УДК 004.724.2 + 004.272.43
ББК 3.9.7.3.02

РАСПРЕДЕЛЕННЫЙ ПОЛНЫЙ КОММУТАТОР КАК «ИДЕАЛЬНАЯ» СИСТЕМНАЯ СЕТЬ ДЛЯ МНОГОПРОЦЕССОРНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

Каравай М. Ф.[1], Подлазов В. С.[2]
(*Учреждение Российской академии наук
Институт проблем управления РАН, Москва*)

Рассматриваются методы построения распределенного полного коммутатора любого размера, составленного из полных коммутаторов и разветвителей каналов малого фиксированного размера. Распределенный коммутатор сохраняет свойства неблокируемости и самомаршрутизируемости полного коммутатора и образует функционально идеальную системную сеть.

Ключевые слова: многопроцессорные вычислительные системы, идеальные системные сети, распределенный полный коммутатор, канальная коммутация, червячная маршрутизация, неблокируемые сети, самомаршрутизируемые сети.

1. Введение

За сетями многопроцессорных вычислительных систем (МВС) в настоящее время утвердился термин системные сети [24]. В литературе по суперкомпьютерным технологиям часто встречается график зависимости времени счета гипотетической задачи от числа задействованных процессоров (рис. 1) [1]. Рост времени выполнения всей задачи при увеличении числа процессоров объясняется простоями процессоров, возникающих в зна-

[1] *Михаил Федорович Каравай, доктор технических наук, доцент* (*mkaravay@ipu.ru, Москва, ул. Профсоюзная, д.* 65*, тел.* (495) 334-90-00).
[2] *Виктор Сергеевич Подлазов, доктор технических наук, доцент* (*podlazov@ipu.ru, Москва, ул. Профсоюзная, д.* 65*, тел.* (495) 334-78-31).

чительной степени из-за задержек доставки пакетов данных по системной сети.

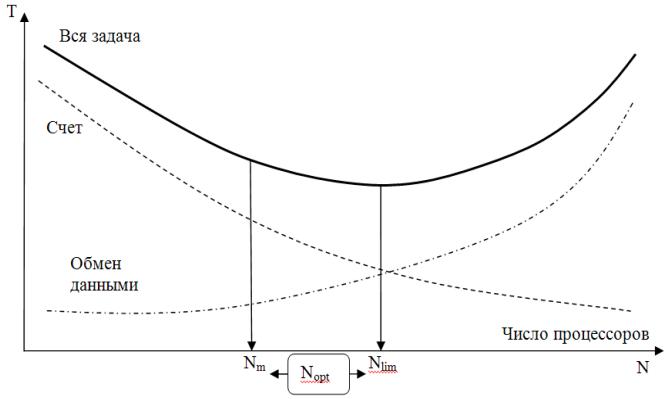

Рис. 1. График времени решения гипотетической задачи на вычислительном кластере

Задержки доставки сильно зависят от числа промежуточных буферизаций пакетов данных в системной сети на пути между источником и приемником. Современная системная сеть (СС) – это часто 2-х, 3-х, 4-хмерные торы [6, 14] или сеть Клоза [25]. Уже на подходе шестимерные торы[3]. Все они не свободны от конфликтов при параллельной передаче пакетов многими абонентами, которые разрешаются посредством буферизации пакетов.

Время доставки отдельного пакета по пустой СС составляет величину $T_0 = \alpha + b / v$, где α (сек) – это латентность сети, v (бит/сек) – скорость передачи и b (бит) – размер пакета. При наличии очередей пакетов в промежуточных буферах время доставки составляет величину $T = T_0 + Q \cdot b / v$, где Q – суммарный размер очередей на отдельном пути от источника к приемнику в СС. Из теории массового обслуживания известно, что при высокой загрузке сети величина $Q \cdot b / v$ может быть много больше T_0. Поэтому в литературе [21] появилось предложение считать идеальной ту СС, в которой обеспечиваются прямые канальные со-

[3] *Ласточка в облаках* // Суперкомпьютеры. 2010. № 2. С. 17 – 19.

единения (без промежуточной буферизации) для любой пары абонентов сети при параллельной передаче пакетов от всех абонентов, т.е. в которой $Q = 0$. Правильнее такую СС называть функционально идеальной, т.к. здесь не учитывается вопрос о ее сложности. Однако для краткости в дальнейшем используется термин «идеальная СС».

Общепринятой моделью параллельной передачи данных по СС является произвольная перестановка пакетов данных между всеми абонентами, поскольку она наилучшим образом характеризует логические возможности сети. Какая СС на N абонентов является функционально идеальной? Очевидно, это СС со структурой полного графа, которая может быть как распределенной, так и сосредоточенной. В первом случае предполагается наличие у каждого абонента $N - 1$ дуплексных портов и использование в СС $N(N - 1)$ дуплексных каналов. Во втором случае предполагается использование СБИС полного коммутатора $N \times N$ с N дуплексными портами. В обоих случаях невозможно построить СС с большим числом абонентов либо вследствие большого числа портов у каждого абонента и каналов между ними, либо из-за невозможности создания СБИС с необходимым числом портов.

Поставим задачу создания функционально идеальной СС, построенной в элементной базе, состоящей из коммутаторов и разветвителей каналов малого размера. Такая элементная база имеется, например, для интерфейса PCI-Express [2] и технологии Space Wire [11, 12].

Пусть имеется исходная идеальная СС (рис. 2), объединяющая K абонентов – ИС(K).

Ставится задача расширить ИС(K) до идеальной сети РС(R), объединяющей $R > K$ абонентов. Более конкретно, мы хотим строить расширенную сеть РС(R), соблюдая следующие условия:

1. РС(R) должна сохранять маршрутные свойства ИС(K).
2. РС(R) должна строиться из тех же схемных компонент, что и ИС(K).
3. РС(R) должна быть расширяема до любого сколь угодно большого R при любом K при сохранении свойств пп.1-2.

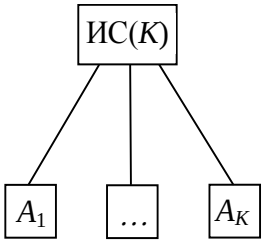

Рис. 2. Исходная сеть на K абонентов

Основными маршрутными свойствами идеальной СС со структурой полного графа являются ее неблокируемость и самомаршрутизируемость. Неблокируемость означает возможность бесконфликтно осуществлять произвольную перестановку пакетов данных между абонентами при параллельной передаче пакетов от всех абонентов, а самомаршрутизируемость – возможность прокладки маршрута при перестановке пакетов каждым абонентом самостоятельно независимо от других абонентов.

