JMR ELECTRONICS INC. WHITE PAPER
8968 Fullbright Avenue, Chatsworth, CA 91311 • 818.993.4801 • www.jmr.com

THE NEED FOR SPEED: USING PCI EXPRESS ATTACHED STORAGE

FOREWORD

The highest performance, expandable, directly attached storage can be achieved at low cost by moving the server or work station's PCI bus "outside" and taking advantage of the benefits of the architecture when adding storage subsystems as needed. This is the basis of JMR's[1] U.S. Patent on a "Large Array of Mass Data Storage Devices Connected to a Computer by a Serial Link"[2].

To make use of high performance, high capacity, low cost SAS[3] and SATA[4] disk drives for mass storage, a bridging interface is required from PCI Express (PCIe)[5] to SAS. Conventionally, that host bus adapter or RAID controller would be installed in the host computing system and the storage array would be connected using multilane (x4) SAS cables, thus limiting the bus transfer bandwidth to 4 x 3Gb/s = 12Gb/s irrespective of the quantity of storage devices used.

This is very restrictive, since current generation PCIe bus bandwidth is 2.5Gb/s per lane, and all modern small computers provide at least x8 (8-lane) PCIe expansion slots, which can provide a simplex operating bandwidth of 20Gb/s, 1.67x the bandwidth available from a 4-lane SAS connection. However, this is the tip of the iceberg: Many modern computers provide x16 PCIe and hardware is transitioning from First Generation PCI Express ("Gen. 1") to Second Generation PCI Express ("Gen. 2"), creating wider and wider expansion pipelines (http://www.pcisig.com/specifications/pciexpress/). Gen. 1 x16 PCIe is capable of transfers at 40Gb/s, and Gen. 2 x16 PCIe doubles that again, to 80Gb/s. Other expansion technologies aren't keeping up with this aggressive pace, as will be discussed shortly.

Switched architecture including PCIe, Fibre Channel[6], Infiniband[7] and even Ethernet can enjoy data communications bandwidth aggregation by using more pipelines; however only PCIe and Ethernet are ubiquitous, and PCIe alone provides the highest possible bandwidth for attachment of all peripherals. For example, it would take at

---

[1] JMR ELECTRONICS INC., Chatsworth, CA.
[2] U.S. Patent #7,000,037
[3] http://www.scsita.org/aboutscsi/sas/tutorials/SAS_General_overview_public.pdf
[4] http://www.serialata.org/
[5] http://www.pcisig.com/specifications/pciexpress/
[6] http://www.fibrechannel.org/technology/overview.html#standards
[7] http://www.intel.com/technology/infiniband/

least four aggregated ports of 10Gb Ethernet[8] to achieve the same bandwidth as a single x16 PCIe attachment (or two x8 PCIe attachments). The added overhead and typical mode delays never allow users to achieve the maximum theoretical bandwidth of 10GE, and the added cost of NICs and switches is a decision point for those who really don't need such a network for other applications. Similar cost obstacles exist for FC (Fibre Channel) and IB (Infiniband) networks.

BRIEF HISTORY

Let's look at comparative technological progress of these popular buses. PCI development and revisions are driven by a very active SIG[9], which currently has 836 corporate members, heavily driven by Intel Corporation. Infiniband has no such committee depth, although the Infiniband Trade Association[10] has 42 current members. Fibre Channel Arbitrated Loop has no such committee depth either, although its T11 (technical standards) committee[11] has 53 members.

The Peripheral Component Interconnect (PCI) standard is now 19 years old; the serial layered protocol PCI Express[12] standard is only five years young but software developed for all previous generation (parallel) PCI is fully compatible with new PCIe devices, making it the most useful and ubiquitous communications bus standard in the world. It has also progressed very rapidly from single-lane first generation (2.5Gb/s) to sixteen-lane second generation (80Gb/s) hardware, over a period of only five years. This is progress (32x enhancement) never before seen in any bus standard. The PCI SIG "roadmap" currently plans for Generation 3 hardware (estimated release 2010) operating at 4GHz clock speed and providing bandwidth of 10Gb/s per lane: a 16-lane connection would be as high as 160Gb/s (16GB/s).

