УДК 004.724.2+004.272.43
ББК 3.9.7.3.02

ТОПОЛОГИЧЕСКИЕ РЕЗЕРВЫ СУПЕРКОМПЬЮТЕРНОГО ИНТЕРКОНЕКТА

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

Рассматриваются простые возможности повышения характеристик интерконекта суперкомпьютеров Gemini (CRAY) и Blue Water (IBM) за счет использования системных сетей с прямыми каналами.

Ключевые слова: параллельные многопроцессорные вычислительные системы,системные сети, самомаршрутизируемые сети, прямые каналы, распределелнные полные коммутаторы, некоммутируемые мультикольца.

1. *Введение*

В основе структуры суперкомпьютеров *Gemini и Blue Water* [12, 13] лежит пара тесно связанных узлов – процессорного узла и высокоинтеллектуального связного узла с большим числом каналов для межпроцессорного интерконнекта. *Gemini* и *Blue Water* имеют процессорный узел с 4 многоядерными процессорами (6 и 8 ядер соответственно). Узлы связи имеют 20 и 47 портов высокоскоростных дуплексных каналов соответственно.

Узлы связи *Gemini* объединены в 3*D*-тор. Измерения *x* и *y* состоят из 4-х идентичных дуплексных колец, измерение z – из 2-х дуплексных колец. Все кольца имеют одинаковую пропускную способность. Общее число пар *N* процессорного и связного

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

узлов составляет величину $N=N_xN_yN_z$, где N_i – число пар узлов в каждом кольце i-го измерения.

Скрытым резервом данной связной системы является неоптимальность использования множества колец. В каждом измерении все кольца имеют одинаковую топологию (последовательность соединения узлов). Использование колец с разной топологией открывает возможность существенного (в разы) повышения пропускной способности множества колец каждого измерения [2, 9, 10]. При этом в узле связи меняется только алгоритм выбора кольца для передачи пакета данных. Особенности и характеристики использования колец разной топологии рассматривается во 2-м разделе статьи, а в 3-м разделе они применяются для 3D-тора.

Каждый узел связи *Blue Water* имеет межузловые каналы трех видов: 7 каналов K_1 максимальной пропускной способности V_1, равной пропускной способности межпроцессорных каналов в процессорном узле, 24 канала K_2 пропускной способности $V_2=V_1/5$ и 16 каналов K_3 пропускной способности $V_3=2V_2$. Каналы K_1 выполнены медным кабелем, а каналы K_2 и K_3 – оптическим кабелем.

32 узла связи образуют суперузел, в котором узлы связаны по схеме полного графа каналами K_1 и K_2. Среди них выделяются 4 группы по 8 узлов, связанных каналами K_1. Остальные узлы связаны каналами K_2.

Каждый суперузел имеет 512 каналов K_3. В максимальной конфигурации *Blue Water* каждый такой канал используется для связи с другим суперузлом по схеме полного графа. В этом случае *Blue Water* содержит 513 суперузлов и в них $513*32*4>64K$ процессоров, связанных каналами разной пропускной способности. Передача пакета между любыми двумя узлами занимает не более 3-х смен каналов с промежуточной буферизацией пакетов (скачков).

Скрытым резервом данной системы связи является, во-первых, неоптимальное использование каналов K_1 максимальной пропускной способности для создания суперузла. Дополнительное использование при каждом узле связи коммутатора 7×7 каналов K_1 открывает возможность построения суперузлов с б**о**льшим числом узлов, связанных только каналами K_1, и осво-

бождения каналов K_2 для связи с дополнительными суперузлами. Так построенные суперузлы имеют топологию распределенного полного коммутатора в виде квазиполного графа или оргграфа [4 – 8]. Эта оптимизация позволяет существенного (в разы) увеличить как общее число узлов и процессоров в системе, так и число узлов, связанных каналами максимальной пропускной способности. При этом в узле связи меняется только алгоритм выбора канала K_1 для передачи пакета данных.

Еще одним резервом является неоптимальное использование каналов K_3 для объединения суперузлов в систему. Дополнительно использование при каждом суперузле коммутатора $M \times M$ каналов K_3, где $M=N^{1/2}$ и N – число суперузлов, позволит уменьшить число каналов K_3 за счет замены полного графа на минимальный квазиполный (ор)граф в виде распределенного полного коммутатора.

Особенности и характеристики использования распределенных полных коммутаторов рассматривается в 4-м разделе статьи, в 5-м разделе они используются для описания системы связи внутри суперузла, а в 6-м разделе – между суперузлами.

Метод расширения полного коммутатора (раздел 4), приводящий к построению распределенного полного коммутатора, в разделе 7 применяется к расширению дуплексного кольца (двух встречных колец). Это приводит к построению мультикольца с разреженными кольцами, состящего из нескольких дуплексных колец, которое имеет большую пропускную способность, чем мультикольцо с полными кольцами (раздел 2). В разделе 8 это мультикольцо прменяется для 3*D*-тора *Gemini.*

2. ***Некоммутируемые мультикольца***

Мультикольцом мы называем набор из $n \geq 2$ кратных колец различной топологии (последовательности соединения узлов). В кратном кольце любой пакет удаляется из канала узлом-получателем, а не узлом-отправителем, освобождая тем самым канал для одновременного и параллельного использования другими узлами. Такую пространственную параллельность обеспечивает не любой способ множественного доступа к каналу. Она возможна в сегментированном кольце и в кольце со вставкой

регистра, но невозможна в кольце с передачей жезла (*FDDI, Token Ring*).

В сегментированном кольце по кольцу циркулируют сегменты равной длины, в которых переносятся пакеты данных. Пакет от любого источника передается только в свободный сегмент и доставляется приемнику безо всякой буферизации в промежуточных абонентах. Поэтому сегментированное кольцо (СК) обеспечивает минимальные времена доставки пакетов по сети.

В кольце со вставкой регистра любой источник всегда передает пакет в канал, а идущий по каналу пакет буферизует у себя, задерживая его доставку по каналу. Кольцо со вставкой регистра (КВР) позволяет использовать пакеты произвольной длины.

