ISSN 2079-3316 **ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ** **т. 13**, № 4(55), с. 47–76
*научная статья* ПРОГРАММНОЕ И АППАРАТНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ СУПЕРЭВМ

УДК 004.724.2 + 004.272.43
10.25209/2079-3316-2022-13-4-47-76

# Самомаршрутизируемая неблокируемая системная сеть с прямыми каналами: *сложность и быстродействие*

Виктор Сергеевич **Подлазов**✉
Институт проблем управления имени В. А. Трапезникова РАН, Москва, Россия
✉*podlazov@ipu.ru* *(подробнее об авторе на с. 72)*

Аннотация. Разработана неблокируемая самомаршрутизируемая сеть с прямыми каналами, в которой конфликты пакетов разрешаются на входе в сеть посредством процедуры борьбы источников за вход в первый каскад сети, обеспечивая пакетную дуальность. Заблокированные при борьбе пакеты повторно передаются источниками с минимальными задержками. Дуальность подразумевает совместное использование шинного (с разведением во времени) и мультиплексного (с разведением по каналам) способов разрешения конфликтов пакетов. Внутри сети возникновение конфликтов предупреждается посредством ее внутреннего распараллеливания, т.е. созданием заведомо бесконфликтных путей. Сеть разработана в 2-, 4-, и 8-каскадном вариантах с масштабированием числа каналов от нескольких сот до многих миллионов при неизменном быстродействии сети. В сети возможно обеспечение 1-, или 2-канальной отказоустойчивости при сохранении ее быстродействия. Накладными затратами на достижение указанных свойств является повышенная сложность сети, которая сопоставима со сложностью теоретического неблокируемого коммутатора Клоза. Хотя его структура известна, но практическая реализация отсутствует вследствие неизвестности процедуры параллельной самомаршрутизации в нем. Практическая ориентация предложенных сетей – это системные сети с передачей маршрутной информации в заголовках пакетов с однократным использованием в каждом каскаде управляющей маршрутной информации для базового полного коммутатора. Предложенные сети выполнены в расширенном схемном базисе, состоящем из полных коммутаторов и отдельных мультиплексоров и демультиплексоров. В работе представлены характеристики построенных сетей при указанном способе представления маршрутной информации.
(*see abstract in English on p. 73*)

Ключевые слова и фразы: полный коммутатор, дуальный коммутатор, пакетная дуальность, мультиплексоры и демультиплексоры, многокаскадный коммутатор, бесконфликтная самомаршрутизация, неблокируемый коммутатор, статическая самомаршрутизация, квазиполный орграф, квазиполный граф, инвариантное расширение сетей, коммутационные свойства, прямые каналы, масштабируемость и быстродействие

Для цитирования: Подлазов В.С. *Самомаршрутизируемая неблокируемая системная сеть с прямыми каналами:* сложность и быстродействие // Программные системы: теория и приложения. 2022. **Т. 13**. № 4(55). С. 47–76. http://psta.psiras.ru/read/psta2022_4_47-76.pdf

## Введение

В работе решается классическая коммутационная задача построения неблокируемого самомаршрутизируемого распределенного коммутатора с децентрализованным управлением и прямыми каналами, который образует системную сеть пакетной коммутации для многопроцессорных вычислительных систем. Эта задача до сих пор не имеет полного решения оптимального по быстродействию и схемной сложности. В работе рассматривается новое решение этой задачи с некоторой оптимизацией по указанным параметрам. Прямой канал между источником и приемником, используемый для доставки пакетов, содержит только комбинационные элементы и не содержит никаких элементов памяти, используемых обычно для буферизации конфликтных пакетов, и поэтому обеспечивает наименьшее время доставки пакетов по сети. Комбинационные элементы в канале образуют $p$-канальные полные коммутаторы $p \times p$, демультиплексоры $1 \times p$ и мультиплексоры $p \times 1$. Они осуществляют маршрутизацию пакетов на основе управляющей маршрутной информации, передаваемой в пакете вместе с данными, и делают это без тактовых задержек – на лету. Это всегда можно сделать в комбинационной схеме без использования внутренней памяти.

В рассматриваемой сети используется статическая маршрутизация пакетов, при которой каждый источник порождает маршрутную информацию, задающую путь через сеть к приемнику независимо от других источников. Эта маршрутная информация порождается для перестановочного трафика, при котором каждый источник передает пакет только одному приемнику, который мы в дальнейшем называем перестановкой пакетов. Маршрутная информация в пакете состоит из нескольких частей, которые в совокупности задают весь путь и которые используются для управления коммутаторами и демультиплексорами при прокладке пути каждого пакета независимо от других пакетов, что и реализует самомаршрутизацию пакетов. Подчеркнем, что под самомаршрутизацией в данной работе понимается «маршрутизация от источника», при которой отправитель в заголовке пакета задает его путь через сеть, а не адрес получателя, по котором выбор пути осуществляет сеть. Перестановочный трафик возникает в разных подзадачах, например при совместном выполнении БПФ произвольными абонентами при их барьерной синхронизации, сортировках, матричных операциях. Кроме того, перестановочный трафик часто имеет место в компьютерах типа *SIMD* или *SPMD* (отечественная

ПС2000). Решение коммутационной задача предполагает построение неблокируемой сети, т.е. сети с бесконфликтными каналами при перестановочном трафике. В рассматриваемой сети бесконфликтная перестановка пакетов осуществляется с некоторыми задержками, которые меньше чем в базовых работах [1, 2].

В работах [1, 2] была развита методика построения неблокируемых отказоустойчивых дуальных фотонных системных сетей широкой масштабируемости. Изначально эти сети были ориентированы на применение в фотонных компьютерах [3], в которых требуются неблокируемые сети с прямыми каналами. Термин «фотонный» подчеркивает возможность построения сети с использованием чисто фотонных однокристальных коммутаторов и прямых каналов между источниками и приемниками без использования в каналах оптической регистровой памяти для промежуточной буферизации пакетов. Конечно, аналогичное построение электронных сетей с прямыми каналами также возможно при цифровом кодировании управляющей маршрутной информации. Такие сети являются системными сетями суперкомпьютеров, которые имеют наибольшее быстродействие.

В основе методики [1, 2] лежит применение в первом каскаде дуального коммутатора, неблокируемого на произвольном трафике. В дуальном коммутаторе сочетаются два способа разрешения конфликтов: шинный (с разведением конфликтующих сигналов во времени) и мультиплексный (с разведением сигналов по разным линиям). Это разведение применяется как к информационным сигналам, такт и к управляющим сигналам маршрутной информации, т.е. во всем частотном диапазоне. В $p$-канальном дуальном коммутаторе неблокируемость достигается посредством увеличения периода $T$ сигналов до $p$ тактов, где такт задает длительность $\tau$ информационного сигнала. На рисунках 1 и 2 представлены две модификации $p$-канального дуального коммутатора с $p = 4-$ первичного (ПК$p$) и расширенного (РК$p$).

В электронном дуальном коммутаторе вместо линий задержки (ЛЗ$\delta$) используются элементы задержки (ЭЗ$\delta$) равной длительности.

Эти дуальные коммутаторы состоят из пар мультиплексор-демультиплексор. Мультиплексор $p$ в каждом входном канале имеет петлю обратной связи через линию задержки ЛЗ$\delta$ длительностью в $\delta = 1$ такт. При конфликте сигналов в любом $Mp$ один сигнал пропускается на выход, а остальные возвращаются на входы через свои ЛЗ$\delta$.

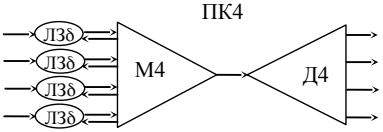



Рисунок 1. 4-канальный первичный дуальный коммутатор ПК4 с маршрутизацией по 4 каналам

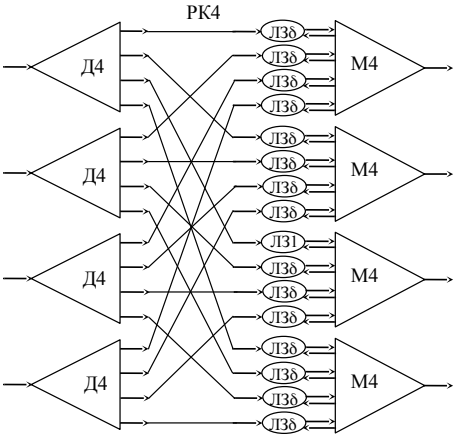
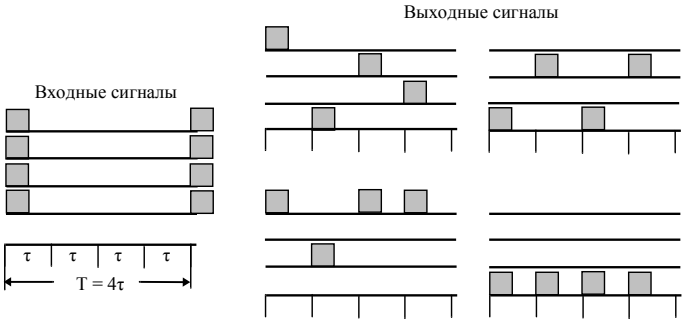


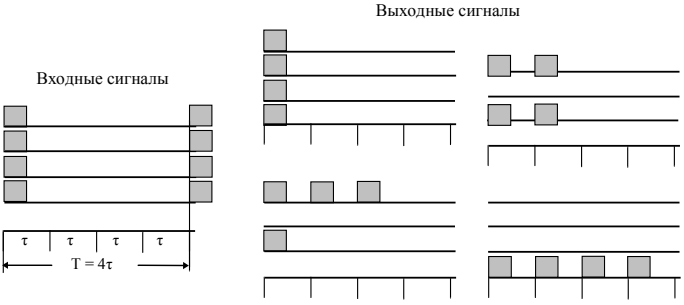

Рисунок 2. 4-канальный расширенный дуальный коммутатор РК4

Маршрутизация сигналов осуществляется демультиплексорами с использованием $\log_2 p$ управляющих сигналов. В фотонном коммутаторе управляющие сигналы передаются в одном такте с информационным сигналом на разных несущих частотах. В электронном коммутаторе они передаются по разным линиям. В первичном коммутаторе ПК$p$ разрешение конфликтов осуществляется только мультиплексором шинным способом, а демультиплексор осуществляет только маршрутизацию сигналов. Расширенный коммутатор РК$p$ выполняет маршрутизацию сигналов параллельно с разведением потенциально конфликтующих сигналов по разным каналам. Сетевые пакеты состоят из $p$-тактных разрядов. Каждый такой разряд имеет вид последовательности $p$ отнотактных информационных и управляющих сигналов с одинаковыми множествами последних во всех тактах, но возможно разных в одинаковых тактах в разных каналах после коммутаторов. Пусть

