StorExpress

Integrating Solid State Storage in a PCI Express Clustering Interconnect

Venkata Krishnan

Lynne Brocco

Revision 1.2

May 2009

Dolphin Interconnect Solutions
www.dolphinics.com
## Contents

Executive Summary .......................................................................................................... 3  
Introduction ....................................................................................................................... 4  
Flash-based Solid State Storage ....................................................................................... 5  
StorExpress ........................................................................................................................ 6  
Usage Models ..................................................................................................................... 7  
  Single server .................................................................................................................. 7  
  Multiple servers ........................................................................................................... 10  
  Shared storage .............................................................................................................. 10  
  Clustering solution ...................................................................................................... 11  
Advanced Usage Models ................................................................................................. 12  
Closing Remarks ............................................................................................................... 12  
Appendix: Case Study .................................................................................................... 13
Executive Summary

Direct PCI Express (PCIe) attached flash storage offers unparalleled performance levels and vastly outperforms SAS/SATA-attached flash storage solutions. However, such PCIe attached flash storage cards, in their existing form, are restricted to the confines of the PCIe slots in the host server. This approach can neither support a disaggregated storage (JBOD/RAID-like) mode that a SAS/SATA-attached storage provides nor can it support a shared storage solution in a clustered environment.

By leveraging the expertise in PCIe based clustering and IO expansion technology, Dolphin, in partnership with Fusion-IO, provides an enhanced PCIe flash storage solution in the form of StorExpress – an appliance that augments the high-performing direct PCIe attached storage from Fusion-IO. StorExpress enables a larger capacity by using Fusion-IO’s PCIe flash storage cards (iodrives) as building blocks and also provides disaggregation capabilities along the lines of SAS/SATA-attached flash storage.

Though the features provided by StorExpress for enhancing a single-host direct attached storage would stand on its own legs, the StorExpress solution goes even further. As a matter of fact, the StorExpress solution is built around Dolphin’s PCIe based interconnect technology – one that integrates PCIe-based clustering/IO expansion on a single platform. On the storage front, StorExpress facilitates a high-performing shared storage configuration in a multiple server environment that includes rackmount as well as blade-servers. On the cluster communication front, StorExpress can be seamlessly integrated into Dolphin’s PCIe-based clustering environment or can even act as a standalone PCIe clustering switch. And with the aid of Dolphin’s software, clustering applications can continue to use legacy APIs (such as TCP/IP-based sockets or MPI) while benefiting from the extremely low-latency and high bandwidth communication afforded by Dolphin’s software/hardware solution. Such a tightly knit configuration offers a great potential for improving the performance of applications that require (a) low-latency/high-bandwidth communication and (b) extremely fast storage.

StorExpress is ideal for deployment in web and database-centric applications such as web caching, web indexing, data warehousing, data mining and OLTP applications. Such applications demand storage capacities and capabilities that can be met only by traditional HDD based solutions but not with current PCI express based solid state storage solutions. At the same time, those applications demand extremely high performance in the form of low latency and high bandwidth – one that can be met easily by using PCI Express based solid state storage devices.

This whitepaper describes some of configurations and usage models supported by Dolphin’s StorExpress solution. A brief summary of the migration of storage from hard disk drives to flash based solid state storage technology is presented, followed by a description of the advantages and disadvantages of direct PCIe attached solid state storage. The paper then goes on to describe the rich feature set offered by StorExpress – this enables it to not only provide an enhanced version of direct PCIe attached storage but also offer additional capabilities that go over and beyond the realm of single host attached storage.
**Introduction**

The advantages of migrating from a traditional hard disk drive (HDD) environment to a flash-based solid state storage (SSS) approach are well known. One of the most important reasons for migrating to flash-based SSS from HDD is the fast access time. Typical access times for SSS\(^1\) devices are in the order of 10-100 µs as opposed to several milliseconds for a HDD. When compared to the nearest level in the storage hierarchy, the access time differential between SSS and DRAM memory is unlike the wide chasm that exists between the access times of HDD and DRAM.

An impediment to wide spread adoption of SSS in storage solutions has been the high costs when compared to that of traditional HDDs. Mostly due to the legacy way of thinking, the conventional way of evaluating storage has been to use the standard metric of cost/GB. Such a comparison, though meaningful from a raw capacity point of view, is irrelevant from a performance standpoint. Indeed, in most enterprises, performance rather than capacity is the need of the hour with significant emphasis being placed on getting higher performance. Performance metrics used in this regard include the maximum sustainable IO operations per second (IOPS) and bandwidth (in MB/s) for a range of read and write operations.

