Data Storage Reliability and Availability

Silverton Consulting, Inc.
StorInt™ Briefing

Introduction

Today's CIO has an ever-changing myriad of products to choose from to optimize his company's data center. Generally large, sophisticated data centers have an intricate, yet integrated, mix of all the technologies available. That is, advanced data centers typically have a blend of server, network and storage capabilities. Such an amalgamation can provide financial advantages but can also result in optimum reliability and availability. Even so, this superior performance often depends heavily on the underlying storage. Optimizing the storage media's effectiveness as to reliability and availability requires careful consideration of the type and criticalness of the stored data and how different storage media can alleviate potential problems.

Storage technology reliability and availability

Historically and even now, data storage has been housed primarily on FC or SAS disks and/or tape. Additionally, SATA disks have secured a firm foothold and are competing with tape for certain applications. Moreover, newer technology like solid state drives (SSDs) are edging into the storage marketplace. However, all of these technologies have characteristic failure modes including:

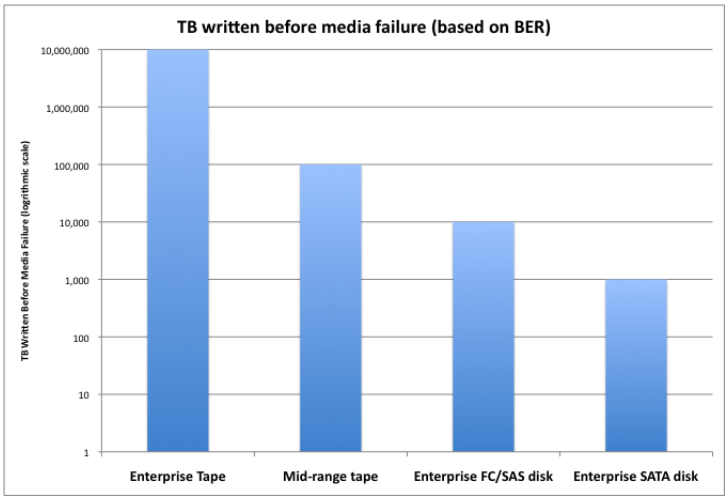
- **Data reliability fault** – defined as the inability to write data to storage media and read that data back without failure. For storage technology this is called Bit Error Rates (BER).
- **Data availability fault**– defined as the inability to still access data when a storage error occurs. Most data availability metrics are system specific, i.e., disk and tape systems each use a distinct measure.

Data reliability

For decades BER has been used to describe raw and correctable error rates for different types of storage media. Generally, this is stated as the number of bits that could be written before an error occurs on read back. BER usually varies significantly by storage technology.

As can clearly be seen in Figure 1 below, enterprise tape reliability is an impressive 1,000 times more reliable than enterprise disk; midrange tape is decisively 10 times more reliable than enterprise disk. As compared to SATA disks, both enterprise and midrange tapes are even more reliable 10,000 and 100 times more reliable than

SATA disk. Thus, using this parameter alone, both enterprise and midrange tape technology would be an attractive storage alternative.



Noticeably, Figure 1 does not provide a comparison of SSD's reliability to the other storage media because NAND reliability characteristics are more complex and currently depend on the technology (MLC or SLC)[1] used and the program/erase cycles expended[2]. However, NAND based SSD BERs generally range from 1,000TB (comparable to SATA disks) down to 100TB.

Figure 1 Storage media TB written to failure

Storage system data availability

As discussed above, the comparison of the various storage media using a data availability parameter is difficult at best. In fact, these metrics vary considerably depending on the storage system design and how the storage media is actually used. At the storage system level, data availability is measured by uptime and is expressed as MTTF (mean time to failure).

Disk data availability

Usually, disk predominant storage systems are designed to employ a combination of redundant hardware, fault tolerant software/firmware, and drive RAID protection. With such complicated systems it's difficult to calculate just how the disk drive reliability impacts the whole system availability but RAID levels are normally the weakest link.

Nonetheless, a number of considerations are used to help determine RAID availability and thus, system availability including:

- **RAID protection level –** Some RAID levels support a single drive failure (e.g., RAID1 or RAID5), some support dual drive failures (RAID6 or RAID-DP), while others can go ever higher but today, most support either a single or dual drive failure.

[1] From http://research.microsoft.com/en-us/um/people/maheshba/papers/hotstorage09-raid.pdf as of 12Apr2012
[2] From http://research.microsoft.com/en-us/um/people/maheshba/talks/eurosys-diffraid.pptx as of 12Apr2012

- **Rebuild time -** This is the duration of the activity to recreate data lost due to a drive failure onto another spare drive. Necessarily, rebuild times rise as disk capacities increase. This measurement is particularly relevant, as a longer rebuild time would expose the entire storage system to data unavailability.
- **Spare drive availability –** Most disk storage systems using RAID protection, support local or global spare drive(s) to hold the data being reconstructed. In addition, for every RAID group at least one local spare disk is usually available to house the reconstructed data arising from the rebuild process. The actual number of spare drives is very system design dependent.

Given these variables, disk storage system availability becomes a complex calculation based on disk drive availability, the probability of another drive failing while a drive (or two) is being rebuilt, and spare availability.[3] The resultant measure is usually specified as system MTTF (mean time to failure).

However, disk MTTF, as compared to system MTTF, is also a function of disk media reliability (BER) and disk hardware/firmware reliability. For instance, enterprise FC/SAS disk drives typically offer a high MTTF of 1.6 million hours and enterprise SATA typically support a MTTF an order of magnitude lower. But, in contrast to tape, disk is not removable and as such, needs many more drives to support equivalent capacity. For example, with a 1000 FC disk drive system, supporting ~600TB raw capacity, the storage will have an disk drive failure every 1,600 hours or **every ~2.2 months**, but with proper RAID grouping that disk failure should not result in a system data availability fault.

Tape Data Availability

Although tape availability is also designated as MTTF, direct comparison with disk MTTF is difficult because of tape's inherent removability and portability. Unlike disk failures, tape failures usually result in no damage to the media and thus, can be easily read on another tape drive.

Tape system data availability is a function of tape drive availability and tape media BER. Midrange (LTO) tape drives have a MTTF of 250,000 hours and enterprise tape drives are even better. Tape library availability can also be a critical factor in determining tape system MTTF. Library robotics experience MTTF similar to tape drives or ~250,000 hours. However, due to the removability of tape, one needs fewer tape drives to support any given capacity. For instance, 1PB of tape storage could be supported by only 8 tape drives which would have a drive failure **every ~3.6 years,** but by moving the media to another drive, that drive failure should not result in tape data un-availability

[3] From http://www.cs.cmu.edu/~garth/RAIDpaper/Patterson88.pdf as of 13Apr2012.

Summary

Clearly, both enterprise and mid-range tape technology is inherently more reliable and more available than pure disk only storage. However, IT needs both high availability and high performance, which cuts across all server, networking and storage technologies. Given those requirements, a combination of technologies using both disk and tape storage could provide an optimum solution. Indeed an archive system combining disk (or SSD's) for frequently accessed data and tape storage as a permanent repository for infrequently accessed data could provide peak performance while maintaining high data availability.