сигналы каждого разряда $p$ разных пакетов поступают на разные входы дуального коммутатора синхронно в первом такте (рисунок 3). При ис-




(*а*) для ПК$p$




(*б*) для РК$p$

Рисунок 3. Распределение сигналов при $p = 4$

пользовании в мультиплексорах ЛЗ1 конфликтующие информационные сигналы разводятся по разным тактам вместе с сопровождающими их управляющими сигналами. Примеры возникающих распределений сигналов по периоду разряда на выходных линиях дуального коммутатора приводятся на рисунке 3*а* для ПК4, а на рисунке 3*б* для РК4. Для ПК$p$ имеет место разведение сигналов по всем $p$ тактам разряда, а для РК$p$ – только по начальным тактам разряда. При этом в разных тактах для разных каналов используются разные наборы управляющих сигналов, которые при дальнейшей маршрутизации уже не участвуют.

В работах [**1**, **2**] представлена методика построения многокаскадных неблокируемых сетей на основе расширенного дуального коммутатора

РК$p$, в которых используются десятки частот для представления управляющей информации. Из них каждая порция в $\log_2 p$ частот используется только однажды в своем каскаде, а передаются все они через все каскады. Таким образом, имеется большая избыточность в требуемой полосе пропускания управляющих сигналов. В работе решается задача сокращения требуемой полосы до минимума – до полосы, требуемой для передачи информационных сигналов и однократной передачи управляющих сигналов. Естественно делать это посредством передачи управляющей информации в заголовке пакета. Однако, при этом возникает принципиальная трудность привязки этой информации к номерам тактов, которые заранее неизвестны (см. рисунок 3). В работе эта трудность преодолевается посредством перехода от поразрядного представления управляющей информации к попакетному представлению, формируя пакетную дуальность.

В разделе 1 представляется метод построения пакетных дуальных коммутаторов с минимальной полосой пропускания, и рассчитываются их характеристики. В разделе 2 представляется базовый метод масштабирования числа каналов неблокируемых коммутаторов с пакетной дуальностью, и рассчитываются их характеристики. В разделе 3 рассматривается протокол маршрутизации для неблокируемых коммутаторов с пакетной дуальностью. В разделе 4 проводится сравнительный анализ характеристик разработанных коммутаторов как системных сетей. В Заключении подводится итог преобразования методики построения неблокируемых самомаршрутизируемых сетей [**1**, **2**] в проект работоспособной сети с сохранением ее исходных свойств, повышением быстродействия и радикальным сокращением полосы пропускания.

## 1. Пакетные дуальные коммутаторы

В неблокируемых сетях, разработанных в [**1**, **2**], в первом каскаде используется дуальный коммутатор РК$p$ с ЛЗ1, а в остальных – обычный коммутатор с ЛЗ0. В этих сетях используется компонентная база, состоящая из дуальных коммутаторов и отдельных мультиплексоров и демультиплексоров. Для маршрутизации пакетов в такой многокаскадной сети могут использоваться десятки управляющих сигналов. При этом весь набор этих сигналов должен передаваться в одном такте с информационным сигналом в каждом разряде пакета. Иначе говоря, для маршрутизации пакетов используется широкий

набор управляющих сигналов, и для их передачи требуется достаточно широкая полоса пропускания. Требуемую полосу пропускания для управляющих сигналов можно существенно уменьшить, если использовать линии задержки ЛЗ$\delta$ (Э3$\delta$) с длительностью пакета, где $\delta = M$, а $M$ – это число разрядов каждого пакета и $T_{\text{П}} = MT$, а $Tp\tau$ – длительность разряда. В этом случае весь набор управляющих сигналов можно передавать только один раз. Поэтому они могут быть переданы в заголовке каждого пакета. Такое представление управляющих сигналов существенно уменьшает необходимую полосу пропускания до величины, которая обеспечивает передачу информационных сигналов малой длительности $\tau$. При использовании длинных ЛЗ$T_{\text{П}}$ задержка каждого конфликтного разряда пакета на $p$ тактов заменяется на задержку каждого конфликтного пакета на время $T_{\text{П}}$. Иначе говоря, разрядный период $p$, выраженный в числе тактов, заменяется на пакетный период $T^*$, выраженный в числе повторных передач пакетов. Для дуального коммутатора ПК$p$ разрядный период и пакетный период совпадают $T^* = p$, а для дуального коммутатора РК$p$ пакетный период $T^{**}$, оказывается в среднем меньше $p$ (см. рисунок 3)! Действительно, на произвольной перестановке пакетов их маршрутизация по разным каналам осуществляется с одинаковой частотой. Тогда среднее значение пакетных периодов $T^{**}$, полученное для РК$p$ посредством имитационного моделирования, задается таблицей 1.

Таблица 1. Пакетные периоды для ПК$p$ и РК$p$

| $p$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| $T^*$ ПК$p$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| $T^{**}$ РК$p$ | 1,49 | 1,89 | 2,13 | 2,29 | 2,41 | 2,51 | 2,60 | 2,67 |

| $p$ | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|
| $T^*$ ПК$p$ | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| $T^{**}$ РК$p$ | 2,75 | 2,82 | 2,88 | 2,93 | 2,99 | 3,03 | 3,08 |

Режим «пакетных» задержек требует, чтобы линии задержки в разных каналах задавались с точностью в один такт длительности $\tau$. Для длинных ЛЗ$T_{\text{П}}$ требуемая точность может оказаться недостижимой. В этом случае неблокируемость дуальных коммутаторов можно обеспечить посредством установления соединений между процессорами-источниками пакетов и дуальным коммутатором первого каскада.

В простейшем дуальном коммутаторе по каждому входному каналу от источника необходимо иметь канал отдельный обратной связи (рисунок 4). По этому каналу передаются сигналы «свободно» и «занято». В начальный момент параллельной передачи пакетов по всем

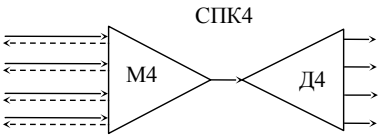



Рисунок 4. Простейший 4-канальный дуальный коммутатор с дополнительными каналами обратной связи (пунктир)

каналам передается «свободно». Если в мультиплексоре параллельно передаваемые пакеты попадают в конфликт, то один из них пропускается на выход, а остальные блокируются на выходе мультиплексора. При этом по всем каналами передается «занято». Сигнал «свободно» по всем каналам обратной связи снова передается после прохождения очередного пакета из мультиплексора в демультиплексор. В результате остальные источники повторно передают свои пакеты. В результате все $p$ источников бесконфликтно передадут свои пакеты за $T^*$ пакетных периодов. Вообще говоря, идея разрешения конфликтов посредством задержек их передачи не нова: например, она используется в методе SCMA/CD для общего канала. Новизна данного подхода состоит в параллельном разрешении конфликта только для небольшого числа пакетов, претендующих на передачу по сети через малые коммутаторы РК$p$, а не всех их как в общем канале, что существенно сокращает время разрешения конфликта. При этом время бесконфликтной передачи пакетов по сети складывается из времени разрешения конфликта и времени параллельной передачи бесконфликтных пакетов по прямым каналам.

К сожалению, разработанный ранее фотонный дуальный коммутатор [**1**, **2**] не может выдавать сигналы обратной связи без существенной своей переработки. Поэтому в дальнейшем рассматриваются только электронные коммутаторы.

В случае расширенного дуального коммутатора каждый демультиплексор для прохождения дополнительных каналов для соединений дополняется встречным мультиплексором (рисунок 5). Этот мультиплексор на выход пропускает сигналы «свободно» и «занято» только с того входа, на выход которого демультиплексор направил очередной пакет. В результате, как и в простейшем коммутаторе, все $p$ источников бесконфликтно передадут свои пакеты через расширенный коммутатор за $T^{**}$ пакетных периодов. Дополнительные схемы и каналы будут называться «схемами и каналами обратной связи (ОС)».

Будем оценивать коммутационную сложность дуальных коммутаторов в числе точек коммутации двумя характеристиками: $S$ для

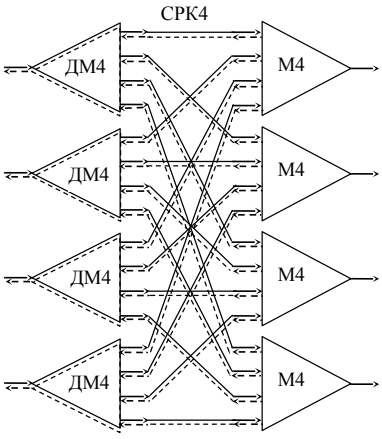



Рисунок 5. Расширенный 4-канальный дуальный коммутатор со схемами и каналами ОС (пунктир)

«чистых» дуальных коммутаторов, обозначаемых дальше как ДК$p$, не содержащих схем и каналов ОС, и $S^*$ для коммутаторов СДК$p$, их содержащих. Тогда их коммутационная сложность задается формулами (1):

$$(1)\qquad \begin{aligned} \text{ПК}p:&\quad S=2p & \qquad \text{РК}p:&\quad S=3p\\ \text{СПК}p:&\quad S^*=3p & \qquad \text{СРК}p:&\quad S^*=3p^2 \end{aligned}$$

В методике [1, 2] канальная отказоустойчивость неблокируемых сетей обеспечивается посредством использования дуальных коммутаторов с топологией квазиполных графов – КПГ$(N_1,p,\sigma)$ [4]. Они строятся на основе $p$-канальных дуальных коммутаторов ДК$p$ и дополнительного каскада демультиплексоров $1\times p$ и мультиплексоров $p\times 1$. Дуальные коммутаторы КПГ$(N_1,p,\sigma)$ имеют $N_1p(p-1)/\sigma+1$ каналов и являются неблокируемыми самомаршрутизируемыми коммутаторами на произвольных перестановках пакетов и обладают $(\sigma-1)$-канальной отказоустойчивостью. Для произвольных значений $p$ и $\sigma$ эти коммутаторы строятся переборными методами [4, 5]. На рисунке 6 приводится пример КПГ(4, 3, 2) со схемами и каналами ОС (пунктир) для установления соединения с источниками.