Traditional HDDs have fared poorly in the above two metrics when compared to SSS. As a means of getting higher performance, one commonly used technique has been to combine multiple HDDs in a large RAID configuration. The RAID approach indeed provides a significant improvement for sequential (i.e. predictable) disk accesses. Typically, the capacity in such setups is always under-utilized. This pushes the cost/GB for HDDs since the storage capacity reflects the utilized storage rather than the raw capacity across all HDDs. Nonetheless, this approach only improves sequential disk access operations. Random read/write accesses, which are critical for a wide range of applications, continue to underperform. Adding multiple HDDs to improve sequential access also comes at the added cost of increased real estate as well as power consumption.

Indeed, the need for higher performance using HDD while trying to minimize power consumption and footprint has resulted in a trend towards using smaller but more expensive HDD drives. These disks, which have lower capacity, have a higher cost/GB than their larger counterparts but can operate with lower power consumption and reduced storage footprint. However, this approach does not solve the performance bottlenecks when the read/write operations are completely random. Applications that do not have a highly sequential (predictable) access pattern continue to fare poorly. And when compared to SSS, the power consumption is still higher in this setup.

Unlike HDDs, SSS offer optimal performance for both sequential as well as random operations. And from a power dissipation point of view, SSS consumes only a fraction of the power of HDDs. The lower heat dissipation also reduces the cooling requirements that would normally be required when deploying HDD-based storage arrays. This eventually leads to a lower total cost of ownership (TCO). Overall, one could argue that,

\(^1\) For the purpose of this white paper, SSS refers to flash-based SSS unless otherwise noted.
rather than continuing to use the legacy storage metric of cost/GB, the metrics that would be relevant in today’s storage environment are IOPs per watt, IOPs/$, bandwidth per watt etc. From a reliability point of view, SSS can offer increased reliability since they have no moving parts and can withstand extreme shock and vibration without danger of data loss. And unlike earlier generation of flash-based SSS solutions which had limited write endurance and hence a short life-span, today’s enterprise based flash-based SSS solutions incorporate sophisticated balancing and wear-leveling algorithms. This provides a high write endurance level resulting in an increased lifespan that ranges from 15 to 25 years. In contrast, wear and tear of mechanical parts in a HDD results in an operational lifespan of typically three years.

And finally, the migration to SSS is also being accelerated by the increasing capacity, higher performance, and dropping prices over time.

Flash-based Solid State Storage

To address the wide disparity between HDD and DRAM memory accesses, two types of SSS solutions are available in the market – one based on flash memory and the other based on DRAM memory.

The lack of high-performing flash-based SSS solutions until recently had meant that enterprise-level SSS solutions resorted to using a DRAM-based solution. This solution, though high-performing, comes at a cost. In addition to the price factor, DRAM-based solutions require either HDDs or flash-storage as a back-up storage as well as battery in the event of power loss. With the vast improvement in flash-based technologies in recent years, flash-based SSS is now able to provide the same level of performance as DRAM-based SSS solutions albeit at a lower cost.

The most common type of flash-based SSS is the SSD (Solid State Drive) – one that provides a traditional disk interface and uses legacy IO protocols such as SAS or SATA. The other emerging type of SSS is a flash-based storage card with a direct PCI express (PCIe) interface. Both approaches have their distinct advantages and disadvantages.

SSS with traditional disk interfaces can seamlessly replace existing storage arrays based on HDDs without disrupting the legacy storage architecture. However, this approach is limited by the performance of the disk controller interface. Using a SSS with a direct PCI Express (PCIe) interface overcomes the performance problem and allows the full potential of the underlying PCIe interconnect to be exploited. With even an x4 Gen1 PCIe interface providing bandwidths of over 800 MB/s, there is indeed a significant performance advantage that PCIe SSS has over traditional disk controller-based SSS solutions. And with today’s servers providing at least one PCIe slot, a PCIe SSS can be integrated as part of the server’s overall storage solution.

