Technical White Paper
Technical Concepts of ETERNUS CS8000

Unified Data Protection with FUJITSU Storage ETERNUS CS8000

Overview

Table of Content

1 Today's Requirements for Data Protection

Investments of IT organizations in data protection software, disk based solutions and tape environments are very costly and are not subject to rapid replacement cycles. Nevertheless storage optimization, consolidation or modernization and easy deployment of innovative solutions into existing data protection infrastructures are subject of efficiency considerations as the storage industry offers more and more attractive technologies. To participate at these fast innovation cycles which promise a reduction of operational expenditures, IT organizations prefer a smooth integration of new technologies. Industry standard interfaces and protocols like VTL (Virtual Tape Library), NFS (SUN Microsystems - Network File System), CIFS (Microsoft - Common Internet File System), and OST (Symantec - OpenStorage) for backup hardware and software enable non-disruptive adoption and coexistence to current data protection strategies with limited capital and without the need to change any proven procedures and policies.

In view of service level aspects and enterprise wide disaster recovery frameworks, strategies for data protection and recovery infrastructures are based on extensive protection measures, improvements of backup window time, fast restore and rapid recovery. The availability of inexpensive disk drives has given some momentum to move away from direct attached tape-based backup and recovery processes. As a result, enterprises are embracing more disk-based technologies including specialized disk based backup appliances for short term protection and fast recovery of data and applications. These appliances include features such as data compression and data deduplication, encryption, local and remote replication. However, tape will still be used and deployed alongside disk appliances, as an archive medium or for keeping large amounts of important enterprise data over long periods. With regard to security and legal requirements for secure data retaining and in case of a unforeseen disaster, specialized disk appliances features a cohesive approach with various replication abilities and seamless policy driven data movement from disk to physical tape such offering the highest level of data availability.

Technology contemplation concerning data protection

- IT organizations keep their backup storage environment for long periods
- New technologies require long periods of testing
- Adoption for archiving solutions is growing
- Long-term data retention is a significant challenge
- Tape Technology:
  - In use for cost efficient long term storage
  - Suitably in completely automated environments
  - Viable, high capacity, long-term offline storage medium
  - Properly for long term archiving deployments
- Nearline - SAS Disk Devices:
  - Inexpensive disk storage with high capacity replaces FC disks
  - Storage of backup data directly on disk improve data transfer rates
  - Deduplication for backup reduces costs and is easy and quick to implement
  - Backup to disk and deduplication replaces manual departmental tape

Requirements for the right solution

IT organizations consider following points for selection of a data protection solution:

- Short and long term data protection with support of all service levels for each class of data
- Backup and archiving with different view on data as not all enterprise data is equal and has to be treated with different service levels (e.g. Recovery Time Objective (RTO), Recovery Point Objective (RPO) and Compliance).
- Highest data availability to protect important enterprise assets
- Strategies for seamless disaster recovery (DR) with remote data center locations
- Guaranteed and easy recovery to support fast and uninterrupted business continuity
- Seamless integration into enterprise environments to reduce capital investments
- High performance with support of each business critical environment
- Easy scalability of capacity and performance for investment protection and at the same time high flexibility to grow with future needs
- Support of legacy investments as long as these environments are necessary
- Tape storage consolidation with newest high capacity technology and easy tape storage integration for Mainframe, UNIX and Windows environments
- Reduction of tape devices and media costs

Meeting Today's Backup and Archive Challenges

A closer view can be found in the IDC Whitepaper #IDCWP36U from October 2012
http://docs.ts.fujitsu.com/dl.aspx?id=1fb2c41d-fcf0-4a67-b3cd-16a61d343263