К сожалению, отмеченные свойства дуальных КПГ$(N_1,p,\sigma)$ выполняются только для СДК$p$ и ДК$p$, выполненных как коммутатор РК$p$. Дело в том, что коммутатор КПГ$(N_1,p,\sigma)$, выполненный на основе ПК$p$, оказывается блокируемым при ЭЗ1 на произвольных перестановках пакетов с сигналами в разных тактах разрядов [6]. Поэтому

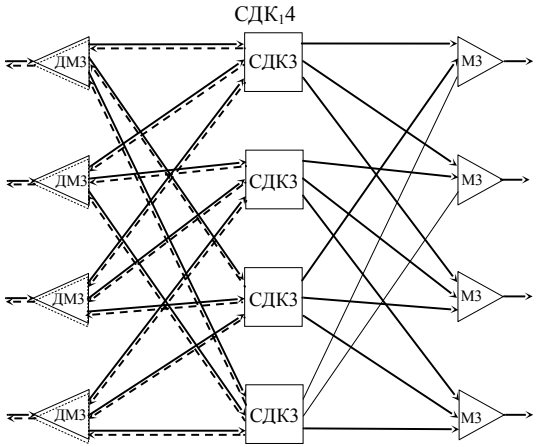



Рисунок 6. Дуальный 4-канальный коммутатор КПГ(4,3,2) со схемами и каналами ОС (штрих)

в дальнейшем рассматриваются только КПГ$(N_1, p, \sigma)$, выполненные на основе РК$p$. Предполагается, что СДК$p$ уже оснащен схемами и каналами ОС (рисунки 4 и 5).

Кроме того, каждый внешний демультиплексор для прохождения каналов ОС дополняется встречным мультиплексором (рисунок 6). Этот мультиплексор пропускает сигналы «свободно» и «занято» на выход только с того входа, на выход которого одноименный демультиплексор направил очередной пакет. Сигнал занято может появляться и в середине пакета, т.к. он только задерживает передачу следующего пакета. В результате, как и в дуальном коммутаторе, все $p$ источников бесконфликтно передадут свои пакеты через расширенный коммутатор за $\tau_{\text{П}}$ пакетных периодов.

Коммутатор КПГ$(N_1, p, \sigma)$ будем также называть дуальным коммутатором СДК$_1N_1$. Дуальный СДК$_1N_1$на базе КПГ$(N_1, p, \sigma)$ кроме канальной отказоустойчивости обеспечивает и начальное масштабирование числа каналов по сравнению с ДК$p$. Еще большего масштабирования числа каналов можно добиться при использовании коммутаторов с топологией квазиполных орграфов – КПОГ$(N_1, p)$, где $N_1 = p^2$ (рисунок 7). Эти коммутаторы изоморфны двумерным обобщенным гиперкубам или двумерным обобщенным мультикольцам, и существуют при любых значениях $p$ и не обладают канальной отказоустойчивостью.

Коммутатор на базе КПОГ$(N_1, p)$ будем также называть дуальным коммутатором СДК$_1N_1$ или ДК$_1N_1$. Как и для КПГ$(N_1, p, \sigma)$,

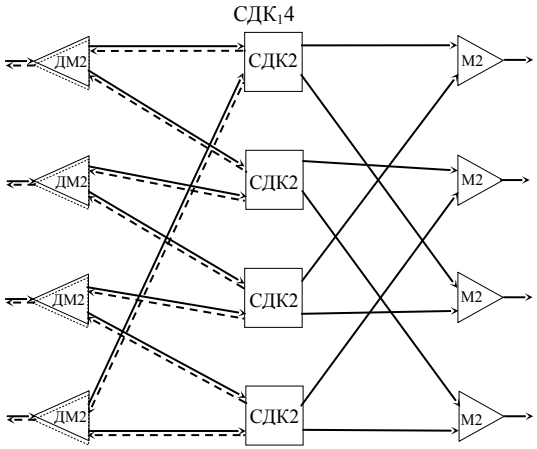



Рисунок 7. Дуальный 4-канальный коммутатор с топологией квазиполного орграфа КПОГ(4,2) со схемами и каналами ОС (штрих)

в дальнейшем рассматриваются только дуальные КПОГ$(N_1, p)$, выполненные на основе РК$p$. Одинаковые обозначения на рисунках 6 и 7 означают, что их внутренняя структура в дальнейшем каскадировании коммутаторов не играет роли.

Неблокируемые коммутаторы с топологией квазиполных графов или орграфов в дальнейшем используются в двух форматах: уже представленный СДК$_1N_1$ и такой же коммутатор без схем и каналов ОС$-$ДК$_1N_1$. Для этих коммутаторов дополнительно к коммутационным сложностям $S_1$ и $S_1^*$ оцениваются и канальные сложности $L_1$ и $L_1^*$, выраженные в числе внутренних каналов. Они задаются формулами (2), где $S$ и $S^*-$ это сложности ДК$p$ и СДК$p$.

$$\begin{array}{rll} \text{ДК}_1N_1: & S_1 = N_1(2p+S) & L_1 = N_1 2p \\ \text{СДК}_1N_1: & S_1^* = N_1(3p+S^*) & L_1^* = N_1 2p \end{array} \tag{2}$$

## 2. Базовое масштабирование числа каналов неблокируемых коммутаторов

Переход от коммутаторов СДК$p$ к коммутаторам СД$K_1N_1$ обеспечивает начальное масштабирование числа каналов неблокируемых коммутаторов. Дальнейшее масштабирование в методике [1, 2] осуществляется базовым методом посредством построения сетей с обменными

связями и их внутренним распараллеливанием. Поясним базовый метод на примере СДК$_1$2 с 1-канальной отказоустойчивостью (рисунок 8).

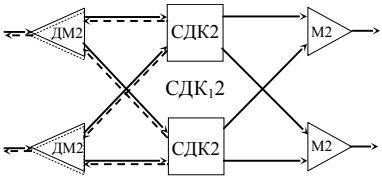



Рисунок 8. Дуальный коммутатор СДК$_1$2 на базе КПГ(2, 2, 2)

Из коммутаторов СДК$_1N$ строится 2-каскадная сеть с обменными связями С$N_2$ (рисунок 9), первый каскад которой содержит $N$ коммутаторов СДК$_1N$, а второй каскад – $N$ коммутаторов ДК$_1N$ без схем и каналов ОС.

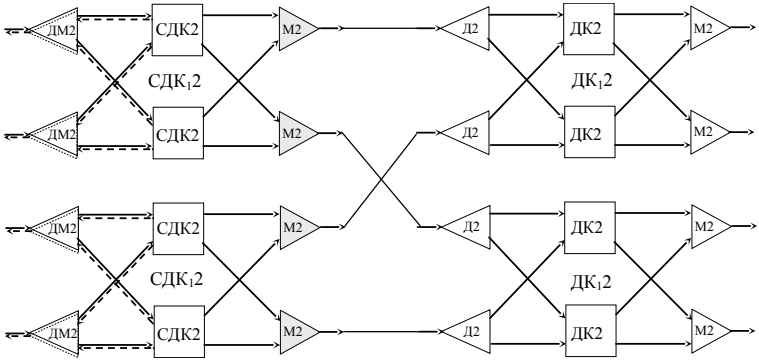



Рисунок 9. Сеть с обменными связями С$_2N_2$

Сеть С$N_2$ имеет $N_2N_1^2$ каналов, но является блокируемой из-за конфликтов на отмеченных заливкой выходных мультиплексорах первого каскада. На этих же мультиплексорах разрушается и канальная отказоустойчивость.

В сети С$N_2$ маршрутизация пакетов осуществляется коммутаторами СДК$_1N$ и ДК$_1N$, а также демультиплексорами ДМ$N$ и М$N$. Для этого они используют маршрутную информацию из заголовков пакетов, т.е. осуществляют самомаршрутизацию пакетов. Протокол самомаршрутизации рассматривается в разделе 3.

