УДК 004.728.3.057.4

Распределённые каналы приёма-передачи данных в сетевых архитектурах многопроцессорных систем

И.М. Косарев

ФГУ ФНЦ НИИСИ РАН, ikosarev@cs.niisi.ras.ru

***Аннотация* — С ростом числа компонентов многопроцессорной системы в числе важнейших выделяются методы обеспечения масштабируемости сетевой топологии и надёжности каналов приёма и передачи данных при сохранении высокой пропускной способности. Мультиплексирование портов коммутатора и резервирование физических линий связи в масштабе единого канала широко применяется для построения гибких и надёжных систем. Главный недостаток такого подхода в том, что суммарная пропускная способность такой сетевой архитектуры снижается пропорционально числу резервных сигнальных пар. В статье предложен метод реализации канального регламентирующего протокола инициализации и контроля полностью дуплексного канала связи с Nx/Mx/1x сигнальными парами, с возможностью мультиплексирования внутренних портов и резервированием передающих линий. Представленные подходы обеспечивают высокую пропускную способность канала в отсутствие многократно повторяющихся ошибок и повышенную надёжность при необходимой пропускной способности в случае повторяющихся ошибок одной или нескольких линий связи.**

***Ключевые слова* — дублирование, высокопроизводительный интерфейс, алгоритм настройки, горячее подключение, кросслинк.**

I. Введение

В состав современных многопроцессорных систем интегрировано большое число процессоров, коммуникационных контроллеров, мостов и т. д. Задачей коммуникационного контроллера является обеспечение производительного соединения с другими узлами вычислительной системы. Большинство архитектур из списка TOP500 используют интерфейсы с открытым стандартом, такие как Infiniband, PCIExpress, Gigabit Ethernet, 10G и другие. Между тем, за последние несколько лет число систем, использующих специально разработанные, не описанные в открытых стандартах архитектуры, неуклонно растёт. Достаточно сказать, что в списке на ноябрь 2015 года из первых в списке наиболее мощных 10 суперЭВМ, 9 используют проприетарную (proprietary) топологию сети [1]. Это объясняется ограниченностью применения таких систем и чёткой направленностью на выполнение определённого типа задач. Они проще в реализации, массивы логики компактнее и отличаются меньшим энергопотреблением. Другими словами, выделенные архитектуры наиболее подходят для встраиваемых систем, систем на кристалле, в производительных микропроцессорах и коммутаторах в составе суперЭВМ [2]. Для повышения надёжности обмена данными требуется реализовать минимально избыточный протокол с возможностью мультиплексирования двух и более портов транспортного уровня, который в отсутствие ошибок, позволял бы использовать пропускную способность всех линий в составе физического канала, обеспечивая передачу и приём данных путём равномерного распределения пакетов данных по всем линиям. В случае многократных ошибок система должна автоматически перестроиться, исключить неисправные линии, выбрать оптимальный режим с резервированием и необходимой пропускной способностью, не допустив при этом потери полезной информации. При реализации требований, предъявляемым к подобным архитектурам [3], становится актуальной и задача обеспечения совместимости реализуемых алгоритмов с требованиями некоторых спецификаций. Поскольку ряд открытых стандартов используют сходные по формату последовательности, а методы реализации ряда основных логических функций находятся вне рамок спецификаций, необходимый функционал, параметры канала, список входных и выходных требований необходимо задать при генерации исходных кодов IP-блоков.

II. Модель приёма-передачи данных

Обратимся к физическому и канальному уровням сетевой модели OSI [4]. Главной особенностью реализации протокола является программируемая ширина канала для каждого из портов транспортного уровня с возможностью автоматической инициализации, в том числе повторной после сбоев, а также обеспечение «горячего подключения». Предполагается синхронная дифференциальная передача данных в каждом направлении. Доступен параллельный режим Nx (N дифференциальных сигнальных пар в каждом направлении) для режима с максимальной производительностью, параллельный режим Mx (используются M дифференциальных сигнальных пар в каждом направлении) и одноканальный 1x режим (используется одна дифференциальная сигнальная пара в каждом направлении). Параметры N и M это целые натуральные числа. При этом справедливы следующие соотношения:

$$N = 2^{i} \quad (1),$$

$$M = 2^{i-k} \quad (2),$$

$$N > M > 1 \quad (3),$$

где i и k – любые целые натуральные числа. Из соотношений (1), (2) и (3) следует:

$$i = \log_2 N \quad (4),$$

$$i > k > 0 \quad (5),$$

$$k < \log_2 N \quad (6).$$

Таким образом, используя выражения (1) — (6), физический канал приёма-передачи данных может быть составлен из 4x/2x/1x, 8x/2x/1x, 8x/4x/1x, 16x/2x/1x, 16x/4x/1x, 16x/8x/1x сигнальных пар и так далее.

