СИСТЕМНАЯ СЕТЬ С МАЛЫМ ДИАМЕТРОМ ИЗ МАЛОПОРТОВЫХ МАРШРУТИЗАТОРОВ

М.Ф. Каравай (mkaravay@ipu.ru), В.С. Подлазов (podlazov@ipu.ru)

(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

Введение.

К системным сетям современных суперкомпьютеров предъявляется требование объединения большого числа связных узлов при минимальном диаметре сети, измеряемой максимально возможным числом скачков между любыми двумя сетевыми узлами. Это требование вызвано необходимостью решения задач с условиях плохого пространственно-временного размещения данных в памяти. В суперкомпьютерах *P775* (*IBM*) и *XC30* (*Cray*) число узлов R измеряется десятками тысяч, а диаметр D – единицами скачков [1, 2]. Указанные значения R и D достигаются за счет использования в качестве связных узлов функционально полных многопортовых маршрутизаторов. В *P775* [1] связной узел имеет 47 дуплексных портов разной пропускной способности (10–50 Гб/с), а в *XC30* (сеть *Dragonfly* [2, 3]) – 40 дуплексных портов близкой пропускной способности.

В РФ в настоящее время нет таких маршрутизаторов. Имеется функционально полный маршрутизатор сети Ангара с 8 дуплексными портами [4]. Он предназначен для построения системной сети в виде 4-мерного тора с R=4096 и D=16. Имеются сведения о намерении увеличить R в 8÷16 раз, но неясно за счет чего. Если просто за счет увеличения числа узлов в измерениях, то это увеличит диаметр до 28÷32 скачков и снизит удельную пропускную способность каждого узла в измерениях с увеличенным числом узлов. И то и другое понизит быстродействие сети. Если же предполагается увеличение числа колец в каждом измерении (как в суперкомпьютере *Gemini* (*Cray*) [5]), то при изменении топологии колец измерений [6] такое значение R достижимо и при меньшем значении D. Для этого потребуется использование двух маршрутизаторов каждом сетевом узле. Однако останется значение $D > 10$ скачков. И это принципиальное ограничение для системных сетей с топологией многомерных торов.

Однако использование двух маршрутизаторов в другой топологии может позволить создать системную сеть с приемлемыми значениями R и D.

Предлагаемое решение.

Объединим в связном узле два маршрутизатора с 8 дуплексными узлами в спарку со следующей структурой (рис. 1 и рис. 2). В ней дуплексные каналы 4-го измерения используются для объединения маршрутизаторов.

Возможно два варианта использования спарки. В первом варианте (рис. 1) оба маршрутизатора с 6 дуплексными портами используются для удваивания числа колец в 3-мерном торе (как в аналогичном торе суперкомпьютера *Gemini*).

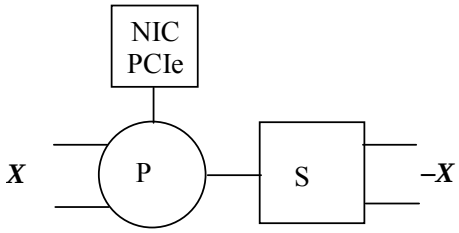

Рис. 1. Спарка маршрутизаторов с удвоенным числом колец в каждом измерении.

В этом варианте за счет изменения топологии колец в каждом измерении (применения дуплексных минимальных коммутируемых мультиколец – ДМКМ [6]) удается как увеличить *R*, так и сократить *D* (Табл. 1).

