СИСТЕМНЫЕ СЕТИ С ВНУТРЕННЕЙ ПАРАЛЛЕЛЬНОСТЬЮ

В.С. Подлазов, М.Ф. Каравай

1. Введение

Системными сетями (*System Area Networks – SAN*) называются сети многопроцессорных вычислительных систем [1]. Базовым блоком больших системных сетей является полный коммутатор, который имеет наименьшие задержки при параллельной передаче пакетов данных между всем портами коммутатора. Он используется в узлах сетей с топологией гиперкуба, тора, решетки, сетей Клоза и т.п. Полный коммутатор, как правило, использует внутреннюю сеть с топологией полного графа (СТПГ) или решетку полных коммутаторов меньшего размера [2]. СТПГ используется и в чистом виде в тех случаях, когда требуется обеспечить наибольшую параллельность и минимальную глубину сети [3].

СТПГ является неблокируемой самомаршрутизируемой сетью с прямыми каналами (СПК). Прямые каналы не содержат промежуточных буферов, но могут использовать промежуточные комбинационные схемы (в составе, например, коммутаторов [2]). В самомаршрутизируемой сети каждый абонент назначает маршрут передаваемого пакета независимо от других абонентов. Бесконфликтная передача пакетов в СПК возможна в неблокируемой и перестраиваемой сети. В первой возможна динамическая бесконфликтная маршрутизация пакетов. Во второй предполагается использование заранее составленных бесконфликтных расписаний – разных для разных перестановок.

Широкое использование СТПГ основано на том, что она имеет наибольшую параллельность, приводящую к наименьшей задержке передачи пакетов. Можно ли получить сеть с прямыми каналами (СПК) с неизменными характеристиками при большем числе узлов и/или с меньшим числом каналов т.е. с «большей параллельностью»? При положительном ответе мы будем называть такую СПК сетью с внутренней параллельностью.

Такая постановка задачи возникает в тех случаях [3, 4], когда возможности СТПГ исчерпаны полностью, но необходимо увеличение числа узлов сети без их изменения (масштабирование сети) или упрощение узлов (снижение энергопотребления сети) без изменения характеристик сети [4].

Будем исследовать СПК, объединяющую N абонентов, в модели сети, синхронной по тактам передачи пакетов одинаковой длины и находящейся в потоковом режиме с перегрузкой. Пусть в каждый абонент на каждом такте поступает для передачи n пакетов данных в течении K тактов подряд ($K >> n$). Каждый абонент имеет m портов и использует m дуплексных каналов. Будем характеризовать СПК двумя параметрами – удельной (на одного абонента) пропускной способностью $\rho(n)$ и средней задержкой пакета $\tau(n)$ при следующих условиях. За K тактов j-й ($1 \le j \le N$) абонент сумеет передать другим абонентам M_j пакетов, причем его i-ый пакет ($1 \le i \le n$)передается за время $t_{i,j}$. Тогда определим указанные параметры для j-го абонента как $\rho_j(n)=M_j/K$ и $\tau_j(n)=\Sigma t_{i,j}/M_j$ и усредним их по абонентам: $\rho(n)=\Sigma\rho_j(n)/N$ и $\tau(n)=\Sigma\tau_j(n)/N$.

Для СПК в виде СТПГ (Full graph) имеем: $m=N–1$ и достигаются характеристики $\rho_F(n)$ и $\tau_F(n)$. Ставится задача построить СПК, для которой $N >> m$ и достигаются характеристики $\rho_Q(n)$ и $\tau_Q(n)$, для которых выполняются условия $\rho_Q(n)\approx\rho_F(n)$ и $\tau_Q(n)\approx\tau_F(n)$. Неформально говоря, необходимо построить СПК с «большей параллельностью», т. е. с большей пропускной способностью, чем у СТПГ. Эта большая параллельность проявляется как внутренняя параллельность, т. к. не меняется степень узлов сети (число портов абонентов). Необходимо найти условия эффективного использования этой внутренней параллельности. В данной работе поставленная задача решается за счет использовании СПК с топологией квазиполного графа (Quasifull graph – СТКГ).

2. Сети с топологией квазиполных графов.