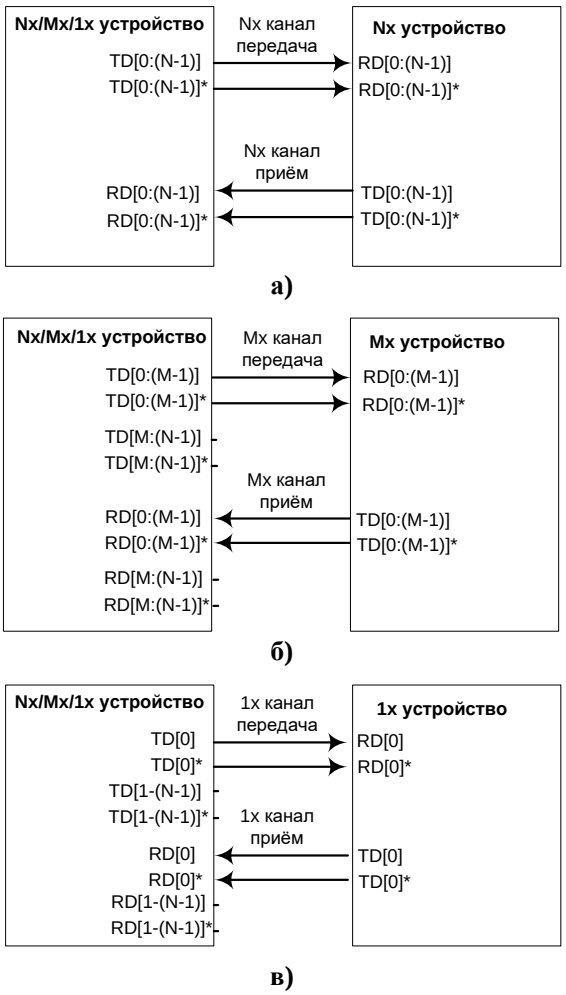

Рис. 1. Схема соединения устройства Nx/Mx/1x: а) с устройством Nx, б) с устройством Mx, в) с устройством 1x

Два устройства соединяются между собой линиями связи на основе оптоволокна или медными проводниками. Линии 0-N приёмника и передатчика подключаются соответственно (0-0, M-M, N-N). На рис. 1 показаны схемы соединения устройства Nx/Mx/1x с устройствами, поддерживающие только отдельные режимы Nx, Mx и 1x. На рис. 2 показана схема соединения двух Nx/Mx/1x устройств в режиме Mx/1x. Здесь каждое Nx/Mx/1x устройство осуществляет передачу данных с использованием двух дублирующих друг друга каналов.

Таким образом, в отдельных режимах Mx/1x может быть обеспечена совместимость с различными типами одно- и многоканальных архитектур, разработанными IP-блоками, контроллерами интерфейсов на основе открытых стандартов, таких, например, как RapidIO LP Serial [5]. Для описания разработанного алгоритма инициализации и контроля, рассмотрим соединение двух Nx/Mx/1x устройств. По умолчанию на обоих устройствах используются режимы Nx.

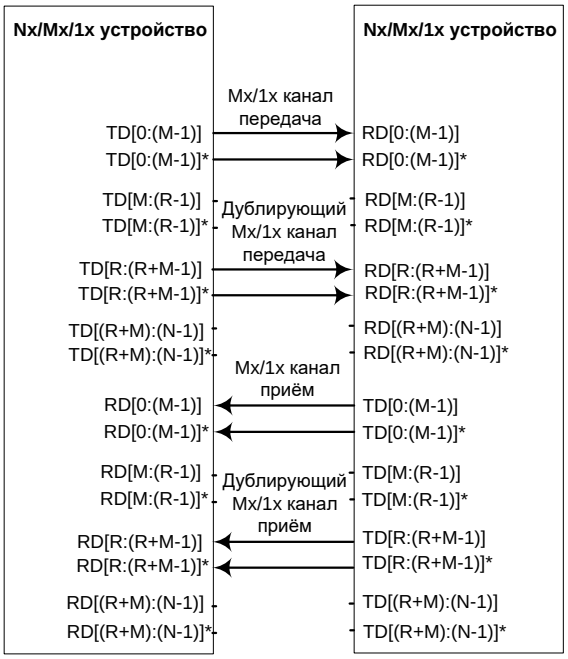

Рис. 2. Схема соединения двух Nx/Mx/1x устройств в режиме Mx/1x

При возникновении многократно повторяющихся ошибок или повреждении линий связи алгоритм инициализации и контроля автоматически перенастраивает ширину канала так, чтобы использовать два и более Mx или 1x каналов на передачу. Здесь, в качестве ошибок рассматриваются только ошибки приёмника, таких как несовпадение чётности, неверная кодировка, сбои в работе дешифратора, дескремблера, нарушение выравнивания Nx/Mx линий между собой, ошибки контрольных сумм в пакетах и контрольных символах. Максимальное число резервированных каналов для режима Mx определяется отношением $\frac{N}{M}$, а для режима 1x это число равно N. Приёмник выбирает один из резервированных наборов линий связи, обеспечивающий число ошибок равному заранее определённому параметром значению или ниже этого

порога. Этот режим с меньшей производительностью, но обеспечивающий большую надёжность. Для продолжения обмена данными достаточно, чтобы из всех Nx дифференциальных линий на приём и передачу работоспособными были бы хотя бы две (одна пара на приём, другая на передачу из установленного набора 0 или R). При этом используется механизм обмена контрольными символами и кодовыми группами. Контрольный символ может быть составлен из нескольких кодовых групп, включающих как полезную, так и служебную информацию. На рис. 3 показан пример передачи потока данных по каналам 1x и 4x.

