УДК 004.724.2+004.272.43

РАСШИРЕНИЕ ВОЗМОЖНОСТЕЙ СИСТЕМНОЙ СЕТИ «АНГАРА»

В.С. Подлазов
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: podlazov@ipu.ru

М.Ф. Каравай
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: mkaravay@ipu.ru

Ключевые слова: Системные сети суперкомпьютеров, сеть «Ангара», инвариантное расширение сетей, масштабирование и быстродействие.

Аннотация: Рассмотрен способ расширения возможностей сети «Ангара» за счет совместного использования собственных 24-портовых маршрутизаторов и рыночных «хабов» 1×3 и 1×4. В различных вариантах их совмещения имеется возможность увеличения масштабируемости сети (повышения числа процессоров), быстродействия сети (сокращения ее диаметра) и ее канальной отказоустойчивости.

1. Введение

Системные сети современных суперкомпьютеров строятся на базе многопортовых коммутаторов-маршрутизаторов – единый однокристальный 48-портовый маршрутизатор *YARK* для 3-мерного тора *Gemini* или для 4-мерного обобщенного гиперкуба *Dragonfly* фирмы *CRAY* [1, 2].

В РФ до последнего времени не было таких маршрутизаторов. Имелся функционально полный однокристальный маршрутизатор сети «Ангара» с 8 дуплексными портами [3, 4]. На его основе был создан однокорпусной маршрутизатор на 24 дуплексных порта интерфейса *PCI-express* за счет сцепления четырех 8-портовых маршрутизаторов. Использование этого маршрутизатора резко упрощает построение сетей самого разного размера – от десятков процессоров до нескольких их тысяч (в топологии 1-мерного или 2-мерного тора) [5, 6].

Заметим, что 24-портовый маршрутизатор имеет внутренний диаметр в 4 скачка: 1 скачок от входного до соединительного порта в 8-портовом маршрутизаторе, 2 скачка между 8-портовыми маршрутизаторами и 1 скачок от соединительного до выходного порта. Однако в сети «Ангара» для связи между 24-портовыми маршрутизаторами используется 4 дуплексных канала: между заданными 8-портовыми маршрутизаторами. Это сокращает проходную задержку маршрутизатора до 1 скачка.

Возможности сети «Ангара» можно существенно расширить, если использовать еще одну внешнюю сцепку маршрутизаторов с «хабом» 1×3. 8-портовый маршрутизатор может быть двумя «хабами» 1×3. Однако проще использовать «хаб» в виде имеющихся на рынке разветвителей 1×3 дуплексных каналов интерфейса *PCI-express*. В иде-

альном случае «хаб» размещается на материнской плате процессора, т.е. используется абонент сети с внутренним «хабом».

2. Инвариантное расширение системных сетей

В свое время авторы решили задачу расширения произвольных системных сетей с сохранением их маршрутных свойств [7]. Указанное расширение осуществляется за счет увеличения числа портов у абонентов до m и использования нескольких копий исходной сети – ИсхС(K), различающихся только непересекающимися наборами абонентов. Решением является расширенная сеть РасС(R), имеющая структуру однородного двудольного графа, в котором все $N(m)$ (определяется дальше) узлов одной доли имеют одинаковые степени K, а все R узлов другой доли – одинаковые степени m. Ребра между узлами разных долей проводятся так, что между любыми двумя узлами одной доли имеются только пути длины 2, каждый из которых проходит только через один узел другой доли. Таких путей должно быть не менее $\sigma \geq 1$, и все они должны проходить через разные узлы другой доли.

В расширенной РасС(R) копии ИсхС(K) трактуются как вершины одной доли двудольного графа, абоненты с m дуплексными портами – как вершины другой его доли, а дуплексные каналы между ними – как его ребра. Иначе говоря, РасС(R) – это сеть, к которой подсоединено R абонентов и которая состоит из N копий ИсхС(K), к каждой из которых подсоединено точно K разных абонентов, и каждый абонент подсоединен точно к N копиям ИсхС(K), и каждый абонент соединяется с любым другим абонентом не менее через σ разных копий ИсхС(K).

Двудольный граф, для которого выполняются равенства $K = m$ и $R = N$, а N задается формулой:

$$N = m(m-1)+1,$$

называется минимальным квазиполным графом. Он задает простейшую системную сеть ПРС(N, m, σ), которая изоморфна симметричным блок-схемам $B(N, m, \sigma)$, изучаемым в комбинаторике [8].