Оба способа в условиях однородных узлов и равномерного распределения длин маршрутов обеспечивают практически одинаковую пропускную W кольца в модели $M/G/1$ теории массового обслуживания [1]. Она определяется как $W=cv$, где $c > 1$ – емкость кратного кольца, а v(бит/сек) – скорость передачи по кольцу. В СК и КВР зависимость задержки передачи пакета (пребывания в буферах) $T(s)$ от загрузки кольца s имеет сходный вид:

(2.1) $$T(s)=F_k/(c-s).$$

Здесь k=(СК, КВР) – вид кольца, s – загрузка кольца $s=\Lambda B$, где Λ(сек$^{-1}$) – суммарная интенсивность генерации пакетов, $B=b/v$(сек) – средняя длительность пакета, а b(бит) – средняя длина пакета. Функция $F_k(s,b,d)$(сек) зависит от загрузки s и ее дисперсии s (для экспоненциального распределения 1-й и 2-й моменты совпадают), средней длины пакетов b и ее дисперсии d, но не зависит от емкости c. Если в КВР используются пакеты одинаковой длины, равной длине сегмента в СК, то имеют место следующие соотношения: $T(0)_{КВР}=0$ и $T(0)_{СК}=B/2$, $F_{КВР}(c/2)\approx 2B$ и $F_{СК}(c/2)\approx 2{,}5B$.

Последние соотношения показывают, что СК и КВР имеют практически одинаковые задержки доставки пакетов, только в СК для каждого источника они состоят из задержек в его выходном буфере, а в КВР – из задержек во вставляемых промежуточных буферах других источников.

Топология любого кольца задается следующим образом. Предположим, что узлы перенумерованы целыми числами из $[0, N-1]$. Пусть номера соседних узлов вдоль направления передачи задаются последовательностью $X_i \in [0, N-1]$ $(i = 0,1,...)$, в которой $X_{i+1} = (X_i + S) \bmod N$, где S называется шагом кольца и $1 \le S \le N-1$, $0 \le X_0 \le S-1$.

Кольцо с шагом $S \ge N/2$ является встречным кольцом с шагом $-(N-S)$. При $X_0 = 0$ и $S = 1$ получаем традиционное кольцо (с шагом 1), а при $X_0 = 0$ и $S = N-1$ – встречное кольцо с шагом -1. В дальнейшем кольцо с шагом S будем называть кольцом S, а его дугу – дугой S.

Определение 2.1. Мультикольцом $\{S_n\} = ({}^1S, {}^2S, ..., {}^nS)$ называется набор из $2 \le n \le N-1$ различных колец ${}^1S = 1$, ${}^2S$, …, ${}^nS$, где ${}^jS \ne {}^kS$.

Рис. 2.1 дает пример мультикольца с 16 узлами и 4 кольцами.

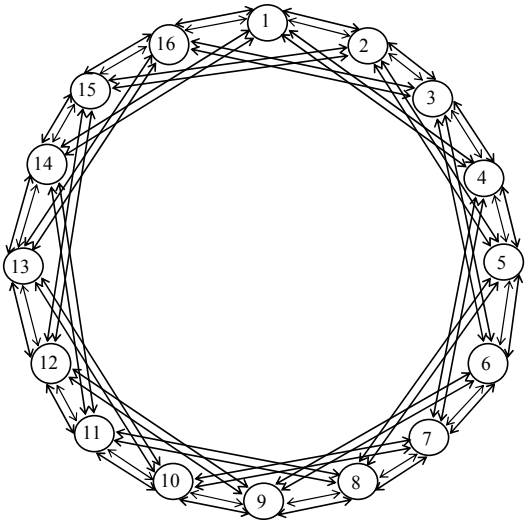

Рис. 2.1 Мультикольцо $\{S_4\} = (1, 3, -3, -1) = (\pm 1, \pm 3)$.

Мультикольцо – это орграф, который обладает центральной симметрией и является симметричным по узлам графом. Это означает, что поворот на любой угол кратный $\delta = 2\pi / N$ сохраняет набор дуг, инцидентных каждому узлу. Такое отображение является автоморфизмом и сохраняет все пути в мультикольце. Поэтому достаточно рассматривать маршруты только из узла с номером 0.

В данном разделе рассматриваемые некоммутируемые сегментированные мультикольца, в которых любой пакет от узла отправления до узла назначения доставляется только по одному кольцу. Выбор кольца для передачи зависит, в первую очередь, от длины пути (числа дуг) между этими узлами. Если в кольце ${}^{j}S$ существует путь от узла отправления ${}^{j}X_{i}$ в узел назначения ${}^{j}X_{i+r}$, то его длина равна r, т.е. длина пути по кольцу 1 называется длиной маршрута между узлами отправления и назначения.

Предполагается, что все узлы генерируют потоки **пакетов одинаковой длины** с одинаковым распределением маршрутов по их длинам. Даже в этом простейшем случае возникает ряд оптимизационных задач. Во-первых, определить какую максимальную пропускную способность имеет заданное мультикольцо, и при каких условиях она достигается. Во-вторых, при заданном числе колец найти мультикольцо с максимальной пропускной способностью. В-третьих, какие наращиваемые наборы мультиколец обеспечивают максимальные пропускные способности при увеличении N.

Пропускная способность кратного кольца наиболее полно характеризуется такой безразмерной величиной как его емкость c, измеряемая как среднее число пакетов, параллельно переданных в одном сегменте за время его прохода по кольцу, в условиях максимальной загрузки, при которой каждый узел передает пакет всегда, когда кольцо на его выходе свободно (свободен проходящий по кольцу сегмент). Для мультикольца $\{S_n\}$ вводятся емкость ${}^{j}c$ каждого кольца ${}^{j}S$ и средняя длина пути ${}^{j}\overline{L}$ в нем. Справедлива формула [1 – 10]:

(2.1) $$ {}^{j}c = N / {}^{j}\overline{L} . $$

Заметим, что для любого некратного кольца, например с передачей жезла, $c \le 1$.

