MAINFRAME TAPE: A TECHNOLOGY FOR THE NEW MILLENNIUM

BY

JON WILLIAM TOIGO

CEO TOIGO PARTNERS INTERNATIONAL
CHAIRMAN DATA MANAGEMENT INSTITUTE

SUMMARY

Tape technology, like the mainframe itself, has been proclaimed a “legacy technology” by analysts and pundits over the past decade. However, tape continues to thrive and to deliver value to organizations that use it. This is in part because of enhancements made to the technology itself in terms of operational performance, capacity and reliability. But it also reflects a steady improvement in disk technologies that have been harnessed to improve tape and to provide additional value.
INTRODUCTION

To some readers, the characterization of tape technology as a Millennium Technology may seem oxymoronical. Over the past decade, tape has been increasingly dismissed by analysts and pundits who see it as a legacy technology soon to be relegated to the dustbin of history.

IT planners may still recall the many issues cited by analysts a decade ago: that tape resiliency is sub par and that tape-based data restores are both slow and unreliable, or that tape technology isn’t keeping pace with the performance or capacity of its cousin magnetic storage technology, the hard disk drive array. These observations, which were of questionable veracity at the time they were originally made, do not jibe with tape technology as it exists today. Not only has tape undergone significant improvements over the past ten years, significant industry effort has been undertaken to complement tape technology with disk technology – yielding so-called “virtual tape” subsystems that have helped sustain the value of the technology as the preferred modality for data protection and archive.

The core value of tape remains unchanged:

1. Tape provides a portable medium for data protection. Data recorded to tape as a backup can be removed from the data center and stored at an alternate location, providing an “air gap” between data itself and those events that would compromise it – especially the 95% of data corruption events that have nothing to do with smoke and rubble facility disasters.

2. Portability also extends to hardware agnosticism. While data may be recorded to tape from disk arrays branded by a particular vendor, or in a particular configuration, tape-based data need not be written back to the same vendor’s gear or hardware configuration. This gives tape a huge advantage in a disaster situation, where identical equipment may not be available for data restore.

3. Tape is also cost-effective: offering high capacity storage at very low cost per GB. Unlike disk, tape has continued to demonstrate an affordability curve that aligns with the falling cost of commodity media. And, from the perspective of both facility space and power demand requirements, the cost of tape is a fraction of comparable disk storage.

Better protection, hardware agnosticism and cost-efficiency give tape certain advantages as a backup medium, and increasingly as an archival medium as resiliency has improved 700% by conservative estimates over the last five years. Bottom line: the future of tape remains bright.
Capacity Improvements are Ongoing

Some analysts continue to raise concerns that tape technology is not keeping pace with disk technology. However, this position does not dovetail with current improvements in both technologies.

In the disk world, from which tape engineering borrows much of its core technology, steady improvements have been seen, including a doubling of capacity (measured as Gb per Inch²) about every 18 months. Technologies such as Gigantic Magnetoresistive Read/Write Heads and Partial Response Maximum Likelihood (PRML) algorithms, and more lately Perpendicular Magnetic Recording (PMR), have helped to give disk its capacity growth. In the next 36 months, disk drives could make their next capacity jump owing to various combinations of bit patterned media coatings and heat-assisted recording.

However, as previously stated, tape continues to grow its capacity as well – often borrowing technologies developed for disk. Currently, PMR technology is being combined with new Barrium Ferrite tape coatings that will deliver tape cartridges starting at 35 TB. IBM and FujiFilm demonstrated the viability of this new technology in early 2010.
If capacity can no longer be referenced as a deficit of tape, neither can performance. The latest generation of tape libraries features multiple robots capable of delivering in excess of 121 TB per hour in throughput. That far exceeds the performance of disk arrays, even with extensive numbers of “short-stroked” drives.

Bottom line: while there will likely always be a game of “leap frog” in magnetic recording, the fundamental advantage of tape remains as a low-cost, high performance, high capacity medium for data storage. This partly explains why tape is now being wedded by vendors to a file system in order to create archival storage repositories that can help firms cope with their file-based data burgeon. Shortly, tape may play an operational role again comparable to its role as primary storage in pre-1970 data center environments.

What About Always-On Applications?

Of course, restoring data from a tape backup remains a time consuming and labor intensive process. Significant improvements have been made in the resiliency of the technology both at the media and library automation level that has led to a reduction in the likelihood of data restoration failures. That said, however, tape-based data restore speeds are still a function of target infrastructure and interconnect technologies. Tape can deliver data at streaming rates far in excess of the write or transport capabilities contemporary disk drives and networks, but the overall performance of data restoral is limited by the overall system rather than by tape streaming rates.
The problem is that some applications cannot wait for tape-based data restore following an interruption event. They sport “always on” time to data requirements that may require near real time data replication between equipment hosted at geographically dispersed facilities. Disk-to-disk data replication via a Wide Area Network has been successfully applied in these cases, which were once limited to the financial and banking sector but that now include a broader range of applications in all industry sectors.