Таблица 1. Межсоединения в простейшей системной сети ПРС(7, 3, 1).

Копии ИсхС(3)	Абоненты		
1	**1**	7	5
2	2	**1**	6
3	3	2	7
4	4	3	**1**
5	5	4	2
6	6	5	3
7	7	6	4

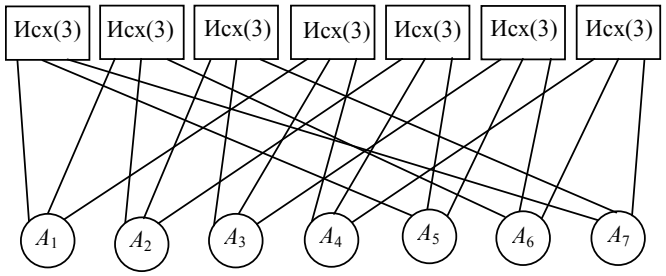

Рис. 1. ПРС(7,3,1) в виде минимального квазиполного графа.

Любая ПРС(N, m, σ) задается таблицей межсоединений состоящей из N строк и m+1 столбцов, в которой 1-ый столбец содержит номера разных копий ИсхС(m), а каждая строка в остальных ячейках – номера абонентов, подсоединенных дуплексным каналом к копии ИсхС(m) в первой ячейке. Например, таблица 1 задает межсоединения для ПРС(7,3,1), а рис. 1 – схему ПРС(7,3,1) в виде минимального квазиполного графа.

Задача построением расширенной сети РасС(R) в общем виде решается следующим образом. Берется N копий ИсхС(K) с $K = rm$, и к ним подсоединяется $R = rN$ абонентов так, что каждый абонент соединяется с любым абонентом последовательно только через одну копию ИсхС(K), и любая пара абонентов соединяется через σ разных ИсхС(K).

Таблица межсоединений для РасС(R) формируется следующим образом. Различные копии ИсхС(K) размещены по портам в M строках таблицы, а различные ПРС – в прямоугольных областях таблицы шириной в m столбцов. В таблице 2 приводится пример расширения ИсхС(15) в РасС(35). Любые два абонента, номера которых не совпадают по mod$N$, соединены друг с другом последовательно только через одну копию ИсхС(K) и используют только ее маршрутные свойства. Кроме того любые два абонента, номера которых совпадают по mod$N$, соединены друг с другом последовательно через m разных копий ИсхС(K) Это обеспечивает как минимум сохранение в РасС(R) маршрутных свойств ИсхС(K). При этом образуется N подмножеств из r абонентов, пропускная способность между которыми в m раз выше.

Таблица 2. Межсоединения в сети РасС(35), составленной из 7 копий ИсхС(15) со встроенными в них 5 копиями ПРС(7,3,1).

Копии ИсхС(15)	Порты копий ИсхС(15)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1-ая ПРС			2-ая ПРС			3-ая ПРС			4-ая ПРС			5-ая ПРС		
1	**1**	**7**	**5**	8	14	12	**15**	**21**	**19**	22	28	26	**29**	**35**	**33**
2	**2**	**1**	**6**	9	8	13	**16**	**15**	**20**	23	22	27	**30**	**29**	**34**
3	**3**	**2**	**7**	10	9	14	**17**	**16**	**21**	24	23	28	**31**	**30**	**35**
4	**4**	**3**	**1**	11	10	8	**18**	**17**	**15**	25	24	22	**32**	**31**	**29**
5	**5**	**4**	**2**	12	11	9	**19**	**18**	**16**	26	25	23	**33**	**32**	**30**
6	**6**	**5**	**3**	13	12	10	**20**	**19**	**17**	27	26	24	**34**	**33**	**31**
7	**7**	**6**	**4**	14	13	11	**21**	**20**	**18**	28	27	25	**35**	**34**	**32**

В настоящей статье рассматриваются возможность расширения сети «Ангара» за счет расширения используемого маршрутизатора.

3. Расширение сети снизу

Простейшая «Ангара» (минимальная 1-мерная решетка) состоит из $M_1 = 2$ 24-портовых маршрутизаторов, объединяет $N_1 = 40$ процессоров и имеет диаметр в $D_1 = 6$ скачков. Ее сложность составляет $S_1 = 2$ маршрутизатора.

