

УДК 004.722

## ТОПОЛОГИЯ "МУЛЬТИТОР" ДЛЯ ВЫСОКОПРОИЗВОДИТЕЛЬНОЙ И УСТОЙЧИВОЙ К НЕИСПРАВНОСТЯМ КОММУНИКАЦИОННОЙ СЕТИ С АРХИТЕКТУРОЙ СМПО-10G

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

Представлено описание топологии *мультитор* для коммуникационной сети, основанной на системе межпроцессорных обменов СМПО-10G. Описаны созданные для этой топологии адаптивные маршрутные алгоритмы, позволяющие строить кратчайшие, свободные от тупиков маршруты передачи сообщений, а также обеспечивающие высокую отказоустойчивость коммуникационной сети и ее сбалансированную загрузку.

*Ключевые слова:* многопроцессорный вычислительный комплекс, коммуникационная сеть, аппаратный модуль СМПО-10GA-1, адаптерный блок СМПО-10GA-AD, коммутаторный блок СМПО-10GA-SW, адаптивные маршрутные алгоритмы, процессорный узел, виртуальный канал.

### Введение

Коммуникационная сеть (КС) высокопроизводительного многопроцессорного вычислительного комплекса (МВК) представляет собой граф, где вершинами являются вычислительные модули (ВМ), а ребрами — каналы связи. ВМ обычно содержит один или несколько процессоров, локальную память и маршрутный коммутатор. Топология КС в значительной степени определяет эффективность и стоимость самой КС и, как следствие, эффективность всего МВК. В свою очередь, на топологию КС напрямую влияет архитектура ее аппаратных средств.

КС должна передавать максимальное число сообщений за единицу времени с высокой надежностью и стоить как можно меньше. Очевидно, что разработка топологии КС — это всегда компромисс между различными противоречащими друг другу требованиями. Ниже приведены наиболее важные требования, предъявляемые к топологиям КС.

Минимизация диаметра КС и средней дистанции — основные требования к топологии КС. Дистанция — это количество каналов связи, пройденных сообщением между ВМ-источником и ВМ-приемником с использованием кратчайшего пути. Средняя дистанция — усредненное

число каналов связи, пройденных сообщениями для всех возможных пар *ВМ-источник—ВМ-приемник*. Диаметр КС — это максимальная дистанция между двумя ВМ. Уменьшение средней дистанции снижает усредненную коммуникационную задержку и занятость каналов связи и буферов, а следовательно, число конфликтов в КС.

Ширина бисекции — один из важнейших параметров топологии КС. Это показатель, определяемый как минимальное количество каналов связи, которые надо удалить из сети передачи данных для ее разделения на две несвязные области одинакового размера. От ширины бисекции во многом зависит реальная производительность КС в приложениях.

Высокая связность и соответственно устойчивость к неисправностям также важны. Связность (*connectivity*) — показатель, характеризующий наличие разных маршрутов передачи данных между узлами сети; этот показатель может быть определен, например, как минимальное количество каналов связи, которые надо удалить из сети передачи данных для ее разделения на две несвязные области. Устойчивость к неисправностям определяется наличием альтернативных путей для транспортировки сообщений в случае неисправностей коммутаторов или

каналов связи, а также в случае перегруженности каналов связи КС. Большие пакеты сообщений могут быть доставлены с большей скоростью, будучи разделенными на маленькие части и переданными по непересекающимся путям.

Топология КС должна поддерживать простую систему маршрутизации, обеспечивающую построение кратчайших маршрутов, поскольку базовые маршрутные алгоритмы реализованы в аппаратуре.

Немаловажным требованием является низкая стоимость КС. Традиционно стоимость КС определяется количеством каналов связи и количеством коммутаторов. Коммутаторы КС имеют постоянную *степень*, т. е. одинаковое количество портов. Коммутаторы с одинаковым количеством портов позволяют строить КС любого размера и более дешевы в изготовлении, что снижает цену разработки всей КС. От числа портов в коммутаторе зависят связность и средняя дистанция КС. КС с топологией *N-мерный тор* строящаяся на коммутаторах с количеством портов  $2N$ , может быть любого размера, а при топологии *гиперкуб* максимальный размер КС зависит от количества портов коммутаторов: чем больше портов в коммутаторе, тем больший размер может иметь КС. Это требование наиболее конфликтно, так как очень трудно совместимо с другими.

### Архитектура СМПО-10G

Аппаратный модуль СМПО-10GA-1 представляет собой двухкомпонентное электронное устройство, позволяющее, ориентируясь на технологии, доступные в России, использовать его как основной элемент КС высокопроизводительных вычислительных систем. Он состоит из коммутаторного блока СМПО-10GA-SW и адаптерного блока СМПО-10GA-AD [1]. Структурные схемы адаптерного и коммутаторного блоков представлены на рис. 1.

