

Использование PCI-Express интерфейса в системах хранения данных

Общая информация о PCI-Express интерфейсе

Для начала необходимо кратко остановиться на основных характеристиках компьютерной шины PCI Express (довольно часто его название сокращают до PCI-E или даже PCIe). Разработка этого интерфейса передачи данных третьего поколения (также иногда обозначаемого как **3GIO** - *3rd Generation I/O*) была начата фирмой Intel после отказа от шины InfiniBand. Его дальнейшее развитие было поручено специальной организации PCI Special Interest Group (PCI-SIG, www.pcisig.com). Первая базовая спецификация PCI-E стандарта появилась в июле 2002 года, в январе 2007 года была выпущена спецификация 2.0, а сейчас идет работа уже над PCI-Express 3.0, утверждение которой предполагается в 2009.

Напомним, что еще сравнительно недавно основным стандартом компьютерной шины был PCI (Peripheral Component Interconnect). Он начал свое широкое распространение с утверждения в 1993 году спецификации 2.0, постепенно вытеснив из обращения ISA. Основные параметры PCI 2.0 - это, конечно, ширина шины, равная 32 битам, соответственно максимальное адресное пространство 4 Гбайт, тактовая частота шины 33 МГц при синхронном обмене данными и результирующая пиковая пропускная способность 133 Мбайт/с. Что еще важно, устройства были рассчитаны на напряжение питания 5 В и 3,3 В. В далеком 1993 году такие величины казались поистине громадными. Далее появился стандарт PCI 2.2, в соответствии с которым ширина шины могла быть увеличена до 64 бит, а тактовая частота до 66 МГц. Еще дальше пошла PCI-X (eXtended) - это ускоренная до 133 МГц шина PCI 2.2 с обязательной 64-битной разрядностью интерфейса. PCI-X допускалась и в более форсированных вариантах, с 266 МГц и 533 МГц тактовой частотой. Как ответвление от линейного развития PCI специально для высокопроизводительных графических адаптеров в 1997 году была создана шина AGP (*Accelerated Graphics Port*). Известны четыре ее вариации, различающиеся пиковой пропускной способностью - AGP 1x, 2x, 4x и 8x. Таким образом, за более чем десятилетнее развитие (а это целый век по компьютерным меркам) была разработано несколько значимых модификаций PCI, существенно продвигающих возможности стандарта. Пиком скоростного развития стандарта PCI можно назвать PCI-X 533 и AGP 8x. Однако всему есть предел, все возможности по наращиванию пропускной способности PCI оказались исчерпанными, дальнейшее ее увеличение стало технологически слишком сложным и дорогим. Потребовалась новая разработка, гарантирующая «безболезненную» масштабируемость и наращиваемость на ближайшие годы и в то же время сохраняющая преемственность с уже имеющимися устройствами. Одним словом, пришло время шины PCI Express, использующей программную модель PCI, но основанной на более высокопроизводительном физическом протоколе.

PCI Express, подобно Infiniband и SAS, использует двунаправленную последовательную шину. У последовательного интерфейса имеется много преимуществ. Во-первых, подключение «точка-точка», исключая арбитраж шины и использование прерываний, когда все устройства подключаются к общей 32/64-разрядной параллельной шине. Во-вторых, нет необходимости в громоздкой синхронизации сигналов. Ведь при параллельной организации передачи биты приходят от источника к приемнику широким «строем», и обработку приходится делать всех разом строго по командам «делай раз, делай два». Что, кстати, с увеличением тактовых частот PCI становилось все более сложным. Наконец, немаловажно упрощение схмотехники и миниатюризация соединений. Последовательный протокол требует относительно небольшого числа проводников и потому допускает намного более высокие тактовые частоты, что и даёт более высокую пропускную способность. Дальнейшее увеличение пропускной способности достигается «связкой» нескольких подобных линий.

PCI Express построен на принципах симплексной технологии, а это означает, что сигналы идут одновременно в противоположных направлениях и по отдельным парам проводов. Итого принципиально требуется всего две пары проводников, вместе называемые **lane** (линия). Соединение между двумя устройствами PCI Express называется **link** (связь, соединение), и состоит из одного (называемого **x1**) или нескольких (**x2**, **x4**, **x8**, **x12**, **x16** и **x32**) двунаправленных последовательных соединений **lane**. При этом каждое устройство должно поддерживать соединение **x1**.

