УДК 004.724.2+004.272.43
ББК 3.9.7.3.02

РАСШИРЕННЫЙ ОБОБЩЕННЫЙ ГИПЕРКУБ КАК ОТКАЗОУСТОЙЧИВАЯ СИСТЕМНАЯ СЕТЬ ДЛЯ МНОГОПРОЦЕССОРНЫХ СИСТЕМ

Каравай М.Ф.[1], Подлазов В.С.[2]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН*, *Москва*)

Рассматривается новая структура системной сети для высокопроизводительных многопроцессорных вычислительных систем. Рассматривается системная сеть в виде расширенного обобщенного гиперкуба, в строках которого связи с топологией полного графа заменены на связи с топологией квазиполного графа, имеющего много меньше ребер.

Ключевые слова: высокопроизводительные многопроцессорные вычислительные системы, системные сети, прямые каналы, неблокируемые самомаршрутизируемые сети, идеальные сети, распределенные полные коммутаторы, обобщенные гиперкубы, сервер *PERCS*.
Keywords: massive parallel multiprocessor computer, system area networks, direct channels, nonblocking self-routing networks, ideal networks, distributed full switches, generalized hypercubes, server PERCS.

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

1. Введение

Одно из направлений построения высокопроизводительных многопроцессорных вычислительных систем (суперкомпьютеров) предполагает использование многопроцессорных и многоядерных (тяжелых) процессорных узлов [8, 9]. Такие процессорные узлы используются в тесной связке с многопортовыми связными узлами. Множество портов в них требуется для возможно большего распараллеливания системной сети, объединяющей связные узлы. Здесь возникает задача эффективного использования заданного множества портов или даже задания этого множества. В работе [1] предложено одно решение этой задачи на основе использования системных сетей с прямыми каналами [1, 7]. Оно основывается на маршрутно-инвариантном расширении таких простейших сетей как кольца и полные коммутаторы и осуществляется посредством замены топологии связей полного графа на квазиполный (ор)граф [2, 3]. Этот подход привел к построению распределенных полных коммутаторов и некоммутируемых мультиколец, пропускная способность которых пропорциональна квадрату числа портов связных узлов сети.

В данной работе вышеупомянутый подход применяется для построения системной сети со структурой расширенного обобщенного гиперкуба. В обычном обобщенном гиперкубе узлы в каждой строке (столбце) любого измерения имеют связи с топологией полного графа. В расширенном обобщенном гиперкубе связи каждого измерения имеют топологию квазиполного (ор)графа. Полный и квазиполный графы имеют одинаковые маршрутные свойства на перестановочном трафике, но квазиполный граф содержит много меньше ребер. Это свойство позволяет многократно увеличивать пропускную способность и отказоустойчивость системной сети при сохранении числа узлов или увеличивать число узлов и отказоустойчивость при сохранении числа каналов в сети. При этом фактически сохраняется диаметр сети и, как следствие, задержка передачи данных по сети.

2. *Обобщенный гиперкуб*

Обобщенный гиперкуб [10, 13] является «кубическим» аналогом многокаскадной сети Клоза, так же как гиперкуб является «кубическим» аналогом многокаскадной сети Бенеша. Как сеть Бенеша является двоичной сетью Клоза, так и гиперкуб является двоичным обобщенным гиперкубом.

Обобщенный гиперкуб обычно задается как d-мерный s-ичный гиперкуб, который имеет $V=s^d$ узлов степени $d(s–1)$ каждый, размещенных в строках (столбцах) по s узлов, задающих «ребра» d-мерного простого куба. «Ребра» здесь понимаются не в графовом, а в геометрическом смысле. Наоборот в графовом смысле «ребро» есть полный граф, т.е. все s узлов одного геометрического «ребра» связаны $s(s–1)$ графовыми ребрами.

Будем обозначать обобщенный d-мерный s-ичный гиперкуб как ОГК(V, d, s). На рис. 1 приведен пример ОГК(16, 2, 4).

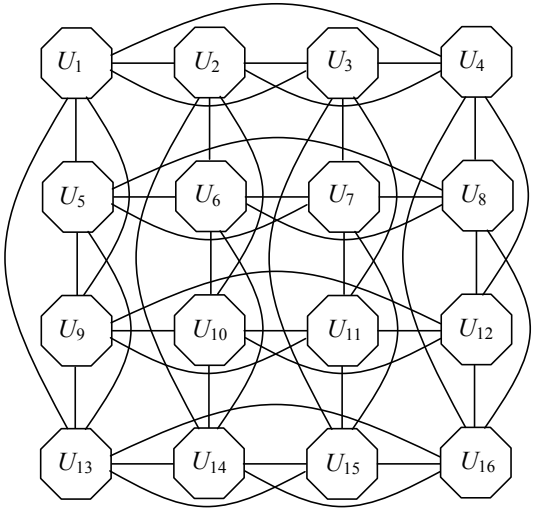

Рис. 1. Структура связей ОГК(16, 2, 4).

Диаметр ОГК(V, d, s) равен $D=d$ и величина бисекции $B=V$.

Для реализации системной сети с топологией ОГК(*V, d, s*) каждый узел должен иметь схемную реализацию, представленную в самом общем виде на рис.2. Узел U_i содержит абонента (процессорный узел) A_i, «коммутатор измерений» $g \times g$ ($g=d+\pi$) и d коммутаторов каналов $1 \times m$ ($m=s-1$) для реализации ребер полного графа в каждой строке (каждом столбце). Здесь π задает число каналов между процессорным и связным узлом. Для простоты в схеме на рис. 2 не показаны входные-выходные буферные очереди между коммутатором измерений и коммутаторами каналов, которые несущественны при описании топологии системной сети. В суперкомпьютерах [8, 9] коммутаторы измерений и каналов входят в состав связного узла.

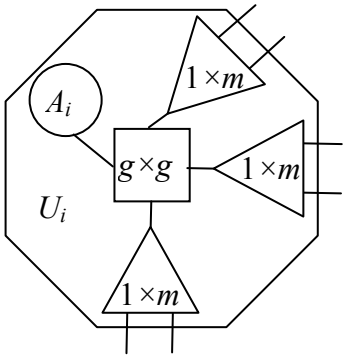

Рис. 2. Схемная реализация узла в ОГК(V, 3, m+1) .

3. Идеальная сеть и распределенный полный коммутатор

Рассмотрим однородный двудольный граф, каждую долю которого составляют N узлов степени m. Значение m выбирается минимальным, при котором любые два узла в одной доле связаны σ путями длины 2 через разные узлы в другой доле. Такой граф мы называем минимальным квазиполным графом [2]. Если он существует, то его параметры связаны соотношением $N=m(m-1)/\sigma+1$.

В данной статье предполагается, что узлами одной доли являются полные коммутаторы $m \times m$, а другой доли – m-

портовые абоненты (связные узлы). Каждый путь между абонентами проходит через один коммутатор, и разные пути проходят через разные коммутаторы. Пример такого графа приведен на рис. 3 для m=4, N=7 и σ=2. На рис. 3 толстыми линиями выделены пути между абонентами, выделенными одинаковой заливкой – их два для каждой пары абонентов.

