УДК 004.722

ПЕРСПЕКТИВНАЯ ГИБРИДНАЯ ТОПОЛОГИЯ KNS ДЛЯ СИСТЕМ МЕЖПРОЦЕССОРНЫХ ОБМЕНОВ НА БАЗЕ СМПО-10G

В. Г. Басалов, А. А. Холостов
(ФГУП "РФЯЦ-ВНИИЭФ", г. Саров Нижегородской области)

Представлено описание гибридной топологии KNS для коммуникационных сетей мультипроцессорных систем, основанных на системе межпроцессорных обменов СМПО-10G. Описаны созданные для этой топологии адаптивные маршрутные алгоритмы, позволяющие строить кратчайшие пути передачи сообщений, свободные от возникновения состояний взаимных блокировок, а также обеспечивающие высокую отказоустойчивость коммуникационной сети и ее сбалансированную загрузку. Приведен сравнительный анализ основных технико-архитектурных характеристик популярных топологий, применяемых при создании коммуникационных сетей высокопроизводительных многопроцессорных вычислительных комплексов.

Ключевые слова: многопроцессорный вычислительный комплекс, коммуникационная сеть, система межпроцессорного обмена СМПО-10G, гибридная топология KNS, адаптивные маршрутные алгоритмы, виртуальный канал.

Введение

На протяжении ряда лет ведутся работы по созданию отечественной системы межпроцессорного обмена (СМПО) для построения многопроцессорных вычислительных комплексов (МВК) различного класса. Отечественные СМПО использовались при создании МВК МП-3, МП-3Т [1], МП-СМПО-2D и МП-СМПО-3D.

Разработанная аппаратно-программная архитектура СМПО-10G является основой создания высокоскоростных коммуникационных сетей (КС) для многопроцессорных вычислительных систем разного уровня производительности — от компактных вычислительных систем до больших МВК, состоящих из тысяч вычислительных узлов. 64-узловая вычислительная система МП-СМПО-3D, созданная на базе архитектуры СМПО-10G с топологией "Мультитор", выдержала приемочные испытания и успешно проходит опытную эксплуатацию [2].

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

- высокая скорость передачи данных (быстродействие);
- расширяемость;
- простота;
- надежность;
- отказоустойчивость;
- низкая стоимость и потребляемая мощность;
- управление загрузкой каналов (адаптируемость).

Все эти требования взаимосвязаны, и удовлетворение только отдельных из них не приведет к успешной коммерциализации КС. Тем не менее самой важной характеристикой при работе в МВК является скорость передачи данных (пропускная способность) КС в условиях реальной нагрузки МВК.

Топология КС в значительной степени определяет интегральное быстродействие и технико-экономическую эффективность СМПО и, как следствие, всего МВК. Топология КС во многом зависит от архитектуры используемых аппаратных средств.

В настоящее время иерархическая организация КС приобретает характер фактического стандарта. Характерная общая черта суперкомпьютеров IBM Power 775, Cray XC30 и Tianhe-2 — их иерархическая архитектура и иерархическая КС. Главный элемент такой сети — одночиповый коммутатор-маршрутизатор с множеством сетевых каналов связи (100 и более высокоскоростных последовательных каналов на кристалл). Через процессорные интерфейсы к таким маршрутизаторам подключаются узлы — серверные платы. Маршрутизаторы обеспечивают соединения ВМ 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 в ней 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 модифицирован и приближен к IBM HUB Chip [2].

Архитектура СМПО-10G основывается на двухкомпонентном аппаратном модуле, позволяющем, ориентируясь на доступные в России технологии, использовать его как основной составной элемент КС высокопроизводительных МВК. Аппаратный модуль состоит из двух компонентов: адаптерного и коммутаторного блоков.

Адаптерный блок представляет собой коммутатор-маршрутизатор для организации одного процессорного интерфейса для связи с ВМ и обеспечения работы четырех сетевых каналов (суммарной пропускной способностью 160 Гбит/с) для внешних связей. Коммутаторный блок представляет собой полный матричный коммутатор-маршрутизатор с десятью сетевыми каналами связи (суммарная пропускная способность — 400 Гбит/с), обеспечивающий передачу сообщений между этими каналами по запрограммированным ранее маршрутным алгоритмам [3]. Каждый порт коммутатора как адаптерного, так и коммутаторного блока имеет два виртуальных канала.