Коммутаторный блок СМПО-10GA-SW предназначен для передачи информации по КС с топологией в виде *N*-мерного тора (*1D*-, *2D*-, *3D*-тор). СМПО-10GA-SW разработан в виде отдельного устройства, включающего 10-портовый полноматричный коммутатор с производительностью каналов до 40 Гбит/с.

Адаптерный блок СМПО-10GA-AD предназначен для связи вычислительного узла с коммутаторным блоком. Наличие в нем 4-портового

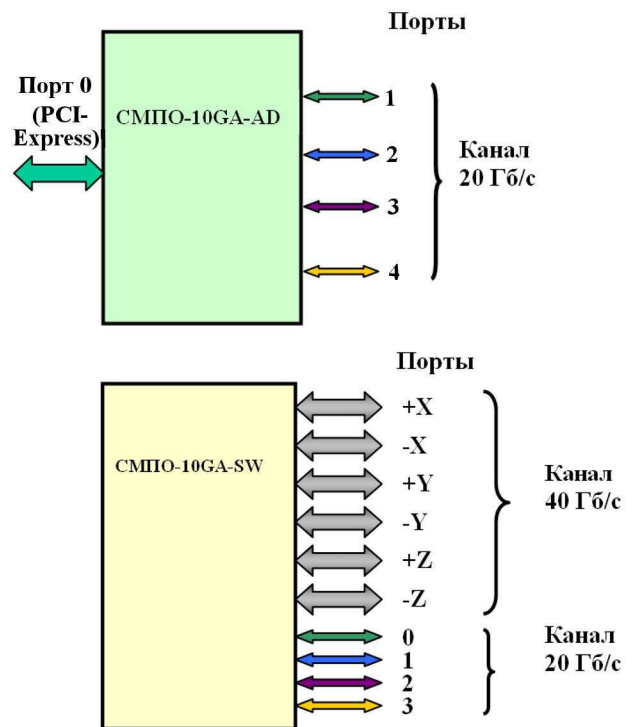


Рис. 1. Структурные схемы компонентов аппаратного модуля СМПО-10GA-1

коммутатора позволяет использовать его в качестве основного коммуникационного элемента для создания небольших систем с топологией КС в виде двумерных торов. Производительность каналов адаптера при реализации на ПЛИС составляет 20 Гбит/с, а при реализации на СВИС — 40 Гбит/с.

В настоящее время каждый порт в адаптерном блоке СМПО-10GA-AD и коммутаторном блоке СМПО-10GA-SW имеет по два виртуальных канала (VC0 и VC1). Виртуальные каналы необходимы для исключения попадания в состояние тупика (deadlock) при передаче сообщений между ВМ вычислительной системы [2].

Особенности архитектуры СМПО-10GA-1 учитывались при разработке топологии КС и маршрутных алгоритмов для достижения максимальной производительности КС и обеспечения ее устойчивости к неисправностям.

### Топология мультитор

Для создания высокопроизводительных и устойчивых к неисправностям КС МВК разработана топология *мультитор*. Базовым элементом этой топологии является процессорный узел.

Структура процессорного узла, состоящего из четырех вычислительных узлов с установленными в них адаптерными блоками СМПО-10GA-AD, объединенных через один уровень коммутации с помощью трех коммутаторных блоков СМПО-10GA-SW, представлена на рис. 2.

Каждый процессорный узел может содержать от одного до четырех коммутаторных блоков СМПО-10GA-SW (количество коммутаторных блоков в процессорном узле ограничено количеством портов в адаптерном блоке), что обеспечивает функционирование от одной до четырех подсетей с топологией *3D-top*. Эти подсети имеют общую производительность по каждому направлению до 160 Гбит/с. Внутри одного процессорного узла коммутаторные блоки СМПО-10GA-SW не имеют непосредственной связи друг с другом. Они обеспечивают передачу сообщений между процессорными узлами, объединенными по топологии *3D-top*. Для связи с коммутаторными блоками разных процессорных узлов каждый коммутаторный блок имеет шесть портов, названных  $+x$ ,  $+y$ ,  $+z$ ,  $-x$ ,  $-y$ ,  $-z$  и

обеспечивающих передачу сообщений в соответствующем названию порта направлении.