Использование подобного подхода имеет следующие преимущества:

- карта PCI Express помещается и корректно работает в любом слоте той же или большей пропускной способности (например, карта x1 будет работать в слотах x4 и x16);

- слот большего физического размера может использовать не все **lane**'ы (например, к слоту x16 можно подвести линии передачи информации, соответствующие x1 или x8, и всё это будет нормально функционировать; однако, при этом необходимо подключить все линии «питание» и «земля», необходимые для слота x16);
- горячая замена карт;
- гарантированная полоса пропускания;
- управление энергопотреблением;
- контроль целостности передаваемых данных.

На шине PCI Express всегда задействовано максимальное количество **lane**'ов, доступных как для карты, так и для слота. Однако это не позволяет устройству работать в слоте, предназначенном для карт с меньшей пропускной способностью шины PCI Express (например, карта x4 физически не поместится в слот x1, несмотря на то, что она могла бы работать в слоте x4 с использованием только одного **lane**).

PCI Express пересылает всю управляющую информацию, включая прерывания, через те же линии, что используются для передачи данных. Последовательный протокол никогда не может быть заблокирован. Во всех высокоскоростных последовательных протоколах (например, Ethernet), информация о синхронизации должна быть встроена в передаваемый сигнал. На физическом уровне, PCI Express использует ставший общепринятым метод кодирования 8B/10B (8 бит данных заменяются на 10 бит, передаваемых по каналу, таким образом 20% передаваемого по каналу трафика является избыточными), что позволяет поднять помехозащищённость.

Пропускная способность PCI-Express

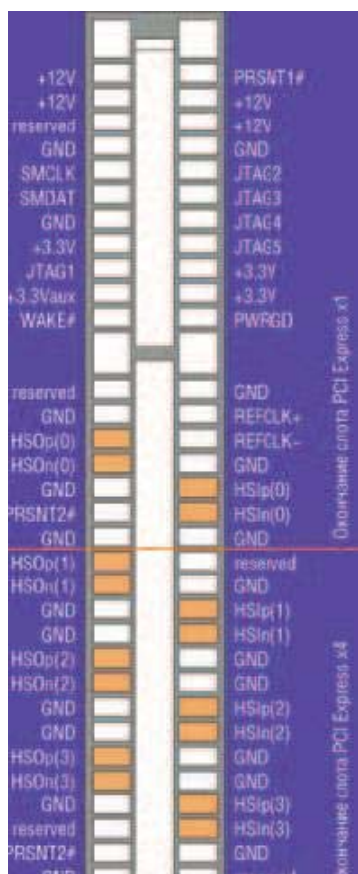
Согласно стандарту 1.0 пропускная способность симплексной линии составляет 2,5 Гбит/сек в одну сторону. Для расчета пропускной способности соединения **link** необходимо учесть то, что в каждом соединении

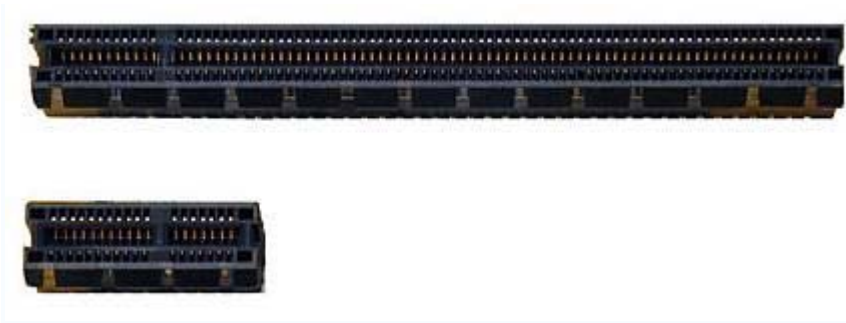
передача дуплексная, а также учесть применение кодирования **8B/10B** (8 бит в 10). Таким образом, полная (дуплексная) пропускная способность соединения **x1** составляет $2,5 \times 2 \times 0,8 / 8 = 0,5$ Гбайт/сек. Понятно, что эффективная скорость передачи информации в одном направлении (только на запись или только на чтение) соответственно в 2 раза ниже.

При построении связи из нескольких линий, которые приходится выстраивать в ряд (допускается **link** из 2, 4, 8, 12, 16 и 32 **lane**) вся последовательность передаваемых данных распределяется на имеющиеся линии «веером» - передача опять параллельная, но все же не синхронная. Если имеется 16 линий, то первый байт блока данных передается по первой линии, второй - по второй, и т. д., а семнадцатый байт - снова по первой. Соответствующая пропускная способность возрастает строго пропорционально, так что для x4 максимальная скорость передачи уже 2 Гбайт/сек, а шина с 32 линиями (x32) способна выдать пропускную способность в 16 Гбайт/сек, или же по 8 Гбайт/сек в каждую сторону.