Nonetheless, direct PCIe attached storage does have some drawbacks. The limited number of PCIe slots on the server limits the scalability of a direct PCIe attached flash-storage solutions. To increase storage capacity, one approach would be to resort to a server providing multiple PCIe slots. Often though, 1U or 2U servers have typically one...
or two available PCIe slots. And populating the slots with high-capacity PCIe SSS cards in the confines of this server may adversely impact its cooling requirements. This may call for a larger server but even a server with a large footprint (4U) typically provides at most four PCIe slots. Moreover, integrating storage within the server and using all available slots drastically impacts the expandability of the flash storage solution. Rather than augmenting the existing storage with higher capacity PCIe SSS cards, low-capacity PCIe SSS cards would need to be replaced to free up the available PCIe slots. Also, given that the PCIe slot is contained within the server, the PCIe based SSS cannot be collocated at distances further away from the server platform. The inability to disaggregate storage from the server makes it difficult to provide a JBOD or RAID solution along the lines of disk-controller based SSDs.

Dolphin’s StorExpress family of products addresses the above problems and combines the best of direct PCIe attached SSS and that of disk-interface based SSS solutions. StorExpress disaggregates the PCIe based flash storage cards away from the server while using only a single PCIe slot in the server. By hosting multiple PCIe storage cards in its chassis, StorExpress eliminates the need for a server that supports several PCIe slots. In addition, when storage is migrated from the server, the storage solution can be easily scaled and also collocated at great distances from the server. The StorExpress solution can also support a multi-host environment wherein the storage can be distributed to each host. With the inclusion of Dolphin’s clustering software drivers, a fully-shared storage solution can also be supported. The rest of this paper describes StorExpress in great detail and also describes the usage models that go over and beyond a traditional direct-attached storage solution.

**StorExpress**

Dolphin’s StorExpress solution is a solid state storage system that provides a direct but disaggregated PCIe connection to a remote chassis containing multiple PCIe flash memory cards. Depending on the configuration, a single chassis can support between 0.5 TB to 4 TB of flash storage. Each PCIe flash card in StorExpress can be made to appear as distinct disk block devices or alternatively can be treated as a single disk block device. This remote PCIe storage solution builds upon Dolphin’s PCI Express expertise and history of PCIe I/O expansion solutions. Using PCIe I/O disaggregation, StorExpress continues to provide the performance, reliability, and power advantages of a direct PCIe attached SSS and at the same time, enhances the solution by bringing it on par with disk-controller based SSS solutions. StorExpress permits a PCIe based storage array to be setup and configured as a JBOD or RAID.

A basic StorExpress solution consists of one storage chassis and PCIe host adapter card. The adapter card has two PCIe x4 connectors, which can also be used for a single x8 connection, and is installed into the server. The StorExpress chassis, which internally hosts multiple PCIe SSS cards, is then connected to the PCIe adapter card using copper or fiber optic cable. Additionally, the PCIe host adapter card comes in different form factors that enable the StorExpress chassis to be integrated not only in a typical rack-mounted server configuration but also in blade server environments.
StorExpress supports a variety of configurations. Table 1 details possible connections between one or multiple StorExpress chassis and one or multiple servers. These configurations are described in greater detail in the subsequent sections.

Table 1. StorExpress can be integrated in a single or multiple servers based configuration.

<table>
<thead>
<tr>
<th>StorExpress Chassis</th>
<th>Server(s)</th>
<th>PCIe connections between StorExpress &amp; Server(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>Single</td>
<td>x4, x8 or dual x8</td>
</tr>
<tr>
<td>Multiple</td>
<td>Single</td>
<td>x4, x8 or dual x8 per chassis</td>
</tr>
<tr>
<td>Single</td>
<td>Multiple</td>
<td>x8 to two servers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x4 to four servers</td>
</tr>
<tr>
<td>Multiple</td>
<td>Multiple</td>
<td>x8 to two servers per chassis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x4 to four servers per chassis</td>
</tr>
</tbody>
</table>

Dolphin currently supports both shared PCIe I/O expansion and inter-server clustering in its interconnect products, and extends this capability to its StorExpress products. In a multi-server configuration, in addition to providing storage I/O access, a StorExpress solution is also capable of functioning as a clustering switch using Dolphin Express software. This eliminates the need for a dedicated clustering fabric for inter-server communication.

Usage Models

Single server
The simplest StorExpress usage model is a single server with an x4 or x8 connection to a StorExpress chassis. An x8 PCIe host adapter provides the cable or fiber connection from the server.

Figure 1. Single Server Configuration

The single server, single chassis usage model provides the following advantages:

- Storage capacity of multiple PCIe flash cards (currently 6), while consuming a single PCIe slot in the server
- Ability to remotely locate the StorExpress chassis from the server
  - Up to 10 meters with copper cable
  - Up to 300 meters with fiber optic cable
- RAID solution supported across the multiple drives

This usage model continues to support the extremely low read access latency (< 50 µs) offered by the underlying direct PCIe attached flash card. For a x8 server-to-chassis connection, random read bandwidths of up to 1500MB/s and random write bandwidths of up to 1250 MB/s are possible (with 16KB block sizes), with over 150,000 read IOPS and 120,000 write IOPS (for 4KB block sizes).