Квазиполным графом мы называем [5] однородный двудольный граф, каждую долю которого составляют N узлов степени m со следующими свойствами. Значение m выбирается минимальным, при котором любые два узла в одной доле связаны σ путями длины 2 через разные узлы в другой доле.

Здесь возникает вопрос о существовании квазиполных графов и об их параметрах. Оказывается, что он уже давно решен в комбинаторике. Такие графы описываются на языке неполных уравновешенных блок-схем, в частности, симметричных блок-схем [5, 6].

Симметричная блок-схема $B(N, m, \sigma)$ состоит из элементов, составляющих одну долю графа, и блоков, составляющих другую долю графа. Число элементов и блоков одинаково и равно N. Параметр m задает число блоков, в которые входит каждый элемент, и число элементов, входящих в каждый блок. Параметр $\sigma < m$ задает число блоков, в которые входит каждая пара элементов. Если $B(N, m, \sigma)$ существует, то ее параметры связаны соотношением $N=m(m–1)/\sigma+1$.

Любая блок-схема описывается таблицей, в которой строчки задают блоки, а ячейки – вхождения элементов. Блоки и элементы задаются своими номерами. Теперь проинтерпретируем блок как коммутатор $m\times m$ с m дуплексными портами, элемент – как абонент с m дуплексными портами, а вхождение элемента в блок – как

подсоединение абонента к коммутатору дуплексным каналом через один из своих портов. Тогда σ интерпретируется как число коммутаторов, через которые любые два абонента соединены разными каналами. При этом все абоненты связаны между собой прямыми каналами через коммутаторы. В отличие от полного графа квазиполный граф может иметь σ независимых путей между любой парой вершин, не являясь при этом *мультиграфом*, поскольку эти пути не параллельны. Вся блок-схема интерпретируется как квазиполный граф, одна доля которого состоит из абонентов, а другая – из коммутаторов. Он описывает простейшую СТКГ с σ-кратным резервированием каналов, которую мы будем обозначать как ПС(N, m, σ). Задающая блок-схему таблица описывает схему межсоединений абонентов и коммутаторов. В табл. 1 приводится описание $B(7, 4, 2)$ и ПС(7, 4, 2), а на рис. 1 – ПС(7, 4, 2) как СТКГ. Толстыми ребрами выделены каналы между абонентами с одинаковой заливкой.

Таблица 1. Схема межсоединений в ПС(7, 4, 2).

Блоки 4×4	$B(7, 4, 2)$ ПС(7, 4, 2)			
0	0	1	2	3
1	0	1	4	6
2	0	2	4	5
3	0	3	5	6
4	1	2	5	6
5	1	3	4	5
6	2	3	4	6

Полный граф содержит $M_F=N_F(N_F-1)$ ребер, а квазиполный граф – $M_Q=mN$ ребер. Обозначим d как $d=M_F/M_Q$. При $N_F=N$ имеем $d=(m-1)/\sigma$. Поэтому СТКГ имеет в d раз меньше дуплексных каналов (кабелей), чем СТПГ.

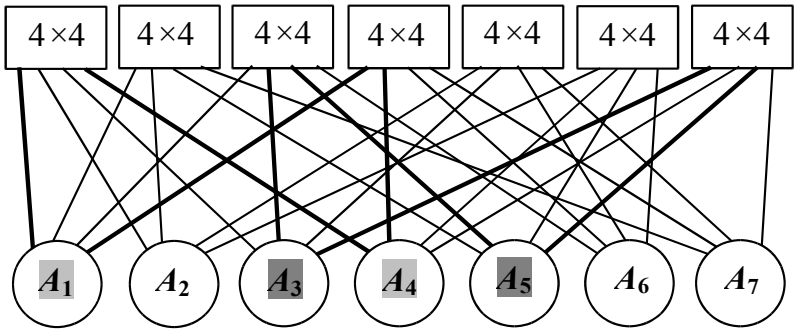

Рис. 1. СТКГ с m=4, N=7 и σ=2.

Уменьшение числа каналов достигается за счет добавления N коммутаторов $m\times m$ суммарной схемной сложности $S=O(Nm^2)=O(N^2)$, т.е. фактически за счет добавления распределенного полного коммутатора $N\times N$.