Расширим описанным выше методом 24-портовый маршрутизатор до 56-портового маршрутизатора с $n_1 = 56$ абонентами за счет использования «хаба» 1×3. Он имеет диаметр в $d_1 = 5$ скачков, содержит 7 24-портовых маршрутизаторов и 56 «хабов», которые размещаются внутри абонентов. Сложность 56-портового маршрутизатора составляет $s_2 = 7$ 24-портовых маршрутизаторов.

Для организации связи между такими маршрутизаторами в одномерном торе исключаются абоненты 1, 15, 29, 43 и 2, 16, 30, 44, а их 24 порта используются для связи

слева и справа. При этом число подключаемых абонентов уменьшается на 8 (до 48). В двумерном торе дополнительно используются порты абонентов 8, 22, 36, 50 и 9, 23, 37, 51, а их 24 порта используются для связи сверху и снизу. Опять число подключаемых абонентов уменьшается еще на 8 (до 40). Увеличение в три раза числа каналов компенсируется их тройной пропускной способностью.

Составим кольцо (1-мерный тор) с «хабами» из m_2 расширенных маршрутизаторов. Оно имеет диаметр в $d_2 = m_2/2 + 5$ скачков, объединяет $n_2 = 48m_2$ абонентов и имеет сложность $s_2 = 7m_2$. Аналогичное кольцо в исходной сети «Ангара» из M_2 24-портовых маршрутизаторов имеет диаметр в $D_2 = M_2/2+4$ скачков, объединяет $N_2 = 16M_2$ процессоров и имеет сложность $S_2 = M_2$.

При $m_2 = M_2$ получаем почти равный диаметр $d_2 = D_2 + 1$, число абонентов $n_2 = 3N_2$ и сложность $s_2 = 7S_2$, т.е. сеть с «хабами» содержит в 3 раз большее число абонентов, но имеет в 7 раз большую сложность.

При равном числе процессоров $n_2 = N_2$ получаем, что $m_2 = M_2/3$, $d_2 = D_2/3 + 3{,}7$ и $s_2/S_2 = 7/3$, т.е. сеть с «хабами» имеет несколько меньший диаметр, но и в 2,3 раза большую сложность. Если $M_2 = 8$ и $D_2 = 8$, то $d_2 \approx 6{,}3$. Если же $M_2 = 16$ и $D_2 = 12$, то $d_2 \approx 7{,}7$.

Наконец, при одинаковой сложности $s_2 = S_2$ имеем $m_2 = M_2/7$ и получаем, что $d_2 = D_2/7 + 4{,}24$ и $n_2 = 48M_2/7 = 3N_2/7$, т.е. сеть с «хабами» имеет несколько меньший диаметр, но и содержит в 2,3 раза меньшее число абонентов. Если $M_2 = 8$ и $D_2 = 8$, то $d_2 \approx 5{,}7$. Если же $M_2 = 16$ и $D_2 = 12$, то $d_2 \approx 6{,}76$.

Составим теперь 2-мерный тор с «хабами» из m_3 расширенных маршрутизаторов в каждом измерении. Он имеет диаметр в $d_3 = m_3 + 6$ скачков, объединяет $n_3 = 40m_3^2$ абонентов и имеет сложность $s_2 = 7m_3^2$.

Аналогичный 2-мерный тор исходной сети «Ангара» состоит из M_3 24-портовых маршрутизаторов в каждом измерении, имеет диаметр в $D_3 = M_3 + 5$ скачков, объединяет $N_3 = 8M_3^2$ процессоров и имеет сложность $S_3 = M_3^2$.

При $m_3 = M_3$ получаем $d_3 = D_3 + 1$ и $n_3 = 5N_3$, т.е. сеть с «хабами» имеет почти одинаковый диаметр и содержит в 5 раз большее число абонентов, но имеет в 7 раз большую сложность.

При равном числе процессоров $n_3 = N_3$ получаем, что $m_3 = M_3/\sqrt{5} = M_3/2{,}2$ и $d_3 \approx D_3/2{,}2 + 4{,}7$, т.е. сеть с «хабами» имеет несколько меньший диаметр, но и в 7/5 раз большую сложность. Если $M_3 = 8$ и $D_3 = 13$, то $d_3 \approx 8{,}1$. Если же $M_3 = 16$ и $D_3 = 21$, то $d_3 \approx 11{,}7$.

