, подсчёта для каждого контрольных сумм и повторной передачи в случае, если контрольная сумма, записанная в последнем флите пакета, не совпадает с вычисленной после передачи. Существует также механизм обхода отказавших каналов связи и узлов с помощью перестройки таблиц маршрутов и использования нестандартных первого/последнего шагов маршрута пакета. Для выполнения различных сервисных операций, включая настройку/перестройку таблиц маршрутов, и выполнения некоторых расчетов может использоваться сервисный процессор, взаимодействующий с адаптером по интерфейсу ELB.

Структурная схема СБИС ЕС8430 сети Ангара изображена на рис. 1.

СБИС ЕС8430 состоит из следующих основных блоков:

- 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;
- кроссбар, 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 обработке.

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 независимых кольцевых буферов для записи пакетов.

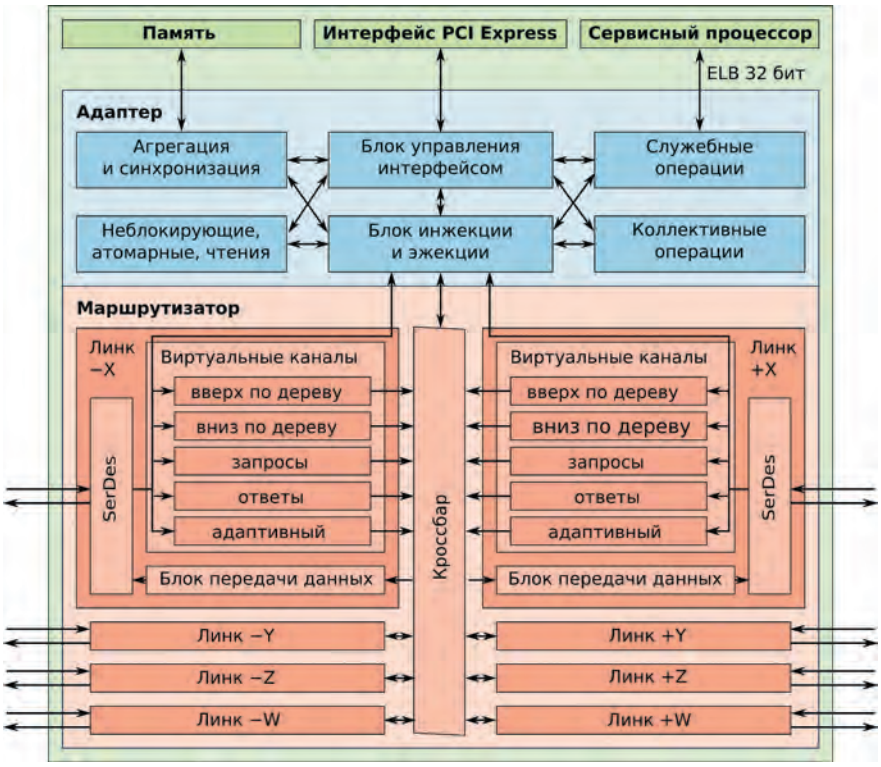

Рис. 1. Схема СБИС ЕС8430 сети Ангара

2.1 Программное обеспечение

Архитектура программного обеспечения сети Ангара представлена на рис. 2. Взаимодействие вычислительного узла, т. е. кода, исполняемого на центральном процессоре, с маршрутизатором осуществляется путем записи данных по адресам памяти, 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 с временем отправки пакета. Для отправки пакетов используется один из регионов памяти маршрутизатора, рассматриваемый как кольцевой буфер. Также имеется отдельный регион для выполнения DMA-операций, когда данные читаются из памяти и записываются в удалённую память напрямую адаптером коммуникационной сети без участия процессора вычислительного узла. Ещё один регион содержит управляющие регистры, обеспечивающие конфигурирование адаптера и получение информации для осуществления мониторинга, диагностики и отладки. Доступ к тем или иным ресурсам маршрутизатора контролируется модулем ядра ОС.

Основным режимом программирования для сети Ангара является совместное использование MPI и OpenMP. Наибольшей эффективности выполнения коммуникаций на прикладном уровне можно достичь с использованием библиотеки SHMEM.

