SATA: A Better Solution for Enterprise Server Storage

Enterprise class servers are designed for more expensive forms of hard disk storage: SCSI, SAS, and SAN storage. Because of misconceptions that will be discussed below, SATA storage is usually not even considered for these servers. Similar misconceptions prefer RAID and in particular RAID 5 to protect data integrity. This study addresses these misconceptions and provides our perspectives on each, and suggests that, in fact, SATA is faster, more robust, more economical, requires less personnel overhead, and provides the best data integrity protection than the more commonly used SAN approach.
Read on.

Misconception 1: *SAN is fast, local storage*. This simply is not true. SAN and NAS share the same issues: Both are complex, with inherent delays and latencies that increase geometrically with the complexity of the network. SATA, on the other hand, has a one-to-one relationship with the HBA port and the target disk and therefore no inherent protocol latency.

Most servers will connect to a set of fibre adapters, usually connecting a smaller set of independent fibre networks. The combined bandwidth of a site's independent fibre networks is the bandwidth of the entire site, not just that of the server. A site with two completely independent fibre networks is rare, and four would be the likely upper limit. Even at this maximum, and assuming 4Gb/s networks, the server would only have to handle 16Gb/s for the entire site; even if this is focused on a single server, this is a relatively small amount of total bandwidth.

A single SATA HBA provides four 3Gb/s ports for a total of 12 Gb/s; two low-cost SATA HBAs in the same server surpass the total bandwidth even of a large complex SAN network. When you factor in the zero protocol latency of SATA, the performance of SATA compared to SAN would be considerably higher.

Misconception 2: *Enterprise servers tend to have huge amounts of data and SATA is often considered too small*. This misconception involves two assumptions that need to be considered separately.

Assumption 1: <u>Enterprise servers have much larger storage capacity than other systems</u>. Several computer generations ago, there was a storage disparity, but today that is no longer the case. My own personal computer has 4 Terabytes of storage for photographs and music. Even 100 Terabyte storage is no longer unusual. Currently SATA technology provides 20 Terabytes of storage per HBA. Linking five SATA HBAs with a single server would match the 100 Terabyte goal of a large enterprise server.

Assumption 2: <u>SAN is local storage.</u> Our attitude is that SAN is not local storage. In fact, it is rare for a server to actually have more than a few hundred gigabytes directly attached to the server.

Misconception 3: *a)SAN is really easy to manage; b) most corporations have whole teams dedicated to managing it.* This self-contradictory statement is made with surprising frequency. SAN networks can become extremely complex, and anything that requires a team to manage also requires a large budget. There may be situations that require this setup, but often SAN systems are much more expensive and complex than the customer needs. The current trend to almost blindly recommend SAN storage to new customers really does the customer a disservice.

Misconception 4: *SATA cannot be virtualized.* Not so. Our SATA is now running on IBM's VIO server and provides a level of functionality equal to any other SCSI, SAS, or fibre-based disk. Alternatively, with lower-cost HBAs, each LPAR can be provided with HBAs and storage directly, rather than going through the VIO server. Indeed, for ISPs wishing to provide independent machines for some of their customers, an LPAR provides this at the cost of disk storage.

It makes sense to make that disk storage as inexpensive as possible, don't you think?

Misconception 5: *SATA can not be used with HACMP.* Eleven years ago, in 1997, Ease Software implemented the first MPIO solutions for AIX by providing EMC with the drivers needed to connect their disk systems to AIX. The key to that solution was simply to provide a port selector.

SATA specified a port selector back in 2005, and is now also available with port multipliers. The port selector allows one host to access a bank of four SATA drives and, with a sub-second protocol command, switch and allow access from the backup or redundant host when the first host fails. This is tested, current technology ready to be deployed.

Misconception 6: *SATA drives are not enterprise quality.* Today, almost the opposite is true. Because the PC market is so intense, most of the research dollars spent today are on low-cost SATA drives. For example, consider research done on the Seagate Barracuda ES line of drives addressing the issue that in large racks of disks, vibrations from other disks interfere with the accessing of data on a given disk. This was true for all disk types, and, according to Seagate documentation, the performance of a disk in that environment may be reduced by as much as half.

The Barracuda ES line of drives uses active feedback to counteract this interference. Much like active noise suppression in headphones, the active feedback counteracts the vibrations allowing the drive to perform nearly at 100% even when placed in a large rack with other drives.

