The Era of Colossal Content Arrives

Fred Moore, President
Horison, Inc.
www.horison.com

Introduction

Did you know that data is growing just over 40% annually and it is estimated that nearly 90% of the data in the world today was created within the past two years? The vast majority of this data reaches tier 3 or archival status in a relatively short period of time. The unprecedented growth of tier 3 digital data such as fixed content, and compliance requirements are demanding significantly more robust, cost effective, secure, and scalable storage solutions than ever before and mandating that new architectures arrive. Tape and disk have historically combined to meet most of the archive demand. However, the numerous enhancements and significant improvements to tape drives, media and libraries that have occurred in the past few years have clearly positioned tape technology as the optimal choice for tier 3 data. The primary focus of this report will be the tier 3 layer of the storage hierarchy.

Unprecedented Demand - Moving Beyond the Big Data Phase into the Era of Colossal Content

Tier 3 storage in the Big Data phase are experiencing a CAGR (Compounded Annual Growth Rate) of 60% annually and include digital archives, fixed content, digital images, multi-media, video, social networks, compliance data and are the fastest growing data category. This growth is being generated by billions of people, in addition to businesses, mandating the emergence of a markedly more effective archiving solution as the need for a repository that makes data easier to find while lowering the cost of ownership increases daily. Just ten years ago, large businesses generated roughly 90% of the world's digital data. Today it is estimated that 75-80% of all digital data is generated by individuals - not by large businesses – however the majority of this data will eventually wind up back in a large business or service provider. This shift in the primary source of data generation is the *fundamental reason* driving the Big Data phase of extensive business analytics and the coming Era of Colossal Content, which will see data growth rates approach and ultimately exceed 100% annually. The chart below highlights what we might expect as the Era of Colossal Content approaches.

Comparing Storage Demand Then and Now

	Then	2012	Future
% of Data Born Digital	1990: 8%	90%	90-95%
New Data Created Each Year	2000: 35-40%	~60%	75-100%
% Structured Data Created	2000: 75%	10%	5-10%
% Unstructured Data Created	2000: 25%	90%	90-95%
# of Photos Uploaded to Facebook Each Month	2000: 0	3 billion	4-5 billion
Daily Video Views on You-Tube	2000: 0	2 billion	>3 billion
Amount of data generated by individuals (vs. businesses)	2000: 10%	75%	80-90%

Source: Horison, Inc.

The Rapid Rise of Fixed Content and Archive

Much of the fixed content, compliance and archive data that makes up tier 3 is referred to unstructured data meaning that naming conventions, a taxonomy, tags, search keys, and indexing capabilities are not present. Most unstructured data is stored as file storage, and industry estimates indicate that over 70% of all digital data is stored in an unstructured format. Delivering an acceptable access time to this data poses special challenges for large archival storage systems, which must rather quickly locate specific files from very large archives.