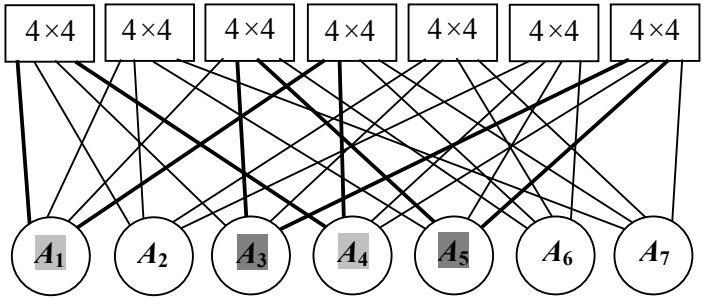

Рис. 3. Минимальный квазиполный граф с m=4, N=7 и σ=2. Простейшая сеть ПС(7, 4, 2).

Здесь возникает вопрос о существовании минимальных 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 [1 – 4].

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 $m \times m$ 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. При этом все абоненты связаны между собой прямыми каналами (через коммутаторы), как в полном графе. В отличие от полного графа рассматриваемый 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 минимальный квазиполный граф, одна доля которого состоит из абонентов, а другая – из коммутаторов. Он описывает «простейшую» [2, 7] системную сеть с σ-кратным резервированием каналов. Задающая блок-схему таблица описывает схему межсоединений абонентов и коммутаторов. На рис. 3 приводится пример ПС(7, 4, 2), в табл. 1 – описание $B(7, 4, 2)$ и ПС(7, 4, 2).

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

ПС(N, m, σ) является «идеальной» сетью, которая имеет возможность использования прямых каналов (без промежуточной буферизации пакетов) для бесконфликтной реализации произвольной перестановки пакетов данных между узлами [2, 3].

Введение в ПС(N, m, σ) коммутаторов $1 \times m$ дуплексных каналов (разветвителей/объединителей каналов – РОКm) превращает ее в распределенный полный коммутатор РК(N, m,

σ), на который у авторов имеется патент [4]. На рис. 4 приводится схема РК(7, 4, 2), состоящая из коммутаторов 4×4 и РОК4.

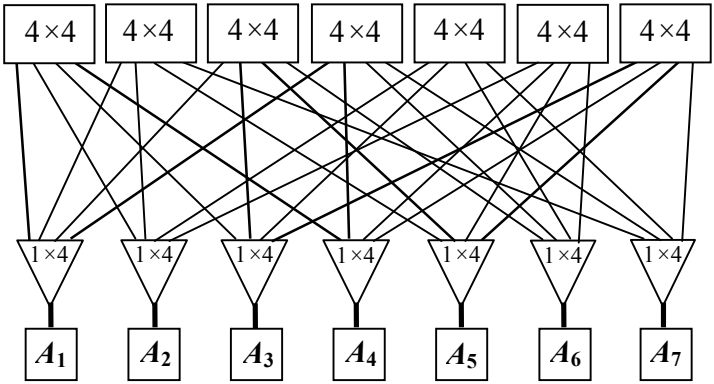

Рис. 4. Схема РК(7, 4, 2) из коммутаторов 4×4 и РОК4.

По построению распределенный полный коммутатор РК(N, m, σ) является неблокируемым самомаршрутизируемым коммутатором $N\times N$ как и исходный коммутатор $m\times m$. Это означает, что произвольная перестановка пакетов данных между абонентами может осуществляться в нем бесконфликтно по прямым (без промежуточной буферизации пакетов) каналам. Каждый абонент прокладывает свой канал самостоятельно – независимо от других абонентов. Обычно предполагается, что прокладка канала осуществляется динамически посредством червячной маршрутизации. Однако возможна и статическая маршрутизация по заранее составленным маршрутным таблицам, которые необходимо хранить в каждом коммутаторе $m\times m$ [7].

С формальной точки зрения диаметр ПС(N, m, σ) и РК(N, m, σ) равен 2 (D=2). Однако диаметр можно выражать и в числе «скачков» D^* (передач по прямым каналам без промежуточной буферизации пакетов). Такой диаметр D^*=1.

В ПС(N, m, 1) и РК(N, m, 1) величина N является нечетным числом. Поэтому величину бисекции B определим как минимальное число дуплексных каналов «точка-точка» между

множествами из $\lceil N/2 \rceil$ и $\lfloor N/2 \rfloor$ абонентов и тогда $B=N+1$. Если же определять величину бисекции B^* в числе прямых каналов, то и в этом случае $B^*=N+1$, т.к. прямой канал является симплексным каналом и два ребра, составляющие путь между абонентами, содержат два встречных прямых канала.

ПС(N, m, 1) имеет топологию квазиполного графа и содержит $w=Nm$ дуплексных каналов «точка-точка» между коммутаторами и абонентами. Сеть с топологией полного графа, содержащая N узлов, имеет $W=N(N-1)$ дуплексных каналов «точка-точка» между абонентами. Легко проверить, что $W/w > \sqrt{N-1}-1$, т.е. имеет место сокращение числа каналов в $\sim\sqrt{N}$ число раз. РК(N, m, 1) содержит $N(m+1)$ каналов за счет использования РОКm.

Квазиполный граф существует не при любых значениях параметров m и σ. Эта проблема может быть разрешена двояко. При $\sigma=1$ можно использовать квазиполный орграф, который существует при любых значениях m и имеет $N=m^2$ [1, 7]. Дополнительным ограничением для его использования в качестве ПС(N, m, 1) служит невозможность использования дуплексных портов. При $\sigma \geq 1$ можно использовать 1-расширенные квазиполные графы, которые удалось построить для всех экспериментально проверенных значений m и σ [5, 7]. В них число узлов N^* каждой доли задается выражением $N^*=N-\delta$, где $\delta < m$. В 1-расширенном квазиполном графе малая часть узлов одной доли связаны $\sigma+1$ путями длины 2, а остальные – σ путями длины 2. В матрице смежности такого графа номера узлов, связанных $\sigma+1$ путями, размещаются на 2δ диагоналях. Системная сеть с топологией 1-расширенгго квазиполного графа обозначается как ПС(N^*, m, $\sigma|\sigma+1$). Дополнительно в ней за счет выбора значения δ можно задавать четность N^*.

Перечисленные выше свойства показывают, что на перестановочном трафике системная сеть с топологией квазиполного графа по пропускной способности и задержкам практически не уступает сети с топологий полного графа, имея много меньшую канальную сложность.

4. Сетевые характеристики простейшей сети ПС(N, m, σ)

По другому может обстоять дело при трафике общесетевого вида, когда несколько источников могут параллельно обращаться к одному приемнику. В сети с топологией полного графа каждый приемник может принять N–1 параллельных пакетов, а в сети с топологией квазиполного графа – только m, т.е. в $\sim\sqrt{N}$ меньше, что приведет к уменьшению пропускной способности и к увеличению задержек передачи. Это уменьшение объяснимо, поскольку в случае полного графа нет конфликта доступа к абоненту при любом распределении адресов назначения. Для квазиполного графа конфликты могут возникать в локальном коммутаторе при одновременном обращении через него разных источников к одному и тому же приёмнику. Возникает вопрос – во сколько раз падает пропускная способность и растет задержка передачи? Эти характеристики можно оценить имитационным моделированием на простой модели с трафиком, состоящим из пакетов одинаковой длины со случайными адресами приемников.