Как уже упоминалось, для каждой линии (**lane**) теоретически необходимо только 2 пары проводников. На практике их конечно заметно больше. На каждую сигнальную пару (на рисунке выделены **желтым**) в разъемах PCI Express приходится по две "земли" (на рисунке GND), экранирующих данную линию. За счет этого контактов в разъеме становится больше, хотя если сопоставить число контактов у шин PCI Express и соответствующих им по быстродействию параллельных шин, выигрыш будет очевидным. Добавление еще одной линии сводится к добавлению еще 2-х пар проводников и 2-х пар земли.

Помимо сигналов, естественно, разводится питание +12 В и +3,3 В (уровень +5 В теперь отсутствует), линия "дежурного" питания (тоже 3,3 В) и некоторые специальные линии - SMBus, JTag. Точно так же, как линии приема и передачи, разводятся дифференциальные линии передачи тактового сигнала REFCLK (два проводника и две "земли", хотя этот сигнал не используется при передаче данных). Для реализации горячего подключения и определения типа установленной карты разводятся линии обнаружения карты PRSNT; для реализации "спящего режима" - линия "пробуждения" WAKE. Все они вынесены в отдельную группу контактов. Ниже приведены фотографии двух популярных разъемов PCI Express x16 и x1.





А в таблице сведены данные об общем числе контактов в слотах различных интерфейсов, также как и максимальных значениях полосы пропускания (суммарно в обоих направлениях).

| Тип интерфейса | Число контактов в разъеме | Теоретическая полоса пропускания, Мбайт/сек |
|-------------------------|---------------------------|---|
| PCI (32 бит, 33 МГц) | 120 | 133 |
| PCI-X (64 бит, 133 МГц) | 184 | 1064 |
| AGP 8x | 124 | 2133 |
| PCI Express x1 | 36 | 500 |
| PCI Express x4 | 64 | 2000 |
| PCI Express x8 | 98 | 4000 |
| PCI Express x16 | 164 | 8000 |
| PCI Express x32 | 294 | 16000 |

Данные скорости справедливы для спецификации 1.1. Однако у PCI Express 2.0, утвержденной в январе 2007, появились нововведения, среди которых наиболее важным является удвоение пропускной способности: одна линия **lane** в каждом из направлений пропускает до 5 Гбит/сек (до 10 Гбит/сек в дуплексном режиме). При этом сохранена совместимость с PCI Express 1.1, так что плата расширения, поддерживающая стандарт PCI-E 1.1 может работать, будучи установленной в слот PCI-E 2.0. Кроме того, внесены некоторые усовершенствования в протокол передачи между устройствами и программную модель.

Внешние PCI Express системы

Логическим дополнением базовой спецификации 2.0 стало утверждение месяцем позже спецификации внешней кабельной системы PCI Express External Cabling 1.0. Регламентируются конструкция и маркировка соединительных разъемов, механические и электрические характеристики медных кабелей. Предусматривается объединение линий шины PCI Express в группы по 1, 4, 8 и 16 (PCI-E x1, x4, x8, x16).



Чем больше линий, тем соответственно больше контактов (18 для x1, 38 для x4, 68 для x8 и 136 для x16), тем сложнее и шире соответствующие разъемы и кабели (см. рисунок). Новая спецификация совместима с PCI-E 1.1 и поддерживает максимальную производительность 2,5 Гбит/сек на линию, что, например, должно обеспечить скорость передачи данных 4 Гбайт/сек по кабелю x8 (суммарно в обе стороны). При этом длина кабеля не должна превышать 10 метров.

Подобный подход позволяет вынести составные части персонального компьютера (например, графическую подсистему) за пределы системного блока, построить пространственно «разнесенную» вычислительную среду. Наиболее продуктивно эта идея стала использоваться для создания внешних систем хранения, видимых операционной системой в качестве обычных локальных устройств, но физически построенных в виде самостоятельных блоков, в отдельном корпусе с собственным процессором, питанием и охлаждением. И при этом находящихся на расстоянии до десяти метров от рабочего компьютера.