Однако сеть $С_2N_2$ можно преобразовать в неблокируемую сеть $СК_2N_2$с тем же числом каналов методом внутреннего распараллеливания[1,2]. В этом методе во второй каскад сети $С_2N_2$ добавляется еще $p$–1 его копий. Все $p$ копий нумеруются от 0 до $p$–1. На рисунке 9 добавляется одна копия, т.к. в $СДК_1N_1$ и $ДК_1N_1$ $p = 2$. Затем $i$-й вход $(0 \leq i \leq p–1)$ $j$-го конфликтного мультиплексора $(1 \leq j \leq N_1)$ в $k$-ом $СДК_1N_1$ $(1 \leq k \leq N_1)$ переключается $k$-й вход $j$-го $ДК_1N_1$ в $i$-й копии второго каскада (рисунок 10). Сами же мультиплексоры перемещаются

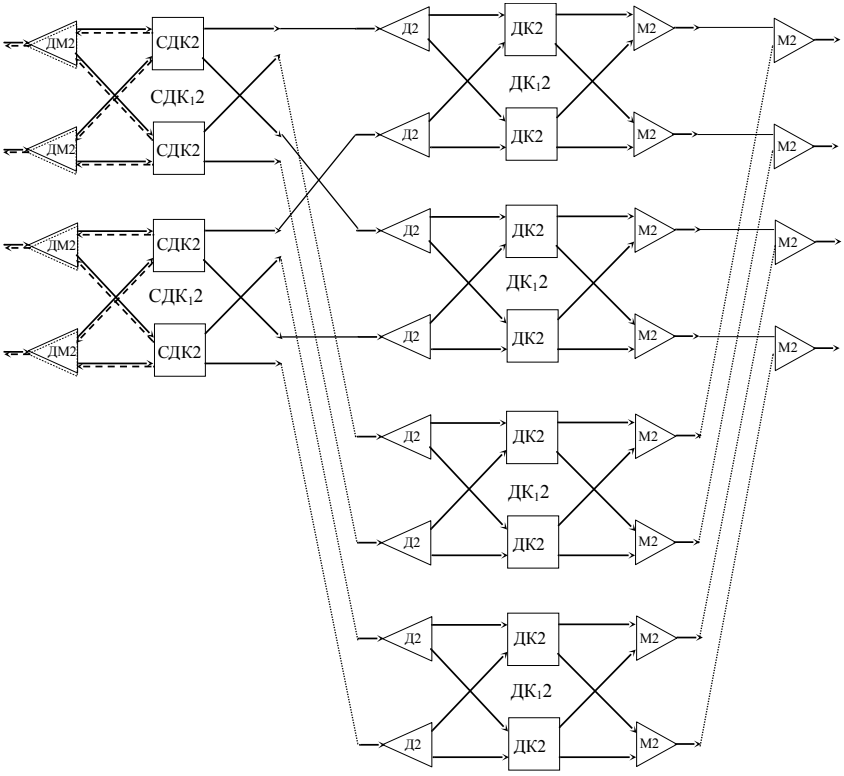



Рисунок 10. Неблокируемый коммутатор $СК_2N_2$ с 1-канальной отказоустойчивостью

так, чтобы объединять одноименные выходы одноименных $ДК_1N_1$ всех копий расширенного второго каскада. Все они вместе образуют неблокируемую схему первого измерения.

На входы всех неблокируемых коммутаторов $\text{ДК}_1N_1$ поступают пакеты разреженных перестановок, которые без конфликтов проходят на их выходы. При этом разные пути проходят через разные коммутаторы $\text{ДК}_1N_1$, что обеспечивает неблокируемость и канальную отказоустойчивость всего коммутатора $\text{СК}_2N_2$. В дальнейшем этот коммутатор будет использовать в двух форматах: как $\text{СК}_2N_2$ (с соединительными схемами и каналами ОС) и как $\text{К}_2N_2$ (без таких схем и каналов).

Коммутатор $\text{СК}_2N_2$ мы называем 2-каскадным, т.к. он построен из двух каскадов дуальных коммутаторов с топологией квазиполных графов или орграфов. Коммутатор $\text{СК}_2N_2$ содержит 4 слоя демультиплексоров $1\times p$, управляющих маршрутизацией пакетов (2 внешних слоя и 2 слоя внутри дуальных коммутаторов) и 2 слоя выходных мультиплексоров $p \times 1$. Сложности коммутаторов $\text{К}_2N_2$ и $\text{СК}_2N_2$ задаются формулами (3).

$$
\begin{aligned}
\text{К}_2N_2: \quad & S_2 = N_1S_1 + pN_1S_1, \\
& L_2 = N_1L_1 + pN_1L_1 + (p+1)N_2 \\
\text{СК}_2N_2: \quad & S^* = N_1S_1^* + pN_1S_1, \\
& L_2^* = N_1L_1^* + pN_1L_1 + (p+1)N_2
\end{aligned}
\tag{3}
$$

Ниже в таблицах 2–4 приводятся характеристики неблокируемых коммутаторов $\text{СК}_2N_2$, рассчитанные по формулам (1)–(3).

Таблица 2. Характеристики коммутаторов $\text{СК}_2N_2$ с 1-канальной отказоустойчивостью

| $p$ | $N_1$ | $N_2N_1^2$ | $T_2^{**}$ | $S_2^*$ | $L_2^*$ |
|---|---|---|---|---|---|
| 2 | 2 | 4 | 1,49 | $N_2^{3,70}$ | $N_2^{3,04}$ |
| 3 | 4 | 16 | 1,89 | $N_2^{2,69}$ | $N_2^{2,24}$ |
| 4 | 7 | 49 | 2,13 | $N_2^{2,39}$ | $N_2^{2,00}$ |
| 5 | 11 | 121 | 2,29 | $N_2^{2,24}$ | $N_2^{1,89}$ |
| 6 | 15 | 225 | 2,41 | $N_2^{2,19}$ | $N_2^{1,84}$ |
| 7 | 21 | 441 | 2,51 | $N_2^{2,13}$ | $N_2^{1,80}$ |
| 8 | 27 | 729 | 2,60 | $N_2^{2,10}$ | $N_2^{1,77}$ |

Масштабирование числа каналов можно продолжать описанным методом на основе уже построенных неблокируемых сетей $\text{СК}_2N_2$ [2].

Таблица 3. Характеристики коммутаторов $\mathrm{CK}_2N_2$ с 2-канальной отказоустойчивостью

| $p$ | $N_1$ | $N_2N_1^2$ | $T_2^{**}$ | $S_2^*$ | $L_2^*$ |
|---|---|---|---|---|---|
| 3 | 3 | 9 | 1,89 | $N_2^{2,60}$ | $N_2^{2,56}$ |
| 4 | 5 | 25 | 2,13 | $N_2^{2,78}$ | $N_2^{2,21}$ |
| 5 | 7 | 49 | 2,29 | $N_2^{2,87}$ | $N_2^{2,10}$ |
| 6 | 11 | 121 | 2,41 | $N_2^{2,77}$ | $N_2^{1,95}$ |
| 7 | 15 | 225 | 2,51 | $N_2^{2,75}$ | $N_2^{1,89}$ |
| 8 | 19 | 361 | 2,60 | $N_2^{2,74}$ | $N_2^{1,86}$ |

Таблица 4. Характеристики коммутаторов $\mathrm{CK}_2N_2$ без канальной отказоустойчивости

| $p$ | $N_1$ | $N_2N_1^2$ | $T_2^{**}$ | $S_2^*$ | $L_2^*$ |
|---|---|---|---|---|---|
| 2 | 4 | 16 | 1,49 | $N_2^{2,35}$ | $N_2^{2,02}$ |
| 3 | 9 | 81 | 1,89 | $N_2^{2,07}$ | $N_2^{1,78}$ |
| 4 | 16 | 256 | 2,13 | $N_2^{1,97}$ | $N_2^{1,70}$ |
| 5 | 25 | 625 | 2,29 | $N_2^{1,93}$ | $N_2^{1,66}$ |
| 6 | 36 | 1296 | 2,41 | $N_2^{1,90}$ | $N_2^{1,64}$ |
| 7 | 49 | 2401 | 2,51 | $N_2^{1,88}$ | $N_2^{1,62}$ |
| 8 | 64 | 4096 | 2,60 | $N_2^{1,87}$ | $N_2^{1,61}$ |

Строится двухкаскадная сеть $\mathrm{C}_4N_4$ с обменными связями, первый каскад которой состоит из $N_2$ сетей $\mathrm{CK}_2N_2$, а второй – из $N_2$ сетей $\mathrm{K}_2N_2$. Поэтому сеть $\mathrm{C}_4N_4$ имеет $N_4N_2^2N_1^4$ каналов. Сеть $\mathrm{C}_4N_4$ является блокируемой, т.к. она в первом каскаде содержит два слоя мультиплексоров $p \times 1$, на которых возможны конфликты и разрушается канальная отказоустойчивость. Для применения метода внутреннего распараллеливания во второй каскад сети $\mathrm{C}_4N_4$ добавляется $p^2$–1 его копий. Описанным выше методом конфликтные мультиплексоры первого слоя используются для построения $p$ неблокируемых схем первого измерения, но составленных уже из сетей $\mathrm{CK}_2N_2$ и $\mathrm{K}_2N_2$. Конфликтные мультиплексоры второго слоя используются для их объединения в неблокируемую схему второго измерения, которая и составляет 2-й каскад неблокируемой сети $\mathrm{CK}_4N_4$, который сохраняет канальную неблокируемость исходных сетей $\mathrm{CK}_2N_2$ и $\mathrm{K}_2N_2$. Все построения сети $\mathrm{C}_4N_4$ и доказательства ее неблокируемости и каналь-

ной отказоустойчивости приведены в [1, 2]. Коммутатор $\mathrm{СК}_4N_4$ мы называем 4-каскадным, т.к. он построен из четырех каскадов дуальных коммутаторов с топологией квазиполных графов или орграфов. Коммутатор $\mathrm{СК}_4N_4$ содержит 8 слоев демультиплексоров $1 \times p$, управляющих маршрутизацией пакетов и 4 слоя выходных мультиплексоров $p \times 1$. Сложности коммутаторов $\mathrm{К}_4N_4$ и $\mathrm{СК}_4N_4$ задаются формулами (4).

(4)

$$\mathrm{К}_4N_4: \quad S_4N_2 = S_2 + p^2N_2S_2 \quad L_4N_2 = L_2 + p^2N_2L_2 + (1+p+p^2)N_4$$

$$\mathrm{СК}_4N_4: \quad S_4^*N_2 = S_2^* + p^2N_2S_2 \quad L_4^*N_2 = L_2^* + p^2N_2L_2 + (1+p+p^2)N_4$$

Ниже в таблицах 5–7 приводятся характеристики неблокируемых коммутаторов $\mathrm{СК}_4N_4$, рассчитанные по формулам (1)–(4).

