13. Коньшин И. Н., Сушко Г. Б., Харченко С. А. Сквозной параллельный алгоритм построения неполного треугольного разложения второго порядка точности с динамическим выбором декомпозиции и упорядочивания // Тез. докл. XIV Международ. конф. «Супервычисления и математическое моделирование». Саров, 2012. / Под ред. Р.М. Шагалиева. Труды XIV Международ. конф. «Супервычисления и математическое моделирование». Саров: РФЯЦ-ВНИИЭФ, 2013. С. 110–111.

14. Ерзунов В. А., Горбунов А. А. Механизм адаптивного выбора решателя в библиотеке PMLP/Parsol // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 2009. Вып. 1. С. 55–62.

15. Капорин И. Е., Милюкова О. Ю. Предобусловливание итерационных методов для эффективного массивно-параллельного решения систем линейных алгебраических уравний / Под ред. Р. М. Шагалиева // Труды XIII Международ. конф. «Супервычисления и математическое моделирование». Саров: РФЯЦ-ВНИИЭФ, 2012. С. 71–72.

16. Smith B. F., Bjorstad P. E., Gropp W. D. Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations. Cambridge: Cambridge University Press, 2004.

17. Бутнев О. И., Пронин В. А., Сидоров М. Л. и др. Пакет программ НИМФА-2 для решения задач многофазной фильтрации с применением суперкомпьютерных технологий / Под ред. Р. М. Шагалиева // Труды XIV Межд. конф. «Супервычисления и математическое моделирование». Саров, 2013. С. 112–119.

18. Saad Y. Iterative Methods for Sparse Linear Systems. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2003.

19. Lukarski D., Anzt H., Tomov S., Dongarra J. *Multi*-Elimination ILU Preconditioners on GPUs: Technical Report UT-CS-14-723 / Innovative Computing Laboratory, University of Tennessee, 2014.

ПЕРСПЕКТИВНАЯ ГИБРИДНАЯ ТОПОЛОГИЯ KNS ДЛЯ КОММУНИКАЦИОННОЙ СЕТИ НА ОСНОВЕ АППАРАТНОГО МОДУЛЯ СМПО-10GA-1

В. Г. Басалов

Российский федеральный ядерный центр –
Всероссийский НИИ экспериментальной физики, г. Саров

Введение

Около двадцати лет в ВНИИЭФ создаются суперкомпьютеры семейства МП-X-Y, как с использованием коммерческих коммуникационных сетей (Myricom, InfiniBand), так и с использованием систем межпроцессорного обмена собственной разработки в суперкомпьютерах МП-3 [1], МП-3Т, МП-СМПО-2D и МП-СМПО-3D.

В настоящее время в ИТМФ разработана отечественная система межпроцессорного обмена СМПО-10GA-1. Областью применения аппаратных компонент СМПО-10GA-1 является создание высокопроизводительных коммуникационных сетей для вычислительных систем разного уровня производительности: от компактных суперЭВМ до больших систем, состоящих из тысяч вычислительных узлов. 64-узловая вычислительная система МП-СМПО-3D, созданная на базе СМПО-10G-1 с топологией MultiTor, выдержала тестовые испытания.

Создание современных высокопроизводительных коммуникационных сетей (КС), объединяющих десятки тысяч вычислительных модулей в единую суперкомпьютерную вычислительную среду, ставит перед разработчиками много сложных технических проблем. Ниже перечислены основные требования к современным КС:

– расширяемость;
– простота;
– надежность;
– отказоустойчивость;
– стоимость и потребляемая мощность;
– управление загрузкой каналов.

Все эти требования имеют сильную взаимосвязь, и поэтому удовлетворение отдельных требований не приведет к достижению приемлемого результата.

Топология коммуникационной сети в значительной степени определяет эффективность и стоимость самой КС, а как следствие и эффективность всего мультипроцессорного вычислительного комплекса. Топология КС напрямую зависит от архитектуры ее аппаратных средств.

В настоящее время иерархическая организация КС приобретает характер фактического стандарта; она предлагалась еще в 2008 году в материалах рабочих групп DARPA по инновационному направлению создания экзафлопсных суперкомпьютеров.