3. Характеристики СТКГ

Характеристики ПС(N, m, σ) будем получать посредством имитационного моделирования для перестановочного и сетевого трафиков в рамках синхронной по тактам (пакетам) модели передачи. При перестановочном трафике i-й ($1 \le i \le n$) пакет в каждом источнике входит в состав произвольной перестановки, т.е. он может передаваться одному и только одному приемнику. При сетевом трафике множество пакетов по всем источникам имеет случайное равномерное распределение адресов приемников.

В модели все источники действуют синхронно по тактам, пытаясь передать в каждом такте все пакеты для которых есть свободные каналы. Если несколько источников адресуются в СТКГ к одному приемнику через один и тот же коммутатор (конфликт), то пакет передает только один из них, а остальные задерживают пакет в очереди до следующего такта. После каждого такта источники заново генерируют n пакетов. Для каждого канала используется отдельная очередь размера N.

Моделирование проводилось в сетевом и перестановочном режимах для m=12. При этом брался полный граф, в котором $N=m$, т.е. любой узел имеет ребра ко всем узлам (и к самому себе). В сетевом режиме все n пакетов каждого источника равномерно распределяются по адресам приемников. Зависимости $\tau_F(n)$ для полного графа (обозначение FN) и $\tau_Q(n)$ для ряда ПС(N, 12, σ) с топологией квазиполных графов (обозначения QN_σ) представлены на рис. 2. Заметим, что в худшем случае (Q23_6) $\tau_Q(n)/\tau_F(n) \le 1{,}31$, а во всех остальных случаях $\tau_Q(n)/\tau_F(n) \le 1{,}15$.

Для каждой сети имеется граничное n_b, такое что при $n=n_b+1$ очередь начинает бесконечно расти (для F12 имеем n_b=10, для – Q133_1 n_b=10, для Q63_2 – n_b=9, для Q33_4 – n_b=8 и для Q23_6 – n_b=6). Здесь необходимо отметить, что для всех $n \le n_b$ наблюдается $\rho(n)=n$. Поэтому вместо $\rho(n)$ надо рассматривать максимальную длину очереди $q(n)$, усредненной по тактам. Она приводится на рис. 3 для рабочей области. Рабочей областью мы считаем такие n, при которых $\tau(n) \le 1{,}5$. Она оказывается $n \le 6$ для всех сетей.

Рис. 2. Зависимость задержки передачи от числа генерируемых пакетов в сетевом режиме для m=12.

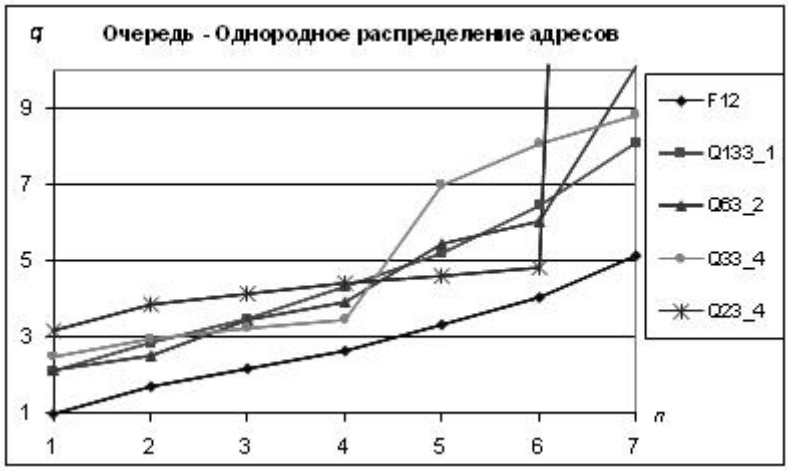

Рис. 3. Зависимость длины очереди от числа генерируемых пакетов в сетевом режиме для m=12 и $\tau(n) \le 1{,}5$.

В рабочей области ПС(N, 12, σ) имеют явное преимущество перед полным графом по числу узлов N, немного уступая по задержке передачи пакетов $\tau_Q(n)$, и имеет очередь $q_Q(n) \le m$.