The point here is not how wonderful Seagate or the Barracuda ES drives are; it's that Seagate and other disk manufactures are placing their research dollars in SATA disks. The enterprise disk market is rather puny compared to the robust consumer disk market, making Enterprise disks a follower rather than a leader in technological innovation. Using SATA drives leverages this research into the more demanding markets of enterprise servers.

Actually, most SAN storage solutions now use SATA disks in order to reduce costs. Enterprise systems have been using non-SCSI disks for a few years now. But, as we mentioned at the beginning, using SATA in a SAN solution forfeits both the speed advantage as well as most of the cost advantage. The key is to use SATA in a pure SATA solution, so that all of its advantages can be utilized.

Actually, the main virtue of SAS is its similarity to SATA: currently, drives are usually sold with both SATA and SAS interfaces. Unfortunately, SAS does not have SATA's volume of sales, and so can never compete in terms of price. At best, the extra expense only purchases a few rarely used features not found in SATA. Almost always, when the customer's needs are analyzed, SAS rarely has any advantage at all.

Misconception 7: *We need RAID 5*. RAID 5 was an elegant disk solution allowing a single drive to fail and still maintain full data integrity. A RAID 5 system used a number of disks all the same size, say four 250G disks, the usable capacity of which will always be one less than the number of disks times the size of the disk: in our example, three times 250G or 750G. As the number of disks increase, the same math applies. For example, a RAID 5 system with eight 250G disks will have a capacity of seven times 250G or 1750G of space. The "extra" disk is a parity disk.

The downside of this approach is access times. A write to the RAID 5 system requires a read, modify, write sequence, because the surplus disk is a parity disk the contents of which must be recomputed and stored after each modification. And, while we have the ability to reconstruct the data if a single disk fails, we still have only one copy of each data block. So a read of a particular block must go to a single source.

RAID 5 is a common engineering compromise. In this case, RAID trades data security for performance. Back when disks drives cost thousands of dollars each, this compromise made sense. Now, when a 250G drive costs $80, it's silly.

Our data integrity approach is to simply use striping along with mirroring. (RAID 0 with RAID 1 or sometimes called RAID 0+1 or RAID 10.) There are several advantages to this approach.

1. A write now puts the same data block in parallel in two different locations. The CPU overhead to set up these writes is trivial.

2. Because all the data is replicated, the server can fetch any block from a choice of two locations. With proper software, the load of accesses can be balanced and the total read throughput increased almost by a factor of two.

3. All data is duplicated, a great win in various ways. For example, an instantaneous snapshot of the data can be taken by "breaking" the mirror. This stops the writes from going to half of the disks. These disks are then backed up via regular means. Later, the mirror is put back in place and AIX will automatically resynchronize the two sides, allowing a consistent backup with the server still running.

4. When a disk does go bad, all other data is still fully redundant. For example, a disk system with eight drives would have two sets of four drives. (Or four pairs of drives). If one drive goes down, that will affect only one pair. The other three pairs are still fully redundant.

5. When a fresh disk replaces the failed one, the system will automatically resynchronize the disks. In the case of RAID 5, this is a very complex process of reading all of the data from all of the other disks in order to compute the proper parity. In the case of RAID 10, it is simply copying one disk image to its mirror.

Misconception 8: *Mirroring is too expensive*. Obviously mirroring doubles the cost of the disk system, but disks are now cheap, so the end cost of a mirrored SATA disk system will still be much lower than any other alternatives we are aware of. Added to the other advantages, of much higher throughput, simpler management, and leading-edge technology, and we think the choice is clear.

We believe we have demonstrated that SATA storage has many benefits and few flaws when compared to SCSI, SAS, SAN, and NAS storage. SATA can be used very effectively in many enterprise class servers. Many of the "knee-jerk" fears that are common with potential customers have been addressed and, we believe, addressed successfully. It may mean a rethinking of current practices. Rather than pooling all the storage in one central disk system there are many advantages to putting twenty to a hundred terabytes of storage directly on each server. Not the least of these advantages is speed. SATA will allow this quantum increase in speed at a reduced overall cost.

Ease Software, Inc.
P. O. Box 201388
Austin, TX 78720-1388
Phone: 877-EASE-AIX
(877-327-3249)
www.easesoftware.com