Для архитектуры СМПО-10G ранее рассматривалось несколько пространственно расширяемых топологий КС: Tree3DTorus, "Мультитор" и HTorus (тор гиперкубов третьей степени) [4]. Перспективность иерархической топологии для архитектуры СМПО-10G невелика, так как небольшое количество портов в коммутаторном блоке, обусловленное возможностями российских технологий, не позволяет создать большую КС с действительно малым диаметром. Дополнительными препятствиями являются разработка и реализация довольно сложного алгоритма маршрутизации для таких сетей.

Гибридная топология KNS на основе архитектуры СМПО-10G

Топологии КС принято разделять на две группы: прямые (direct) и непрямые (indirect).

Непрямая топология ограничивает возможность поэтапного расширения МВК и не обеспечивает линейной зависимости стоимости МВК от ее производительности, что ограничивает ее применение в действительно больших КС, объединяющих сотни тысяч ВМ. КС с непрямой топологией имеет небольшой диаметр, зависящий от количества уровней сети, и соответственно высокие технико-архитектурные характеристики. Однако дальнейшее уменьшение диаметра КС обеспечивается только уменьшением количества уровней, которое достигается ростом количества портов в коммутаторах, а следовательно, значительным увеличением их стоимости и использованием дорогостоящих технологий. К непрямым топологиям относится, например, часто используемая в существующих МВК, как зарубежных, так и отечественных, топология Fat Tree.

Прямые топологии, наоборот, отличаются дешевизной комплектующих (коммутаторы, как правило, имеют небольшое число портов) и хорошей масштабируемостью, в основном не зависящей от количества портов. Но это обусловливает значительное увеличение диаметра КС и средней длины пути сообщения, повышает вероятность конфликтов сообщений и соответственно снижает пропускную способность, т. е. ухудшает технико-архитектурные характеристики всей КС в целом. К прямым топологиям относятся широко известные топологии Mesh, Tor, Hypercub и TOFU.

Гибридная топология KNS, предложенная испанскими учеными на международной конференции в сентябре 2013 г. в Барселоне [5] для реализации ВС эксафлопсной производительности, представляет собой попытку совместить в одной сети лучшие черты прямых и непрямых топологий, т. е. добиться значительного уменьшения

диаметра больших КС, стоимости, энергозатрат при обеспечении высокой пропускной способности.

Свое название топология KNS получила от трех основных своих параметров: K — количество ВМ в одном измерении (координатном направлении), N — количество измерений в прямой топологии, S — количество уровней в непрямом участке топологии. ВМ в этой топологии располагаются (нумеруются) ортогонально измерениям, аналогично прямым топологиям Mesh или Torus. Все ВМ, имеющие одинаковое значение одной из координат (например все ВМ с координатами (0, 0), (1, 0), (2, 0), (3, 0), т. е. принадлежащие оси координат X), объединяются между собой с помощью либо одного полноматричного коммутатора, как представлено на рис. 1 (см. также цветную вкладку), либо при недостаточном количестве портов в коммутаторе — в непрямые топологии, например Fat Tree, как показано на рис. 2.

КС с топологией KNS, построенная на основе аппаратного модуля СМПО-10G, может иметь от одного до четырех измерений. Максимальное количество измерений КС с топологией KNS определяется количеством портов адаптерного блока в СМПО-10G, внутренний коммутатор которого имеет четыре внешних сетевых порта и