Характерная общая черта суперкомпьютеров IBM Power 775, Cray XC30 и Tianhe-2 – их иерархическая архитектура и иерархическая коммуникационная сеть. Главный элемент такой сети – одночиповый маршрутизатор с множеством связей (линков). В IBM Power 775 это IBM HUB Chip с четырьмя процессорными интерфейсами (суммарная односторонняя пропускная способность 768 Гбит/с) и с 47 сетевыми линками трех типов (суммарная пропускная способность 4704 Гбит/с). В Cray XC30 это Cray Aries Chip с четырьмя процессорными интерфейсами (512 Гбит/с) и с 40 сетевыми линками трех типов (965 Гбит/с). В Tianhe-2 – маршрутизатор NRC Chip, суммарная пропускная способность однотипных линков составляет 1280 Гбит/с, из них два используются как интерфейсы с процессорами, а остальные – как сетевые.

Через процессорные интерфейсы к таким маршрутизаторам подключаются узлы – серверные платы. Маршрутизаторы обеспечивают соединения узлов на трех уровнях иерархии, реализуя при этом и переход с одного уровня на другой. Для IBM Power 775 и Cray XC30 реализованы соединения типа «каждый с каждым», и в перспективе такую же сеть следует ожидать в Tianhe-2, хотя сейчас используется топология «толстого» дерева, но с соединением «каждый с каждым» внутри процессорной стойки. Маршрутизатор NRC Chip уже сейчас позволяет перейти к иерархической сети, но, скорее всего, он будет переделан и приближен к IBM HUB Chip [2].

Архитектура, разрабатываемой в ИТМФ, системы межпроцессорных обменов СМПО-10GA-1 представляет собой двухкомпонентное электронное устройство, позволяющее, максимально ориентируясь на технологии доступные в России, использовать его, как основной составной элемент коммуникационной среды высокопроизводительных вычислительных систем. Аппаратный модуль состоит из двух компонент:

– коммутаторного блока СМПО-10GA-SW;
– адаптерного блока СМПО-10GA-AD.

Коммутаторный блок СМПО-10GA-SW представляет собой независимое устройство с 10 сетевыми линками (суммарная пропускная способность которых 400 Гбит/с).

Адаптерный блок СМПО-10GA-AD имеет один процессорный интерфейс для связи с вычислительным модулем и 4 сетевых линка (суммарной пропускной способностью 160 Гбит/с).

Для такой архитектуры СМПО-10GA-1 уже было ранее предложено несколько возможных топологий коммуникационной сети: TreeTor, MultiTor и ТорГК, но поиск новых перспективных топологий продолжался. Попытка разработать иерархическую топологию для архитектуры СМПО-10GA-1 не увенчалась успехом. Небольшое количество портов в коммутаторном блоке не позволяло создать расширяемую, с действительно малым диаметром топологию коммуникационной сети. Также препятствием явился довольно сложный алгоритм маршрутизации.

Гибридная топология KNS на основе архитектуры СМПО-10GA-1

Классические топологии коммуникационных сетей делятся на две группы: прямые (direct) и непрямые (indirect). К прямым топологиям относят такие широко известные топологии как Mesh, Tor, Hypercub и TOFU, а к непрямым топологиям относят, например, часто используемую в существующих вычислительных системах (ВС) как зарубежных, так и отечественных, топологию FatTree.

Эти топологии имеют серьезные ограничения, не позволяющие применять их в действительно больших коммуникационных сетях, объединяющих сотни тысяч вычислительных модулей.

Прямые топологии отличает дешевизна комплектующих (коммутаторы, как правило, имеют небольшое число портов), но это обуславливает значительное увеличение средней длины пути сообщения и соответственно снижает характеристики всей КС в целом.

Непрямые топологии имеют небольшую среднюю длину пути сообщения, зависящую от количества уровней и соответственно высокие характеристики, но снижение диаметра КС требует снижения количества уровней, которое достигается ростом количества портов в коммутаторах и, следовательно, значительном увеличении их стоимости.

Гибридная топология KNS, предложенная испанскими учеными на международной конференции в сентябре 2013 года в Барселоне [3], для вычислительных систем экзафлопсной производительности. Она представляет собой попытку совместить в одной топологии лучшие черты прямых и непрямых топологий, то есть добиться значительного снижения диаметра действительно больших КС за меньшие деньги, не допустив при этом значительного увеличения потребляемой мощности.