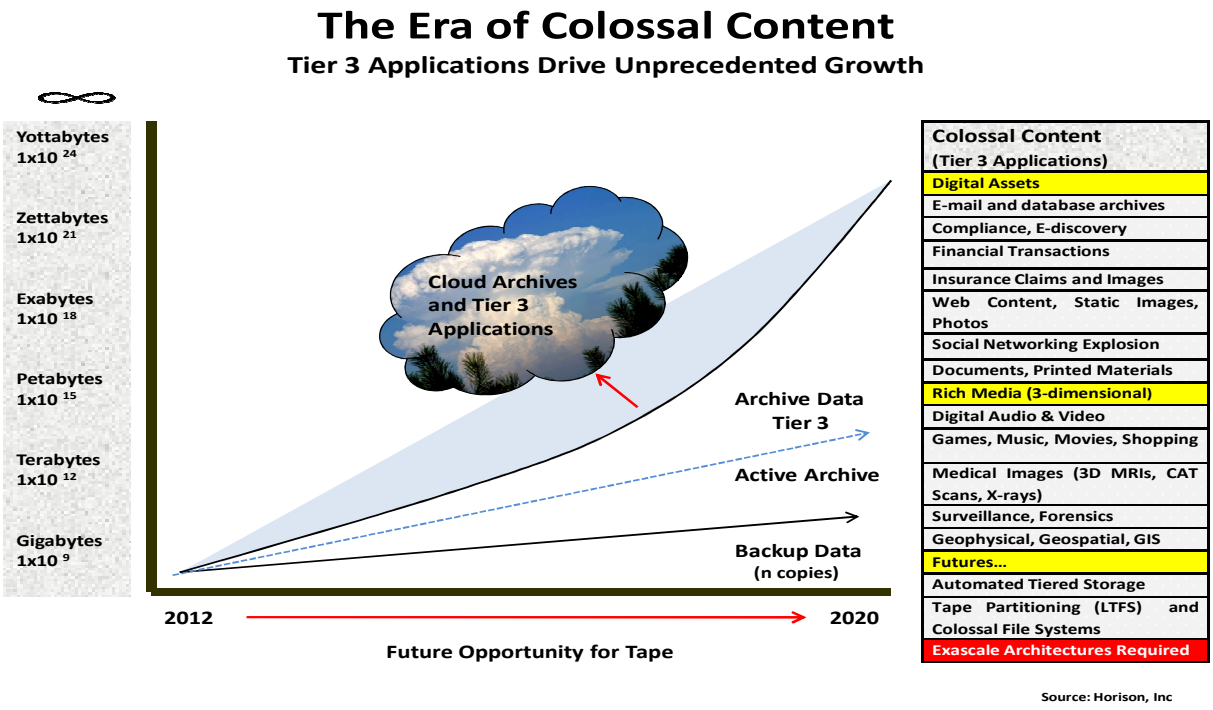
Colossal Content Trends

A growing list of government and legal regulations worldwide now describe the way data must be managed and stored and protected throughout its lifetime. For many applications and data files, the lifetime requirement for data preservation has become infinite as this data will *never be deleted*. As an example, the retention period for medical records such as X-rays and MRI images generally need to be kept for the lifetime of the patient though they may be rarely referenced after they are generated. Largely driven by individuals, the social networking wave has quickly become a major generator of massive amounts of low-activity unstructured digital data that is kept (archived) indefinitely. Most video, television programming, sports events, movies, and scientific data will be kept forever and for most of this data, frequency of access will steadily decline over time.

Electronic discovery (E-discovery) refers to any process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case. Much of the data used for E-discovery is physically stored as archival data, normally used infrequently, but is used more frequently during the E-discovery search process making it an ideal application for tape archives. Internet click-streams, mobile data rich with geospatial information, remote sensing data, and social-network content are among the many forms of information ushering in the Colossal Content Era. High performance data analytics capabilities will be needed to examine massive amounts of data to uncover

hidden patterns, unknown correlations and other useful information in order to gain competitive advantage.

For some organizations, facing hundreds of terabytes or several petabytes of archive data for the first time may trigger a need to redesign their entire storage infrastructure. Primary drivers of tier 3 storage requirements are listed in the "Era of Colossal Content" chart below. All of this data will have to be stored somewhere. Do you have a game plan for building a scalable, cost effective storage infrastructure to manage the tremendous growth of tier 3 archival data that lies ahead?