Пусть теперь идеальная исходная СС – это полный коммутатор $m \times m$ с m дуплексными портами. Обычный метод расширения полного коммутатора на большее число абонентов – это построение многокаскадной сети Клоза [23] или гиперкуба [18]. Они привлекательны тем, что имеют меньшую сложность, чем полный коммутатор того же размера, и обладают свойством перестраиваемости. Перестраиваемая сеть имеет отдельное бесконфликтное расписание для любой перестановки данных между входными и выходными портами. Однако его построение требует много больше времени, чем реализация самой перестановки. Поэтому на практике для произвольной перестановки обычно используется самомаршрутизация, например червячная маршрутизация [19, 20]. Она допускает возникновение конфликтов, снижающих пропускную способность сети и увеличивающих задержки передачи пакетов. В итоге оказывается, что сеть Клоза и гиперкуб не сохраняют идеальности полного коммутатора.

Известно, что расширение полного коммутатора в виде двумерного обобщенного гиперкуба или двумерного полного мультикольца [8, 25] сохраняет свойства неблокируемости и самомаршрутизируемости, т.е. обеспечивает сохранение идеальности расширенной сети. Однако при этом выполняются только

пп. 1-2. Отметим, что [25] – это единственная иностранная публикация, известная авторам, в которой имеется частичное пересечение с рассматриваемым методом.

Кроме того, авторами был разработан новый метод построения РС(R), удовлетворяющий условиям 1-2 [3, 4, 7-10], основанный на математической теории неполных уравновешенных блок-схем, исследуемых в комбинаторике.

В дальнейшем эти методы рассматриваются отдельно и дополняются каскадным их применением, которое позволяет удовлетворить условию 3.

Методы, рассматриваемые в разделах 2.1 и 2.3, являются полностью оригинальными, а метод из раздела 2.2 частично пересекается с методом построения самой большой коммутаторной СБИС в [25].

2. Распределенные полные коммутаторы

2.1. РАСПРЕДЕЛЕННЫЙ ПОЛНЫЙ КОММУТАТОР НА БАЗЕ МУЛЬТИКОЛЬЦА

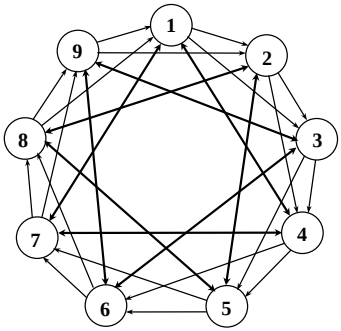

Рис. 3. Полное двумерное мультикольцо с $N = 9$ узлами

Двумерное полное мультикольцо определяется как кольцевой орграф с $N = m^2$ узлами, в котором из каждого узла выходят набор дуг с длинами $(1, 2, \ldots, m-1, m, 2m, \ldots, (m-1)m)$. Длиной дуги мы называем разницу номеров по $\mathrm{mod}\, N$ инцидентных ей узлов. В таком мультикольце все узлы имеют одинаковую степень $2(m-1)$.

На рис. 3 приводится пример мультикольца с девятью узлами. Оно имеет дуги с набором длин {1, 2, 3, 6}. Дуги с длинами 3 и 6 обозначены двунаправленными стрелками. Каждый его узел содержит абонента с $m = 3$ портами и коммутатор $m \times m$ (рис. 4).

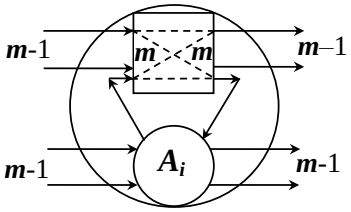

Рис. 4. Коммутатор m×m и абонент A_i (процессор) в составе i-го узла

Маршрутизация в двумерном мультикольце осуществляется как червячная маршрутизация, т.е. путем прокладки прямого канала между абонентом-источником и абонентом-приемником через промежуточный коммутатор. Эта прокладка осуществляется путем посылки пилотного пакета, содержащего адрес абонента-приемника. Она осуществляется в два этапа – сначала по дугам малых длин (от 1 до $m - 1$), а затем – по дугам больших длин (от m до $(m - 1)m$). На первом этапе используются каналы от абонентов к коммутаторам, а на втором – от коммутаторов к абонентам (или наоборот). Нельзя только смешивать в одном этапе передачи по каналам малых и больших длин, т.к. это может привести к возникновению конфликта и, как следствие, к возникновению тупиковой ситуации.

Схема подсоединений дуг двумерного мультикольца может быть перерисована в виде двудольного орграфа (рис. 5). Одну его долю составляют абоненты, а другую – коммутаторы. Степень всех вершин в каждой доле одинакова и равна m. Значение m выбирается минимальным, при котором любые два абонента связаны одним путем длины два через один и только коммутатор в другой доле. В этом случае число вершин в каждой доле N задается равенством $N = m^2$. Такой орграф авторы называют минимальным квазиполным орграфом [5]. На рис. 4 приведен пример этого орграфа для $m = 3$ ($N = 9$).

Схема соединений между коммутаторами и абонентами при $N = 9$ задается в табл. 1. В любой СС со структурой минимального квазиполного орграфа любые два абонента связаны одним путем длины два через один и только один коммутатор.

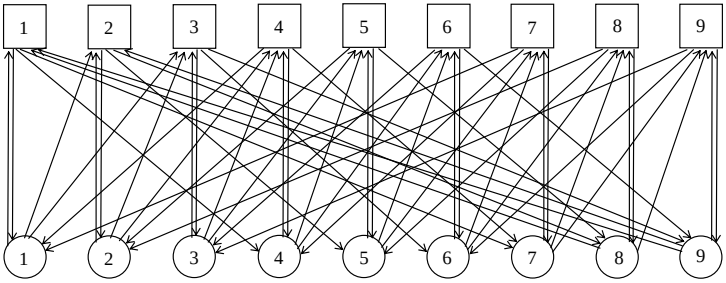

Рис. 5. Минимальный квазиполный орграф для двумерного мультикольца при $m = 3$

Можно сформулировать следующую теорему, которая приводится без формального доказательства, но фактически доказывается всем текстом данного раздела.