Таблица 1. Характеристики системных сетей некоторых суперкомпьютеров

Суперкомпьютер	Сеть	***R***	***D***	Число портов связного узла
Gemini [5]	3D-тор	16384	40	20
Ангара [4]	4D-тор	4096	16	8
Ангара [4]	4D-тор	32768	28	8
Ангара со спаркой и ДМКМ [6]	3D-тор	4096	12	12
Ангара со спаркой и ДМКМ [6]	3D-тор	32768	18	12
Gemini с ДМКМ [6]	3D-тор	16384	16	20
Gemini с разреженным ДМКМ [6]	3D-тор	20736	14	20
P775 [1]	2-уровневый полный граф	16416	3-5	47
*XC*30 [2]	4D обобщенный гиперкуб	11616	2-4	40

Во втором варианте маршрутизаторы спарки используются для образования разных измерений в многомерном обобщенном гиперкубе (рис. 3) В спарке первичный маршрутизатор *P* подсоединяется к сетевой карте (*NIC*) *PCIe,* а вторичный маршрутизатор *S* – только к первичному через дуплексный канал 4-го измерения. Предполагается, что передача пакета из *NIC PCIe* в первичный и вторичный маршрутизатор происходит с одинаковыми задержками.

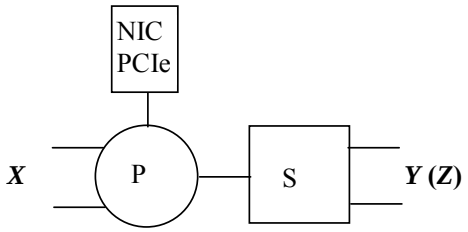

Рис. 2. Сетевая карта *PCIe* (*NIC*) и маршрутизаторы (*P, S*). Сетевой узел – это спарка маршрутизаторов (*P* – первичный, *S* – вторичный).

Предлагаемая системная сеть использует топологию многомерного обобщенного гиперкуба $GGC(R, N, n)$, где R – общее число связных узлов, N – число узлов в каждом измерении, n – число измерений и $R=N^n$. На рис. 3 приведен пример $GGC(64, 4, 3)$. Предполагается что измерения гиперкуба являются подсетями с малым диаметром d. Подчеркнем, что ребра в измерениях на рис. 3 обозначают не каналы между узлами, а только принадлежность этих узлов к одной подсети. Для построения подсетей измерения X используются порты только первичных маршрутизаторов, а подсетей измерений Y, Z и т.д. – порты только вторичных маршрутизаторов.

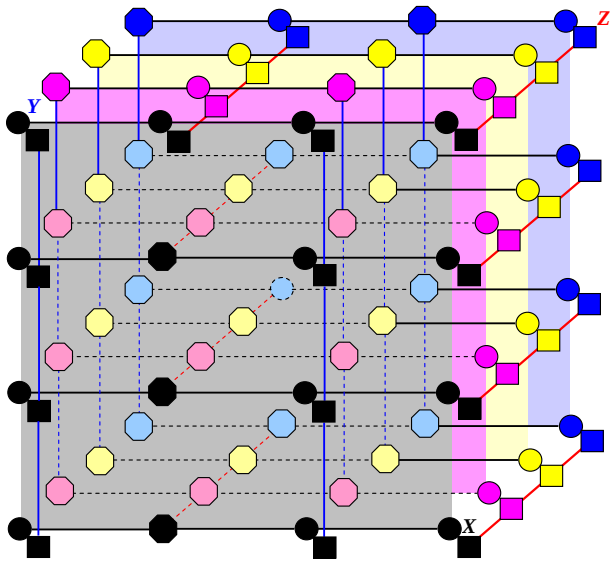

Рис. 3. 3-мерный обобщенный гиперкуб $GGC(64, 4, 3)$с разреженными связями по Y и Z измерениям. Первичные и вторичные маршрутизаторы нарисованы отдельно только в «удобных» местах. В остальных местах спарки нарисованы единым блоком в виде восьмиугольников.

Если используется 2-мерный гиперкуб, то его диаметр $D=d_X+d_Y=2d$. Если используется 3-мерный гиперкуб, то связи измерения Y осуществляются только через спарки с нечетными номерами в измерении X, а связи измерения Z – только через спарки с четными номерами в измерении X (см. рис. 3). В этом случае $D=2d_X+d_Y+d_Z=4d$. Наконец, если используется 4-мерный гиперкуб, то связи измерения Y осуществляются только через спарки с номерами $a \equiv 1 \ (mod 3)$ в измерении X, связи измерения Z – только через спарки с номерами $b \equiv 2 \ (mod 3)$ в измерении X, а связи 4-го измерения U только через спарки с номерами $c \equiv 0 \ (mod 3)$ в измерении X. В этом случае $D=3d_X+d_Y+d_Z+d_U=6d$.