К каждому коммутаторному блоку одного процессорного узла может быть подключено до четырех вычислительных узлов с установленными в них адаптерными блоками СМПО-10GA-AD. Порты коммутаторных блоков, предназначенных для связи с адаптерными блоками, имеют номера 0, 1, 2, 3. Порты адаптерных блоков, связанных с коммутаторными блоками, имеют номера 1, 2, 3, 4. Порт адаптерного блока, связанный с ВМ ("PCI-Express"), имеет номер 0 (см. рис. 1). На рис. 2 показано, как должна осуществляться связь между портами коммутаторных и адаптерных блоков в процессорном узле. Порты первого коммутаторного блока, входящего в первую подсеть, подключаются к порту 1 адаптерных блоков всех вычислительных узлов, входящих в процессорный узел. Порты второго коммутаторного блока, входящего во вторую подсеть, подключаются к порту 2 адаптерных блоков всех вычислительных узлов, входящих в процессорный узел. Соответственно порты третьего и четвертого коммутаторных

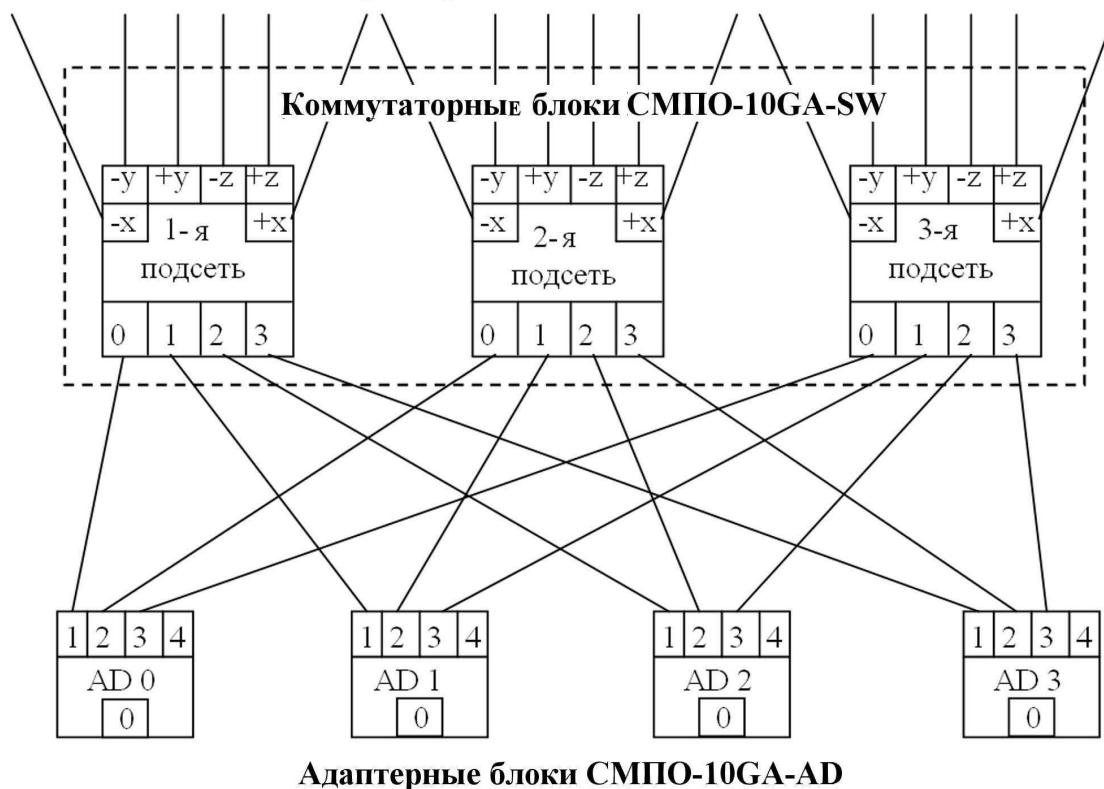


Рис. 2. Схема соединения портов процессорного узла в топологии *мультитор* третьей степени

блоков из третьей и четвертой подсетей подключаются к портам 3 и 4 адаптерных блоков.

На рис. 3 приведен пример КС МВК с топологией *мультитор* третьей степени. Это означает, что топология *мультитор* состоит из трех подсетей. Для того чтобы не загромождать рисунок, обратные связи показаны только в одном месте.

Хотя коммутаторные блоки одного процессорного узла не связаны напрямую, благодаря коммутаторам адаптерных блоков между ними можно передавать сообщения, не занимая топологические порты. Сообщение, передающееся по какой-либо подсети, в случае появления у нее неисправности может сменить подсеть и продолжить движение к получателю по другой, исправной подсети. При этом не требуется усложнения коммутаторных и адаптерных блоков в виде дополнительных виртуальных каналов, так как переход между подсетями только увеличивает длину маршрута на два транзитных участка (hops), но не создает дополнительных циклов (замкнутых путей). Возможность перехода из одной подсети в другую позволяет обойти множественные неисправности оборудования, что повышает надежность КС с топологией *мультитор*.

Размер идентификаторов узлов КС СМПО-10G составляет четыре байта. При использо-

вании топологии *мультитор* коммутаторные и адаптерные блоки имеют разную идентификацию. Форматы этих идентификаторов приведены на рис. 4.

