

# Survey of the Storage Evolution

Mike Knowles

Army Research Laboratory (ARL)/Raytheon Corporation, Aberdeen Proving Ground, MD  
knowles@arl.army.mil

## Abstract

*The storage market in 2003 is in a state of transition on several fronts. This seems to be a time where noteworthy technologies are evolving that will have long-term impacts on Information Technology and High Performance Computing (HPC). Two of the more important evolutions in the storage arena are the serialization of storage access mechanisms, and efforts to access storage via IP networks. While these two efforts are not necessarily related, they are both seen by many storage vendors as defining trends in the industry. This paper focuses on these two IO bus level evolutionary trends in particular, and how we can expect them to affect HPC users.*

## 1. Serialization

Historically, storage subsystem bandwidth, including memory, has not kept pace with processing capacity of CPUs (Moore's Law: processing capacity doubles every 18 months) and network bandwidth growth (doubling every 12 months). While the capacity of storage subsystems has shown remarkable continued growth (doubling every 12 months), the bandwidth available to those storage devices has not kept pace. One of the main issues for this lag in bandwidth from storage has been the bus based access mechanisms of the storage devices themselves. These devices are typically accessed via shared parallel buses that limit bandwidth in several ways. Parallel-shared storage buses typically require arbitration for bus access, are half duplex, are limited in frequency by cross talk between adjacent wires, and are restrictive in length and number of supported connections. To address these concerns, system and storage vendors are converging on serial storage buses. In particular, Serial Advanced Technology Attachment (SATA), and Serial Attached SCSI (SAS) storage buses are now evolving into product lines. SATA 1.0 devices are available today while

SAS devices should be available 1H04. These new point-to-point access mechanisms can then be aggregated into low-level switched storage networks. This type of aggregation has already occurred with Ethernet infrastructure, Fibre Channel (FC), and other high-speed system interconnects. A switched architecture promises better performance at the individual device level (by dedicated connections), and at the subsystem level (less bus arbitration, full duplex, increased clock rates). Other benefits of this architecture include better scaling, more total devices supported, better fault tolerance (no single point of failure), improved cooling (due to reduced cabling), lower power requirements, and more distance between devices. What is even more encouraging from a buyer's standpoint is that the organizations that are developing SATA and SAS are actually working together so that physical interconnects between the devices are the same, and the SATA protocol can be tunneled over SAS. Therefore, buyers can choose based on price/performance characteristics which type of device they want in a particular storage subsystem and use the appropriate driver. These serial interconnect methods are described below.

### 1.1. Serial Advanced Technology Attachment (SATA).

Advanced Technology Attachment (ATA) devices have been with us for years in desktop and laptop systems. Since the target market for these devices is the desktop/laptop they have been designed and manufactured with significantly reduced duty cycles, and performance specifications compared to Small Computer System Interface (SCSI) and FC disks. Therefore, they are significantly less expensive than SCSI and FC disks. Economies of scale have further helped to reduced costs of ATA disks relative to enterprise type disks (SCSI, FC) since the enterprise unit sales make up approximately one twelfth of the ATA unit sales. Capacity of ATA disk devices is high, relative to enterprise disks, since they typically run at slower RPM. The slower RPM has

hindered bandwidth capability. However, these devices have been accessed via parallel buses that have steadily increased in bandwidth, such that current interface speeds are a respectable 133 Mega-Byte/second (MB/s). A group of SATA disks will typically be serially attached to a Host Bus Adapter (HBA) in a hub or star configuration (Figure 1). The HBA will be attached via a PCI-X system bus. This configuration allows for dedicated channel bandwidth with the ability to have several outstanding ATA commands to different devices. In summary, ATA disks are inexpensive; they have great capacity characteristics, and decent performance. These facts, coupled with the point-to-point serial access mechanism of SATA, have created a very dense (GB/ft<sup>2</sup>), respectably performing, low cost storage solution. Vendors are positioning this capability at the enterprise and HPC markets, as large capacity, moderate performance, low cost RAID storage, one level above near-line tape.