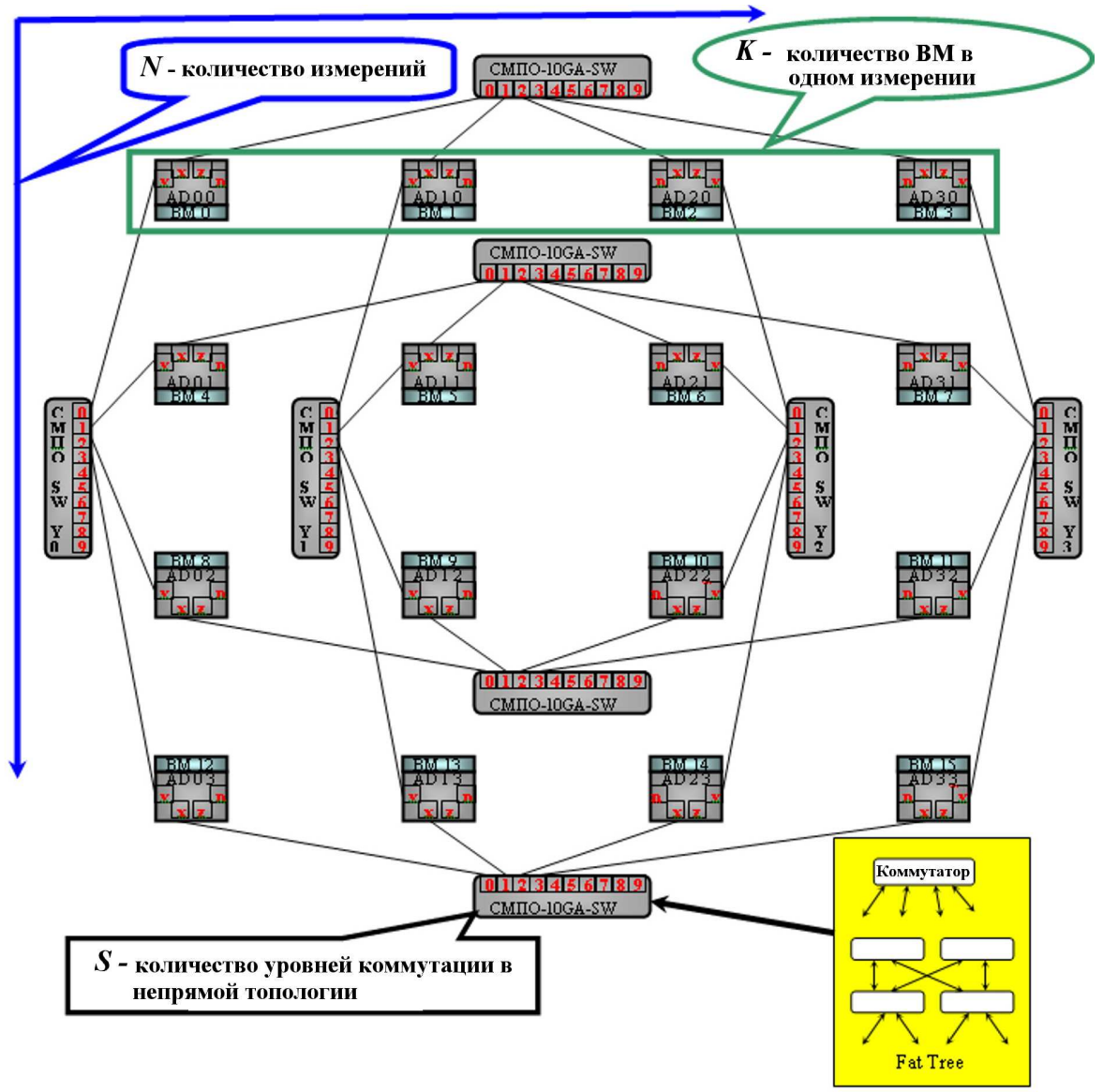

Рис. 1. Гибридная топология 2D KNS (4×2×1) на основе архитектуры СМПО-10G

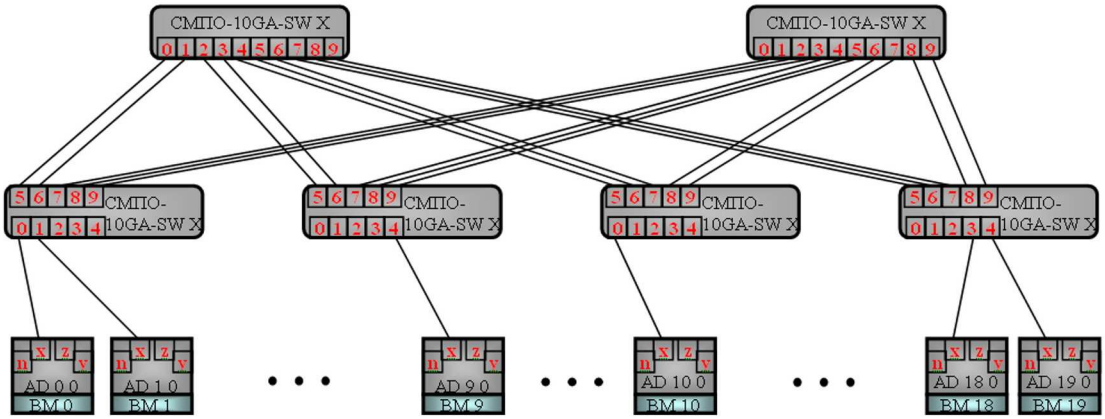

Рис. 2. Гибридная топология 1D KNS (20×1×2) на основе архитектуры СМПО-10G

один процессорный интерфейс. Соединение ВМ одного измерения посредством одного коммутаторного блока из СМПО-10G, т. е. создание топологии 4D KNS с размерами 10×4×1, позволит объединить до 10 000 узлов с диаметром КС, равным 8.

Для топологий с количеством ВМ в одном направлении более 10 можно соединять ВМ в одном измерении с использованием нескольких коммутаторных блоков, которые, в свою очередь, объединены с помощью непрямых топологий (например Fat Tree или некоторого *обезжиренного дерева*), и таким образом создавать КС для МВК любого размера. Разработанная система маршрутизации аппаратного модуля СМПО-10G с топологией KNS позволяет это сделать. Она допускает использование разного количества ВМ в различных измерениях, а также применение для различных измерений как разных непрямых топологий, так и разного количества уровней этих топологий для объединения ВМ.

Для топологии KNS были разработаны принципы идентификации коммутаторных и адаптерных блоков СМПО-10G. Размер идентификаторов узлов КС СМПО-10G традиционно, как и для других топологий, составляет четыре байта. Однако коммутаторные и адаптерные блоки имеют разную структуру идентификации. В идентификаторах адаптерных блоков отводится по одному байту на каждое измерение (рис. 3), что позволяет идентифицировать более 4 млрд ВМ. Идентификаторы коммутаторных блоков, принадлежащих одному измерению, одинаковы и имеют значение номера байта, соответствующего этому измерению в идентификаторе адаптерного блока (рис. 4).

