

Extreme Scalability and Flexibility for Access to 100 Percent of Your Data

Oracle Optimized Solution for Secure Tiered Storage
Infrastructure

ORACLE WHITE PAPER | MAY 2016





Table of Contents

Introduction	1
Solution Objectives	2
Architecture Overview	2
Data Management with Oracle HSM	3
Oracle HSM	3
OpenStack Swift Support	4
Improvements in Oracle HSM	4
Tiered Storage Options	5
Server Infrastructure	6
SPARC T-Series Servers	6
Oracle Solaris Cluster	7
Sizing the Solution	8
Capacity Considerations	8
Storage Hardware Considerations	10
Small Configuration	11
Medium Configuration	13
Large Configuration for Small Files	15
Large Configuration for Large files	17
SPARC Server Cluster Configuration	18
Security Through Oracle Multitenant and Oracle HSM	19
Security Throughout the Solution Stack	20

Oracle FS1-2 Configuration and Performance Testing	22
Oracle Flash FS1-2 Configuration Best Practices	22
Configuring Oracle FS1-2 and Oracle HSM	22
Configuring LUNs on Oracle FS1-2	23
Oracle FS1-2 and Oracle HSM Test Results	25
Throughput for File Ingest	25
Linear Scalability for Dump and Ingest	26
Ingest Testing	27
Workload, Performance, and Architecture	29
Capability to Archive 10 Times More Data per Day	31
Oracle ZFS Storage Appliance Configuration and Performance Testing	32
Oracle ZFS Storage Appliance Configuration Best Practices	32
Best Practices for Configuring Oracle ZFS Storage Appliance and Oracle HSM	34
Oracle ZFS Storage Appliance and Oracle HSM Test Results	34
Oracle's Modular Tape Systems Configuration	35
Library Management Applications	36
Data Integrity Validation Process for Write, Read, and Validate	37
Performance Implications of Data Integrity Validation	38
Conclusion	41
References	41
Appendix I: Details for Reference Configurations	42
Appendix II: Software Revisions and Testing Tools	43



ORACLE®



Introduction

Exponential growth in digital data and tight regulatory compliance requirements make data retention and data access a great challenge. At the same time, many companies find value in being able to analyze historical data for collaboration and business intelligence, thus making it appealing to save operational data for several years. However, IT budgets are simply not growing fast enough to meet today's increasing storage capacity and performance requirements with only disk-based solutions. Offline tape archiving is not a viable solution, since compliance requirements and business usage requests mean that data must be accessible on demand.

This paper focuses on how to implement a secure, scalable, flexible, and yet cost-effective tiered storage solution that assures the integrity and accessibility of your data for the duration of its lifecycle by using the architecture and best practices defined in Oracle Optimized Solution for Secure Tiered Storage Infrastructure. This solution takes advantage of Oracle's broad portfolio of storage products, including intelligent flash storage systems, high-performance disk storage systems, and tape systems—with all the data managed by Oracle Hierarchical Storage Manager (Oracle HSM). The solution also utilizes the compute power, security, and I/O features in Oracle's SPARC servers to provide a robust platform for managing unstructured content from small implementations to very large implementations with billions of files. The resulting architecture keeps storage costs low, integrates with open standards such as OpenStack, and provides dynamic access and reliable, secure data protection over many years.

Oracle Optimized Solution for Secure Tiered Storage Infrastructure also is designed to greatly simplify deployment and management and to provide guidelines for component selection based on performance and capacity requirements.



Solution Objectives

Oracle Optimized Solution for Secure Tiered Storage Infrastructure orchestrates a tiered storage infrastructure consisting of high-performance storage area network-attached storage systems, extreme-performance flash storage systems, and tape archive systems. All of the data on the storage infrastructure is managed by Oracle HSM. The solution is designed to accomplish a range of objectives:

- » **Ensure data security and protection.** Increased illegal data access and use require a new level of data security and protection to assure that not only is data safe from hackers but also that the data put into the archive is the same data you access from the archive.
- » **Increase storage efficiency.** Organizations need to decrease overall storage costs over the lifetime of data. They need dynamic access to data from any storage tier, enabling valuable collaboration and reuse.
- » **Manage complex data.** Explosive data growth is challenging organizations' ability to cope and respond. Organizations need to free valuable IT staff from working on low-value, manual data management tasks so that IT can concentrate on higher-value, strategic, and transformational projects.
- » **Automate data placement.** Organizational decision-making requires accelerating the valuable data discovery process for making both tactical and strategic decisions based on current and historical data. Collaboration is vital, and users need to be able to reuse and share data more expediently by eliminating the time-consuming search-and-restore process from a backup.
- » **Survive IT transformations.** Data-driven organizations need to ensure that valuable data can be accessed in the future regardless of changes in technology or IT staff. Open formats are essential to maintain data accessibility with the capability for scaling expansively in both performance and capacity.

Architecture Overview

Oracle Optimized Solution for Secure Tiered Storage Infrastructure (shown in Figure 1) takes advantage of the robust capabilities of Oracle HSM for managing content on storage tiers that include the following:

- » Oracle FS1-2 flash storage system for the most intelligent and most scalable converged flash storage system in the industry
- » Oracle ZFS Storage Appliance with hybrid pools of hard disk drives (HDDs) and solid-state drives (SSDs)
- » Oracle's StorageTek modular tape library systems with Oracle's StorageTek LTO 6 or StorageTek T10000D tape drives
- » The OpenStack Swift RESTful interface for enabling private cloud infrastructures

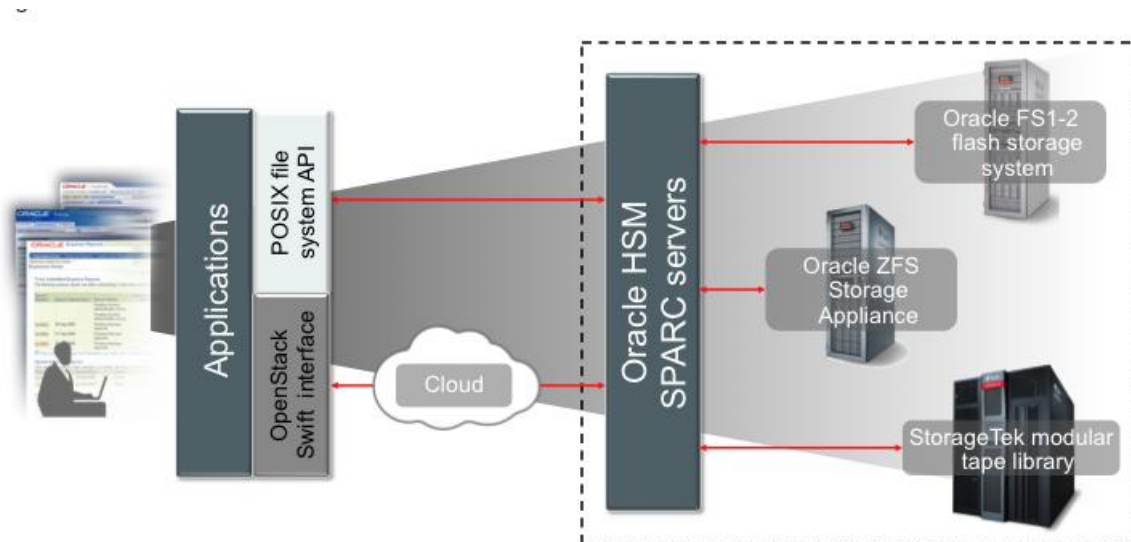


Figure 1. Oracle HSM manages data stored on flash storage, primary drives, tape archive, and cloud.

This solution provides a very scalable, flexible, and cost-effective storage platform for many use cases. Such broad use demands an infrastructure that scales for both performance and capacity. The use of Oracle's disk storage systems and SPARC servers running Oracle Solaris delivers deployment flexibility and high performance. The StorageTek modular library systems provide long-term storage preservation enabling nondisruptive expansion.

The architecture can be divided logically into the following categories:

- » **Data management.** Oracle HSM is a powerful policy engine that enables organizations to automatically move data to the appropriate storage tier based on access requirements.
- » **Server infrastructure.** SPARC servers provide the best platform to run Oracle HSM.
- » **Tiered storage.** Tiered storage includes a range of storage area network-connected storage devices consisting of Oracle ZFS Storage Appliance, the Oracle FS1-2 intelligent flash system, a StorageTek modular tape archive system, and a private cloud.


The following subsections provide an overview of these three major components of the architecture.

Data Management with Oracle HSM

Oracle HSM is the critical component of a tiered storage infrastructure because it unifies the different tiers of storage while giving applications a simple, file-structured view of the data. It provides automatic and dynamic access to content from any storage device by combining Oracle HSM data management with Oracle's StorageTek QFS advanced file system to present a single file system view to applications and users. As a result, Oracle HSM hides the complexity of tiered storage and provides transparent access regardless of where data is stored. Abstracting the storage from the application simplifies management, while providing all the benefits of scalability and flexibility across multiple storage tiers.

Oracle HSM

Oracle HSM is a storage software application that runs on Oracle Solaris, optionally utilizing Oracle Solaris Cluster for an active/passive high-availability (HA) environment. Although it is recommended that Oracle HSM run on Oracle's SPARC-based servers, it is also supported on Oracle's Sun X86 systems. In an HA configuration HA NFS



enables applications to access data on the active node while providing the ability to fail over to the passive node if the active node fails.

Oracle HSM accesses content from the primary disk, which is referred to as disk cache, based on preset policies and creates copies on archive disks, tape devices, or both. If content is released from the disk cache, Oracle HSM then dynamically accesses the content from any device. Up to four data copies can be created during the archive process. These copies ensure the security of the data from corruption as it is not online and available. Copies can be made locally and remotely. Remote copies are remote disk archives, and they also are the Oracle HSM file systems that allow additional tape copies to be made at the remote site.

Each Oracle HSM disk cache can scale to 32 PB and support more than 1 billion files. Moreover, the capacity under Oracle HSM management can reach hundreds of PB through the use of tape media. The archiving file system policies automatically manage the lifecycle of the archive data through four features of Oracle HSM:

- » **Archive.** The archiving process transparently archives data from disk cache to archive disk, tape, or both without operator intervention. The Oracle HSM archive process uses policies based on file system characteristics, such as path name, wildcard, size, age, owner, group, or date, to automatically manage the copies.
- » **Release.** The releasing process automatically manages the disk cache and releases files from the disk cache that are archived when the high-capacity threshold is reached on the primary storage or according to policy. The list of files eligible to be released is prioritized based on policies such as archive status, size, release status, and age.
- » **Stage.** The staging process automatically stages released files back to disk cache or directly to the requesting application when files are accessed. Staging options include prestaging and bypassing the disk cache. Removable media access is optimized for mounting and positioning.
- » **Recycle.** The recycling process repacks archive media onto new media in order to reclaim space. The recycling process can be used to migrate from older to newer technology; however, a new capability—media migration—now simplifies moving data to new tape media.

Throughout a file's lifecycle, the Oracle HSM metadata remains online and available to the content management application. All files appear to be located on the disk cache when they might be only on tape. The result is cost-effective management and use of tiered storage while dynamic and immediate direct access is provided to 100 percent of the data without operator intervention or human knowledge about where the data resides. As a result, users have access to data that might be many years old or might not have been accessed in many years.

OpenStack Swift Support

OpenStack Swift is an open source object storage system used to create scalable cloud infrastructures. Oracle now provides an OpenStack Swift interface to the Oracle HSM software, making the deployment of private cloud infrastructures efficient and massively scalable. Oracle HSM enables storage tiering, which offers organizations the freedom to choose how data is stored while they keep all of the data available for access by applications. Data can be kept on flash or disk storage devices for fast access or on digital tape for low-cost long-term archiving.

Tape now can become part of the infrastructure supported in the cloud with the benefit of much higher reliability and lower cost than traditional disk devices. Users have the flexibility to determine how much of their data they want on fast-access storage and how much they want on low-cost storage, and they can tailor their systems based on their specific needs. By deploying Oracle HSM as an OpenStack Swift cloud server with Oracle's broad portfolio of storage products, you can rely on this solution to store massive amounts of unstructured data reliably and cost effectively. The solution also is optimized for multitenancy and high concurrency to meet high-performance requirements.

Improvements in Oracle HSM

Oracle HSM 6.1 provides a number of significant advances that enhance the integrity of operations and data as well as availability. Improvements include the following:

- » **Fixity.** Fixity is a term used by data archivists and preservationists to ensure the integrity of a digital object by making sure it remains unchanged, constant, and stable. Oracle HSM enables the use of several fixity algorithms including SHA-1, SHA-256, SHA-384, SHA-512, and MD5 as well as the original proprietary algorithm. The hash value and algorithm also can be accepted from an application through the Oracle HSM API. Accepting a hash value from outside the Oracle HSM environment ensures the data stored in the application storage is the same data stored in the archive. As the file is copied throughout the archive tiers, the hash is recalculated and compared with the original hash as each copy is created. The hash calculation takes place in the server processor chip, not in software, providing data preservation while not affecting performance.
- » **Accelerated media migration.** Media changes as technology changes; however, data does not change. Therefore, migration of data through technology changes is an important process in an archive strategy. In the past, staging data back into the disk cache was required in order to rearchive to new media. Today, this is accomplished through two new Oracle HSM features:
 - » **StorageTek Direct Copy:** Data on one media is copied directly to a StorageTek T10000D media. There are two options in this process. One is to repack the content, only copying an archive image that has at least one active file, or copy all archive images, even those with no active files, from one media to the newer media.
 - » **Server copy:** Data on one media is copied to the server memory and then copied to the new media. The repack option is always used with this method of media migration.

Following the migration of data using either method, the metadata is updated with the new location of the data. Optionally, a log file entry is written indicating the source and the destination of the file to help you maintain a list of all data activity.


- » **Expanded LUN size.** The disk cache now supports a single LUN size of 128 TB. With support of up to 250 LUNs, the supported disk cache capacity is 32 PB. Disk storage systems now have the ability to create very large LUNs, and this simplifies the creation of an Oracle HSM file system because fewer LUNs are required to reach the required capacity. As with previous Oracle HSM releases the LUNs non-disruptively increase in capacity. Following an upgrade to Oracle HSM 6.1, it is possible to grow the original LUN to 128 TB.
- » **Extended metadata performance improvement.** Oracle HSM has supported extended attributes in previous releases; however, these attributes were stored with the file content, which is generally on slower capacity disk storage devices. All utilities or access to these attributes through the Oracle HSM API required reading from these slower storage devices. Oracle HSM 6.1 moves these attributes into the metadata storage devices—usually SSD—resulting in large performance improvements when extended metadata is used. This improvement applies to the new fixity capability, which stores the hash in extended metadata.

More information on Oracle HSM can be found at: <http://www.oracle.com/us/products/servers-storage/storage/storage-software/storage-archive-manager/overview/index.html>.