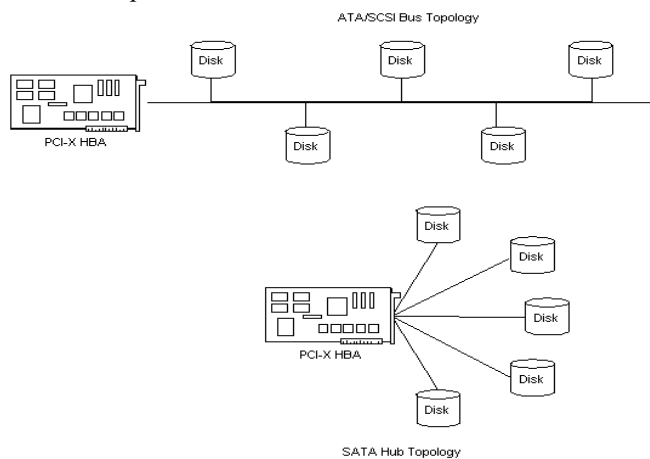


Figure 1.

## 1.2. Serial Attached SCSI (SAS).

The SAS specification is now under review, but there seems to be a marked migration by many system and storage vendors towards implementing SAS devices and interconnects. The ability to support both SAS and SATA via the same back plane and interconnect seems to be a primary driver for this enthusiasm. Traditionally, SCSI drives have been used in servers and workstations, instead of the desktop. This differentiation remains with SAS. Indeed, there will probably be more of an overlap in device capabilities between SAS devices and FC devices, when SAS devices start appearing in 2004. The latencies, capacities, high availability, and bandwidth characteristics are very similar between proposed SAS devices and their FC counterparts. Vendors, however, maintain that FC will still be the primary high-end storage subsystem. SAS devices will differentiate themselves from SATA devices by bringing to market the robust SCSI command set, a

higher duty cycle design and improved mean time between failure (MTBF), a targeted interface bandwidth of 300 MB/s (with projected increases to 600 MB/s), more on-device buffering capability, deeper command queuing/ordering, and dual access ports (for high availability). The current bandwidth of SCSI (ULTRA320) is 320MB/s (640MB/s has been tested, but is not generally available) so SAS devices will actually have slower interfaces than current parallel SCSI devices, but will not have to share a bus. The maximum number of devices that can be addressed in a SAS design is also increased to 16000 per domain (parallel SCSI buses support 16). These theoretical device counts can be attained via expander cards that daisy chain the switched serial connections (Figure 2). The expander cards allow for multiple hosts to connect with SAS disks, they are essentially the switches in the serial switch interconnect. They will have up to 128 ports available for host or device connections. The expanders will also allow SATA devices to physically attach. Logical compatibility between SAS and SATA disks is provided by a SATA Tunneling Protocol (STP) that will allow SATA disks to be plugged into expander devices in a SAS interconnect. Therefore, there is one-way compatibility: SAS supports SATA.

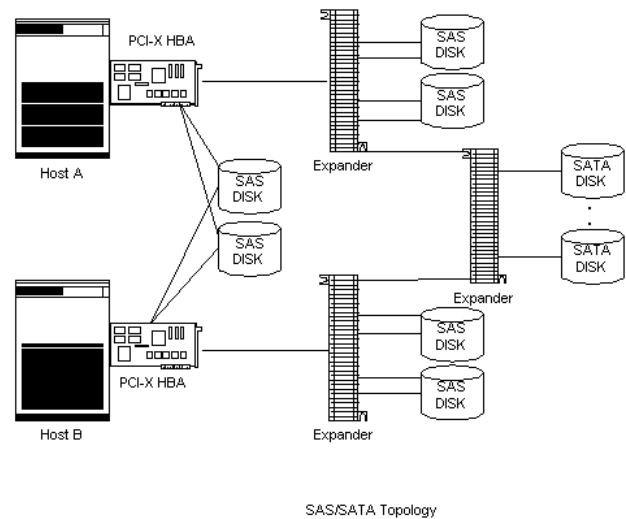


Figure 2.