Идентификаторы всех коммутаторных блоков одного процессорного узла идентичны, они занимают три байта (четвертый байт зарезервирован), которые содержат трехмерную координату этого процессорного узла.

Идентификаторы адаптерных блоков занимают четыре байта. Первые три байта определяют координату процессорного узла, к которому относится адаптерный блок, а два младших бита четвертого байта — номер адаптерного блока в процессорном узле.

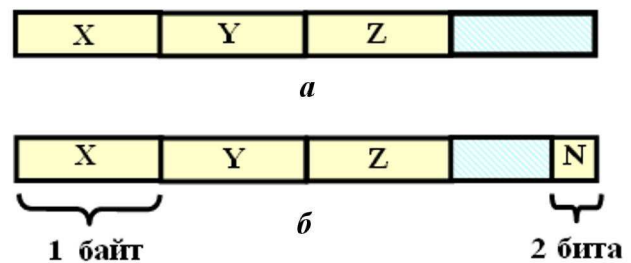


Рис. 4. Форматы идентификаторов коммутаторного (а) и адаптерного (б) блоков аппаратного модуля СМПО-10GA-1

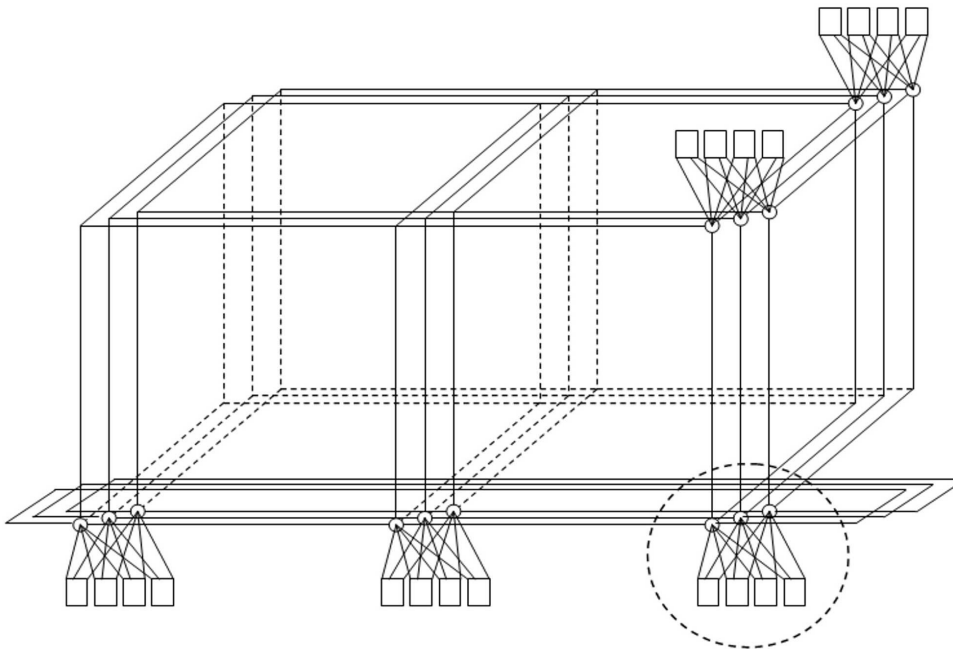


Рис. 3. Топология *мультитор* третьей степени на основе архитектуры СМПО-10GA-1

## Адаптивный метод выбора оптимального маршрута сообщения

Идея этого метода заключается в том, что вычисление оптимального выходного порта для передачи каждого транзитного информационного сообщения осуществляется коммутатором непосредственно в момент передачи [3].

Наиболее перспективным направлением при создании маршрутных алгоритмов для КС СМПО-10G с топологией *мультитор* является объединение возможностей двух механизмов: детерминированного при полностью исправной коммуникационной среде в виде  $n$ -мерного тора и адаптивного при возникновении неисправностей. Это позволит наиболее полно использовать сильные стороны каждого механизма и увеличить надежность и производительность КС с топологией *мультитор*, создаваемых на базе СМПО-10G.

Разработка маршрутных алгоритмов для архитектуры СМПО-10G осложняется ее особенностями, в частности тем, что аппаратный модуль СМПО-10GA-1 представляет собой двухкомпонентное электронное устройство. Маршрутные алгоритмы для разных устройств должны совместно обеспечивать правильную и надежную работу всей КС.

### Адаптивный маршрутный алгоритм коммутатора адаптерного блока

В адаптерном блоке для разных портов коммутатора разработаны разные маршрутные алгоритмы. Это сделано с целью увеличения пропускной способности и повышения надежности КС в целом.