При моделировании каждый источник генерирует пакет с адресом приемника, который (адрес) распределен по степенному закону, т.е. с вероятностью p_i выбора i-го приемника, задаваемой как $p_i=g(a,N)(i/N)^{a-1}$ $(1 \le i \le N)$, где $g(a,N)$ – нормировочный множитель, который определяется соотношением $g(a,N)\sum_{i=1}^{N} p_i = 1$. С увеличением a все большая часть источников адресуются к N-му приемнику. Это распределение включает: равномерное распределение – a=1 и $g(1,N)=1$, линейное распределение – a=2 и $g(2,N)=2/(N+1)$, параболическое распределение – a=3 и $g(3,N)=6/[(N+1)(2N+1)]$, и т.д.

В модели случайный адрес при заданном a находится как $\max(u_1, ..., u_i, ..., u_a)$, где u_i $(1 \le i \le a)$ – случайное целое с равномерным распределением на [0, N–1].

Каждый источник может иметь пакет не более чем к одному приемнику, т.е. моделируется случай σ=1 с одним путем между любой парой абонентов. В модели все источники

действуют синхронно по тактам, передавая в каждом такте по одному пакету. Если несколько источников адресуются к одному приемнику через один и тот же коммутатор, то пакет передает только один из них, а остальные задерживают передачу до следующего такта. После каждого такта источники, которые не имеют пакетов, заново их генерируют с заданным распределением адресов приемников.

Исследовались два базовых режима: с постоянным распределением адресов (с постоянным a – режим ПР) по источникам и со случайным выбором закона распределения для каждого источника на каждом такте (с равномерным распределением a на [1, N] – режим СР). В каждом режиме еще предусматривается перемешивание адресов приемников в виде случайного выбора их адресов на каждом Δ-ом такте (случайный приемник – СПΔ). Он осуществляется как сдвиг назначенных адресов на одинаковый случайный шаг (с равномерным распределением на [0, N–1]). При Δ=0 изменений адресов приемников фактически не производится, при Δ=1 оно осуществляется на каждом такте. По мнению авторов режим СРСПΔ соответствует условиям плохого (случайного) пространственно-временного размещения данных по узлам сети. Подобное размещение данных по оперативной памяти вычислительной системы с кэш памятью приводит к сильной деградации её производительности.

В сети с топологией полного графа за каждый такт передается N–1 пакетов, а в сети с топологией квазиполного графа – случайное число пакетов η ($0< \eta < N$). Измерялось среднее значение η^* за большое число тактов в установившемся режиме. Результаты моделирования в режиме ПР для малых m в представлены в графиках на рис. 5 и 6.

Ось абсцисс задает значение показателя степени a, а ось ординат – отношение $\rho=\eta^*/N$. Рис. 5 задает графики режима ПРСП0, а рис.6 – режима ПРСП1.

Выбор малых m объясняется тем, что для них легко построить простейшие и 1-расширенные простейшие сети ПС(N, m, 1) и ПС(N^*, m, $\sigma|\sigma+1$). Здесь и далее для m=7 и m=11 используются 1-расширенные ПС(39, 7, 1|2) и ПС(95, 11, 1|2),

т.к. ПС(42, 7, 1) и ПС(111, 11, 1) или не существует или еще не построена соответственно.

Если обозначить пропускную способность сети с топологией полного графа как W, а сети с топологией квазиполного как w, то рис.5 и рис.6 показывают, что при малых m в режиме ПРСП0 $w/W=\rho \geq 0{,}3$, а в режиме ПРСП1 $w/W=\rho \geq 0{,}89$.

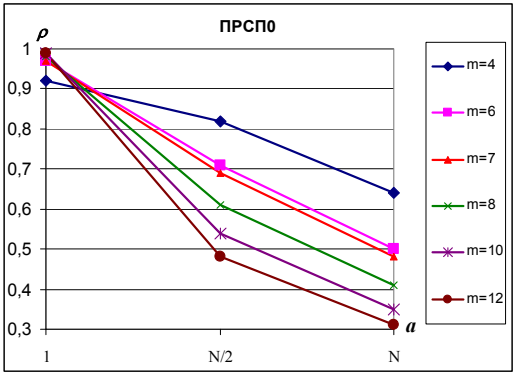

Рис. 5. Отношение w/W пропускных способностей сетей с топологией квазиполного и полного графов в режиме ПРСП0.

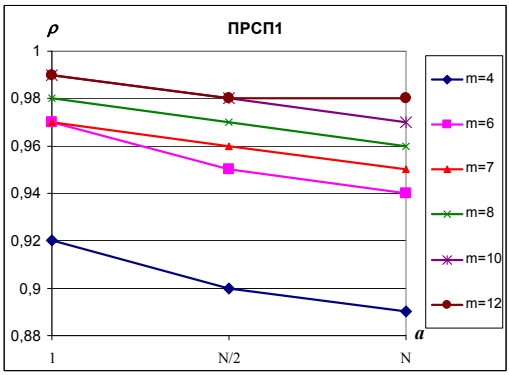

Рис. 6. Отношение w/W пропускных способностей сетей с топологией квазиполного и полного графов в режиме ПРСП1.

В [1, 7] и в разделе 7 рассматриваются сети с m=7 и с m=38. На рис. 7 представлены результаты моделирования в режиме ПРСПΔ для m=38. Они показывают, что варианты с Δ=0 и с Δ >> 1 (две нижние кривые) практически совпадают.

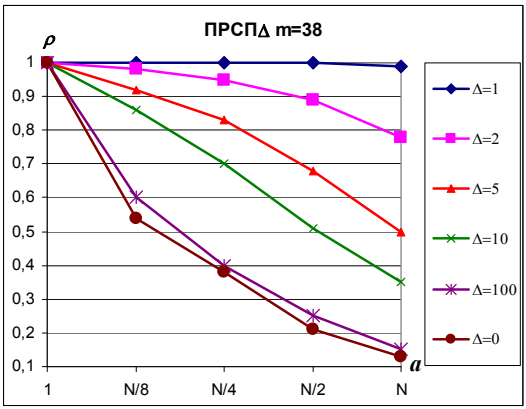

Рис. 7. Отношение w/W пропускных способностей сетей с топологией квазиполного и полного графов в режиме ПРСПΔ для m=38.

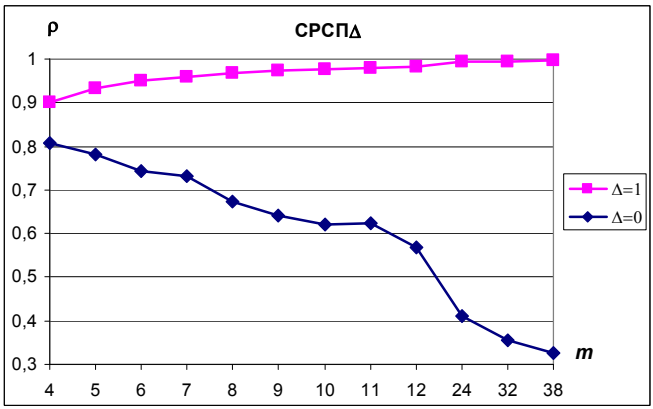

Рис. 8. Отношение w/W пропускных способностей сетей с топологией квазиполного и полного графов в режиме СРСПΔ.

