Топология MKNS для коммуникационных сетей вычислительных систем

В. Г. Басалов, Ю. А. Малых

Федеральное государственное унитарное предприятие
«Российский федеральный ядерный центр - Всероссийский научно-исследовательский институт экспериментальной физики» (ФГУП «РФЯЦ-ВНИИЭФ»), г. Саров, Россия

Аннотация. В высокопроизводительных параллельных вычислительных системах, называемых суперкомпьютерами, коммуникационная сеть играет ключевую роль в производительности и стоимости вычислительной системы. Эффективность и стоимость коммуникационной сети в значительной степени определяется сетевой топологией. В этой статье представлено описание топологии MKNS, разработанной для коммуникационных сетей параллельных вычислительных систем эксафлопсной производительности. Приведен сравнительный анализ основных коммуникационных характеристик топологий MKNS, 4D-Torus, 6D-Torus и Fat Tree, используемых при создании коммуникационных сетей параллельных вычислительных систем.

Ключевые слова: коммуникационная сеть, вычислительная система, топология MKNS.

DOI 10.14357/20718632210302

Введение

Компьютерное моделирование является одним из основных средств исследований в различных областях естествознания (физика высоких энергий, астрофизика, метеорология, конструирование новых лекарств, сложные инженерные проекты, цифровые устройства или «цифровые двойники», искусственный интеллект и д.р.). Сложность этих исследований требует применения высокопроизводительной вычислительной техники, в частности параллельных вычислительных систем, производительность самых мощных из которых приближается к эксафлопсному уровню. Производительность параллельной вычислительной системы на задачах с интенсивным обменом данными между вычислительными модулями (ВМ) (задачи моделирования, задачи на графах и нерегулярных сетках, вычисления с использованием разреженных матриц) во многом определяется параметрами коммуникационной сети (КС). Ключевым параметром КС, определяющим её эффективность, является топология связей[1]. Очевидно, что разработка топологии КС - это всегда компромисс между различными противоречащими друг другу требованиями.

Минимизация такой коммуникационной характеристики, как диаметр - основное требование к топологии КС. Дистанция - это количество каналов связи, пройденных сообщением между источником и приемником с использованием кратчайшего пути. Диаметр топологии КС - это максимальная дистанция между двумя вычислительными узлами, содержащими несколько ВМ с несколькими процессорами. Уменьшение диаметра снижает коммуникационную задержку и занятость каналов связи и буферов, а, следовательно, число конфликтов в КС.

Операция широковещания (один – многим) это операция, разработанная для упрощения и улучшения эффективности взаимодействия, между группой ВМ. Широковещательный маршрут представляет собой остовное дерево, построенное на графе текущей топологии КС. Скорость выполнения операции широковещания напрямую зависит от такой коммуникационной характеристики, как диаметр остовного дерева.

Ширина бисекции - одна из важнейших коммуникационных характеристик топологии КС, определяемая как минимальное количество двунаправленных каналов связи, которое надо удалить из КС для её разделения на две несвязные области одинакового размера. Максимизация ширины бисекции максимизирует бисекционную (глобальную) пропускную способность КС.

Связность - коммуникационная характеристика топологии КС, определенная как минимальное количество каналов связи, которое надо удалить из КС для её разделения на две несвязные области. Увеличение связности увеличивает количество альтернативных путей для транспортировки сообщений в КС, что повышает устойчивость к неисправностям.

Немаловажным требованием к топологии КС является уменьшение такой коммуникационной характеристики, как стоимость КС. Традиционно стоимость КС определяется количеством кабелей связи, а также количеством и сложностью коммутаторов, выражаемую через количество портов. Это требование наиболее конфликтно, так как очень трудно совместимо с другими требованиями.

Вычислительная система содержит $N = N_{BM}/m$ вычислительных узлов, где N_{BM} – количество ВМ в вычислительной системе, m – количество ВМ в вычислительном узле для КС с топологией Fat Tree $m = 1$.

В Табл. 1 приведены формулы вычисления численных значений коммуникационных характеристик для топологий nD-Top и Fat Tree, где n – количество измерений в тороидальной топологии, ki – количество вычислительных узлов в i-ом измерений тора, d - количество сетевых портов в коммутаторах коммуникационной сети.

Коммуникационное оборудование современных КС можно разделить на две большие группы: коммутаторы с «большим» количеством сетевых портов – больше 36 («high radix» (HR) коммутаторы) и коммутаторы с «небольшим» количеством сетевых портов – 15 и меньше («low radix» (LR) коммутаторы). К первой группе относятся коммутаторы Infiniband [2] с 36- и 40-сетевыми портами, разработанные компанией Mellanox, и недавно разработанный коммутатор Rosetta [3] с 64-сетевыми портами, разработанный фирмой Cray. Ко второй группе относятся 10-портовые коммутаторы IBM для вычислительных систем серии BlueGene/Q [4], 10-портовые коммутаторы КС TofuD для суперкомпьютера Fugaku фирмы Fujitsu [5], 8-портовые коммутаторы КС «Ангара» [6] разработки АО «НИЦЭВТ» и 10-портовые коммутаторы КС СМПО-10С [7] разработки РФЯЦ-ВНИИЭФ.