Возможны различные правила выбора кольца для передачи по заданному маршруту (с заданными узлами отправления и назначения). В силу узловой симметрии и сохранения путей результат этого выбора будет одинаковым во всех узлах с одинаковыми маршрутами. Применение некоторого правила создает расписание маршрутов $R(r) = {}^{j}p_r$, где ${}^{j}p_r$ – вероятность назначения пакета в кольцо ${}^{j}S$, а $r = {}^{1}L$ – длина маршрута по кольцу 1. Пусть заданный маршрут в кольце ${}^{j}S$ имеет длину ${}^{j}r$. Простейшее правило в состоит в назначении для него колец, в которых заданный маршрут имеет наименьшую длину, т.е. ${}^{j}p_r = 1/n_r$, если ${}^{j}r = \min\limits_{1 \le k \le m} {}^{k}r$ и n_r – число таких колец, и ${}^{j}p_r = 0$ в противном случае.

В случае равномерного распределения длин маршрутов и использования простейшего правила величины ${}^{j}\overline{L}$ для (2.1) вычисляются следующим образом. Сначала строится расписание маршрутов $R(r)$. В нем величина ${}^{j}l_r$ задает длину пути маршрута длины r в j-ом кольце, а величина ${}^{j}L_r = {}^{j}l_r \, {}^{j}p_r$ является взвешенной длиной. Величина ${}^{j}d = \sum\limits_{r=1}^{N-1} {}^{j}p_r$ задает число маршрутов, назначенных в j-ое кольцо ($\sum\limits_{j=1}^{m} {}^{j}d = N-1$). Тогда

$$ {}^{j}\overline{L} = \frac{\sum\limits_{r=1}^{N-1} {}^{j}L_r}{{}^{j}d}. \qquad (2.2) $$

Емкость одиночного (симплексного) кольца для любого распределения длин маршрутов задается асимптотическим выражением c=2. Емкость одиночного дуплексного кольца, т.е. мультикольца $\{S_2\}$=(±1) зависит от распределения длин маршрутов [2, 9, 10].]. Для равномерного распределения емкость каждого симплексного кольца в $\{S_2\}$ задается асимптотическим

выражением c=4, а суммарная емкость дуплексного кольца – выражением C=8.

При $m > 2$ емкость C мультикольца $\{S_n\}$, рассчитанная как $C = \sum_{j=1}^{n} {}^{j}c$, не подтверждается результатами моделирования. Это происходит из-за простоев колец, возникающих при их неоднородной загрузке, которая, в свою очередь, является следствием их разной емкости и заданного потока пакетов. Результаты моделирования задают некоторую эффективную емкость мультикольца $\widetilde{C}$, которая аналитически рассчитывается следующим образом [2, 9, 10].

Обозначим ${}^{j}q = {}^{j}c / {}^{j}d$ и $q = \min {}^{j}q$, тогда эффективная емкость $\widetilde{C}$ это:

(2.3) $$\widetilde{C} = (N-1)q .$$

Обозначим ${}^{j}L = \sum_{r} {}^{j}L_r$ и $L = \max {}^{j}L$, тогда (2.3) можно переписать в виде:

(2.4) $$\widetilde{C} = N(N-1)/L .$$

В табл. 2.1 дается пример построения расписания маршрутов по простейшему правилу для мультикольца на рис. 2.1.

Таблица 2.1 Назначение колец в $\{S_4\} = (\pm 1, \pm 3)$ *при* $N = 16$.

S	r	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Взвешенные длины				
1	0	***1***	***2***	3	***4***	**5**	6	7	***8***	9	10	11	12	13	14	15	1	2	2	5	2
3	0	***3***	***6***	***9***	***12***	15	2	5	***8***	11	14	1	4	7	10	13	1	2	3	2	2
-3	0	***13***	***10***	**7**	***4***	1	14	11	***8***	5	2	15	12	9	6	3	1	2	3	2	2
-1	0	***15***	***14***	13	***12***	***11***	10	9	***8***	7	6	5	4	3	2	1	1	2	2	5	2

Такое расписание в дальнейшем называется кратчайшим. Верхняя строка в левой части таблицы дает длины маршрутов (по кольцу 1). Первый столбец задает кольца. В каждой строке задаются последовательности прохождения узлов в каждом кольце, считая узлом отправления маршрутов узел 0. Жирным

шрифтом выделены узлы назначения, к которым ведет кратчайший путь. В правой части таблицы приводятся взвешенные длины соответствующих маршрутов. В табл. 2.2 приводится расписание маршрутов $R(r)$, построенное по табл. 2.1.

Таблица 2.2. Кратчайшее расписание $R(r)$ *для* $\{S_4\} = (\pm 1, \pm 3)$.

Эффективная емкость мультикольца $\widetilde{C} = 20$.

S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	0	0,5	1	0	0	0,25	0	0	0	0	0	0	0
3	0	0	1	0	0	1	0	0,25	1	0	0	0,5	0	0	0
-3	0	0	0	0,5	0	0	1	0,25	0	1	0	0	1	0	0
-1	0	0	0	0	0	0	0	0,25	0	0	1	0,5	0	1	1

В общем случае кратчайшее расписание, сохраняет неоднородность загрузки колец, что снижает эффективную емкость мультикольца. Поэтому применяется эвристическая процедура выравнивания загрузки колец. В ней проводится многократное выравнивание эффективных емкостей колец с минимальным и максимальным значениями за счет расщепления некоторых маршрутов на несколько долей для передачи по разным кольцам. Она имеет сложность $O(n)$ и приводит к выровненному расписанию $\widetilde{R}(r)$. В таб. 2.3 приводится такое расписание.

Таблица 2.3. Выровненное расписание $\widetilde{R}(r)$ *для* $\{S_4\} = (\pm 1, \pm 3)$.

Эффективная емкость мультикольца $\widetilde{C} = 21{,}8$.