Таблица 5. Характеристики коммутаторов $\mathrm{СК}_4N_4$ с 1-канальной отказоустойчивостью

| $p$ | $N_2$ | $N_4N_2^2$ | $T_4^{**}$ | $S_4^*$ | $L_4^*$ |
|---|---|---|---|---|---|
| 2 | 9 | 81 | 1,49 | $N_2^{2,93}$ | $N_2^{2,63}$ |
| 3 | 49 | 2 401 | 1,89 | $N_2^{2,26}$ | $N_2^{2,04}$ |
| 4 | 169 | 28 561 | 2,13 | $N_2^{2,06}$ | $N_2^{1,87}$ |
| 5 | 441 | 194 481 | 2,29 | $N_2^{1,96}$ | $N_2^{1,79}$ |
| 6 | 961 | 923 521 | 2,41 | $N_2^{1,93}$ | $N_2^{1,76}$ |
| 7 | 1 521 | 2 313 441 | 2,51 | $N_2^{1,88}$ | $N_2^{1,72}$ |
| 8 | 3 249 | 10 55 6001 | 2,60 | $N_2^{1,86}$ | $N_2^{1,70}$ |

Таблица 6. Характеристики коммутаторов $\mathrm{СК}_4N_4$ с 2-канальной отказоустойчивостью

| $p$ | $N_2$ | $N_4N_2^2$ | $T_4^{**}$ | $S_4^*$ | $L_4^*$ |
|---|---|---|---|---|---|
| 3 | 9 | 81 | 1,89 | $N_2^{2,27}$ | $N_2^{2,31}$ |
| 4 | 25 | 625 | 2,13 | $N_2^{2,33}$ | $N_2^{2,05}$ |
| 5 | 49 | 2 401 | 2,29 | $N_2^{2,36}$ | $N_2^{1,97}$ |
| 6 | 121 | 14 641 | 2,41 | $N_2^{2,26}$ | $N_2^{1,85}$ |
| 7 | 225 | 50 625 | 2,51 | $N_2^{2,23}$ | $N_2^{1,81}$ |
| 8 | 361 | 130 321 | 2,60 | $N_2^{2,23}$ | $N_2^{1,79}$ |

Сравнение таблиц 2–4 с таблицами 5–7 показывает, что при квадратичном увеличении числа каналов 4-каскадных коммутаторов

Таблица 7. Характеристики коммутаторов $\mathrm{CK}_4N_4$ без канальной отказоустойчивости

| $p$ | $N_2$ | $N_4 = N_2^2$ | $T_4^{**}$ | $S_4^*$ | $L_4^*$ |
|---|---|---|---|---|---|
| 2 | 16 | 256 | 1,49 | $N_2^{1,96}$ | $N_2^{1,82}$ |
| 3 | 81 | 6 561 | 1,89 | $N_2^{1,79}$ | $N_2^{1,66}$ |
| 4 | 256 | 65 536 | 2,13 | $N_2^{1,74}$ | $N_2^{1,61}$ |
| 5 | 625 | 390 625 | 2,29 | $N_2^{1,72}$ | $N_2^{1,59}$ |
| 6 | 1 296 | 1 679 616 | 2,41 | $N_2^{1,70}$ | $N_2^{1,57}$ |
| 7 | 2 401 | 5 764 801 | 2,51 | $N_2^{1,69}$ | $N_2^{1,56}$ |
| 8 | 4 096 | 16 777 216 | 2,60 | $N_2^{1,69}$ | $N_2^{1,56}$ |

по сравнению с 2-каскаднми коммутаторами в них сохраняется период $T_4^*$, канальная отказоустойчивость, но резко уменьшаются удельная коммутационная и канальная сложности. Последнее свойство впервые получено для 4-х каскадных неблокируемых самомаршрутизируемых коммутаторов. Масштабирование числа каналов можно продолжать еще дальше описанным методом на основе построенных неблокируемых сетей $\mathrm{CK}_4N_4$. Строится двухкаскадная сеть $\mathrm{C}_8N_8$ с обменными связями, первый каскад которой состоит из $N_4$ сетей $\mathrm{CK}_4N_4$, а второй – из $N_4$ сетей $\mathrm{K}_4N_4$. Поэтому сеть $\mathrm{C}_8N_8$ имеет $N_8N_4^2N_2^4N_1^8$ каналов. Сеть $\mathrm{C}_8N_8$ является блокируемой, т.к. она в первом каскаде содержит четыре слоя мультиплексоров $p \times 1$, на которых возможны конфликты и разрушается канальная отказоустойчивость. Для применения метода внутреннего распараллеливания во второй каскад сети $\mathrm{C}_8N_8$ добавляется $p^4 - 1$ его копий. Описанным выше методом конфликтные мультиплексоры первого слоя используются для построения $p^3$ неблокируемых схем первого измерения, но составленных уже из сетей $\mathrm{K}_4N_4$. Конфликтные мультиплексоры второго слоя используются для их объединения в $p^2$ неблокируемых схем второго измерения. Конфликтные мультиплексоры третьего слоя используются для их объединения в $p$ неблокируемых схем третьего измерения. Наконец, конфликтные мультиплексоры четвертого слоя используются для их объединения в неблокируемую схему четвертого измерения, которая и составляет второй каскад неблокируемой сете $\mathrm{CK}_8N_8$, которая сохраняет канальную неблокируемость исходных сетей в $\mathrm{CK}_4N_4$ и $\mathrm{K}_4N_4$. Все построения сети $\mathrm{C}_8N_8$ и доказательства ее неблокируемости и канальной отказоустойчивости приведены в[1,2]. Коммутатор $\mathrm{CK}_8N_8$ мы называем 8-каскадным, т.к.

он построен из восьми каскадов дуальных коммутаторов с топологией квазиполных графов или орграфов. Коммутатор $\mathrm{CK}_8N_8$ содержит 16 слоев демультиплексоров $1\times p$, управляющих маршрутизацией пакетов. Сложности коммутаторов $\mathrm{K}_8N_8$ и $\mathrm{CK}_8N_8$ задаются формулами (5).

$$\begin{aligned}\mathrm{CK}_8N_8: \quad & S_8^* = N_4S_4^* + p^4N_4S_4,\\ & L_8^* = N_8L_8^* + p^4N_4L_4 + (1+p+p^2+p^3)N_8\end{aligned} \tag{5}$$

Ниже в таблицах 8–10 приводятся характеристики неблокируемых коммутаторов $\mathrm{CK}_8N_8$, рассчитанные по формулам (1)–(5).

Таблица 8. Характеристики коммутаторов $\mathrm{CK}_8N_8$ с 1-канальной отказоустойчивостью

| $p$ | $N_4$ | $N_8N_4^2$ | $T_8^{**}$ | $S_8^*$ | $L_8^*$ |
|---|---|---|---|---|---|
| 2 | 16 | 256 | 1,49 | $N_2^{2,48}$ | $N_2^{2,33}$ |
| 3 | 256 | 65 536 | 1,89 | $N_2^{2,03}$ | $N_2^{1,92}$ |
| 4 | 2 401 | 5 764 801 | 2,13 | $N_2^{1,88}$ | $N_2^{1,79}$ |
| 5 | 14 641 | 2,14E+08 | 2,29 | $N_2^{1,82}$ | $N_2^{1,73}$ |
| 6 | 50 625 | 2,56E+09 | 2,41 | $N_2^{1,80}$ | $N_2^{1,71}$ |
| 7 | 194 481 | 3,78E+10 | 2,51 | $N_2^{1,76}$ | $N_2^{1,68}$ |
| 8 | 531 441 | 2,82E+11 | 2,60 | $N_2^{1,75}$ | $N_2^{1,67}$ |

Таблица 9. Характеристики коммутаторов $\mathrm{CK}_8N_8$ с 2-канальной отказоустойчивостью

| $p$ | $N_4$ | $N_8N_4^2$ | $T_8^{**}$ | $S_8^*$ | $L_8^*$ |
|---|---|---|---|---|---|
| 3 | 81 | 6 561 | 1,89 | $N_2^{2,14}$ | $N_2^{2,16}$ |
| 4 | 625 | 390 625 | 2,13 | $N_2^{2,10}$ | $N_2^{1,96}$ |
| 5 | 2 401 | 5 764 801 | 2,29 | $N_2^{2,09}$ | $N_2^{1,90}$ |
| 6 | 14 641 | 2,14E+08 | 2,41 | $N_2^{2,00}$ | $N_2^{1,80}$ |
| 7 | 50 625 | 2,56E+09 | 2,51 | $N_2^{1,98}$ | $N_2^{1,78}$ |
| 8 | 130 321 | 1,7E+10 | 2,60 | $N_2^{1,97}$ | $N_2^{1,75}$ |

Сравнение таблиц 5–7 с таблицами 8–10 показывает, что при квадратичном увеличении числа каналов 8-каскадных коммутаторов по сравнению с 4-каскаднми коммутаторами в них сохраняется период $T_8^*$, канальная отказоустойчивость, но имеет место дальнейшее уменьшение удельной коммутационной и канальной сложностей.

Таблица 10. Характеристики коммутаторов $СК_8N_8$ без канальной отказоустойчивости

| $p$ | $N_4$ | $N_8 = N_4^2$ | $T_8^{**}$ | $S_8^*$ | $L_8^*$ |
|---|---|---|---|---|---|
| 2 | 256 | 65 536 | 1,49 | $N_2^{1,74}$ | $N_2^{1,66}$ |
| 3 | 6 561 | 430 46 721 | 1,89 | $N_2^{1,65}$ | $N_2^{1,58}$ |
| 4 | 65 536 | 4,29E+09 | 2,13 | $N_2^{1,62}$ | $N_2^{1,55}$ |
| 5 | 390 625 | 1,53E+11 | 2,29 | $N_2^{1,61}$ | $N_2^{1,54}$ |
| 6 | 1 679 616 | 2,82E+12 | 2,41 | $N_2^{1,60}$ | $N_2^{1,54}$ |
| 7 | 5 764 801 | 3,32E+13 | 2,51 | $N_2^{1,60}$ | $N_2^{1,53}$ |
| 8 | 16 777 216 | 2,81E+14 | 2,60 | $N_2^{1,59}$ | $N_2^{1,53}$ |

Последнее свойство впервые получено для 8-х каскадных неблокируемых самомаршрутизируемых коммутаторов.