Свое название топология KNS получила от трех основных своих параметров: K – количество ВМ в одном измерении, N – количество измерений и S – количество уровней в непрямой топологии. ВМ в этой топологии располагаются ортогонально, как в прямых топологиях Mesh и Tor. Между собой они объединяются как в непрямых топологиях, с помощью либо одного полноматричного коммутатора, либо объединяя их в непрямые топологии, например, Fat Tree.

В настоящее время в мире не существует архитектурного решения системы межпроцессорного обмена, на котором возможно создание КС с топологией 4D KNS, за исключением архитектуры СМПО-10GA-1.

На рис. 1 представлена топология 2D KNS, построенная с использованием компонент аппаратного модуля СМПО-10GA-1: адаптерных блоков СМПО-10GA-AD и коммутаторных блоков СМПО-10GA-SW.

КС с топологией KNS, построенная на основе СМПО-10GA-1, может иметь от одного до четырех измерений. Максимальное количество измерений КС с топологией KNS определяется количеством портов (линков) адаптерного блока СМПО-10GA-AD, коммутатор которого имеет четыре сетевых порта и один процессорный интерфейс.

Система маршрутизации аппаратного модуля СМПО-10GA-1 с топологией KNS допускает разное количество ВМ в разных измерениях.

Соединение ВМ одного измерения осуществляется посредством одного коммутаторного блока СМПО-10GA-SW (4D KNS топология позволяет с использованием одиночного коммутатора СМПО-10GA-SW объединить до 10000 узлов), либо нескольких коммутаторных блоков СМПО-10GA-SW, объединенных с помощью непрямых топологий, например, Fat Tree или некое «обезжиренное» дерево. Объединение коммутаторов СМПО-10GA-SW в деревья позволит строить вычислительные системы любой размерности.

Система маршрутизации аппаратного модуля СМПО-10GA-1 с топологией KNS допускает для одного измерения применение как разных непрямых топологий, так и разных уровней этих топологий.

На рис. 2 представлена топология 1D KNS 20 1 2, соединяющая 20 ВМ с помощью группы коммутаторных блоков, объединенных в топологию Fat Tree.