Анализ КС вычислительных систем входящих в первую десятку TOP 500 и используемого в них коммуникационного оборудования

Табл. 1. Коммуникационные характеристики

Топология	nD-Top	Fat Tree
Диаметр	$D = \sum_{i=1}^{n} \frac{ki}{2}$	$D = 2\,log_{d/2}\,N$
Диаметр остовного дерева	$D_{o.д.} = \sum_{i=1}^{n}(ki - 1)$	$D_{o.д.} = D$
Ширина бисекции	$B = 2N/\,max_i(ki)$	$B = N/2$
Связность	$C = 2n$	$C = d/2$
Количество кабелей связи	$NCC = n* \prod_{i=1}^{n} ki$	$NCC = N\,log_{d/2}\,N$
Количество портов	$NP = 2nN$	$NP = d\,\frac{N}{d/2}\,log_{d/2}\,N$

показал, что использование LR коммутаторов в вычислительных системах стремительно снижается. В 2015 году больше половины вычислительных систем, входящих в первую десятку TOP 500, имели в составе коммуникационного оборудования LR коммутаторы, а в 2020-м только одна из десяти вычислительных систем создана с использованием LR коммутаторов. Стоит отметить, что из первых пятидесяти вычислительных систем из списка TOP-500 за июль 2020 только четыре имеют КС, построенные с использованием LR коммутаторов[8].

КС с коммуникационным оборудованием на LR коммутаторах ориентированы, как правило, на использование топологий многомерных торов. КС вычислительной системы Fugaku имеет топологию TOFU (модифицированный 6D-Top) [5], КС вычислительных систем серии BlueGene/Q имеет топологию 5D-Top [4], КС с коммуникационным оборудованием «Ангара» имеет топологию 4D-Top [6], а КС с обновлённым коммуникационным оборудованием «Ангара 2» ориентирована на топологию 6D-Top [9].

Применение в КС коммуникационного оборудования с HR коммутаторами позволяет значительно расширить список используемых топологий. Топологии Fat Tree, Dragonfly, HyperX разработаны с ориентацией на использование HR коммутаторов. Применение этих топологий улучшает коммуникационные характеристики КС в сравнении с тороидальными топологиями.

Например, для вычислительной системы, содержащей примерно от 32000 до 40000 вычислительных узлов, диаметр КС с топологией 5D-Torus будет равен примерно 20 хопам (хоп - это участок КС, расположенный между двумя коммуникационными устройствами), диаметр КС с топологией Fat Tree будет равен 8 хопам, лучшие вычислительные системы Китая имеют диаметр КС в 7 хопов [10], диаметр КС с коммутаторами Aries и топологией Dragonfly будет равен 5 хопам [11]. КС с коммутаторами Rosetta и топологиями Dragonfly и HyperX имеют диаметр в 3 хопа [12].

Гибридная топология k-ary n-direct s-indirect (KNS), предложенная испанскими учеными для реализации вычислительных систем эксафлопсной производительности, представляет собой попытку добиться значительного уменьшения диаметра больших КС, стоимости, энергозатрат при обеспечении высокой пропускной способности[13].

Сравнение коммуникационных характеристик топологий [14], показало, что КС с коммуникационным оборудованием с LR коммутаторами и топологией k-ary n-direct s-indirect имеют характеристики, в основном, сравнимые с характеристиками КС, создаваемых на базе коммуникационного оборудования с HR коммутаторами и топологией Fat Tree.

Однако в процессе разработки алгоритмов маршрутизации для КС с топологией k-ary n-direct s-indirect были выявлены следующие недостатки этой топологии:

- главным недостатком является то, что КС с топологией k-ary n-direct s-indirect, имеют низкую связность, равную количеству измерении n. Например, у КС с двумерной топологией k-ary n-direct s-indirect значение связности равно двум. В случае отказа любого канала связи в КС возникает «висячая вершина». Алгоритм обхода таких неисправностей, предложенный в работе [13], довольно сложен и труднореализуем на практике. Увеличение связности при фиксированном количестве вычислительных узлов достигается за счет увеличения числа измерений в КС, что ведет к увеличению диаметра КС;
- минимальная дистанция в КС с топологией k-ary n-direct s-indirect равна двум хопам, а значит, сообщение проходит минимум три коммуникационных устройства, что увеличивает среднюю дистанцию и задержку при передаче сообщений;
- при конкурентоспособных коммуникационных характеристиках КС с топологией k-ary n-direct s-indirect имеет довольно высокую стоимость, выраженную в количестве кабелей связи и коммутаторов.

В этой статье представлена модернизированная топология MKNS, разработанная для КС с коммуникационным оборудованием на LR коммутаторах для вычислительных систем эксафлопсной производительности.

Топология MKNS

Название топологии MKNS означает «модернизированная KNS». Одновременно «M»

в названии определяет количество ВМ в вычислительном узле, K –количество вычислительных узлов в измерениях топологии MKNS. В первом измерении значение максимального количества вычислительных узлов равно d-2 (d - количество сетевых портов в коммутаторах коммуникационной сети). В остальных измерениях количество вычислительных узлов неограниченно. В разных измерениях количество вычислительных узлов может быть разным. Нумерация вычислительных узлов и соответствующих адаптерных блоков в каждом измерении начинается с 0. N определяет количество используемых измерений (координатных направлений) в топологии MKNS (до четырех измерений), S –количество уровней в непрямом участке топологии MKNS: один уровень при использовании коммутаторных блоков, два - при использовании коммутаторной сборки. Каждый вычислительный узел идентифицируется его координатой в КС.