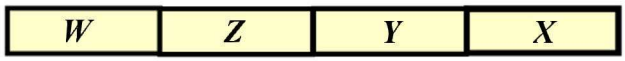

Рис. 3. Формат идентификаторов адаптерных блоков

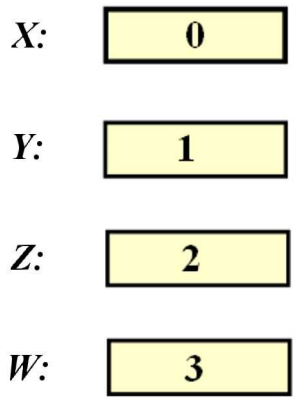

Рис. 4. Форматы идентификаторов коммутаторных блоков для разных измерений

Адаптивный метод выбора оптимального маршрута сообщения

Применение топологии KNS позволяет значительно упростить процесс выбора оптимального маршрута сообщения по сравнению с тороидальными топологиями, например Torus или "Мультитор" [4]. Это упрощение достигается за счет уменьшения количества типов циклов, угрожающих возникновением состояний взаимных блокировок. В топологии KNS отсутствует тип циклов, образуемый обратной связью между узлами одного измерения. Здесь остались

только циклы, образуемые решетчатой структурой расположения ВМ. Для устранения возникновения состояний взаимной блокировки при передаче сообщений традиционно используется специализированный алгоритм DOR (Dimension-Order Routing), или, как его еще называют, XY-routing-алгоритм [6]. В случае отказа части оборудования происходит автоматический переход к адаптивной маршрутизации за счет использования двух виртуальных каналов VC0 и VC1 адаптерных блоков. Следует заметить, что в коммутаторных блоках СМПО-10G нет необходимости вводить виртуальные каналы для предотвращения состояний взаимной блокировки, так как участки непрямой топологии, где бы они ни находились, заведомо топологически свободны от циклов.

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

В работе [7] был представлен метод алгоритмически коммутируемой маршрутизации для прямых топологий на базе СМПО-10G. 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 адаптирован к топологии KNS.