2 FUJITSU Storage ETERNUS CS Data Protection Appliances

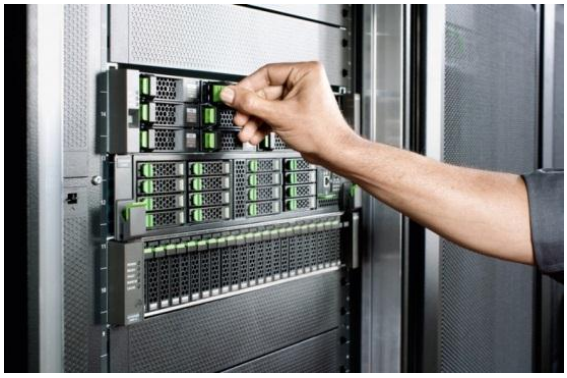
With the ETERNUS CS800 and ETERNUS CS8000 data protection appliance series Fujitsu offers data protection for each enterprise infrastructure. By aligning disk, deduplication and tape technology into an easy to manage Information Lifecycle Management (ILM) infrastructure, enterprise data protection can be flexibly aligned to different SLAs (Service Level Agreements). The entire ETERNUS CS range is well suited for every IT infrastructure and all major backup and archive software solutions. Advanced technology enables the consolidation of storage devices and simplifies each company's data protection efforts. In addition to saved energy and lower total cost of storage, IT organizations benefit from the efficient use of the most cost-effective technologies.

ETERNUS CS8000 has been the first choice to drive efficient data protection for more than 10 years. The Fujitsu appliance is used around the world by the largest data centers in data-intensive segments such as financial services, telecommunications, public administration agencies, industry, and IT service provisioning. The ETERNUS CS8000 platform is designed on the basis of real-world best practices. From the very beginning, the appliance was developed in close cooperation with users.

With the considerations of newest technologies and continuous changing customer requirements Fujitsu has further evolved its long-term vision for data protection. This not only embraces backup and recovery but also archiving to create a singular platform for all data protection needs. Fujitsu also follows a service-oriented view of the data-protection operation which includes the integrated management of disk and tape, automated migration of data across media types and systems based on policies and features and functions for high data availability.

This white paper highlights the technical concepts of the ETERNUS CS8000 Data Protection Appliance and is intended for everyone being interested in a comprehensive understanding of this solution.

2.1 ETERNUS CS800



ETERNUS CS800 is a data protection appliance optimized for environments where IT-organizations want replace backup to traditional tape by backup to disk. Utilizing leading deduplication and compression technology, the disk capacity requirements can be reduced by up to 95%. Integrated replication functionality copies data automatically on remote sites to protect them against disasters. The combination of data deduplication with replication reduces the amount of data which has to be transferred over longer distances by a factor up to 20 allowing consolidating remote backups into central systems with reduced network costs. For long term retention requirements ETERNUS CS800 provides also an integrated export to tape option.

As an integrated ready-to-use appliance ETERNUS CS800 is easy to use, efficient in operations and caused by high efficient deduplication more economically compared with pure software solutions in combination with standard disk systems. The appliance operates together with all commercially available open systems backup software which enables easy integration into existing environments.

Further details of ETERNUS CS800 are described in dedicated white papers.
More information is available at: http://www.fujitsu.com/fts/eternus-cs800

Main Features	Benefits
■ High capacity and performance	■ Backup with disk performance for fastest data protection ■ Highest scalability range from entry up to 240 TB ■ Data deduplication technology reduces disk capacity requirements such enabling large cost savings
■ Improved and automated disaster recovery protection	■ Encrypted replication to remote sites with reduced network bandwidth ■ Integrated Export to Tape Option for long-term-retention ■ Reduced media handling and easy administration
■ Integration into dynamic infrastructures	■ Turnkey solution for easy and cost-efficient deployment ■ Flexible options with VTL and NAS interfaces for seamless integration ■ OST optimized operation with Symantec OST and CommVault Simpana

2.2 ETERNUS CS8000

ETERNUS CS8000 is a unified data protection platform which simplifies backup and archiving for open systems and mainframes. Industry standard VTL and NFS/CIFS interfaces allow one unified data protection approach and optimize backup and archive infrastructures. The modular cluster architecture allows seamless scalability of capacity and performance during the server and storage lifecycles. Automated policy-based operations for managing data on disk, data reduction with deduplication or long-term storing on high capacity tape simplifies storage administration tasks. With a fine-graded centralized data lifecycle management, flexible service levels regarding storage capacity, transfer speed and cost can be provided. Different data protection requirements at the same time in terms of disaster resiliency, security, and retention periods can be served with a single solution. Integrated data replication features over short and long distances facilitate comprehensive disaster recovery architectures.