В настоящее время разработаны следующие реализации библиотеки MPI: MPICH 3.0.4 (стандартная и оптимизированная), MPICH 3.2, OpenMPI 1.10. Также ведутся работы над поддержкой библиотеки libfabric программной платформы Open Fabric Interface [17], которая разработана и продвигается компанией Intel для упрощения интерфейса между приложениями и системным программным обеспечением и высокоскоростными средствами. Реализация libfabric для сети Ангара позволит обеспечить поддержку на сети Ангара функционирования библиотек Intel MPI, MPICH 3.3, Open MPI, GasNET, а также более эффективную поддержку технологии программирования Charm++ и языков программирования Clang, UPC.

Программное обеспечение сети Ангара протестировано с операционными системами OpenSUSE/SLES 11, 12 и 15, CentOS 6.0-7.5, ALT Linux Server 8, Scientific Linux 7.5, Astra Linux SE 1.3-1.6, ЗОСРВ «Нейтрино» (QNX 6.5), ОС «Эльбрус», поддерживаются версии ядра Linux от 2.6.21 до 4.15.3.

При эксплуатации суперкомпьютеров в условиях наличия отказов (каналов связи или узлов) и занятых другими задачами вычислительных узлов необходимо предоставлять возможность выделения для задачи пользователя связного множества узлов, отвечающего запрашиваемым вычислительным мощностям, а также задавать для этих множеств таблицы маршрутов. При этом для сетей с топологией «многомерный тор» необходимо учитывать специфические требования маршрутизации, равномерности распределения сетевого трафика, минимизации фрагментации и числа возможных транзитных узлов. В силу взаимоисключающего характера ряда требований важен выбор разумного компромисса.

Для решения данных задач разработан набор сервисных утилит ANSU (англ. Angara Node Selection Utility), которые позволяют определять связность произвольно-

го множества вычислительных узлов и создавать таблицы маршрутизации для приближенного решения задачи равномерного распределения трафика внутри заданного множества [18], находить множества вычислительных узлов с учетом фрагментации вычислительной системы с максимизацией утилизации вычислительной системы при потоке задач [19]. Исследование характеристик эффективности использования ANSU на реальной вычислительной системе и наборе синтетических систем в среднем показало улучшение утилизации вычислительных систем на 7% и понижение среднего времени ожидания задания в очереди на 36% по сравнению с базовым вариантом программного обеспечения [20].

2.2 Варианты исполнения

СБИС маршрутизатора сети Ангара ЕС8430 изготовлена на фабрике TSMC с использованием технологического процесса 65 нм и содержит приблизительно 180 млн транзисторов, имеет размеры 13,0×10,5 мм; корпусировка FCBGA, 1521 вывод в виде массива 39×39 контактов с шагом 1 мм, подложка имеет размеры 40×40 мм. СБИС в максимально нагруженном режиме потребляет 20 Вт энергии.

Плата сетевого адаптера изготавливается на собственном производстве в АО «НИЦЭВТ».

Для использования сети Ангара в составе вычислительных комплексов имеется полноформатный сетевой адаптер ЕС8431, сетевые адаптеры при этом объединяются напрямую. Для удобства пользователей и обеспечения возможности применения сети Ангара в вычислительных комплексах с требованиями по компактности используемых серверов разработан вариант сетевого оборудования на основе 24-портового коммутатора ЕС8433 и низкопрофильного адаптера ЕС8432, для коммутации возможно использовать оптические и медные кабели с коннекторами CXP.

Рис. 2. Программное обеспечение сети Ангара

3. Достигнутые результаты

В данной работе приводятся результаты исследования двух вычислительных систем, построенных с использованием сети Ангара: «Ангара-К1» в АО «НИЦЭВТ» и Desmos в ОИВТ РАН. Вычислительная система Desmos [21] состоит из 32 узлов, включающих процессоры Intel 1650v3 и ускорители компании AMD FirePro S9150. Память каждого узла – 16 ГБ. Топология сети Ангара 4x2x2x2.

Вычислительный кластер «Ангара-К1» состоит из 36 узлов, объединенных сетью Ангара. Кластер состоит из двух типов узлов с процессорами Intel Xeon E5-2630 и Intel Xeon E5-2660. Память каждого узла – 64 ГБ. Узлы объединены сетью Ангара с топологией 3D-тор 3x3x4. В исследовании использовалась библиотека MPI (MPICH 3.0.4).