Теорема 1. *Системная сеть, построенная по схеме минимального квазиполного орграфа, является неблокируемой и самомаршрутизируемой посредством червячной маршрутизации на произвольной перестановке пакетов данных за счет прокладки отдельных каналов между любыми двумя абонентами.*

Минимальный квазиполный граф описывает схему расширения идеальной СС ИС(m) до идеальной СС РС(m^2), в которой ИС(m) представляет собой полный коммутатор $m \times m$. В табл. 2 приводится схема межсоединений абонентов и коммутаторов при произвольном m. Здесь на пересечении i-ой строки ($1 \le i \le N$) и j-го столбца ($1 \le j \le m$) в левой части таблицы содержится номер $(i - j) \bmod (N + 1)$, а в правой – номер $(i + (j - 1)\, m) \bmod (N + 1)$. При этом приемные и передающие порты в каждом абоненте или в каждом коммутаторе могут принадлежать разным дугам. Поэтому дуплексные порты здесь использовать невозможно.

Таблица 1. Схема соединений в квазиполном графе мультикольца при m = 3.

Коммутаторы	Симплексные каналы от абонентов			Симплексные каналы к абонентам		
1	1	9	8	**1**	**4**	**7**
2	2	1	9	**2**	**5**	**8**
3	3	2	1	**3**	**6**	**9**
4	4	3	2	**4**	**7**	**1**
5	5	4	3	**5**	**8**	**2**
6	6	5	4	**6**	**9**	**3**
7	7	6	5	**7**	**1**	**4**
8	8	7	6	**8**	**2**	**5**
9	9	8	7	**9**	**3**	**6**

Таблица 2. Схема соединений в квазиполном графе мультикольца при произвольном m.

$m\times m$	Входы коммутаторов $m\times m$					Выходы коммутаторов $m\times m$				
1	1	N	N–1	…	N–(m–2)	**1**	**1 + *m***	**1 + 2*m***	**…**	**1 + (*m*–1)*m***
2	2	1	N–2	…	N–(m–3)	**2**	**2 + *m***	**2 + 2*m***	**…**	**2 + (*m*–1)*m***
…	…	…	…	…	…	**…**	**…**	**…**	**…**	**…**
m	m	m–1	m–2	…	1	***m***	**2*m***	**3*m***	**…**	***N***
m + 1	m + 1	m	m–1	…	2	***m* + 1**	**2*m* + 1**	**3*m* + 1**	**…**	**1**
…	…	…	…	…	…	**…**	**…**	**…**	**…**	**…**
N–1	N–1	N–2	N–3	…	N–m	***N*–1**	***m*–1**	**2*m*–1**	**…**	***N*–*m*–1**
N	N	N–1	N–2	…	N–(m–1)	***N***	***m***	***2m***	**…**	***N*–*m***

Отличительным свойством идеальной СС РС(m^2) является то, что она является неблокируемой и самомаршрутизируемой на любой однородной m-перестановке, при которой (по определению) в портах каждого абонента имеются пакеты, адресованные только разным абонентам или поступившие от разных або-

нентов. Это свойство позволяет каждому абоненту параллельно передавать и принимать до m разных пакетов. Это же свойство позволяет реализовать групповую операцию «все – всем» за два сеанса [26, 22] с суммарной длительностью, равной времени передачи $m + 1$ пакета. Во время первого сеанса каждый источник передает свой пакет с каждого порта m разным приемникам, а каждый приемник получает m пакетов на каждый порт от разных источников. Это занимает время передачи одного пакета. Во время второго сеанса каждый источник передает каждый пакет, принятый в первом сеансе, так же как в первом сеансе. Это занимает время передачи m пакетов. В результате каждый абонент получит **все** пакеты других абонентов.

Выделим у каждого абонента схему формирования m портов (рис. 6), которую составляет разветвитель/объединитель m симплексных каналов. Будем различать два вида таких схем. Первая (рис. 6а) позволяет всем портам работать параллельно и независимо. Она обозначается РОКm* и изображается с заливкой. Это схема типа многопортовой сетевой карты в PCI-Express. Вторая (рис. 6б) позволяет работать только одному порту. Это схема демультиплексора-мультиплексора. Такие схемы есть в технологии Space Wire.

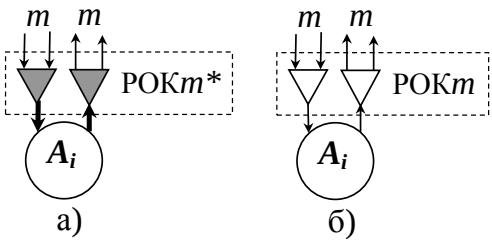

Рис. 6. Разветвитель/объединитель m симплексных каналов РОКm (а) и РОКm (б)*

Теперь схему СС РС(m^2) (рис. 5) можно преобразовать в схему СС с двумя коммутирующими каскадами – коммутаторов $m \times m$ и РОКm* или РОКm. Пример такой схемы при $m = 3$ приведен на рис. 7. Часть такой схемы СС, располагающаяся выше интерфейса абонент-РОКm* (РОКm) и заключенная в пунктир-

ный прямоугольник, представляет собой полный распределенный коммутатор $N \times N$, где $N = m^2$. Он обладает свойствами неблокируемости и самомаршрутизируемости. При этом неблокируемость достигается на любой однородной m-перестановке при использовании схемы РОКm* и на любой обычной 1-перестановке при использовании схемы РОКm.

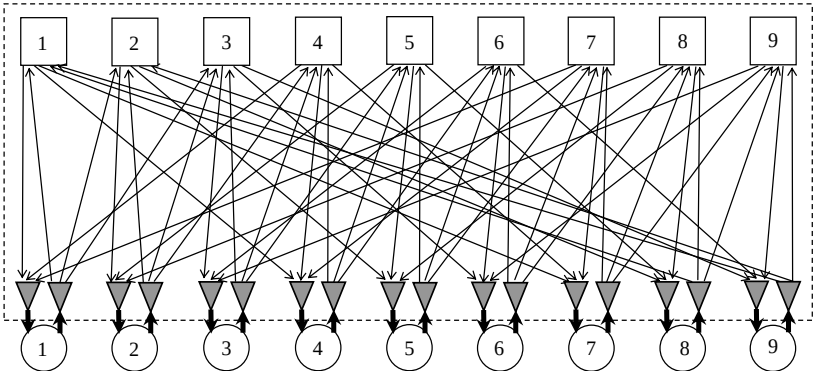

Рис. 7. Распределенный полный коммутатор при $N = 9$

Рассмотрим две такие характеристики распределенного полного коммутатора $\text{РС}(N) = \text{РС}(m^2)$ как схемную и портовую (канальную) сложности. Базовый способ создания СС в виде полного коммутатора – это соединение абонентов прямыми каналами. При этом схемная сложность S составляет $S = N(N-1)$. Схемная сложность однокаскадного распределенного коммутатора s составляет $s = 2mN + Nm^2 = 2mN + N^2$. Их отношение: $s/S \approx 1 + 2/N^{1/2}$. Поэтому распределенный коммутатор немного сложнее. Совсем другая картина складывается по числу портов или каналов. Для СС в виде полного коммутатора число портов составляет величину $W = N(N-1)$, а для распределенного коммутатора – $w = mN$. Их отношение: $w/W \approx 1/N^{1/2}$. Поэтому распределенный коммутатор имеет много меньше каналов и занимает много меньшую площадь при реализации в СБИС или в ПЛИС.