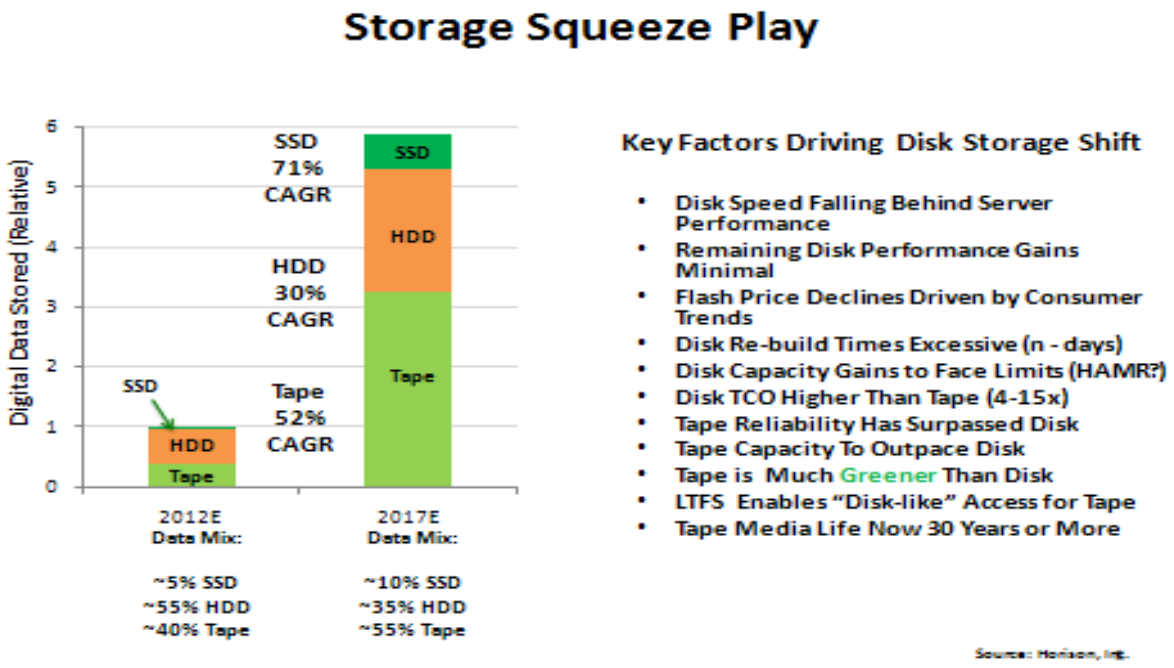
Key Points: *Archive, fixed content, compliance, E-discovery, entertainment, video, scientific, social networks, data analytics and unstructured data requirements are soaring defining the Era of Colossal Content and have become the primary drivers for future archival storage demand. Much of this growth will reside on tape technology as tape is expected to become the primary tier 3 technology.*

The Storage Squeeze Play – Critical Storage Issues Surface as Data Growth Escalates

Tiered storage is a data storage environment consisting of two or more kinds of storage, delineated by differences in four primary attributes: price, performance, capacity and functionality. To implement an optimal data lifecycle management strategy from a technology perspective, the industry de-facto standard four-tiered storage hierarchy model is used and consists of storage tiers 0, 1, 2, and 3. Each tier has its own set of attributes ranging from the highest performance and most expensive technology (tier 0) to the largest capacity and lowest cost (tier 3). These tiers are further explained later in the Looking Inside the Storage Tiers section.
Technology roadmaps and technology changes have positioned HDD (hard disk drive or just disk) storage for a squeeze play in the data center. HDDs are increasing in capacity but not in performance as the

IOPS (I/Os Per Second) for HDDs have basically leveled off. As HDDs increase in capacity, their ability to deliver performance declines. The potential for more concurrently active data sets increases as HDD capacity grows and the increasing contention for the single actuator arm causes response time delays. This concern has fueled a trend of re-allocating high-performance or response time critical data to tier 0 (SSD) in order to deliver an acceptable service level. HDDs also face increasing RAID rebuild times, a higher TCO, and use considerably more energy than tape or SSD.

A common objective for many CIOs today is that *"if data isn't used, it shouldn't consume energy".* In response to this directive, the movement of archival data from HDD to more reliable, much more energy efficient and more cost-effective tape storage is well underway at the other end of the storage hierarchy. Tape capacities will continue to grow and total costs will decline, while HDD performance is flat and capacity growth will begin to slow facilitating a "storage squeeze play" – see chart below.



The HDD market will continue to grow but growth rates will slow pushing data from HDDs to SSD and tape. Over the next five years, forecasts for growth in the total amount of data stored on HDDs is expected to fall from 40%-45% annually to 30%, below the 52% CAGR for tape and 71% CAGR for flash data. This forecast assumes the primary driver of tier 0 storage is high-performance data and the primary driver for tape growth is the explosion in archived data. As a result, HDDs are caught in the middle as storage administrators strive to optimize their storage infrastructure.

Not all storage vendors offer all four tiers of the storage hierarchy however. Many have a tiered storage portfolio using only different speed (RPM) and varying cost disk drives. This approach forces all data to a variety of HDDs yielding a much more expensive storage solution and doesn't address tier 0 high-performance data or the tier 3 Colossal Content requirement effectively.