В перестановочном режиме исследовался случай разных и одинаковых перестановок пакетов данным между абонентами. В первом случае i-е ($1 \le i \le n$) пакеты всех абонентов образуют одну независимую перестановку. Во втором случае все эти перестановки в каждом такте одинаковы. В разных тактах все перестановки генерируются независимо.

Для первого случая на рис. 4 представлены зависимости $\tau(n)$ для $n \le n_b$, при том что для всех $n \le n_b$ наблюдается $\rho(n)=n$. Отметим несколько различий по сравнению с рис. 2. Граничная отсечка (n_b) и рабочая область стали меньше и уменьшилось преимущество полного графа по $\tau(n)$ в рабочей области. Значения n_b для GN стали меньше, чем для QN_σ, а рабочие области во всех случаях совпадают.

Для случая одинаковых перестановок различия еще обостряются: рабочая область сужается до $n \le \sigma$, зато в ней $\tau_F(1)=1$ и $\tau_Q(n)=1$, а также $q_F(1)=1$ и $q_Q(n)=1$, при этом по прежнему $\rho(n)=n$.

Полученные результаты были проверены и для частных случаев при m=24, а именно: ПС(553, 24, 1) и ПС(47, 24, 12). Дело в том, что для квазиполных графов имеется проблема их существования и точного или приближенного построения [5, 8]. Моделирование проводилось для тех ПС(N, 24, σ), которые на настоящее время удалось построить.

На рис. 5 представлены $\tau_F(n)$ и $\tau_Q(n)$ в сетевом режиме для m=24. Отметим, что здесь в худшем случае $\tau_Q(n)/\tau_F(n) \le 1{,}15$. При этом для рабочей области $\tau(n) \le 1{,}5$ и очередь $q(n) < m/2$.

На рис. 6 представлены $\tau_F(n)$ и $\tau_Q(n)$ в перестановочном режиме для m=24. Отметим, что здесь рабочие области и задержки для сетей с топологией полного графа и квазиполного графа с σ =1 практически совпадают, а при $\sigma > 1$ наблюдается преимущество квазиполного графа – тем большее, чем больше σ.

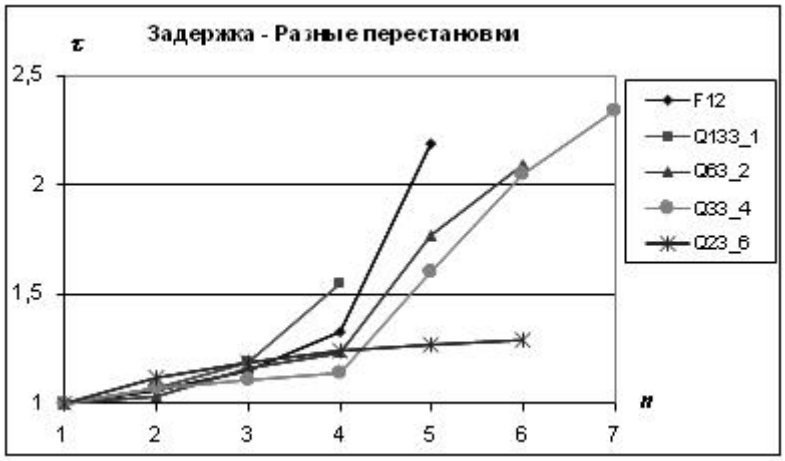

Рис. 4. Зависимость задержки передачи от числа пакетов в перестановочном режиме для m=12