Отправка пакетов сообщений, источником которых является ВМ, т. е. порт "PCI-Express" адаптера, осуществляется циклически через все функционирующие порты адаптера; при этом всегда выбирается виртуальный канал VC0. В результате пакеты равномерно распределяются по коммуникационным подсетям, обеспечивая сбалансированную загрузку всей КС.

Для принятия решения о маршрутизации пакета в портах 1, 2, 3, 4 адаптер сравнивает идентификатор получателя  $\{X_D, Y_D, Z_D, N_D\}$  с идентификатором текущей адаптерной платы  $\{X_C, Y_C, Z_C, N_C\}$ . Сравнение производится вычитанием значений полей  $\{X_D, Y_D, Z_D, N_D\}$  из значений полей  $\{X_C, Y_C, Z_C, N_C\}$ .

Если  $\{X_D, Y_D, Z_D, N_D\}$  и  $\{X_C, Y_C, Z_C, N_C\}$  совпали, пакет передается в порт "PCI-Express", а затем в узел. В противном случае коммутатор адаптерной платы выдает пакет в порт со смещением 1 от номера текущего порта (например: текущий порт 1, искомый — 2; текущий порт 2, искомый — 3; текущий порт 4, искомый — 1). Если найденный выходной порт не способен передавать пакеты, то коммутатор адаптерной платы выдает пакет в порт со смещением 2 от номера текущего порта (например: текущий порт 1, искомый — 3; текущий порт 2, искомый — 4; текущий порт 3, искомый — 1). Если и этот порт нерабочий, то делается вывод о неисправности сети. Для передачи пакета используется виртуальный канал VC1.

### Адаптивный маршрутный алгоритм коммутатора коммутаторного блока

Адаптивный маршрутный алгоритм коммутатора коммутаторного блока одинаков для всех его портов.

Для принятия решения о маршрутизации пакета коммутатор коммутаторной платы (СМПО-10GA-SW) сравнивает идентификатор получателя  $\{X_D, Y_D, Z_D\}$  с идентификатором текущей коммутаторной платы  $\{X_C, Y_C, Z_C\}$ . Сравнение производится вычитанием значений полей  $\{X_D, Y_D, Z_D\}$  из значений полей  $\{X_C, Y_C, Z_C\}$ .

Если  $\{X_D, Y_D, Z_D\}$  и  $\{X_C, Y_C, Z_C\}$  совпали, пакет передается в порт с номером  $N_D$  и виртуальным каналом VC0. Если передача пакета в порт с номером  $N_D$  невозможна, то выбирается первый способный к передаче пакетов порт с виртуальным каналом VC1, ведущий в адаптерные блоки (порты 1, 2, 3, 4). Если такого нет, то делается вывод о неисправности сети.

Если  $\{X_D, Y_D, Z_D\}$  и  $\{X_C, Y_C, Z_C\}$  не совпали, для передачи пакета выбирается оптимальный выходной порт и виртуальный канал, ведущий в следующий коммутаторный блок. В случае, если передача пакета через выбранный выходной порт невозможна, выбирается первый способный к передаче пакетов порт с виртуальным каналом VC1, ведущий в адаптерные блоки (порты 1, 2, 3, 4). Если такого нет, то делается вывод о неисправности сети.

Топологии *мультитор* присуще наличие большого количества замкнутых участков сети (циклов), приводящих к возможности возникновения состояний *взаимных блокировок* при

обмене сообщениями. Выбор оптимального выходного порта в алгоритме коммутатора коммутаторного блока осуществляется в соответствии с хорошо известным алгоритмом DOR (Dimension Order Routing), или, как его еще называют, XY-routing алгоритмом [4]. Изменение измерений при прохождении пакетом сообщения маршрута происходит в фиксированном порядке: пакет сначала направляется до соответствующей координаты в  $X$ -измерении в положительном или отрицательном направлении, затем обрабатывается измерение  $Y$  и в конце — измерение  $Z$ .

Использование имеющегося разделения каждого физического канала на два виртуальных канала со своими буферными пространствами позволяет исключить в топологии *мультитор* циклы, образованные обратной связью. Применение двух виртуальных каналов разделяет коммуникационную среду на две виртуальных КС с ациклическими графами зависимостей каналов. Если в маршруте между источником и приемником пакета используется обратная связь, то пакет передается по виртуальному каналу VC0, если обратная связь не задействована — по VC1. Смена виртуальных каналов происходит при прохождении сообщением коммутатора с нулевым значением координаты в каждом измерении топологии *мультитор*.

Граф зависимостей каналов, получающийся в результате работы маршрутного алгоритма коммутатора коммутаторного блока, не содержит циклов. Поэтому возможность возникновения состояний взаимных блокировок исключается.

Высокая отказоустойчивость обеспечивается обходом неисправных участков КС через коммутаторы адаптерных блоков со сменой подсети. При этом увеличивается длина маршрута, но не образуются дополнительных циклов.