## 3. Протокол передачи пакетов для дуальных неблокируемых коммутаторов с разным числом каскадов

Самомаршрутизация пакетов через коммутаторы $СДК_1N_1$ или $ДК_1N_1$ с топологией $КПГ(N_1, p, \sigma)$ или $КПОГ(N_1, p)$ осуществляется с использованием в заголовке пакета двух адресов $A_1$ и $A_2$ из $\log_2 p$-разрядных двоичных чисел. Адрес $A_1$ используется для маршрутизации пакетов входным демультиплексором коммутаторов $КПГ(N_1, p, \sigma)$ или $КПОГ(N_1, p)$. Адрес $A_2$ используется для маршрутизации пакетов демультиплексором дуального коммутатора $ДКp$. Эти демультиплексоры осуществляют коммутацию пакетов по текущему адресу одновременно с его уничтожением, т.е. без дальнейшей передачи по каналам коммутатора. В результате каждый пакет покидает коммутаторы $СДК_1N_1$ или $ДК_1N_1$ без адресов $A_1$ и $A_2$ в своем заголовке. Заголовок коммутатора $СК_2N_2$ содержит четыре таких адреса – $A_{1,1}$, $A_{1,2}$, $A_{2,1}$, $A_{2,2}$. Из них первые два используются коммутаторами $СДК_1N_1$, а вторые два – коммутаторами $ДК_1N_1$. Заголовок коммутатора $СК_4N_4$ содержит восемь таких адресов – $A_{1,1}$, $A_{1,2}$, $A_{2,1}$, $A_{2,2}$, $A_{3,1}$, $A_{3,2}$, $A_{4,1}$, $A_{4,2}$. Из них первые четыре используются коммутаторами $СК_2N_2$, а вторые четыре – коммутаторами $К_2N_2$. Заголовок коммутатора $СК_8N_8$ содержит шестнадцать таких адресов – $A_{1,1}$, $A_{1,2}$, $A_{2,1}$, $A_{2,2}$, $A_{3,1}$, $A_{3,2}$, $A_{4,1}$, $A_{4,2}$, $A_{5,1}$, $A_{5,2}$, $A_{6,1}$, $A_{6,2}$, $A_{7,1}$, $A_{7,2}$, $A_{8,1}$, $A_{8,2}$. Из них первые восемь используются коммутаторами $СК_4N_4$, а вторые восемь – коммутаторами $К_4N_4$.

## 4. Сравнительный анализ результатов

В построенной сети каждый разряд пакета имеет длительность в 1 сигнал-такт. Передача пакетов источниками ведется по прямым каналам за минимальное время. Возможные конфликты источников разрешаются посредством организации борьбы между ними за доступ к сети и повторных передач пакетов от источников, проигравших эту борьбу. Максимальное число повторных передач равно $p$, но в среднем оно значительно меньше при больших $p$ (см. таблицу 1). Доступ к сети обеспечивается посредством организации соединений между источниками и первым каскадом сети за минимальное время. Сами конфликты возможны только в смежных группах по $p$ источников в каждой. Внутри сети возникновение конфликтов предотвращается посредством ее внутреннего распараллеливания. В результате, произвольная перестановка реализуется бесконфликтно по сети за время последовательной передачи не более $p$ пакетов. Близкими свойствами обладают перестраиваемые сети [**8**–**10**], в которых можно составить бесконфликтное расписание для каждой конкретной перестановки. Однако эти сети оказываются блокируемыми на произвольных перестановках.

В перестраиваемых сетях заблокированные пакеты повторно передаются после передачи бесконфликтных пакетов. В сетях с прямыми каналами повторные передачи осуществляются на основе установления соединений между источниками и приемниками и отсутствия подтверждения бесконфликтной передачи пакетов. Ожидание таких подтверждений занимает много больше времени, чем установление соединений в рассматриваемой сети. Кроме того, самих блокированных пакетов и повторных их передач может быть значительно больше вследствие возможности конфликтов не только между смежными источниками в группах. Сети с топологией обобщенных гиперкубов [**11**] в общем случае не являются даже перестраиваемыми сетями [**12**]. Однако их можно сделать перестраиваемыми посредством увеличения числа каналов в некоторых измерениях [**13**, **14**]. Кроме того, двумерные гиперкубы и мультикольца являются неблокируемыми самомаршрутизируемыми сетями [**1**, **2**]. Они имеют топологию квазиполных орграфов в схемной базе коммутаторов $p \times p$ демультиплексоров $p \times 1$ и мультиплексоров $1 \times p$. Они имеют малое число каналов $N_1 = p^2$ и коммутационную сложность больше сложности полного коммутатора, а также и существенно меньшую канальную сложность. Такую топологию имеют многоканальные коммутаторы *YARK* и *ROSETTA*[**9**, **15**–**17**]

в сетях *Dragonfly* и *Slingshot*. Сети с топологией квазиполных орграфов допускают масштабирование числа каналов с сохранением свойств неблокируемости и самомаршрутизируемости, но при быстро растущей схемной сложности. В частности, при одинаковом числе каналов их коммутационная сложность в несколько раз больше сложности построенных 2-каскадных сетей $\mathrm{K}_2N_2$, на порядок больше сложности 4-каскадных сетей $\mathrm{K}_4N_4$, и на несколько порядков больше сложности 8-каскадных сетей $\mathrm{K}_8N_8$ [2]. Близкими свойствами обладают сети с топологией квазиполных графов, изоморфных неполным уравновешенным симметричным блок-схемам [18]. Они содержат $N_1 = p(p - -1)/\sigma + 1$ каналов и обладают свойством $(\sigma - 1)$-канальной отказоустойчивости. Эти сети значительно хуже поддаются масштабированию [2].

В некотором смысле промежуточными свойствами обладают многокаскадные неблокируемые сети Клоза [19, 20]. Разработана их структура, но не предложено никаких процедур параллельной самомаршрутизации. Эти сети имеют элементную базу из квадратных коммутаторов $N \times N$ и трапециевидных коммутаторов $N \times 2N$ и $2N \times N$. Реализуем такой 5-каскадный коммутатор с использованием коммутатора *YARK* [9] с $N_0 = 64$ каналами и коммутационной сложностью $S_0 = 64^2$. Для реализации трапециевидного коммутатора $32 \times 64$ используется один коммутатор *YARK*, и он же используется для реализации 2-х квадратных коммутаторов $32 \times 32$.

В 3-каскадном варианте неблокируемого коммутатора входной и выходной каскады содержат $N_1 = 32$ трапециевидных коммутаторов, хребет содержит $2N_1$ квадратных коммутаторов, все они содержат по 32 коммутатора *YARK*. 3-каскадный вариант имеет $N_3 = 1024$ канала и коммутационную сложность $S_3N_3^{1,86}$.

В 5-каскадном варианте входной и выходной каскады содержит $N_3$ трапециевидных коммутаторов, а хребет содержит $2N_1$ 3-каскадных коммутаторов. 5-каскадный вариант имеет $N_5 = 32$–$678$ каналов и коммутационную сложность $S_5 = 2N_3S_0 + 2N_1S_3N_5^{1,67}$. Таким образом, теоретический неблокируемый коммутатор Клоза при сравнимом числе каналов (см. таблицы 1 и 5) имеет сопоставимую сложность с неблокируемыми сетями $\mathrm{K}_4N_4$ и $\mathrm{K}_8N_8$ и несколько большее быстродействие. Однако он практически не реализуем из-за отсутствия процедур параллельной самомаршрутизации. Можно создать неблокируемые самомаршрутизируемые сети на любое число каналов с номинальным быстродействием. Они создаются посредством каскадирования сети

с топологией квазиполного графа (двумерного гиперкуба таблица 4). Однако эти сети имеют существенно большую коммутационную сложность. Так при одинаковом числе каналов они в несколько раз сложнее построенной 2-каскдной сети, на 1,5 порядка сложнее 4-каскадной сети и на 3 порядка сложнее 8- каскадной сети [2]. Автору не известны другие неблокируемые сети с прямыми каналами. В большинстве системных сетей передача пакетов осуществляется скачками с их буферизацией в узлах сети. Максимальное число скачков от источника до приемника задает диаметр сети. Подчеркнем, что в сетях с прямыми каналами передача пакетов осуществляется за один скачок, т.е. их диаметр равен 1. Известны сети с малым диаметром в 3 скачка. Это сети со структурой иерархии полных или квазиполных орграфов [15, 16, 21]. Известны простые системные сети с большими диметрами, измеряемых десятками скачков. Это сети со структурой многомерных торов [22–24], в частности это 3-мерная сеть *TOFU* с сотнями тысяч абонентов. Канальной отказоустойчивостью считается возможность сохранения полнодоступности сети при отказах каналов с сохранением ее исходных характеристик (неблокируемость сети, задержки передачи или диаметра сети). В чистом виде канальной отказоустойчивостью обладают, по-видимому, только сети с топологией квазиполных графов и построенные на них сети в [1, 2].

В других сетях восстановление полнодоступности сети при отказах каналов достигается посредством ухудшения тех или них характеристик. Так, в перестраиваемой сети Клоза при отказах каналов нагрузка на оставшиеся каналы увеличивается, что увеличивает число конфликтов и увеличивает задержки передачи части пакетов. Аналогично для сетей со структурой обобщенных гиперкубов. В сетях с диаметром в 3 скачка [15–17] отказы каналов приводит к увеличению диаметра сети до 5 скачков. В сети *TOFU* обеспечение полнодоступности сети достигается посредством увеличения числа ее измерений до 6, а диаметра сети на 1 скачок. Таким образом, предложенные в данной работе сети при широкой масштабируемости обладают набором свойств, недостижимых ни в одной из известных сетей [25, 26].