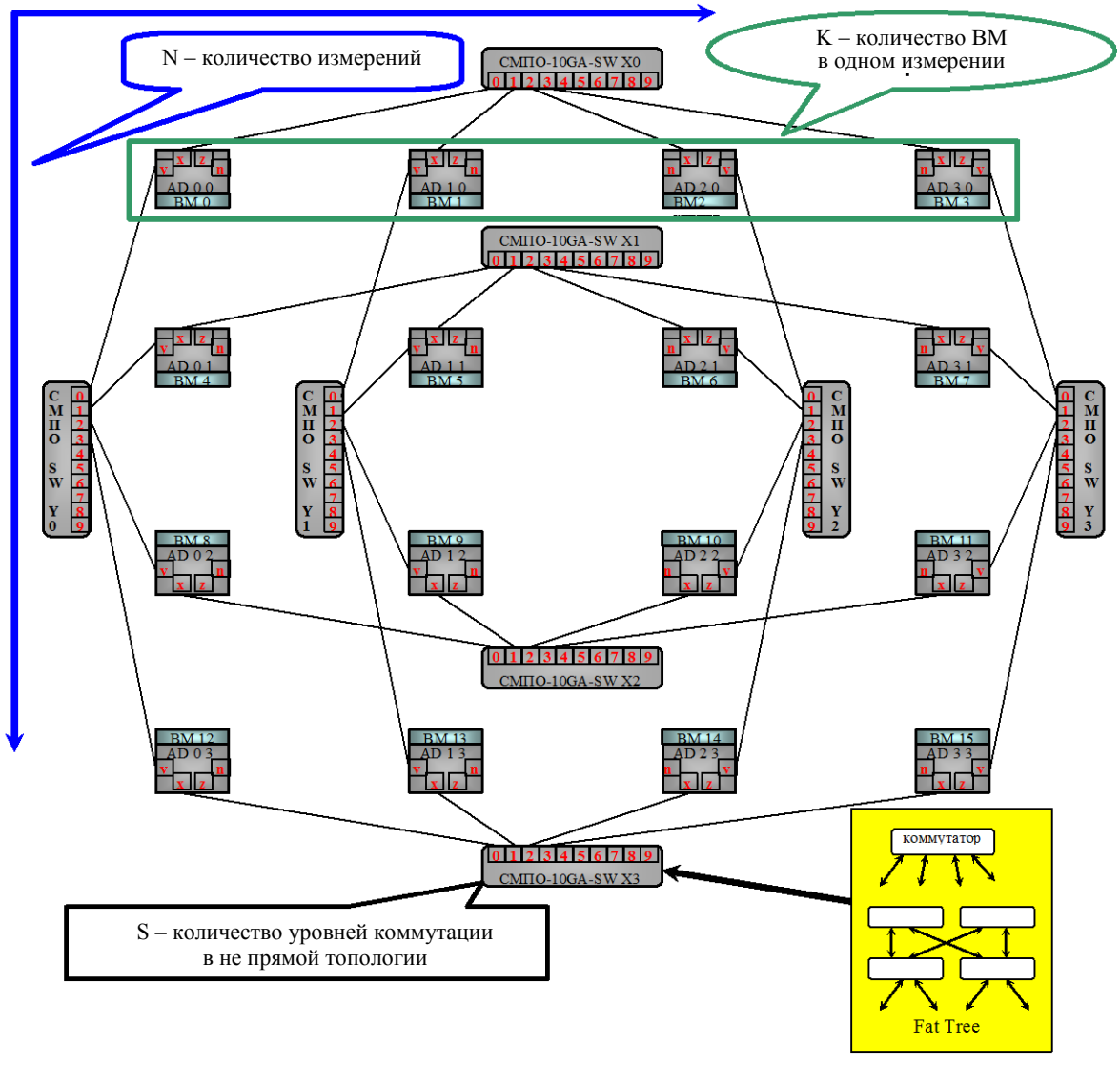

Рис. 1. Гибридная топология 2D KNS 4 2 1 на основе архитектуры СМПО-10GA-1

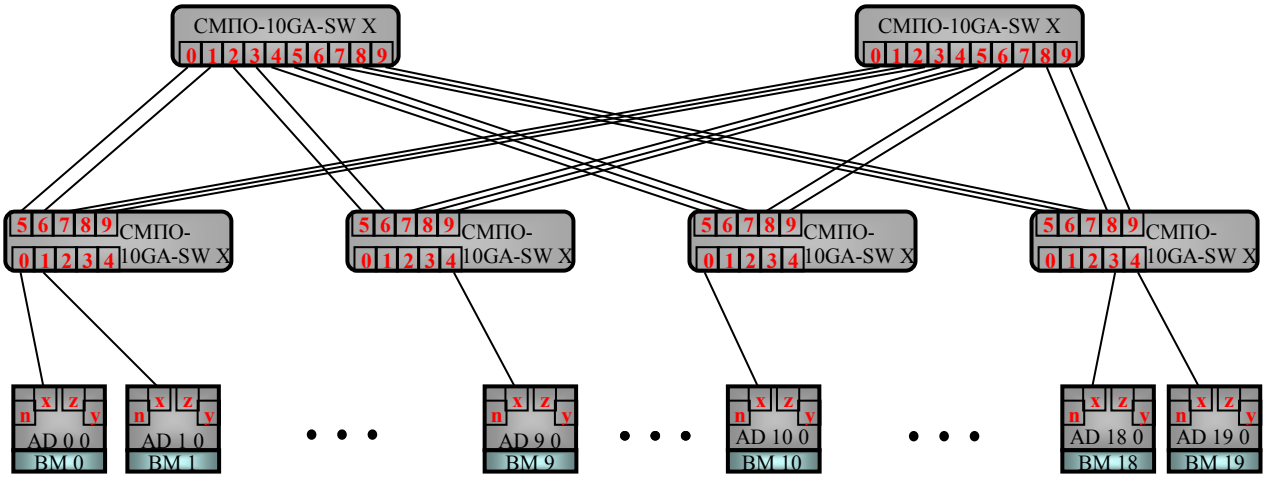

Рис. 2. Гибридная топология 1D KNS 20 1 2

Для топологии KNS были разработаны принципы идентификации коммутаторных и адаптерных блоков СМПО-10GA-1. Размер идентификаторов узлов коммуникационной сети СМПО-10GA-1 традиционно составляет четыре байта. При использовании топологии KNS коммутаторные и адаптерные блоки имеют разную идентификацию.

Идентификаторы адаптерных блоков занимают четыре байта, по одному байту на каждую размерность, что позволяет идентифицировать более четырех миллиардов ВМ.

Идентификаторы коммутаторных блоков, принадлежащих одному измерению, одинаковы и имеют значение номера байта этого измерения в идентификаторе адаптерного блока.