### Характеристики КС с топологией *мультитор*

Количество процессорных узлов в МВК, использующем КС с топологией *мультитор*, определяется по формуле

$$NPU = \prod_{i=1}^k z_i,$$

где  $k$  — количество измерений (размерность) тора,  $1 \leq k \leq 3$ ;  $z_i$  — количество процессорных узлов в  $i$ -м измерении тора.

МВК содержит  $N$  ВМ и соответственно  $N$  адаптерных блоков СМПО-10GA-AD, количество которых определяется по формуле

$$N = m \cdot NPU,$$

где  $m$  — количество ВМ в каждом процессорном узле,  $m \leq 4$ .

Количество задействованных при этом коммутаторных блоков СМПО-10GA-SW можно определить по формуле

$$SW = s \cdot NPU,$$

где  $s$  — степень топологии *мультитор*,  $1 \leq s \leq 4$ .

Поскольку топология *мультитор* образована на основе трехмерного тора, уместно предположить, что главной составляющей диаметра КС будет диаметр тора. К этому диаметру необходимо добавить два транзитных канала связи: первый — от адаптерного блока источника сообщения до смежного коммутаторного блока, второй — от коммутаторного блока до адаптерного блока приемника сообщения. Полный диаметр КС с топологией *мультитор* определяется по формуле

$$D = \sum_{i=1}^k \left( \frac{z_i}{2} \right) + 2.$$

Стоит также отметить, что минимальная дистанция в топологии *мультитор* равна 2.

Значение связности для топологии *мультитор* равно  $s$ . Оно определяется количеством связей между адаптерным блоком СМПО-10GA-AD ВМ и коммутаторными блоками СМПО-10GA-SW одного процессорного узла.

Ширина бисекции для обычного трехмерного тора определяется по формуле

$$BW_{tor} = 2 \min(z_1 z_2, z_1 z_3, z_2 z_3).$$

В топологии *мультитор* между процессорными узлами используется по  $s$  связей, поэтому ширина бисекции для них будет больше в  $s$  раз:

$$BW = s \cdot BW_{tor}.$$

Легко заметить, что ширина бисекции в топологии *мультитор* увеличивается с увеличением ее степени.

Стоимость — показатель, который может быть определен, например, как общее количество линий передачи данных в МВК. Стоимость обычного тора можно оценить по формуле

$$COST_{tor} = k \prod_{i=1}^k z_i.$$

Для вычисления стоимости топологии *мульти-тор* необходимо  $COST_{tor}$  увеличить в  $s$  раз — используемую степень данной топологии — и прибавить стоимость древовидной составляющей (с количеством связей между адаптерными и коммутаторными блоками всех процессорных узлов), вычисляемую по формуле

$$COST_{tree} = ms \prod_{i=1}^k z_i.$$

В итоге стоимость топологии *мульти-тор* равна

$$COST = s \cdot COST_{tor} + COST_{tree}.$$

Отсюда следует, что стоимость топологии *мульти-тор* увеличивается с увеличением ее степени.

Устойчивость к неисправностям — способность КС функционировать при наличии одной или более неисправностей. Система маршрутизации КС с топологией *мульти-тор* неустойчива к неисправностям, когда ее степень равна 1. Устойчивость к неисправностям повышается с увеличением степени  $s$  топологии *мульти-тор*. Это достигается за счет увеличения количества альтернативных маршрутов и реализации адаптивных маршрутных алгоритмов, обеспечивающих обход неисправных участков.

### Моделирование топологий КС

Авторами проведено моделирование двух топологий КС: *4D-тор* (используемой при создании высокоскоростной сети "Ангара" [5], разработанной в ОАО "НИЦЭВТ", г. Москва) и *мульти-тор* с разным количеством подсетей, от одной до трех. Количество ВМ в моделируемых вычислительных системах достигает 4096 и ограничено возможностями используемого программного обеспечения и оборудования. Моделирование проводилось с использованием открытой программы Booksim [4], дополненной топологией *мульти-тор* и соответствующими маршрутными алгоритмами. Программа Booksim разработана в Стэнфордском технологическом университете и представляет собой консольное приложение для моделирования трафика вычислительных сетей.

На основе входных параметров Booksim выводит информацию о средней коммуникационной задержке доставки пакета сообщения в сети и среднем количестве транзитных участков.