2.2. РАСПРЕДЕЛЕННЫЙ ПОЛНЫЙ КОММУТАТОР НА БАЗЕ ОБОБЩЕННОГО ГИПЕРКУБА

Мультикольцо является не единственной сетевой структурой, которая обеспечивает неблокируемость и самомаршрутизируемость СС. Другой такой структурой является двумерный m-ичный (обобщенный) гиперкуб. В нем каждая строка или столбец из m узлов образует полный граф. Он имеет $N = m^2$ узлов. Пример двумерного обобщенного гиперкуба с $N = 9$ узлами приведен на рис. 8. Хотя он выглядит как граф, но его можно представить как орграф, если учесть, что каждый узел содержит кроме абонента еще и коммутатор $m \times m$ (рис. 9) [5, 25]. Здесь хорошо видно, что каналы могут быть дуплексными, а порты – только симплексными.

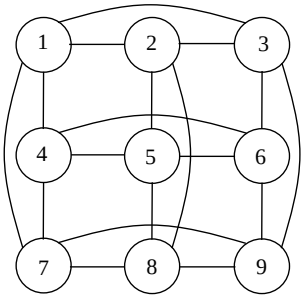

Рис. 8. Двумерный троичный гиперкуб как граф

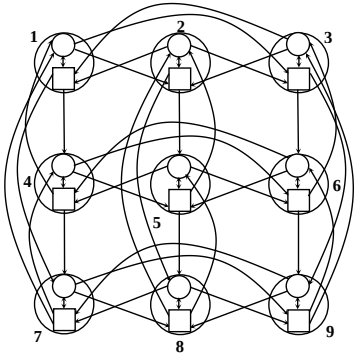

Рис. 9. Двумерный троичный гиперкуб как орграф

Его также можно представить в виде минимального квазиполного орграфа, одну долю которого составляют абоненты, а другую – коммутаторы. Степень всех вершин в каждой доле одинакова и равна m. Число вершин в каждой доле N задается равенством $N = m^2$. В этом орграфе любые два абонента связаны одним путем длины 2 (через один и только один коммутатор). Схема соединений в нем между коммутаторами и абонентами задается в табл. 3. Она отличается от аналогичной схемы для мультикольца.

Таблица 3. Схема соединений в квазиполном орграфе обобщенного гиперкуба при $m = 3$.

Коммутаторы	Симплексные каналы от абонентов			Симплексные каналы к абонентам		
1	1	2	3	**1**	**4**	**7**
2	1	2	3	**2**	**5**	**8**
3	1	2	3	**3**	**6**	**9**
4	4	5	6	**1**	**4**	**7**
5	4	5	6	**2**	**5**	**8**
6	4	5	6	**3**	**6**	**9**
7	7	8	9	**1**	**4**	**7**
8	7	8	9	**2**	**5**	**8**
9	7	8	9	**3**	**6**	**9**

При произвольном m на пересечении i-й строки ($1 \le i \le N$) и j-о столбца ($1 \le j \le m$) в левой части таблицы содержится номер $\lfloor i/m \rfloor m + j$, а в правой – номер $(i) \bmod(m + 1) + (j - 1)m$.

Любой минимальный квазиполный граф описывает схему расширения идеальной СС ИС(m) до идеальной СС РС(m^2), в которой ИС(m) представляет собой полный коммутатор $m \times m$. За счет использования разветвителей/объединителей каналов РОКm* или РОКm СС РС(m^2) можно представить как полный распределенный коммутатор $N \times N$, где $N = m^2$. В данном случае такой распределенный коммутатор представлен на рис. 10. Такой коммутатор будем называть однокаскадным (по числу каскадов схем РОКm* или РОКm) и обозначать как $РК_1(N)$.

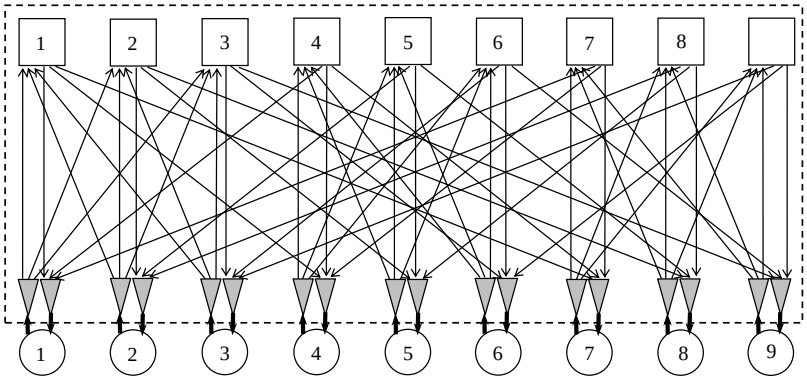

Рис. 10. Еще один распределенный полный коммутатор при $N = 9$

Здесь возникает вопрос: можно ли его расширять дальше за счет увеличения числа каскадов в постоянном схемном базисе коммутаторов $m \times m$ и схем РОКm* или РОКm. Ответ положительный. Делается это следующим образом. Пусть в качестве исходного коммутатора выступает k-каскадный распределенный полный коммутатор $R_k \times R_k$ на R_k абонентов, где $R_k = m^{k+1}$ и $R_1 = N$. Берем N таких коммутаторов. Разобьем каждый из них на m^k равных частей – каждая из m портов. К первой части всех коммутаторов $R_k \times R_k$ подсоединяется абоненты с номерами от 1 до R_k как при построении коммутатора $R_1 \times R_1$. К i-ой части ($1 < i \le m^k$) всех коммутаторов $R_k \times R_k$. подсоединим абонентов с номерами от $1 + iR_1$ до $(i + 1)R_1$. При этом абонент с номером $j + iR_1$ ($1 \le j \le R_1$) подсоединяется к тем и только тем коммутаторам, что и абонент с номером $j + R_k$. Так создается распределенный коммутатор $R_{k+1} \times R_{k+1}$ на $R_{k+1} = (R_k / m)R_1 = m^{k+2}$ абонентов.