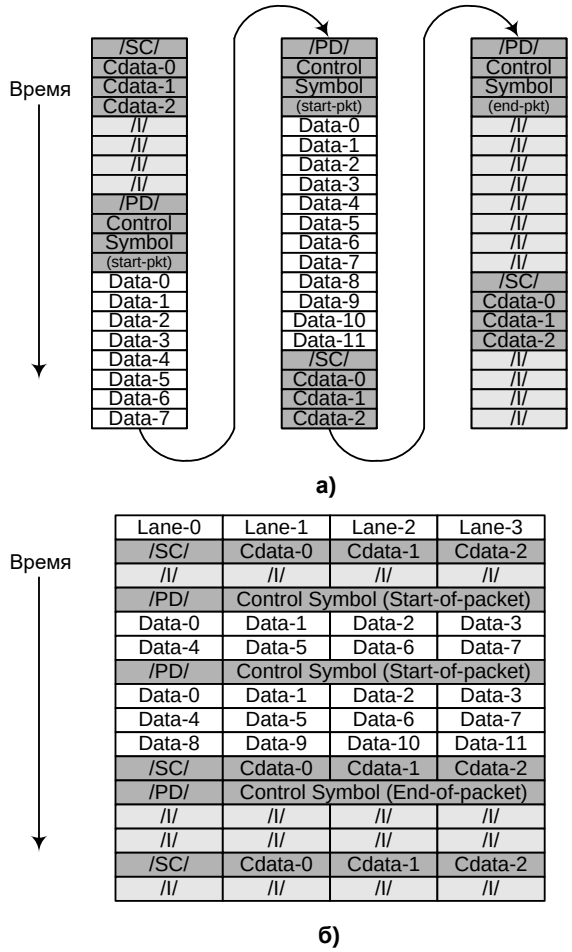

Рис. 3. Пример передачи потока данных: а) по каналу 1x, б) по каналу 4x

В режиме 1x данные последовательно передаются по одной из линий. Символом /SC/ отмечена кодовая группа, обозначающая начало контрольного символа, /PD/ - начало пакета. Символом /I/ отмечена синхронизационная (idle) последовательность данных, которая передаётся при начальной инициализации, а также в отсутствие пакетов и контрольных символов. Под синхронизацией подразумеваются процессы настройки приёма по каждой из линий, обеспечивающие распознавание входных кодовых групп и минимальное число ошибок. В режиме 4x пакеты данных и контрольные символы распределяются и передаются последовательно по всем четырём линиям, а синхронизационные кодовые группы /I/ передаются по всем линиям параллельно [6]. Такой способ передачи данных в Nx и Mx режимах обеспечивает возможность реализации схемы выравнивания линий между собой. Допустимая возможная задержка данных между линиями может достигать десятков наносекунд, что соответствует разности длин отдельных линий связи внутри единого канала порядка 1-10 метров.

III. Управляющий алгоритм

На рис. 4 представлена схема состояний и переходов управляющего алгоритма инициализации и контроля канального уровня. Для реализации описанного механизма обмена для каждого Nx/Mx/1x устройства необходим один управляющий набор логики (машина состояний) инициализации и контроля, а также служебные машины состояний синхронизации (отдельно для каждой из линий) и единый блок выравнивания линий. При сбросе устройство переходит в состояние ВСЁ_ОТКЛ, передача и приём данных не осуществляется, все счётчики и внутренние состояния сбрасываются. Переход из этого состояние в состояние ПОИСК происходит по событию внутреннего таймера. Длительность таймера составляет величину порядка сотен микросекунд. Этот интервал необходим, чтобы устройство на противоположной стороне, будучи в любом другом состоянии, получило ошибки данных, потеряло синхронизацию и также перешло в состояние ВСЁ_ОТКЛ. В состоянии ПОИСК выходной порт передаёт по линиям 0 и R последовательность данных, состоящую из кодовых групп, содержащих comma-символы для выравнивания данных внутри одной линии, skip-символы для обеспечения компенсации возможной разности частот приёмника и передатчика, а также alignment-символы для обеспечения выравнивания Nx или Mx линий между собой. В случае если устройству не требуется использовать только 1x режим, а также любая из линий 0 или R синхронизованы (выполнено выравнивание данных внутри линии и отсутствие ошибок кодировки), то устройство переходит в режим ОТКРЫТИЕ_0. Значение R выбирается исходя из требований по резервированию линий. Это может быть как одна линия (в случае простого дублирования), так и набор линий в случае многократного их резервирования. В состоянии ОТКРЫТИЕ_0 последовательность idle одновременно передаётся по всем линиям Nx/Mx, в этом же состоянии приёмник осуществляет выравнивание всех линий между собой, используя alignment-символы в составе idle. Максимальная длительность пребывания в этом состоянии ограничена таймером. Длительность этого интервала может составлять десятки миллисекунд. Такой временной интервал необходим, чтобы гарантировать завершение выравнивания линий. В случае успешного выравнивания линий Nx, в отсутствие ошибок и требований перехода в режим Mx, устройство

переходит в состояние Nx_РЕЖИМ. Это рабочий режим с максимальной производительностью. Устройство обеспечивает передачу и приём пакетов, контрольных символов по всем линиям, контролируя наличие ошибок приёма и выравнивание линий. В случае достижения таймером порогового значения, а также успешного выравнивания только набора линий Mx_0 или Mx_R, либо в случае, когда требуется использовать только режимы Mx, машина состояний переходит из состояния ОТКРЫТИЕ_0 в Mx_РЕЖИМ_0 или Mx_РЕЖИМ_R. Иначе, по истечению таймера в состоянии ОТКРЫТИЕ_0, если произошло несколько ошибок приёма данных по линиям, либо выравнивание между линиями Nx, Mx_0 или Mx_R не достигнуто, но при этом линии 1x_0, либо 1x_R синхронизованы, устройство переходит в состояние 1x_РЕЖИМ_0 или 1x_РЕЖИМ_R, соответственно.