К входным параметрам моделирования относятся: параметры расширения топологии; функция трафика, задающая распределение для генерации номера источника и приемника для пакетов; функция маршрутизации; интенсивность внедрения информационного потока. Средняя коммуникационная задержка измеряется циклами (итерациями) выполнения программы Booksim. На одном цикле происходит создание пакетов, внедрение их в сеть, продвижение через выходные каналы коммутаторов и, если пакеты были доставлены до адресата, их извлечение из сети. Коммуникационной задержкой одного пакета сообщения считается количество циклов выполнения программы, которое данный пакет находился в сети, с момента его создания до момента его извлечения. Средняя коммуникационная задержка получается путем вычисления среднего арифметического задержек всех сгенерированных пакетов. По аналогии рассчитывается среднее количество транзитных участков.

В ходе каждого эксперимента авторы изменяли параметр расширения топологии, а именно параметр  $K$  — количество узлов сети в каждом измерении топологий *4D-тор* и *мульти-тор*, и получали результаты моделирования для сетей с разным количеством ВМ. Неизменными оставались параметры функции маршрутизации и трафика. Интенсивность внедрения пакетов в сеть была выбрана на уровне 50%. Интенсивность внедрения пакетов — это параметр, определяющий количество пакетов, которые будут создаваться на одном цикле моделирования. Если значение данного параметра равно 50%, это означает, что на одном цикле моделирования число созданных пакетов составит ровно половину от количества ВМ в МВК. Функция маршрутизации представлена алгоритмом Dimension Order Routing [6], а функция трафика — нормальным распределением для каждого генерируемого пакета.

На рис. 5 показаны результаты моделирования средней дистанции для топологий *4D-тор* и *мульти-тор* в зависимости от числа ВМ в МВК\*. Средняя дистанция для топологии *мульти-тор* при разном числе подсетей одинакова, и, как видно из рисунка, она больше, чем средняя дистанция для топологии *4D-тор*. Стоит отметить, что результаты моделирования средней дистан-

\* При моделировании топологии *4D-тор* использовалась модель, в которой коммутатор и адаптер ВМ совмещены, т. е. минимальная дистанция равна 1.

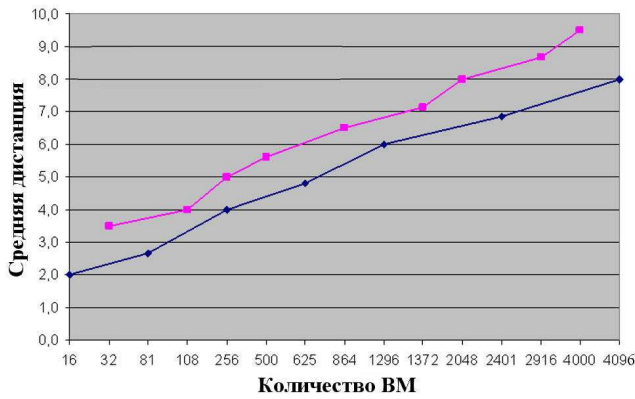


Рис. 5. Зависимости средней дистанции от количества узлов для топологий *4D-тор* (—◆—) и *мультистор* с одной подсетью (—■—)

ции для этих топологий полностью совпали с расчетами, проведенными по формуле из [4]:

$$H_{\min} = \frac{1}{N^2} \sum_{x,y \in N} H(x,y),$$

где  $N$  — количество VM в МВК;  $H(x,y)$  — дистанция между всеми парами VM в МВК.

На рис. 6 (см. также цветную вкладку) приведены результаты моделирования зависимости средней коммуникационной задержки от количества VM МВК для топологий *4D-тор* и *мультистор* с одной, двумя и тремя подсетями. Из рисунка видно, что средняя задержка для топологии *4D-тор* меньше, чем для топологии *мультистор* с одной подсетью, но больше, чем для топологии *мультистор* с двумя и тремя подсетями. Также можно заметить, что при интенсивности внедрения пакетов в сеть на уровне 50% применение топологии *мультистор* с тремя подсетями практически не уменьшает средней задержки по сравнению с топологией *мультистор* с двумя подсетями.

Средняя задержка доставки пакета для топологии *мультистор* с тремя подсетями становится заметно меньше по сравнению с топологией *мультистор* с двумя подсетями при большей загрузке КС. Это хорошо видно из рис. 7, который отображает результаты моделирования при интенсивности внедрения пакетов на уровне 99%.