В этом варианте можно добиться уменьшения D за счет использования подсетей со структурой 2-мерных p-ичных коммутируемых мультиколец [7, 8], которые являются распределенными полными коммутаторами на $(p+1)^2$ узлов и имеют топологию квазиполных орграфов [11]. На рис. 4 приведен пример 2-мерного 3-чного

коммутируемого мультикольца на 9 узлов, состоящее из колец с шагами 1, 2 (сплошные дуги) и 3, 6 (пунктирные дуги).

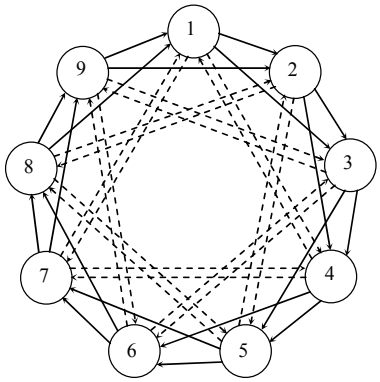

Рис. 4. 2-мерное 3-ичное мультикольцо с шагами 1, 2 и 3, 6.

На m=7 дуплексных портах при p=5 (или p=4) можно построить подсеть с N=20. Она состоит из 7 симплексных колец с шагами 1, 2, 3, 4 и 5, 10, 15 (или 1, 2, 3 и 4, 8, 12, 16), является самомаршрутизируемым неблокируемым распределенным коммутаторов и поэтому имеет диаметр d=1. Общие характеристики такой системной сети задаются в табл. 2.

Таблица 2. Характеристики сети для спарки маршрутизаторов: n – число измерений обобщенного гиперкуба, m – число сетевых портов на карте, N – число узлов в каждом измерении, d – диаметр каждого измерения, R – общее число узлов, D – диаметр сети.

n	m	N	d	$R=N^n$	D	Число портов сетевого узла
2	7	20	1	400	2	14
3	7	20	1	8000	4	14
4	7	20	1	160000	6	14

В данном случае неприменима сеть *flattened butterfly* [9], т.к. она принципиально предполагает использованием маршрутизаторов с 2 и более портами для подключения процессоров. Кроме того, эта сеть не является ни перестраиваемой, ни неблокируемой, что увеличивает ее диаметр по сравнению с номинальным из-за конфликтов даже на перестановочном трафике.

Возможность развития

Характеристики системной сети можно еще улучшить при построение спарки за счет использования сетевой карты с двумя портами *PCIe,* связанных мостом внутри карты (рис. 5). В этом случае как первичный так и вторичный маршрутизаторы подсоединяются через свой разъем *PCIe,* а переход между измерениями гиперкуба осуществляется через мост в *NIC PCIe*.

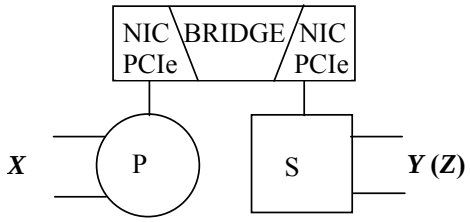

Рис. 5. Рис. 2. Двухпортовый узел *NIC PCIe* и маршрутизаторы (*P, S*).

В этом случае m=8 и из p-ичного мультикольца при p=5 можно построить подсеть с N=25. Она состоит из симплексных 8 колец с шагами 1, 2, 3, 4 и 5, 10, 15, 20 и является неблокируемой подсетью и поэтому имеет d=1. Общие характеристики такой системной сети задаются в табл. 3.

Таблица 3. Характеристики сети для спарки маршрутизаторов: n – число измерений обобщенного гиперкуба, m – число сетевых портов на карте, N – число узлов в каждом измерении, d – диаметр каждого измерения, R – общее число узлов, D – диаметр сети.

n	m	N	d	$R=N^n$	D	Число портов сетевого узла
2	8	25	1	625	2	14
3	8	25	1	15625	4	14
4	8	25	1	390625	6	14

Такие же характеристики будет подсеть в виде мультикольца, состоящего из 4 дуплексных колец с шагами ±1, ±2 и ±5, ±10.