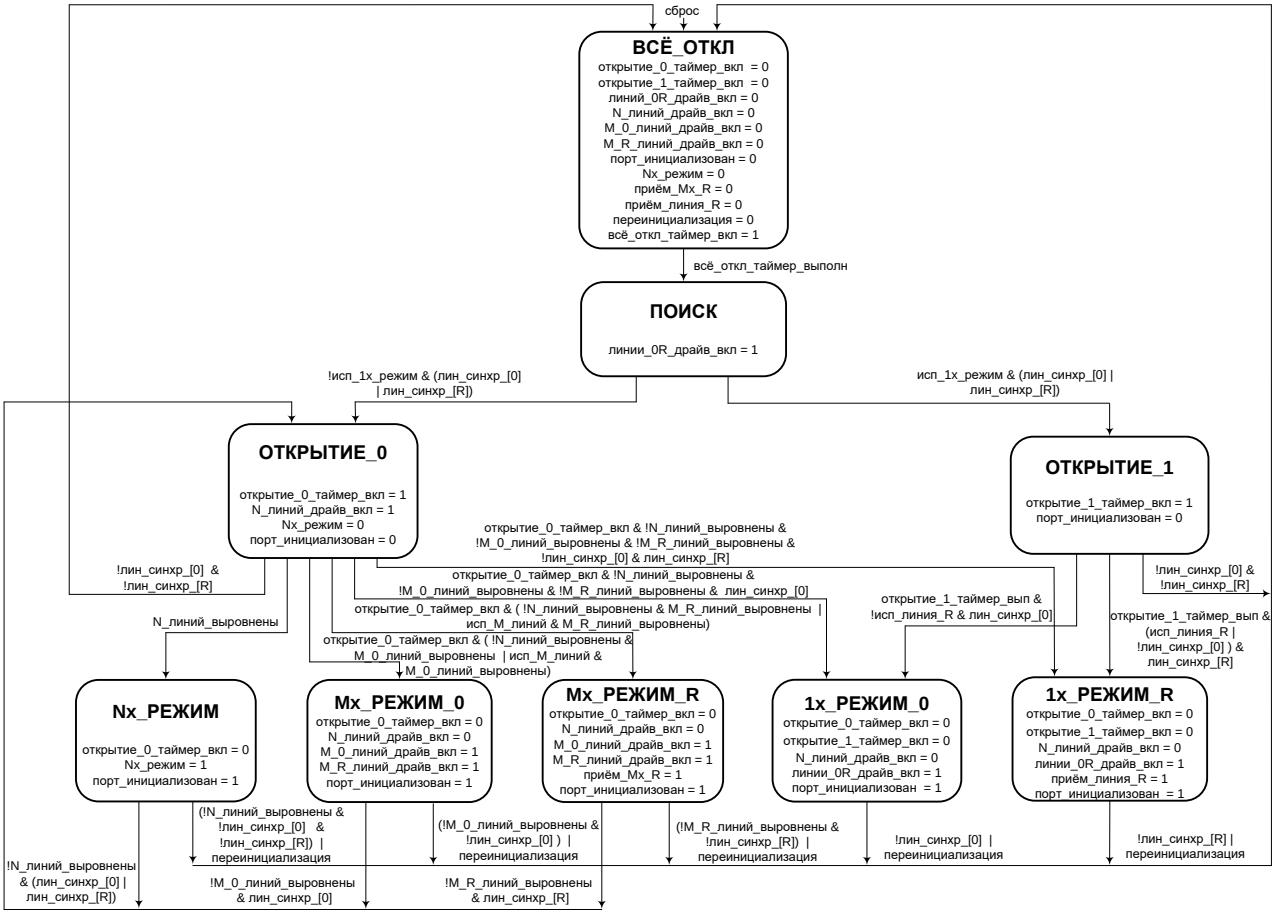

Рис. 4. Схема состояний и переходов управляющего алгоритма инициализации и контроля для устройства Nx/Mx/1x

