Wednesday, February 21. 2007

Common wisdom in rotating rust

In IT many decisions are based on common wisdom. In the field of storage for example: SATA fails more often than FC is the classic one. Or: After a burnin disks runs almost without problems before their crash years leater.
Bianca Schroeder wrote the paper "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?" for the fifth USENIX Conference on File and Storage Technologies.Especially as an avid fan of the Sun Fire X4500 i´ve got much critics for it´s SATA disks as more error-prone than FC-variants. But:In our data sets, the replacement rates of SATA disks are not worse than the replacement rates of SCSI or FC disks. This may indicate that disk-independent factors, such as operating conditions, usage and environmental factors, affect replacement rates more than component specific factorsFurthermore she obliterates the myth that harddisks have a steady failure rate after burnin. The reality is that after they wear out and the discs fail more frequent with time. For me this leads to an interesting consideration. It may be nescessary to replace all discs at certain intervals when you want to ensure, that the increasing failure rate doesn´t lead to a time between failures smaller than the time needed to resync the discs in a RAID-5 set.
Robin Harris of StorageMojo comes in "Everything You Know About Disks Is Wrong"to another, but related conclusion: One implication of Schroeder’s results is that big iron arrays only appear more reliable. How? Using smaller “enterprise” drives means that rebuilds take less time. That makes RAID 5 failures due to the loss of a second disk less likely. So array vendors not only get higher margins from smaller enterprise disks, they also get higher perceived reliability under RAID 5, for which they also charge more money.
So the higher reliability of the large storage arrays may be only a pseudo correlation. Based on this data i would tend to think similar with one important limitation: "as long you talk about the dics" because the big iron storage often protects you better against failures of the electronics of the enclosure as well by redundant caches of raid controller or redundant raid controller as well.
After all, one of my assumptions was confirmed: MTBF numbers seems to be bullshit ... or a little bit more sophisticated:For drives less than five years old, field replacement rates were larger than what the datasheet MTTF suggested by a factor of 2-10. For five to eight year old drives, field replacement rates were a factor of 30 higher than what the datasheet MTTF suggested. At any case, the paper of Mrs. Schroeder is a must read for everybody in professional contact with hard discs.

Posted by Joerg Moellenkamp in English at 02:24

I also recommend reading http://labs.google.com/papers/disk_failures.html
Anonymous on Feb 21 2007, 10:20