В табл. 4 приводится пример схемы межсоединений для двухкаскадного распределенного коммутатора 27×27 при $m = 3$. Он расширен из полного коммутатора на рис. 10.

Теорема 2. *Расширенная сеть* $\text{РС}_k(m^{k+1})$ *в виде построенного выше* k*-каскадного распределенного коммутатора* $\text{РК}_k(m^{k+1})$ *является неблокируемой и самомаршрутизируемой СС. В ней любые два порта любых абонента связаны только одним путем длины* $2k$*, проходящим последовательно только*

через один коммутатор $m \times m$ *и разные* РОКm. *Таким образом* $РС_k(m^{k+1})$ *является идеальной СС.*

Теорема 2 доказывается по индукции. Основанием индукции является теорема 1. При переходе от k-каскадного коммутатора к $(k + 1)$-каскадному коммутатору неблокируемость и самомаршрутизируемость обеспечивается в каждой группе по теореме 1 и по построению, а между группами – их свойствами по индуктивному предположению.

Таблица 4. Схема межсоединений в двухкаскадном коммутаторе при $m = 3$.

3×3	Входы коммутаторов 3×3									Выходы коммутаторов 3×3									
1	1	2	3	10	11	12	19	20	21	**1**	**4**	**7**	**10**	**13**	**16**	**19**	**22**	**25**
2	1	2	3	10	11	12	19	20	21	**2**	**5**	**8**	**11**	**14**	**17**	**20**	**23**	**26**
3	1	2	3	10	11	12	19	20	21	**3**	**6**	**9**	**12**	**15**	**18**	**21**	**24**	**27**
4	4	5	6	13	14	15	22	23	21	**1**	**4**	**7**	**10**	**13**	**16**	**19**	**22**	**25**
5	4	5	6	13	14	15	22	24	21	**2**	**5**	**8**	**11**	**14**	**17**	**20**	**23**	**26**
6	4	5	6	13	14	15	22	24	21	**3**	**6**	**9**	**12**	**15**	**18**	**21**	**24**	**27**
7	7	8	9	16	17	18	25	26	27	**1**	**4**	**7**	**10**	**13**	**16**	**19**	**22**	**25**
8	7	8	9	16	17	18	25	26	27	**2**	**5**	**8**	**11**	**14**	**17**	**20**	**23**	**26**
9	7	8	9	16	17	18	25	26	27	**3**	**6**	**9**	**12**	**15**	**18**	**21**	**24**	**27**

В k-каскадном полном коммутаторе червячная маршрутизация осуществляется по адресу, состоящему из цуга $(k + 1)$-го локальных адресов. Первый адрес используется для выбора порта в первом каскаде, i-й $(1 < i \le k)$ – для выбора выходного порта в i-м каскаде, а последний – для выбора выходного порта в хребте. На обратном пути от хребта к абоненту адреса не используются. Однако здесь в общем случае приходится осуществлять множественный доступ к одному выходному порту [9].

Рассмотрим еще один пример построения двухкаскадного распределенного полного коммутатора при $m = 2$, который приводится на рис. 11-12. Из них видно, что в распределенном k-

каскадном полном коммутаторе имеется один хребтовый каскад, который состоит только из коммутаторов $m \times m$, и k каскадов, каждый из который состоит только из схем РОКm* или РОКm.

В заключение данного раздела еще раз подчеркнем, что отличительным свойством идеальной СС $РС_k(m^{k+1})$ в виде распределенного полного коммутатора $РК_k(m^{k+1})$ является то, что она является неблокируемой и самомаршрутизируемой на любой однородной m-перестановке, при которой во всех портах каждого абонента имеются пакеты, адресованные разным абонентам или поступившие от разных абонентов.

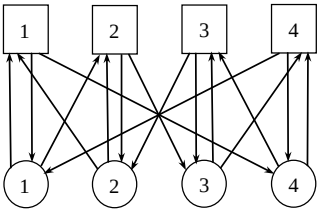

Рис. 11. Минимальный квазиполный орграф для $m = 2$

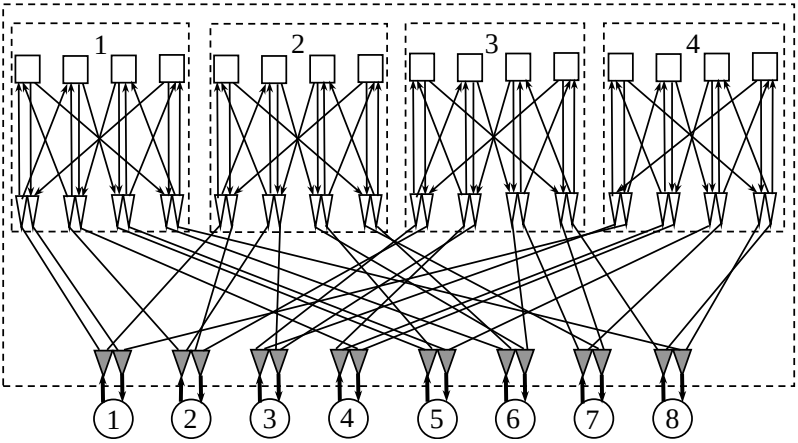

Рис. 12. Двухкаскадный распределенный полный коммутатор 8×8

2.3. РАСПРЕДЕЛЕННЫЙ ПОЛНЫЙ КОММУТАТОР НА БАЗЕ СИММЕТРИЧНЫХ БЛОК-СХЕМ

Рассмотрим еще одну модель расширения идеальной СС ИС(m) (рис. 2). Ее задает однородный двудольный **граф**, одну

долю которого составляют коммутаторы $m \times m$, а другую – m-портовые абоненты. Значение m выбирается минимальным, при котором любые два узла в одной доле связаны σ путями длины два через разные узлы в другой доле. В одной доле имеется N коммутаторов, а в другой – N абонентов. Каждый такой путь проходит через один коммутатор, и разные пути проходят через разные коммутаторы. Двудольный однородный граф с описанными свойствами мы называем минимальным квазиполным графом [3]. Пример такого графа приведен на рис. 13 для $m = 4$, $N = 7$ и $\sigma = 2$. На рис. 13 толстыми линиями выделены пути между абонентами, выделенными одинаковой заливкой. Нетрудно видеть, что их два для каждой пары абонентов.