Main Features	Benefits
■ One uniform data protection appliance as single target system for backup and archiving	■ VTL and NFS/CIFS interfaces enables flexible consolidation in SAN and NAS environments ■ Consolidation of target storage devices reduces the overall data center costs and administration efforts ■ Enables easy step-by-step consolidation projects
■ Centralized integrated management of tape, disk and deduplication	■ Flexible service levels regarding capacity, performance and media costs
■ Virtualization on top of a modular uniform cluster architecture	■ Seamless and highest scalability of storage capacity and performance during the system lifecycle ■ Decoupled technology lifecycles of server and storage enables faster innovations and investment protection
■ Data mirroring, multiple safe, replication and cascading	■ High availability of data for backup and archive operations ■ Enhanced disaster recovery scenarios

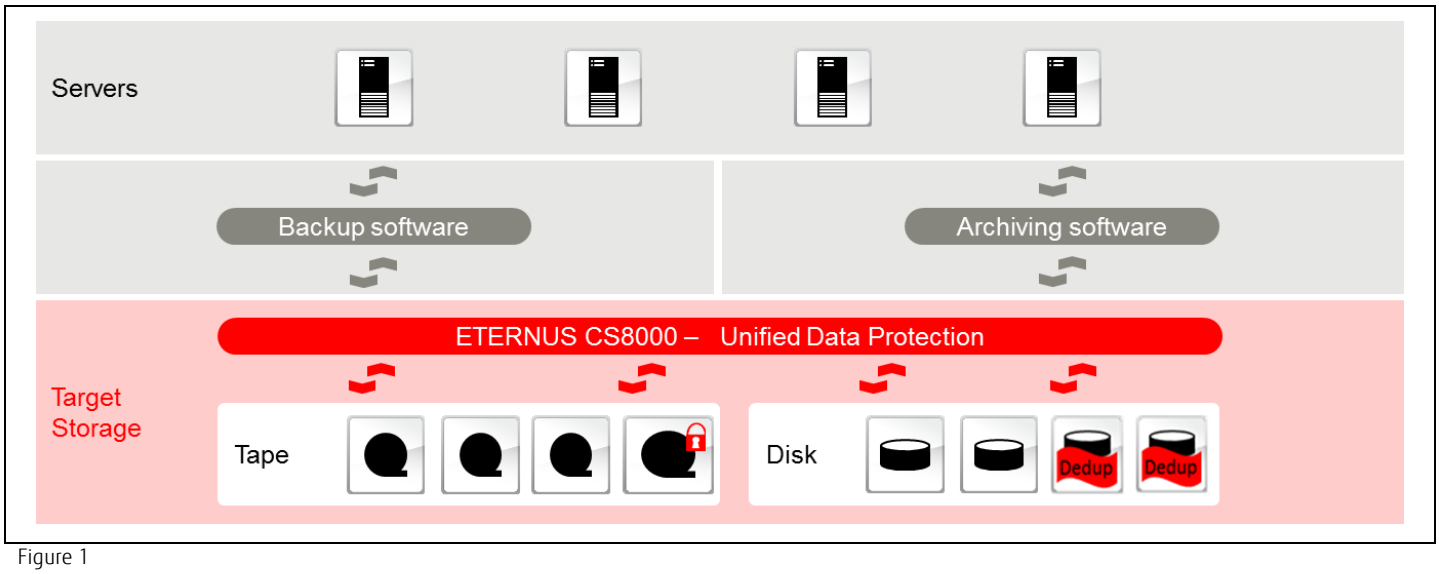


Figure 1

3 ETERNUS CS8000 Overview