Key points: *HDDs are caught in the middle as storage administrators strive to optimize their storage infrastructure to better address high performance applications with SSD and archival demands with tape.*

Looking Inside the Storage Tiers - Pricing Models and TCO Comparisons

The storage industry has historically focused on the ASP (Average Selling Price) of storage to determine the lowest cost solution. Though this is still used for comparison purposes, the TCO (Total Cost of Ownership) provides a much more accurate picture of what a storage system will cost you over it's useful life. Don't confuse ASP with TCO. If you only look at the ASP, you will likely miss out on how much storage is actually costing your business and you're only getting part of the story.

Storage Tier	Avg. Data Dist. by Tier	CAGR	Storage Device(s)	ASP Range $/GB (Average Selling Price)
Tier 0 Very high performance applications	(1-3%)	~71%	Flash memory SSDs and DRAM SSDs	<$50 (current Flash ASPs) <$300 (current DRAM ASPs) ASP Trend – steady decline
Tier 1 Mission-critical enterprise-class applications, OLTP, high-performance databases	(12-20%)	~30%	Enterprise mission-critical high functionality disk arrays	Enterprise arrays - $5.00-$10 Includes controller, cache & drives. Add-on storage modules with no controller and cache have a lower $/GB. ASP Trend – declining slowly
Tier 2 Mid-range arrays, general applications, NAS, disk backup, reference data	(20-25%)	~30%	Midrange disk, NAS, VTLs, VSM/VLE, cloud for backup data, low cost disk	Midrange arrays - $1.00-$4 Includes controller, cache & drives. Add-on storage modules with no controller and cache have a lower $/GB. ASP Trend – declining slowing
Tier 3 Archive, fixed content, compliance, tape backup, unstructured data, infinite retention	(43-60%)	~52%	Tape libraries, , tape in the cloud for archive data	Tape libraries - $<.15 Includes tape drives, media and library and uses a 2:1 compression to determine library capacity (See pricing notes below). ASP Trend – steady decline

Source: Horison, Inc. (Aug. 2012)

Tiered Storage and Pricing Notes:

1) The **Average Data Distribution** by tier is an industry average that represents the amount of digital data a typical business has in each tier. Remember these are averages; some

businesses may have more or less data than the range indicated for each tier based on the type of IT applications they have. For example, some businesses have much more than 60% of their data in tier 3. In any case, the sum of all data across the tiers for a given business must add up to 100%.

2) The figures above are ***hardware ASPs*** are for working storage subsystems only; no storage management SW, personnel, facility costs, energy or maintenance costs are used as these are optional and vary widely negating any apples-to-apples comparisons. Working disk subsystems ASPs include cache, controller and drives. Tape library ASPs include a library, drives and media costs using a 2-x compression for tape capacity calculations. Deals and discounting vary by vendor, customer account level, preferred status, and potential sales opportunity. Averages are used in all cases.

3) The ***key factor*** used to calculate the purchase price per gigabyte for an automated tape library subsystem is the ratio of the number of tape cartridges to the number of tape drives in the library. ***The price of tape decreases as the ratio of cartridges to drives increases***. Higher capacity cartridges will lower the $/GB. Non-automated (offline) tape storage is not included in the table because the cartridge-per-drive ratio is meaningless as the number of cartridges and therefore tape capacity is unbounded.

4) Average ***storage administrator costs*** for disk and tape have been used in the past but formal industry averages are no longer available. These also vary by operating system and skill set as mainframe storage administrators can manage considerably more disk and tape storage than non-mainframe storage administrators given that more robust storage management tools are available. As a rule of thumb, the average storage administrator can manage far more tape data than disk data as TBs for disk and PBs for automated tape libraries represent a good baseline.

Acquisition Costs – ASP versus TCO

The growing need for business analytics has significantly increased the value of data stored for long periods of time (digital archives). Much of this data will optimally be stored on tape given the significant TCO benefits and other advantages of tape over disk. A comprehensive TCO study by ESG (Enterprise Strategies Group) [1] comparing an LTO-5 tape library system with a SATA disk system for backup using de-duplication (best case for disk) shows that using deduplication has a 2-4x higher TCO than the tape system for backup in several use cases. A variety of other studies comparing tape and disk for backup and archiving (best case for tape) conclude that disk has a TCO ranging from 2 to 15x higher than tape! Tape has been expanding its historical role as a backup solution to a much broader set of requirements including data archives and disaster recovery services. The advent of the cloud, and the inherent consolidation of unstructured data into large-scale archive repositories in the cloud signals that another category of storage is emerging – – which will further expand the economic model for tape in archiving data. The TCO advantage for tape is expected to become even more compelling with future technology developments.

