

All about the Benjamins

Written by Max Kalashnikov

Tuesday, June 09, 2009 - Last Updated Monday, June 08, 2009

The choice of the unit of measure of storage is interesting to me because it's otherwise tough to measure price for performance.

I remain agape at the price tag on high-end, supposedly high-performance, storage systems. Connected by FibreChannel or gigabit Ethernet, that's a limit of 400 and 110 MB/s, respectively. (Yes, I know of 8Gb/s FC and 10GE, but these are prohibitively expensive, if supported. Even link-aggregated GigE practically tops out at 880MB/s) I'm thinking that writes across 40 7200RPM disks could saturate an FC link, and it would take fewer than 20 15k disks. Neither of these strikes me as impractical or unusual sizes of storage arrays, even doubling those numbers for RAID 1. More importantly, such arrays don't strike me as high performance.

Particularly shocking is that a brand name "SAN" solution of such a size would cost in the neighborhood of a quarter million dollars and be at its performance limit. Granted, it might be half that price without fancy management and replication software. whereas the less fancy alternative, at one tenth to one fifth the cost, would still be expandable from a performance standpoint. How much does the Veritas database suite cost these days?

The cheaper alternative, which I have implemented and benchmarked, is using Serial-Attached SCSI (SAS) instead of FibreChannel and commodity SATA disks instead of 10k or 15k spindles. Although it's not necessarily "SAN" in the marketing sense, SAS readily supports multiple hosts per bus. It's also typically implemented as 4x 300MB/s channels on one connector for interfacing to expanders (a rough equivalent to FC switches). An x4 PCIe slot is actually the limiting throughput factor for one of these, as each x1 lane is only 250MB/s. Even with RAID1, rolling my own array would cost \$25k (including labor), maybe double that for Dell brand MD1000s. One could then spend twice again the same amount to get triple the throughput on the same server(s), before running up against the limit. Additional fanciness can be gained from 3rd-party storage software vendors, especially in this economy, for under 6 figures.

That's for truly random I/O. For sequential I/O, such as for logs, the situation is even more egregious: only 4 7.2k spindles would saturate a (dedicated) FC link. If it's paired for redundancy, one would need a second pair for the non-sequential, perhaps introducing some management complexity, unless FC link aggregation becomes common enough to be standardized.

All about the Benjamins

Written by Max Kalashnikov

Tuesday, June 09, 2009 - Last Updated Monday, June 08, 2009

Another issue I've had come up in conversation is reliability and/or maintenance. This [Usenix paper](#) belies the notion that SATA disks are any less reliable than others. With a 3-6% annualized replacement rate, that's 2-5 disks per year, or about 15% or 12 disks over 2.5 years, on an 80-disk array. I've actually already included this (4 spares per 20 non-spares) in the \$25k above.

Somewhere between 2 and 3 years, you're going to have to bite the bullet, spend another \$25k for twice as much space, and migrate the old data, assuming you're not already upgrading for other reasons. Woe is you. You'll just have to resort to drowning your sorrows in the hundreds of grand you saved, never mind the headache of shipping disks back and forth.

The Storage Emperor's new clothes are looking mighty skimpy, indeed.