Tiered Storage Options

Tiered storage is critical because content must be kept for long periods, yet some use cases require fast ingest as well as fast access for recently ingested data. Oracle has two tier 1 storage products that meet the requirements of a tiered storage environment. Both the Oracle FS1-2 flash storage system and Oracle ZFS Storage Appliance are well suited to meet the requirements of Oracle Optimized Solution for Secure Tiered Storage Infrastructure. In addition, Oracle offers the most complete line of tape systems for low-cost, high-reliability archiving.

- » **Oracle FS1-2 flash storage system.** The Oracle FS1-2 flash storage system is the most intelligent flash storage in the industry, and it is used in tiered storage environments for the most demanding and highest performance workload requirements. The system dynamically responds to various and changing application I/O requirements based on usage patterns with smart quality of service (QoS) policies that reorder I/O operations for multiple access patterns. This rapid learning feature delivers the highest ingest or access performance using flash media when needed, and yet it also incorporates low-cost capacity disk media in the same storage system. Dynamic auto-tiering is also engaged to intelligently restructure storage pools based on business priorities, to further aid in using this system for a multiapplication, multiworkload environment. More information on Oracle FS Series can be found at <http://www.oracle.com/storage/san/fs1/index.html>.

- 
- » **Oracle ZFS Storage Appliance.** An alternative to the Oracle FS1-2 flash storage system, Oracle ZFS Storage Appliance can be used to store both the recently stored data and the most active data, no matter the age of the data. This storage supports mixed combinations of high-performance SSD caching and various-speed, large capacity HDD storage for Oracle HSM. Flexible archive policies provided by Oracle HSM keep copies of the data on the most appropriate storage tier. The Oracle ZFS Storage ZS3-2 entry-level engineered storage system delivers extreme efficiency and reduces cost, complexity, and risk while meeting high ingest and access requirements, thereby providing primary storage as well as disk archive for smaller configurations. Oracle ZFS Storage ZS3-4 is a large-scale engineered NAS storage system that delivers reduced complexity and risk for enterprise customers demanding high-performance storage with extreme efficiency and low TCO for the largest and most demanding workloads. More information on Oracle ZFS Storage Appliance can be found at <http://www.oracle.com/storage/nas/index.html>.
 - » **Oracle's StorageTek tape and library systems.** Proven StorageTek tape and library systems help organizations maximize secure data access, manage complexity, and control costs. Tape is used for archival of data and provides efficient access to all of the data, regardless of where it is stored or the retention period, while also providing data security and data protection through multiple copies. The Data Integrity Validation feature of the StorageTek T10000D tape drive is based on ANSI standard cyclic redundancy checks (CRCs). This capability provides additional security and data protection by validating that what was sent to tape is what was actually written. If inactivity or environmental factors deteriorate the media, Oracle HSM is notified and a new archive copy is created from an alternate copy. More information on StorageTek modular library systems and tape drives can be found at <http://www.oracle.com/goto/tape>.

Server Infrastructure

The server infrastructure deployed for Oracle Optimized Solution for Secure Tiered Storage Infrastructure is based on Oracle's SPARC T-Series servers, with high availability provided by Oracle Solaris Cluster and security features available in Oracle Solaris.

SPARC T-Series Servers


SPARC T-Series servers lend themselves to the most demanding cloud and enterprise applications deployed in Oracle Optimized Solution for Secure Tiered Storage Infrastructure. Built-in zero-overhead virtualization provides real-time scaling for Oracle HSM. SPARC T-Series servers also are available with high-speed networking, which aids in removing I/O bottlenecks in highly virtualized environments.

Oracle's latest SPARC CPU is incorporated into these servers to provide the industry's most highly integrated "system on a chip" and to supply the most high-performance threads of any multicore processor available. The SPARC core architecture provides best-in-class systems that are engineered and optimized to accelerate Oracle software and business-critical applications with security, extreme performance, mission-critical reliability, and scalability. SPARC servers also provide up to three times faster security with CPU-integrated, zero-overhead encryption accelerators.

Many organizations discover that the first step to managing complexity is to host as many users' applications as possible on each of as few as possible hardware platforms. Oracle's StorageTek QFS file system multitenant architecture now enables multiple users' applications to share a single file system under Oracle HSM using any supported protocol, including NFS, CIFS, FTP, and OpenStack Swift.

SPARC T-Series servers represent scalable building blocks that are configured into four reference configurations for sizing an Oracle HSM deployment. The server models recommended for the solutions include Oracle's SPARC T7 processor-based servers with the appropriate number of cores required for Oracle HSM and any additional cores needed for additional application consolidation.

In addition to the features that support an archive environment for unstructured data, new and innovative features also are available for the structured environment. These features make it possible for some of Oracle Database



functions to run in the SPARC chip—technology known as Oracle's Software in Silicon. Further, these features enable world record-setting performance, all while maintaining the highest level of security. Oracle's SPARC technology and the Oracle Solaris operating system provide the best tools and infrastructure for processing both structured and unstructured data. Only a vendor, such as Oracle, that develops all the layers together can deliver these advanced products.

Additional information on SPARC servers can be found at <http://www.oracle.com/us/products/servers-storage/servers/sparc/oracle-sparc/overview/index.html>.

Oracle Solaris Cluster

In today's global 24/7 economy, keeping enterprise applications up and running is more important—and can be more complex—than ever. Tiered storage is normally associated with dark archives that are rarely accessed and often high availability is not considered a requirement. However, this archived data provides an advantage in product development and corporate strategy, and the “new normal” is continuous access to 100 percent of all of your data, including old data on paper and microfilm that is digitized, indexed, and archived.

Government regulations, corporate financial goals, and evolving requirements to address new opportunities mean IT systems need to be constantly available, which can be a challenge with today's complex solution stacks and unique business requirements. Recovery time objectives (RTOs) must be determined to decide whether a cluster environment is required. Oracle HSM runs in an active/passive environment utilizing Oracle Solaris Cluster, which provides broad advantages to the solution, including the following:

- » **High-availability framework.** This software framework detects node failures quickly and activates resources on another node in the cluster. The framework includes a cluster membership monitor, which is a distributed set of algorithms and agents that exchange messages over the cluster interconnect. The exchange enforces a consistent membership view in order to synchronize reconfiguration, handle cluster partitioning, and help maintain full connectivity among all cluster members.
- » **Virtualization support.** Oracle Solaris Cluster provides comprehensive support for the following Oracle virtualization software: Oracle Solaris Zones, Oracle VM Server for SPARC (also called logical domains or LDoms), and the Dynamic Domains feature (available on Oracle's SPARC Enterprise M-Series servers). This enables flexible HA for server consolidation efforts. Applications can run unmodified in a virtualized environment.
- » **Flexible storage support.** Oracle Solaris Cluster can be used with different storage technologies such as Fibre Channel, SCSI, iSCSI, and NAS storage on Oracle or non-Oracle storage products. There is also broad file system and volume manager support.
- » **High-availability Oracle Solaris ZFS.** With the virtually unlimited scalability of ZFS, Oracle Solaris Cluster offers a file system solution with exceptional availability, data integrity, and flexibility for growth.
- » **Component monitoring.** Oracle Solaris Cluster provides extensive monitoring of application processes, disk path integrity, and network availability. For example, all disk paths can be monitored and configured to automatically reboot a node in the event of a multiple-path failure.
- » **Failover and scalable agents.** Software programs that support Oracle or third-party applications can take full advantage of Oracle Solaris Cluster features.
- » **Security.** Security is integrated within the complete solution through the provision of functional security guidelines and best practices as well as utilization of the tools with the hardware and software that are part of each component. Preservation assures the data is always secure and available.

The implementation guide for Oracle Optimized Solution for Secure Tiered Storage Infrastructure includes specific configuration instructions for Oracle HSM in an Oracle Solaris Cluster environment.

The following web page provides additional detailed information about Oracle Solaris Cluster:
<http://www.oracle.com/technetwork/server-storage/solaris-cluster/documentation/index.html>.

Sizing the Solution

Two main points must be taken into consideration when selecting a hardware configuration for Oracle Optimized Solution for Secure Tiered Storage Infrastructure: capacity and ingest performance. Ingest performance is important for use cases that have instrumentation that generates TB of data per hour. Capacity is important for use cases that might have a smaller daily ingest requirement but have a very long retention period (often forever) for millions or even billions of files. All use cases have specific requirements for both performance and capacity.

Oracle has performed testing to provide guidance for selecting a configuration that most closely meets current requirements. This testing has resulted in the following configuration categories that match a range of ingest performance and capacity requirements:

- » Small configurations range from 198 TB to 3.65 PB of archive capacity.
- » Medium configurations range from 7.7 PB to 50.3 PB of archive capacity.
- » Large configurations range from 15 PB to 50.3 PB of archive capacity.
- » Capacity configurations range from 15 PB to 860 PB of archive capacity with extended capacity on primary disk, allowing for an increase in capacity for active data.

Capacity Considerations

When capacity is calculated based on ingest rates running 24/7, 365 days per year, it quickly grows beyond the typical projection based on 50 percent growth per year. Table 1 represents capacity of content with a retention of seven years based on proven continuous ingest rates of 24 hours per day and seven days per week, ingesting 3.2 GB/sec. Following this rule of thumb, the solution will store 10 percent of the total content on primary disk for frequent access, 30 percent on disk archive, and 200 percent on tape, representing data protection copies as well as archive, ensuring the security of your data. This sample allocation reaches capacities much higher than the 50 percent growth projection.

TABLE 1. SEVEN-YEAR TABLE USING MAXIMUM INGEST RATE RUNNING 24/7 AND 365 DAYS PER YEAR

Ingest GB/sec	Ingest Capacity per Day (TB)	Ingest Capacity per Year (PB)	Capacity in Seven Years (PB)	10% Primary Disk (PB)	30% Archive Disk (PB)	200% Tape Archive (PB)
3.20	270.00	96.24	673.7	67.4	202.1	1,347.4

Realistically, the ingest rates noted in Table 1 are experienced only at peak times, not 365 days per year; however, they must be processed with little or no impact to users during those peak times. The total ingest capacity is more likely a much smaller percentage of the totals shown. To achieve a solution that meets both performance and capacity requirements, the small, medium, large, or capacity configuration should be selected based on the peak ingest and access rates required, and total capacity should be selected based on actual expected growth and retention time. The scalability and flexibility of the solution enables the infrastructure to grow capacity or performance or both to meet new requirements.

An example of capacity requirements, Table 2 starts with 50 TB of capacity for a small configuration, 500 TB of capacity for a medium configuration, and 1,000 TB (1 PB) of capacity for a large configuration with an expected growth of 50 percent per year through seven years. As stated previously, estimated capacity for each tier of storage is 10 percent for primary disk, 30 percent for disk archive, and 200 percent for tape archive to provide secure, data protection copies.

TABLE 2. SEVEN-YEAR CAPACITY GROWTH, INCREASING 50 PERCENT PER YEAR

		Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
Small configuration's capacity growth for seven years	Total content capacity (TB)	50	75	113	169	253	380	570
	Primary disk (10%) (TB)	5	8	11	17	25	38	57
	Archive disk (30%) (TB)	15	23	34	51	76	114	171
	Archive tape (200%) (TB)	100	150	225	338	506	759	1,139
Medium configuration's capacity growth for seven years	Total content capacity (TB)	500	750	1,125	1,688	2,531	3,797	5,695
	Primary disk (10%) (TB)	50	75	113	169	253	380	570
	Archive disk (30%) (TB)	150	225	338	506	759	1,139	1,709
	Archive tape (200%) (TB)	1,000	1,500	2,250	3,375	5,063	7,594	11,391
Large configuration's capacity growth for seven years	Total content capacity (PB)	1.0	1.5	2.3	3.4	5.1	7.6	11.4
	Primary disk (10%) (PB)	0.1	0.2	0.2	0.3	0.5	0.8	1.1
	Archive disk (30%) (PB)	0.3	0.5	0.7	1.0	1.5	2.3	3.4
	Archive tape (200%) (PB)	2.0	3.0	4.5	6.8	10.1	15.2	22.8

The graphs in Figure 2 and Figure 3 make it clear that even with dramatic growth in overall capacity, the primary disk storage tier, which is the most expensive tier, remains relatively small, thus keeping costs low. The most cost-effective storage tier—tape—carries the largest capacity.

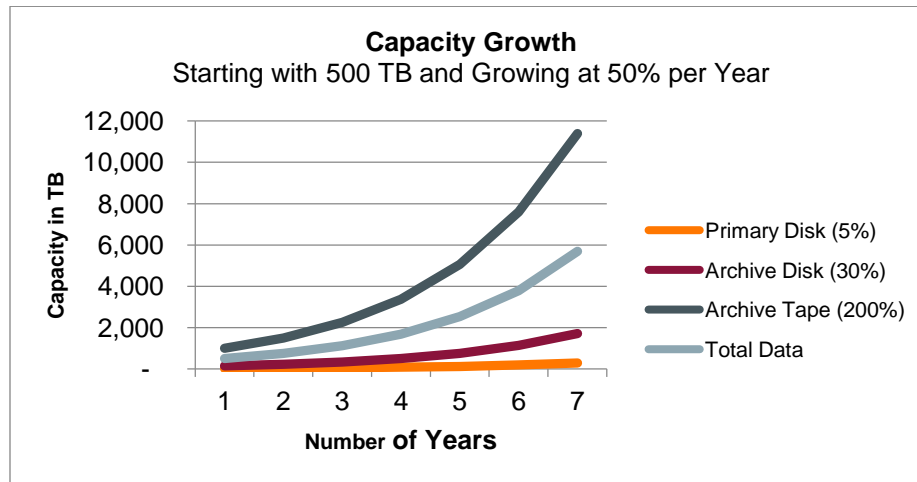


Figure 2. Capacity starting with 500 TB and growing 50 percent per year.

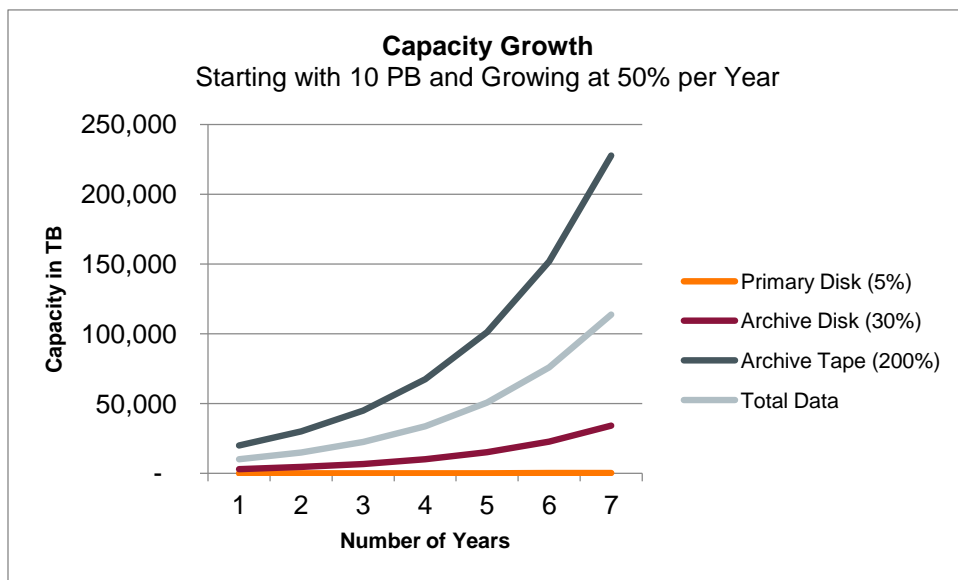


Figure 3. Capacity starting with 10 PB and growing at 50% per year.