Адаптивный метод выбора оптимального маршрута сообщения

Применение топологии KNS позволяет значительно упростить процесс выбора оптимального маршрута сообщения, по сравнению с тороидальными топологиями, например, Tor или MultiTor. Это упрощение достигается за счет уменьшения количества типов циклов. В топологии KNS отсутствует цикл, образуемый обратной связью между узлами одного измерения. В топологии KNS остались только циклы, образуемые решетчатой структурой расположения вычислительных узлов. Для устранения возникновения состояний «взаимной блокировки» при передаче сообщений, традиционно используем специализированный алгоритм DOR (Dimension-order routing) или, как его еще называют XY-routing алгоритм. В случае отказа части оборудования происходит автоматический переход к адаптивной маршрутизации, за счет использований виртуальных каналов адаптерных блоков СМПО-10GA-AD. Следует заметить, что в коммутаторных блоках СМПО-10GA-SW нет необходимости вводить виртуальные каналы для предотвращения состояний «взаимной блокировки», так как участки непрямой топологии, где они находятся, и так свободны от циклов.

Поскольку аппаратный модуль СМПО-10GA-1 состоит из двух типов устройств адаптерных блоков СМПО-10GA-AD и коммутаторных блоков СМПО-10GA-SW, необходима разработка двух соответствующих алгоритмов маршрутизации, обеспечивающих гарантированную доставку пакета из ВМ источника в ВМ приемник.

В работе [4] был представлен метод алгоритмически коммутируемой маршрутизации. Идея этого метода заключается в том, что вычисление оптимального выходного порта для передачи каждого транзитного информационного сообщения осуществляется коммутатором непосредственно в момент передачи. В данной работе этот метод был адаптирован к топологии KNS.

Адаптивный маршрутный алгоритм коммутатора адаптерного блока СМПО-10GA-AD

Для принятия решения о маршрутизации пакета коммутатор адаптерной платы СМПО-10GA-AD сравнивает адрес получателя D (X, Y, Z, W) со своим идентификатором C (X, Y, Z, W). Сравнение производится вычитанием значений полей D (X, Y, Z, W) из значений полей C (X, Y, Z, W) и получением разности R (X, Y, Z, W).

Далее выполняется алгоритм – DOR. Выполнение алгоритма всегда начинается с R(X).

Если R(X) не равен 0, то пакет отправляется в порт с номером X и виртуальный канал VC0, если передача пакета в порт с номером X невозможна, то выбирается первый способный к передаче пакетов порт с виртуальным каналом VC1, если их нет, то делаем вывод о неисправности сети.

Иначе переход к рассмотрению R(Y), если R(Y) не равен 0, то пакет отправляется в порт Y и виртуальный канал VC0. Если передача пакета в порт с номером Y невозможна, то выбирается первый, способный к передаче пакетов порт, с виртуальным каналом VC1, если их нет, то делаем вывод о неисправности сети.

Иначе переход к рассмотрению R(Z), если R(Z) не равен 0, то пакет отправляется в порт Z и виртуальный канал VC0. Если передача пакета в порт с номером Z невозможна, то выбирается первый, способный к передаче пакетов порт, с виртуальным каналом VC1, если их нет, то делаем вывод о неисправности сети.

Иначе переход к рассмотрению R(W), если R(W) не равен 0, то пакет отправляется в порт W и виртуальный канал VC0. Если передача пакета в порт с номером W невозможна, то выбирается первый, способный к передаче пакетов порт, с виртуальным каналом VC1, если их нет, то делаем вывод о неисправности сети.

В завершении, если D (X, Y, Z, W) и C (X, Y, Z, W) совпали, пакет передается в ВМ получатель.

Адаптивный маршрутный алгоритм коммутатора коммутаторного блока СМПО-10GA-SW

Архитектура аппаратного модуля СМПО-10GA-1 позволяет с использованием одиночного коммутатора СМПО-10GA-SW построить топологией 4D KNS вычислительную систему, содержащую до 10000 вычислительных модулей.

При этом алгоритм маршрутизации очень прост и заключается в том, что коммутатор выбирает из четырехбайтового адреса пакета номер байта, равный его идентификатору. Значение этого байта и содержит оптимальный выходной порт. Если этот порт неспособен к передаче пакета, выбираем первый порт, через который можно отправить пакет.

Реализация этого алгоритма не требует никаких подготовительных действий.