Адаптивный маршрутный алгоритм адаптерного блока

Для принятия решения о дальнейшей маршрутизации пакета адаптерный блок сравнивает идентификатор ВМ-приемника $\{X_D, Y_D, Z_D, W_D\}$ с идентификатором текущей адаптерной платы $\{X_C, Y_C, Z_C, W_C\}$. Сравнение производится вычитанием значений полей X_D, Y_D, Z_D, W_D из значений полей X_C, Y_C, Z_C, W_C соответственно. В результате получается разность $\{X_R, Y_R, Z_R, W_R\}$.

Если $\{X_D, Y_D, Z_D, W_D\}$ и $\{X_C, Y_C, Z_C, W_C\}$ совпали, т. е. все значения полей X_R, Y_R, Z_R, W_R равны 0, пакет достиг ВМ-приемника, сообщение передается процессорному интерфейсу. В противном случае пакет считается транзитным и начинается поиск оптимального выходного порта с помощью алгоритма DOR.

Выполнение алгоритма всегда начинается с анализа значения поля X_R. Если оно не равно 0, то пакет отправляется в порт, соответствующий координате X, с виртуальным каналом VC0. Если передача пакета в соответствующий порт невозможна, то выбирается первый способный к передаче пакетов порт с виртуальным каналом VC1. Если таких нет, то делается вывод о неисправности сети.

Если значение поля X_R равно 0, проводится такой же анализ для значений полей Y_R, Z_R, и W_R, пока не будет найден оптимальный выходной порт.

Адаптивный маршрутный алгоритм коммутаторного блока

Как упоминалось ранее, архитектура СМПО-10G позволяет с использованием одиночного коммутаторного блока на непрямом участке топологии 4D KNS (размером 10×4×1) построить КС, объединяющую до 10 000 ВМ. При этом алгоритм маршрутизации очень прост и заключается в том, что маршрутизатор, анализируя четырехбайтовый адаптерный идентификатор (он же адрес ВМ-приемника), находит номер байта, совпадающий с идентификатором коммутаторного блока, и значение этого байта является оптимальным номером выходного порта. Если этот порт не способен к передаче пакета, выбирается любой порт, через который можно отправить пакет. Реализация данного алгоритма не требует никаких подготовительных действий.

Для МВК большого 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 коммутаторные блоки СМПО-10G маршрутных таблиц. Таблицы, содержащие до 30 элементов, позволят создавать вычислительные системы до 810 000 ВМ.

Таблица представляет собой массив данных, где индексом является значение байта, соответствующего измерению (координате) коммутаторного блока в адресе пакета, а результатом является номер оптимального выходного порта. Например, в коммутаторных блоках измерения X индексом таблицы будет значение нулевого байта в адресе пакета, в коммутаторных блоках измерения Y — значение первого байта в адресе пакета и т. д.

Маршрутные таблицы рассчитываются заранее с учетом параметров МВК. При топологии KNS таблицы для соответствующих коммутаторных блоков одного измерения будут одинаковы. Малое разнообразие и небольшой размер таблиц значительно упрощают процесс их создания.

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

Характеристики КС с топологий KNS

Количество ВМ в МВК, использующем КС с топологией KNS, определяется по формуле

$$N_{\text{ВМ}} = \prod_{i=1}^{n} K_i,$$

где n — количество измерений в топологии KNS (для СМПО-10G $1 \leq n \leq 4$); K_i — количество ВМ в i-м измерении.

МВК, содержащий $N_{\text{ВМ}}$ ВМ, включает в себя соответственно и $N_{\text{ВМ}}$ адаптерных блоков.

Общее количество коммутаторных блоков в МВК с топологией KNS определяется по формуле

$$V_{\text{КМ}} = \sum_{i=1}^{n} \left(V_i \prod_{\substack{j=1,\\ j\neq i}}^{n} K_j \right),$$

где V_i — количество коммутаторных блоков в дереве непрямой топологии i-го измерения.

Приведем другие важные характеристики КС, описание которых дано в работе [4].

Диаметр топологии KNS определяется по формуле

$$D = \sum_{i=1}^{n} 2 + 2(S_i - 1),$$

где S_i — количество уровней непрямой топологии в i-м измерении. Величина D может охарактеризовать максимально необходимое время для передачи данных между ВМ, поскольку время передачи обычно прямо пропорционально длине пути. Отметим также, что минимальная дистанция в топологии KNS равна 2.

Топология KNS обладает свойством полной бисекции, и ее ширина равна $N_{\text{ВМ}}/2$ при любом разделении вычислительной системы пополам.