All requirements, starting with performance and capacity, must be taken into consideration when selecting a small, medium, or large configuration. The Oracle Optimized Solutions team has proven that the components work together, and the test results provide guidelines for size selection. With Oracle HSM, Oracle FS1-2 or Oracle ZFS Storage Appliance, and StorageTek tape systems providing tools for migrating data nondisruptively, it is possible to begin with one configuration size and easily and confidently move to the next size.


Storage Hardware Considerations

Oracle FS1-2 storage drive enclosures can be configured with a range of SSD flash drives and HDD disk drives to meet business needs and performance needs. Drive enclosure media options include 400 GB performance SSDs, 1.6 TB capacity SSDs, 1.2 TB performance disk drives, and 8 TB capacity disk drives. A single Oracle FS1-2 flash storage system supports any combination of these drives. By scaling out either SSD or disk drive enclosures, a single Oracle FS1-2 flash storage system can support up to 912 TB of flash or 5.76 PB of disk-based storage or a combination in a maximum of 30 disk enclosures.

An additional decision is selection of HDD disk drives for the Oracle FS1-2 system. The file size that is ingested and archived is an important factor. Smaller files, less than 200 MB, require performance HDD drives to meet performance expectations. Capacity HDD drives can be used for ingest and archive of large files. The SSD drives are always used for the Oracle HSM metadata.

Oracle ZFS Storage Appliance systems are configured with tiers of solid state storage including large dynamic random access memory (DRAM) pools, and both read and write cache areas use flash memory. The Oracle ZFS Storage Appliance systems are available with single- or dual-controller options in two basic models:

- » The Oracle ZFS Storage ZS3-2 appliance is an entry-level engineered storage system for smaller deployments equipped with 1 TB DRAM and up to 12.8 TB read flash and as much as 3.1 PB of capacity per cluster.
- » The Oracle ZFS Storage ZS3-4 appliance is an engineered storage system for larger deployments, with up to 3 TB DRAM and 12.8 TB read flash, and it scales up to 6.9 PB of raw uncompressed capacity per cluster.



Through the use of hybrid storage pools in Oracle ZFS Storage Appliance, all writes initially go to SSD drives and are immediately destaged to HDD. Therefore, in most cases, the capacity HDD drives can be used in this configuration and still meet performance expectations.

The highly scalable StorageTek tape libraries ensure data availability in heterogeneous tape storage environments of any size, including those that comprise Oracle Applications, Microsoft Windows desktops, mainframes, and supercomputers. The StorageTek modular library systems that are proposed in the small, medium, large, and capacity tiered-storage solution scale from 30 to 100,000 slots (75 TB to 850,000 TB), meeting virtually all capacity, archive, and access requirements.

The following sections provide capacity comparisons for small, medium, large, and capacity configurations for Oracle Optimized Solution for Secure Tiered Storage Infrastructure. This discussion can help you select the appropriately sized configuration to meet initial capacity requirements, and the information shows the nondisruptive scalability of the solution as business needs grow and data capacity increases. Ease of scalability is discussed in further detail in the configuration and best practices sections that follow.

Small Configuration

Figure 4 and Figure 5 show the small reference configuration with its starting storage capacity for Oracle Optimized Solution for Secure Tiered Storage Infrastructure. Four cores in each SPARC T7-1 server are dedicated to running Oracle HSM with Oracle Solaris Cluster. Up to 32 cores are available in each processor for application consolidation or future expansion. The small configuration provides a choice of the following (see Appendix I for more details):

- » The Oracle FS1-2 system and Oracle's StorageTek SL150 modular tape library with:
 - » Initial capacity of 48 TB Oracle HSM disk cache and 150 TB tape archive
 - » Nondisruptive expansion to 2.9 PB tape archive and 750 TB Oracle HSM disk cache
- » The Oracle ZFS Storage ZS3-2 appliance and the StorageTek SL150 modular tape library with:
 - » Initial capacity of 163.2 TB Oracle HSM disk cache and 150 TB tape archive
 - » Nondisruptive expansion to 2.9 PB tape archive and 1.5 PB Oracle HSM disk cache

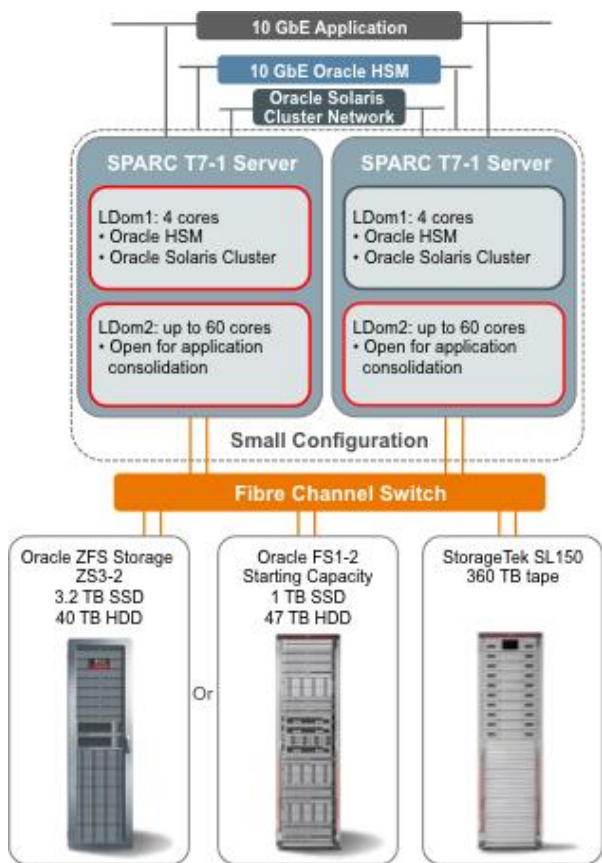


Figure 4. Small reference configuration with the Oracle ZFS Storage ZS3-2 appliance or the Oracle FS1-2 and the StorageTek SL150 modular tape library shows capacity with the physical-to-logical LDom configuration for best performance with Oracle HSM.

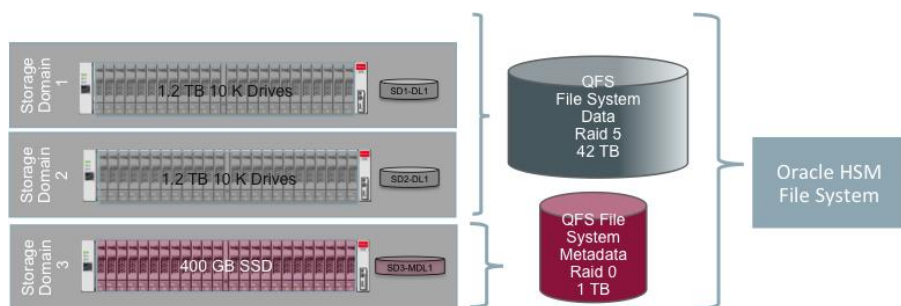


Figure 5. Small reference configuration for the Oracle FS1-2 disk layout and LUN assignment. Additional LUNs can be created within each storage domain for multiple file systems.

Medium Configuration

Figure 6 and Figure 7 show the medium reference configuration with its starting storage capacity. Like the small configuration, four cores in each SPARC T7-2 server are dedicated to running Oracle HSM with Oracle Solaris Cluster for high availability. Up to 28 cores in each server are available for application consolidation or expansion. The medium configuration provides a choice of the following (see Appendix I for more details):

- » The Oracle FS1-2 system and Oracle's StorageTek SL3000 modular library system with:
 - » Initial capacity of 91.2 TB Oracle HSM disk cache and 1,700 TB tape archive
 - » Nondisruptive expansion to 2.9 PB Oracle HSM disk cache and 47.4 PB tape archive
- » The Oracle ZFS Storage ZS3-2 appliance and the StorageTek SL3000 modular library system with:
 - » Initial capacity of 231.2 TB Oracle HSM disk cache and 1,700 TB tape archive
 - » Nondisruptive expansion to 1.5 PB Oracle HSM disk cache and 47.4 PB tape archive

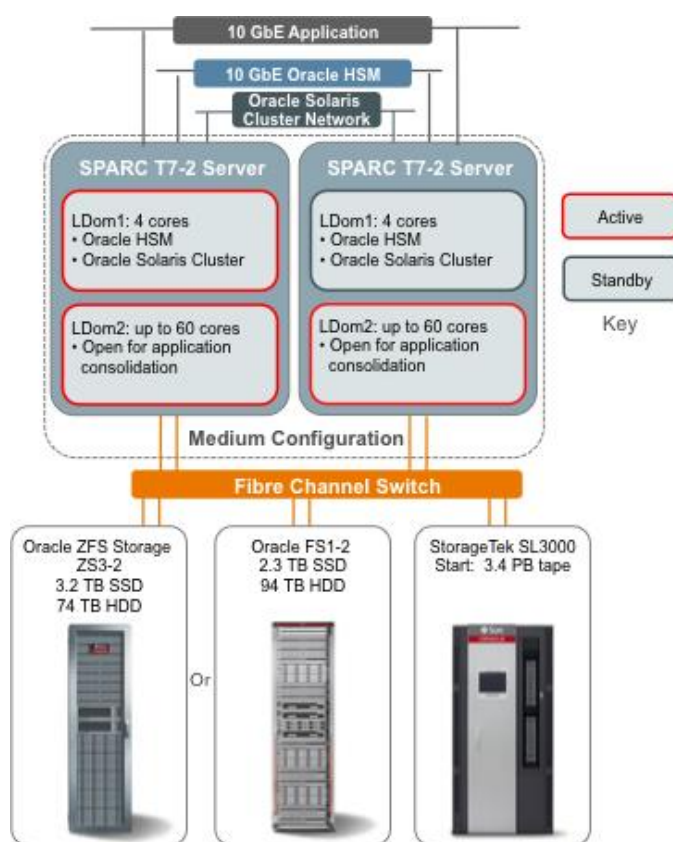


Figure 6. Medium reference configuration with the Oracle ZFS Storage ZS3-2 appliance or the Oracle FS1-2 system and the StorageTek SL3000 modular library system shows capacity with the physical-to-logical LDom configuration for best performance with Oracle HSM.

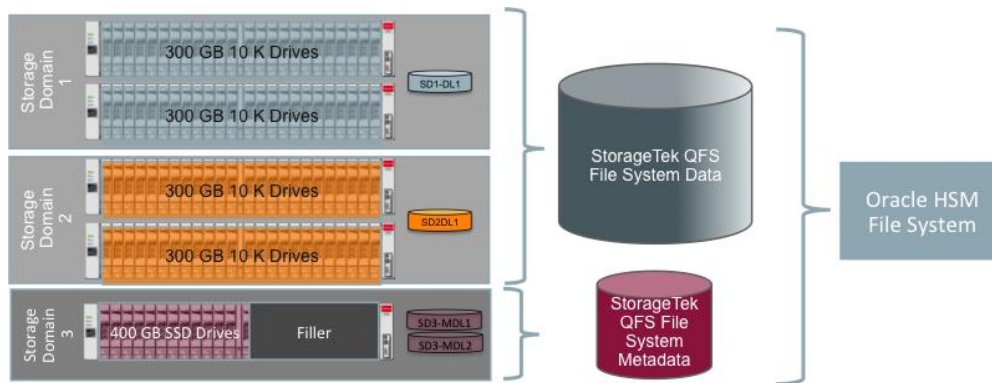


Figure 7. Medium reference configuration for the Oracle FS1-2 disk layout and LUN assignment. Additional LUNs can be created in each storage domain for multiple file systems.

Large Configuration for Small Files

Figure 8 and Figure 9 show the large reference configuration with its starting storage capacity utilizing performance disks for the disk cache. This is appropriate for file ingest of files less than 200 MB. Eight cores in each SPARC T7-2 server are used for Oracle HSM with Oracle Solaris Cluster for high availability. The large configuration provides a choice of the following (see Appendix I for more details):

- » The Oracle FS1-2 system and Oracle's StorageTek SL8500 modular library system with:
 - » Initial capacity of 139.2 TB Oracle HSM disk cache and 5,950 TB tape archive
 - » Nondisruptive expansion to 2.9 PB disk and 47.4 PB tape archive
- » The Oracle ZFS Storage ZS3-4 appliance and the StorageTek SL3000 modular library system with:
 - » Initial capacity of 462.4 TB Oracle HSM disk cache and 14,875 TB tape archive
 - » Nondisruptive expansion to 1.5 PB Oracle HSM disk cache and 857.5 PB tape archive

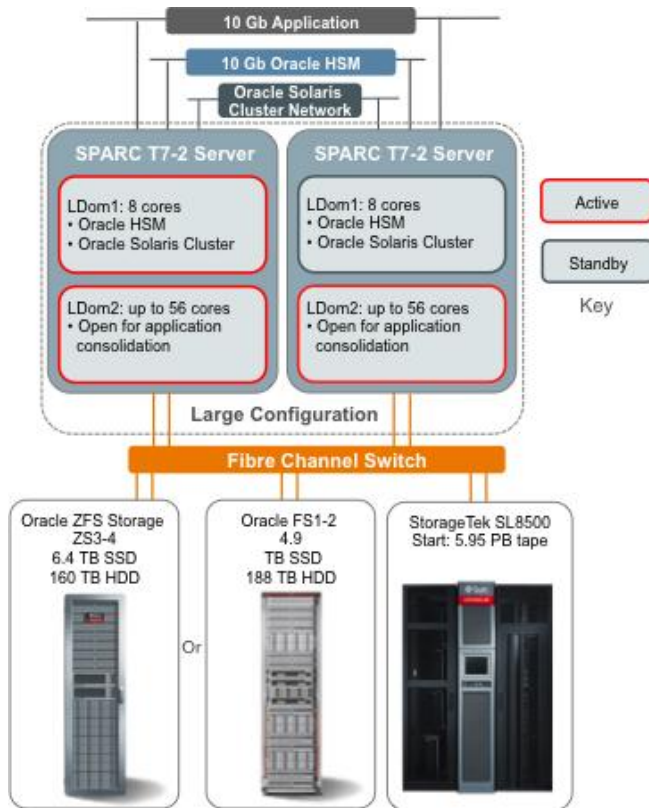


Figure 8. Large reference configuration with the Oracle ZFS Storage ZS3-4 appliance or the Oracle FS1-2 system and the StorageTek SL8500 modular library system shows capacity with the physical-to-logical LDom configuration for best performance with Oracle HSM.