The ETERNUS CS8000 Data Protection Appliance provides VTL interfaces (Virtual Tape Library) for storing block based tape volumes and a NAS interface (Network Attached Storage) for storing file based data structures. The **CS VTL subsystem** is designed to store and retrieve within a SAN (Storage Area Network) sequential block-based tape volume data. Its behavior for backup software appears exactly like addressing physical tape libraries. The **CS ViNS subsystem** (Virtual Network Storage) provides within a TCP/IP network a NAS based file data archive with a virtual extremely high capacity.

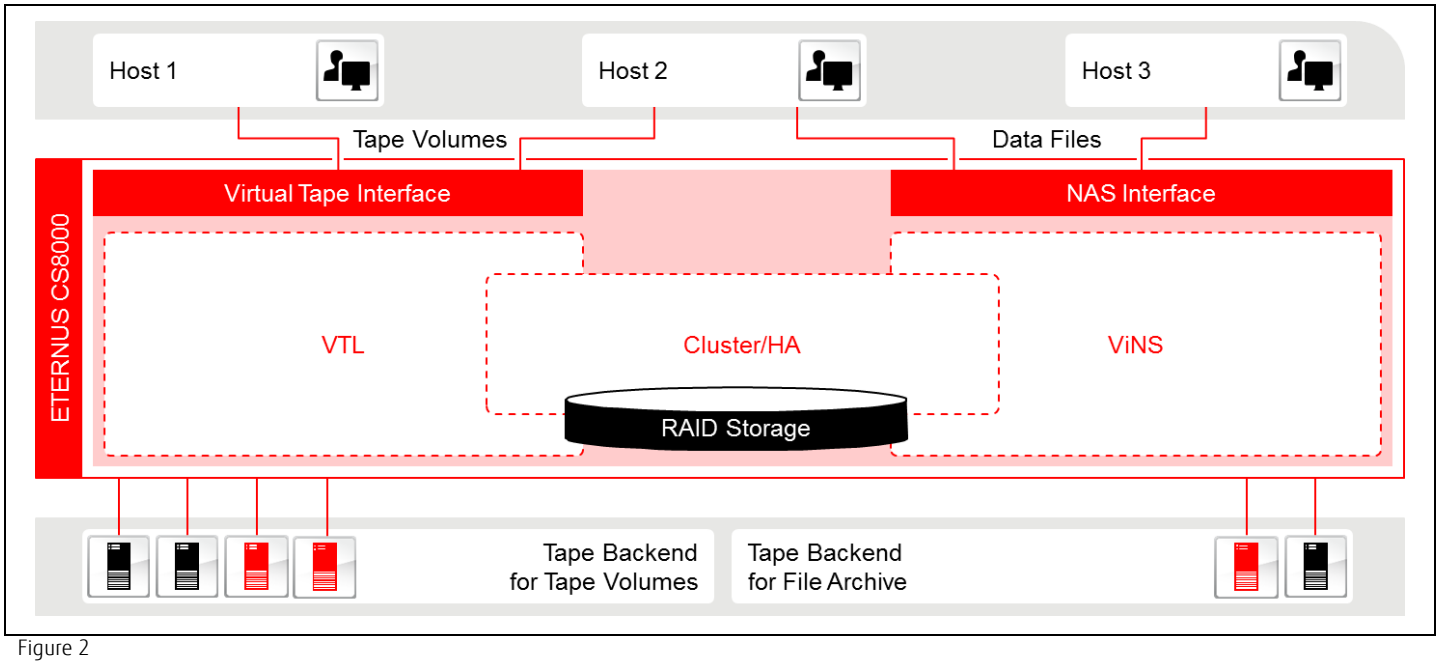


Figure 2

Regarding different data structures of tape and file the two autonomous storage virtualization subsystems CS VTL and CS ViNS are operating in parallel within the appliance on base of one single unified high available cluster architecture. Beside a joint cluster configuration as a shared CS VTL / CS ViNS appliance, the subsystems can also be concatenated to utilize CS VTL as virtual tape storage for CS ViNS. For homogeneous requirements the appliance cluster can also be operated with only one subsystem, CS VTL or CS ViNS.