By comparison, the FC-AL standard was released in 1994 as a 1Gb/s bus. 2Gb/s emerged in 2001, then 4Gb/s in 2005 and 8Gb/s in 2008. The progress over a period of 14 years has resulted in a bandwidth increase of x8. FC-AL host adapters are not found as a standard on mother boards, and a typical 8Gb/s FC HBA such as the Qlogic QLE2560CK[13] has a cost of roughly $1,100. To aggregate to 32Gb/s requires a 4-port device such as the Emulex LPE11004-M4[14], which has a cost of roughly $2,000. To network this aggregation to other users requires at minimum an 8-port 8Gb/s switch such as the HP AQ233A[15], which has a cost of approximately $3,500.

---

[8] http://www.10gea.org/
[9] PCI SIG = Peripheral Component Interconnect Special Interest Group http://www.pcisig.com/home
[10] http://www.infinibandta.org/
[11] http://www.t11.org/index.html
[12] http://www.interfacebus.com/Design_Connector_PCI_Express.html
[13] http://www.qlogic.com/Products/SAN_products_FCHBA_QLE2560.aspx
[14] http://www.emulex.com/products/host-bus-adapters/emulex-branded/lightpulse-lpe11004/overview.html
[15] http://h10010.www1.hp.com/wwpc/uk/en/sm/WF06b/12169-304608-3659972-3659972-3659972-3662821-3832302.html

Infiniband has been with us since 1999 and SDR signal rates have progressed from 2Gb/s (1x) to 24Gb/s (12x) over a period of ten years. This 12x enhancement is certainly notable, as is the robust QoS (Quality of Service)[16] and failover capability contained within the standard; but these, too, come at very substantial cost to the user.

There are advantages to Fibre Channel, Infiniband and Ethernet protocol networks with regard to resource sharing, and these are well known so we don't dwell on them here. However, in many applications such resource sharing is not required and DAS (direct attached storage) will meet the needs. Even when resource sharing via a network is a prerequisite, there can be a need to simply add more storage using a wide-bandwidth, low-cost interconnection. Herein lies the beauty of PCIe attached storage.

PCIe Attached Benefits

The PCIe standard, as mentioned, is robust and constantly progressing while remaining fully backwards compatible. A x1, x2, x4 or x8 card can plug directly into an x16 PCIe slot and will be fully functional at reduced bandwidth. Likewise, Gen. 1 hardware will work in Gen. 2 slots, and going forward, also in Gen. 3 slots. PCI software developed many years ago are fully supported by PCIe. Virtually every modern peripheral device for small computers today are PCIe bus components.

The standard also calls for the ability to "hot plug" PCIe cards and devices, a new benefit not previously found in other processor I/O architectures. Bus expansion is possible without losing any features or bandwidth of the original motherboard bus slot.

A Simple Example

As such, small, medium or even quite large data storage arrays may be configured using inexpensive switches located in the storage chassis themselves, and operating bandwidth expands via link bonding (aggregation) until the lanes used are fully saturated. A systematic example follows, as outlined in Diagram 1.

Say we have a host (server or work station) with only one available x8 or x16 PCIe slot, as all others are used by required peripherals. This is a very common problem with small computers including both PC and Mac platforms. We desire to attach ~40TB of storage to that system, either due to raw data requirements or to gain the bandwidth achievable by a large quantity of rotating disks, and we would like to have RAID protection of that storage. We will use RAID 5[17] (striping with parity) as a common example. (Note PCIe can address 256 targets, and if each target is a RAID

---

[16] http://ieeexplore.ieee.org/Xplore/login.jsp?url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F35%2F34938%2F01668378.pdf&authDecision=-203

[17] http://searchstorage.techtarget.com/sDefinition/0,,sid5_gci214332,00.html

controller with 16TB attached, such a system seamlessly scales to 4,096TB raw capacity.)

So, we might need 48 x 1TB disks built as six 8-disk RAID 5 sets to yield 42TB usable capacity. Additionally, we'll point out why "8-disk RAID sets" was chosen in a moment -- all connected to a single x8 PCIe slot. Here's an easy way to do it.