Заключение

Рассмотрена возможность объединения малопортовых сетевых маршрутизаторов для построения сетевых узлов системных сетей с улучшенными характеристиками.

Предложена топологии системной сети на основе обобщенного гиперкуба с разреженными измерениями, в котором каждое измерение представляет подсеть с единичным диаметром.

Рассмотрены построения варианты подсетей в виде 2-мерных мультиколец и характеристики системных сетей с числом узлов в десятки тысяч и диаметром в несколько (4-6) скачков .

ЛИТЕРАТУРА

1. *Arimili B., Arimili R., Chung V., et al.* The PERCS High-Performance Interconnect // 18th IEEE Symposium on High Performance Interconnects. 2009. P. 75–82.

2. *Alverson R., Roweth D., Kaplan L. and Roweth D.* Cray XC® Series Network // URL: http://www.cray.com/Assets/PDF/products/xc/CrayXC30Networking.pdf.

3. *Kim J., Dally W. J., Scott S. and Abts D.* Technology-driven, highly-scalable dragonfly topology // Proceedings of the 35th annual international symposium on computer architecture (Proceeding ISCA'2008). P. 77–88. URL: http://users.ece.gatech.edu/~sudha/academic/class/Networks/Lectures/4%20-%20Topologies/papers/dragonfly.pdf.

4. *Михеев В.А. и др.* Реализация высокоскоростной сети для суперкомпьютерных систем: проблемы, результаты, развитие // URL: http://2013.nscf.ru/TesisAll/Section%201/12_2761_SimonovAS_S1.pdf.

5. *Alverson R., Roweth D. and Kaplan L.* The Gemini System Interconnect // 18th IEEE Symposium on High Performance Interconnects. 2009. P. 3–87.

6. *Подлазов В.С.* Повышение характеристик многомерных торов // Управление большими системами: сборник трудов (электронный журнал). М.: Учреждение Российской академии наук ИПУ им. В.А.Трапезникова РАН. 2014. выпуск 51. С. 60-81. URL: http://ubs.mtas.ru/upload/library/UBS5103.pdf.

7. *Алленов А.В., Подлазов В.С., Стецюра Г.Г.* Пропускная способность набора кольцевых каналов I. Класс наборов колец. Наборы с простыми узлами // АиТ. 1996. № 3. С. 135–144.

8. *Алленов А.В., Подлазов В.С.* Пропускная способность набора кольцевых каналов II. Кольцевые коммутаторы // АиТ. 1996. № 4. С. 162–172.

9. *Kim J., Dally W. J., and Abts D.* Flattened Butterfly: A Cost-Efficiently Topology for High-Radix Networks // URL: http://www.cs.berkeley.edu/~kubitron/courses/cs258-S08/handouts/papers/ISCA_FBFLY.pdf.

10. *Каравай М.Ф., Подлазов В.С.* Метод инвариантного расширения системных сетей многопроцессорных вычислительных систем. Идеальная системная сеть. // АиТ. 2010. №. 10. С. 166–176.

11. *Каравай М.Ф., Подлазов В.С.* Распределенный полный коммутатор как «идеальная» системная сеть для многопроцессорных вычислительных систем // Управление большими системами: сборник трудов (электронный журнал). М.: Учреждение Российской академии наук ИПУ им. В.А.Трапезникова РАН. 2011. вып. 34. С. 92-116. URL: http://ubs.mtas.ru/upload/library/UBS3405.pdf .

12. *Каравай М.Ф., Подлазов В.С.* Топологические резервы суперкомпьютерного интерконнекта // Управление большими системами: сборник трудов (электронный журнал). М.: Учреждение Российской академии наук ИПУ им. В.А.Трапезникова РАН. 2013. вып. 41. С. 395–423. URL: http://ubs.mtas.ru/upload/library/UBS4114.pdf .