На рис. 8 представлены результаты моделирования в режиме СРСПΔ для разных m. В режиме с Δ=1 для всех проверенных m ($4 \le m \le 38$) имеет место соотношение $\rho > 0{,}91$, т.е. пропускные способности сетей с топологией полного и квазиполного графов практически совпадают. В то же время, в режиме без перемешивания адресов приемников (Δ=0) имеет место вырождение пропускной способности, как и в режиме ПРСП0.

Для задержек передачи картина выглядит не столь оптимистично даже в режиме СРСП1. Сначала дадим полную картину задержек, задаваемую нашей моделью, в затем выделим область эквивалентности сетей с топологией полного и квазиполного графов.

В каждом такте измеряется задержка передачи δ (в тактах) для каждого источника. Случайная величина δ усредняется в каждом такте по источникам, осуществившим передачу пакета в данном такте, а затем – по всем тактам. В результате формируется средняя задержка передачи τ (в тактах), которая и представлена на последующих графиках.

Сначала рассмотрим режим с постоянным распределением адресов приемников для всех источников – ПРСПΔ. На рис. 9 и 10 представлены графики зависимости $\tau(a)$ задержки передачи от степени распределения в случае малых m. Эти графики соответствуют графикам зависимости на рис. 5 и 6. Здесь надо иметь ввиду, что $\tau(a)=1$ для сетей с топологией полного графа. Разрывы на графиках возникли потому, что для представленных m нет значений a с выбранной дискретностью представления. Отметим три вывода из представленных графиков. Задержки в режиме ПРСП1 растут с ростом a, при том что пропускная способность остается высокой при всех значениях a: $\rho(a) > 0{,}89$ (рис. 6). Задержки в режиме ПРСП1 существенно меньше, чем в режиме ПРСП0. В режиме ПРСП1 $\tau(a) < 1{,}5$, а в режиме ПРСП0 $\tau(a) < 2$ при $a < 10$.

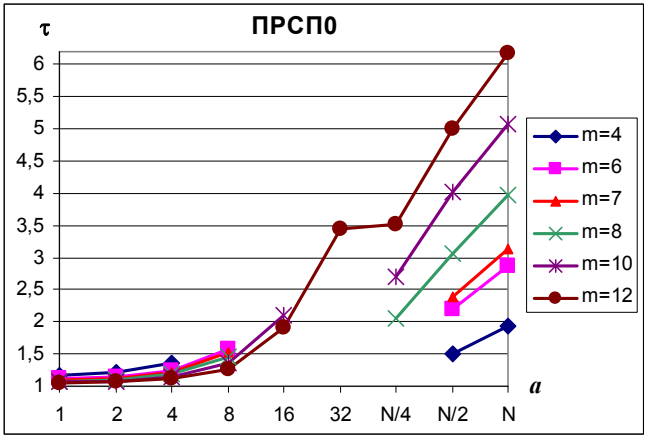

Рис. 9. Задержки сетей с топологией квазиполного графа в режиме ПРСП0.

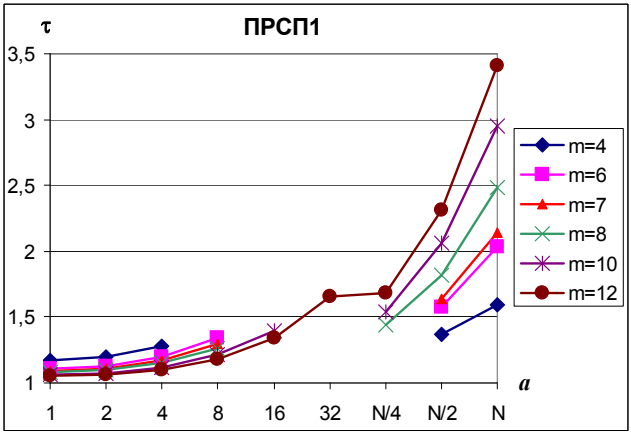

Рис. 10. Задержки сетей с топологией квазиполного графа в режиме ПРСП1.

На рис. 11 и 12 представлены графики зависимости $\tau(a)$ для m=38. Существенно то, что в режиме ПРСП1 $\tau(a) < 1{,}3$ при $a <$

N/32=43, т.е. диапазон появления малых задержек оказался существенно шире, чем при малых *m*.

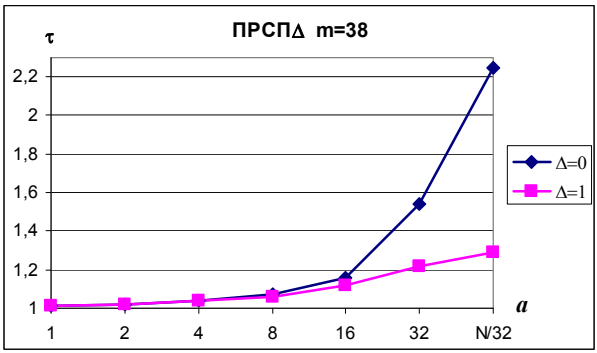

Рис. 11. Задержки сети с топологией квазиполного графа с m=38 в режиме ПРСПΔ при малых a.

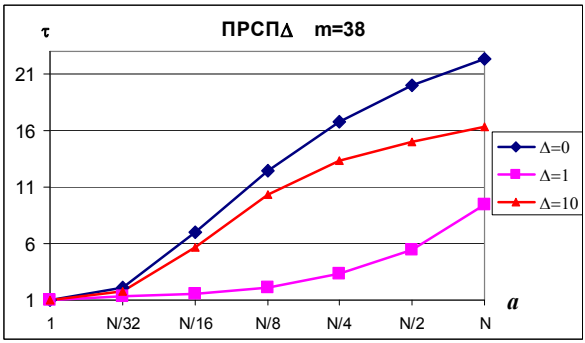

Рис. 12. Задержки сети с топологией квазиполного графа с m=38 в режиме ПРСПΔ при больших a.

В целом по режиму ПРСПΔ можно сделать вывод, что имеется диапазон распределений с небольшим показателем *a*, в котором сеть с топологией квазиполного графа имеет задержки только на 10÷30% больше, чем сеть с топологией полного графа.

Теперь рассмотрим задержки в режимах со случайным распределением адресов приемников для всех источников –

СРСПΔ. На рис. 13 представлены графики зависимости τ(*m*) задержки передачи при *a*=*N* для всех проверенных *m*.

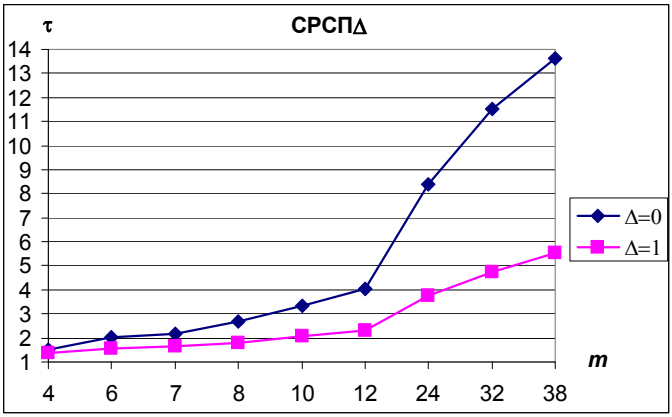

Рис. 13. Задержки сети с топологией квазиполного графа в режиме СРСПΔ для всех m.