Let's install a x8 I-O card in the existing PCIe slot to bring that slot to the "outside world," and expand from it. Such a card is an inexpensive (< $200) device. We attach that port to a multi-slot PCIe bus extender contained in a common enclosure (3U rack) with a 16-bay RAID subsystem[18], thus preserving space compared with using two separate "boxes" to accomplish this task. We then plug in the x8 PCIe cable[19] to that extender rack and create five new PCIe slots from the single port, using standard PCIe switch silicon on the extender's internal backplane. Into two of those "new" slots we install high performance SAS (hardware) RAID controllers each providing x4 SAS performance of 12Gb/s, and connect each of those controllers internally to 8 disk drives, via a x4 SAS Expander[20] also located within the box.

Now, we have an x8 PCIe cable that is connected to 16 disk drives via eight 3Gb/s SAS lanes, so we can enjoy up to 24Gb/s bandwidth, provided each disk device is capable of I-Os of 24Gbs/16, or 1.5Gb/s. In reality, inexpensive disk drives cannot support such data transfers, so the disk drives themselves will be the limitation.

If we use inexpensive 7,200 rpm SATA-2 3Gb/s disk drives which currently cost about $0.20/GB (for Enterprise class devices)[21] and thus represent a "max for min" bargain, we are limited to about 80MB/s peak data transfer rates; thus, an array of 16 devices can achieve perhaps 1.28 GB/s peak performance. Sustained transfer rates are closer to 50MB/s per disk, or 800 MB/s. Still, not bad for only 16 inexpensive disks, and we are nowhere near maximizing performance from an x8 PCIe slot, yet.

Let's now install an inexpensive (~$500) 2-port x8 PCIe switch[22] in two more of the remaining available slots of the PCIe bus extender/storage system rack chassis. This gives us four more x8 connections for additional storage units. Each of those can be simple RAID subsystem units and will not require bus extender backplanes, provided our array will scale to only 48TB raw capacity in this example.

In each RAID subsystem unit are two additional SAS RAID controllers, each addressing 8 drives. We can therefore utilize eight SAS 3Gb/s lanes with each controller and achieve another 800-1,280MB/s throughput, aggregated with the initial

---

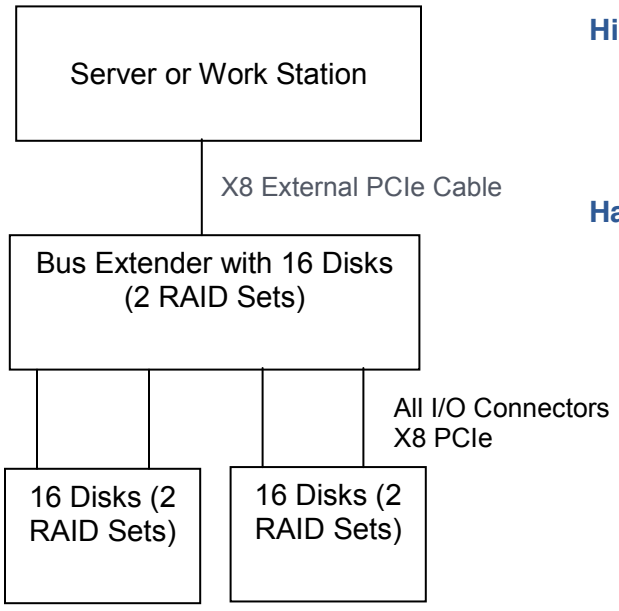
[18] http://jmr.com/storagetechnology_html/images/collateral/NewBluestorPCIe_X8_ExtenderDatasheet%20Rev%20C.pdf
[19] http://www.molex.com/cmc_upload/0/000/381/103/350.pdf
[20] http://en.wikipedia.org/wiki/Serial_Attached_SCSI#SAS_Expanders
[21] http://www.hitachigst.com/portal/site/en/products/ultrastar/A7K1000/
[22] http://jmr.com/storagetechnology_html/images/collateral/NewBluestorPCIe_X4_X8_Internal SwitchesDatasheet%20Rev%20C.pdf