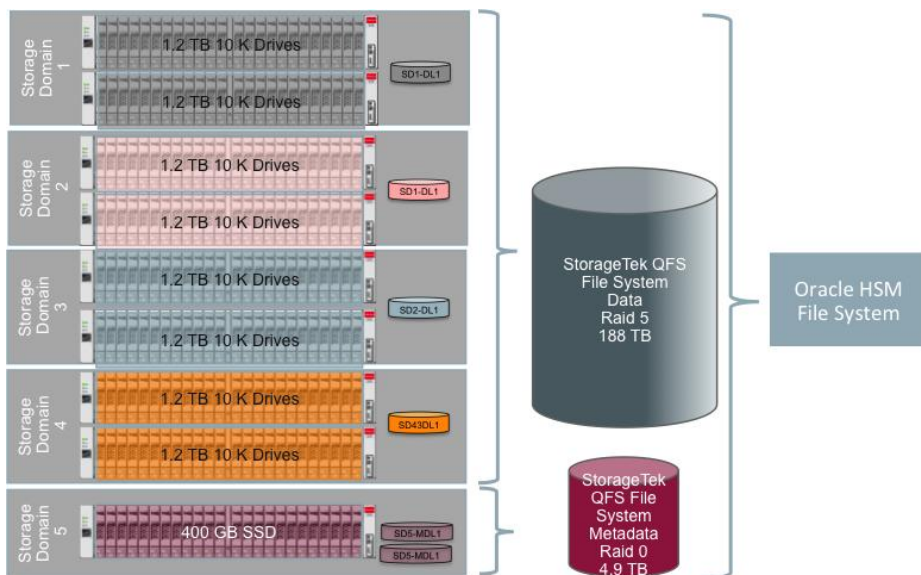


Figure 9. Large reference configuration for the Oracle FS1-2 disk layout and LUN assignment. Additional LUNs can be created in each storage domain for multiple file systems.

Large Configuration for Large files

Figures 10, 11, and 12 show the capacity reference configuration with more starting disk storage capacity than the previous configuration. For ingest and archive of files greater than 200 MB, the capacity disk drives can be used. For the case of a very active ingest and archive, when additional tape drives are required, Oracle's SPARC T7-4 with two processors provides 16 PCI slots compared to 8 slots in Oracle's SPARC T7-2. This provides support for additional FC HBA cards and additional tape drives. Another option to reach the required ingest and archive performance is to use the scale-out feature of Oracle HSM and include additional servers to share the ingest and archive or staging load requirements.

Like the large configuration, the capacity configuration provides a choice of the following (see Appendix I for more details):

- » The Oracle FS1-2 system and the StorageTek SL8500 modular library system with:
 - » Initial capacity of 244.8 TB disk and 25,075 TB tape archive
 - » Nondisruptive expansion to 2.9 PB disk and 857.5 PB tape archive
- » The Oracle ZFS Storage ZS3-4 appliance and the StorageTek SL8500 modular library system with:
 - » Initial capacity of 462.4 TB disk and 25,075 TB tape archive
 - » Nondisruptive expansion to 1.5 PB disk and 857.5 PB tape archive

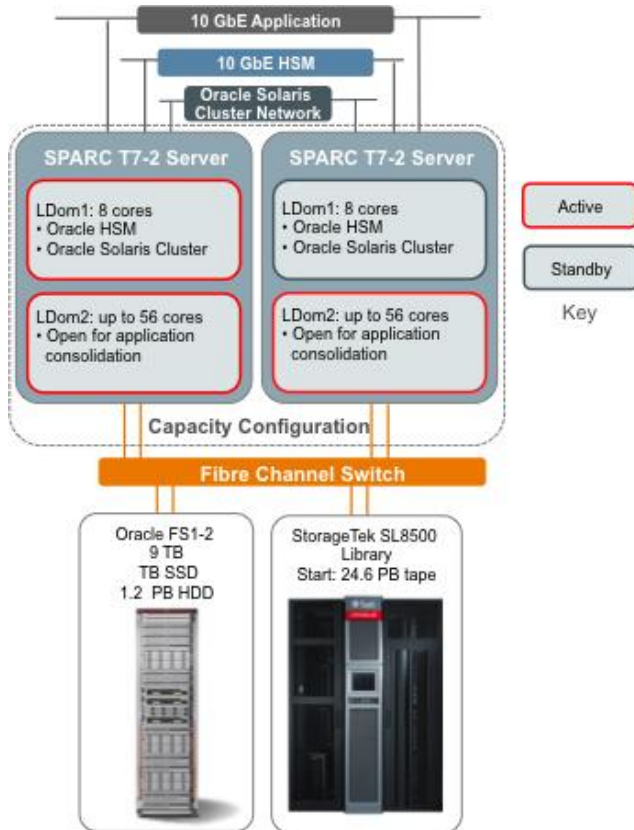


Figure 10. Capacity reference configuration with the Oracle ZFS Storage ZS3-4 appliance or the Oracle FS1-2 system and the StorageTek SL8500 modular library system shows capacity with the physical-to-logical LDom configuration for best performance with Oracle HSM.

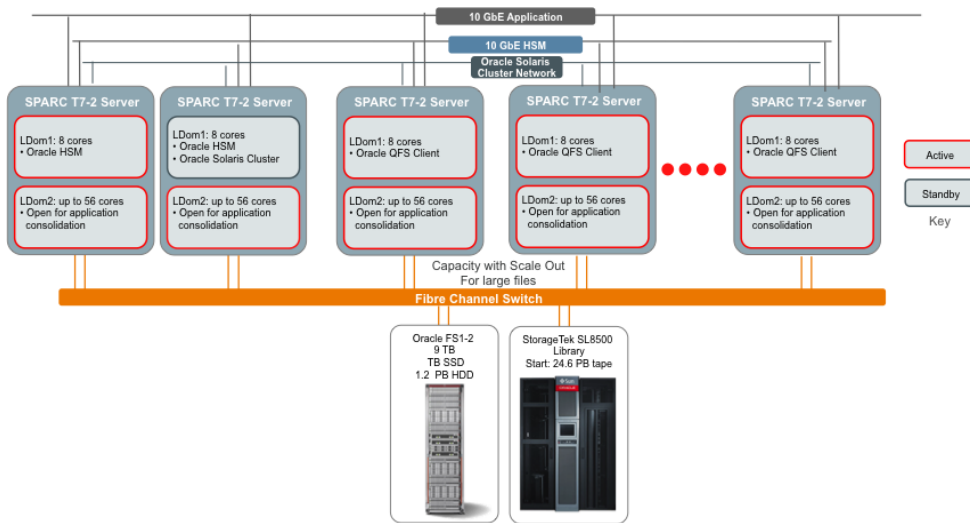


Figure 11. Capacity reference configuration for large files with the Oracle ZFS Storage ZS3-4 appliance or the Oracle FS1-2 system and the StorageTek SL8500 modular library system shows capacity with the physical-to-logical LDom configuration for best performance with Oracle HSM. This uses the scale-out feature of Oracle HSM.

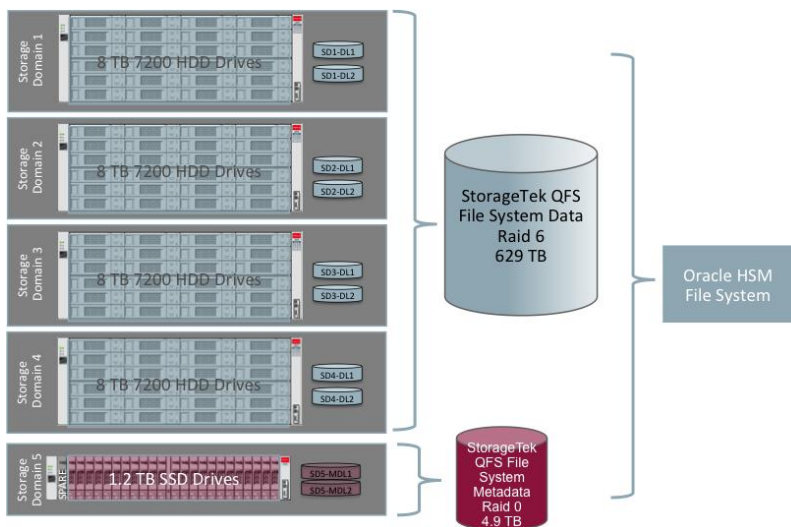


Figure 12. Large reference configuration for the Oracle FS1-2 disk layout and LUN assignment when ingesting and archiving large files. Additional LUNs can be created in each storage domain for multiple file systems.

SPARC Server Cluster Configuration

The SPARC T7 processor offers a multithreaded hypervisor—a small firmware layer that provides a stable virtual machine architecture that is tightly integrated with the processor. Corresponding layers of virtualization technology are built on top of the hypervisor. The strength of Oracle's approach is that all the layers of the architecture are fully multithreaded, from the processor up through applications that use the fully threaded Java application model. In

addition to the processor and hypervisor, Oracle provides fully multithreaded networking and the fully multithreaded Oracle Solaris ZFS file system.

Oracle VM Server for SPARC (also called logical domains or LDomS), Oracle Solaris Zones, and multithreaded applications are able to have allocated exactly the resources they need. In this solution, Oracle HSM is supported in an LDom, giving flexibility to the number of threads allocated. For the small and medium configurations, four cores are adequate for Oracle HSM, while eight cores are recommended for both the large and capacity configuration. This flexible allocation of resources helps control license costs for applications that are based on number of cores. Resources then are available to run additional applications in the same physical server.

Security Through Oracle Multitenant and Oracle HSM

Oracle Multitenant, an optional feature of Oracle Database, is used to configure and manage a StorageTek QFS multitenant environment and includes zero, one, or many customer-created pluggable file systems under Oracle HSM using any supported IP protocol, including NFS, CIFS, FTP, and OpenStack Swift. A multitenant environment is initiated by creating an LDom with the primary and secondary domain resources. Shared disks then are assigned to both the primary and secondary domains using the Oracle Solaris Zones special file system. Oracle HSM now can be installed on primary and secondary domains, followed by configuring the shared Oracle HSM file system as displayed in Figure 13.

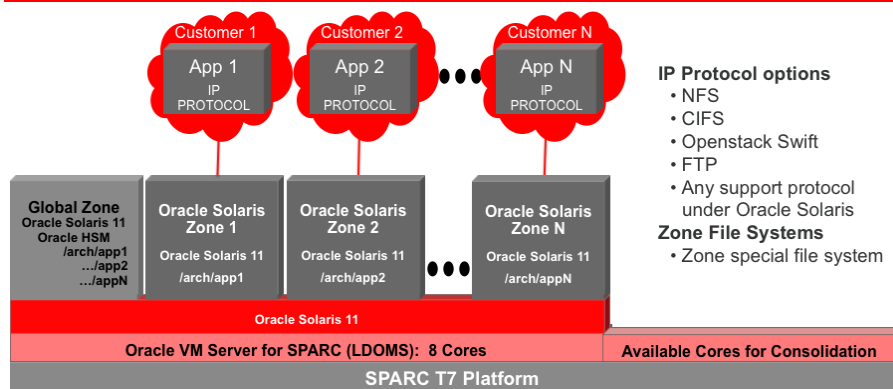


Figure 13. The Oracle HSM multitenant architecture is designed for mission-critical applications and delivers the industry's highest levels of scalability and investment protection.

- » Oracle Solaris Zones, including Kernel Zones (a feature of Oracle Solaris), and Oracle Solaris 11 provide a flexible, cost-efficient, cloud-ready solution for the data center, with the following attributes:
 - » Only one nonglobal zone per file system is supported.
 - » For Oracle Solaris Cluster, Oracle Real Application Clusters (Oracle RAC) is supported in a zone cluster.
- » Oracle VM Server for SPARC is supported with StorageTek QFS, with the following restrictions:
 - » A minimum of four cores should be assigned to the domain.
 - » The suggested minimum amount of RAM is 24 GB.
- » Oracle Solaris Cluster has the following requirements:
 - » Virtual storage devices used by shared StorageTek QFS must be backed by whole SCSI FC LUN disk arrays.
 - » Virtual storage devices must not be shared with any other guest domain on the same server.
 - » Virtualized partial LUN disk arrays are not supported.

» In addition, the following conditions apply:

- » The Oracle HSM-StorageTek QFS metadata server (MDS) must boot from a physical device and, therefore, it must have at least one PCI root complex.
- » Disk I/O virtualization is not supported for LUNs that are used in a QFS file system.
- » Network virtualization is supported.
- » Tape devices must be attached via nonvirtualized PCI slots attached to the Oracle HSM MDS server.
- » Oracle's StorageTek QFS client may boot from a virtualized disk; however, they still need a PCI root complex to access file system devices via PCI controllers (FC, SAS, and so on).

Security Throughout the Solution Stack

The following table identifies where security fits in the complete solution.

TABLE 3. SECURITY IN THE SOLUTION STACK

Function	Benefit
Oracle Solaris and Oracle Solaris Cluster	Provides security by default during installation through disabling a large set of network services
Oracle VM for SPARC and Oracle Solaris Zones	Provides separate execution environments called domains Establishes each domain as an independent instance Follows existing Oracle Solaris security guidelines to harden Oracle Solaris OS
Oracle Solaris Zones	Creates a single instance of Oracle Solaris OS within an application execution environment Isolates processes for an application from the system
Network partitioning	Combines network partitioning with Oracle VM Server for SPARC and Oracle Solaris Zones Enforces administration rights based on strictly restricted roles
Network isolation with partitioning	Creates a virtual local area network (VLAN) at the datalink layer Compartmentalizes data traffic through assignment of groups of users to VLANs, improving security per VLAN
Isolation of storage data traffic, application data traffic, and user data traffic	Isolates the storage network to the application Isolates application and database communication Isolates user access to the application
Physically separated management network	Isolates the complete network, switches, cables, and ports from other network traffic Prevents access from outside the data center (this is often a nonroutable address)
NFS file system exceptions	Restricts data access to specific hosts
Target groups and initiator groups used to secure block devices	Maps target LUNs to specific host initiators to ensure only appropriate clients are granted access to specific block devices
Use of physically separate networks and interfaces for data and administrative traffic	Isolates data traffic and administrative traffic to restrict a possible breach in security
Data integrity validation within the tape systems	Validates the integrity of the data on tape at creation time, access time, and as a scheduled validation process
Use of T10-PI for end-to-end data integrity	Prevents silent data corruption, ensuring that incomplete and incorrect data cannot overwrite good data and silent corruption is



Function

Benefit

identified on reads

Oracle FS1-2 Configuration and Performance Testing

The sections that follow detail best practices and performance testing results that were determined as a part of Oracle's testing.

Oracle Flash FS1-2 Configuration Best Practices

The following section describes the architecture of Oracle FS1-2 and its configuration with Oracle HSM.

Configuring Oracle FS1-2 and Oracle HSM

Oracle's patented QoS Plus technology, a feature of Oracle FS1 Series, provides a big differentiator over traditional controller-based disk storage. QoS Plus is a policy-based virtualization feature that incorporates business priority I/O queue management fused with subLUN auto-tiering into one simple management framework. It is delivered by prioritizing data access and ingest for different LUNs based on an assigned level of business priority.