Видно, что в режиме СРСП1 τ > 1,5, при том что пропускная способность сети остается максимально высокой: ρ > 0,97 (рис. 8). В режиме СРСП0 задержки оказываются многократно больше. В попытке снизить задержки было введено ограничение на максимальное значение степени распределения *a*. На рис. 14 представлены графики зависимости τ(*a*) для *m*=38.

Здесь в режиме СРСП1 τ(*a*) < 1,3 при *a* < *N*/16=87 и τ(*a*) < 1,6 при *a* < *N*/8=175. Это означает, что по сравнению с режимом ПРСП1 диапазон малых задержек оказался в 2÷4 раза шире. В режиме СРСП1 для *m* ≤ 12 уже τ(*a*) < 1,2 при *a* < *N*/8 и τ(*a*) < 1,4 при *a* < *N*/4.

Таким образом, сети с топологией квазиполного графа могут иметь задержки только на 20÷30% большие, чем сети с топологией полного графа при ограничении максимального значения степени показателя значениями *a* < *N*/8.

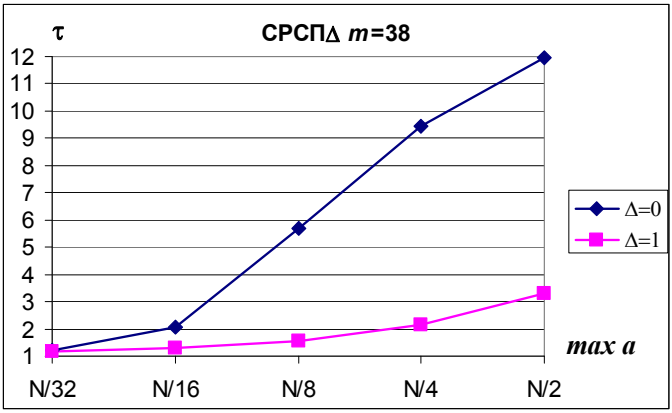

Рис. 14. Задержки сети с топологией квазиполного графа с m=38 в *режиме СРСПΔ при ограничении a.*

В заключение раздела специально отметим, что при равномерном распределении (a=1) адресов приемников пропускные способности и задержки передачи сетей с топологией полного и квазиполного графов практически совпадают (рис. 15).

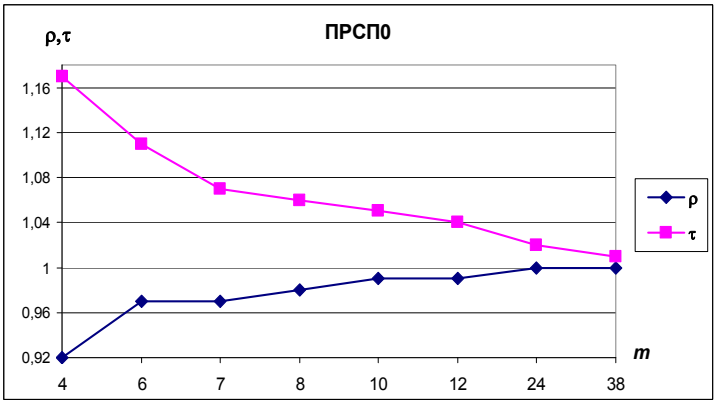

Рис. 15. Характеристики сети с топологией квазиполного графа при a=1 для разных m в наихудшем режиме.

5. Расширенный обобщенный гиперкуб

Как упоминалось в начале статьи (рис. 1), обобщенный гиперкуб имеет по строкам и столбцам топологию связей полного графа. Расширение обобщенного гиперкуба можно осуществить за счет замены топологии связей строки (столбца) каждого измерения – с топологии полного графа на топологию квазиполного графа.

Пусть имеется ОГК(*V, d, s*). Добавим к каждому узлу *d* коммутаторов $m \times m$ ($m=s-1$). Для объединения узлов в каждой строке (столбце) будем использовать ПС(N, m, σ), построенную за счет использования одного коммутатора $m \times m$ при каждом узле. ПС(N, m, σ) позволяет объединить в строке N узлов, где $N=m(m-1)/\sigma+1$. Увеличим число узлов в строке каждого измерения до этой величины и объединим их посредством ПС(*N, m*, σ). Тем самым удаётся "расширить" ОГК с *s*-ичности до *N*-ичности.

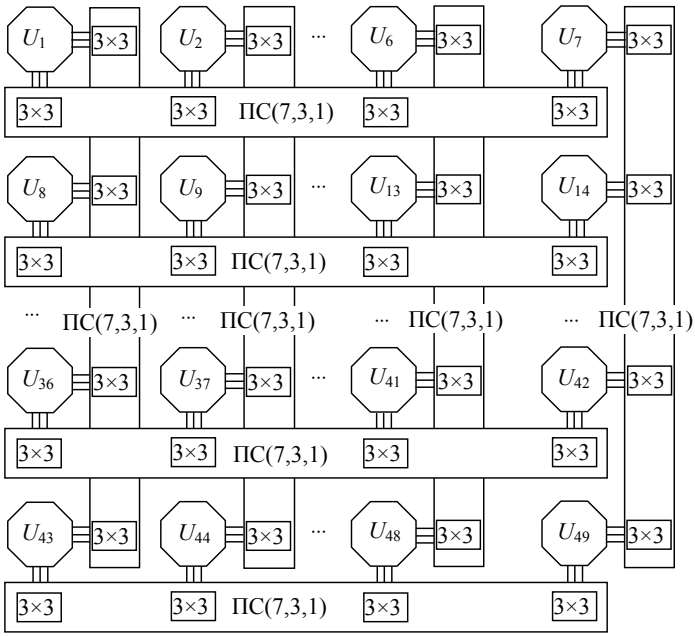

Рис. 16.Структура связей РОГК(49, 2, 7, 3, 1).

В результате получим расширенный обобщенный гиперкуб *с* $R=N^d$ узлами, в котором узлы любой строки (столбца) каждого измерения связаны σ прямыми каналами. Такой гиперкуб будем назвать расширенным обобщенным *d*-мерным *N*-ичным гиперкубом и обозначать РОГК(*R, d, N, m,* σ). На рис. 16 для примера показан РОГК(49, 2, 7, 3, 1).

В описанном построении ПС(*N, m*, σ) может быть заменена на 1-расширенную ПС(*N*, m,* σ|σ+1) (см. раздел 3), и тогда расширенный гиперкуб обозначается как РОГК(*R, d, N*, m,* σ|σ+1).

Напомним, что 1-расширенная простейшая сеть ПС(*N*, m,* σ|σ+1) порождается при уменьшении числа узлов сети ПС(*N, m*, σ) до значения $N^*=N-\delta$, где $\delta < m$. При этом появляются узлы одной доли, связанные σ+1 путями длины 2, а остальные – σ путями длины 2. Сеть ПС(*N*, m,* σ|σ+1) удается построить для любых *m* и σ.

Эти дополнительные пути потребуются в следующем разделе 4 для формирования отказоустойчивой конфигурации РОГК по связным узлам. Далее, всюду где упоминается РОГК(*R, d, N*, m,* σ|σ+1) имеется ввиду он или РОГК(*R, d, N, m,* σ), если последний существует для заданных *m* и σ.