Unified Data Protection with CS VTL and CS ViNS

The tight integration of hardware components within a modular high scalable cluster architecture offers a disaster resilient unified data protection platform for backup and archive purposes. Based on a highly flexible frontend/backend design, the hosts are directly connected with the CS VTL or CS ViNS frontend interfaces. Completely decoupled from the hosts, the appliance controls backup and archive data within their internal RAID (Redundant Arrays of Independent Disks) storage and the migration of data to tape or disk storage devices connected at its backend. Optional deduplication of tape data on disk reduces disk capacity requirements. Provisioning of resources like frontend connectivity, appliance performance or aggregate storage capacity is freely configurable.

The virtualization subsystems eliminate the dependencies between the data accessed at the host interfaces and the target storage devices where the files are physically stored. One of the major benefits of abstracting the hosts from the actual storage is the ability of an independent and centralized storage device based hierarchical storage management (HSM) while maintaining concurrent I/O access. Storage device-based HSM provides the automated determination of storage location for backup and archive data across different types of storage devices and media. The non-disruptive data movement according performance, capacity and costs requirements embraces state of the art tape, disk, compression, and deduplication technologies.

The unified appliance form factor offers several advantages:

- Host independent unified backup and archive data management on disk and tape
- Industry standard hardware components within one single scale-out architecture
- Disaster resiliency and failover between sites based on a unified cluster technology
- Common subsystem control with central appliance control units
- Common user interface and service functions
- Advanced CS ViNS tape data management by internal concatenation option with CS VTL

3.1 CS VTL Subsystem Overview

Tape storage provides the highest storage capacity with lowest total cost of ownership for securely storing long-term backups, for record keeping and disaster recovery. For large IT-organizations tape processing and life-cycle management of tape data is still a key part of their data protection approach. To overcome specific obstacles with direct attached physical tape, the virtualization of tape storage allows integration of VTLs with existing backup and recovery processes. The benefits of such virtualization include improved data center policies and cost savings through storage consolidation.

In a simplified view, tape virtualization enables disk storage usage with a behavior like tape libraries and tape drives. In some cases tape data is completely stored on a disk library (DL) embedded disk array only (called as backup to disk or B2D) in other cases tape data stored on embedded disk is exported for disaster recovery or archive purposes to other media, such as physical tapes (called backup-to-disk-to-tape, or B2D2T). In case of B2D2T the shift to a VTL eliminates streaming problems that often impair efficiency in tape drives as the internal disk technology does not rely on streaming. The VTL embedded tape migration can write more effectively data to tape regardless of hosts data transfer speeds.

Tape virtualization of the CS VTL subsystem facilitates backup and restore operations for B2D operations in conjunction with a high efficient data deduplication technology or for B2D2T operations enriched with a fine graded automated tape processing. To cope with data storage growth and complex heterogeneous tape environments CS VTL reduces backup and restore times with less tape management overhead for hosts. Within the CS VTL subsystem the implementation of a fully integrated tiered storage hierarchy of disk and tape takes advantage of the benefits of both technologies to enhance performance and provide the capacity needed for today's tape processing requirements.

CS VTL offers simultaneously a multitude of different possibilities for automatic processing and secure storing of tape data (logical volumes). Corresponding to policies, options and different workloads the virtualization with CS VTL allows storage of tape data:

- On physical tape cartridges within backend attached physical tape libraries. Different tape technologies and different tape library types from different vendors are supported simultaneously.
- Instead storing on physical tape, storing permanently on RAID systems within backend attached ETERNUS CS systems (optionally with data deduplication in the backend system).
- Storing permanently on internal RAID systems only in native format without backend attached storage system.
- Storing permanently on internal RAID systems only with deduplication of logical volumes.