**Заключение**

В работе проведено воплощение методики построения неблокируемой самомаршрутизируемой сети широкой масштабируемости [1, 2] в проект самой сети. В этой сети сведена к минимуму полоса пропускания, необходимая для передачи маршрутной информации. Эта

минимизации достигнута посредством перехода от разрядно–дуального способа разрешения конфликтов пакетов к пакетно-дуальному способу без изменения структуры сети. При пакетно-дуальном способе маршрутная информация передается однократно в заголовках пакетов в виде для каждого каскада в виде $\log_2 p$-разрядных двоичных чисел. При этом дуальность выражается в повторных передачах пакетов, конфликтующих в первом каскаде сети. Переход к пакетной дуальности осуществляется посредством увеличения элементов задержки в каналах обратной связи первого каскада до длины пакета данных. Трудности с реализацией длинных линий задержки с необходимой точностью привели к их исключению из сети и к реализации их функций борьбой конфликтующих источников за вход в сеть в процессе их соединения с ней. Разработана неблокируемая самомаршрутизируемая сеть широкой масштабируемости, обладающая канальной отказоустойчивостью. Она обеспечивает бесконфликтную передачу пакетов по прямым каналам при произвольных перестановках. Такая передача осуществляется после разрешения конфликтов источников в процессе их соединения с сетью за минимальное время из-за отсутствия тактовых задержек и сопровождается повторными передачами пакетов от проигравших борьбу источников. Число повторных передач не превосходит аналогичных передач в перестраиваемых сетях, но выполняется за меньшее время. Предложенная сеть выполнена в схемной базе малоканальных коммутаторов $p \times p$, демультиплексоров $1 \times p$ и мультиплексоров $p \times 1$. Сеть имеет самоподобную структуру: бесконфликтная 8-каскадная сеть строится из параллельно включенных бесконфликтных 4-каскадных сетей, которые состоят из параллельно включенных бесконфликтных 2-каскадных сетей, составленных из параллельно включенных сетей с топологией квазиполных графов или орграфов. При этом во всех этих сетях возникновение повторных конфликтов предотвращается посредством их внутреннего распараллеливания. Предложенная сеть может работать в условиях любого трафика. Для этого достаточно, чтобы выходные мультиплексоры пропускали только один из входных пакетов, а остальные блокировали. При этом повторная передача блокированных пакетов должна осуществляться источниками за большее время, на основе уже следующего второго уровня соединений межу источниками и приемниками как в перестраиваемых сетях. При этом задержки передачи пакетов в построенной сети станут сопоставимыми с задержками в перестраиваемых сетях [**9**, **10**]. Характеристики построенных сетей рассчитаны для случай синхронной

передачи пакетов одинаковой длительности. Однако предложенная сеть, функционирующая на основе соединений источников с сетью, остается работоспособной и в случае асинхронных передач пакетов разной длительности . Однако возможен другой вариант функционирования предложенной сети – как распределенного вычислителя с выполнением функций АЛУ выходными мультиплексорам на основе методики вычислений в общем канале (ВОК) [27]. Построение сети с такими функциями – это следующая задача автора.

## Список литературы


[1] Барабанова Е. А., Вытовтов К. А., Подлазов В. С. *Неблокируемые отказоустойчивые двухкаскадные дуальные фотонные коммутаторы* // Проблемы управления.– 2021.– № 4.– с. 82–92. ↑49, 51, 52, 54, 55, 57, 62, 66, 68

[2] В. С. Подлазов *Неблокируемые отказоустойчивые дуальные фотонные коммутаторы широкой масштабируемости* // Пробл. управл.– 2021.– № 5.– с. 70–87. ↑49, 51, 52, 54, 55, 57, 60, 62, 66, 67, 68

[3] Stepanenko S. *Structure and Implementation Principles of a Photonic Computer* // EPJ Web of Conferences.– 2019.– Vol. **224**.– 04002.– 7 pp. ↑49

[4] Каравай М. Ф., Подлазов В. С. *Метод инвариантного расширения системных сетей многопроцессорных вычислительных систем. Идеальная системная сеть* // Автомат. и телемех.– 2010.– № 12.– с. 166–177. ↑55

[5] Каравай М. Ф., Подлазов В. С. *Расширенные блок-схемы для идеальных системных сетей* // Пробл. управл.– 2012.– № 4.– с. 45–51. ↑55

[6] Барабанова Е. А., Вытовтов К. А., Подлазов В. С. *Многокаскадные коммутаторы для оптических и электронных суперкомпьютерных систем* // *Материалы 8-го Национального Суперкомпьютерного Форума*, НСКФ-2019 (26–29 ноября 2019 года, ИПС имени А.К. Айламазяна РАН, Переславль-Залесский, Россия). ↑55

[7] Newman P. *Fast packet switching for integrated services*, A dissertation submitted for the degree of Doctor of Philosophy.– Wolfson College, University of Cambridge.– 1988.– 159 pp. ↑

[8] Pipenger N. *On rearrangeable and non-blocking switching networks* // J. Comput. Syst. Sci.– 1978.– Vol. **17**.– No. 2.– pp. 145–162. ↑66

[9] Scott S., Abts D., Kim J., Dally W. *The Black Widow High-radix Clos Network* // *Proc. 33rd Intern. Symp. Comp. Arch.*, ISCA'2006 (17–21 June 2006, Boston, MA, USA).– IEEE.– 2006.– ISBN 0-7695-2608-X.– pp. 16–28. ↑66, 67, 69

[10] *Mellanox OFED for Linux User Manual*, Rev 2.3–1.0.1.– Melanox Technologies, Ltd..– 2014.– 208 pp. ↑66, 69

[11] Bhuyan L. N., Agrawal D. P. *Generalized hypercube and hyperbus structures for a computer network* // IEEE Trans. on Computers.– 1984.– Vol. **C-33**.– No. 4.– pp. 323–333. ↑66

[12] Lubiw A. *Counterexample to a conjecture of Szymanski on hypercube routing* // Inform. Proc. Let.– 1990.– Vol. **35**.– No. 2.– pp. 57–61. ↑66

[13] Gu Q.-P., Tamaki H. *Routing a permutation in hypercube by two sets of edge-disjoint paths* // J. of Parallel and Distributed Comput.– 1997.– Vol. **44**.– No. 2.– pp. 147–152. ↑66

[14] Efe K. *A variation on the hypercube with lower diameter* // IEEE Trans. Computers.– 1991.– Vol. **40**.– No. 11.– pp. 1312–1316. ↑66

[15] Alverson B., Froese E., Kaplan L., Roweth D. *Cray® XC™ Series Network*, WP-Aries01-1112.– Cray Inc..– 28 pp. ↑66, 68

[16] Kim J., Dally W. J., Scott S., Abts D. *Technology-driven, highly-scalable dragonfly topology* // *Proceedings of the 35th annual International Symposium on Computer Architecture*, ISCA 2008 (21–25 June, 2008, Beijing, China).– 2008.– ISBN 978-0-7695-3174-8.– pp. 77–88. ↑66, 68

[17] De Sensi D., Di Girolamo S., McMahon K. H., Roweth D., Hoefler T. *An in-depth analysis of the slingshot interconnect*.– 2020.– 13 pp. arXiv:2008.08886 [cs.DC] ↑66, 68

[18] Холл М. *Комбинаторика*, Главы 10–12.– М.: Мир.– 1970.– 424 с. ↑67

[19] Clos C. *A study of non-locking switching networks* // Bell System Tech. J.– 1953.– Vol. **32**.– No. 2.– pp. 406–424. ↑67

[20] Бенеш В. Э. // *Математические основы теории телефонных сообщений*, М.: Связь.– 1968.– с. 83–150. ↑67

[21] Arimili B., Arimilli R., Chung V., Clark S., Denzel W., Drerup B., Hoefler T., Joyner J., Lewis J., Li J., Ni N., Rajamony R. *The PERCS high-performance interconnect* // *HOTI '10: Proceedings of the 2010 18th IEEE Symposium on High Performance Interconnects* (18–20 August 2010, Mountain View, CA, USA).– 2010.– ISBN 978-0-7695-4208-9.– pp. 75–82. ↑68

[22] Alverson R., Roweth D., Kaplan L. *The Gemini system interconnect* // *HOTI '10: Proceedings of the 2010 18th IEEE Symposium on High Performance Interconnects* (18–20 August 2010, Mountain View, CA, USA).– 2010.– ISBN 978-0-7695-4208-9.– pp. 83–87. ↑68

[23] Stegailov V., Agarkov A., Biryukov S., Ismagilov T., Khalilov M., Kondratyuk N., Kushtanov E., Makagon D., Mukosey A., Semenov A., Simonov A., Timofeev A., Vecher V. *Early performance evaluation of the hybrid cluster with torus interconnect aimed at molecular dynamics simulations* // *PPAM 2017: Parallel Processing and Applied Mathematics*, Lecture Notes in Computer Science.– vol. **10777**, Cham: Springer.– 2017.– ISBN 978-3-319-78023-8.– pp. 327–336. ↑68

[24] Ajima Y., Inoue T., Hiramoto Sh., Shimiz T. *Tofu: Interconnect for the K computer* // FUJITSU Sci. Tech. J.– 2012.– Vol. **48**.– No. 3.– pp. 280–285. ↑68

[25] Flajslik M., Borch E., Parker M. A. *Megafly: A topology for exascale systems* // *ISC High Performance 2018: High Performance Computing*, Lecture Notes in Computer Science.– vol. **10876**, Cham: Springer.– 2018.– ISBN 978-3-319-92039-9.– pp. 289–310. ↑68

[26] De Sensi D., Di Girolamo S., Ashkboos S., Li Sh., Hoefler T. *Flare: flexible in-network allreduce* // *SC '21: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (14–19 November, 2021, St. Louis, Missouri, USA), New York: ACM.– 2021.– ISBN 978-1-4503-8442-1.– 16 pp. ↑68

[27] Прангишвили И. В., Подлазов В. С., Стецюра Г. Г. *Локальные микропроцессорные вычислительные сети*, Гл. 6.– М.: Наука.– 1984.– 175 с. ↑70

*Поступила в редакцию* 05.05.2022;
*одобрена после рецензирования* 11.07.2022;
*принята к публикации* 19.09.2022.