Сравним некоторые характеристики трехмерных обобщенных гиперкубов ОГК(*V,* 3*, s*) и РОГК(*R,* 3*, N*, m,* σ|σ+1). Сначала оценим фактор *R/V* увеличения числа узлов при одинаковых параметрах узла, который приводится в табл. 2 при малых *m* и σ.

Таблица 2. Фактор R/V при m=s–1

σ \ *m*	4	6	8	10	12
1	17,6	96	254	566	1071
2	2,7	9,8	27,0	56	114
3	1,0	3,9	9,4	18,3	36,2
4	–	1,5	4,6	8,0	16,4

Табл. 2 показывает, что за счёт увеличения “-ичности” РОГК(*R,* 3*, N*, m,* σ|σ+1) по сравнению с ОГК(*V,* 3*, s*) может

иметь во много раз большее число узлов и/или в σ раз большую пропускную способность каждого измерения.

Таблица 3. Значение s при R=V (s=N)*

σ \ *m*	4	6	8	10	12
1	13	32	57	91	133
2	7	15	27	42	63
3	5	11	19	29	43
4	–	8	15	22	33

В табл. 3 показана степень узла *s* по каждому измерению в ОГК(*V, d, s*) при *V=R* из РОГК(*R, d, N*, m,* σ|σ+1). Табл. 3 показывает, что РОГК(*R, d, N*, m,* σ|σ+1) по сравнению с ОГК(*V, d, s*) может иметь во много раз меньшее число портов в каждом узле и/или в несколько раз большую пропускную способность каждого измерения.

6. Отказоустойчивость расширенных обобщенных гиперкубов

Полная однородность каждой строки (столбца) в РОГК(*R, d, N*, m,* σ|σ+1) по узлам обеспечивает его κ-отказоустойчивость по процессорным узлам [6] при любом $1 \le \kappa \le \lfloor N/2 \rfloor$. Наличие в каждой строке σ независимых путей между любой парой процессорных узлов абонентов обеспечивает (σ–1)-отказоустойчивость по каналам и коммутаторам $m \times m$ каждой ПС(*N**, *m*, σ|σ+1) [6, 7]. Однако 1-скачковая связь любых строк (столбцов) разных измерений только через один связной узел не обеспечивает отказоустойчивости РОГК(*R, d, N*, m,* σ|σ+1) по связным узлам.

Для обеспечения 1-отказоустойчивости по узлам можно использовать большую избыточность РОГК(*R, d, N*, m,* σ|σ+1) по числу узлов следующим образом. Разобьем узлы каждой строки на пары четный-нечетный. Каждая такая пара должна входит в состав строки каждого измерения. Это приведет к уменьшению вдвое числа строк всех измерений кроме первого, но обеспечит два варианта перехода между строками разных

измерений – через четный и нечетный узлы. Такой 1-отказоустойчивый расширенный обобщенный гиперкуб будем означать РОГК(R^*, d, N^*, m, σ|σ+1, 1). Он содержит $(N^*/2)^d$ пар узлов, т.е. $R^*=2(N^*/2)^d=N^*(N^*/2)^{d-1}$ узлов. В принятых ранее обозначениях РОГК(R, d, N^*, m, σ|σ+1) обозначается как РОГК(R, d, N^*, m, σ|σ+1, 0).

На рис. 17 показан 2-мерный РОГК(18, 2, 6, 3, 1, 1). Для построения ПС(6, 3, 1) использован 1-расширенный гиперкуб, в котором пары абонентов (i, j) с номерами $j=(i\pm 3)\bmod 6$ связаны двумя путями, а остальные – одним путем. На рис. 18 этот гиперкуб расширен до трех измерений – РОГК(54, 3, 6, 3, 1, 1).

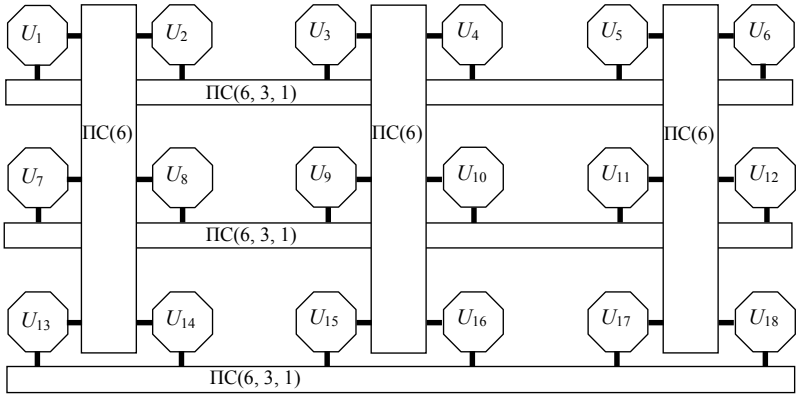

Рис. 17. Топология связей в 2-мерном РОГК(18, 2, 6, 3, 1, 1).

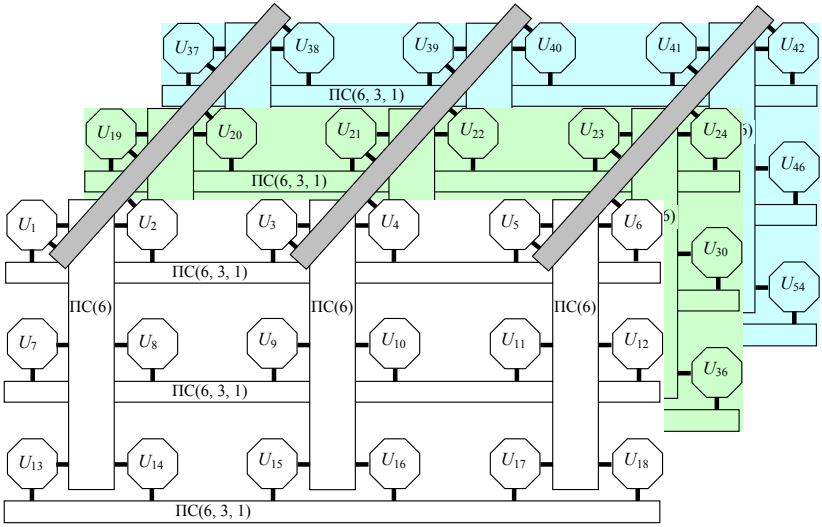

Рис. 18. Топология связей в 3-мерном РОГК(54, 3, 6, 3, 1, 1).

На этих рисунках толстыми линиями показаны наборы из 3 ребер. Если заменить их на наборы из 4 ребер и использовать ПС(6, 4, 2) для объединения узлов в строках, то эти же рисунки зададут РОГК(18, 2, 6, 3, 2, 1) и РОГК(54, 3, 6, 3, 2, 1). соответственно. В ПС(6, 4, 2) пары абонентов (i, j) с номерами $j = (i \pm 1) \bmod 6$ связаны тремя путями, а остальные – двумя путями.

Аналогичным образом можно строить и μ-отказоустойчивые расширенные обобщенные гиперкубы РОГК(R^*, d, N^*, m, $\sigma|\sigma+1$, μ). Для этого строки должны разбиваться на группы по $(\mu+1)$ узлов, а гиперкуб будет содержать $R^*=N^*(N^*/(\mu+1))^{d-1}$ узлов. В этом случае отказоустойчивость можно разменивать на пропускную способность между измерениями.