На рис. 3 представлена задержка передачи сообщений при помощи теста OSU Micro-Benchmarks на суперкомпьютере Desmos. Архитектура маршрутизатора сети Ангара оптимизирована для суперкомпьютерных приложений, задержка на один хоп передачи, включающий передачу через маршрутизатор и канал связи, составляет 129 нс. С учетом этого и высокой частоты процессора на суперкомпьютере Desmos задержка передачи сообщения длиной 16 байт составляет 0.85 мкс с использованием библиотеки MPI и 0.7 мкс с использованием библиотеки SHMEM.

Сравнительное оценочное тестирование проводилось на кластере «Ангара-К1» и на 36 узлах суперкомпьютера МВС-10П, установленном в МСЦ РАН и включающем по сравнению с «Ангара-К1» процессоры Intel Xeon E5-2690 того же поколения SandyBridge, но более производительные. Узлы МВС-10П объединены сетью Mellanox Infiniband 4xFDR. Сравнение результатов тестирования на синтетических коммуникационных тестах, тестах NAS Parallel Benchmarks и модели прогноза погоды ПЛАВ показало, что сеть Ангара не уступает по производительности сети Mellanox Infiniband 4xFDR, а в ряде случаев превосходит ее [22, 23]. В частности, чем большая доля коммуникаций в тесте или приложении, тем больше становится роль сети. На более слабых процессорах кластера «Ангара-К1» на 32 узлах удалось получить в абсолютных показателях больший результат по сравнению с МВС-10П на тесте сортировки целых чисел (NPB Integer Sort).

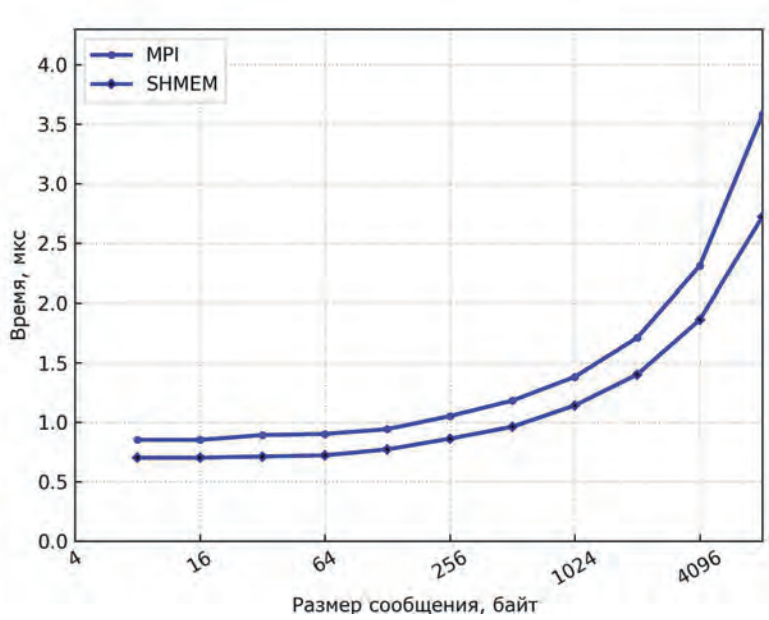

Рис. 3. Задержка передачи сообщения на суперкомпьютере Desmos.

На кластере «Ангара-К1» проводилось исследование производительности задачи газодинамики программного комплекса FlowVision, при этом рассматривалось несколько тестовых задач, имеющих 260 тысяч, 5.5 млн и 26.8 млн расчетных ячеек. Сравнение полученных результатов с двумя кластерами на основе сети Mellanox Infiniband 4xFDR показало, что сеть Ангара позволяет получить масштабируемость в ряде случаев лучше, чем Infiniband 4xFDR [24, 25].

Оценочное тестирование на суперкомпьютере Desmos на программном пакете VASP для расчетов методами квантовой молекулярной динамики проводилось также в сравнении с суперкомпьютером Fisher на основе сети Mellanox Infiniband 4xFDR, оснащенных, более мощными процессорами AMD EPYC 7301. Сравнение показало превосходство масштабируемости суперкомпьютера Desmos по сравнению с Fisher [26].

Также подтверждена совместимость и работоспособность сети Ангара с программными продуктами ANSYS 18.2 (Mechanical, Fluent, LS-DYNA, CFX), OpenFOAM, Gromacs, CP2K, LAMMPS, GAMESS и другими программными пакетами и параллельными программами, при помощи которых различными научными группами и коллективами ведутся научные расчеты и научная работа в области системного программного обеспечения [27-29].