[1] A Comparative TCO Study: VTLs and Physical Tape, By Mark Peters, ESG, Feb., 2011

Key points: *Don't confuse ASP with TCO. If you only consider the ASP, you might miss out on the most cost effective solution. The tape TCO is expected to improve faster than disk for the foreseeable future.*

Tape Today – Oracle Prepares for the Era of Colossal Content

By 2000, the tape industry was facing the growing perception that disk had become cheaper and more reliable than tape. In response to this challenge, the tape industry had begun re-architecting itself and results of this process are now evident. Key technology improvements were borrowed from the disk industry that yielded several tape developments including much longer media life, improved drive reliability, higher duty cycles, vastly improved bit error rates and much faster data rates than any previous tape technology. Today there are two companies offering enterprise-class tape drives; Oracle with the StorageTek T10000 drive family and IBM with the TS11x0 drive family while the increasingly popular LTO drive family is the de-facto standard format in the midrange market.

The StorageTek T10000C tape drive was announced with Fujifilm's advanced media providing a native capacity of 5 TB, making it the industry's largest capacity tape cartridge ever announced. The next generation Oracle StorageTek tape technology will further boost capacity and performance, using Fujifilm's advanced tape media coating technology. Fujifilm has pushed tape capacity, reliability and media life to new levels. Reliability for the StorageTek T10000C tape drive and its media is three orders of magnitude higher than the most reliable Fibre Channel disk drive and offers a media life of 30 years or more. Honestly, did you realize that tape has a higher capacity than disk and is more reliable than disk?

Technology and Reliability	BER (Bit Error Rate)
Enterprise Tape (T10000x, TS11xx)	1 x $10E^{19}$ bits
Midrange Tape LTO-5, LTO-6	1 x $10E^{17}$ bits
Enterprise HDD (FC/SAS)	1 x $10E^{16}$ bits
Enterprise HDD (SATA)	1 x $10E^{15}$ bits
Desktop HDD (SATA)	1 x $10E^{14}$ bits

The StorageTek T10000C enterprise tape drive provides the industry's highest capacity, performance, reliability, and data security capabilities to support 24/7 enterprise data center operations and can store over 1 Exabyte of data in Oracle's StorageTek SL8500 Library to address the requirements of both Big Data and Colossal Content. The eco-efficient StorageTek T10000C tape drive ensures the lowest TCO (total cost of ownership) providing ongoing compatibility with StorageTek enterprise libraries.

Key points: *Significant developments from Oracle have resulted in tape having a higher capacity than disk and tape being more reliable than disk! Expect this trend to continue.*

Comparing Tape and Disk for Building an Effective Digital Archive

The chart below compares key tier 3 archival storage requirements that are addressed by tape and disk to yield an optimized infrastructure. Though challenging and becoming increasingly complex as storage requirements grow, carefully designing the digital archive yields much improved operational efficiencies and sizeable cost savings.

Note: With the storage challenge steadily increasing, the time has arrived for businesses to begin re-architecting and optimizing their archives before the task becomes overwhelming.