Значение связности в топологии KNS равно n. Оно определяется количеством измерений топологии KNS, а следовательно, количеством задействованных сетевых каналов связи адаптерных блоков.

Устойчивость к неисправностям повышается с увеличением количества измерений в топологии KNS. Это достигается за счет увеличения количества альтернативных маршрутов и реализации адаптивных маршрутных алгоритмов, обеспечивающих обход неисправных участков.

Стоимость — показатель, который может быть определен, например, как общее количество каналов связи в КС МВК. Стоимость КС с топологией KNS можно определить по формуле

$$F_{\text{КМ}} = \sum_{i=1}^{n} \left(F_i \prod_{\substack{j=1,\\ j\neq i}}^{n} K_j \right),$$

где F_i — количество связей в дереве непрямой топологии i-го измерения.

Сравнение характеристик топологий КС

На рис. 5—7 представлено сравнение основных технико-архитектурных характеристик (диаметр, ширина бисекции и стоимость, выраженная в количестве соединительных кабелей) двух различных топологий КС на базе архитектуры СМПО-10G, топологий КС на базе архитектуры InfiniBand (с 36-портовыми коммутаторами) и топологии КС на базе архитектуры "Ангара" для МВК, объединяющих разное количество ВМ.

Из рис. 5, 6 видно, что диаметры (лучшие показатели имеют меньшие значения) и бисекции (лучшие показатели — с большими значениями) КС на базе архитектуры СМПО-10G с топологией KNS незначительно уступают только КС на базе архитектуры InfiniBand с топологией Fat Tree, но при этом значительно опережают другие топологии.

Из рис. 7 видно, что стоимость КС (выраженная через количество каналов связи) с архитектурой СМПО-10G и топологией KNS ниже, чем стоимость КС с архитектурой СМПО-10G и топологией "Мультитор", практически совпадает со стоимостью КС с архитектурой "Ангара" и топологией 4D Torus и выше, чем стоимость КС с архитектурой InfiniBand, особенно с топологией 3D Torus. При этом важно отметить, что коммутаторные блоки InfiniBand более чем в 3 раза

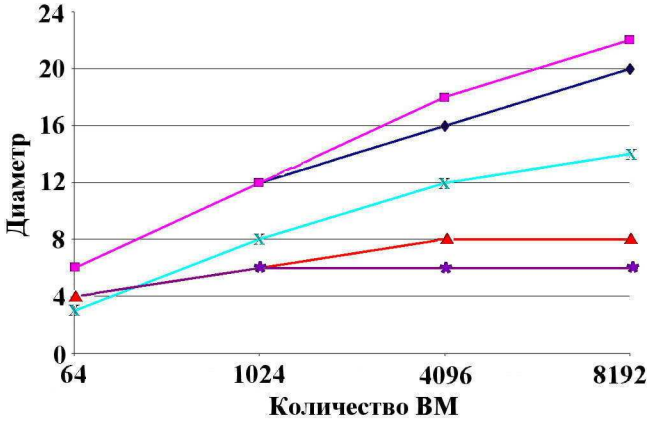

Рис. 5. Зависимость диаметра КС разных топологий от количества ВМ: —♦— — "Ангара", 4D Torus; —■— — СМПО-10G, "Мультитор"; —▲— — СМПО-10G, KNS; —×— — InfiniBand, 3D Torus; —*— — InfiniBand, Fat Tree

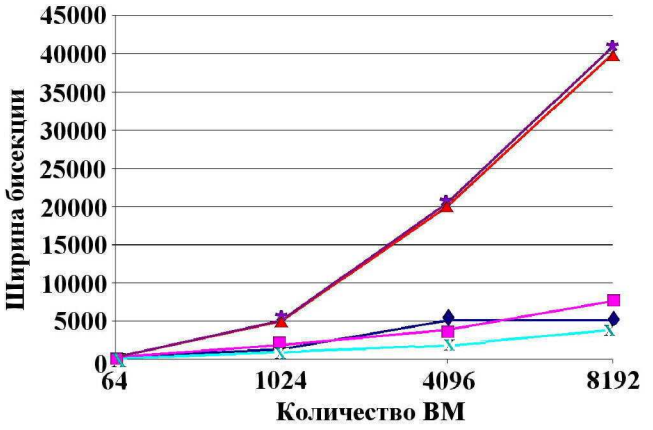

Рис. 6. Зависимость ширины бисекции КС разных топологий от количества ВМ: —♦— — "Ангара", 4D Torus; —■— — СМПО-10G, "Мультитор"; —▲— — СМПО-10G, KNS; —×— — InfiniBand, 3D Torus; —*— — InfiniBand, Fat Tree

сложнее (по числу портов), чем коммутаторные блоки СМПО-10G.