На Рис. 1 представлена топология MKNS при M = 2, K = 4, N = 2 и S = 1. Коммуникационное оборудование, включающее адаптерные

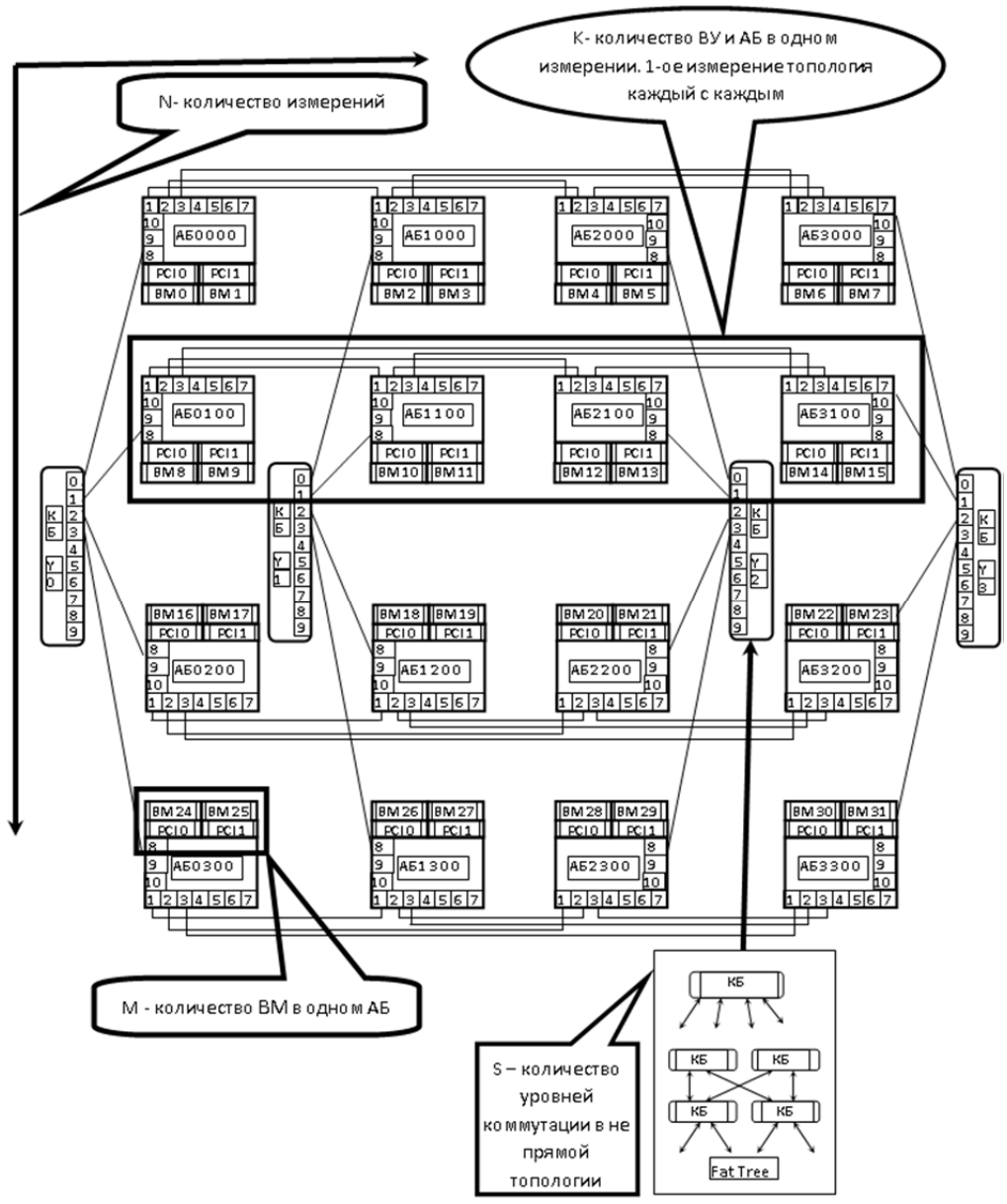

Рис. 1. Блок-схема топологи MKNS(2, 4, 2, 1)

блоки и коммутаторные блоки, выполнено с использованием LR коммутаторов с 10 сетевыми портами.

КС с топологией MKNS организуется таким способом, при котором вычислительные узлы, включающие адаптерный блок (АБ) и один или несколько ВМ с арифметическими процессорами, располагаются ортогонально в нескольких измерениях, подобно топологиям nD-Top. Связь между вычислительными узлами, находящимися в первом измерении, происходит с помощью АБ, объединенных двунаправленными каналами связи по полносвязной топологии (каждый с каждым) без использования коммутаторных блоков (КБ). В других измерениях вычислительные узлы и соответствующие им АБ объединяются с помощью КБ или с помощью коммутаторных сборок (КСб), представляющих собой набор КБ, объединенных по топологии Fat Tree, как показано на Рис. 2.

Каждый АБ содержит полноматричный коммутатор, обеспечивающий коммутацию m контроллеров PCI Express и d сетевых портов. АБ предназначен для реализации интерфейса между ВМ, входящими в вычислительный узел и КС. Также АБ обеспечивает смену измерений при передаче транзитных сообщений между вычислительными узлами.