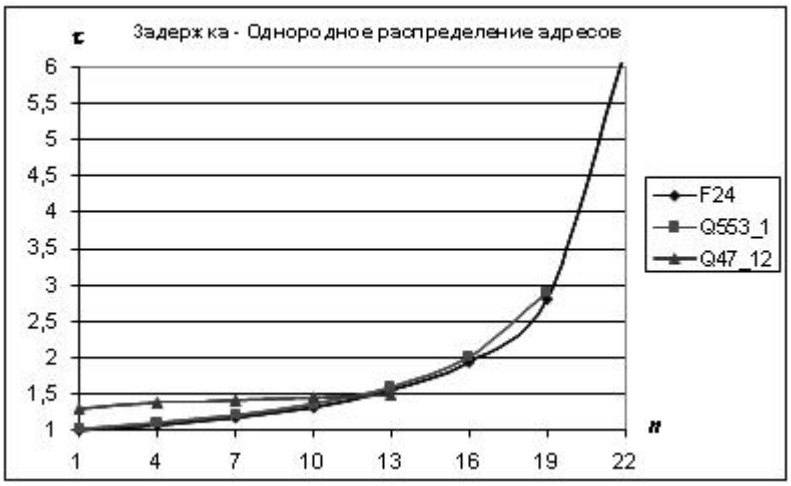

Рис. 5. Зависимость задержки передачи от числа генерируемых пакетов в сетевом режиме для m=24

Рис. 6. Зависимость задержки передачи от числа пакетов в перестановочном режиме для m=24

4. Использование сетей с внутренней параллельностью

Часть современных суперкомпьютеров строится на основе пары узлов: процессорный узел – связной узел [2, 10]. В них связной узел каждой пары содержит многопортовый коммутатор-маршрутизатор, к которому подсоединяются местный процессорный узел и связные узлы других пар. Системная сеть, объединяющая связные узлы, использует прямые каналы «каждый с каждым», т.е. использует топологию полного графа. Каждый канал подсоединяется к своему порту маршрутизатора, а его порты разбиты на группы: для ближних связей (внутри шасси), для средних связей (между шасси и между ближними шкафами) и для дальних связей (между дальними шкафами). Ближние связи осуществляются электрическим кабелем, средние связи – электрическим [9] или оптическим [2] кабелем и дальние связи – только оптическим кабелем. В [2] эти группы содержат 7, 24 и 16 дуплексных портов, а в [9] эти группы содержат 15, 5 и 10 дуплексных портов. В этих системных сетях любой маршрут занимает не более 3-х скачков (передач по прямым каналам с промежуточной маршрутизацией). В [9, 10] минимизируется число оптических кабелей, которые являются более дорогими, и только один скачок осуществляется по оптическому кабелю.

Замена сети с топологией полного графа на сеть с топологией квазиполного графа [4, 11] позволяет резко увеличить число связных узлов, не увеличивая число портов соответствующей группы и число скачков. При этом может оказаться, что не все они могут быть связаны электрическим кабелем. В этом случае необходимо использовать внутреннюю параллельность, которая позволяет иметь независимые каналы при сохранении числа портов и уменьшении числа узлов. Это в свою очередь повышает отказоустойчивость системной сети и ее пропускную способность.

Рассмотрим это на примере системной сети *Dragonfly* суперкомпьютера *CrayXC30* [9, 10]. Она имеет топологию 3-мерного обобщенного гиперкуба. 15 электрических каналов 1-го измерения связывают 16 связных узлов в шасси по топологии полного графа. 5 электрических каналов 2-го измерения связывают 6 шасси в двух шкафах по топологии полного графа. 10 оптических каналов 3-го измерения соединяют шкафы также по топологии полного графа. Заменим полный граф в каждом измерении на квазиполный граф [11]. При этом отсутствие внешних коммутаторов предполагает использование имеющихся, что требует сокращения вдвое числа портов, используемых для построения связей каждого измерения. В качестве портов абонентов (рис. 1) выступают первая половина портов данного измерения, а в качестве портов коммутаторов – вторая их половина.

Чтобы не менять состав шасси оставим для 1-го измерения 6 портов, для 2-го – 1» портов и для 3-го – 10 портов. В 1-ом измерении останется 16 узлов, объединенных квазиполным графом со структурой 2-мерного 4-ичного обобщенного гиперкуба [12]. Из связных узлов 2-го измерения строятся ПС(N, 7, σ), которые объединяют одноименные связные узлы в разных шасси. При σ =1 получаем N=39 шасси. Их вряд ли удастся покрыть электрическими кабелями. Поэтому можно уменьшить число шасси за счет использования большей внутренней параллельности ПС(N, 7, σ) при σ > 1. При σ =2 получаем N=22 шасси, при σ =3 – N=15 шасси и при σ =4 – N=11 шасси, в которых каждая пара связных узла связана 2, 3 и 4 независимыми прямыми каналами соответственно. Эти каналы удобно использовать для обеспечения отказоустойчивости системной сети и для распараллеливания неоднородного трафика без использования промежуточных связных узлов и увеличения числа скачков (как это делается в [9]). В третьем измерении строятся ПС(N, 5, σ), в которых N=21 при σ =1 и N=11 при σ =2.