Для вычислительных систем большего размера, когда непрямая часть топологии KNS представляет собой многоуровневое дерево, разработана универсальная табличная маршрутизация, т. е. выходной порт для пакета выбирается из заранее созданных и загруженных в коммутаторы таблиц. Таблицы размерностью до 30 элементов позволят создавать вычислительные системы до 810000 вычислительных модулей.

Таблица представляет собой массив, где индексом является значение соответствующего измерению коммутаторного блока номера байта в адресе пакета, а значением является номер оптимального выходного порта. Например, в коммутаторных блоках измерения X индексом таблицы будет значения 0-го байта в адресе пакета, в коммутаторных блоках измерения Y индексом таблицы будет значения 1-го байта в адресе пакета и так далее.

Маршрутные таблицы рассчитываются заранее с учетом параметров вычислительной системы. При топологии KNS таблицы для соответствующих коммутаторных блоков одного измерения будут одинаковы. Небольшой размер таблиц и их небольшое разнообразие значительно упрощает их создание.

Для увеличения надежности системы возможно использование адаптивной маршрутизации с увеличением таблицы в два раза для хранения основного и альтернативного выходного порта.

Сравнение характеристик топологий коммуникационных сетей

В таблице приведены основные характеристики (диаметр, ширина бисекции и стоимость, выраженная в количестве необходимого оборудования: адаптеров, коммутаторов и соединительных кабелей) различных топологий коммуникационных сетей на базе архитектуры СМПО-10GA-1, топологии коммуникационной сети «Ангара» и различных топологий коммуникационных сетей на базе архитектуры InfiniBand для вычислительной системы, содержащей примерно 8192 вычислительных модуля.

Название топологии	«Ангара» 4D Tor (16, 8, 8, 8)	СМПО-10GA-1 1C MultiTor (16, 16, 8)	СМПО-10GA-1 3C MultiTor (16, 16, 8)	СМПО-10GA-1 3C MultiTor 10*10*10*8 ВМ	InfiniBand 3D Tor (8, 8, 8) 16 ВМ	InfiniBand Fat Tree 36 портов $S = 3$
Диаметр	$\sum_{i=1}^{k}(z_i/2)$ 20	$\sum_{i=1}^{k}(z_i/2)+2$ 22	$\sum_{i=1}^{k}(z_i/2)+2$ 22	$\sum_{i=1}^{k}(z_i/2)+2(S_i-1)$ 8	$\sum_{i=1}^{k}(z_i/2)+2$ 14	$2S$ 6
Стоимость (количество ребер)	$k*\prod_{i=1}^{k}z_i$ 32768	$k*\prod_{i=1}^{k}z_i+N$ 14336	$3*k*\prod_{i=1}^{k}z_i+$ $+12*\prod_{i=1}^{k}z_i$ 14336	$\sum_{i=1}^{N'}F_i\prod_{i=1,\,i\neq j}^{N'}K_j$ 34000	$3*k*\prod_{i=1}^{k}z_i+N$ 12800	$S*N$ 24576
Ширина бисекции	$\mathrm{Min}(z_0*z_1*z_2;$ $z_0*z_2*z_3;$ $z_0*z_1*z_3;$ 512	$\mathrm{Min}(z_0*z_1;$ $z_0*z_2;z_1*z_2)*2$ 256	$\mathrm{Min}(z_0*z_1;$ $z_0*z_2;z_1*z_2)*6$ 768	$N/2$ 4000	$\mathrm{Min}(z_0*z_1;z_0*z_2;$ $z_1*z_2)*3$ 384	$N/2$ 4095
Количество адаптерных блоков	N 8192	N 8192	N 8192	N 8000	N 8192	N 8190
Количествао коммутаторных блоков	0	$\prod_{i=1}^{k}z_i$ 2048	$3*\prod_{i=1}^{k}z_i$ 6144	$\sum_{i=1}^{N'}V_i\prod_{i=1,\,i\neq j}^{N'}K_j$ 3400	$\prod_{i=1}^{k}z_i$ 512	$2{,}5*N/18$ 1138

Примечание. N – количество вычислительных модулей, k – мерный тор T (z_1, z_2, …, z_k) с длинами сторон z_1, z_2,…, z_k, C – степень MultiTor, S – количество уровней в Fat Tree, F_i – количество связей в дереве i-го измерения и V_i – количество коммутаторных блоков СМПО-10GA-SW в дереве i-го измерения.