Каждый КБ содержит полноматричный коммутатор, обеспечивающий коммутацию d сетевых портов. КБ обеспечивает транзит сообщений между вычислительными узлами, находящимися в одном измерении.

Для организации в первом измерении полносвязной топологии задействованы с первого по d-3 сетевые порты каждого АБ, поэтому максимальное количество вычислительных узлов в этом измерении равно d-2. Для организации передачи информационных пакетов между АБ вычислительных узлов, расположенных в первом измерении, используется (d-3)×(d-2)/2 двунаправленных каналов связи, подключенных определенным образом.

Порт с номером d-2 каждого АБ используется для связи с соответствующим КБ или КСб, которые обеспечивают передачу сообщений во втором измерении топологии MKNS. Порт с номером d-1 каждого АБ используется для связи с соответствующим КБ или КСб, которые обеспечивают передачу сообщений в третьем измерении топологии MKNS. Порт с номером d каждого АБ используется для связи с соответствующим КБ или КСб, которые обеспечивают передачу сообщений в четвёртом измерении топологии MKNS.

В одном измерении топологии MKNS могут быть использованы либо КБ, либо КСб. В разных измерениях топологии MKNS могут использоваться как КБ, так и КСб.

Передача информации между ВМ, подключенными к одному вычислительному узлу, означает передачу информации только через АБ этого вычислительного узла. Например, передача информации из ВМ 25 в ВМ 24 осуществляется следующим образом: из ВМ 25 через соответствующий порт PCI 1 информация передается в АБ (0 3 0 0) и далее через порт PCI 0 в ВМ 24.

Передача информации между ВМ, подключенными к разным вычислительным узлам,

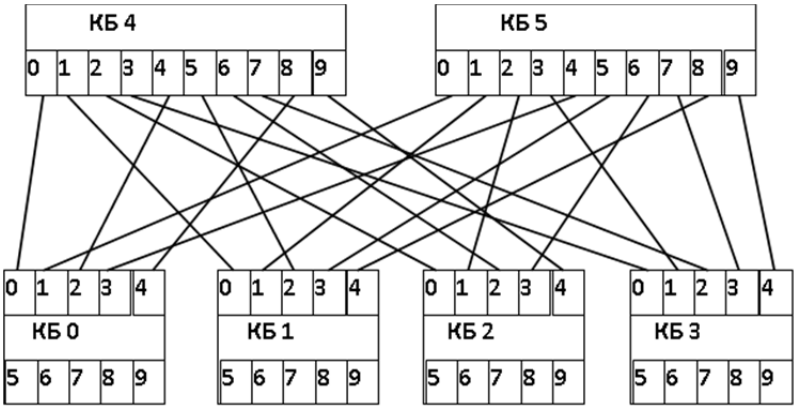

Рис. 2. Вариант реализации двухуровневой КСб по топологии Fat Tree

значения координат АБ которых отличаются только в первом значении, происходит следующим образом: информация из ВМ источника одного вычислительного узла через соответствующий порт PCI0 или PCI1 передается в его АБ, затем через соответствующий сетевой порт этого АБ информация передается в соответствующий сетевой порт АБ другого вычислительного узла, а затем через порт PCI0 или PCI1 в ВМ приемник. Например, передача информации из ВМ 0 в ВМ 5 осуществляется следующим образом: из ВМ 0 через порт PCI 0 информация передается в АБ (0 0 0 0), далее через сетевой порт 2 информация передается в сетевой порт 1 АБ (2 0 0 0) и далее через порт PCI 1 в ВМ 5.

Передача информации между ВМ, подключенными к разным вычислительным узлам, значения координат АБ которых отличаются либо только во втором, либо только в третьем, либо только в четвёртом значении, происходит следующим образом: информация из ВМ источника через соответствующий порт PCI0 или PCI1 передается в АБ, затем через соответствующий данному измерению сетевой порт этого АБ информация передается в соответствующий сетевой порт КБ, из которого через сетевой порт передается в сетевой порт другого АБ, а затем через порт PCI0 или PCI1 в ВМ приемник. Например, передача информации из ВМ 2 в ВМ 19, координаты АБ которых отличаются только во втором измерении, осуществляется следующим образом: из ВМ 2 через порт PCI 0 информация передается в АБ (1 0 0 0), далее через сетевой порт 8 информация передается в сетевой порт 0 КБ Y1, далее через его сетевой порт 2 информация передается в сетевой порт 8 АБ (1 2 0 0) и затем через порт PCI 1 в ВМ 19.