Напомним, что классическая схема подключения внешней системы хранения к компьютеру (хосту) требует сначала преобразования данных для передачи по внешнему интерфейсу (выполняется

специальным компьютерным контроллером, НВА – Host Bus Adapter), а затем обратного преобразования при приеме и последующей передаче данных уже в системе хранения. Использование контроллера внешнего интерфейса, безусловно, удобнее с точки зрения стандартизации и мобильности системы хранения, но есть и много минусов. Во-первых, за контроллеры надо платить и немало, от \$250 (SAS) до \$1000 (Fibre Channel). К этому надо добавить соединительные кабели, стоящие в зависимости от длины до \$100. А во-вторых, предел скорости обмена данными в системе хранения определяется именно контроллерами и для SAS/Fibre Channel не превышает 300/400 Мбайт/сек. Поскольку в современных RAID контроллерах, являющихся ядром любой внешней системы хранения, для передачи данных между дисками и процессором используется PCI-E шина, то подключение подобных систем к компьютеру по PCI-E позволяет устранить пару лишних преобразований информации и, как следствие, существенно увеличить результирующую скорость потока данных.

Первопроходцем, на практике реализовавшим идею использования PCI-Express в качестве внешнего интерфейса, стала хорошо известная на российском рынке компания Maxtronic International (www.maxtronic.com.tw, www.maxtronic.ru). Уже осенью 2007 года она выпустила 8-дисковые RAID системы SA-3378S и SA-4378S соответственно в настольном и стоечном исполнении. Чуть позднее эта идея была подхвачена и многими другими производителями систем хранения.

Сегодня продуктовая линейка Maxtronic включает не только 8-ми, но и 12-ти, 16-ти и даже 24-х дисковые устройства с PCI-E интерфейсом, причем как одно- так и двухконтроллерные. В качестве примера можно привести SA-8808D с двумя независимыми контроллерами на базе современного



процессора Intel i80341. Каждый из контроллеров управляет обработкой данных с 12 дисков и обеспечивает скорость передачи в канале PCIe x4 до 750 Мбайт/сек. При прямом подключении данной системы к компьютеру через соответствующий 2-х канальный адаптер эффективный поток данных достигает уже 1,4 Гбайт/сек. При этом необходимо

учитывать, что общая стоимость SA-8808D заметно ниже аналогичных систем на базе SAS или FC интерфейсов. Это стало веским основанием для компании Maxtronic выделить этот класс решений особым названием **ExaRAID**. Более того, не останавливаясь на достигнутом, компания стала успешно развивать данную технологию для построения SAN.

SAN на базе PCI-Express

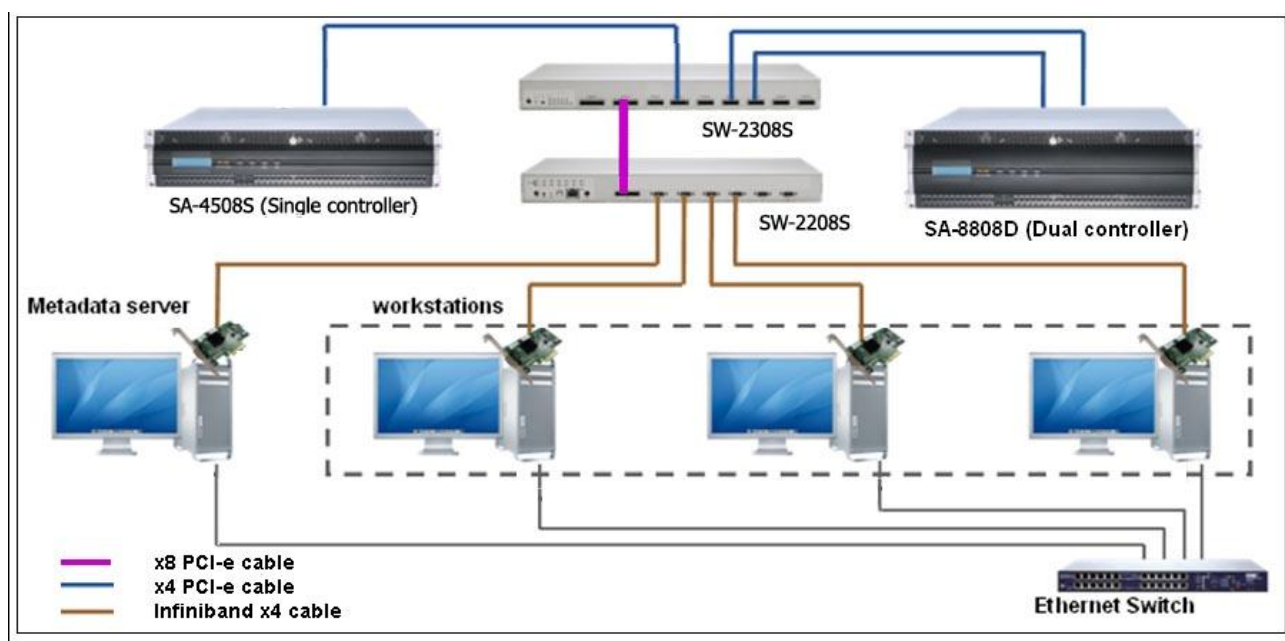
Для корпоративных систем хранения большой емкости (в сотни терабайт) и высокого быстродействия с возможностью высокоскоростного доступа к общим данным одновременно многим пользователям отдельные устройства хранения объединяют в самостоятельные сети, т.е. SAN (Storage Area Network). Таким образом SAN – это логическое и физическое объединение через специальные коммутаторы нескольких RAID массивов в единую систему с возможностью доступа с любого сервера (рабочей станции) к любому из подключенных RAID массивов. Де-факто, в настоящее время самый популярный интерфейс для построения SAN – это Fibre Channel (FC). Он имеет массу преимуществ, о которых уже много раз рассказывалось. Достаточно напомнить, что он обеспечивает последовательный канал передачи данных с возможностью коммутации и маршрутизации потоков и работой на больших расстояниях (до 300 метров на оптических многомодовых кабелях, до 10 километров на одномодовом кабеле). Но у него есть один существенный недостаток - высокая стоимость реализации. Например, FC контроллер, который должен быть установлен в каждый из компьютеров, подключенных к SAN, стоит около \$1000. Да и FC коммутаторы далеко не дешевы (от \$2500).