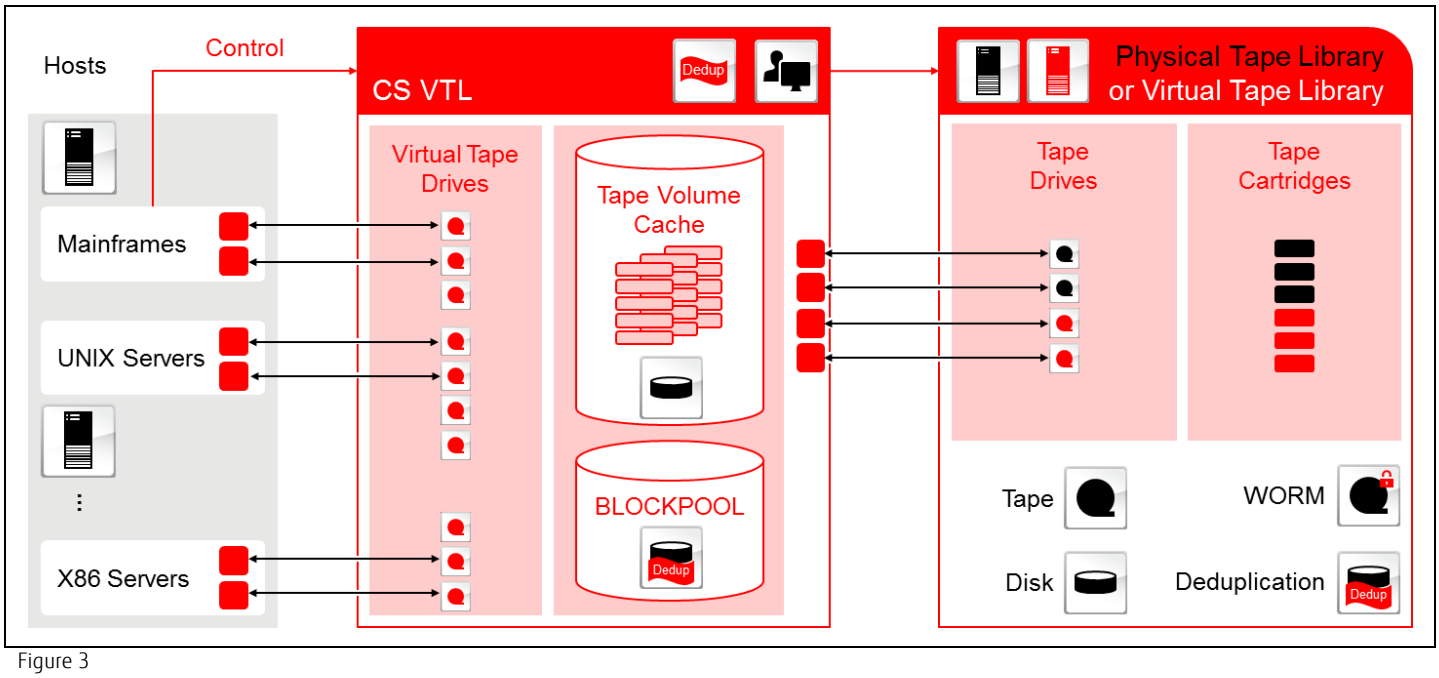


Figure 3

At the frontend CS VTL improves the efficiency of host tape operations by storing tape data written over the frontend interfaces into internal RAID storage. The interfaces support numerous backup applications for a multitude of mainframe, UNIX and x86 system connections. CS VTL emulates multiple tape libraries and multiple tape drive industry standards with a virtualization of all essential physical tape library objects which consist mainly of three parts:

- Virtual library control interface
- Virtual tape drive interfaces for the I/O data path of tape volumes
- Logical tape volumes stored either in the internal Tape Volume Cache (TVC) for optimized backend processing or alternatively in a deduplicated format as unique data blocks in the so called BLOCKPOOL.