Рекомендовал к публикации *д.ф.-м.н. С. А. Амелькин*

## Информация об авторе:

Виктор Сергеевич Подлазов

Д. т. н., гл.н.с. Института проблем управления им. В.А. Трапезникова РАН, научные интересы: архитектуры интерконнекта и маршрутизация в суперкомпьютерных системах

0000-0002-9175-1138

*e-mail:* podlazov@ipu.ru
podlazov@gmail.com

*Автор заявляет об отсутствии конфликта интересов.*

ISSN 2079-3316 **PROGRAM SYSTEMS: THEORY AND APPLICATIONS** **vol. 13**, No 4(55), pp. 47–76
*Research Article* HARDWARE AND SOFTWARE FOR SUPERCOMPUTERS

UDC 004.724.2 + 004.272.43
10.25209/2079-3316-2022-13-4-47-76

# Multichannel non-blocking system area network with direct channels

Viktor Sergeevich **Podlazov**
V. A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia
✉*podlazov@ipu.ru* *(learn more about the author in Russian on p. 72)*

Abstract. An unblockable non-blocking self-routing network with direct channels has been developed, in which packet conflicts are resolved at the entrance to the network by the procedure of connecting sources to the first cascade of the network , providing a packet duality. Packets blocked during conflict resolution are retransmitted by sources with minimal delays. Within the network, the occurrence of conflicts is prevented by means of parallelization of the network itself. The network is designed in 2-, 4- and 8-stage versions with scaling of the number of channels from several hundred to many millions, keeping the same network performance. The network can provide 1- or 2-channel fault tolerance at the same link rate. The overhead cost of achieving these properties is higher complexity of the network, that becomes comparable to the complexity of the theoretical non-blocking Clos switch, which has no known practical implementation.The purpose of the proposed networks is photonic networks with routing information transmission in packet headers represented as one-time-use binary numbers. The proposed networks are made in an extended circuit basis, consisting of switches and separate multiplexers and demultiplexers. The paper presents the characteristics of the constructed networks with the specified method of presenting routing information. (*In Russian*).

Key words and phrases: dual switch, packet duality, multiplexers and demultiplexers, multi-stage switch, conflict-free self-routing, non-blocking switch, static self-routing, quasi-complete digraph, quasi-complete graph, switching properties, direct channels, scalability and speed

2020 *Mathematics Subject Classification*: 65Y05; 68Q10

For citation: Viktor S. Podlazov. *Multichannel non-blocking system area network with direct channels* // Program Systems: Theory and Applications, 2022, **13**:4(55), pp. 47–76. (*In Russian*). http://psta.psiras.ru/read/psta2022_4_47-76.pdf

## References


[1] E. A. Barabanova, K. A. Vytovtov, V. S. Podlazov. "Non-blocking fault-tolerant two-stage dual photon switches", *Control Sciences*, 2021, no. 4, pp. 67–76. doi ↑49, 51, 52, 54, 55, 57, 62, 66, 68

[2] V. S. Podlazov. "Non-blocking fault-tolerant dual photon switches with high scalability", *Control Sciences*, 2021, no. 5, pp. 61–76. doi ↑49, 51, 52, 54, 55, 57, 60, 62, 66, 67, 68

[3] S. Stepanenko. "Structure and Implementation Principles of a Photonic Computer", *EPJ Web of Conferences*, **224** (2019), 04002, 7 pp. doi ↑49

[4] M. F. Karavay, V. S. Podlazov. "An invariant extension method for system area networks of multicore computational systems. An ideal system network", *Autom. Remote Control*, **71**:12 (2010), pp. 2644–2654. doi ↑55

[5] M. F. Karavay, V. S. Podlazov. "Expanded block-diagrams for "ideal" system area networks", *Automation and Remote Control*, **74**:12 (2013), pp. 2180–2188. doi ↑55

[6] Ye. A. Barabanova, K. A. Vytovtov, V. S. Podlazov. "Multistage switches for optical and electronic supercomputer systems", *Materialy 8-go Natsional'nogo Superkomp'yuternogo Foruma*, NSKF-2019 (26–29 noyabrya 2019 goda, IPS imeni A.K. Aylamazyana RAN, Pereslavl'-Zalesskiy, Rossiya) (in Russian). ↑55

[7] P. Newman. *Fast packet switching for integrated services*, A dissertation submitted for the degree of Doctor of Philosophy, Wolfson College, University of Cambridge, 1988, 159 pp. URL ↑

[8] Pipenger N.. "On rearrangeable and non-blocking switching networks", *J. Comput. Syst. Sci.*, **17**:2 (1978), pp. 145–162. doi ↑66

[9] S. Scott, D. Abts, J. Kim, W. Dally. "The Black Widow High-radix Clos Network", *Proc. 33rd Intern. Symp. Comp. Arch.*, ISCA'2006 (17–21 June 2006, Boston, MA, USA), IEEE, 2006, ISBN 0-7695-2608-X, pp. 16–28. doi ↑66, 67, 69

[10] *Mellanox OFED for Linux User Manual*, Rev 2.3–1.0.1, Melanox Technologies, Ltd., 2014, 208 pp. URL ↑66, 69

[11] L. N. Bhuyan, D. P. Agrawal. "Generalized hypercube and hyperbus structures for a computer network", *IEEE Trans. on Computers*, **C-33**:4 (1984), pp. 323–333. doi ↑66

[12] A. Lubiw. "Counterexample to a conjecture of Szymanski on hypercube routing", *Inform. Proc. Let.*, **35**:2 (1990), pp. 57–61. doi ↑66

[13] Q.-P. Gu, H. Tamaki. "Routing a permutation in hypercube by two sets of edge-disjoint paths", *J. of Parallel and Distributed Comput.*, **44**:2 (1997), pp. 147–152. doi ↑66

[14] Efe K.. "A variation on the hypercube with lower diameter", *IEEE Trans. Computers*, **40**:11 (1991), pp. 1312–1316. doi ↑66

[15] B. Alverson, E. Froese, L. Kaplan, D. Roweth. *Cray® XC™ Series Network*, WP-Aries01-1112, Cray Inc., 28 pp. URL ↑66, 68

[16] J. Kim, W. J. Dally, S. Scott, D. Abts. "Technology-driven, highly-scalable dragonfly topology", *Proceedings of the 35th annual International Symposium on Computer Architecture*, ISCA 2008 (21–25 June, 2008, Beijing, China), 2008, ISBN 978-0-7695-3174-8, pp. 77–88. doi ↑66, 68

[17] De Sensi D., Di Girolamo S., K. H. McMahon, D. Roweth, T. Hoefler. *An in-depth analysis of the slingshot interconnect*, 2020, 13 pp. arXiv 2008.08886 [cs.DC] ↑66, 68

[18] M. Jr. Hall. *Combinatorial Theory*, Blaisdell Publishing Company, Waltham–Toronto–London, 1967, ISBN 978-0471315186, 310 pp. ↑67

[19] C. Clos. "A study of non-locking switching networks", *Bell System Tech. J.*, **32**:2 (1953), pp. 406–424. doi ↑67

[20] V. E. Benes (ed.). *Mathematical Theory of Connecting Networks and Telephone Traffic*, Mathematics in Science and Engineering, vol. **17**, Bell Telephone Laboratories. Academic Press, New York–London, 1965, ISBN 9780080955230, 319 pp. ↑67

[21] B. Arimili, R. Arimilli, V. Chung, S. Clark, W. Denzel, B. Drerup, T. Hoefler, J. Joyner, J. Lewis, J. Li, N. Ni, R. Rajamony. "The PERCS high-performance interconnect", *HOTI '10: Proceedings of the 2010 18th IEEE Symposium on High Performance Interconnects* (18–20 August 2010, Mountain View, CA, USA), 2010, ISBN 978-0-7695-4208-9, pp. 75–82. doi ↑68

[22] R. Alverson, D. Roweth, L. Kaplan. "The Gemini system interconnect", *HOTI '10: Proceedings of the 2010 18th IEEE Symposium on High Performance Interconnects* (18–20 August 2010, Mountain View, CA, USA), 2010, ISBN 978-0-7695-4208-9, pp. 83–87. doi ↑68

[23] V. Stegailov, A. Agarkov, S. Biryukov, T. Ismagilov, M. Khalilov, N. Kondratyuk, E. Kushtanov, D. Makagon, A. Mukosey, A. Semenov, A. Simonov, A. Timofeev, V. Vecher. "Early performance evaluation of the hybrid cluster with torus interconnect aimed at molecular dynamics simulations", *PPAM 2017: Parallel Processing and Applied Mathematics*, Lecture Notes in Computer Science, vol. **10777**, Springer, Cham, 2017, ISBN 978-3-319-78023-8, pp. 327–336. doi ↑68

[24] Y. Ajima, T. Inoue, Sh. Hiramoto, T. Shimizu. "Tofu: Interconnect for the K computer", *FUJITSU Sci. Tech. J.*, **48**:3 (2012), pp. 280–285. ↑68

[25] M. Flajslik, E. Borch, Parker M. A.. "Megafly: A topology for exascale systems", *ISC High Performance 2018: High Performance Computing*, Lecture Notes in Computer Science, vol. **10876**, Springer, Cham, 2018, ISBN 978-3-319-92039-9, pp. 289–310. ↑68

[26] De Sensi D., Di Girolamo S., S. Ashkboos, Sh. Li, T. Hoefler. "Flare: flexible in-network allreduce", *SC '21: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (14–19 November, 2021, St. Louis, Missouri, USA), ACM, New York, 2021, ISBN 978-1-4503-8442-1, 16 pp. ↑68

[27] I. V. Prangishvili, V. S. Podlazov, G. G. Stetsyura. *Local Microprocessor Computer Networks*, Ch. 6, Nauka, M., 1984 (in Russian), 175 pp. ↑70