A comparison of different disk access technologies is shown in Table 1:

Table 1.

	SATA 1	SATA 2	SCSI	SAS	FC
Duplex Seek Time	Half 9ms	Half 9ms	Half 3-7ms	Full 3-7ms	Full 3-7ms
Interface BW (MB/s)	150	150 (300 planned)	(shared, 320 tested) 640	300 (600 planned)	200 (400 announced)
Connectivity	1	1	16	16k	15M
Capacity (GB)	250	250	146	146	146
MTBF (hrs)	600k	1000k	1200k	1200k	1200k
RPM	7200	10k	10-15k	10-15k	10-15k
Command Set	ATA	ATA	SCSI	SCSI	SCSI
Port Count	1	1	1	2	2
Relative Cost	\$	\$\$	\$\$\$	\$\$\$	\$\$\$
Availability	Today	1H04	Today	1H04	Today

Table 2.

	SCSI	FC	GbE
Channel Speed (MB/s)	320 (640 coming)	200 (400 & 1000 coming)	100 (1000 coming)
Clock Rate (Gbps)	—	2.12	1.25
Max Bus Distance (m)	15	300 multi-mode, 10k (single mode)	100 (cat-5), 500 (multi-mode), 10k (single mode)
Topologies	bus	pt-to-pt, loop, fabric	switched
Media Duplex	copper half	copper, fibre full	copper, fibre full
Max Payload (Bytes)	—	=< 2112	=< 1500,

## 2. Storage Over IP

The other major evolutionary capability currently occurring in the storage industry is the push to incorporate IP into storage at the bus level. Storage access mechanisms have historically been extremely limited in the distances devices can be away from host controllers. This fact has constrained storage configurations in capacity, and scalability at a time when data growth is exploding. Fibre Channel (FC) Storage Area Networks (SANs) were designed to provide some recourse to this problem. FC has been largely successful on this front. But, due to the inability of the FC device vendors to work together to promote their industry, incompatibilities plagued FC technology for years and slowed its adoption. At the same time IP network bandwidths were increasing, and switching technologies solved contention and latency issues in networks. Much of this was in fact based on leveraging emerging FC technology. FC and Gigabit Ethernet (GbE), the current enterprise IP network of choice, in fact share the same underlying media (table 2), and borrow extensively from each other in the higher layers of their protocols.

The ubiquity of IP-based networks also has driven interoperability between host and network device vendors, and comprehension by large numbers of support personnel, compared to FC. Economies of scale also have lowered the GE price per port below FC, and driven efforts in further embellishing IP-based hardware (10 Gigabit Ethernet) and software solutions (security, and management). Flexibility in storage is required to add resources and/or reallocate resources without affecting ongoing operations. This type of flexibility and ease of use were major advantage to Network Attached Storage (NAS). A NAS storage array could be wheeled in the door, powered on, plugged into the IP network, allocated to clients, and be providing storage in a matter of minutes. The desire to easily share storage from multiple clients also has contributed to NAS success. While the ability to share data directly through the SAN has remained elusive. Data replication and backup have become ever more important in this data-centric world, while backup windows are closing. Therefore, new flexible methods to backup data have become necessary, which would replicate data and/or move it offsite at the same time in an automated way. All of these factors have contributed to the idea that IP-based networks are the means by which data should be accessed, and as a means to eliminate storage distance limitations. To this end there have been several IP based storage access protocols defined. All of these build on well the well known SCSI, IP, and/or FC protocols. The storage industry's primary IP based protocols are described below.