Передача информации между ВМ, подключенными к разным вычислительным узлам, значения координат АБ которых отличаются в нескольких значениях, происходит следующим образом: информация из ВМ источника через порт PCI0 или PCI1 передается в АБ, в котором происходит сравнение координат текущего АБ и АБ приемника. Сравнение начинается с первого по порядку измерения. Если не совпали значения координат в первом измерении, то через соответствующий сетевой порт этого АБ информация передается в соответствующий сетевой порт АБ другого ВУ. Иначе, если не совпали значения координат для второго, или третьего, или четвертого измерения, то через соответствующий данному измерению сетевой порт этого АБ информация передается в соответствующий сетевой порт соответствующего КБ, из которого через соответствующий сетевой порт передается в соответствующий сетевой порт АБ другого ВУ. В достигнутом АБ опять происходит процесс сравнения координат уже этого текущего АБ и АБ приемника. Процесс передачи информации от АБ к АБ повторяется, пока координата достигнутого АБ не совпадет с координатой АБ приемника и тогда через порт PCI0 или PCI1 информация будет отправлена в ВМ приемник. Например, передача информации из ВМ 0 в ВМ 21, имеющих координаты, отличающиеся в двух измерениях, осуществляется следующим образом: из ВМ 0 через порт PCI0 информация передается в АБ(0 0 0 0), далее через сетевой порт 2 информация передается в сетевой порт 1 АБ(2 0 0 0). В АБ(2 0 0 0) происходит смена измерения передачи информации, через сетевой порт 8 информация передается в сетевой порт 0 КБ Y2, далее через сетевой порт 2 в сетевой порт 8 АБ(2 2 0 0) и далее через порт PCI1 в ВМ 21.

Основные коммуникационные характеристики топологии MKNS

Количество ВМ в вычислительной системе, использующей КС с топологией MKNS, определяется по формуле:

$$N_{BM} = m * \prod_{i=1}^{n} ki,$$

где вместо n – количество измерений в топологии MKNS $1 \le n \le 4$, вместо ki – количество вычислительных узлов в i-ом измерении, вместо m – количество ВМ в вычислительном узле.

Вычислительная система содержит N_{BM} ВМ и соответственно $N = N_{BM} / m$ вычислительных узлов и адаптерных блоков.

Общее количество КБ в КС с топологией MKNS определяется по формуле:

$$N\text{кб} = \sum_{i=2}^{n} (vi \prod_{j=1, i \neq j}^{n} kj),$$

где вместо vi - количество КБ в КСб i-ого измерения: vi = 1, если в i-ом измерении исполь-

зуется КБ, vi = 6, если в *i*-ом измерении используется 20-портовая КСб, показанная на Рис. 2, вместо *kj* – количество вычислительных узлов в *j*-ом измерении.

В сравнении с КС с топологией KNS при создании КС с топологией MKNS требуется на четверть меньше КБ.

Приведем другие важные характеристики КС, описание которых дано в статье [14]. Диаметр топологии MKNS определяется по формуле:

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

где вместо Si - 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. В сравнении с КС с топологией KNS у КС с топологией MKNS диаметр меньше на 1 хоп. Отметим также, что минимальная дистанция в топологии MKNS равна одному транзитному коммутатору.

Диаметр остовного дерева, построенного на топологии MKNS, зависит от количества измерений топологии MKNS. Топология 1D-MKNS представляет собой полносвязную топологию с диаметром остовного дерева $D_{o.д.}$ = 2. Для других измерений топологии MKNS диаметром остовного дерева рассчитывается по формуле:

$$D_{o.д.} = 4(n\text{-}1),$$

где вместо n – количество измерений в топологии MKNS $2 \le n \le 4$.

Топология MKNS обладает свойством полной бисекции, её ширина бисекции равна *N*/2 при любом разделении КС пополам.

Связность *C* в топологии MKNS вычисляется по следующей формуле:

$$C = (k1 - 1) + (n - 1),$$

где вместо $k1$- количество вычислительных узлов в первом измерении.

КС с топологией MKNS имеет высокую устойчивость к неисправностям, поскольку топология MKNS даже с одним измерением имеет связность равную $k1 - 1$. Устойчивость к неисправностям 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 кабелей связей в КС.

Общее количество кабелей связей *NCC* в КС с топологией MKNS можно определить по формуле:

$$NCC = (k1 - 1) * \frac{k1}{2} * \prod_{i=2}^{n} ki + (n - 1) * \prod_{i=1}^{n} ki.$$

Общее количество сетевых портов *NP* в КС с топологией MKNS вычисляется по формуле:

$$NP = d(N\text{кб} + N),$$

где вместо d - количество сетевых портов в коммутаторе.

Увеличение связности, уменьшение диаметра и количества КБ в топологии MKNS по сравнению с топологией KNS, потребовало увеличения количества кабелей связи для реализации в первом измерении полносвязной топологии, но организация каналов связи первого измерения топологии MKNS на многослойной печатной плате позволит значительно сократить их количество.

Оценка топологии MKNS

В этом разделе приводится оценка топологии MKNS путём сравнения её основных коммуникационных характеристик для КС с коммуникационным оборудованием на основе 10-портовых LR коммутаторов при подключении к коммутатору двух ВМ, с другими топологиями. Сравнение проводилось с коммуникационными характеристиками топологии 4D-Тор, применяемой в КС с коммуникационным оборудованием на основе 8-портовых LR коммутаторов с подключением к коммутатору одного ВМ, характеристиками топологии 6D-Тор, применяемой в КС с коммуникационным оборудованием на основе 12-портовых LR коммутаторов с подключением к коммутатору двух ВМ, и характеристиками топологии Fat Tree, применяемой в КС с коммуникационным оборудованием на основе 40-портовых HR коммутаторов.

В Табл. 2 приведены значения диаметров топологий MKNS, 4D-Torus, 6D-Torus и Fat Tree для вычислительных систем, содержащих в своем составе от 16 до 16000 ВМ.