S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	0	0,5	1	0	0	0,125	0	0	0	0	0	0	0
3	0	0	1	0	0	1	0	0,375	1	0	0	0,5	0	0	0
-3	0	0	0	0,5	0	0	1	0,375	0	1	0	0	1	0	0
-1	0	0	0	0	0	0	0	0,125	0	0	1	0,5	0	1	1

Было проведено исследование условий достижения мультикольцами $\{S_n\}$ с числом колец m максимальной эффективной емкости при $N \leq 124$. Оказалось, что при равномерном распределении длин маршрутов для любого N всегда существует хотя

бы одно мультикольцо с дуплексными каналами (с парами встречных каналов одинакового шага) и выровненным расписанием, для которого

(2.5) $\widetilde{C}(N) = k(N)n^2$, где $k(N) = O(1)$.

Среди таких мультиколец были выделено семейство наращиваемых мультиколец, для которых справедливость формулы (2.5) при увеличении N обеспечивается только добавлением новых дуплексных колец. Оно представлен в таб.2.4.

В таб.2.4 для каждого N приводится только два набор колец, для которых. $\widetilde{C}(N) \approx KN$, где $0{,}5 \le K \le 2$.

Таблица. 2.4. Семейство наращиваемых квазиоптимальных мультиколец.

*N**n*	4	6	8	10	12
16	±1,±3	±1,±2,±3			
32		±1,±2,±3	±1,±2,±3,±7		
64			±1,±2,±3,±7	±1,±2,±3,±5,±7	
128				±1,±2,±3,±5,±7	±1,±2,±3,±5,±7,±9

Оценим расход кабеля $E(n)$ на наращиваемое мультикольцо при условии, что все кабели прокладываются вдоль колец (±1). Примем, что $E(n) = 4 + \sum_{i=1}^{n/2-1} 2(2i-1)$. Тогда $E(n) \approx n^2/2 - 4n$.

3. ***Варианты использования полных мультиколец в 3D-торе***

Сначала рассмотрим 3*D*-тор из *Gemini* [12] с N=16K узлами (N_z=16, N_x=N_y=32) при равномерном распределении длин маршрутов.

Если измерение z состоит, как в [12], из двух идентичных дуплексных колец ±1, то его суммарная емкость составляет C_z=15. Здесь и далее приводятся значения эффективных емкостей, полученные имитационным моделированием. Мультикольцо $\{S_4\} = (\pm 1, \pm 3)$, состоящее из двух дуплексных колец (±1) и (±3), обеспечит эффективную емкость $\widetilde{C}_z = 21{,}8$ (см.

таб.2.3). Таким образом имеет место увеличение пропускной способности колец измерения z в 1,5 раза.

Если измерения x и y состоят, как в [12], из четырех идентичных дуплексных колец с шагом ±1, то их емкость составляет C_x=C_y=30,5. Мультикольцо $\{S_8\} = (\pm 1, \pm 2, \pm 3, \pm 7)$, состоящее из четырех дуплексных колец ±1, ±2, ±3 и ±7 обеспечит при выровненном расписании эффективную емкость $\widetilde{C}_x = \widetilde{C}_y = 65$ (с округлением до целого). Таким образом имеет место увеличение пропускной способности колец этих измерений более чем в 2 раза. Заметим, что кратчайшее расписание обеспечивает только $\widetilde{C}_x = \widetilde{C}_y = 58$.

Рассмотренное мультикольцо $\{S_8\}$ обладает некоторой неоднородностью, состоящей в том, что кольца ±2 расщепляются на два миникольца, проходящие через четные и нечетные узлы. Эту неоднородность можно устранить выбрав простые N_x=N_y=37. В этом случае при выровненном расписании $\widetilde{C}_x = \widetilde{C}_y = 66$.

Такая же картина сохраняется при N_x=N_y=64. В этом случае при выровненном расписании $\widetilde{C}_x = \widetilde{C}_y = 66$. Аналогично, при простых N_x=N_y=67 получается $\widetilde{C}_x = \widetilde{C}_y = 65$.

Главное преимущество увеличения пропускной способности колец состоит в уменьшении очередей в буферах узлов, что приводит к значительному (в разы) сокращению времен доставки пакетов по сети в условиях ее высокой загрузки. Пусть s ($0 < s < C$, где C=C_x=C_y) задает загрузку 4-х пар дуплексных колец ±1. Тогда для измерений x и y сокращение времени доставки τ можно грубо оценить согласно формуле (2.1) выражением τ=(64–s)/(31–s)≈1+1/(31–s). При $s > 0{,}5C$ сокращение времени доставки сокращается в $\tau > 3$ раз, а при $s > 0{,}8\ C$ – в $\tau > 6$ раз.

Главным недостатком мультикольца является повышенный расход связного кабеля. Если прокладывать все кольца мультикольца $\{S_8\} = (\pm 1, \pm 2, \pm 3, \pm 7)$ параллельно кольцу ±1, то расход кабеля по сравнению с 4 дуплексными кольцами ±1 увеличится в

3,25 раз. Расход кабеля можно снизить при хордой прокладке дуг колец (±2, ±3, ±7).

4. *Распределенные полные коммутаторы*

Распределенным полным коммутатором (РПК) мы называем сеть на N абонентов, обладающая функциональным свойством полного графа, но имеющая существенно меньше каналов M ($M << N$) и не являющаяся однокристальным полным коммутатором $N \times N$. Функциональное свойства полного графа – это его неблокируемость и самомаршрутизируемость на произвольной перестановке пакетов данных между абонентами. Это свойство означает возможность соединения произвольных пар источник-приемник «прямыми» каналами без промежуточной буферизации пакетов, которые выбираются каждым источником независимо от других источников.