### 2.1. Internet Small Computer System Interface (iSCSI).

iSCSI is a block oriented storage protocol that runs on TCP/IP networks. No specialized networking equipment is required in an iSCSI environment, but it would help performance. iSCSI targets can be placed on

existing networks, and accessed through host IP-based Network Interface Cards (NIC). iSCSI provides a mechanism to implement SCSI commands, control, and data over existing networks. It is very similar to the FC Protocol (FCP, FC encapsulated SCSI command, control and data) except that iSCSI uses generic IP networks not FC switches. A block oriented protocol allows initiators to place data and metadata, on target devices in locations of the initiator's choosing. The iSCSI initiator must be able to interpret the underlying block layout of the iSCSI target devices to make the stored data meaningful (i.e. to access file systems or databases). In contrast to block oriented protocols NAS uses file oriented protocols such as Network File System (NFS). NAS clients access data by a file handle provided by a NAS server. The actual placement and management of the data is the responsibility of the NAS device, not the client. A NAS server will actually create a file system and present it to a client, while a client (initiator) in an iSCSI configuration is presented with a raw device (target) and interprets/creates the data structures on it. The primary advantages of block oriented access to storage are performance and data sharing. Servers in iSCSI are actually the storage and/or gateway devices (targets), while the clients are IP based host systems (initiators). iSCSI must rely on several underlying protocols to provide many of its basic services. These services include GbE for the physical link between devices, IP for packet forwarding and routing, TCP for flow control, and error recovery. The security, integrity, authenticity of iSCSI packets must also be maintained by the IPsec utilities, and other protocols are used for device discovery, login, and status maintenance operations.

The topologies that will be prevalent with the advance of iSCSI are client to appliance type connections, and client to iSCSI gateway. The client to appliance configuration allows client SCSI commands to be routed directly to, and interpreted by, embedded storage. An iSCSI gateway configuration provides for an iSCSI to FCP translation, such that the actual storage devices are attached to FC fabrics. This translation replaces one transport mechanism for another, while the package contents remain fixed. Many initiators to one iSCSI target device and one initiator to many iSCSI target devices are anticipated. Note however that each individual session to a unique device requires the maintenance of a TCP/IP session. So if there are several devices an initiator wishes to use, several TCP/IP sessions have to be maintained at both ends of the connection. When one considers there will be a major increase in the number of network interfaces per host, there is a large amount of TCP/IP overhead to keep track of. To support the increased processing loads of TCP/IP based network stacks, iSCSI will require TCP/IP Offload Engines (TOE), on the NIC. The TOEs will take care of most of the burden of the end-

to-end communication. This function has been moving to the NIC, but the process has to be accelerated to support iSCSI. There are now iSCSI adapters that offload not only the TCP/IP protocol processing, but a large portion of the iSCSI protocol processing as well.

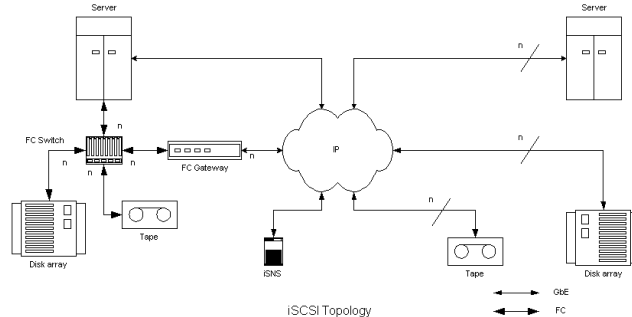


Figure 3.

iSCSI requires a name server capability so that host systems can query a known address and find out what devices are available to login to and use. This alleviates the problem of polling large address ranges during device initialization, and when devices attach to a network. Therefore, an internet Storage Name Server (iSNS) protocol has been defined to provide the lookup service. The iSNS will provide device registration, zoning and state change management for IP SANs. iSCSI supports a system boot capability so hosts will need iSNS to find their boot devices. The iSNS will also be able to perform security key management routines that iSCSI devices will require to access each other. Security in an iSCSI environment is extremely important to provide data encryption and integrity for the payload, authentication between devices involved in transactions, and authority to grant various access types to remote client devices. iSCSI depends on IPsec to provide the diverse set of tools for all these types of security mechanisms.