4. Заключение

В данной статье представлена архитектура высокоскоростной сети Ангара с топологией 4D-тор. В работе описаны используемые алгоритмы маршрутизации, методы обеспечения отказоустойчивости, организация низкоуровневого и сервисного системного программного обеспечения, варианты поддерживаемых библиотек передачи сообщений MPI, направления развития системного программного обеспечения. В статье представлены результаты исследования производительности сети Ангара на двух вычислительных системах: «Ангара-К1» в АО «НИЦЭВТ» и Desmos в ОИВТ РАН. На различных тестах и приложениях сеть Ангара позволяет достичь высокого уровня производительности и масштабируемости, которые не уступают или опережают соответствующие характеристики на вычислительных системах с использованием сети Mellanox Infiniband 4xFDR.

Литература

1. Левин В.К. и др. Коммуникационная сеть МВС-Экспресс // Информационные технологии и вычислительные системы. 2014. №.1. С. 10–24.
2. Басалов В.Г., Вялухин В. М. 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.
3. Адамович И.А. и др. Опыт разработки коммуникационной сети суперкомпьютера «СКИФ-Аврора» // Программные системы: теория и приложения. 2010. Т. 1. №. 3. С. 107-123.
4. Климов Ю.А. и др. Паутина: высокоскоростная коммуникационная сеть // Программные системы: теория и приложения. 2015. Т. 6. №. 1 (24). С. 109–120.