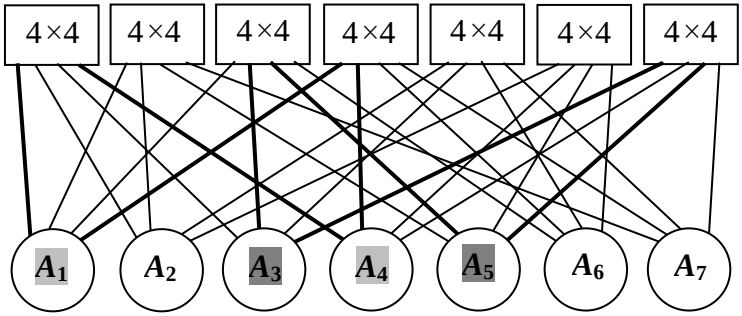

Рис. 13. Минимальный квазиполный граф с $m = 4$, $N = 7$ *и* $\sigma = 2$

Здесь возникает вопрос о существовании минимальных квазиполных графов и об их параметрах. Оказывается, что он уже давно решен в комбинаторике [3, 13]. Такие графы описываются на языке неполных уравновешенных блок-схем, в частности, симметричных блок-схем [3, 4, 7-10].

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. Вхождение некоторого элемента в некоторый блок задает ребро на двудольном графе между соответствующими вершинами разных долей. Параметр $\sigma < m$ задает число блоков, в которые входит

каждая пара элементов. Указанные параметры связаны соотношением $N = m\,(m-1)\,/\,\sigma + 1$.

Следует отметить, что попарно сбалансированное расположение элементов и блоков, задаваемое симметричными блок-схемами, уже нередко используется в вычислительной технике для описания различного рода взаимодействий процессоров и блоков памяти [17, 22, 26].

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 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. Вся блок-схема интерпретируется как минимальный квазиполный граф, одна доля которого состоит из абонентов, а другая – из коммутаторов. Он описывает идеальную системную сеть с σ-кратным резервированием каналов – РС(N, m, σ). Задающая блок-схему таблица описывает схему межсоединений абонентов и коммутаторов. В таблице 5 приводится пример $B(7, 4, 2)$ и РС(7, 4, 2).

Таблица 5. Схема межсоединений в РС(7, 4, 2).

Блоки 4×4	$B(7, 4, 2)$ РС(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

Таблица 6. Параметры N и m при $\sigma = 1$.

$B(N, m, 1)$ и ПРК(N, m, 1)											
m	2	3	4	5	6	7	8	9	10	**11**	12
N	3	7	13	21	31	43	57	73	91	**111**	133

Таблица 7. Параметры N и m при σ = 2 и σ = 3.

B(*N*, *m*, 2) и ПРК(*N*, *m*, 2)										
m	2	3	4	5	6	7	8	9	10	**11**
N	2	4	7	11	16	22	29	37	46	**56**
B(*N*, *m*, 3) и ПРК (*N*, *m*, 3)										
m	3	4	5	6	7	8	9	10	11	**12**
N	3	5	–	11	15	–	25	31	–	**45**

Для блок-схем существует проблема их построения [3, 13]. В таблицах 6 и 7 приводятся параметры блок-схем $B(N, m, \sigma)$ при малых m и σ. Светлой заливкой выделены блок-схемы, которые не существуют по теории [13], а темной заливкой – блок-схемы, которые еще не построены.

Введение в РС(N, m, σ) разветвителей/объединителей каналов РОКm* или РОКm превращает ее в однокаскадный распределенный полный коммутатор $РК_1(N, m, \sigma)$. На рис. 14 приводится схема $РК_1(7, 4, 2)$, состоящая из коммутаторов 4 × 4 и разветвителей/объединителей дуплексных каналов (РОК4*).

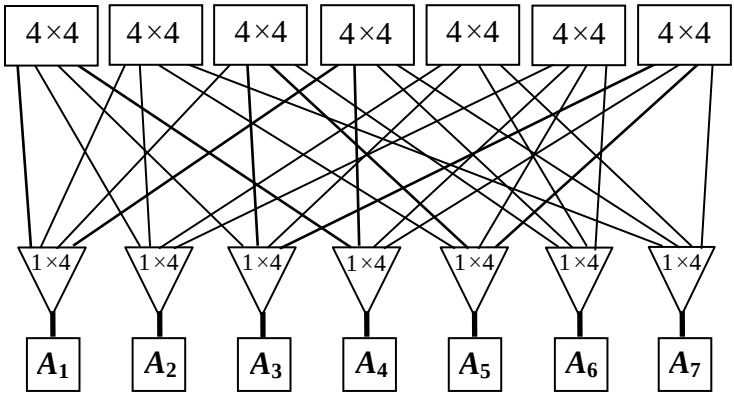

Рис. 14. Схема $РК_1(7, 4, 2)$ *из коммутаторов* 4×4 *и* РОК4*.

$РК_1(N, m, \sigma)$ может быть расширен в k-каскадный распределенный коммутатор $РК_k(R_k, m, \sigma)$ с $R_k > R_1 = N$, в котором любые два абонента связаны путями длины $2k$ не меньше чем σ разными путями через разные коммутаторы $m \times m$. Делается это тем

же методом, что и в предыдущем разделе [4, 7]. Единственное отличие состоит в том, что последняя группа содержит меньше m столбцов. Поэтому в ней размещаются только те абоненты, которые помещаются в имеющихся столбцах и «вручную» добавляется еще один абонент согласно определению блок-схемы.

В таблице 8 приводится пример построения двухкаскадного распределенного полного коммутатора при $m = 4$ и $\sigma = 2$. Здесь «вручную» удалось подключить одиннадцатого абонента. Видно, что некоторые абоненты (например 2 и 8) связаны $\sigma = 2$ путями через однокаскадные коммутаторы $R_1 \times R_1$ (7×7), а некоторые абоненты (3 и 10) – $m = 4$ путями. Можно доказать [10, 7], что m путями связаны те и только те абоненты, номера которых равны по mod N. Это заставляет обозначать k-каскадный распределенный коммутатор $R_k \times R_k$ как $РК_k(R_k, m, \sigma\backslash m)$. Поскольку $\sigma < m$, то число линий ввода/вывода, по которым соединены любые два абонента, не стало меньше σ.

Таблица 8. Табличное описание двухкаскадного $РК_2(11, 4, 2\backslash 4)$.