At the backend CS VTL optimizes tape volume data transfer to the attached tape storage targets. Host interfaces and backend drive interfaces are completely separated and kept independent. CS VTL requires no adaptations of host software and backend operations are completely autonomous.

In summary CS VTL is the only virtual tape library on the market with an extensive flexibility regarding heterogeneous support and consolidation potential. Virtualization and the technical decoupling of hosts from physical storage devices offer a couple of advantages:

Host applications

- Simultaneous use of many virtual tapes drives in parallel.
- Storage on RAID systems reduces time window for backup and enables fast recovery of data.
- Storage on RAID systems offers essential RTO (Recovery Time Objective) and RPO (Recovery Point Objectives) improvements.
- Deduplication of logical volumes which are stored on RAID systems only, offers significant reduction of disk capacity requirements.
- Automated migration of logical volumes to physical tape simplify tape processing for hosts and release them from intensive physical tape maintenance operations.
- Tape technology migrations at the backend without changes to applications and without affecting application availability.

Disaster Protection

- Various synchronous and asynchronous data replication features for backup data on disk and tape provide highest data availability.
- Redundant components, dual-site configurations and comprehensive failover mechanisms provide disaster resiliency of the data protection appliance.

Backend storage

- Physical tape at the backend and heterogeneous technologies in parallel operations are possible.
- Efficient tape maintenance and streaming mode prevent failure with physical tape storage.
- Support of different vendors and heterogeneous tape technologies at the backend.
- Take advantage of fast technology lifecycles with exploitation of performance and capacity from newest tape technologies.
- Easy migration from an older tape library to the latest generation fully transparent for the backup application.
- ETERNUS CS disk only appliances at the backend; in parallel with physical tape is possible.

Consolidation

- Heterogeneous frontend attachments for opens systems and mainframes
- Virtualization allows consolidation of tape processing in one backend technology for the whole data center.
- Sharing of a centralized automated high reliable tape processing appliance.

Cost

- IT department can separate investment decisions regarding hosts and applications from storage devices and the associated management tasks.
- Investment protection for existing tape library systems.
- Different disk technologies in parallel for the internal RAID system, including deduplication reduces overall costs for disk storage capacity.
- The multitude of connectivity options enables tape storage consolidation in the data center, reduces the effort required for company-wide administration and cuts the overall storage costs.
- Compared with a high number of expensive physical drives to reach a respective data throughput a bunch of virtual drives can be operated very cost efficient.
- Highest tape transfer rates and using the entire tape capacity reduces the number of cartridges, the size of the tape libraries, floor space, power consumption, and cooling.

3.2 CS ViNS Subsystem Overview

The CS ViNS (Virtual Network Storage) subsystem is a unified disk and tape storage target to archive file data, including files with compliance demands. It is also an ideal secondary storage target to relieve a primary host storage system by moving less used or outdated data from the primary storage to ETERNUS CS8000. Based on predefined storage classes and retention periods files are stored primary on disk and optionally secondary on tape. As an alternative, files can be replicated to disk of a secondary CS ViNS system.

ViNS Frontend

Frontend attachments to CS ViNS behave like connecting a logical NAS system storage over standardized NFS (Network File System) and CIFS (Common Internet File System) interfaces. The NAS interfaces provide file-based data storage of virtually unlimited size to the network. Although it is technically possible to access the CS ViNS subsystem as a NAS unit, it is not designed to be a general purpose file server for simply sharing computer files over a network. CS ViNS is designed to store archive data for short term access and long-term retention, such using integrated policy-based automation for moving the archive data from disk to physical tape.