Из Табл. 2 видно, что для всего диапазона вычислительных систем от 16 до 16000 ВМ диаметр топологии MKNS в 2-3 раза меньше, чем диаметр топологии 4D-Top, в 1,4 - 2 раза меньше, чем диаметр топологии 6D-Top, и даже меньше, чем диаметр топологии Fat Tree, несмотря на использование в этой КС коммуникационного оборудования на основе HR коммутаторов.

В Табл. 3 приведены значения диаметров остовных деревьев, построенных на топологиях MKNS, 4D-Torus, 6D-Torus и Fat Tree для вычислительных систем, содержащих в своем составе от 16 до 16000 ВМ.

Из Табл. 3 видно, что диаметр остовного дерева, построенного на топологии MKNS, в 2 – 3,4 раза меньше, чем диаметр оставного дерева, построенного на топологии 4D-Torus, в 1,5 – 2 раза меньше, чем диаметр остовного дерева, построенного на топологии 6D-Torus, и соизмерим с диаметром остовного дерева, построенного на топологии Fat Tree.

В Табл. 4 приведены значения связности топологий MKNS, 4D-Torus, 6D-Torus и Fat Tree для вычислительных систем, содержащих в своем составе от 16 до 16000 ВМ.

В топологии MKNS значение связности в 2 - 3 раза больше, чем в топологии KNS и не уступает связности топологии 4D-Top. Топологии 6D-Torus и Fat Tree имеют большую связность, чем связность топологии MKNS, за счет использования коммутаторов с большим числом портов.

В Табл. 5 приведены значения ширины бисекции топологий MKNS, 4D-Torus, 6D-Torus и Fat Tree для вычислительных систем, содержащих в своем составе от 16 до 16000 ВМ.

Табл. 2. Значения диаметра в зависимости от количества ВМ

Количество ВМ	MKNS	4D Torus	6D Torus	Fat Tree
16	1	4	3	2
64	2	6	5	4
160	3	7	6	4
1600	5	12	7	6
16000	7	21	14	8

Табл. 3. Значения диаметра остовного дерева в зависимости от количества ВМ

Количество ВМ	MKNS	4D-Torus	6D-Torus	Fat Tree
16	2	4	3	2
64	3	8	5	4
160	4	11	8	4
1600	8	22	13	6
16000	12	41	19	8

Табл. 4. Значения связности в зависимости от количества ВМ

Количество ВМ	MKNS	4D Torus	6D Torus	FatTree
16	7	8	6	16
64	10	8	10	16
160	8	8	10	20
1600	9	8	12	20
16000	10	8	12	20

Из Табл. 5 видно, что ширина бисекции топологии MKNS с двумя ВМ в вычислительном узле в 2 раза меньше, чем ширина бисекции у топологии Fat Tree с одним ВМ в вычислительном узле. Для небольших вычислительных систем, содержащих от 16 до 160 ВМ, ширина бисекции топологии MKNS меньше, чем ширина бисекции топологий 4D-Torus и 6D-Torus. Для средних вычислительных систем, содержащих 1600 ВМ, ширина бисекции топологии MKNS меньше, чем ширина бисекции топологии 4D-Torus, и больше, чем ширина бисекции топологии 6D-Torus. Для больших вычислительных систем, содержащих 16000 ВМ, ширина бисекции топологии MKNS в 1,5 - 2 раза больше, чем ширина бисекции топологий 4D-Torus и 6D-Torus соответственно.

В Табл. 6 приведены значения стоимости (выраженные через количество кабелей связи) топологий MKNS, 4D-Torus, 6D-Torus и Fat Tree для вычислительных систем, содержащих в своем составе от 16 до 16000 ВМ.

Из Табл. 6 видно, что для больших вычислительных систем, содержащих 16000 ВМ, количество кабелей связи в топологии MKNS на 6% больше, чем у топологии 6D-Torus, на 8% больше, чем у топологии Fat Tree и на 22% меньше, чем у топологии 4D-Torus.

В Табл. 7 приведены значения стоимости (выраженной через количество сетевых портов) топологий MKNS, 4D-Torus, 6D-Torus и Fat Tree для вычислительных систем, содержащих в своем составе от 16 до 16000 ВМ.

Из Табл. 7 видно, что для всех вычислительных систем количество сетевых портов в топологии MKNS в среднем на 44% меньше, чем количество сетевых портов в топологи 4D-Torus. Для диапазона вычислительных систем,

Табл. 5. Значения ширины бисекции в зависимости от количества ВМ

Количество ВМ	MKNS	4D Torus	6D Torus	FatTree
16	4	16	8	8
64	16	32	32	32
160	40	64	32	80
1600	400	420	320	800
16000	4000	2662	2000	8000

Табл. 6. Значения стоимости (выраженной через количество кабелей связи) в зависимости от количества ВМ

Количество ВМ	MKNS	4D Torus	6D Torus	FatTree
16	28	32	12	16
64	160	192	80	128
160	360	640	336	320
1600	4400	6720	4374	3200
16000	52000	63888	49152	48000

Табл. 7. Значения стоимости (выраженной через количество сетевых портов) в зависимости от количества ВМ

Количество ВМ	MKNS	4D Torus	6D Torus	FatTree
16	80	128	96	56
64	320	512	384	264
160	880	1280	960	640
1600	9600	12800	9600	9600
16000	104000	128000	96000	128000

содержащих от 16 до 160 ВМ, количество сетевых портов в топологии MKNS в среднем на 16,3% меньше, чем количество сетевых портов в тополоти 6D-Torus, и в среднем на 33% больше, чем количество сетевых портов в тополоти Fat Tree. Для вычислительных систем, содержащих 1600 ВМ, количество сетевых портов одинаково для топологий MKNS, 6D-Torus и Fat Tree. Для вычислительных систем, содержащих 16000 ВМ, количество сетевых портов в топологии MKNS на 8% больше, чем количество сетевых портов в тополоти 6D-Torus, и на 23% меньше, чем количество сетевых портов в тополоти Fat Tree.