Some analysts have taken this subset of application requirements as a harbinger of the demise of tape. If companies are deploying WAN-based disk-to-disk replication methods for some application data, the success of this method will eventually displace tape altogether, they argue.

This argument errs on several practical points. For one, WANs are expensive and their transport capacities tend to be quite minimal. Using an expensive data replication technique for data and applications that do not require “always on” is rarely cost-justifiable. Adding to the costs of the solution are a combination of other factors including the requirement that source and target gear must usually be of the same brand and configuration (contributing to cost and management requirements) and the additional requirement that investments must be made in redundant facilities and equipment that in the best of circumstances may never need to be used (few organizations can afford a full redundancy strategy to protect against the 5% likelihood of a facility failure).

As a best practice, data protection capabilities should match the restoral priority of an application and its data as determined by an assessment of the criticality of the business process that they serve. Mission critical applications may need an “always on” service; most applications do not require such services.
AUGMENTING TAPE WITH VTS

The above logic does not extend to local disk-based copies of data. Disk offers one thing that tape does not: faster access to discrete data files. Writing a backup set to local disk enables a firm to access a safety copy of specific data sets or files that may have become corrupted by machine or human processes and to restore these files for use in a more timely way than would be the case with tape. The rotational nature of disk has its advantages.

This has led to the widespread adoption of a range of technologies that represent themselves as virtual tape subsystems (VTS) or virtual tape libraries (VTL). Since the meaning of VTS/VTL has undergone significant changes over the years, it is useful to quickly summarize their evolution.

The VTS/VTL was originally introduced as a disk-based buffer intended to provide a temporary location where data could be staged prior to writing it to tape -- mainly as a tape media cost-savings measure. Mainframes were notorious for their inefficient use of tape media capacity in early libraries, writing small data sets to tape, then ejecting the cartridge and writing small data sets to the next piece of media. The idea of the first VTS was to stage backup data on disk in order to aggregate multiple small jobs so they would fully fill a tape cartridge – reducing media costs.

The second generation of VTS technology sought to address slow tape backup performance by emulating many tape drives using software and disk, providing multiple targets for concurrent backup streams, usually in excess of the number of physical drives actually available in a physical library. Once staged to disk, writing backup sets to tape was accomplished in a predictable way (in terms of speed) and in a way that would not impact normal application performance.
Reducing media costs and finessing operational constraints may have defined the first and second generation VTS, but the current generation subsystems have been targeted for a broader set of functionality. As a solid data platform, vendors have sought to make the VTS a location from which discrete backup data sets can be restored quickly in the event of a corruption event and also as a location where additional services can be applied to data such as encryption, de-duplication, indexing, validation, and WAN-based replication.

“Value-add” functionality did much to increase the appeal of VTS and disk-to-disk-to-tape strategies generally. Unfortunately, it also provided more credence to vendors promoting a “diskless backup” strategy – code for the elimination of tape and the replication of all data over a WAN. The deficits of such a solution are not theoretical, but empirical as one recent case underscores.

**TAPELESS BACKUP? A FACT-BASED ANALYSIS**

Planners at a major telco made a decision in 2008 to use a de-duplicating VTS to replace its tape subsystems. As a telco with low cost access to the interexchange carrier network, the issue of WAN bandwidth cost was less important than it might be for those of us who must pay full price for WAN services, so a strategy was formulated to backup data to disk targets – a VTS – then to reduce the data using de-duplication services on that platform prior to replicating it across the
network pipe to the disaster recovery data center. It was believed that this approach would be less expensive than using traditional tape backup and would provide greater efficiency by eliminating manual processes associated with tape.

Planners soon discovered that the strategy was not as efficient as originally thought. For one thing, that the capacity of the VTS platform (a specialized appliance with its own disk) that would be used to cache backup data was not horizontally scalable. When the capacity of one VTS appliance was fully leveraged, another needed to be installed.

That wasn’t cheap. By adding value – in the form of VTS and de-duplication software -- to what would otherwise be described as a box of commodity disk drives, the vendor charged the customer $10,000 per TB for disk that could be purchased for less than $100 per TB. Even with the telco’s volume purchasing discounts, this mark-up seemed a bit exorbitant. Moreover, since the appliances did not scale, the telco discovered that it needed to stand up additional boxes and to hire additional staff to manage each stovepipe array, increasing labor costs.

Additionally, the IT planners reported that they hadn’t realized how much work would be needed to classify data prior to using the appliance at all. The vendor’s promised de-duplication/compression ratios weren’t realized in the case of most of the data that the telco was backing up. Thus, planners reported that they needed to segregate their data prior to writing them to the VTS so that only the data that would de-duplicate to the promised ratios occupied the drives.