В случае, если устройству требуется использовать только 1x режим, а также любая из линий 0 или R синхронизованы (выполнено выравнивание данных внутри линии и отсутствие ошибок кодировки), то устройство переходит в режим ОТКРЫТИЕ_1. В состоянии ОТКРЫТИЕ_1 idle последовательность одновременно передаётся только по линиям 0 и R. Максимальная длительность пребывания в этом состоянии также ограничена таймером. Определённый временной интервал необходим для гарантированного завершения синхронизации линий 0 и R. Число режимов 1x может варьироваться, максимальное значение определяется числом N. В случае достижения таймером порогового значения, если синхронизация достигнута линией 0, то выполняется переход в состояние 1x_РЕЖИМ_0, иначе, если синхронизация достигнута линией R, то выполняется переход в состояние 1x_РЕЖИМ_R. В режимах 1x_РЕЖИМ_0 и 1x_РЕЖИМ_R приём и передача данных осуществляется только по линиям 0 и R, соответственно. При повторной инициализации, либо если обнаружены ошибки любой из линий в составе прошедшего инициализацию Nx/Mx/1x канала, устройство возвращается в состояние ВСЁ_ОТКЛ. В случае если машина состояний и инициализации принимает значения Nx_РЕЖИМ, Mx_РЕЖИМ_0 или Mx_РЕЖИМ_R, то в случае нарушения выравнивания Nx, Mx_0 или Mx_R соответственно, устройство переходит обратно в состояние ОТКРЫТИЕ_0.

При анализе состояний и их переходов видно, что состояние Nx_РЕЖИМ обладает наивысшим приоритетом, далее следуют последовательно в порядке снижения приоритета состояния Mx_РЕЖИМ_0, Mx_РЕЖИМ_R, 1x_РЕЖИМ_0 и 1x_РЕЖИМ_R, где также происходит обмен пакетами и контрольными символами. В случае повреждения

линий связи, многократно повторяющихся ошибок, при отключении и повторном включении, при «горячем подключении» происходит автоматическая перенастройка. Все неподтверждённые и потерянные данные передаются повторно в соответствии с методами управления потоком данных на канальном и транспортном уровне.

IV. Оценка эффективности метода

Наивысшая пропускная способность канала достигается в режиме Nx, минимально необходимая в режиме 1x. Рассмотрим зависимость пропускной способности полностью резервированного канала связи от числа линий. Полностью резервированным каналом связи назовём такую Nx/Mx/1x конфигурацию, где R есть множество в интервале $\{1, \frac{N}{M}\}$ для конфигурации Mx, а для режима 1x R равно N. По оси абсцисс отложим принятую за 100% пропускную способность Nx линка (link), а по оси ординат число ошибочных или отключённых линий. На рис. 5 представлены графики зависимостей приведённой пропускной способности от числа неисправных линий для конфигураций 8x/4x/1x и 8x/2x/1x.

Для конфигурации 8x/4x/1x при выходе из строя любой из линий пропускная способность канала падает в два раза. В наихудшем случае, когда выходят из строя 2 линии, по одной линии в каждом 4x наборе, включается режим 1x, пропускная способность падает в 8 раз. Наилучший случай отличается тем, что очередная по счёту неисправная линия находится в том же 4x наборе линий. В обоих случаях, когда работоспособна всего одна линия из восьми, осуществляется приём и передача данных в состояниях 1x_РЕЖИМ_0 или 1x_РЕЖИМ_R. Для конфигурации 8x/2x/1x при выходе из строя любой из линий пропускная способность канала падает в четыре раза. В наихудшем случае, когда выходят из строя 4 линии, по одной линии в каждом 2x наборе, включается режим 1x, пропускная способность падает в восемь раз. Полученные характеристики наглядно показывают, что работоспособность линка сохраняется до тех пор, пока работоспособны всего по одной линии в каждом направлении из N дуплексных линий, составляющих полный канал. Главный недостаток рассмотренного метода в том, что при выходе из строя всего одной линии пропускная способность снижается в два и более раз. На графиках это проявляется в сильном отклонении полученной кривой от идеальной характеристики. Идеальная характеристика представляет собой прямую, соединяющую точку при отсутствии ошибок линий и максимальной пропускной способностью в режиме Nx и точку с нулевой пропускной способностью при полностью неисправном канале. Чем ближе полученная характеристика надёжности к идеальной, тем эффективнее канал использует возможности каждой сигнальной пары.

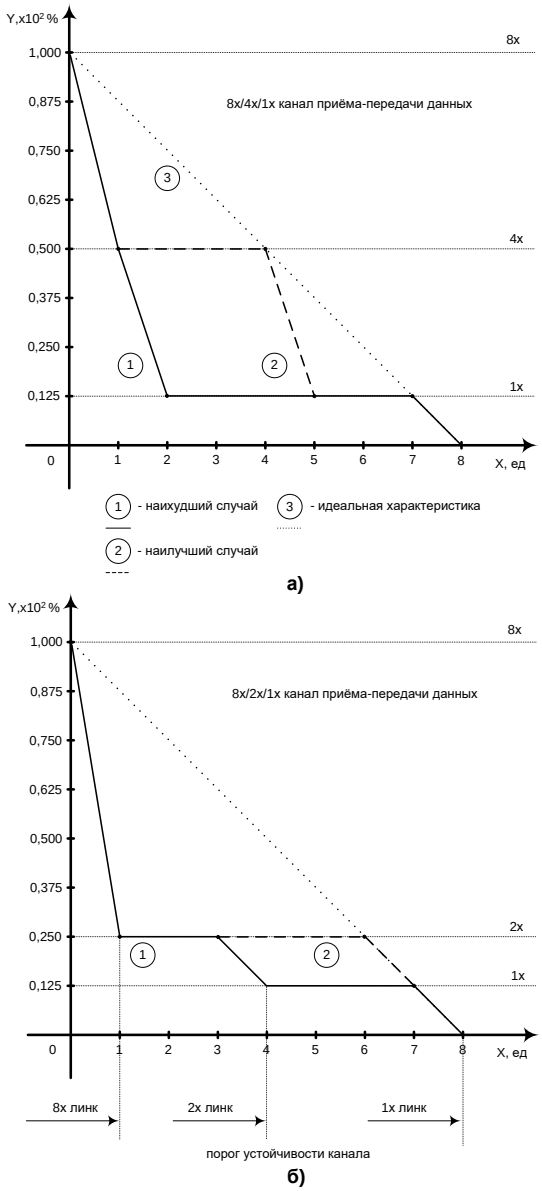

Рис. 5. График зависимости приведённой пропускной способности канала от числа неисправных линий: а) для 8x/4x/1x, б) для 8x/2x/1x