5. Alverson R., Roweth D., Kaplan L. The Gemini System Interconnect // 2010 18th IEEE Symposium on High Performance Interconnects. – IEEE. 2010. – P. 83-87.
6. Faanes G. et al. Cray Cascade: a Scalable HPC System Based on a Dragonfly Network // Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. – IEEE Computer Society Press. 2012. – 9 p.
7. Haring R. et al. The IBM Blue Gene/Q Compute Chip // IEEE Micro. 2011. –Vol. 32, no. 2. P. 48–60.
8. Ajima Y., Sumimoto S., Shimizu T. Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers. // IEEE Computer. 2009. Vol. 42, no. 11. P. 36—40. URL: https://doi.org/10.1109/MC.2009.370.
9. Tofu Interconnect 2: System-on-Chip Integration of High-Performance Interconnect / Y. Ajima, T. Inoue, S. Hiramoto et al. // International Supercomputing Conference / Springer. 2014. P. 498—507. URL: https://doi.org/10.1007/978-3-319-07518-1_35.
10. Ajima Y. et al. The Tofu Interconnect D // 2018 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, 2018. P. 646—654.
11. Scalable communication architecture for network-attached accelerators / S. Neuwirth, D. Frey, M. Nuessle, U. Bruening // High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium / IEEE, 2015. P. 627—638. URL: https://doi.org/10.1109/HPCA.2015.7056068.
12. Silicon Cube: A Supercomputer Specially Designed for Meteorological Applications / C. Sha, P. Yan, D. Qin et al. URL: http://meetings.wmo.int/CBS-16/TECO/Presentations/Background-to-presentations/1D3-SUGON-SiliconCube.pdf
13. Adaptive Bubble Router: a Design to Improve Performance in Torus Networks /
14. V. Puente, R. Beivide, J. Gregorio et al. // Parallel Processing, 1999. Proceedings. 1999 International Conference on / IEEE, 1999. P. 58—67. URL: https://doi.org/10.1109/ICPP.1999.797388.
15. The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus / S. Scott et al. 1996. URL: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=7B4A3FAE3AC52B4E4141206064C00923?doi=10.1.1.126.3882&rep=rep1&type=pdf
16. Blue Gene/L Torus Interconnection Network / N. Adiga, M. Blumrich, D. Chen et al. // IBM Journal of Research and Development. 2005. Vol. 49, no. 2.3. P. 265–276.
17. Реализация аппаратной поддержки коллективных операций в маршрутизаторе высокоскоростной коммуникационной сети с топологией «многомерный тор» / Сыромятников Е.Л., Макагон Д.В., Парута С.И., Румянцев А.А. // Успехи современной радиоэлектроники. 2012. №1. С. 11–15.
18. A Brief Introduction to the Openfabrics Interfaces - A New Network API for Maximizing High Performance Application Efficiency / Grun P. et al. // 2015 IEEE 23rd Annual Symposium on High-Performance Interconnects. IEEE, 2015. P. 34–39.
19. Мукосей А.В., Семенов А.С. Приближенный алгоритм выбора оптимального подмножества узлов в коммуникационной сети Ангара с отказами // Вычислительные методы и программирование. 2017. Т. 18, № 1. С. 53–64.
20. Мукосей А.В., Семенов А.С. Оптимизация фрагментации при выделении ресурсов для высокопроизводительных вычислительных систем с сетью Ангара // Вестник ЮУрГУ, серия «Вычислительная математика и информатика». 2018. Т. 7, № 2. С. 50–62. DOI: 10.14529/cmse180204.
21. Мукосей А.В., Симонов А.С., Семенов А.С. Оптимизация утилизации при выделении ресурсов для высокопроизводительных вычислительных систем с сетью Ангара // Вестник ЮУрГУ, серия «Вычислительная математика и информатика». 2019. Т. 8, № 1. С. 5–19.
22. Angara Interconnect Makes GPU-based Desmos Supercomputer an Efficient Tool for Molecular Dynamics Calculations / Stegailov V., Semenov A., et al. // The International Journal of High Performance Computing Applications. 2019. Vol. 33, no. 3. P. 507–521. URL: https://journals.sagepub.com/doi/abs/10.1177/1094342019826667.
23. Результаты оценочного тестирования отечественной высокоскоростной коммуникационной сети Ангара / А.А. Агарков, Т.Ф. Исмагилов, Д.В. Макагон и др. // Суперкомпьютерные дни в России: Труды международной конференции (26-27 сентября 2016 г., г. Москва). М.: Изд-во МГУ, 2016. С. 626–639.
24. Tolstykh, M., Goyman, G., Fadeev, R., Shashkin, V.: Structure and Algorithms of SLAV Atmosphere Model Parallel Program Complex. Lobachevskii Journal of Mathematics. 2018. Vol. 39, no. 4. P. 587–595.
25. Акимов В.С. и др. Исследование масштабируемости FlowVision на кластере с интерконнектом Ангара // Вычислительные методы и программирование. 2017. Т. 18. №.4. С. 406–415.
26. Akimov V. S. et al. FlowVision Scalability on Supercomputers with Angara Interconnect // Lobachevskii Journal of Mathematics. 2018. V. 39, no. 9. P. 1159–1169.
27. Stegailov V., Smirnov G., Vecher V. VASP Hits the Memory Wall: Processors Efficiency Comparison. Concurrency and Computation: Practice and Experience. 2019. URL: https://doi.org/10.1002/cpe.5136.
28. Polyakov S., Podryga V., Puzyrkov D. High Performance Computing in Multiscale Problems of Gas Dynamics. Lobachevskii Journal of Mathematics. 2018. Vol. 39, no. 9. P. 1239–1250.
29. Ostroumova G., Orekhov N., Stegailov V. Reactive Molecular-Dynamics Study
30. of Onion-like Carbon Nanoparticle Formation. Diamond and Related Materials 94. 2019. P. 14–20.
31. Khalilov M., Timofeev A. Optimization of MPI-Process Mapping for Clusters with Angara Interconnect. Lobachevskii Journal of Mathematics. 2018. Vol. 39, no. 9. P. 1188–1198.

ANGARA INTERCONNECT: ARCHITECTURE AND PERFORMANCE RESULTS

Simonov A.[7], Zhabin I.[8], Kushtanov E.[9], Makagon D.[10], Semenov A.[11], Shcherbak A.[12]

Purpose: *Presentation of the architecture, hardware options, system and service software and the sustained performance results of the Angara interconnect.*

Results: *This paper presents the architecture of the Angara interconnect with 4D torus topology and describes its implementation in ASIC. We present the performance results of the Angara interconnect on two computing systems: Angara-C1 at JSC NICEVT and Desmos at JIHT RAS. On the different benchmarks and applications the Angara interconnect allows to obtain high performance and scalability, which are often better than corresponding characteristics on systems with the Mellanox Infiniband 4xFDR interconnect.*

Keywords: Angara interconnect, high performance computing, torus topology, Infiniband, MPI library

References:

1. Levin V. et al. Kommunikacionnaya set› MVS-Ekspress // Informacionnye tekhnologii i vychislitel›nye sistemy. 2014. No. 1. P. 10–24. [In Russ.]
2. Basalov V., Vyalukhin V. Adaptivnaya sistema marshrutizacii dlya otechestvennoj sistemy mezhprocessornyh obmenov SMPO-10G // Voprosy atomnoj nauki i tekhniki. Seriya: Matematicheskoe modelirovanie fizicheskih processov. 2012. No. 3. P. 64–70. [In Russ.]
3. Adamovich I. et al. Opyt razrabotki kommunikacionnoj seti superkomp'yutera «SKIF-Avrora» // Programmnye sistemy: teoriya i prilozheniya [Program Systems: Theory and Applications]. 2010. V. 1. No. 3. P. 107–123. [In Russ.]
4. Klimov Y. et al. Pautina: vysokoskorostnaya kommunikacionnaya set' // Programmnye sistemy: teoriya i prilozheniya [Program Systems: Theory and Applications]. 2015. V. 6, no. 1 (24). P. 109–120. [In Russ.]
5. Alverson R., Roweth D., Kaplan L. The Gemini System Interconnect // 2010 18th IEEE Symposium on High Performance Interconnects – IEEE. 2010. – P. 83-87.
6. Faanes G. et al. Cray Cascade: a Scalable HPC System Based on a Dragonfly Network // Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. – IEEE Computer Society Press. 2012. – 9 p.
7. Haring R. et al. The IBM Blue Gene/Q Compute Chip // IEEE Micro. 2011. –Vol. 32, no. 2. P. 48–60.
8. Ajima Y., Sumimoto S., Shimizu T. Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers. // IEEE Computer. 2009. Vol. 42, no. 11. P. 36—40. URL: https://doi.org/10.1109/MC.2009.370.
9. Tofu Interconnect 2: System-on-Chip Integration of High-Performance Interconnect / Y. Ajima, T. Inoue, S. Hiramoto et al. // International Supercomputing Conference / Springer. 2014. P. 498—507. URL: https://doi.org/10.1007/978-3-319-07518-1_35.
10. Ajima Y. et al. The Tofu Interconnect D // 2018 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, 2018. P. 646—654.
11. Scalable communication architecture for network-attached accelerators / S. Neuwirth, D. Frey, M. Nuessle, U. Bruening // High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium / IEEE, 2015. P. 627—638. URL: https://doi.org/10.1109/HPCA.2015.7056068.
12. Silicon Cube: A Supercomputer Specially Designed for Meteorological Applications / C. Sha, P. Yan, D. Qin et al. URL: http://meetings.wmo.int/CBS-16/TECO/Presentations/Background-to-presentations/1D3-SUGON-SiliconCube.pdf
13. Adaptive Bubble Router: a Design to Improve Performance in Torus Networks /
14. V. Puente, R. Beivide, J. Gregorio et al. // Parallel Processing, 1999. Proceedings. 1999 International Conference on / IEEE, 1999. P. 58—67. URL: https://doi.org/10.1109/ICPP.1999.797388.
15. The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus / S. Scott et al. 1996. URL: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=7B4A3FAE3AC52B4E4141206064C00923?doi=10.1.1.126.3882&rep=rep1&type=pdf
16. Blue Gene/L Torus Interconnection Network / N. Adiga, M. Blumrich, D. Chen et al. // IBM Journal of Research and Development. 2005. Vol. 49, no. 2.3. P. 265–276.
17. Realizatsiya apparatnoy podderzhki kollektivnykh operatsiy v marshrutizatore vysokoskorostnoy kommunikatsionnoy seti s topologiyey «mnogomernyy tor» / Makagon D.V., Syromyatnikov E.L., Paruta S.I., Rumyantsev A.A. // Uspekhi sovremennoy radioelektroniki [Progress of the Modern Radioelectronics]. 2012. No. 1. P. 11–15. [In Russ.]