Заключение

В результате проведенной работы были проанализированы основные мировые тенденции развития коммуникационных сетей суперкомпьютеров и исследована возможность использования для коммуникационных сетей на базе аппаратного модуля СМПО-10GA-1 новых конкурентоспособных топологий, например, KNS.

В настоящее время в мире не существует архитектурного решения системы межпроцессорного обмена, на котором возможно создание КС с многомерной топологией KNS, за исключением архитектуры СМПО-10GA-1.

Было выяснено, что построение вычислительных систем по этой топологии не требует никаких изменений в механизмах, алгоритмах и аппаратуре компонент аппаратного модуля СМПО-10GA-1. Изменения коснутся только маршрутного алгоритма системы маршрутизации.

На этом этапе работ над топологией KNS разработаны:

– принципы идентификации адаптерных и коммутаторных блоков, входящих в распределенную коммуникационную среду СМПО-10GA-1 с топологией KNS;

– адаптивные маршрутные алгоритмы адаптерных блоков СМПО-10GA-AD для топологии KNS;

– адаптивные маршрутные алгоритмы коммутаторных блоков СМПО-10GA-SW для топологии KNS;

– приведены основные характеристики коммуникационных сред с топологией KNS.

Проведенное сравнение основных характеристик различных топологий коммуникационных сетей на базе системы межпроцессорных обменов СМПО-10GA-1, топологии коммуникационной сети «Ангара» и различных топологий коммуникационных сетей на базе архитектуры InfiniBand показало, что КС на архитектуре СМПО-10GA-1 с топологией KNS имеют параметры, сравнимые с параметрами КС на архитектуре InfiniBand с топологией Fat Tree.

Литература

1. Вихарев В. М., Сапронов С. И. Принципы программной организации коммуникационной системы мультипроцессора МП-3 // Вопросы атомной науки и техники. Сер. Матиматическое моделирование физических процессов. 1997. Вып. 2. С. 79–84.

2. Горбунов В., Елизаров Г., Эйсымонт Л. Экзафлопсные суперкомпьютеры: достижения и перспективы // Открытые системы. 2013. № 7.

3. García P. J., Escudero-Sahuquillo J., Quiles F. J., Duato J. High-Performance Interconnection Networks on the Road to Exascale HPC: Challenges and Solutions // 12th HPC Advisory Council Conference. September, 2013, Barcelona, Spain.

4. Басалов В. Г., Вялухин В. М. Адаптивная система маршрутизации для отечественной системы межпроцессорных обменов СМПО-10G // Вопросы атомной науки и техники. Сер. Матиматическое моделирование физических процессов. 2012. Вып. 3. С. 64–70.

УНИФИЦИРОВАННАЯ СИСТЕМА «GEOMGRID2» ДЛЯ ОБЕСПЕЧЕНИЯ ПРЕПРОЦЕССИНГА ДВУМЕРНЫХ ЗАДАЧ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

О. В. Беломестных, С. В. Гагарин, Е. А. Приб, А. А. Ушкова

Российский федеральный ядерный центр –
Всероссийский НИИ технической физики им. Е. И. Забабахина, г. Снежинск

Введение

На протяжении многих лет в РФЯЦ-ВНИИТФ проводятся расчеты задач механики сплошной среды (МСС) при помощи различных методик численного моделирования института. С каждым годом растут требования к скорости и качеству проведения расчетов методиками численного моделирования. Также с увеличением вычислительных мощностей института и повышением сложности задач, рассчитываемых программными комплексами математического моделирования РФЯЦ-ВНИИТФ, остро встал вопрос о повышении эффективности подготовки данных для расчета и необходимости в разработке новых современных средств для более качественного и точного задания геометрии и формирования начальных данных.

В 2010 году в РФЯЦ-ВНИИТФ была начата разработка системы «GeomGrid2» [1] для обеспечения двумерных программных комплексов математического моделирования института всей функциональностью, необходимой для эффективного и качественного расчета начальных данных.

Система «GeomGrid2» обеспечивает функциональностью все звенья технологической цепочки этапа подготовки и расчета начальных данных (РНД). Функциональность системы позволяет: формировать геометрию задачи, осуществлять построение сеток и полей частиц, задавать газодинамические параметры, распределять вещества по ячейкам сеток и формировать начальные файлы-разрезы в различных форматах данных.