Одним из наиболее эффективных методов решения данного ограничения является объединение на канальном уровне двух и более портов транспортного уровня, соединённых с коммутационным ядром контроллера. Объединенный мультиплексированный канал связи или кросслинк (crosslink) содержит две и более машины состояния инициализации по числу портов (одна машина состояния на каждый порт) и одну общую машину состояний выбора режима для формирования предварительных установок и коммутирования линий на физическом уровне [7]. При мультиплексировании, к примеру, двух портов преимуществом является то, что помимо отдельных настроек Nxa/Nxb/Mxa/Mxb/1xa/1xb, данный метод позволяет задействовать комбинированные режимы

(Mxa+1xb)/(Mxb+1xa)/(1xa+1xb). На рис. 6 приведены графики зависимостей приведённой пропускной способности мультиплексированного канала связи от числа неисправных линий для конфигураций 8x/4x/1x и 8x/2x/1x. При сравнении графиков на рис. 5 и рис. 6 видно, что мультиплексированный канал связи эффективнее использует пропускную способность дуплексного канала, поскольку при ошибках машины состояний позволяют перевести систему в один из комбинированных режимов.

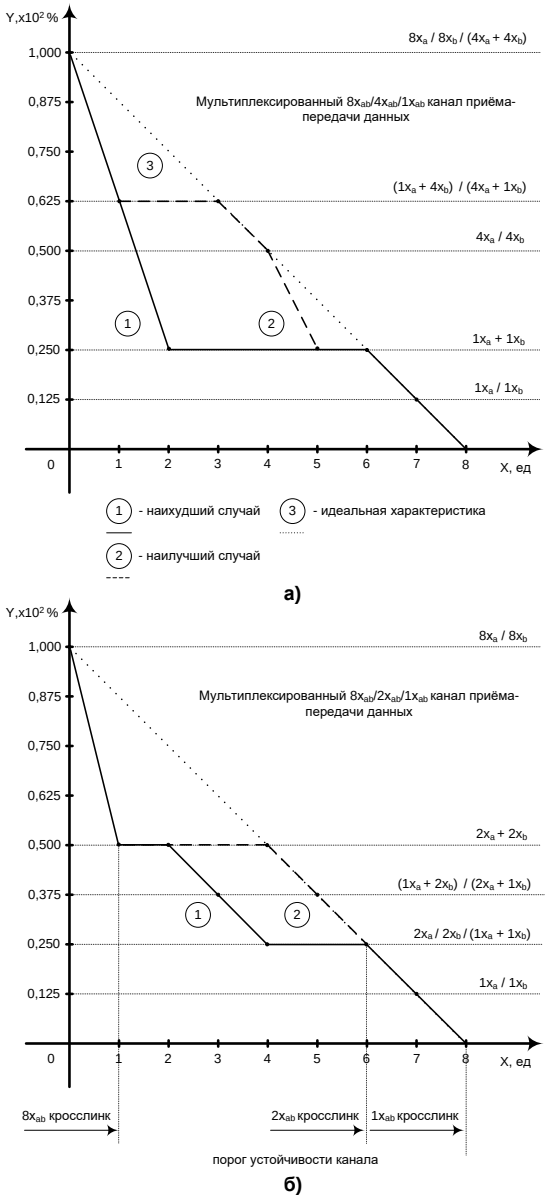

Рис. 6. График зависимости приведённой пропускной способности мультиплексированного канала от числа неисправных линий:
а) для $8x_{ab}/4x_{ab}/1x_{ab}$, б) для $8x_{ab}/2x_{ab}/1x_{ab}$

При использовании полностью резервированного мультиплексированного 8x/4x/1x канала при выходе из строя одной из линий пропускная способность падает всего на 37,5% и составляет 62,5% от максимальной величины. Для конфигурации 8x/2x/1x это значение равно 50%. Прирост суммарной пропускной способности справедлив для двух представленных конфигураций для всех значений на оси абсцисс.

На представленных графиках настройки канального уровня предполагают использование одного 8x канала порта a или порта b или двух 4x каналов порта a и порта b. В случае нарушения целостности канала или при смене конфигурации машина состояния выбора режима перенастраивает коммутационную логику так, чтобы задействовать дополнительные режимы. При выходе из строя одной или нескольких линий, включаются комбинированные режимы с использованием 4x, 2x или 1x дуплексных линий порта a или порта b. Отметим, что в этом случае именно конфигурация 8x/2x/1x предпочтительнее, поскольку в наихудшем случае обеспечивает более высокий темп приёма-передачи данных при выходе из строя двух и более линий связи.