7. Системная сеть на основе расширенного обобщенного гиперкуба

В данном разделе рассматривается возможность расширения системной сети суперкомпьютера, разработанного

IBM для проекта *Blue Waters* [9]. Теперь его принято именовать как сервер *PERCS* (*Productive, Easy-to-use, Reliable Computer System*) или система *P*775 [12]. Одна попытка такого расширения была предпринята авторами в [1]. Для возможностей последующего сравнения изложим кратко ее суть.

Системная сеть в [9] имеет топологию двухуровневого полного графа (рис. 19).

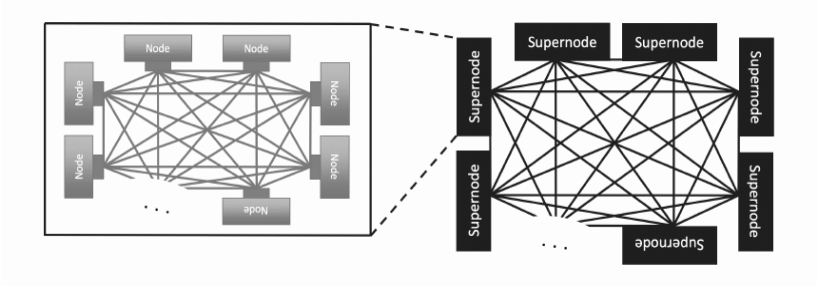

Рис. 19. Узлы; суперузлы и связи между ними.

Каждый узел связи (*node*) имеет межузловые каналы трех видов: 7 высокоскоростных каналов K_1 с пропускной способностью V_1; 24 канала низкоскоростных K_2 с пропускной способностью $V_2 = V_1/5$ и 16 среднескоростных каналов K_3 с пропускной способностью $V_3 = 2V_2$. Каналы K_1 выполнены медным кабелем, а каналы K_2 и K_3 – оптическим кабелем. На рис. 19 приведена структура связей этой системной сети.

32 узла связи образуют суперузел (*supernode*), 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. Суперузлы связаны каналами K_3 также по схеме полного графа.

Каждый суперузел имеет 512 каналов K_3. В максимальной конфигурации суперкомпьютера каждый такой канал используется для связи с другим суперузлом по схеме полного графа. В этом случае он содержит 513 суперузлов. Передача пакета между любыми двумя узлами занимает не более трёх смен каналов с промежуточной буферизацией пакетов (скачков).

В [1] был рассмотрен подход, основанный на замене топологии связей полного графа в узлах и суперузлах на топологию квазиполного графа. В нем оставались неизменными следующие параметры сети: число портов в узле связи и максимальное число скачков между любыми узлами сети.

Таблица 4. Параметры системной сети суперузла

Параметры	Новая сеть	Исходная сеть
Коммутатор каналов K_1	7×7	–
Число узлов	39	32
Фактор межузловой пропускной способности f_1	$2 < f_1 < 5$	$f_1=1$
Каналы K_2	0	24
Фактор энергопотребления узла e_1	$1 \approx e_1 < 2$	$e=1$

Для модификация сети внутри суперузла (левая часть рис. 11) к каждому узлу добавлялся коммутатор 7×7 каналов K_1 и узлы связывались сетью с топологией 1-расширенного квазиполного графа ПС(39, 7, 1|2). В результате образовывался суперузел из 39 узлов, связанных прямыми высокоскоростными каналами K_1. Каналы K_2 не использовались. Параметры полученной сети представлены в табл. 4. В ней фактор f_1 оценен по результатам моделирования, представленным на рис. 8.

Для модификации сети между суперузлами (правая часть рис. 11) к каждому суперузлу добавлялось 16 коммутаторов 38×38 каналов K_3 и суперузлы связывались 16 сетями ПС(1407, 38, 1). Параметры полученной сети представлены в табл. 5. В ней фактор f_3 оценен по результатам моделирования (рис. 8).

В результате предложенной модификации получена расширенная системная сеть с б***о***льшим числом узлов, с б***о***льшей пропускной способностью и с многократной отказоустойчивостью 3-скачковых путей. В ней остались

неиспользованными 24 канала K_2, которые оказалось невозможно использовать в описанной топологии связей.

Таблица 5. Параметры системной сети между суперузлами

Параметры	Новая сеть	Исходная сеть
Коммутаторы каналов K_3	38×38	–
Число коммутаторов в суперузле	16	0
Число суперузлов	1407	513
Число узлов всей сети	54873	16416
Число путей между суперузлами	16	1
Отказоустойчивость 3-скачковых путей	есть	нет
Фактор пропускной способности между суперузлами f_3	$2 < f_3 < 16$	$f_3=1$
Фактор энергопотребления суперузла e_3	$1 < e_3 < 2$	$e_3=1$

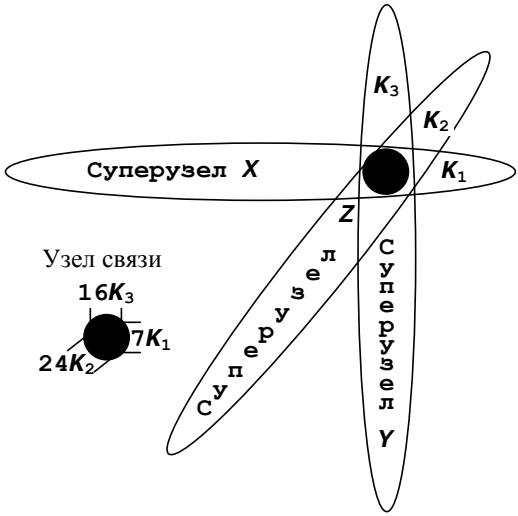

Рис. 20. Суперузлы разных измерений.

Однако их можно использовать в топологии трехмерного расширенного обобщенного гиперкуба – точнее 3-мерного параллелепипеда, в каждом измерении которого узлы имеют разные степени узлов. В измерении X она равна 7, в измерении Y – 16 и в измерении Z – 24.

Пусть строки в измерении X составляют рассмотренные выше модифицированные суперузлы с коммутаторами 7×7 каналов K_1 при каждом узле. Тогда строки в измерении Y могут составлять «новые» суперузлы с коммутаторами каналов K_2 при каждом узле, а строки в измерении Z могут составлять «новые» суперузлы с коммутаторами каналов K_3 при каждом узле. Общая топология такой системной сети представлена на рис. 20.

В рассматриваемой топологии возможны различные варианты построения суперузлов каждого измерения. Первый вариант – это использование малых коммутаторов 8×8 в суперузлах измерений Y и Z. Для построения суперузлов измерения Y к каждому узлу необходимо добавить 2 таких коммутатора, а измерения Z – 3 таких коммутатора. В данном случае в суперузлах каждого измерения удобно использовать топологию квазиполного орграфа [1 – 3].

Таблица 6. Параметры 1-отказоустойчивой по связным узлам системной сети с топологией 3-мерного параллелепипеда.