Comparing Key Factors for Tape and HDD

Tier 3 Capability	Tape	HDD (Disk)
Suppliers	Five major tape system suppliers remain (IBM, HP, Oracle, Quantum, Spectra) Est. $3B rev. in 2012	Three HDD suppliers remain (Seagate, WD, Toshiba) Est. ~$30B rev. in 2012
Price/TCO (5 years)	TCO for HDD ranges 2-15X higher than tape for backup and archiving. Price per GB much lower for tape	HDD purchase price typically 20% of 5-year HDD TCO. Disk has more frequent conversions and upgrades
Long-life media	30 years or more on all new tape media	~4 years for most HDDs before upgrade or replacement, 7 years or more is typical for tape drives
Maximum Capacity	Current max. tape cartridge is Oracle/StorageTek's T10000C 5 TB (10 TB compressed @ 2:1)	Max. HDD capacity is 4 TB
Reliability	Enterprise tape BER (Bit Error Rate) has surpassed disk – 3 orders of magnitude higher than HDD	Disk BER not improving as fast as tape
DR - Move data to remote location for with or without electricity	Yes, can move data remotely with or without electricity. Natural disasters can require physical media movement	Difficult to move disk data to a remote location for DR purposes without requiring electricity
Inactive data does not consume energy	Yes, this is becoming a goal for most data centers. "If the data isn't being used, it shouldn't consume energy"	Rarely for disk, potentially in the case of "spin-up spin-down" disks *Note: data striping in arrays often negates the spin-down function*
Provide the highest security levels	Yes, encryption and WORM capability available on essentially all midrange and enterprise tape	WORM and encryption available on selected disk products but usage not yet widespread
Capacity growth rates	Roadmaps favor tape over disk with 35 TB capability jointly demonstrated by Fujifilm and IBM	Continued steady capacity growth - roadmaps project disk to lag tape
RAID re-build times becoming prohibitive	NA	Can take several days to rebuild a nTB HDD in a busy RAID array
Key Applications	Large files and large streaming data applications, active archives, less frequently used data, data that must be securely stored indefinitely	Active, frequently accessed data, mission critical data, databases, OLTP, revenue generating applications, replication
Labor costs	Typical tape administrator can manage PBs of tape	Typical disk administrator can manage TBs of disk

Source: Horison, Inc.

The general perception of tape is often quite different than today's reality as the magnetic tape industry has significantly improved its technology in the past 10 years. The latest technology developments in the tape industry indicate that tape is the most cost-effective technology for the fast growing tier 3 archival opportunities, whether on-site or in the cloud. While some backup applications are moving to disk in non-mainframe systems, long-term storage applications like digital archives, fixed content, multi-media and compliance are economically better suited for tape storage. Tape's role is significantly expanding from primarily a backup solution to that of a premier long-term archival storage technology.

Key points: *Tape vendors continue to innovate and deliver compelling new features with lower economics and higher reliability which have positioned tape as the optimal choice for long-term archiving as well as continuing to play a key role for backup.*

Oracle Combines SSD, Disk and Tape to Optimize the Storage Infrastructure

More than 40 years of storage expertise and innovative thinking have gone into creating Oracle's Pillar Axiom, StorageTek, and Sun Storage product portfolios. Utilizing SSD, HDD, tape and virtualization as core building blocks, Oracle is one of the few storage vendors who has developed a complete tiered storage hierarchy offering some of the most reliable, scalable, easy to manage, and energy-efficient systems in the industry. Optimal data placement is best achieved with a tiered storage implementation. At the highest level, tiered storage refers to an infrastructure capable of optimally aligning storage service levels with application requirements. The business case for implementing tiered storage is good and becomes more compelling as the storage pools get larger. The most effective tiered storage implementations need hardware and storage management software to work together in order to truly provide a seamless active operation and for customers to realize the substantial TCO and ROI benefits.

Oracle Tiered Storage - Building Optimized Storage

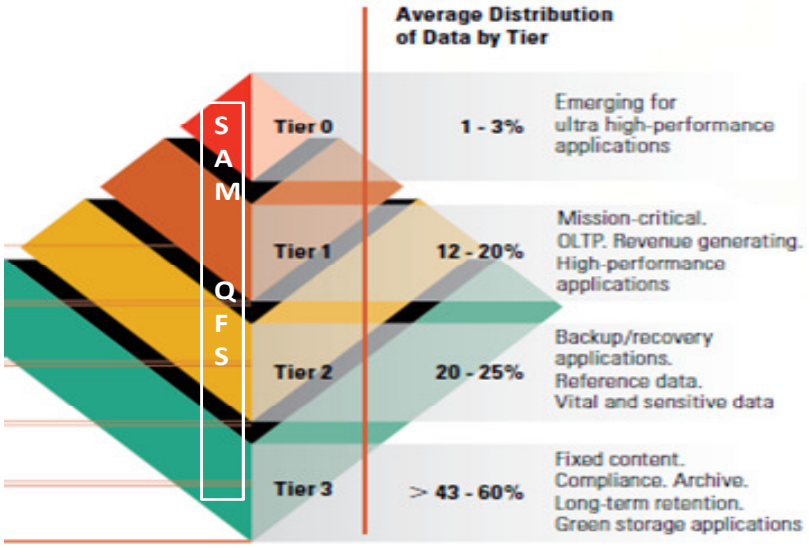
Integrated Storage Systems

Tier 0 (SSD)
Sun F5100 Flash Array, Sun SSD, Sun F20 PCIe Card

Tier 1 (Primary Disk)
SAN Pillar Axiom 600

Tier 2 (Secondary Disk)
Sun Storage 2500M-2 Fibre Channel Array, NAS Sun ZFS Storage Appliance, VLE, VSM

Tier 3 (Archival Storage)
Libraries – SL8500, SL3000, SL150
Drives – STK T10000, STK LTO



Source: Horison, Inc.