## 2.2. Fibre Channel over IP (FCIP).

Fibre Channel over IP is basically an encapsulation, or tunneling, of FC packets in IP packets. This is typically done to take advantage of wide area network (WAN) connections as a transport mechanism between FC islands. Both endpoints of this type of connection will usually be FC switches, so there is basically one topology supported with FCIP (Figure 4). The connections between the switches are viewed as local FC Inter Switch Link (ISL), a common FC fabric entity. A FC packet leaving one SAN will be compressed, encrypted, segmented, and encapsulated by IP headers to be routed across network links. At the destination SAN, the reverse procedure presents a native FC packet to a FC switch port. This

process is implemented through FC bridges. There are a surprising number of FC bridges available today. The most common FC bridges are FC to GbE bridges, but there are bridges available for FC to SONET, FC to ATM, FC to dark fiber, and other wide area transports. FCIP usually refers to FC to GbE bridging.

As the requirement for storage at a distance has evolved, specialized FCIP bridges have been developed that translate FC addresses between distant SANs so there are no domain collisions (FC allows only one domain in a fabric). Filtering of local FC device reconfiguration notices is common, so that remote SANs are not unnecessarily informed of local SAN device reconfiguration. Zoning local initiators to remote targets is now supported with FC bridges, so only specific devices are visible in distinct SANs. Specialized bridge software also exists that differentiates traffic type, or patterns. Disk traffic differs from tape traffic characteristics so there are specialized bridge modules to accommodate both. Most of the functionality of these specialized bridges resides at a level above the FCP that is encapsulated. The FC packets remain intact while intelligence in the bridges handle special requirements.

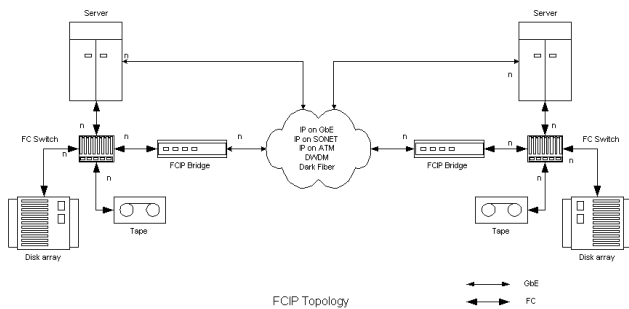


Figure 4.

### 2.3. Internet Fibre Channel Protocol (iFCP).

iFCP is a TCP/IP based protocol for interconnecting FC storage devices, or FC SANs, using an IP infrastructure to augment FC flow control and routing. iFCP was developed to provide IP for link control and routing, and TCP for congestion control, frame recovery, frame ordering, and bandwidth allocation for FC packets. All of these characteristics are especially applicable to distance related problems associated with the FCP protocol. Unlike FCIP, iFCP does not simply join distant FC SANs together. It actually provides the capability to use IP based SANs instead of FC based SANs. This has actually slowed iFCP support by FC vendors. iFCP is a block based protocol very similar to iSCSI. There is a wide range of topologies supported (Figure 5). iFCP is primarily a gateway-to-gateway protocol that allows distinct SANs to appear as one fabric. However, the iFCP protocol implemented in a IP NIC allows initiators to

attach directly into existing IP networks, and access FC targets in a FC SAN via gateways. It could also be implemented in a storage controller interface, so that target devices can be directly attached to IP networks. Another supported topology of iFCP allows IP based applications to be run over FC SANs. For example, NFS file server access can be implemented by assigning IP addresses to both client and server FC HBAs in a FC SAN. The NFS client could then mount and read/write a file system from the server across the FC links.