Advanced QoS Plus software manages system resources (CPU, cache, flash, and capacity) to automate storage provisioning based on business priority for the components of Oracle Optimized Solution for Secure Tiered Storage Infrastructure that provide ingest, search, and access of content. QoS Plus performs data collection, evaluation, and movement based on the most efficient data granularity in the storage industry, making Oracle FS1-2 the most efficient auto-tiering system in the market. It is this flexibility that makes Oracle FS1-2 an excellent storage solution with Oracle HSM for managing unstructured data.

Oracle FS1-2 is designed to scale performance along with capacity. Unlike most storage systems, which have a fixed number of storage controllers (usually a maximum of two), Oracle FS1-2 can be scaled in multiple dimensions by independently adding more storage controllers or more trays of disks, SSDs, or both, as needed. Oracle FS1-2 is built on four intelligent hardware assemblies, as described below and shown in Figure 14.

- » Two pilot nodes per system, which provide an easy-to-use interface for managing physical and virtual configurations
- » Up to eight controller nodes per system in sets of two controller nodes
- » Up to 30 drive enclosures (DEs) per system; six strings with five drive enclosures per string
- » Up to eight optional replication engines per system; up to four engines can be installed and shipped with racked systems

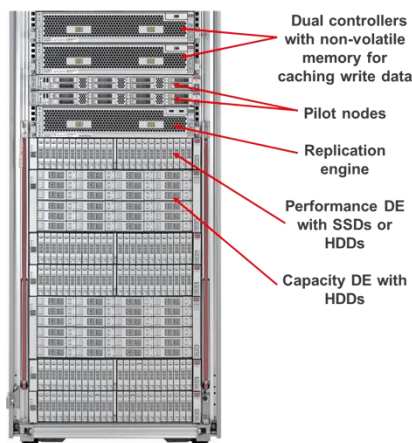


Figure 14. The Oracle FS1-2 components.

The Oracle FS1-2 flash storage system components can be flexibly combined to meet unique application performance and storage capacity requirements. This flexibility is especially valuable to an application that takes advantage of tiered storage.

Figure 15 shows the physical connection of the SPARC servers to the Oracle FS1-2 controllers and the Oracle tape archive system for a high-availability configuration.

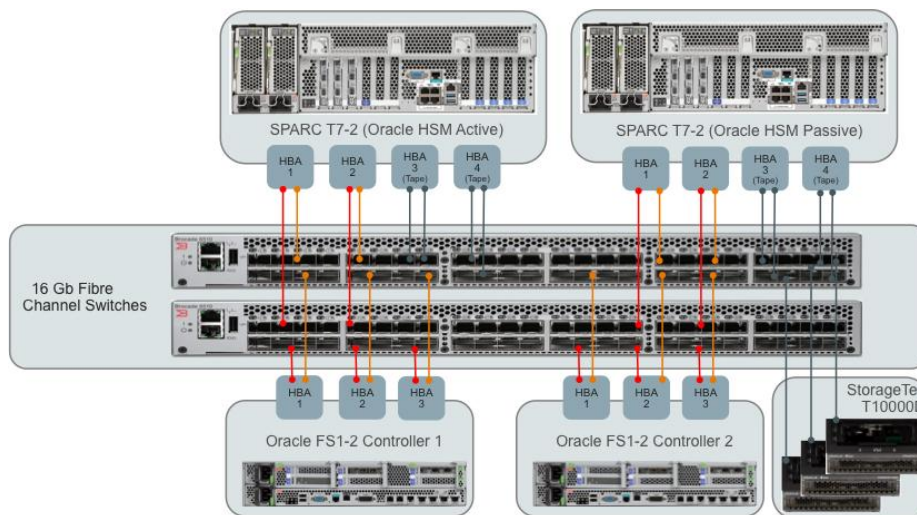


Figure 15. Physical configuration of Oracle FS1-2 for high availability.

LDoms provide applications with highly efficient, enterprise-class virtualization capabilities that provide fully dynamic resource management on supported SPARC servers. For this testing, server resources are allocated to Oracle HSM software and for Oracle Solaris Cluster, using LDoms, to meet the ingest requirements for the small, medium, and large reference configurations. This core assignment was varied in order to identify the least number of cores possible and still get the best performance.

Configuring LUNs on Oracle FS1-2

This section describes the best practices for configuring LUNs for Oracle HSM and Oracle FS1-2.

Figure 16 describes the physical-to-logical configuration of the Oracle FS1-2 storage trays in the test environment. The logical components are created using the Oracle FS1-2 management user interface. The Oracle HSM file system uses the option of metadata separation, meaning the metadata and content are on separate volumes. The Oracle HSM metadata resides on one logical device, and the primary content resides on a different logical device.

This configuration enables the storing of the metadata, which is small in size, on the highest performing storage device, and the storing of the primary content, which is the actual data, on the next highest performing storage device. The test of the 1 billion files resulted in the metadata consuming about 8 GB in capacity. The number of files, not the capacity, determines the size of the metadata. The logical LUNs to be used for the Oracle HSM data files are mapped to the server and all LUNs then are configured into the Oracle HSM file system and presented to the application as a single POSIX-compliant file system.

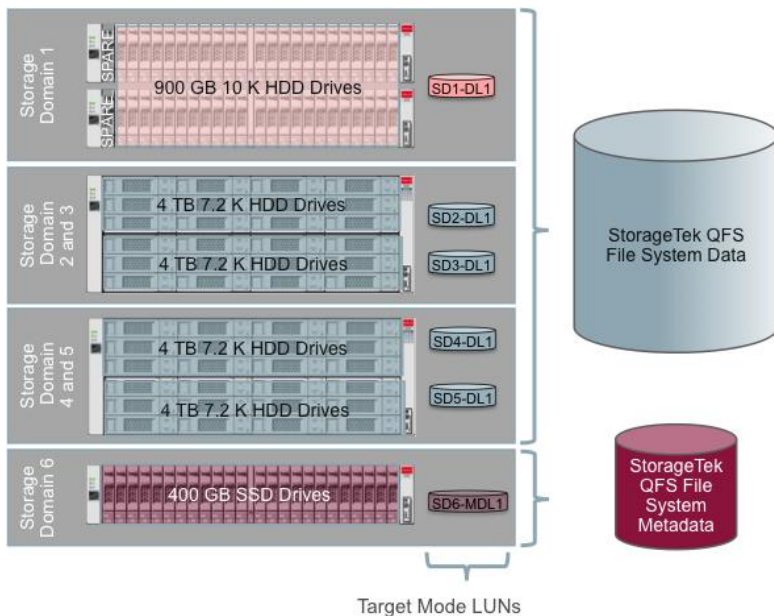


Figure 16. The Oracle FS1-2 pools and LUN allocation as configured for the functional and comparison testing for Oracle HSM primary storage show how the storage pools can be configured for the primary content and metadata and presented to Oracle HSM.

Configuration guidelines for the content pools and for the metadata pools are as follows:

- » **Disk allocation unit (DAU) in Oracle HSM.** The Oracle HSM DAU (stripe width) is important for write performance. The DAU setting is the minimum amount of contiguous space that is used when a file is written. Each file system has its own DAU. The stripe width specifies the size of the blocks to be written to a single LUN before switching to the next LUN. The Oracle HSM DAU has been found to work best when set to 128 K when using the Oracle FS1-2 flash storage system for disk cache.
- » **Metadata LUNs.** In the past, the scalability of an Oracle HSM file system to achieve the maximum number of files was greatly impacted by the time required to run the `samfsdump` command to back up the metadata. As a result, file systems were intentionally restricted in size in relation to the number of files in a single file system in order to achieve the required performance. The combination of Oracle FS1-2 and Oracle HSM removes that restriction. Testing proves it is possible to grow an Oracle HSM file system to greater than 1 billion files, and the `samfsdump` command as well as the ingest of daily files, is not impacted.
- » **Data LUNs per file system.** Within a storage domain, the Oracle FS1-2 writes to device groups within a storage domain. As a result, the number of LUNs for a single file system within that storage domain should be one in order to avoid drive contention. If the storage domain is larger than 128 TB, create multiple LUNs of equal size but as few as possible, mapping them to different controllers. Additional LUNs can be created within this storage domain for other file systems or applications, and the Oracle FS1-2 QOS Plus feature manages the performance. All LUNs created in multiple storage domains for a single Oracle HSM file system then are defined to a single Oracle HSM file system, spreading the I/O across all of the disks.
- » **Oracle FS1-2 stripe width.** Oracle FS1-2 stripe width determines how many device groups within a storage domain are used for a single LUN. The default *Auto-select* sets it based on the QOS Plus setting. For example, if the QOS setting is *high*, it uses four device groups. If QOS is *medium*, it uses three. If one or two device groups are in a storage domain, a QOS setting of *high* sets it to four and force writes to reuse a device causing disk contention. As a result, it is best to define a storage profile for the LUNs in the Oracle FS1-2 and set the stripe width to *all* or set it to the exact number of device groups within the storage domain. If there are two device groups, set it to two.

» **Oracle Solaris format command.** The format command is used to label the LUNs presented from Oracle FS1-2. For each LUN, the starting sector of the first partition is always 1,280 for Oracle FS1-2. The remaining capacity is to be allocated for the next partition, leaving the default end partition alone. It is necessary to use EFI labels (format -e) due to the size of the LUNs.

Additional information for configuring LDOMs for Oracle HSM can be found at <http://docs.oracle.com/cd/E19604-01/821-0406/>.

Oracle FS1-2 and Oracle HSM Test Results

Figure 17 shows the configuration used for the baseline performance tests that follow in this section. The multiple LUNs from all the SSD and HDD storage devices are mapped to the Oracle HSM server using a stripe width of two, to create the Oracle HSM file system with three content domains.

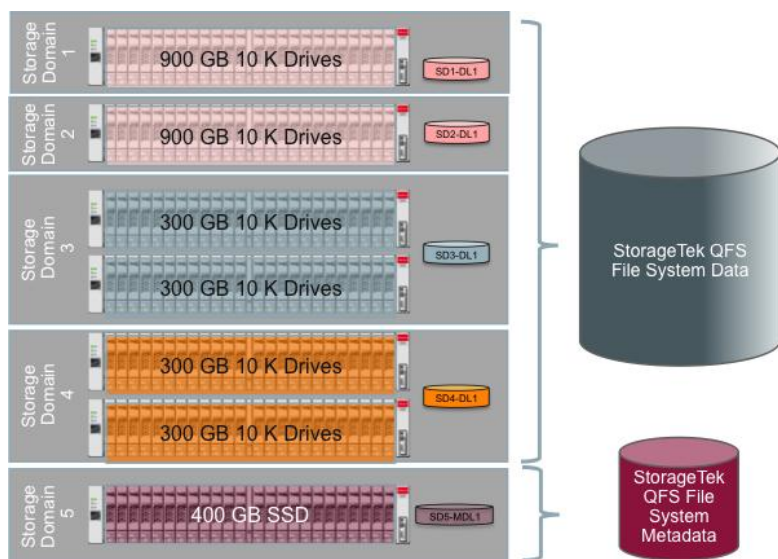


Figure 17. The Oracle FS1-2 baseline performance-testing configuration shows the physical and logical configuration for the best performance with Oracle HSM.

All testing and test results are intended for use in configuring Oracle HSM and Oracle FS1-2 for Oracle Optimized Solution for Secure Tiered Storage Infrastructure. The results can be used as a guideline for similar performance and capacity requirements. The tests are not intended to be full-performance testing for the purpose of general Oracle HSM and Oracle FS1-2 use or for pushing servers and storage to their maximum performance.

The reported test results are based on writing directly to the Oracle HSM file system and recording the highest rates achieved with the configuration available. The performance is optimal for the number of paths configured, reflecting that the storage side of the controllers allows the ingest rates to run at optimal speed.

Throughput for File Ingest

The `fstest` command results shown in Figure 18 with a workload that ingests 1,028 million records indicate that the CPU running the Oracle HSM software is not dependent on the number of cores. Consequently, unless ingest requirements are pushing the limits of the system, the use of either four or eight cores provides the needed performance for the reference configurations and reduces license costs.

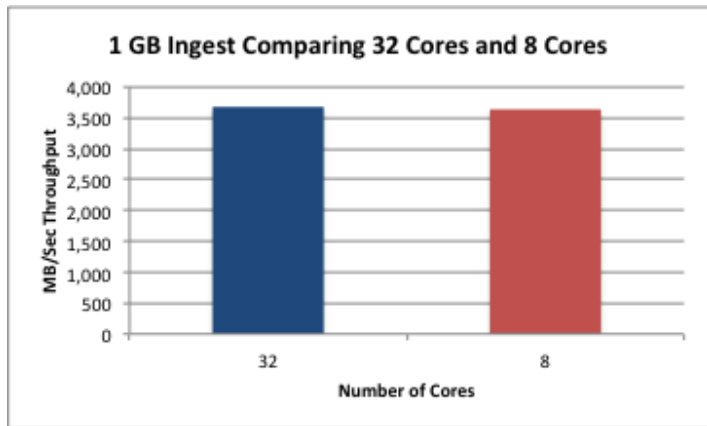


Figure 18. Comparison of ingesting 1,028 MB files and comparison of 8-core with 32-core LDOMs while increasing the workload.

Linear Scalability for Dump and Ingest

Metadata performance has a significant impact on most file systems. However, with the intrinsic high performance provided by Oracle FS1-2, the Oracle HSM `samfsdump` command test results shown in Figure 19 demonstrate that there is linear scalability for the time required—when additional files are added to the Oracle HSM file system. This scalability is tested for up to 1 billion files.

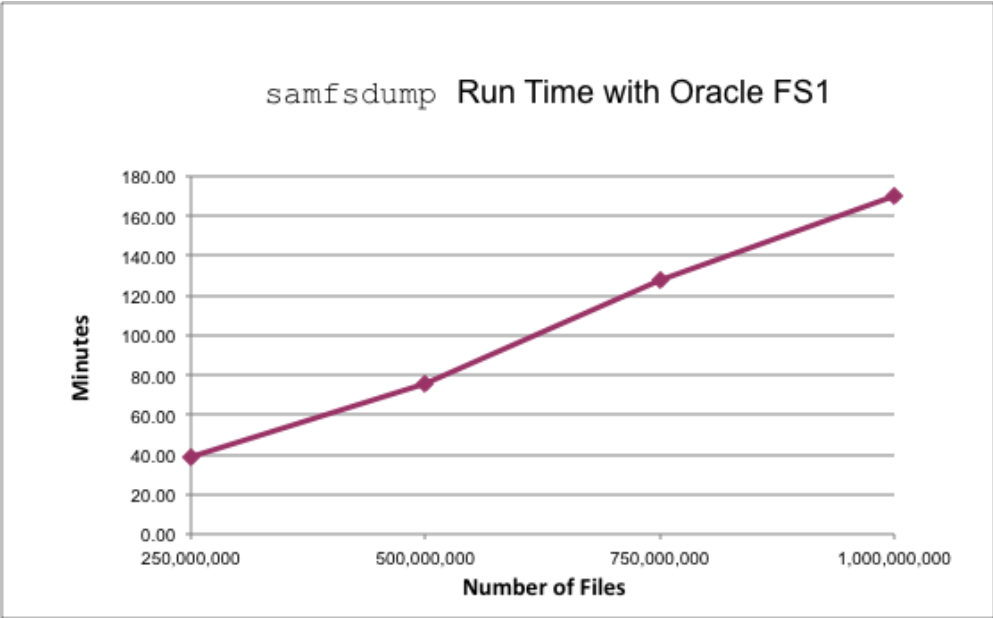


Figure 19. Linear scalability for a 100-million file `samfsdump` command test is achieved even as the file system grows to 1 billion files.