Рассмотрим основные показатели надёжности для различных конфигураций физического канала связи. Для расчета вероятностей безотказной работы P и возникновения критической ошибки системы Q воспользуемся соотношениями общего логико-вероятностного метода [8]. Так, канал Nx можно представить как систему с последовательной структурой, где отказ любой из дуплексных линий приводит к отказу канала в целом. Для упрощения расчётов примем за единицу вероятность безотказной работы логической схемы блока коммутации, машин состояния, наборов контрольной логики и прочих компонентов системы. Для такого подхода справедливы следующие соотношения:

$$P_s = \prod_{i=1}^{N}(1-2\times q_i) \quad (7),$$

$$Q_s = \sum_{i=1}^{N} 2\times q_i \quad (8),$$

где qi – вероятность ошибки линии связи в каждом направлении. Конфигурации Nx/Mx, Nx/Mx/1x и Nxab/Mxab/1xab можно рассматривать как системы смешанной параллельно-последовательной структуры. При этом для представленных вариантов справедливы следующие уравнения:

$$P_p = 1-\prod_{i=1}^{\frac{N}{M}}(1-\prod_{j=1}^{M}(1-2\times q_{ij})) \quad (9),$$

$$Q_p = \prod_{i=1}^{\frac{N}{M}}(\sum_{j=1}^{M} 2\times q_{ij}) \quad (10),$$

где qij – вероятность ошибки линии связи в каждом направлении.

При расчёте показателей P и Q воспользуемся соотношениями (7) – (10). Вероятность возникновения ошибки линии связи в каждом направлении q (отдельно на приём и передачу) оценим как величину порядка 10^{-5}, что соответствует значению, полученному при моделировании работы портов микросхемы коммутатора, работающего по стандарту RapidIO [9]. В табл. 1 представлены результаты расчётов для конфигураций 8x, 8x/4x, 8x/2x и 8x/1x.

Таблица 1

Значения показателей надёжности для конфигураций канала связи

Конфигурация канала связи	Показатели надёжности	
	P	Q
8x	0,9998400111	1,60*10-4
8x/4x	0,9999999936	6,40*10-9
8x/2x	> 0,9999999999	2,56*10-18
8x/1x	> 0,9999999999	2,56*10-38

Значения показателей надёжности рассмотренных конфигураций Nx/Mx/1x(Nx_{ab}/Mx_{ab}/$1x_{ab}$) соответствуют величинам, рассчитанным для канала 8x/1x, поскольку состояние Mx(Mx_{ab}) рассматривается как промежуточное между Nx(Nx_{ab}) и 1x($1x_{ab}$), а вероятность безотказной работы внутренних логических схем принята нами ранее за единицу.

V. Заключение

Таким образом, применение описанных методов обеспечивает:

1). Реализацию канального регламентирующего протокола инициализации и контроля полностью дуплексного канала связи с Nx/Mx/1x сигнальными парами, автоматической начальной и повторной инициализацией в случае многократно повторяющихся ошибок.

2). Возможность реализации на основе предложенного алгоритма схемы мультиплексирования двух и более внутренних портов транспортного уровня с возможностью реконфигурации в процессе работы без потери полезных данных.

3). Обеспечение наивысшей пропускной способности канала, определяемой изначально заданным числом сигнальных пар в отсутствие повторяющихся ошибок линий связи.

4). Обеспечение повышенной надёжности канала приёма-передачи данных путём многократного резервирования передающих линий связи при сохранении необходимой пропускной способности в случае многократно повторяющихся ошибок одной или нескольких линий связи.

К недостаткам данной методики можно отнести:

1). Отсутствие открытого стандарта.

2). Увеличение числа логических элементов в системе.

3). Повышение задержек при приёме и передаче данных.

4). Уменьшение числа доступных портов внутреннего коммутационного ядра при использовании мультиплексированных каналов.

Описанные подходы не ограничены использованием исключительно в специально разработанных блоках под конкретный тип задач. Представленный метод может быть использован при разработке как блоков высокопроизводительных интерфейсов с открытыми стандартами [10], так и отдельных IP-блоков для сетевых устройств, не находясь в противоречии с их спецификациями и расширяя возможности при реализации их требований.

Литература

[1] Top500 Supercomputer Sites. URL: http://www.top500.org/statistics/sublist/ (дата обращения: 30.03.2016).

[2] Бобков С.Г. Высокопроизводительные микропроцессоры для суперЭВМ эксафлопсного диапазона // Электроника, микро- и наноэлектроника: сб. науч. тр., 14-я Рос. науч.-технич. конф. (2012, г. Суздаль); [под ред. В.Я. Стенина]. М.: Изд-во МИФИ, 2012. С. 129–141.

[3] Задябин С.О., Косарев И.М. Архитектура перспективного коммутатора высокоскоростных каналов ввода-вывода с изменяемой шириной // Электроника, микро- и наноэлектроника: сб. науч. тр., 12-я Рос. науч.-технич. конф. (2010, г. Суздаль); [под ред. В.Я. Стенина]. М.: Изд-во МИФИ, 2010. С. 130–134.

[4] ГОСТ Р ИСО/МЭК 7498-1-99. Информационная технология. Взаимосвязь открытых систем. Базовая эталонная модель. Ч. 1. Базовая модель. Москва, Стандартинформ, 2006. С. 45–51.