Параметры	Измерение X: каналы K_1	Измерение Y: каналы K_3	Измерение Z: каналы K_2	Общее число узлов
Число и вид коммутаторов при узле	1 – 7×7	2 – 8×8	3 – 8×8	$R^*=$ $=N_X N_Y N_Z/4=$ $=50176$
Число путей в суперузле	1	2	3	
Число узлов в суперузле	$N_X = 48$	$N_Y = 32$	$N_Z = 32$	

В результате узлы в суперузлах измерения X связываются ПС(49, 7, 1), в суперузлах измерения Y – двумя независимыми ПС(64, 8, 1) и в суперузлах измерения Z – тремя независимыми ПС(64, 8, 1). Это позволяет построить 1-отказоустойчивую по узлам системную сеть, параметры которой приведены в табл. 6.

В другом варианте можно иметь более высокую пропускную способность системной сети в суперузлах за счет использования топологии квазиполного графа с $\sigma > 1$. Для этого в узлах для создания суперузлов измерений Y и Z потребуется использовать коммутаторы 16×16 и 24×24 соответственно. Конкретно: узлы суперузлов измерения X свяжем 1-расширенной ИС(20, 7, 2|3), узлы суперузлов измерения Y свяжем 1-расширенной ПС(60, 16, 4|5) и узлы суперузлов измерения Y свяжем 1-расширенной ПС(70, 24, 8|9). Количество каналов в суперузлах выбиралось из соображения приблизительного равенства суммарной пропускной способности между узлами суперузлов. В результате можно построить 1-отказоустойчивую по узлам и по каналам системную сеть, параметры которой приведены в табл. 7.

Таблица 7. Параметры 1-отказоустойчивой по связным узлам системной сети с топологией 3-мерного параллелепипеда с повышенной пропускной способностью.

Параметры	Измерение X: каналы K_1	Измерение Y: каналы K_3	Измерение Z: каналы K_2	Общее число узлов
Число и вид коммутаторов при узле	1 – 7×7	1 – 16×16	1 – 24×24	$R^*=$ $=N_XN_YN_Z/4=$ $=21000$
Число путей в суперузле	2	4	8	
Число узлов в суперузле	$N_X = 20$	$N_Y = 60$	$N_Z = 70$	

Если не повышать пропускной способности сети в суперузлах измерения X, то получим вариант системной сети, представленной в табл. 8. Причем этот вариант является ориентировочным, т.к. 1-расширенные квазиполные графы для ПС(80, 16, 3|4) и ПС(94, 24, 6|7) еще не построены. Здесь значение числа узлов в суперузлах после построения может измениться на несколько единиц.

Таблица 8. Параметры 1-отказоустойчивой по связным узлам системной сети с топологией 3-мерного параллелепипеда с повышенной пропускной способностью.

Параметры	Измерение X: каналы K_1	Измерение Y: каналы K_3	Измерение Z: каналы K_2	Общее число узлов
Число и вид коммутаторов при узле	1 – 7×7	1 – 16×16	1 – 24×24	$R^*=$ $=N_X N_Y N_Z/4=$ $=73320$
Число путей в суперузле	1	3	6	
Число узлов в суперузле	$N_X = 39$	$N_Y = 80$	$N_Z = 94$	

8. Заключение

В работе предложена новая топология системной сети в виде расширенного обобщенного гиперкуба. Эта топология разрабатывалась с ориентацией на суперкомпьютер фирмы *IBM*, разрабатывавшийся в рамках проекта *Blue Waters* [9]. Теперь этот суперкомпьютер именуют как сервер *PERCS* или система *P*775 [12]. Эта система первая пробила «стену памяти», т.е. показала выдающиеся характеристики в условиях плохой пространственно-временной локализации данных в памяти. В частности, на тестах “Graph 500” [11] она в сильно усеченном варианте имела лучшую удельную производительность. Эти характеристики достигнуты, в частности, за счет высокой параллельности и малой глубины системной сети. В данной работе указанные свойства системной сети сохранены, но повышены ее пропускная способность, масштабируемость и отказоустойчивость, за счет использования системной сети в виде расширенного обобщенного гиперкуба в форме 3-мерного параллелепипеда. В нем строки (столбцы) разных измерений имеют не только разные простейшие сети, объединяющие их узлы, но и разные скорости передачи по каналам этих сетей.

Литература

1. КАРАВАЙ М.Ф., ПОДЛАЗОВ В.С. *Топологические резервы суперкомпьютерного интерконнекта* // Управление большими системами. Вып. 40. – М:. ИПУ РАН. – 2012. – С. 395–423. – URL: http://ubs.mtas.ru/upload/library/UBS4114.pdf.
2. 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.
3. КАРАВАЙ М.Ф., ПОДЛАЗОВ В.С**.** *Распределенный полный коммутатор как «идеальная» системная сеть для многопроцессорных систем* // Управление большими системами. Вып. 34. – М:. ИПУ РАН. – 2011. – С. 92–116. – URL: http://ubs.mtas.ru/upload/library/UBS3405.pdf.
4. КАРАВАЙ М.Ф., ПОДЛАЗОВ В.С., СОКОЛОВ В.В. *Способ построения неблокируемого самомаршрутизируемого расширенного коммутатора* // Патент на изобретение № 2435295 РФ от 06.09.2009. Зарегистрирован 03.08.2011.
5. КАРАВАЙ М.Ф., ПОДЛАЗОВ В.С**.** *Расширенные блок-схемы для идеальных системных сетей* // Проблемы управления. – № 4. – 2012. С. 45–51.
6. КАРАВАЙ М.Ф., ПОДЛАЗОВ В.С. *Сетецентрический подход к обеспечению отказоустойчивости многопроцессорных систем реального времени* // Четвертая всероссийская мультиконференция по проблемам управления (МКПУ-2011). – Окт. 2011. – Дивноморское. т. 1. – С. 305–308.
7. ПОДЛАЗОВ В.С., КАРАВАЙ М.Ф. *Системные сети с прямыми каналами для многопроцессорных вычислительных систем – идеальные системные сети* // Palmarium Academic Publishing. – 2012. – 168 С. – URL: http://www.ipu.ru/sites/default/files/publications/18125/3747-18125.pdf – интернет-магазин www.ljubljuknigi.ru.

8. ALVERSON R., ROWETH D. AND KAPLAN L. *The Gemini System Interconnect* // 18th IEEE Symposium on High Performance Interconnects. – 2009. – P. 83–87.
9. ARIMILI B., ARIMILI R., CHUNG V., ET AL. *The PERCS High-Performance Interconnect* // 18th IEEE Symposium on High Performance Interconnects. – 2009. – P. 75–82.
10. BHUYAN L. N. AND AGRAWAL D. P. *Generalized Hypercube and Hyperbus Structures for a Computer Network* // IEEE Transaction on Computers. – Vol C-33. – No 4. April 1984. – P. 323–333.
11. *Graph 500 List* // June. 2012. – URL: http://www.graph500.org/results_june_2012 .
12. HRUSKA J. *After Years of Work IBM, NCSA Cancel «BlueWaters» Supercomputer* // Aug. 2011. – URL: http://hothardware.com/News/After-Years-of-Work-IBM-NCSA-Cancel-Blue-Waters-Supercomputer/ .
13. ZIAVRAS S. G. AND KRISHNAMURTHY S. *Evaluating the communications capabilities of the generalized hypercube interconnection network* // Concurrency: Practice and Experience. – Vol 11. – No 6. – 1999. – P. 281–300.