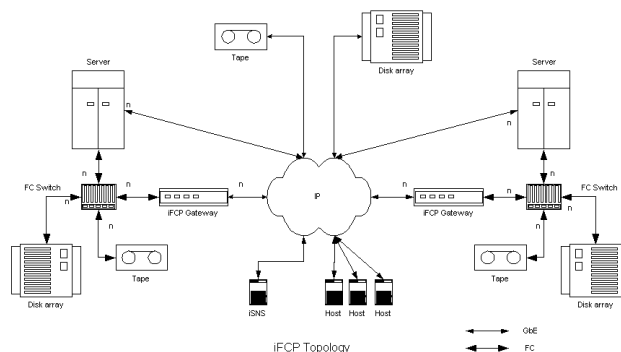


Figure 5.

FC compatibility is maintained via FC emulation at the initiator, and IP to FC translation at the gateway. This translation is implemented via lookup tables that map FC device addresses to IP device addresses. To attached FC switches, iFCP translators appear as FC switches with a shared ISL. To a connected IP based router, the iFCP translator appears as another IP router. The contents of the packets must be examined and manipulated by iFCP based devices to provide routing, traffic control, and fault tolerance.

Similar to current FCIP bridges, filtering of local FC information is provided at the iFCP gateways to minimize remote device notifications. Zoning capabilities are also supported across distinct fabrics. Unlike FCIP, congestion problems will not affect all communications between SANs, but only those that have contributed to the congestion. Device discovery and control are based upon the iSNS mechanism. As with iSNS device naming, capability, status and state changes will be maintained via a dedicated/known IP accessible server. There are also similar security issues. iFCP requires iSNS ticket handling features for authentication, and packet encryption across public networks.

The amount of processing required to handle the various protocol functions, translations, and logic dictates intelligence in the iFCP device interface. Therefore, TOE will be required for iFCP interfaces to minimize host processing requirements. This situation also occurs with iSCSI. However, there are much broader based requirements, and ongoing development, for TOE in generic NICs than for iFCP based devices. With the IP

vendors moving toward iSCSI, and the FC reluctance to support iFCP, this protocol will therefore probably fade away as iSCSI devices become more available.

The following table contrasts FC to TCP/IP storage access:

**Table 3.**

	FC	TCP/IP
Error Recovery Granularity	sequence of frames	frame
Number of Devices	15 Million per fabric	unlimited
Routing	dynamic FSPF	dynamic OSPF
SAN Security	physical, protocol isolation, password (more coming), zoning	many tools and mechanisms HW and SW, VLAN, VPN
Data Security Flow Control	HW now available credit based link level	HW & SW available end to end sliding window
Upper Layer Protocols	SCSI-3, IP,VI	TCP, SCSI-3, VI,...
Primary Packet Class	class 3, datagram, connectionless noack	connection oriented with ack
Distance Limitations	based on "droop" in flow control, flow control limits, and retransmit limits	none, designed for unreliable connections
QoS High Availability	best effort supported	selectable supported
Device Discovery & State Management	fabric SNS, RSCN	dedicated/known iSNS host, polling
	intelligence in switch, inband, network, tools becoming abundant	inband, network, tools abundant, intelligence in end points

	FC	TCP/IP
Pros	high performance, low overhead, host enabling, proven	ubiquitous, less expensive per port, lots of experienced admins, economics, interoperability
Cons	non standard protocols and devices, non-trivial learning curve, more expensive per port, interoperability	high overhead, need TOE to enable host, not proven

### 3. Conclusions

The progression in the storage industry towards serializing bus access to storage will provide HPC users with a cost effective, deep archival type storage capability. This capability will enable more online data and limit tape access wait times. The major benefit of this trend seems to be monetary so that more storage is available for less. Also, there will be more choices in storage characteristics, to better match storage to storage access patterns. With the continued march toward IP based storage solutions, the most dramatic effect to HPC users will be in data sharing. High performance storage will remain on FC devices for some time, however, various shared access options will become available via IP based protocols. The distance at which data can be shared is quickly increasing. The ability to replicate data over distance for disaster recovery purposes also helps maintain access to data.