Указанным свойством обладает сеть в виде минимального квазиполного графа, одну долю которой составляют коммутаторы $m \times m$, а другую – m-портовые абоненты. В одной доле имеется N коммутаторов, а в другой – N абонентов. В ней выбирается минимальное m такое, что между любыми двумя узлами в каждой доле можно было проложить ровно σ разных путей длины 2, проходящих через разные узлы в другой доле. Каждый такой путь между любой парой абонентов проходит через один коммутатор, и разные пути проходят через разные коммутаторы. Пример такой сети приведен на рис.4.1 для m=4, N=7 и σ=2. На рис. 4.1 толстыми линями выделены пути между абонентами, выделенными одинаковой заливкой. Нетрудно видеть, что их два для каждой пары абонентов.

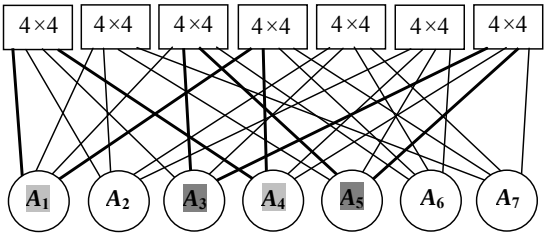

Рис. 4.1. Минимальный квазиполный граф с m=4, N=7 и σ=2.

Здесь возникает вопрос о существовании минимальных квазиполных графов и об их параметрах. Оказывается, что он уже давно решен в комбинаторике. Такие двудольные графы описываются на языке неполных уравновешенных блок-схем, в частности, симметричных блок-схем [4 – 8].

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 $N=m(m-1)/\sigma+1$.

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 $m\times m$, 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 дуплексными каналами. Вся блок-схема интерпретируется как минимальный квазиполный граф, одна доля которого состоит из абонентов, а другая из коммутаторов. Он описывает распределенный полный коммутатор с σ-кратным резервированием каналов – РПК(N, m, σ). По построению он наследует маршрутные свойства полного коммутатора m×m, т.е. является неблокируе-

мым и самомаршрутизируемым. Задающая блок-схему таблица описывает схему межсоединений абонентов и коммутаторов. В табл. 4.2 приводится пример $B(7, 4, 2)$ и РПК(7, 4, 2).

Для блок-схем существует проблема их построения. В табл. 4.3 и 4.4 приводятся параметры блок-схем $B(N, m, \sigma)$ при малых m и σ. Светлой заливкой выделены блок-схемы, которые не существуют по теории, а темной заливкой – блок-схемы которые еще не построены.

Таблица 4.2. Схема межсоединений в РК(7, 4, 2).

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

Таблица 4.3. Параметры N и m при σ=1.

$B(N, m, 1)$ и РК(N, m, 1)											
									0	1	2
			3	1	1	3	7	3	1	11	33

Таблица 4.4. Параметры N и m при σ=2.

$B(N, m, 2)$ и РК(N, m, 2)										
									0	1
				1	6	2	9	7	6	6

Возможностью соединения произвольных пар источник-приемник «прямыми» каналами обладает также полное 2-мерное коммутируемое мультикольцо [3, 4, 7].

Мультикольцо $\{S_M\} = ({}^1S, {}^2S, ..., {}^MS)$ является коммутируемым, если каждый его узел содержит полный коммутатор $m \times m$ ($1 < m < M$) который позволяет перенаправлять пакеты между некоторыми кольцами. Коммутируемое мультикольцо с $N = p^r$ узлами называется полным r-мерным мультикольцом. В нем длины шагов колец задаются цифрами в p-ичной системе счисления, а число колец M задается выражением $M = (p-1)r$.

При r=2 число колец $M = 2(p-1)$, шаги ${}^iS = i$ при $1 \le i \le p-1$ и ${}^jS = (j-p)(p-1)$ при $p \le j \le M$, а в узле используется коммутатор $p \times p$. На рис.4.2 приведен пример полного двумерного мультикольца $\{S_4\} = (1, 2, 3, 6)$ с с **N**=9 узлами.

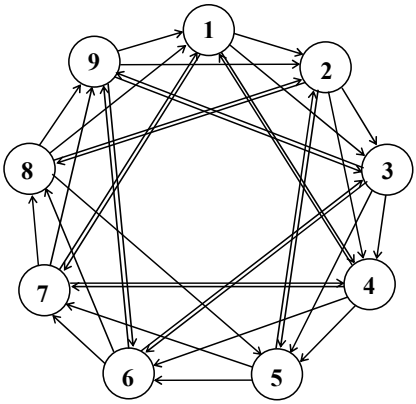

*Рис. 4.2. Полное 2-мерное мультикольцо с **N**=9 узлами.*

Пусть каждый узел полного двумерного мультикольца содержит абонента с $m=p$ входными портами и $m=p$ выходными портами коммутатор $m \times m$ (рис. 4.3).

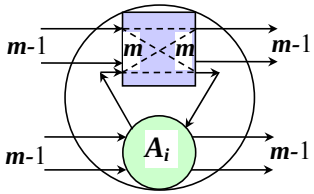

Рис. 4.3. Коммутатор m×m и абонент A_i в составе i-го узла.

Маршрутизация в полном 2-мерном мультикольце осуществляется методом червоточины, т.е. путем прокладки прямого канала между источником и абонентом через промежуточный коммутатор. Эта прокладка осуществляется путем посылки пилотного пакета, содержащего адрес абонента-приемника. Она осуществляется в два этапа – сначала по дугам малых длин ${}^{i}S = i$, а затем – по дугам больших длин ${}^{j}S = (j - p)(p - 1)$. Дуги малых длин идут от абонентов к коммутаторами, а дуги больших длин – от коммутаторов к абонентам. Любой путь по 2-мерному мультикольцу проходит только через один коммутатор. При этом маршрутизация на произвольной перестановке пакетов является неблокируемой и самомаршрутизируемой, и пилотный пакет может быть частью заголовка пакета данных.

Схема межсоединений дуг 2-мерного мультикольца может быть перерисована в виде двудольного орграфа. Одну его долю составляют абоненты, а другую – коммутаторы. Полустепень всех узлов 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). Такой орграф можно назвать минимальным квазиполным орграфом. На рис. 4.3 приведен пример этого орграфа для $m=3$ ($N=9$).