Заключение

В статье представлена модернизированная топология MKNS, разработанная для коммуникационных сетей вычислительных систем эксафлопсной производительности. КС с топологией MKNS организуется таким образом, что вычислительные узлы, включающие адаптерный блок и один или несколько вычислительных модулей с арифметическими процессорами, располагаются ортогонально в нескольких измерениях. Связь между вычислительными узлами, находящимися в первом измерении, происходит с помощью адаптерных блоков, объединенных двунаправленными каналами связи по полносвязной топологии (каждый с каждым) без использования коммутаторных блоков. В других измерениях вычислительные узлы и соответствующие им адаптерные блоки объединяются с помощью коммутаторных блоков или с помощью коммутаторных сборок, представляющих собой набор коммутаторных блоков, объединенных по топологии Fat Tree.

За счет объединения в первом измерении вычислительных узлов двунаправленными каналами связи по полносвязной топологии достигается уменьшение диаметра и диаметра остовного дерева топологии КС, следовательно, и коммуникационной задержки, увеличение связности топологии КС, что повышает отказоустойчивость за счет большего количества альтернативных маршрутов и уменьшение количества коммутаторных блоков в КС.

Сравнение основных коммуникационных характеристик топологий MKNS, 4D-Torus, 6D-Torus для КС с коммуникационным оборудованием на базе LR коммутаторов и топологии Fat Tree для с КС с коммуникационным оборудованием на базе HR коммутаторов показало, что диаметр и диаметр остовного дерева топологии MKNS в 1,5 - 3 раза превосходят эти характеристики топологий 4D-Top и 6D-Top и, в основном, сравнимы с этими характеристиками топологии Fat Tree.

К недостаткам топологии MKNS можно отнести большое количество используемых кабелей связи, которое может быть более чем в два раза уменьшено за счет реализации каналов связи первого измерения топологии MKNS на объединительной плате.

В настоящий момент проводится натурное тестирование экспериментальной вычислительной системы (ЭВС). Коммуникационная сеть ЭВС имеет топологию 2D-MKNS с диаметром 3 хопа и в ней используется коммуникационное оборудование СМПО-10С.

Литература

1 Степаненко С.А. Мультипроцессорные среды супер-ЭВМ. Масштабирование эффективности. - М.: ФИЗМАТЛИТ, 2016. – 312 с.

2 InfiniBand Trade Assosiation, InfiniBand Architecture Specification Volume 1 – Release 1.2.1. [Электронный ресурс] - Режим доступа: https://www. infiniband.com.

3 T.P.Morgan., How cray makes Ethernet Suited For HPC And AI With Slingshot. [Электронный ресурс] - Режим доступа: https://nextplatform/2019/08/16/.

4 D.Chen, et al., The IBM Blue Gene/Q Interconnection Network and Message Unit. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC), pp.1-12, 2012.

5 Y. Ajima, T. Inoue, S. Hiramoto, and T. Shimizu, “Tofu: Interconnectfor the K computer”, Fujitsu Scientific and Technical Journal vol. 48,no.3, pp. 280-285, 2012.

6 Симонов А.С., Макагон Д.В., Жабин И.А., Щербак А.Н., Сыромятников Е.Л., Поляков Д.А. Первое поколение высокоскоростной коммуникационной сети «Ангара». Наукоемкие технологии. – 2014. – Т. 15, №1. – С. 21-28.

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

8 Top500 list [Электронный ресурс] - Режим доступа: http://www.top500.org

9 Степин А. Российский интерконнект Ангара-2: 200 Гбит/с при задержках до 0,8 мкс. [Электронный ресурс] - Режим доступа: https://servernews.ru/1033236

10 Dongarra J. Report on the Sunway TaihuLight System. University of Tennessee Department of Electrical Engi-

neering and Computer Science Tech. Report UT-EECS-16–742. June 20, 2016.
11 Kim J., Dally W.J., Scott S., Abts D. Technology-Driven, Highly-Scalable Dragonfly Topology. Computer Architecture. - 2008. - № 08. - P 77 - 88.
12 Cray XC30 System. [Электронный ресурс] - Режим доступа: https://www.nersc.gov
13 Roberto Peñaranda, Crispín Gómez, María E. Gómez, Pedro López & Jose Duato// The K-ary N-direct S-indirect family of topologies for large-scale interconnection network, 05.02.2016, The Journal of Supercomputing volume 72, pages1035–1062, [ресурс интернет] http://doi.org/10.1007/s11227-016-1640-z).
14 Басалов В. Г., Холостов А. А. Перспективная гибридная топология KNS для систем межпроцессорных обменов на базе СМПО-10G. // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. - 2016. - Вып.3. - С. 62-69.