Managing the Storage Tiers in the Mainframe

Oracle has developed several unique enhancements to support its tier 3 storage solutions. The StorageTek VSM 5 system consists of a server, disk storage, and software that together provide a buffer or cache between the mainframe servers and the tape storage systems. View this as disk that appears to the system as multiple physical tape drives, but the tape drives are virtual tape drives. Instead of first writing directly to a physical tape drive, mainframe applications read and write to virtual tape drives created and maintained on disk. As data becomes less active over time, the probability of re-use diminishes and the data will either migrate to tape for long-term retention.

Another unique and innovative offering, the StorageTek VLE (Virtual Library Extension) integrates a scalable tier 2 disk array with the StorageTek VSM 5 combining disk and tape storage giving customers the optimal tier 3 performance and capacity solution. View StorageTek VLE as the *"Active Archive"* portion of the archive. StorageTek VLE addresses data types which often need to be stored on disk longer because of their usage patterns, from 45 to 90 days for example, before the probability of re-use diminishes and the data then migrates to tape for long term storage. Without the VLE there is more activity for tape storage, which can create situations where data is being constantly migrated and recalled back and forth from tape. A combination of StorageTek VSM 5, StorageTek VLE, and Oracle's StorageTek tape drives provides the highest performance, and the most reliable and economical enterprise storage system available for tier 3 data.

Managing the Storage Tiers in Open Systems

Oracle's Sun Storage Archive Manager (SAM) software provides policy-based storage tiering, comprehensive data protection, and data archiving combine with automatic data management to help reduce storage cost and complexity. SAM software provides an abstraction layer between applications and underlying storage tiers so that applications remain unaware of the physical location of data. As information is written by applications, it is automatically and transparently placed on the most cost-effective storage based on user specified retention and retrieval policies to align storage and archiving costs with business priorities and ease the management burden. Managed files appear to exist in the topmost directory of the storage hierarchy, no matter where they actually reside on physical storage. SAM provides HSM (Hierarchical Storage Management) services for non-mainframe systems and is the engine for a tiered storage implementation.

Improving Tape Administrator Efficiency

Oracle's StorageTek Tape Analytics software captures library, drive, and media health metrics in its dedicated server database and runs analytical calculations on these data elements to produce proactive recommendations for tape storage administrators. StorageTek Tape Analytics can eliminate library, drive, and media errors thorough an intelligent monitoring software application available exclusively for Oracle's StorageTek tape libraries and significantly improves storage administrators tape management capability, an increasingly important requirement for large archival storage systems.

Key points: *A combination of StorageTek VSM 5, StorageTek VLE, and tape storage (StorageTek T10000 drives) provides the highest performance, the most reliable and most economical enterprise storage system for tier 3 data in mainframe operating environments.*

Summary

Are you ready to address the tier 3 archive challenges that lie ahead? Data archiving supports many requirements, including E-discovery, regulatory compliance, managing test data, data analytics, and historical preservation. Tape densities will continue to grow and costs will decline, while disk drive performance remains flat and reliability has fallen behind tape. The opportunities for tape storage solutions either on-premise or in the cloud, have grown considerably and are being fueled by a plethora of significant technology advancements positioning tape systems to address much of the inevitable Colossal Content explosion. It really shouldn't matter which technology is the best for digital archiving, it just happens that the numerous improvements in tape have made it the optimal choice for archiving for the foreseeable future. Oracle's tiered storage solutions are engineered and integrated to enable a complete storage infrastructure with user defined policies and management capabilities to address tier 0 through tier 3 requirements. Building the archive of the future is attainable – now is the time to develop a solid and sustainable game plan.

End of report