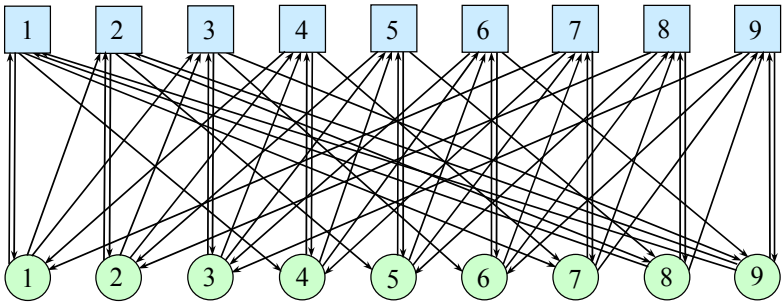

Рис. 4.3. Минимальный квазиполный орграф для полного 2-мерного мультикольца при m=p=3.

Схема межсоединений в полном 2-мерном мультикольце при N=9 задается в таб. 4.1.

Аналогичная регулярная схема межсоединений существует для любого N. В ее таблице на пересечении i-ой строки ($1 \le i \le N$)и j-го столбца ($1 \le j \le m$) в левой части таблицы межсоединений содержится номер $(i–j)\mathrm{mod}N+1$, а в правой – номер $(i+(j–1)m)\mathrm{mod}N+1$. Все соединения задаются симплексными каналами и через симплексные порты.

Таблица 4.1. Таблица межсоединений для полного 2-мерного мультикольца

Коммутаторы 3×3	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**

Каналы в полном мультикольце можно сделать дуплексными, если кольца с шагами в смещенной системе счисления заменить на кольца с шагами в несмещенной системе счисления. Кольца с шагами (1, 2, …, p–1) заменяются на кольца с шагами (±1, ±2, …, $\pm1\lfloor$(p–1)/2$\rfloor$), а кольца с шагами (p, $2p$, …, (p–1)p) – кольцами с шагами ($\pm p$, $\pm2p$, …, $\pm\lfloor(p-1)/2\rfloor p$). Однако при этом порты остаются симплексными, т.к. приемные и передающие порты каждого конца любого дуплексного канала находятся на разных частях сети: один на абоненте, а другой на коммутаторе.

5. *Применение квазиполных (ор)графов в суперузле BLue Water*

Предположим, что каждый связной узел в *Blue Water* [13] дополнен одним полным коммутатором 7×7 для каналов K_1. Пусть выбор выходных портов этого коммутатора осуществляется по их номерам в заголовке каждого пакета. Тогда можно создать суперузел с использованием только каналов K_1 связного узла и этих дополнительных коммутаторов. Такой суперузел имеет структуру квазиполного графа или орграфа (см. рис.4.1 и рис.4.4), в которой связной узел является абонентом.

Сначала попробуем построить суперузел в виде распределенного коммутатора РПК(43, 7, 1) со структурой квазиполного графа. «В чистом виде» эта попытка неосуществима, т.к. блок-схема B(43, 7, 1) (см. таб.4.3) не существует. Однако, если допустить, что некоторые абоненты связаны параллельно более чем через один коммутатор 7×7, то можно построить РПК(39, 7, 1\2). Схема дуплексных межсоединений для него обладает тем свойством, что каждый абонент i связан дополнительным путем через один коммутатор с 4 абонентами, чьи номера задаются как ($i\pm1$)mod39 и ($i\pm2$)mod39.

Таким образом РПК(39, 7, 1\2) позволяет создать суперузел из 39 связных узлов, объединенных неблокируемыми каналами K_1. Это позволит освободить каналы K_2 каждого узла в каждом суперузле для связи с другими суперузлами. Их можно использовать двояко – оставить по одному каналу K_2 или K_3 между любой парой суперузлов или удвоить число таких каналов. В пер-

вом случае число суперузлов увеличится с 513 до 1561=39*(16+24)+1, т.е. более чем втрое. Во втором случае число суперузлов увеличится только до 781, но появляется возможность снизить «длину» резервного пути между любыми связными узлами с 5 до 3 скачков.

Теперь, если построить суперузел в виде распределенного коммутатора со структурой квазиполного орграфа, то число узлов в нем возрастет до 49, а число суперузлов – до 1960 или до 980 при отсутствии или при наличии резервирования каналов соответственно. При этом схема межсоединений абонентов и коммутаторов задается таблицей, аналогичной табл. 4.1, в которой на пересечении i-ой строки ($1 \le i \le 49$) и j-го столбца ($1 \le j \le 7$) в левой части таблицы межсоединений содержится номер $(i-j) \bmod 49+1$, а в правой – номер $(i+7(j-1)) \bmod 49+1$.

6. *Применение квазиполных (ор)графов для связи между суперузлами в Blue Water*

Аналогично можно оптимизировать структуру связей между суперузлами посредством замены полного графа на минимальный квазиполный (ор)граф. При этом «длина» пути между любыми узлами любых суперузлов остается не превосходящей 3-х скачков. Для этого достаточно дополнить каждый суперузел с M узлами коммутатором $M \times M$ каналов K_3 и подсоединить каждый узел в суперузле одним (!) каналом K_3 к этому коммутатору. При этом число суперузлов N может достичь величины $N=M^2$. Правда, для квазиполного графа возникает проблема его построения – для больших M она решена только для случая, когда $M-1$ является простым числом [5]. Поэтому для случая суперузла с 39 узлами придется использовать коммутатор с M=42. У авторов имеются блок-схемы таких квазиполных графов до M=32, и имеется возможность построить их для больших M (38, 42, 44 и т.д.).

В отличие от предыдущего раздела, здесь возникает ряд проблем, для решения которых авторы недостаточно компетентны.

Во-первых, коммутатор $M \times M$ должен быть оптоэлектронным – оптическим по внутренним каналами и электронным по

управлению их коммутацией. Существуют ли такие коммутаторы достаточно большого размера (на 30 – 50 портов) – это открытый вопрос, но он уже в «повестке дня» современной технологии [14].