The planners noted that other data also needed to be segregated from de-dupe because of concerns about the acceptability of de-duplicated data to SEC regulators. The SEC requires publicly traded firms to submit certain data on a routine basis and specifies that the data must be in a full and unaltered state. To date, there hasn’t been a court case or administrative hearing to decide whether de-duplication materially alters data. The telco had decided that it didn’t want to become a test case.

The lessons learned by the telco are summarized here and help to frame a fact-based discussion of the merits of tapeless backup, instead of another noisy contest of opinions around tape versus disk.

1. De-dupe can be an effective service, especially in WAN-based replication scenarios where reducing payload size can be beneficial to WAN service optimization. However, with all uses of de-duplication today, reduction ratios vary widely by data type. This underscores the broader point that you must have a highly granular understanding of your data and its business context before you select or apply any data protection methodology – both to ensure the right assets are being protected and also to get the best mileage out of the data protection technology you are using.
2. Leveraging any data protection technology requires granular consideration of data itself and its business context (criticality factors, regulatory factors and cost factors). There is no “one size fits most” data protection strategy.

3. WAN-based replication not appropriate for all data based on criticality, volume and cost of repository, services and bandwidth: recovery service level objectives define protection options. Sometimes tape is the appropriate choice.

4. Proprietary protection service delivery mechanisms are generally a cost-accelerator in DR/BC plans. Replication must usually be performed between two products from the same hardware vendor. Moreover, scalability issues can drive up labor costs.

In the end, the telco rededicated itself to tape-based data protection for all but its most mission critical applications. They may have reached this conclusion more quickly had they not enjoyed low cost access to high bandwidth networks. In another case, involving a financial institution that was interested in eliminating tape in favor of disk, the following costs were identified in conjunction with a strategy that the vendor called “multi-hop mirroring.”

The vendor suggested to the financial firm to purchase three copies of its storage array, two for local synchronous mirroring and a third for remote asynchronous mirroring. Additionally, they recommended that the client make system images every couple of hours to enable data restoral to specific points in time – protection against a fault in its transaction-oriented operations. Specialized software, called point-in-time mirror splitting, was required for that functionality.

The total cost in hardware alone was $1.3 million. Software costs added another quarter million to the solution, and WAN costs ranged from $500 per month for T-1 speed interconnects to $45,000 per month for OC-3 pipes – and significantly more for bigger pipes!
Adding in other cost-of-ownership and recurring cost items, like warranty and maintenance agreements, labor and administration expense and energy costs, planners were soon looking at a significant expense – several orders of magnitude more than what tape backup would cost the financial firm.

Finally, it should be noted that the absence of tape in the solution compromised data protection significantly given that most database corruption events are not discovered for 24 to 48 hours after they occur. By then, all of the split image copies of data – both local and remotely replicated – would likely be unusable. A recent tape backup might be the only solution.
CONCLUSION

On its own, tape technology still has an excellent story to tell as a data protection strategy. Performance, capacity, reliability and cost characteristics are keeping pace with the needs of most data protection and data recovery needs in most business applications today.

Tape technology can be effectively augmented by disk, especially in a mainframe environment, provided that the strategy is appropriate to the business process and applications that are being served. If you are considering a VTS to augment your mainframe tape strategy, the following criteria should be considered.

1. To optimize resource utilization, the VTS product that you consider should take advantage of specialty processor engines. “zIIP eligibility” helps to isolate value-add work performed on the data hosted on the VTS from production work.

2. The VTS should provide capabilities to customize storage groups and to segregate data based on business criteria. That way, you can mirror data that you need to mirror to support “always on” applications, while applying appropriate services and tape backup for the remainder.

3. NFS streaming support is valuable, since it will enable the broadest connectivity to data and the greatest flexibility in data replication.

4. The VTS itself should be manageable via zOS features, including SMS DataClass management of DASD buffer disk. This capitalizes on management skills and tools you already have and enables you to use existing resources better.

5. The VTS should be capable of using any size buffer, providing a fit with your existing processor, memory and disk resources.

6. The VTS should feature tight integration with backup, security, key management and encryption systems you already have.

7. The VTS should be hardware agnostic so that it works with existing DASD rather than requiring the purchase of overpriced stovepipe array-appliances.

8. The VTS should work with the broadest number of data protection scenarios. For example, it should optimize tape operations, while supporting WAN-based replication processes for those applications that need them.
As suggested by the illustration above, there are often a number of data protection requirements in a contemporary enterprise. Neither disk-based nor tape-based data protection strategies provide an exclusive solution. In truth, a combination of both disk and tape are generally required to meet business continuity needs.

It is time to end the tape-versus-disk debate once and for all and to use the right technology for the right data based on the careful consideration of business factors.