Заманчивую идею использования интерфейса PCI-E и для построения SAN сдерживало невозможность коммутации по PCI-E, а также отсутствие длинных (хотя бы до 100 метров) кабелей. Даже в небольших видеостудиях и телецентрах крайне важно "оттащить" шумную

систему хранения от рабочего компьютера, ведь работать среди шума некомфортно, а заниматься творчеством просто невозможно. Компания Maxtronic International первой преодолела эти ограничения и весной 2009 года представила принципиально новое решение, названное **ExaSAN**.

Реализация ExaSAN основана на использовании двух специальных свитчей (коммутаторов), а именно **Storage Switch SW-2308S** для подключения систем хранения и **Host Switch SW-2208S** для подключения компьютеров (хостов). SW-2308S оснащен семью портами x4 для массивов, одним портом PCI-E x8 для каскадирования (соединение в цепочку до четырех коммутаторов SW-2308S) и еще одним портом PCI-E x8 для соединения с коммутатором хостов SW-2208S. У SW-2208S есть 6 портов x4 PCI-E соответственно для 6 компьютеров. Если требуется подключить более 6 компьютеров, то можно использовать **Host Switch SW-2508S** уже на 12 портов. Таким образом, данное решение допускает объединение в единую рабочую среду по PCI-Express интерфейсу до 28 систем хранения и до 12 компьютеров. В дальнейшем будут разработаны и другие устройства, поддерживающие инфраструктуру ExaSAN.

Принципиальная структура ExaSAN показана на рисунке ниже.



Если все описать словами, то в компьютеры устанавливаются NT Card (Non-transparent card), по сути представляющие из себя простые Infiniband контроллеры (их цена в пределах \$200). NT Card соединяется с коммутатором хостов Infiniband оптическим кабелем длиной до 100 метров. Коммутаторы хостов и систем хранения соединяется между собой PCI-E x8 кабелем. Системы хранения подключаются к своему коммутатору с помощью своих штатных PCI-E x4 кабелей. Собственно, все.

Поскольку ExaSAN является классической SAN системой, то программная поддержка SAN может осуществляться с помощью широко используемых программ. Для Windows, Linux и Mac – это, например, MetaSAN от Tiger Technology. Исключительно под Mac - XSAN от Apple. Для их эффективного функционирования рекомендуется выделять специальный сервер метаданных, и охватывать все компьютеры обычной Ethernet сетью.

Производительность ExaSAN явно заслуживает уважения. По каждому из каналов PCIe x4 достигаются скорости записи/чтения до 700 Мбайт/сек (против 400 у конкурирующего решения на базе FC 4G интерфейса). Более чем убедительно. И это только начало пути!