$РК_2(11, 4, 2\backslash 4)$							
7×7	1-ый $РК_1(7, 4, 2)$				2-ый $РК_1(7, 4, 2)$		
1	1	2	3	4	**8**	**9**	**10**
2	1	2	5	7	**8**	**9**	**11**
3	1	3	5	6	**8**	**10**	**11**
4	1	4	6	7	**8**		
5	2	3	6	7	**9**	**10**	
6	2	4	5	6	**10**		**11**
7	3	4	5	7	**10**		**11**

В таблице 9 приводится пример построения трехкаскадного распределенного полного коммутатора при $m = 4$ и $\sigma = 2$. Здесь «вручную» удалось подключить восемнадцатого абонента. Любые два абонента связаны не менее чем двумя путями длины 6.

Описанная картина сохраняется при добавлении каждого следующего каскада. Поэтому в k-каскадном коммутаторе любые два абонента связаны только путями длины $2k$, проходящими через разные коммутаторы. Число таких путей, проходящих через разные коммутаторы $m \times m$, задается набором $\sigma^r m^{t-r}$

($1 \le r \le k$), в котором нет членов меньше σ. Можно показать [7, 4], что число абонентов R_k задается выражением $R_k \approx m\lfloor (N/m)^k \rfloor \approx m\lfloor (m-1)/\sigma)^k \rfloor$, которое имеет точность в несколько процентов при $k > 2$.

Таблица 9. Табличное описание трехкаскадного $РК_3(18, 4, 2\backslash 4)$.

$РК_3(18, 4, 2\backslash 4)$											
11×11	1-ый ПРК(7, 4, 2)				2-ый ПРК(7, 4, 2)				3-ый ПРК(7, 4, 2)		
1	1	2	3	4	**8**	**9**	**10**	**11**	15	16	17
2	1	2	5	7	**8**	**9**	**12**	**14**	15	16	18
3	1	3	5	6	**8**	**10**	**12**	**13**	15	17	18
4	1	4	6	7	**8**	**11**	**13**	**14**	15		
5	2	3	6	7	**9**	**10**	**13**	**14**	16	17	
6	2	4	5	6	**9**	**11**	**12**	**13**	16		18
7	3	4	5	7	**10**	**11**	**12**	**14**	17		18

Полные распределенные коммутаторы на основе минимального квазиполного **графа** и на основе минимального квазиполного **орграфа** отличаются числом абонентов, используемыми каналами и числом разных путей между любыми двумя абонентами.

Число абонентов в первых в $\sim[m/(m-1)]^k$ раз меньше, чем во вторых. Они используют дуплексные порты вместо симплексных. И, самое главное, они могут иметь σ разных канальных путей между любыми двумя абонентами. Правда, при этом число абонентов уменьшается в $\sim\sigma$ раз.

Последнее свойство можно использовать как для обеспечения отказоустойчивости идеальной СС, так и для повышения ее коммутационных возможностей. Такая идеальная СС является неблокируемой и самомаршрутизируемой при произвольной (неоднородной) σ-перестановке пакетов данных между абонентами, т.е. позволяет любому абоненту вести обмен пакетами данных одновременно по σ каналам с любыми другими абонентами.

3. Заключение

В работе рассмотрен метод построения функционально идеальной СС на любое число абонентов в виде распределенного многокаскадного полного коммутатора. Он обладает свойством неблокируемости и самомаршрутизируемости на произвольной перестановке пакетов данных. Он имеет квадратичную схемную сложность, но содержат значительно меньшее число проводников, чем цельный полный коммутатор.

Построенные идеальные СС ориентированы на две основных области применения – в многоядерных СБИС и в отказоустойчивых МВС реального времени. В первом случае квадратичная схемная сложность не играет особой роли, т.к. сложность коммутатора много меньше сложности ядра, но существенно важно значительное сокращение числа проводников. Во втором случае особую роль выполняет полная однородность СС, которая позволяет иметь любое число резервных процессоров, и возможность резервирования каналов.

Для создания подлинно идеальной СС необходимо сокращение схемной сложности до сложности сетей Клоза или многомерных обобщенных гиперкубов при сохранении свойства неблокируемости и самомаршрутизируемости. Классические сети Клоза не обеспечивают выполнения одновременно обоих этих свойств. Надежду на успех поддерживает наличие подобной сети в виде d-мультиплицированной сети Бенеша (двоичной сети Клоза) [18]. Она, однако, малопригодна для практических применений из-за малого размера коммутаторов ($2d \times 2d$) и, как следствие, сравнительно большой глубины сети ($\sim 2\log_2 N$ каскадов), тогда как современные тенденции [25] требуют построения системных сетей с малой глубиной за счет использования связных СБИС максимально больших коммутаторов.

Отражением этой же тенденции является также сетевая структура самого нового суперкомпьютера Blue Water (IBM) [15], в которой связь между процессорными узлами выполняется не более чем за 3 скачка между связными узлами с промежуточной буферизацией пакетов. Эта сетевая структура строится на основе двухуровневой иерархии полных графов (!) для системы с десятками тысяч процессоров. Замена полных графов на

квазиполные (ор)графы, как в идеальной СС, может дать значительный эффект в части увеличения числа процессоров и/или снижения числа используемых каналов. Это – новая и неожиданная для авторов область применимости идеальных СС.

Литература

1. ГОРБУНОВ В.С. *Архитектура хорошо масштабируемого вычислительного кластера* // Труды международной научно-технической конференции «Суперкомпьютерные технологии: разработка, программирование, применение» (СКТ-2010) Дивноморское. Сентябрь 2010. Т.1. С. 48 – 54.
2. ГОРБУНОВ В.С., ЛАЦИС А.О., ИВАНОВ А.Н. *О построении суперкомпьютеров на основе интерфейса PCI-Express.* // Труды международной научно-технической конференции «Суперкомпьютерные технологии: разработка, программирование, применение» (СКТ-2010) Дивноморское. Сентябрь 2010. Т.1. С. 55 – 57.
3. КАРАВАЙ М.Ф., ПАРХОМЕНКО П.П., ПОДЛАЗОВ В.С. *Комбинаторные методы построения двудольных однородных минимальных квазиполных графов (симметричных блок-схем)* // Автоматика и телемеханика. 2009. №. 2. С. 153 – 170.
4. КАРАВАЙ М.Ф., ПОДЛАЗОВ В.С. *Метод инвариантного расширения системных сетей многопроцессорных вычислительных систем* // Автоматика и телемеханика. 2010. №. 12. С. 166 – 176.
5. КАРАВАЙ М.Ф., ПАРХОМЕНКО П.П., ПОДЛАЗОВ В.С. *Универсальная сетевая структура для отказоустойчивых многопроцессорных систем реального времени* // Труды конференции «Технические и программные средства систем управления, контроля и измерения» (УКИ’10). М. 2010. С. 583 – 597. URL: http://cmm.ipu.ru/proc/index.html (дата доступа – 26.09.2011)..
6. КОРЖ А.А., МАКАГОН Д.В., БОРОДИН А.А. и др. *Отечественная коммуникационная сеть 3D-тор с поддержкой глобально адресуемой памяти для суперкомпьютеров транспетафлопсного уровня производительности* // Междуна-