array. The total 48TB system now has six inexpensive PCIe hardware RAID controllers, one 5-slot PCIe bus extender, one x8 PCIe host I-O adapter, two x8 PCIe switches, and five x8 PCIe external cables. None of this hardware is expensive, and no bridging to Fibre Channel, Infiniband, 10GE or any other bus is required. As a result, we have built six 8-drive RAID 5 sets, for 42TB usable capacity, and we ask the host computer to stripe those six RAIDs (RAID 0) to create an overall RAID 50 system which can transfer data at the aggregated bandwidth of the 48 disk drives, which is between 2.4GB/s and 3.84GB/s. This is all done with common hardware and inexpensive 7,200 rpm disk drives, and as few drives as possible. It's also done without a "network," or any expensive, space-consuming network switches.

Using 10k rpm or 15k rpm disk drives enhances performance, but at an increase in cost, until the x8 PCIe bus is fully saturated at the host slot connection. PCIe can be aggregated and simple RAID storage systems operating at 4 GB/s are very achievable.

There are other ways to achieve this performance, but none at the relatively low cost of direct-attached PCIe. The "cost per GB (storage)" model for complete turnkey systems is in the $1/GB range at this time -- for systems that outperform other approaches.

Diagram 1: Block diagram of above example:



Highlights:

- 48TB Raw Capacity
- 6 RAID 5 Sets
- 42TB Usable Capacity

Hardware Requirements:

1 – PCIe Bus Extender/RAID Unit

2 – PCIe RAID Units

1 – PCIe Host I/O Card

2 – PCIe Switch Cards

48 – 1TB SATA-2 Disk Drives

5 – X8 PCIe Cables

Total Throughput > 2.4GB/sec (3.8GB/sec Pk)

Total Cost: $48,000.00 (Typical)

PERFORMANCE TESTING

This chart provides real-world test results using Iometer[23] and displays data transfer rates (throughput) under various test conditions for a 48-disk array similar to that described in the preceding "Simple Example." In this case, the server was a JMR BlueStor[24] unit with two Dual Core 3GHz Xeon CPUs, 4GB memory, running Windows Server 2003 R2 (32 bit). All disk drives were 15k rpm Enterprise 3.5" form factor SAS[25] devices, to maximize performance to near the PCIe 8-lane bus (Gen. 1) limitation.

OPERATION	MB/s	Read MB/s	Write MB/s
Sequential 64K write only	3275	0	3275
Sequential 64K read only	1415	1415	0
Sequential 64K read 67%/write 33%	295	195	100
Random 64K write only	97	0	97
Random 64K read only	305	305	0
Random 64K read 67%/write 33%	189	127	62
Sequential 512K write only	2241	0	2241
Sequential 512K read only	3308	3308	0
Sequential 512K read 67%/write 33%	1355	894	462
Random 512K write only	332	0	332
Random 512K read only	791	791	0
Random 512K read 67%/write 33%	556	373	184
Sequential 2M write only	2654	0	2654
Sequential 2M read only	3270	3270	0
Sequential 2M read 67%/write 33%	2103	1388	718
Random 2M write only	702	0	702

[23] http://www.iometer.org/
[24] http://jmr.com/storagetechnology_html/server.html
[25] http://www.fujitsu.com/us/services/computing/storage/hdd/enterprise/

Random 2M read only	1367	1367	0
Random read 67%/write 33%	1026	683	343
Sequential 30M write only	3283	0	3283
Sequential 30M read only	2724	2724	0
Sequential 30M read 67%/write 33%	2560	1678	883
Random 30M write only	1688	0	1688
Random 30M read only	2376	2376	0
Random 30M read 67%/write 33%	2106	1409	697

SUMMARY

The results indicate reasonably high performance for a low-cost storage array which can be directly attached to any modern server or work station. Resource sharing, if needed, can be done via 10GbE using appropriate NICs in each server or work station and a 10GbE switch, at some sacrifice in overall performance compared with direct attachment; but as previously addressed, in many cases the host computing systems are already networked and the user only needs to add more storage – in which case DAS serves the purpose. If resource sharing is not required, then each computer/storage system can stand alone, making PCIe DAS a very viable, high-performance, low-cost alternative to other approaches.