Capacity and reliability can be further increased by connecting a second StorExpress chassis to the server. The connection can use the same host adapter as with the single chassis model if x4 performance is sufficient, or a second adapter can be installed in the server to support an x8 connection to each chassis.

The advantages of the single host / single StorExpress chassis are retained, while doubling the storage capacity and allowing RAID configurations across the two chassis, as shown in Figure 2.

**Figure 2. Single Server Dual Chassis Configuration. Three different RAID configurations are shown in the figure.**

Further performance enhancements can be obtained with a “dual x8” connection to the server, shown in Figure 3. This model requires a second PCIe slot in the server, for a second PCIe host adapter. This second adapter is connected to the StorExpress chassis, providing a parallel x8 connection.
One subset of flash memory drives in the chassis is assigned to the first adapter, and another subset is assigned to the second adapter. By default, assignment of the storage cards is evenly split, but can be configured in other proportions if desired.

With this configuration, random read bandwidths of 2900 MB/s, random write bandwidths of 2800 MB/s, and over 270,000 random read and write IOPS (for 4KB block sizes) can be achieved. Figure 4 charts out the performance for varying block sizes ranging from 512 bytes to 1MB for such a dual-x8 configuration.

Not only is performance greatly increased, but reliability can also be enhanced as this model supports a high performance RAID1 configuration across the two sets of drives, with redundant cables and adapter cards. In Figure 3, devices 0:2 use one host adapter.
and devices 3:5 use the other host adapter. Data is mirrored across the two sets of devices using RAID 1.

**Multiple servers**

**Shared storage**
Leveraging on Dolphin’s proven multi-server PCIe I/O expansion technology, a StorExpress chassis can also support connections to multiple servers. This configuration allows storage resources to be shared among multiple physical systems.

Up to four servers may be connected to a single chassis. Although there is physically a single box, the devices appear as direct PCIe attached storage to the servers. Currently, the StorExpress model supports the allocation of a fixed amount of storage to each server as one or more logical drives.

A dual server configuration supports an x8 connection to the chassis from each server. Each server has the same x8 PCIe performance advantages as with the x8 single server configuration, while sharing the capacity (See Figure 5). Extending this even further, up to four servers can be supported in this model wherein each server would each have an x4 connection to the chassis, and PCIe x4 performance. In all cases, the same access latency benefit is provided as with the single server model.

![Figure 5. Dual Server Single Chassis Configuration](image)

The ability to connect a single server to two chassis can be extended to multiple servers as well. For instance, two servers can each have a connection to two chassis, with RAID configured across each servers assigned drives in both chassis (See Figure 6). This provides the storage capacity of a single chassis, but provides the reliability of two chassis connections.
The configurations shown in Figure 5 and Figure 6 are only two among the several multi-host configurations possible. The inclusion of a Dolphin switch (DXS410) as part of the DX interconnect can allow several hosts to be connected to a single interconnect and enable each host to carve out a portion of flash storage from the pool of StorExpress resources. See Figure 7.

**Figure 7. A multi-host/StorExpress configuration**

---

**Clustering solution**

The StorExpress solution uses the same underlying technology as the Dolphin Express product line, which integrates PCIe I/O expansion and inter-processor communication over the same physical interconnect. Using Dolphin Express software, servers connected
to the same StorExpress chassis can communicate with each other through the StorExpress switch while also taking advantage of PCIe flash SSS performance. No additional hardware is required. With the aid of Dolphin’s software drivers, existing applications using APIs such as sockets or MPI will continue to run without the need for code modification or recompilation. Figure 8 illustrates a 2-server setup where in the servers not only own part of the storage in StorExpress but applications on each server can also communicate using sockets or MPI.

Figure 8. Clustering and Storage I/O Solution

Advanced Usage Models

The integration of clustering and Storage I/O as part of single interconnect technology sets the stage for providing a full-fledged shared storage solution. With the aid of the PCIe-based interconnect that provides extremely low latency/high bandwidth and software that supports all traditional APIs, it is straightforward to implement a storage solution that goes beyond direct-attached storage. Indeed, with the current infrastructure, a traditional client-server based network file system or network block storage can be easily put in place. For instance, in Figure 7, Servers A, B and C could function as file or block device clients while Server D could own the StorExpress device and function as a file or block device server. Dolphin will soon provide a truly shared storage solution that removes the need for upper layer software or involvement of a server’s CPU/memory subsystem. Such a solution would allow multiple storage servers to host a subset of the overall flash storage while providing a SAN-like solution based on PCIe technology.