Historically, in traditional file systems, performance decreases as more files are added to a file system, but this is not true with Oracle HSM and Oracle FS1-2. Based on the tests performed by Oracle, linear scalability is achieved no

matter how many files are in the file system, and the ingest time is always the same. The benefit is that performance does not degrade as the file system increases in file count, as shown in Figure 20.

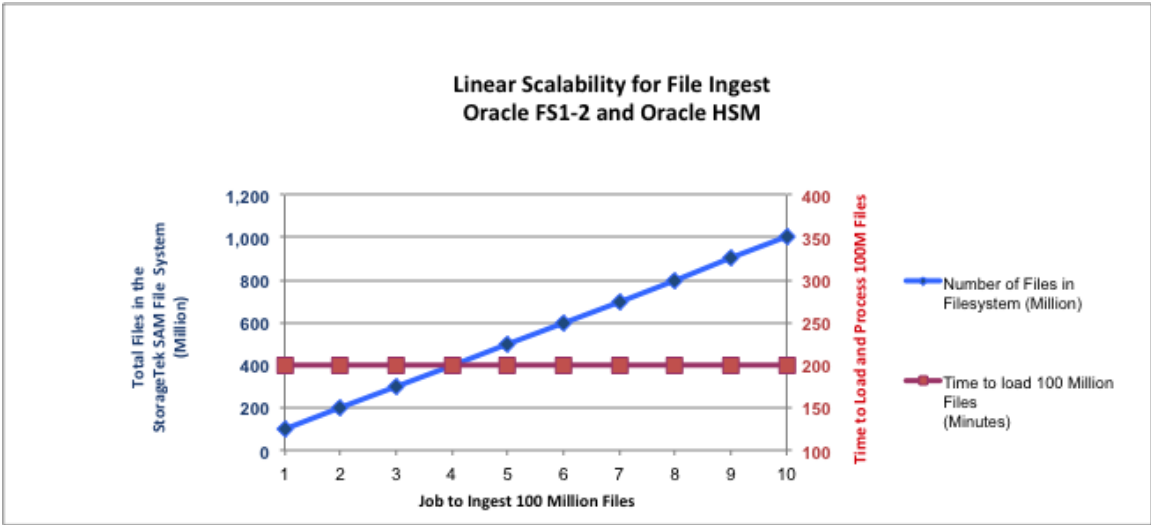
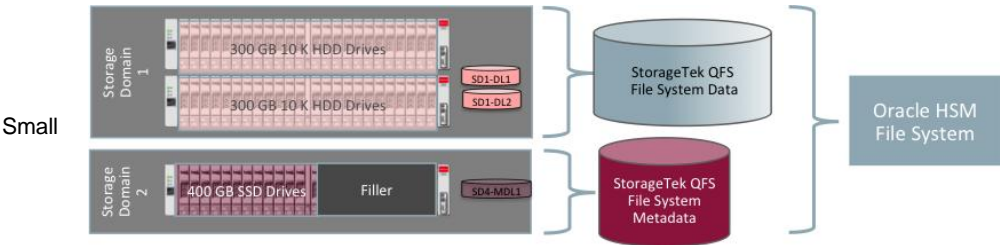


Figure 20. Time (in red) to ingest 100 million files as the file system (blue line) grows to 1 billion files.

Ingest Testing

For small- and medium-capacity disk cache, Figure 21 illustrates the configuration for the Oracle HSM metadata and content for the testing. For both sizes, the metadata is on a disk enclosure of seven performance SSD drives configured as RAID 10. For the small content part of the disk cache, two disk enclosures of 900 GB HDDs are configured as RAID 5 and two LUNs are created, one mapped to each controller. For the medium, the metadata is identical to the small configuration and the content part of the file system is four disk enclosures of 900 GB HDDS configured as RAID 5, and eight LUNs are created for the file system. For the large disk cache, the metadata is 13 performance SSD disk enclosures using four LUNs, two on each disk group. The content part of the large disk cache is six 900 GB HDD disk enclosures configured as RAID 5 with 12 LUNs that make up the Oracle HSM file system.



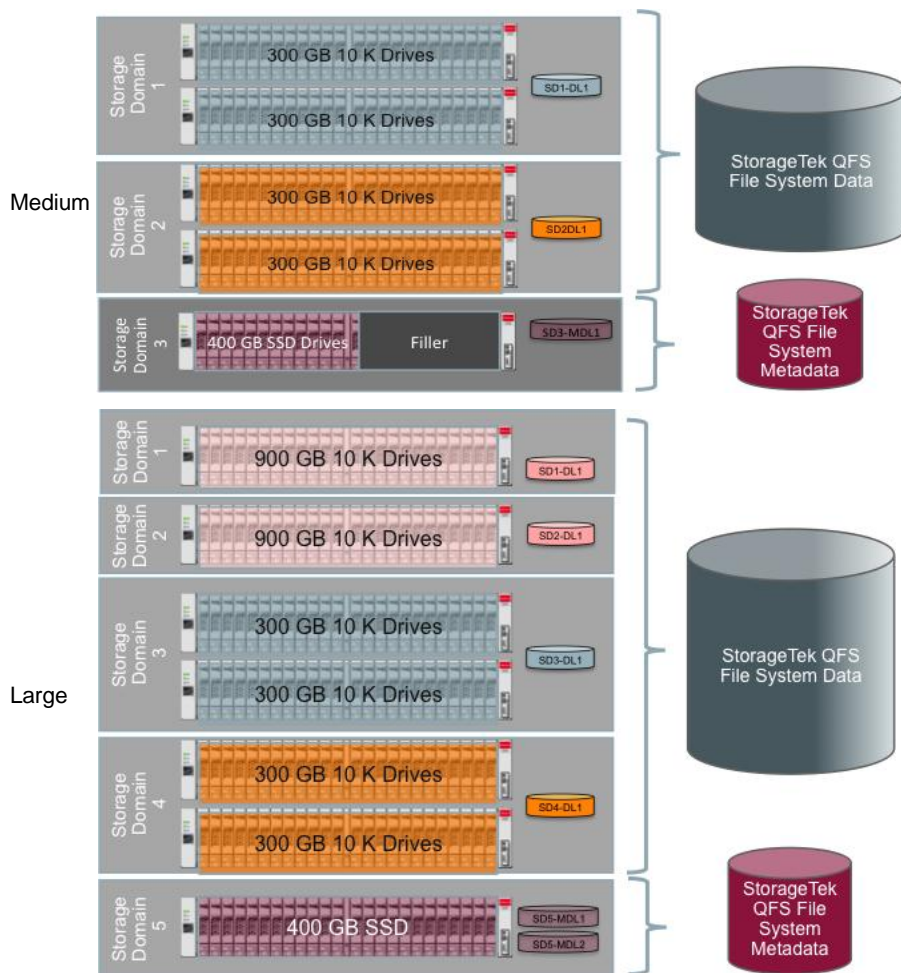


Figure 21. The Oracle FS1-2 disk configuration for Oracle FS1-2 flash storage small-, medium-, and large-capacity Oracle HSM disk cache.

The test results of ingesting 1 GB files are shown in Figure 22 below. As the number of storage devices increases, the performance also increases. This is a testament to both the power of Oracle's SPARC T7 systems and the efficiency of the StorageTek QFS file system. What clearly does make a difference is the number of spindles employed to store the data. In this case, more is better.

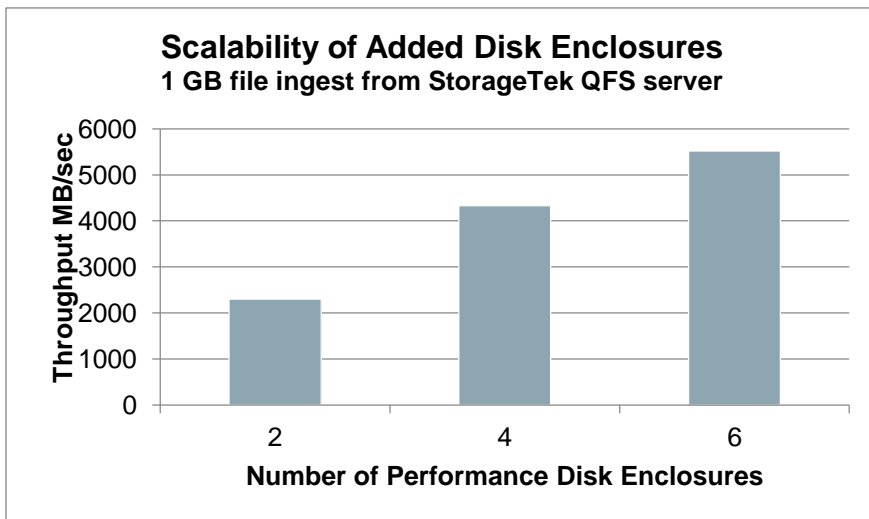


Figure 22. Test result of ingesting 1 GB files into small, medium, and large Oracle HSM disk caches from Oracle's StorageTek QFS Client illustrates scalability as the number of disk enclosures increases.

Workload, Performance, and Architecture

Testing with various workloads and architectures identifies specific configurations for specific workloads. With ingestion of small files that are less than 100MB, the performance is best when ingesting on the Oracle HSM metadata server (MDS) and not running in a shared QFS configuration. The reason for this is that for each file written by Oracle's StorageTek QFS Client, the metadata server must write the metadata. This instruction is sent from the client to the metadata server over the IP network, causing a latency. For large files, the majority of the throughput takes place by StorageTek QFS Client. For small files, more instructions are sent over the IP network to the metadata server to write the metadata.

For the large files, there is little impact to performance when running in a shared or unshared environment.

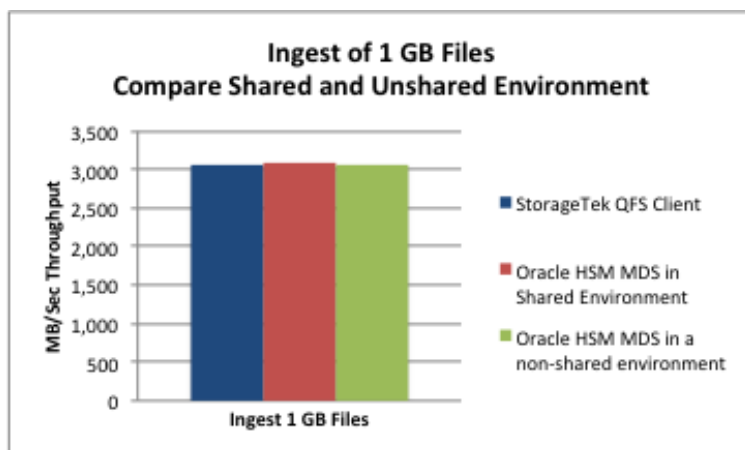


Figure 23. Comparison of the ingest of 1 GB files on the Oracle HSM metadata server in a non-shared environment, Oracle HSM metadata server in a shared configuration and StorageTek QFS Client.

For small files that are 100 KB in size, there is not a big impact on performance when running on the StorageTek QFS Client compared to the Oracle HSM metadata server running in a shared configuration. However, the greatest impact is when running on the Oracle HSM metadata server when running in a non-shared configuration. The following graph shows a 10x difference between shared and non-shared configuration.

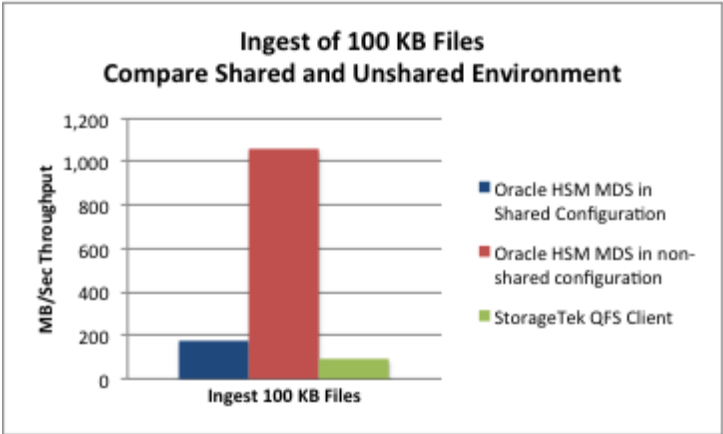


Figure 24. Comparison of the ingest of 100 KB files on the Oracle HSM metadata server in a shared environment, Oracle HSM metadata server in a non-shared configuration and StorageTek QFS Client.

The use of NFS adds another level of protocol to ingest files into the Oracle HSM file system; however, this is often the easiest method to access the file system. Through spreading the workload across multiple NFS clients, the ingest rate of four clients almost reaches the same workload running on a single Oracle HSM server. In the figure below, the Oracle HSM metadata server performance is achieved through using a thread count of 32. For the NFS client result in Figure 24, each NFS test used eight threads.

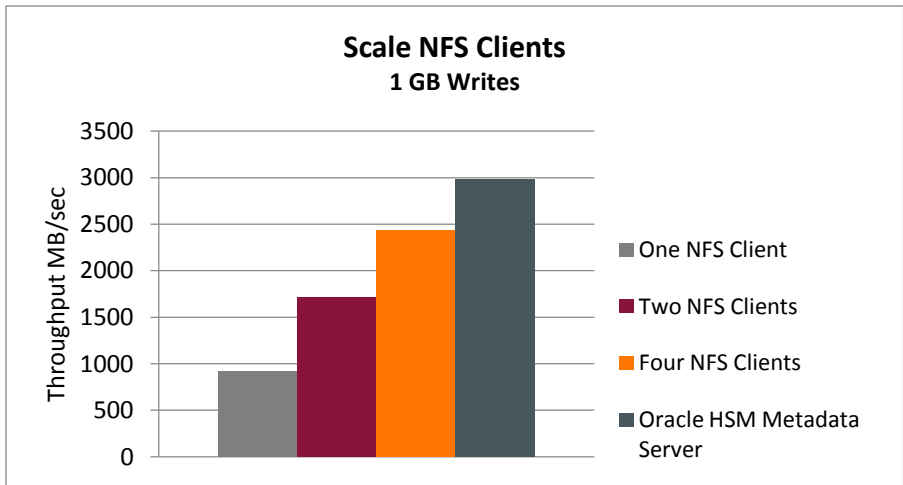


Figure 24. Comparison of ingest of 1 GB files on the Oracle HSM metadata server and one, two, and four NFS clients.