[5] RapidIO Interconnect Specification. Part 6: LP-Serial Physical Layer Specification. Revision 2.2. URL: http://www.rapidio.org/specs/current/RapidIO_Rev_2.2_Specification. zip (дата обращения: 30.03.2016).

[6] Sanjeeb M., Neeraj K.S., Vijayakrishnan R. System on Chip Interfaces for Low Power Design. Morgan Kaufman Publ., 2015, pp. 114–116.

[7] PCI Express Base Specification. Revision 3.0. URL: http://pcisig.com/specifications/pciexpress/PCI_Express_Base_r3.0_10Nov10.pdf (дата обращения: 30.03.2016).

[8] Рябинин И.А., Черкесов Г.Н. Логико-вероятностные методы исследования надёжности структурно-сложных систем. М.: Радио и связь, 1981. С. 19–28.

[9] Задябин С.О., Косарев И.М. Методы увеличения пропускной способности каналов последовательных интерфейсов // Электроника, микро- и наноэлектроника: сб. науч. тр., 10-я Рос. науч.-технич. конф. (2008, г. Петрозаводск); [под ред. В.Я. Стенина]. М.: Изд-во МИФИ, 2008. С. 85–88.

[10] Козлов Н.А., Бобков С.Г. Высокопроизводительный блок интерфейса RapidIO для создания многоядерных микропроцессоров с виртуальными каналами RapidIO // Программные продукты и системы. 2015. №4 (112). С. 88–92.

Distributed multi lane serial links for multiprocessor systems interconnects

I.M. Kosarev

Scientific Research Institute of System Analysis (SRISA RAS), ikosarev@cs.niisi.ras.ru

***Keywords* — lane backup, high-speed serial bus, training state machine, hot plug, crosslink.**

ABSTRACT

The technique of port multiplexing and overlapping of multi lane data channel is widely used to build multi-configuration interconnects and robust data links for parallel computing environments. The main disadvantage of this approach is that the total channel bandwidth is reduced proportionally to the number of duplicated links. This article describes a common approach for building data link layer protocol including link initialization and training algorithm for reliable full-duplex serial interconnects having multiple Nx/Mx/1x data lanes providing port multiplexing and flexible configuration of physical channel with different link widths. Operational aspects of the training algorithm provide high data rate for the data channel having a plurality of physical pipelines in absence of multiple bit lane and link errors and split up port on two or more different serial links for higher reliability in case of multiple bit data errors or link failures.

REFERENCES

[1] Top500 Supercomputer Sites. Available at: http://www.top500.org/statistics/sublist/ (accessed: 30.03.2016).

[2] Bobkov S.G. High performance microprocessors for exaflop computing systems. *Sbornik nauchnykh trudov 14 Ross. nauch.-tekhnich. konf. Elektronika, micro i nanoelectronika* [Proc. of the 14th Russian Science and Tech. Conf. Electronics, micro- and nanoelectronics]. Moscow, MEPHI Publ., 2012 pp. 129-141 (in Russian).

[3] Zadyabin S.O., Kosarev I.M. Architecture of promising multi port switch with high speed serial links with variable width. *Sbornik nauchnykh trudov 12 Ross. nauch.-tekhnich. konf. Elektronika, micro i nanoelectronika* [Proc. of the 12th Russian Science and Tech. Conf. Electronics, micro- and nanoelectronics]. Moscow, MEPHI Publ., 2010, pp. 130-134 (in Russian).

[4] GOST R ISO/IEC 7498-1-99. Information technology. Open systems interconnection. Basic reference model. Part 1: The basic model. Moscow, Standartinform, Publ., 2006. pp. 45-51 (In Russian).

[5] RapidIO Interconnect Spec. Part 6: LP-Serial Physical Layer Specification. Revision 2.2. Available at: http://www.rapidio.org/specs/current/RapidIO_Rev_2.2_Specification (accessed: 30.03.2016).

[6] Sanjeeb M., Neeraj K.S., Vijayakrishnan R. System on Chip Interfaces for Low Power Design. Morgan Kaufman Publ., 2015, pp. 114-116.

[7] PCI Express Base Spec. Revision 3.0. Available at: http://pcisig.com/specifications/pciexpress/PCI_Express_Base_r3.0_10Nov10.pdf (accessed: 30.03.2016).

[8] Ryabinin I.A., Cherkesov G.N. The logic-probabilistic research methods of structure-complex systems reliability. Moscow, Radio and communication Publ., 1981, pp. 19-28 (in Russian).

[9] Zadyabin S.O., Kosarev I.M. Methods of increasing bandwidth of serial data channel interconnects. *Sbornik nauchnykh trudov 10 Ross. nauch.-tekhnich. konf. Elektronika, micro i nanoelectronika* [Proc. of the 10th Russian Science and Tech. Conf. Electronics, micro- and nanoelectronics]. Moscow, MEPHI Publ., 2008, pp. 85-88 (in Russian).

[10] Kozlov N.A., Bobkov S. G. High performance RapidIO block to create multi-core microprocessors with RapiIO virtual links. *Programmnye produkty i sistemy*. [Software & Systems]. 2015, no. 4 (112), pp. 88-92 (in Russian).