Наконец, при одинаковой сложности $s_3 = S_3$ имеем $m_3 = M_3/\sqrt{7} = M_3/2{,}6$ и получаем, что $d_3 \approx D_3/2{,}6 + 4{,}9$ и $n_3 = 40M_3^2/7 = 5N_3/7$, т.е. сеть с «хабами» имеет несколько меньший диаметр и содержит в 1,4 раз меньшее число процессоров. Если $M_3 = 8$ и $D_3 = 13$, то $d_3 \approx 9{,}9$. Если же $M_3 = 16$ и $D_3 = 21$, то $d_3 \approx 13$.

В таблице 3 сравниваются характеристики сети «Ангара» с «хабами» 1×3 и без них. Расширенный маршрутизатор сравнивается с минимальной 1-мерной решеткой. Таблица 3 показывает, что диаметр сети можно несколько уменьшить, а число абонентов увеличить в несколько раз за счет некоторого увеличения сложности сети.

Таблица 3. Характеристики сети «Ангара» с внутренними «хабами».

Исходная «Ангара»	D	N	S	$s = S/N$
Расширенный маршрутизатор	$5D/6$	$7N/3$	$7S/2$	$3s/2$
1-мерный тор с расширенным маршрутизатором	$M_2 = m_2$ $\boldsymbol{D}+1$	$3N$	$7S$	$7s/3$
	$\sim D/3+3{,}7$	$\boldsymbol{N}$	$7S/3$	$7s/3$
	$\sim D/7+4{,}2$	$3N/7$	$\boldsymbol{S}$	$7s/3$
2-мерный тор с расширенным маршрутизатором	$M_3 = m_3$ $\boldsymbol{D}+1$	$5N$	$7S$	$7s/5$
	$\sim D/2{,}2+4{,}7$	$\boldsymbol{N}$	$7S/5$	$7s/5$
	$\sim D/2{,}6+4{,}9$	$5N/7$	$\boldsymbol{S}$	$7s/5$

В рамках данного подхода можно создать расширенный коммутатор и в отдельном корпусе на 36, 56 и 78 портов. Для этого достаточно сцепку из четырех 8-портовых маршрутизаторов использовать совместно с отдельными «хабами» 1×2, 1×3 и 1×4, размещенными между указанной сцепкой и внешними портами корпуса. При этом отпадает необходимость использования абонентов с внутренним «хабом». Однако при использовании последних рассмотренные возможности позволяют осуществить дополнительное сокращение диаметров торов и увеличение числа процессоров, объединяемых ими.

4. Заключение

Рассмотрен способ расширения возможностей сети «Ангара» за счет совместного использования собственных 24-портовых маршрутизаторов и рыночных «хабов» 1×3 и 1×4. В различных вариантах их совмещения имеется возможность увеличения масштабируемости сети (повышения числа процессоров), быстродействия сети (сокращения диаметра) и ее канальной отказоустойчивости.

Список литературы

1. Alverson R., Roweth D., Kaplan L. The Gemini System Interconnect // 18th IEEE Symposium on High Performance Interconnects. 2009. P. 3-87.
2. Alverson R., Froese E., Kaplan L., Roweth D. Cray XC® Series Network // URL: http://www.cray.com/Assets/PDF/products/xc/CrayXC30Networking.pdf.
3. Михеев В.А. и др. Реализация высокоскоростной сети для суперкомпьютерных систем: проблемы, результаты, развитие // URL: http://2013.nscf.ru/TesisAll/Section%201/12_2761_SiNonovAS_S1.pdf.
4. Симонов А.С. и др. Первое поколение высокоскоростной коммуникационной сети «Ангара» // Наукоемкие технологии. 2014. Т. 15, №1. С. 21–28.
5. Stegailov V. et al. Early Performance Evaluation of the Hybrid Cluster with Torus Interconnect Aimed at Molecular Dynamics Simulations // International Conference on Parallel Processing and Applied Mathematics. Springer, Cham, 2017. P. 327-336. URL: https://link.springer.com/chapter/10.1007/978-3-319-78024-5_29 (accessed: 6.11.2018).
6. Агарков А.А. и др. Предварительные результаты оценочного тестирования отечественной высокоскоростной коммуникационной сети Ангара // Параллельные вычислительные технологии (ПаВТ'2016): труды международной научной конференции (2016 г., Архангельск). Челябинск: Издательский центр ЮУрГУ. 2016. С. 42-53.
7. Каравай М.Ф., Подлазов В.С. Метод инвариантного расширения системных сетей многопроцессорных вычислительных систем. Идеальная системная сеть. // Автоматика и телемеханика. 2010. №. 10. С. 166-176.
8. Холл М. Комбинаторика // М.: Мир. 1970. Гл. 10-12.