Заключение

Гибридная топология KNS полностью адаптирована к архитектуре СМПО-10G. Для адаптерных и коммутаторных блоков разработаны механизмы идентификации и соответствующие адаптивные маршрутные алгоритмы, обеспечивающие устойчивое к неисправностям функционирование КС.

Сравнение основных технико-архитектурных характеристик популярных топологий, применяемых при создании СМПО суперЭВМ, показало, что КС на базе архитектуры СМПО-10G с топологией KNS имеют характеристики, в основном сравнимые с характеристиками КС, создаваемых на базе архитектуры InfiniBand с топологией Fat Tree.

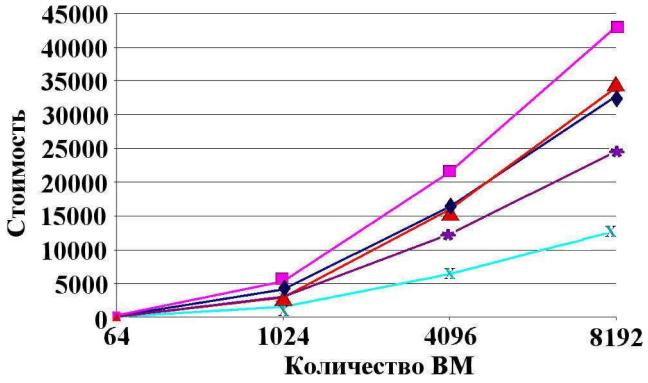

Рис. 7. Зависимость стоимости (количества соединительных кабелей) КС разных топологий от количества ВМ: —♦— — "Ангара", 4D Torus; —■— — СМПО-10G, "Мультитор"; —▲— — СМПО-10G, KNS; —×— — InfiniBand, 3D Torus; —*— — InfiniBand, Fat Tree

Список литературы

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

2. *Горбунов В., Елизаров Г., Эйсымонт Л.* Эксафлопсные суперкомпьютеры: достижения и перспективы // Открытые системы. 2013. № 7. С. 10—14.

3. *Холостов А. А.* Масштабируемая система межпроцессорных обменов 10G // Второй национальный суперкомпьютерный форум. Переславль-Залесский, 26—29 ноября 2013 г.

4. *Басалов В. Г., Козлов Д. О., Холостов А. А.* Топология "Мультитор" для высокопроизводительной и устойчивой к неисправностям коммуникационной сети с архитектурой СМПО-10G // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 2015. Вып. 3. С. 76—84.

5. *Garcna P. J., Escudero-Sahuquillo J., Quiles F. J., Duato J.* High-performance interconnection networks on the road to exascale HPC // Challenges and Solutions. HPC Advisory Council Conference. Barcelona, Spain. September 12, 2013.

6. *Dally W., Towles B.* Principles and Practices of Interconnection Networks. San Francisco: Morgan Kaufmann Publishers, 2004.

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

Статья поступила в редакцию 09.11.15.

THE PROMISING HYBRID TOPOLOGY KNS FOR INTERPROCESSOR COMMUNICATION SYSTEMS ON THE BASIS OF SMPO-10G / V. G. Basalov, A. A. Kholostov (FSUE "RFNC-VNIIEF", Sarov, Nizhny Novgorod region).

The paper describes the hybrid topology KNS for communication networks of multiprocessor systems on the basis of SMPO-10G system for interprocessor communications. The adaptive routing algorithms developed for this topology are described, which allow building the shortest message paths free of interlocking states and provide a high level of fault tolerance of the communication network and its balanced loading. Comparative analysis of the main architectural and technical characteristics of most popular topologies used to create communication networks for high-performance multiprocessor computing systems is presented.

Keywords: multiprocessor computing system, communication networks, interprocessor communication system SMPO-10G, hybrid topology KNS, adaptive routing algorithms, virtual channel.