Closing Remarks

Dolphin’s StorExpress enables direct attached PCIe flash storage to be disaggregated from the server thereby supporting traditional direct attached storage models that is in line with HDD based solutions. The solution also goes beyond just providing an expansion platform for PCIe flash storage cards and can indeed support a multi-host environment wherein the storage can be apportioned to different hosts. With the aid of Dolphin’s software drivers, the StorExpress solution can be further expanded to support a fully shared storage solution.
Appendix: Case Study using DRBD®

A usage case for StorExpress is described in this section. The application under consideration is DRBD® (Distributed Replicated Block Device) from LINBIT technologies (www.linbit.com, www.drbd.org). DRBD is a distributed high availability storage system for Linux. Any write performed on the local storage subsystem is synchronously and transparently written to another system’s storage through a fast and reliable interconnect.

Figure 9 shows a DRBD based MySQL database installation. In the figure, DRBD is used to replicate the active master to a standby master. Dolphin Express is the interconnect solution between the two servers and provides a high performance connection. The replicated master is connected to slaves. The slaves are used for read queries while incoming writes go to the master node.

Figure 9. DRBD support on a platform using Dolphin Express components

For the configuration shown in Figure 9, the storage attached to the primary and standby master could be based on either traditional HDDs or StorExpress. It must be mentioned that even though the figure illustrates DRBD as used in a MySQL database, the use of DRBD is not limited to a MySQL environment. Indeed, DRBD can be deployed in any environment where there is a need for high-availability storage.

From a performance perspective, the efficacy of DRBD is influenced by the storage subsystem as well as the latency/bandwidth of the underlying interconnect. For this reason, the following two configurations were considered:

(a) Gigabit Ethernet/10 Gigabit Ethernet/Dolphin Express as the interconnect between primary/standby master and a traditional HDD used for storage
(b) Gigabit Ethernet/10 Gigabit Ethernet/Dolphin Express as the interconnect between primary/standby master and StorExpress for storage
Note that DRBD can run unmodified over Dolphin Express interconnect using Dolphin’s SuperSockets software. The performance of the above two configurations was evaluated using **dd** – a commonly used UNIX program for low-level copying of data at the device block level.

**Figure 10. Performance of dd block level transfers using different interconnect technologies with storage based on (a) HDD and (b) StorExpress.**

![DRBD Performance (HDD)](image)

![DRBD Performance (StorExpress)](image)

Figure 10 (a) and (b) illustrate the performance, in terms of the DRBD bandwidth, for varying block sizes for the two different storage configurations. Obviously, the higher the bandwidth offered by the underlying interconnect, the better the DRBD performance. At the same time, DRBD also requires application writes to be completed both at the local as
well as the remote system storage subsystem. This involves additional hand-shaking involving small messages between DRBD running on the primary and standby master. Rather than being bandwidth centric, such messages are very latency centric and can also impact the overall performance. Nonetheless, for a low-performance storage subsystem using HDDs, the interconnect performance in terms of latency and bandwidth plays a minimal role. Indeed, a Gigabit Ethernet interconnect can achieve the same level of performance as a 10 Gigabit Ethernet or Dolphin Express. See Figure 10(a). Here the HDD performance is the bottleneck.

However, when the storage bottleneck is addressed by removing the HDD and replacing it with StorExpress, the emphasis moves to the interconnect performance. As can be seen in Figure 10(b), there is a significant performance improvement when moving from a Gigabit Ethernet to a 10 Gigabit Ethernet. This is result of the additional bandwidth afforded by the 10 Gigabit Ethernet over Gigabit Ethernet. At the same time, there is also a latency component involved in the DRBD solution. And low latency is not the forte of 10 Gigabit Ethernet.

The combination of Dolphin’s StorExpress and its low latency/high bandwidth communication model using Dolphin Express/SuperSockets software provides an ideal environment for supporting an extremely efficient replicated block storage mechanism. As a result, it shows up to 50% improvement in performance over a 10 Gigabit Ethernet solution for a wide ranging set of block sizes. Overall, Dolphin Express along with StorExpress based storage provides the best replication performance for DRBD.