Во-вторых, использование в узле одного канала K_3 для связей между суперузлами явно недостаточно для сохранения высокой пропускной способности всей системы связи. Число таких каналов можно увеличивать вместе с добавлением коммутаторов $M \times M$. Но их число и их распределение между суперузлами – это открытый вопрос. Здесь есть возможности увеличения числа каналов K_3, используемых параллельно: для увеличения числа суперузлов, связанных к каждым суперузлом, для увеличения пропускной способности связей между суперузлами и для повышения отказоустойчивости этих связей.

7. *Мультикольца с разреженными кольцами*

Построение сетей на основе квазиполного графа позволяет расширять с сохранением маршрутных свойств любую сеть [6, 11], а не только полный коммутатор, как в разделе 3. В данном разделе рассматривается способ расширения сети, состоящей из дуплексного кратного кольца – двух встречных колец с шагом (±1). Он приводит к построению мультиколец, состоящих из дуплексных колец, к каждому из которых подсоединена только часть абонентов. Они называются мультикольцами с разреженными кольцами, в противоположность мультикольцам из раздела 2, которые содержат полные кольца, к каждому из которых подсоединены все абоненты.

Сам способ расширения сети, состоящей из дуплексного кольца, состоит в следующем.

Пусть к дуплексному кольцу подсоединено K абонентов. Каждый абонент, подсоединяется к каждому однонаправленному кольцу через дуплексный порт. Берется $N=m(m-1)+1$ дуплексных колец, к которым подсоединяются абоненты с $2m$ дуплексными портами. Каждое из N дуплексных колец разбивается на равные части по m портов. Каждая часть нумеруется в каждом дуплексном кольце в диапазоне $1 \le j < \lceil K/m \rceil$. Все j-ые части в

каждом дуплексном кольце составляют *j*-ое простейшее мультикольцо– ПМК(*N* ,*m* ,1). В нем подсоединение к каждому однонаправленному кольцу в дуплексном кольце описывается блок-схемой *B*(*N,m,*1) и задается соответствующим квазиполным графом. На рис. 7.1 приводится ПМК(7, 3, 1).

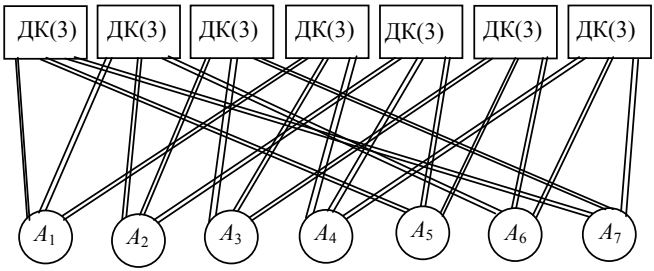

Рис.7.1. ПМК(7, 3, 1). ДК(3) – дуплексное кольцо с 3 абонентами.

1-ай ПМК(*N* ,*m* ,1)должен иметь стандартную структуру. К *j*-му ПМК(*N* ,*m* ,1) подсоединяются абоненты с номерами от *jN* до *N*(*j*+1)–1 так, чтобы абоненты на одинаковых позициях имели номера на *N* больше, чем номера в (*j*–1)-ом ПМК(*N* ,*m* ,1).

В табл. 7.1 приводится схема подсоединений абонентов к дуплексным кольцами при *K*=10 и *m*=2, а в табл. 7.2 – при *K*=16 и *m*=4.

Таблица 7.1. Таблица подсоединений абонентов к кольцам при K=10 *и m*=2 *(ДК – дуплексные кольца, П – их порты, ПМК – ПМК(3,2,1))*

	1-й ПМК		2-й ПМК		3-й ПМК		4-й ПМК		5-й ПМК	
ДК/П	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	1	3	**4**	**6**	7	9	**10**	**12**	13	15
2	2	1	**5**	**4**	8	7	**11**	**10**	14	13
3	3	2	**6**	**5**	9	8	**12**	**11**	15	14

Таблица 7.2. Таблица подсоединений абонентов к кольцам при K=16 и m=4 (*ДК – дуплексные кольца, П – их порты)*

	1-й ПМК(13,4,1)				2-й ПМК(13,4,1)				3-й ПМК			4-й ПМК	
ДК/П	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	...	*15*	*16*
1	1	13	11	5	**14**	**26**	**24**	**18**	27	39	...	50	44
2	2	1	12	6	**15**	**14**	**25**	**19**	28	27	...	51	45
3	3	2	13	7	**16**	**15**	**26**	**20**	29	28	...	52	46
4	4	3	1	8	**17**	**16**	**14**	**21**	30	29	...	40	47
5	5	4	2	9	**18**	**17**	**15**	**22**	31	30	...	41	48
6	6	5	3	10	**19**	**18**	**16**	**23**	32	31	...	42	49
7	7	6	4	11	**20**	**19**	**17**	**24**	33	32	...	43	50
8	8	7	5	12	**21**	**20**	**18**	**25**	34	33	...	44	51
9	9	8	6	13	**22**	**21**	**19**	**26**	35	34	...	45	52
10	10	9	7	1	**23**	**22**	**20**	**14**	36	35	...	46	40
11	11	10	8	2	**24**	**23**	**21**	**15**	37	36	...	47	41
12	12	11	9	3	**25**	**24**	**22**	**16**	38	37	...	48	42
13	13	12	10	4	**26**	**25**	**23**	**17**	39	38	...	49	43

Пропускная способность мультикольца с разреженными кольцами зависит от его емкости C, которая определяется как $C \approx 8N$ и совпадает с эффективной емкостью. Число абонентов R этого мультикольца зависит от числа абонентов K исходного дуплексного кольца, расширением которой и было построено мультикольцо, т.е.

(7.1) $$R \approx NK/m \approx Km.$$

Здесь для целей сравнения приходится выбирать $K=4m$. Длина кабеля в мультикольце с разреженными кольцами составляет $E=2N$ длин кольца (1) или (–1).