Малых Юрий Алексеевич. Федеральное государственное унитарное предприятие «Российский федеральный ядерный центр-всероссийский научно-исследовательский институт экспериментальной физики» (ФГУП «РФЯЦ-ВНИИЭФ»), г. Саров, Россия. Начальник научно-исследовательского отдела. Область научных интересов высокопроизводительные вычислительные системы, коммуникационные сети, коммуникационное оборудование. E-mail: YAMalykh@vniief.ru

Басалов Владимир Геннадиевич. Федеральное государственное унитарное предприятие «Российский федеральный ядерный центр-всероссийский научно-исследовательский институт экспериментальной физики» (ФГУП «РФЯЦ-ВНИИЭФ»), г. Саров, Россия. Старший научный сотрудник. Количество печатных работ: 7. Область научных интересов высокопроизводительные вычислительные системы, коммуникационные сети, программное обеспечение. E-mail: VGBasalov@vniief.ru

MKNS Network Topology for High-Performance Computer

V. G. Basalov, Y. A. Malykh

Federal State Unitary Enterprise “Russian Federal Nuclear Center – All-Russian Research Institute of Experimental Physics (FSUE "RFNC-VNIIEF"), Sarov, Russia

Abstract. Communication networks play a key role in parallel high-performance computers, called supercomputers, in terms of their performance and cost. Efficiency and cost of communication networks mostly depend on the network topology. In this paper, we describe the MKNS topology developed for communication networks of exascale parallel computers. We provide a comparative analysis of the basic communication characteristics of the MKNS, 4D-Torus, 6D-Torus and Fat Tree topologies, which are used to build communication networks for parallel high-performance computers.
Keywords: network, high-performance computer, MKNS topology

DOI 10.14357/20718632210302

References

1 Stepanenko S. A. 2016. Multiprotsessornye sredy super-EVM. Masshtabirovanie effektivnosti. [Multiprocessor supercomputer environments. Efficiency scaling] Moscow: Fizmatlit. 312 p.
2 InfiniBand Trade Assosiation, InfiniBand Architecture Specification Volume 1 – Release 1.2.1. Available at: https://www.infiniband.com (accessed November 10, 2019).
3 T. P. Morgan, How Cray Makes Ethernet Suited for HPC and AI with Slingshot. Available at: https://nextplatform/2019/08/16/(accessed December 6, 2019).
4 D. Chen, et al. 2012. The IBM Blue Gene/Q Interconnection Network and Message Unit. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC):1-12.
5 Y. Ajima, T. Inoue, S. Hiramoto, and T. Shimizu. 2012. Tofu: Interconnect for the K computer. Fujitsu Scientific and Technical Journal vol. 48, no.3:280-285.
6 Simonov A. S., Makagon D. V., Gabin I. A., Shcherbak A. N., Syromyatnikov E. L., Polyakov D. A. 2014. Pervoe pokolenie vysokoskorostnoy kommunikatsionnoy seti «Angara» [The first generation of the “Angara” high-speed communication network.] Science technology. V. 15, 1: 21-28.

7 Kholostov A. A. 2013. Masshtabiruemaya sistema mezhprotsessornykh obmenov 10G. [10G scalable interconnection system]. Trudy 2-go natsionalnogo superkompyuternogo foruma. [Proceedings of the Second National Supercomputing Forum.] Pereslavl-Zalessky. 126 -134.
8 Top500 list. Available at: http://www.top500.org (accessed February19, 2021).
9 Stepin A. Rossiskiy interkonnekt «Angara-2» [Russian interconnect "Angara-2"]. Available at: https://servernews.ru/1033236 (accessed February 11, 2021).
10 Dongarra J. 2016. Report on the Sunway TaihuLight System. University of Tennessee Department of Electrical Engineering and Computer Science Tech. Report UT-EECS-16–742.
11 Kim J., Dally W. J., Scott S., Abts D. 2008. Technology-Driven, Highly-Scalable Dragonfly Topology. Computer Architecture 08:77 – 88.
12 Cray XC30 System. Available at: https://www.nersc.gov (accessed February 01, 2021).
13 Roberto Peñaranda, Crispín Gómez, María E. Gómez, Pedro López & Jose Duato 2016. The K-ary N-direct S-indirect family of topologies for large-scale interconnection network. The Journal of Supercomputing volume 72:1035–1062, Available at: http://doi.org/10.1007/s11227-016-1640-z) (accessed November 02, 2019).
14 Basalov V. G., Kholostov A. A. 2016. Perspektivnaya gibridnaya topologiya KNS dlya system mezhprotsessornykh obmenov na baze SMPO-10G. [A promising hybrid topology KNS for interconnection systems based on SMPO-10G] Voprosy Atomnoy Nauki i Tekhniki. Matematicheskoe modelirovanie fizicheskikh protsessov. [Mathematical Modeling of Physical Processes] v.3:62-69.

Malykh Y. A. Head of research department. Federal State Unitary Enterprise "Russian Federal Nuclear Center – All-Russian Research Institute of Experimental Physics (FSUE "RFNC-VNIIEF"), Sarov, Russia. E-mail: YAMalykh@vniief.ru

Basalov V. G. Senior researcher. Quantity of printed work: 7. Federal State Unitary Enterprise "Russian Federal Nuclear Center – All-Russian Research Institute of Experimental Physics (FSUE "RFNC-VNIIEF"), Sarov, Russia. E-mail: VGBasalov@vniief.ru