7 Simonov Alexey, PhD in computer science, the first deputy of the general director of JSC NICEVT, Moscow, Russina. E-mail: simonov@nicevt.ru
8 Zhabin Ivan, deputy head of a department, JSC NICEVT, Moscow, Russia. E-mail: zhabin@nicevt.ru
9 Kushtanov Evgeny, head of a division, JSC NICEVT, Moscow, Russia. E-mail: evgeny.kushtanov@nicevt.ru
10 Makagon Dmitry, head of a department, JSC NICEVT, Moscow, Russia. E-mail: makagond@nicevt.ru
11 Semenov Alexander, head of a department, JSC NICEVT, Moscow, Russia. E-mail: semenov@nicevt.ru
12 Shcherbak Andrey, head of a department, JSC NICEVT, Moscow, Russia. E-mail: andrey.shcherbak@nicevt.ru

18. A Brief Introduction to the Openfabrics Interfaces - A New Network API for Maximizing High Performance Application Efficiency / Grun P. et al. // 2015 IEEE 23rd Annual Symposium on High-Performance Interconnects. IEEE, 2015. P. 34–39.
19. Mukosey A., Semenov A. Priblizhennyj algoritm vybora optimal›nogo podmnozhestva uzlov v kommunikacionnoj seti Angara s otkazami // Vychislitel›nye metody i programmirovanie. 2017. V. 18, no. 1. P. 53–64. [In Russ.]
20. Mukosey A., Semenov A. Optimizaciya fragmentacii pri vydelenii resursov dlya vysokoproizvoditel›nyh vychislitel›nyh sistem s set›yu Angara // Vestnik YUUrGU, seriya «Vychislitel›naya matematika i informatika» [Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering]. 2018. V. 7, no 2. P. 50–62. DOI: 10.14529/cmse180204. [In Russ.]
21. Mukosey A., Simonov A., Semenov A. Optimizaciya fragmentacii pri vydelenii resursov dlya vysokoproizvoditel›nyh vychislitel›nyh sistem s set›yu Angara // Vestnik YUUrGU, seriya «Vychislitel›naya matematika i informatika» [Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering]. 2019. V. 8, no 1. P. 5–19. [In Russ.]
22. Angara Interconnect Makes GPU-based Desmos Supercomputer an Efficient Tool for Molecular Dynamics Calculations / Stegailov V., Semenov A., et al. // The International Journal of High Performance Computing Applications. 2019. Vol. 33, no. 3. P. 507–521. URL: https://journals.sagepub.com/doi/abs/10.1177/1094342019826667.
23. Performance Evaluation of the Angara Interconnect / Agarkov A.A., Ismagilov T.F., Makagon D.V. et al. // Superkomp'yuternye dni v Rossii: Trudy mezhdunarodnoi konferentsii (Moskva, 26–27 sentyabrya 2016) [Russian Supercomputing Days: Proceedings of the International Scientific Conference (Moscow, Russia, September, 26–27, 2016)]. Moscow, Publishing of Moscow State University. 2016. P. 626–639. [in Russ.]
24. Tolstykh, M., Goyman, G., Fadeev, R., Shashkin, V.: Structure and Algorithms of SLAV Atmosphere Model Parallel Program Complex. Lobachevskii Journal of Mathematics. 2018. Vol. 39, no. 4. P. 587–595.
25. Issledovanie masshtabiruemosti FlowVision na klastere s interkonnektom Angara / Akimov V. et al. // Vychislitel›nye metody i programmirovanie. 2017. V. 18, no. 4. P. 406–415. [In Russ.]
26. Akimov V. S. et al. FlowVision Scalability on Supercomputers with Angara Interconnect // Lobachevskii Journal of Mathematics. 2018. V. 39, no. 9. P. 1159–1169.
27. Stegailov V., Smirnov G., Vecher V. VASP Hits the Memory Wall: Processors Efficiency Comparison. Concurrency and Computation: Practice and Experience. 2019. URL: https://doi.org/10.1002/cpe.5136.
28. Polyakov S., Podryga V., Puzyrkov D. High Performance Computing in Multiscale Problems of Gas Dynamics. Lobachevskii Journal of Mathematics. 2018. Vol. 39, no. 9. P. 1239–1250.
29. Ostroumova G., Orekhov N., Stegailov V. Reactive Molecular-Dynamics Study
30. of Onion-like Carbon Nanoparticle Formation. Diamond and Related Materials 94. 2019. P. 14–20.
31. Khalilov M., Timofeev A. Optimization of MPI-Process Mapping for Clusters with Angara Interconnect. Lobachevskii Journal of Mathematics. 2018. Vol. 39, no. 9. P. 1188–1198.