Отметим, что в рассмотренном примере остались неизменными: пара процессорный–системный блоки, их число в шасси, число кабелей при каждом связном узле и максимальное число скачков в системной сети. При этом число узлов в сети увеличилось в ~4 и более раз и соответственно выросла пропускная способность системной сети.

5. Заключение

Проведено сравнение функциональных характеристик системных сетей с топологией полного и квазиполного графов, имеющих одинаковые узлы-абоненты. Последняя рассматривается как сеть с внутренней параллельностью, расширяющей ее прикладные возможности. Характеристики (пропускная способность, задержки и очереди) измерялись на имитационной модели функционирования сетей в потоковом режиме на сетевом и перестановочном трафиках.

Рассмотренные сети являются неблокируемыми и самомаршрутизируемыми. Сети с топологией квазиполного графа могут иметь значительно больше узлов и независимых каналов между узлами. При сетевом трафике сети с топологией квазиполного графа имеют на 10÷30% большие задержки и очереди при одинаковой пропускной способности, чем сети с топологией полного графа. При перестановочном трафике сети с топологией квазиполного графа по всем характеристикам превосходят сети с топологией полного графа.

Полученные результаты обосновывают возможность эффективного использования сетей с топологией квазиполного графа для расширения (масштабирования) многопроцессорных систем и повышения их отказоустойчивости [4, 9].

ЛИТЕРАТУРА:

1. L. Rzymianowicz Designing efficient network interfaces for system area networks // URL: http://bibserv7.bib.uni-mannheim.de/madoc/volltexte /2002/54/pdf/54_1.pdf.
2. S. Scott, D. Abts, J . Kim, and W. Dally The black widow high-radix Clos network // Proc. 33rd Intern. Symp. Comp. Arch. (ISCA'2006). 2006. URL: http://cva.stanford.edu/people/jjk12/isca06.pdf.
3. B. Arimili, R. Arimili, V. Chung, et al. The PERCS High-Performance Interconnect // 18th IEEE Symposium on High Performance Interconnects. 2009. p.75-82.
4. М.Ф. Каравай, В.С. Подлазов Топологические резервы суперкомпьютерного интерконнекта // Управление большими системами. 2013. вып. 41. с.395-423. URL: http://ubs.mtas.ru/upload/library/UBS4114.pdf.
5. М.Ф. Каравай, П.П. Пархоменко, В.С. Подлазов Комбинаторные методы построения двудольных однородных минимальных квазиполных графов (симметричных блок-схем) // АиТ. 2009. №2. с.153-170.
6. М. Холл Комбинаторика // Главы 10-12. М.: Мир. 1970. 424с.
7. М.Ф. Каравай, В.С. Подлазов, В.В. Соколов Метод расширения полных коммутаторов в фиксированном схемном базисе // Труды 5-й международной конференции «Параллельные вычисления и задачи управления» (PACO'2010) М. ИПУ РАН. Окт. 2010. с.295-305.
8. М.Ф. Каравай, В.С. Подлазов Расширенные блок-схемы для идеальных системных сетей // Проблемы управления. №4. 2012. с.45-51.
9. B. Alverson, E. Froese, L. Kaplan and D. Roweth Cray XC® Series Network // URL: http://www.cray.com/Assets/PDF/products/xc/CrayXC30Networking.pdf.
10. J. Kim, W. J. Dally, S. Scott, D. Abts 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.
11. М.Ф. Каравай, В.С. Подлазов Расширенный обобщенный гиперкуб как отказоустойчивая системная сеть для многопроцессорных систем // Управление большими системами. 2013. Сдана в печать.
12. М.Ф. Каравай, В.С. Подлазов Распределенный полный коммутатор как «идеальная» системная сеть для многопроцессорных вычислительных систем // Управление большими системами. 2011. вып. 34. с.92-116. URL: http://ubs.mtas.ru/upload/library/UBS3405.pdf.