### Заключение

Разработаны новая топология *мультистор*, а также адаптивные маршрутные алгоритмы ком-

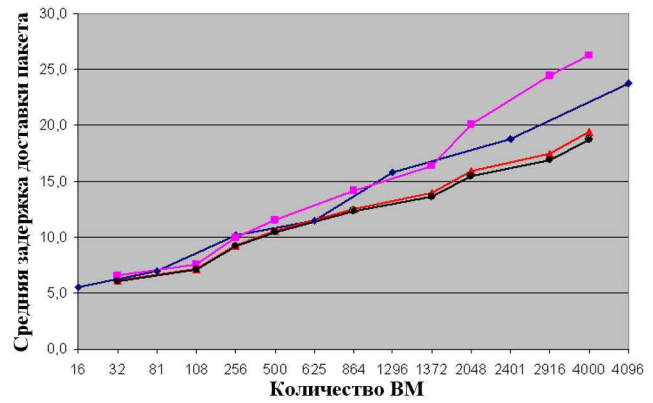


Рис. 6. Зависимости средней коммуникационной задержки от количества узлов для топологий *4D-тор* (—◆—) и *мультистор* с одной (—■—), двумя (—▲—) и тремя (—●—) подсетями

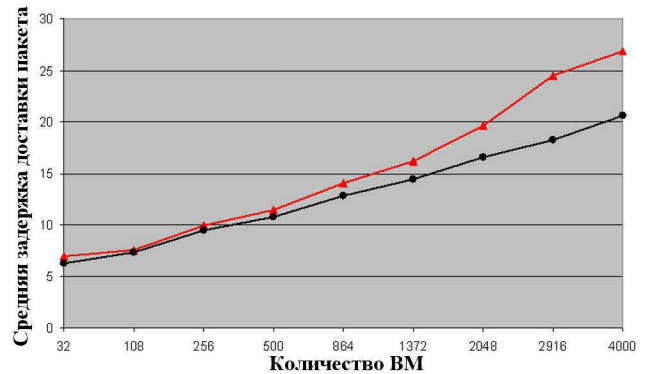


Рис. 7. Зависимости средней коммуникационной задержки от количества узлов для топологий *мультистор* с двумя (—▲—) и тремя (—●—) подсетями при интенсивности внедрения пакетов  $\sim 99\%$

мутаторов адаптерных и коммутаторных блоков для КС на базе системы межпроцессорных обменов СМПО-10G, которые обеспечивают ее устойчивое к неисправностям функционирование.

На базе КС СМПО-10G с топологией *мультистор* третьей степени и описанными в статье маршрутными алгоритмами создан МВК МП-СМПО-3D, объединяющий 64 VM (768 ядер). Полученная производительность 7,5 Тфлопс на тесте Linpack составила 79,4% относительно пиковой производительности 9,4 Тфлопс.

К достоинствам предложенной топологии *мультистор* и разработанных для нее маршрутных алгоритмов относятся:

- устойчивость (при  $s > 1$ ) к деградации КС;
- хорошая геометрическая масштабируемость МВК, имеющих КС с данной топологией;



- увеличение производительности КС за счет передачи пакетов данных от источника к приемнику параллельно по  $s > 1$  независимым путям.

Недостаток топологии *мультитор* традиционен для тороидальных топологий: это большой диаметр КС и небольшая ширина бисекции.

Дальнейшее направление работ — разработка и применение для аппаратного модуля СМПО-10GA-1 новых топологий и маршрутных алгоритмов, которые позволят строить более эффективные КС.

### Список литературы

1. Холостов А. А. Масштабируемая система межпроцессорных обменов 10G // Второй национальный суперкомпьютерный форум. г. Переславль-Залесский, 26–29 ноября 2013 г.
2. Duato J. A theory of fault-tolerant routing in wormhole networks // IEEE Trans. Parallel and Distributed System. 1997. Vol. 8, No 8. P. 790–802.
3. 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.
4. Dally W., Towles B. Principles and Practices of Interconnection Networks. San Francisco: Morgan Kaufmann Publishers, 2004.
5. Жабин И., Макагон Д., Симонов А. и др. Кристалл для "Ангары" // Суперкомпьютеры. 2013. № 4 (16). С. 46–49.
6. Khan M. A., Ansari A. Q. A Quadrant-XYZ Routing Algorithm for 3-D Asymmetric Torus Network-on-Chip. The Research Bulletin of Jordan ACM. 2011. Vol. II. <http://www.ijj.acm.org/volum2/no2/ijjvol2no2p2.pdf>

Статья поступила в редакцию 11.12.14.

MULTITORUS TOPOLOGY FOR HIGH PERFORMANCE AND FAULT-SAFE COMMUNICATION NETWORK WITH THE SMPO-10G ARCHITECTURE / V. G. Basalov, D. O. Kozlov, A. A. Kholostov (FSUE RFNC-VNIIEF, Sarov, Nizhny Novgorod region).

The paper considers the multitorus topology for the communication network based on the system of interprocessor exchanges SMPO-10G. The paper describes the adaptive routing algorithms developed specially for this topology, which enable building the shortest and free-from-dead-ends message transmission routs, as well as providing high fault tolerance of the communication network and its balanced load.

*Keywords:* multiprocessor computation complex, communication network, SMPO-10GA-1 hardware module, СМПО-10GA-AD adapter unit, СМПО-10GA-SW switch unit, adaptive routing algorithms, processor node, virtual channel.

---