Рассмотренные характеристики сведены в табл. 7.3. Видно, что в равных условиях мультикольцо с разреженными кольцами в два раза превосходит мультикольцо с полными кольцами по емкости при немного большем расходе кабеля.

Таблица 7.3. Характеристики двух видов мультиколец.

Характеристика	Полные кольца	Разреженные кольца
Число портов абонента	$2m$	$2m$
Число абонентов	$4m^2$	$4m^2$

Эффективная емкость	$\sim 4m^2$	$\sim 8m(m-1)$
Длина кабеля	$\sim 2m(m-4)$	$\sim 2m(m-1)$

Результаты сравнения показывают, что способ инвариантного расширения произвольных сетей оказался весьма эффективным не только для коммутаторов (раздел 4), но и для некоммутируемых кольцевых сетей.

8. *Варианты использования разреженных мультиколец в 3D-торе*

Мультикольца с разреженными кольцами могут применяться в 3D-торе *Gemini* [12] (N_z=16, N_x=N_y=32).

Для измерения *z* пусть m=2. Возьмем *3* дуплексных кольца на 10 абонентов. Тогда мультикольцо объединит 15 абонентов (табл. 7.1), а его емкость составит 3×7,5=22,5. Это в 1,5 раз больше, чем в измерении *z Gemini,* и чуть больше, чем в мультикольце с полными кольцами (раздел 3).

Для измерений *x* и *y* пусть m=4. Возьмем 13 дуплексных колец на 16 абонентов. Тогда мультикольцо объединит 35 абонентов (табл. 7.2), а его емкость составит 13×7,5=97,5. Это в 3 с лишним раза больше, чем в измерениях *x* и *y Gemini,* и в 1,5 раза больше, чем в мультикольце с полными кольцами (раздел 3).

Общее число процессорных и связных узлов в 3*D*-торе рассмотренной структуры может достигать N=$N_zN_xN_y$=18375, т.е. быть больше, чем в 3*D*-торе *Gemini*. При этом число портов связных узлов остается неизменным, т.е. равным 20.

9. *Заключение*

Некоммутируемое мультикольцо и распределенный полный коммутатор разрабатывались авторами как самостоятельные системными сети для МВС с несколькими десятками или сотнями абонентов. Они разрабатывались как системные сети с минимальными временами доставки пакетов данных по прямым каналам. Данная работа показывает, что их можно эффективно

использовать в качестве компонент в более масштабных и изощренных системных сетях.

Литература

1. АНДРЕЕВ Л.В. *Однонаправленные кольцевые сети связи с коммутацией пакетов // Проблемы передачи информации. 1982. Т. 18. Вып. 4. С. 85 – 103.*
2. АЛЛЕНОВ А.В., ПОДЛАЗОВ В.С., СТЕЦЮРА Г.Г. *Пропускная способность набора кольцевых каналов. I. Класс наборов колец. Наборы с простыми узлами // Автоматика и телемеханика. 1996. №. 3. С. 135 – 144.*
3. АЛЛЕНОВ А.В., ПОДЛАЗОВ В.С. *Пропускная способность набора кольцевых каналов II. Кольцевые коммутаторы // Автоматика и телемеханика. 1996. №. 4. С. 162 – 172.*
4. 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. http://cmm.ipu.ru/proc/index.html.*
5. 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.*
6. 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.*
7. КАРАВАЙ М.Ф., ПОДЛАЗОВ В.С**.** *Распределенный полный коммутатор как «идеальная» системная сеть для многопроцессорных вычислительных систем // Управление большими системами. Выпуск 34. М:. ИПУ РАН. 2011. С. 92 – 116. URL: http://ubs.mtas.ru/upload/library/UBS3405.pdf*

8. КАРАВАЙ М.Ф., ПОДЛАЗОВ В.С., СОКОЛОВ В.В. *Метод расширения полных коммутаторов в фиксированном схемном базисе // Труды 5-й международной конференции «Параллельные вычисления и задачи управления» (PACO'2010). М. 2010. С. 295 – 305.* http://paco.ipu.ru/pdf/A205.pdf.
9. ПОДЛАЗОВ В.С., ПОДЛАЗОВА А.В. *Обеспечение наращиваемости отказоустойчивых многопроцессорных систем с общей памятью с использованием многокольцевых некоммутируемых сетей связи с неоднородными узлами // Труды Института проблем управления РАН. 2002. т. XVIII. С. 164 – 181.*
10. ПОДЛАЗОВ В.С. *Наращиваемые многокольцевые некоммутируемые сети связи для многопроцессорных вычислительных систем // Проблемы управления. 2006. №. 2. С. 50 – 57.*
11. ПОДЛАЗОВ В.С., СОКОЛОВ В.В. *Метод однородного расширения системных сетей многопроцессорных вычислительных систем // Проблемы управления. 2007. №. 2. С. 22 – 27.*
12. ALVERSON R., ROWETH D. AND KAPLAN L., CRAY INC. *The Gemini System Interconnect // 18th IEEE Symposium on High Performance Interconnects. 2009. P. 83 – 87.*
13. ARIMILI B., ARIMILI R., CHUNG V., et al. *The PERCS High-Performance Interconnect // 18th IEEE Symposium on High Performance Interconnects. 2009. P. 75 – 82.*
14. RITER M.B., VLASOV Y., KASH J.A., AND BENNER A. *Optical technologies for data communication in large parallel systems // Topical Workshop on Electronics for Particle Physics (TWEPP-10). 2010. Aachen. Germany. URL: http://iopscience.iop.org/1748-0221/6/01/C01012/pdf/1748-0221/6/01/C01012.pdf.*

TOPOLOGICAL RESERVE OF SUPERCOMPUTER INTERCONNECT

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: *Consider the simple capabilities of supercomputer interconnect characteristics improvement owing to utilization of the direct channels system area networks. Consider the interconnect of supercomputers* Gemini *(CRAY) and Blue Water (IBM).*

Keywords: massive parallel multiprocessor computer, system area networks, self-routing networks, direct channels, distributed full switches, nonswitch multirings.