To summarize the discussion about workloads and architectures, the best practice is to understand the expected workload and method of ingesting the data into the Oracle HSM disk cache. For large files, scaling horizontally results in the best performance. If the application does not run on one of the following operating systems or cannot

run on the same server with the StorageTek QFS client, then connecting through an increased number of TCP/IP servers helps the system reach the needed performance:

- Oracle Linux 6.5, 6.4, 6.3, 5.10, 5.9 (with default kernel)
- RedHat 6.5, 6.4, 6.3, 5.10, 5.9 SMP RHEL AS and ES (via OL)
- SUSE 11 Service Pack 1 smp sles

Figure 22 can be referenced for horizontal scaling for ingest and access workloads as well as for distributed I/O for tape archive and staging.

Capability to Archive 10 Times More Data per Day

Oracle HSM archive and staging processes are now horizontally scalable and enable you to increase the total throughput of the archive by tenfold compared to previous releases of Oracle HSM and to archive multiple PB of data per day. I/O from StorageTek QFS Client servers can be used to archive and stage files to and from the disk cache to tape, preventing the metadata server (MDS) from becoming a bottleneck. Up to nine instances of StorageTek QFS Client running Oracle Solaris, in addition to the MDS, can be used to archive data to tape and stage data from tape. Figure 25 is an architecture diagram using scale-out tape for additional performance when you archive and stage data. This configuration uses Oracle HSM manual failover and not Oracle Solaris Cluster for the Oracle HSM MDS.

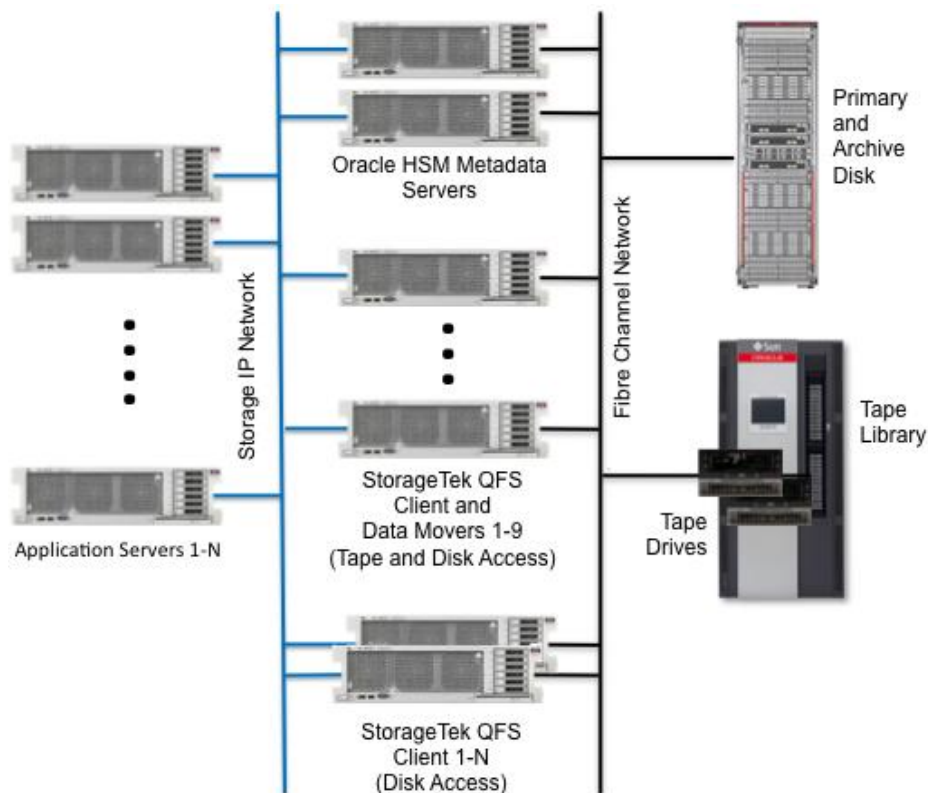


Figure 25. Oracle HSM with tape I/O horizontal scalability.

Access from a server using the NFS or CIFS protocols have varying results due to the performance characteristics of these protocols. Additional application servers can be implemented to add additional throughput as needed.

Oracle ZFS Storage Appliance Configuration and Performance Testing

Oracle ZFS Storage Appliance Configuration Best Practices

Oracle ZFS Storage Appliance radically simplifies the complexity of and reduces time spent on storage management for enterprises requiring both high-performance storage as well as high capacity for disk archive by providing a dramatically easy and fast way to manage and scale the storage. The administration toolset features simple installation and configuration, capacity scaling, tuning, and problem-solving capabilities. The intuitive browser user interface makes it possible to rapidly deploy powerful advanced data services such as snapshots, clones, thin provisioning, four different compression algorithms, and replication. The DTrace Analytics feature of Oracle ZFS Storage Appliance provides real-time analysis and monitoring functionality, enabling unparalleled fine-grained visibility into statistics for disk, controller CPU, networking, cache, virtual machine, and other items, in a way that uniquely ties client network interface activity back to the disks with everything in between.

In a tiered storage environment, for the highest tier, the Hybrid Storage Pool feature of Oracle ZFS Storage Appliance introduces an easy method for writing data to the correct high-speed storage without the need for users or storage administrators to make decisions. The Oracle ZFS Storage Appliance file system—Oracle Solaris ZFS—seamlessly optimizes performance by recognizing I/O patterns automatically and places data on the best storage media using Hybrid Storage Pool technology. The combination of read- and write-optimized flash accelerators with high-performance HDDs delivers optimal performance for primary storage for Oracle HSM, while higher capacity HDDs meet the requirements of a disk archive.

For example, Oracle Solaris ZFS transparently executes writes to low-latency SSD media so that writes can be acknowledged quickly, allowing the application to continue processing. Oracle Solaris ZFS then automatically flushes the data to HDDs as a background task. This dynamic use of SSDs by Oracle ZFS Storage Appliance adds to the range of tiered storage that is managed by Oracle HSM. Figure 26 shows the combination of these disk architectures. This configuration of storage also provides both SAN and NAS interfaces to meet the requirements of Oracle HSM as it manages multiple copies of the data on multiple tiers of storage.

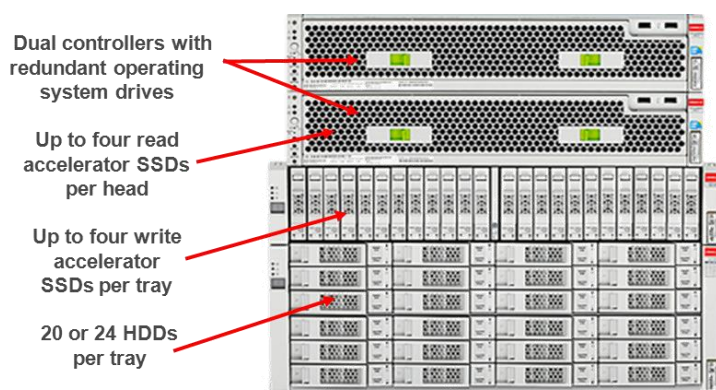



Figure 26. The Oracle ZFS Storage ZS3-2 system with dual controllers and a mix of SSDs and HDDs deliver the flexibility and performance required for a tiered storage environment.



Configuration tasks also are dramatically simplified in Oracle ZFS Storage Appliance through the browser user interface, which takes the guesswork out of system installation, configuration, and tuning. While testing Oracle Optimized Solution for Secure Tiered Storage Infrastructure, the Oracle ZFS Storage Appliance analytics were used extensively as different number of LUNs, different zpool configurations, different paths, and different Oracle HSM configuration parameters were used.

Best Practices for Configuring Oracle ZFS Storage Appliance and Oracle HSM

Figure 27 shows how the storage pools can be configured for the primary storage for Oracle HSM.

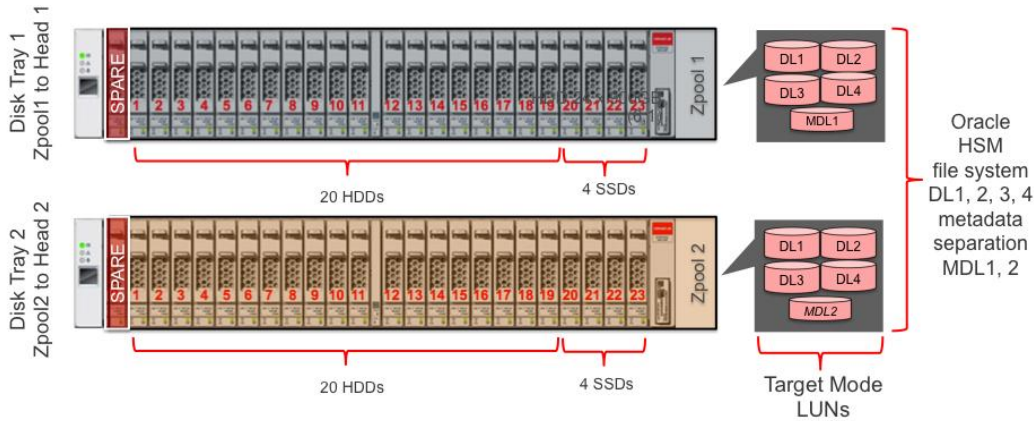


Figure 27. The Oracle ZFS Storage Appliance pools and LUN allocation for Oracle HSM primary storage.

Configuration settings for the content pools and for the metadata pools are as follows:

- » Four pools with five disks from each tray
- » Four data LUNs per pool
 - » 128 kB record size
 - » secondarycache = none
 - » Throughput for direct writes to the StorageTek QFS file system
- » One metadata LUN per pool
 - » 16 kB record size
 - » logbias = latency
 - » secondary cache = all

Block storage is required for Oracle HSM primary storage; therefore, LUNs on Oracle ZFS Storage Appliance are created and mapped to the server. Oracle HSM then is used as a volume manager and configures 1 to 4 LUNs into a single small file system for the metadata and configures up to 16 LUNs into a single file system for the content.

Oracle ZFS Storage Appliance and Oracle HSM Test Results

The following results apply to the small, medium, and large configurations when you use Oracle ZFS Storage Appliance as the primary storage for Oracle HSM.

Testing on a medium and a large configuration indicates that for this workload of writing 100 K files, a zpool that is created using SSD and 15 K drives and configured as RAID Z provides similar performance as the same combination of SSD and 15 K drives using RAID 10; therefore, a RAID Z environment is ideal for also achieving the best available capacity. The comparison shown in Figure 28 supports this.

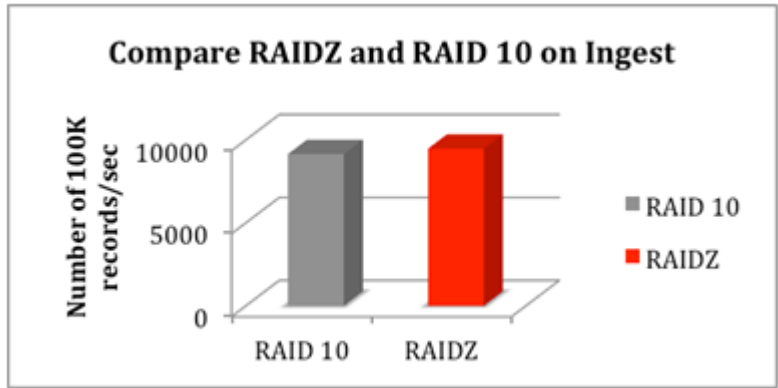


Figure 28. Comparison of RAID Z and RAID 10 for a workload of ingesting 100 K records results in slightly better performance and a nice side effect of better disk usage with RAID Z.

Oracle's Modular Tape Systems Configuration

Oracle's tape system strategy is to provide read access for up to three generations of tape media from the newest tape drives. This feature gives businesses many years of data access without migrating the content and provides the ability for you to take advantage of the latest tape drive technology and media for new archives. As tape media becomes more dense and appealing to a data center, Oracle HSM provides the tools to migrate the content to this new technology. Table 4 provides an overview of the configuration choices for the three libraries that are available from Oracle.

TABLE 4. LIBRARIES AT A GLANCE




	StorageTek SL150 Modular Tape Library	StorageTek SL3000 Modular Library System	StorageTek SL8500 Modular Library System
			
Number of cartridge slots	30 to 300	200 to 5,925	1,450 to 100,000
StorageTek T10000D capacity	n/a	1,700 TB to 50,362 TB	12,325 TB to 850,000 TB
StorageTek LTO 6 capacity	75 TB to 750 TB	500 TB to 14,812 TB	3,625 TB to 250,000 TB
Maximum number of tape drives	20	56	640
Maximum native throughput (TB/hr.)	11.5	48.4	552.9

Table 5 provides a description of the features of the two tape drives tested and recommended for Oracle Optimized Solution for Secure Tiered Storage Infrastructure.

TABLE 5. STORAGETEK TAPE DRIVES AT A GLANCE

	StorageTek T10000D Tape Drive	StorageTek LTO 6 Tape Drive
Media capacity	8.5 TB	2.5 TB
Throughput	252 Mb/sec	160 Mb/sec
Generations of media support	3	3
Data integrity validation	Yes ¹	Yes ²

When selecting a library and tape drives for Oracle Optimized Solution for Secure Tiered Storage Infrastructure, collect the following information

- » Retention period of the content, which contributes to capacity requirements
- » Number of copies on tape (a minimum of two is recommended), which contributes to capacity requirements
- » Current content capacity to be archived
- » Daily content capacity to be archived, which contributes to performance requirements
- » Estimated yearly growth of content
- » Other applications that share the library
- » Estimated activity of staging from tape to disk
- » Scale-out tape requirements; deployment of up to 10 servers for running archive and stage processes
- » Requirements for data integrity validation (DIV)


Library Management Applications

The following tape library management applications are deployed separately from Oracle HSM:

- » Oracle's StorageTek Tape Analytics software proactively monitors StorageTek library environments to ensure the health and data availability of the global tape infrastructure. StorageTek Tape Analytics software is changing the way the world manages tape by moving from a reactive approach to a proactive, predictive approach. This 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. A proactive approach to managing the health of a tape environment improves the performance and reliability of existing tape investments.
- » Oracle HSM uses a high-performance file system and includes a solution to automatically audit the archive to verify the integrity of all data. Data is often archived for several years or in some cases forever. It is a requirement to be able to verify that the data archived many years ago is still accessible. The integration of Oracle HSM with the Data Integrity Validation feature of the StorageTek T10000D tape drive is a policy-driven activity that periodically loads the media into a drive and validates the accessibility and validity of the data. This test is performed within the tape drive itself and does not require data to be sent back to the host server. Data Integrity Validation automatically self-heals if a problem is found, using alternate copies if necessary, and provides reports detailing what is discovered and when it is corrected.
- » Oracle's StorageTek Automated Cartridge System Library Software is used to configure and control the tape drives.

¹ StorageTek T10000D uses CRC as defined in ANSI X3.139 and is calculated in the chip for no impact to performance.

² LTO uses the Reed Solomon CRC calculation in software at an 88 percent decrease in performance. This is supported in IBM Tivoli.



It is critical to ensure that stored data is recorded accurately and just as important to be sure it remains unchanged through its retention time. The Data Integrity Validation feature takes this one step further by validating CRC checksums generated at the host. This integrity check for write, read, and validate has the highest level of importance when storing data that might be kept forever and yet has low access requirements.