родная научная конференция «Параллельные вычислительные технологии 2010» г.Уфа, март – апр. 2010. С. 227 – 237.
7. НИКОЛАЕВ А.Б., ПОДЛАЗОВ В.С. *Отказоустойчивое расширение системных сетей многопроцессорных вычислительных систем //* Автоматика и телемеханика.. 2008. № 1. С. 162 – 170.
8. ПОДЛАЗОВ В.С., СОКОЛОВ В.В. *Однокаскадные коммутаторы большой размерности для многопроцессорных и многомашинных вычислительных систем //* Проблемы управления. 2006. № 6. С. 19 – 24.
9. ПОДЛАЗОВ В.С., СОКОЛОВ В.В. *Схемотехника однокаскадных коммутаторов большой размерности //* Датчики и системы. 2006. № 9.С. 12 – 17.
10. ПОДЛАЗОВ В.С., СОКОЛОВ В.В. *Метод однородного расширения системных сетей многопроцессорных вычислительных систем* // Проблемы управления. 2007. № 2. С. 22 – 27.
11. СОЛОХИНА Т.В., ПЕТРИЧКОВИЧ Я.Я., ШЕЙНИН Ю.Е. *Технология Space Wire и бортовых распределенных комплексов //* ЭЛЕКТРОНИКА: Наука, Технология, Бизнес. 2007. №. 1. С. 38 – 49.
12. ШЕЙНИН Ю.Е., СОЛОХИНА Т.В., ПЕТРИЧКОВИЧ Я.Я. *Технология Space Wire и бортовых распределенных комплексов //* ЭЛЕКТРОНИКА: Наука, Технология, Бизнес. 2006. №. 6. С. 64 – 75.
13. ХОЛЛ М. *Комбинаторика. //* М.: Мир. 1970. 421 С.
14. ALVERSON R., ROWETH D. AND KAPLAN L., CRAY INC. *// The Gemini System Interconnect //* 18th IEEE Symposium on High Performance Interconnects. 2009. P. 83 – 87.
15. ARIMILLI B. ARIMILLI R., CHUNG V., et al, *The PERCS High-Performance Interconnect //* 18th IEEE Symposium on High Performance Interconnects. 2009. P. 75 – 82.
16. ARORA S., LEIGHTON F.T., MAGGS B.M. *On-line algorithm for path selection in nonblocking network //* SIAM Journal of Computing 1996. 25(3). P. 600 – 652.
17. BERCOVICH E., BERCOVICH S. *A combinatorial architecture for instruction-level parallelism //* Microprocessors and Microsystems. 1998. V. 32. P. 23 – 31.

18. GU Q.P., TAMAKI H. *Routing a permutation in hypercube by two sets of edge-disjoint paths //* Journal of parallel and distributed computing. 1997. V. 44. No. 2. P. 147 – 152.
19. NI L.M., MCKINLEY P.K. *A survey of wormhole routing techniques in direct networks //* IEEE Computer 1993. V.26. No. 2. P. 62 – 73.
20. *Guide to myrinet-2000 switches and switch networks //* URL: http://www.myti.com/myrinet/m3switch/guide/ (дата доступа – 26.09.2011)
21. KUMAR A., PEH L-S., KUNDU P., JHA N.K. *Toward ideal on-chip communication using express virtual channels //* IEEE Micro. 2008. Jan/Feb. P. 80 – 90.
22. OKBIN L., SANGHO L., SEONGYEOL K., ILYONG CH. *An Efficient Load Balancing Algorithm Employing a Symmetric Balanced Incomplete Block Design //* Lecture Notes on Computer Science 3046. 2004. P. 647 – 654.
23. PIPENGER N. *On rearrangeable and non-blocking switching networks //* Journal Of Computer and Systems Science. 1978. V. 17. P. 307 – 311.
24. RZYMIANOWICZ L. *Designing efficient network interfaces for system area networks //* URL: http://bibserv7.bib.uni-mannheim.de/madoc/volltexte/2002/54/pdf/54_1.pdf (дата доступа – 26.09.2011).
25. SCOTT S., ABTS D., KIM J., AND DALLY W. *The black widow* high-radix Clos network // Proc. 33rd International Symposium on Computer Architecture. (ISCA'2006). 2006. Рукопись доступна на сайте URL: http://cva.stanford.edu/people/jjk12/isca06.pdf (дата доступа – 26.09.2011).
26. YOUNGJOO CH., CHANGKYUN CH., ILYONG CH. *An Efficient Conference Key Distribution System Based on Symmetric Balanced Incomplete Block Design //* Lecture Notes on Computer Science 2657. 2003. P. 147 – 154.

DISTRIBUTED FULL SWITCH AS IDEAL SYSTEM AREA NETWORK FOR MULTIPROCESSOR COMPUTERS

Mikhail Karavay, Institute of Control Sciences of RAS, Moscow, Doctor of Science, assistant professor (mkaravay@ipu.ru, Moscow, Profsoyuznaya st., 65, (495)334-90-00).
Viktor Podlazov, Institute of Control Sciences of RAS, Moscow, Doctor of Science, assistant professor (podlazov@ipu.ru, Moscow, Profsoyuznaya st., 65, (495)334-78-31).

Abstract: *We consider a way to build distributed full switches of arbitrary size consisting of fixed-size switches and channel splitters. The distributed switch preserves nonblocking and self-routing properties of a complete switch and forms an ideal system area network.*

Keywords: massive parallel multiprocessor computer, ideal system area networks, distributed full switch, channel switching, wormhole routing techniques, non-blocking networks, self-routing networks.

Статья представлена к публикации
членом редакционной коллегии В.Н. Лебедевым

3-я Российская конференция
с международным участием
«Технические и программные средства
систем управления, контроля и измерения»
(*УКИ-12*)

ИПУ РАН, Москва, 16-19 апреля 2012г.
http://cmm.ipu.ru