Oracle HSM's fixity capability augments the Data Integrity Validation feature. Data Integrity Validation is calculated at the *block* level. Fixity is calculated at the *file* level and has the ability to receive the hash from the application that originally stored the data. Both can be used in conjunction with each other to ensure the integrity of the data from intended and unintended data corruption.

Historically, data written to tape is verified after it is written or validated at the inefficient full-file level. Data Integrity Validation starts this process at the server on a record level and continues the CRC check throughout the write process until the data is written to tape media. StorageTek T10000D again validates the data at the drive and at the server when it is read.

Performance on the server during the validation-only step is not affected because the validation does not require data to be staged back to the server for the CRC check process. The validation is executed on the drive in the background, and the server is notified to take action only if an error is detected. This process ensures all media, even media that contains dark archive files and is rarely if ever accessed, is loaded into the drive and read on a schedule, such as yearly or every six months. A migration to new media after many years of being stored in a library slot is not the first time a file is read.

Data Integrity Validation Process for Write, Read, and Validate

The step-by-step write, read, and validate processes are as follows:

- » Write steps (from server to the StorageTek T10000D media):
 - » File is written by the application to an Oracle HSM file system.
 - » Policy states that the file should be archived on tape and the archiver goes through the archive process.
 - » Oracle HSM calculates the 32-bit CRC specified in ANSI X3.139 on each 2 MB record.
 - » This four-byte CRC is added to each 2 MB record (configurable size) as it is sent to the StorageTek T10000D tape drive.
 - » The StorageTek T10000D tape drive receives and recalculates the CRC and compares it to the CRC on the record.
 - » No match = StorageTek T10000D sends a request to the server to resend the record.
 - » Match = StorageTek T10000D writes the record to tape.
- » Read steps (from StorageTek T10000D to server):
 - » The server sends a request for the file.
 - » Media is loaded into the StorageTek T10000D tape drive.
 - » The StorageTek T10000D drive reads the records and recalculates the CRCs and compares them to the CRC attached to each record.
 - » No match = StorageTek T10000D sends a message to Oracle HSM, which stages the data from a second archive file and schedules the file for rearchiving.
 - » Match = StorageTek T10000D sends the record to the server.
 - » Server receives the record and recalculates the CRC and compares it with the CRC attached to record.
 - » No match = StorageTek T10000D sends message to Oracle HSM; Oracle HSM stages the data from a duplicate archive image and rearchives it.
 - » Match = StorageTek T10000D sends the record to the application.
- » Validate steps:

- » Based on policy, Oracle HSM sends a command to validate a range of StorageTek T10000D media.
- » Media is loaded into the StorageTek T10000D tape drive as a background task.
- » The StorageTek T10000D tape drive recalculates the CRC and compares it to four bytes on the record.
 - » No match = StorageTek T10000D sends a message to Oracle HSM, which stages the data from another archive image and rearchives it.
 - » Match = StorageTek T10000D sends a success status to Oracle HSM and unloads the media.

Performance Implications of Data Integrity Validation

Both SPARC- and x86-based servers have the option to generate the required CRCs in the chip versus in software, requiring very little processor overhead. Test results from driving 10 StorageTek T10000D tape drives at optimal speed prove this. Figure 26, which compares using Oracle's Data Integrity Validation to not using data integrity validation, shows not only that data integrity validation has zero impact on performance but also the scalability of adding StorageTek T10000D tape drives.

The test environment for all the performance results graphed in Figure 29 uses three StorageTek T10000D tape drives per 8-Gb FC HBA port. Data was read from two of Oracle's dual-controller Pillar Axiom 600 storage systems and written to tape in parallel. Data integrity validation mode was turned on. The first bar shows the zero-performance impact of Data Integrity Validation on Oracle HSM archiving to 10 StorageTek T10000D tape drives.

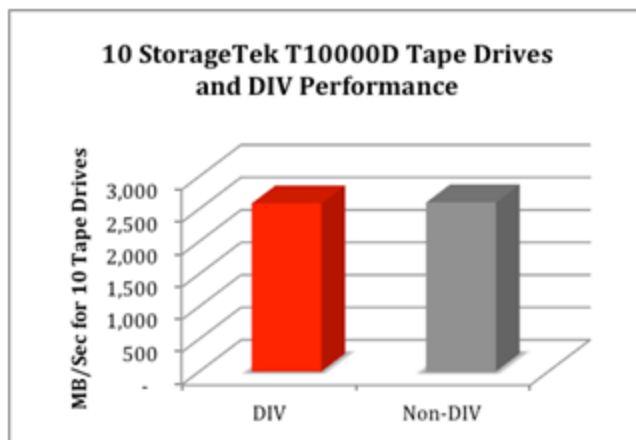


Figure 29. Writing to 10 StorageTek T10000D tape drives in parallel with data integrity validation (DIV) mode turned on shows no performance impact.

As tape drives are added, the amount of data archived by Oracle HSM scales perfectly with the number of tape drives, as Figure 30 shows.

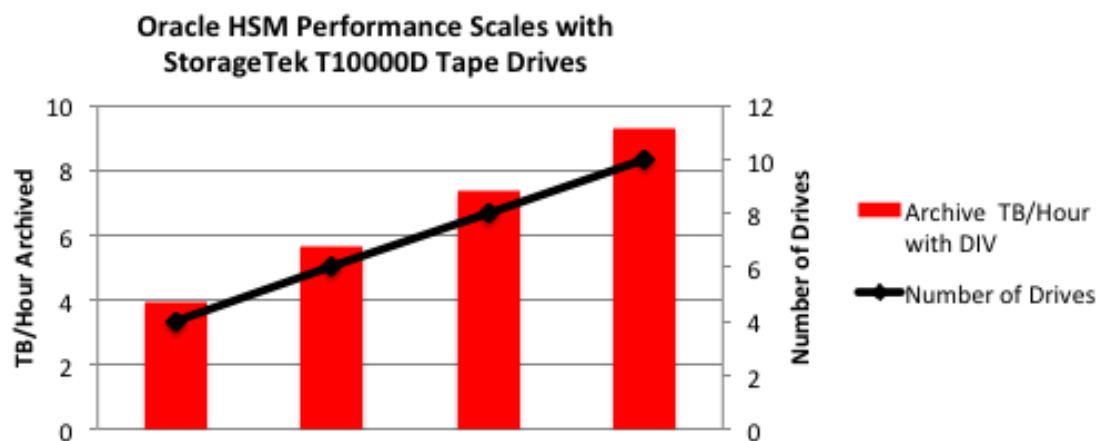


Figure 30. Adding two StorageTek T10000D tape drives at a time provides perfect scalability.

Looking at the data in another context in Figure 31—in addition to perfect scalability, as new drives are added, there is no impact on the performance of the drives already running as long as the disk system can read the data as fast as the tape drives can write the data. A drop in performance is more a function of the disk system than the tape system.

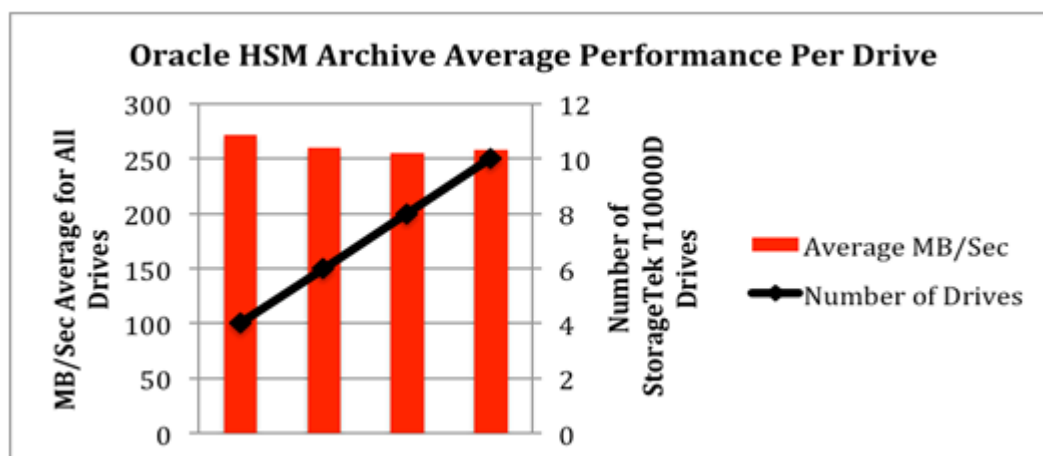



Figure 31. Additional StorageTek T10000D tape drives do not affect the performance of any tape drive.



With vast amounts of critical data being stored digitally, it is essential that the data remain unchanged during transfer from server to tape media and back as well as when it is stored for long periods of time. For legal and preservation purposes, the fixity of this data must be verified. The Data Integrity Validation feature of the StorageTek T10000D tape drive allows Oracle HSM to use CRC checksums to ensure the integrity of archived data is preserved.

Conclusion

Designed to address the challenges of rapid data growth and data management challenges associated with active archiving, Oracle Optimized Solution for Secure Tiered Storage Infrastructure automates data management processes to help organizations save time and money. The solution employs powerful, policy-based storage tiering and automated data movement to increase storage efficiency while reducing the risk of data loss and the risk of loss of access to data.

Oracle Optimized Solution for Secure Tiered Storage Solution optimizes storage efficiency by ensuring that data is always kept on the storage tier that best matches the current access and retrieval requirements of that data. This automatic migration of data within and across the storage system tiers more accurately aligns the current business value of the data to the cost of its storage.

TABLE 6. WEB RESOURCES FOR FURTHER INFORMATION

Web Resource Description	Web Resource URL
Oracle HSM	http://www.oracle.com/us/products/servers-storage/storage/storage-software/storage-archive-manager/overview/index.html
SPARC servers	http://www.oracle.com/technetwork/server-storage/Oracle-sparc-enterprise/overview/index.html
Oracle Solaris operating system	http://www.oracle.com/technetwork/server-storage/solaris/overview/index.html
Oracle Solaris Cluster	http://www.oracle.com/technetwork/server-storage/solaris-cluster/overview/index.html
StorageTek tape libraries and drives	http://www.oracle.com/technetwork/server-storage/Oracle-tape-storage/overview/index.html
Oracle Optimized Solution for Secure Tiered Storage Infrastructure	http://www.oracle.com/technetwork/server-storage/hardware-solutions/oo-soln-tiered-storage-1912233.html
Hierarchical Storage Manager and StorageTek QFS Software Installation and Configuration Guide Release 6.1	http://docs.oracle.com/cd/E71197_01/SAMIC/title.htm

References

For more information, visit the web resources listed in Table 6.

Appendix I: Details for Reference Configurations

Configuration Guideline for Oracle Optimized Solution for Secure Tiered Storage Infrastructure Components					
Device	Component	Quantity			
		SMALL	MEDIUM	LARGE	CAPACITY
Servers					
SPARC T7-2	Server for Oracle HSM	2	2	2	2
	StorageTek QFS tape scale-out	0	0	0	2
	Server for StorageTek QFS Client	0	0	0	0
	Memory	512	512	512	512
	16 Gb FC HBA per server for disk	1	2	3	3
	16 GB FC HBA per server for tape	1	2	3	3
	10 GbE cards per server	0	0	1	1
	Oracle HSM license for each server running Oracle HSM metadata server ³	2	2	4	4
	StorageTek QFS Client license	0	0	0	2
	Oracle Solaris Cluster license per server	2	2	4	0
Disk Storage					
Oracle FS1-2	7 devices 1.6 TB SSD DE	1	1	1	0
	13 devices 1.6 TB SSD DE	0	0	0	1
	1.2 TB performance HDD DE	2	4	6	0
	8 TB capacity DE	0	0	0	4
	Front-end 16 Gb FC HBA	6	6	6	6
	Est. usable active data capacity	47 TB	94 TB	189 TB	629 TB
Oracle ZFS3-2 Storage	Performance DE with two SSD 20 HDD	2	4	8	
	Front-end 16 Gb FC HBA	4	4	4	
	Est. usable active data capacity	40 TB	74 TB	160 TB	
Tape Systems					
	StorageTek SL150	1	0	0	0
	StorageTek SL3000	0	1	1	0
	StorageTek SL8500	0	0	0	1
	Number of licensed slots	60	400	700	2,950

³ The failover server does not require the Oracle HSM license because it is not running Oracle Solaris Cluster.

Configuration Guideline for Oracle Optimized Solution for Secure Tiered Storage Infrastructure Components					
Device	Component	Quantity			
		SMALL	MEDIUM	LARGE	CAPACITY
	StorageTek T10000D tape drives	0	8	12	16
	StorageTek LTO 7 tape drives	4	0	0	0
	Est. configured archive data capacity	360 TB	3.4 PB	5.95 PB	24.6 PB
	Library grows to max. capacity	750 TB	50 PB	50 PB	850 PB
Brocade FC Switch					
	Total FC connections required across two switches with Oracle FS1-2	20	28	36	40
	Brocade 6510 16 GB FC Switch	2	2	2	2
	Licensed ports each switch	24	24	24	24
	Total ports licensed	48	48	48	48

Appendix II: Software Revisions and Testing Tools

Software Revisions

Table 7 provides an overview of the Oracle HSM components and the specific version numbers used in Oracle Optimized Solution for Secure Tiered Storage Infrastructure.


TABLE 7. ORACLE HSM SOFTWARE COMPONENTS

	Software	Release
Oracle HSM servers	Oracle Solaris	Oracle Solaris 11.3
	Oracle HSM	6.1
	Oracle Solaris Cluster	4.2

Testing Tools

The test applications that were used during this testing are the file system test (also called FS test or `fstest` command), and the `samfsck`, `samfsdump`, and `saamfsrestore` commands.

The FS Test called Multiprocess File System Performance Test Tool (`mpfstest` command) writes to an Oracle HSM file system and aims to test a file system's basic functionality and I/O performance. Unlike `fstest` command, which is a single process and a single-file test tool, `mpfstest` command is a multiprocess file system performance



test tool. It is able to generate a multiprocess workload to measure and report a file system's read and write performance.

The following areas are tested and measured by `mpfstest` command:

- » Multiprocess write performance for files of fixed size
- » Multiprocess write performance for files of random sizes within a given range
- » Multiprocess read performance for files of any size

The following commands dump or restore Oracle HSM file control structure data:

- » The `samfsdump` command creates a dump file containing metadata control structure information for each specified file.
- » The `samfsrestore` command uses the contents of the dump file to restore control structures for all the files in the `samfsdump` command metadata dump file or for each specified file.

The `samfsck` command checks and optionally repairs a StorageTek QFS file system or an Oracle HSM file system.

For more information, refer to the Oracle HSM and StorageTek QFS Software Customer Documentation Library at this website: <http://www.oracle.com/technetwork/documentation/tape-storage-curr-187744.html#samqfs>

**Oracle Corporation, World Headquarters**

500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries

Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US

blogs.oracle.com/oracle



facebook.com/oracle



twitter.com/oracle



oracle.com

Integrated Cloud Applications & Platform Services

Copyright © 2016, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group. 0615

Extreme Scalability and Flexibility for Access to 100 Percent of Your Data
June 2016
Author: Donna Harland



Oracle is committed to developing practices and products that help protect the environment