Client application data is directly stored on a cluster file system using the internal RAID storage. Within the CS ViNS subsystem a clustered file system acts as ViNS Online Store (VOS) allowing very high throughput and disaster resiliency. CS ViNS manages the file system transparently to the applications and keeps track where data reside within the system, in order to maintain accessibility to any given file in a file system, regardless of the media on which it is resting. Requirements for compliant archiving are fulfilled on the disk layer with WORM functionality as well as with support of WORM tape. Various high availability and disaster recovery scenarios ensure data availability.

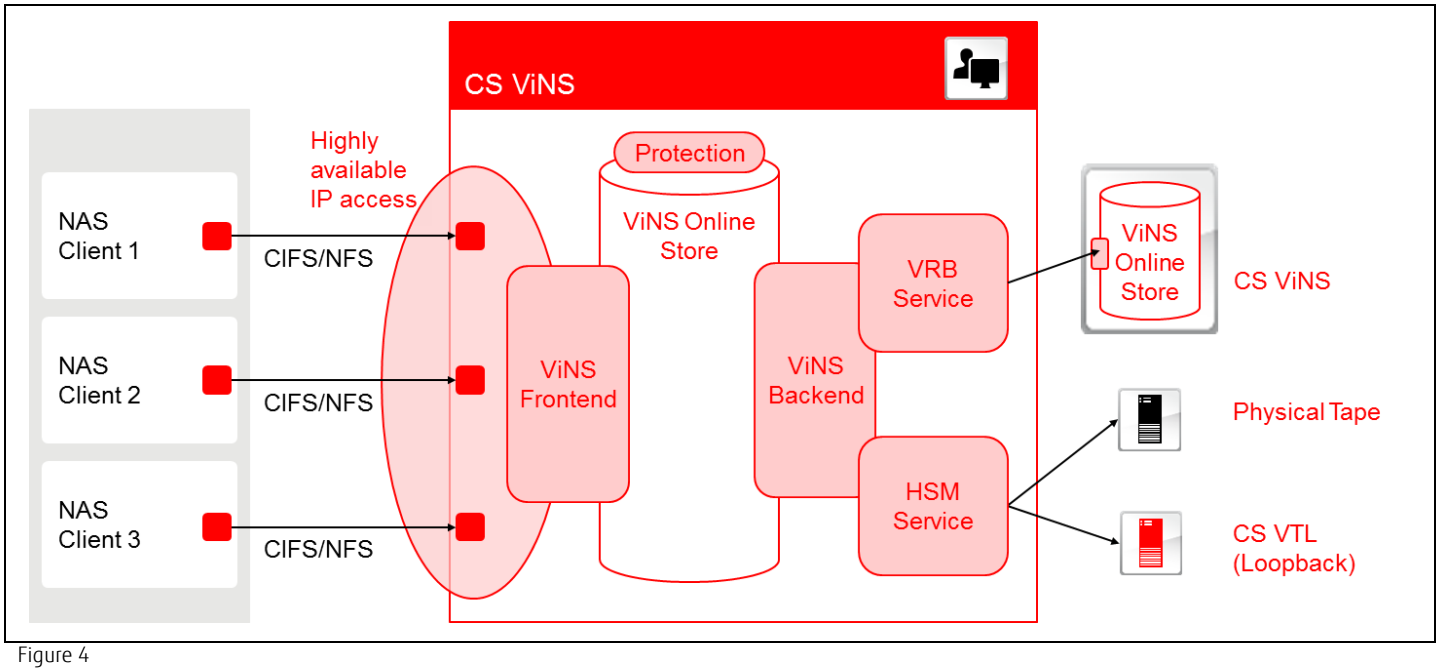


Figure 4

ViNS backend

The ViNS replication and backup (VRB service) offers a data protection method by CS ViNS which performs the replication of file systems kept in the ViNS Online Store (VOS) to another CS ViNS system.

Alternatively the ViNS backend offers hierarchical storage management (HSM service) for migration of file systems to tape for long term archiving purposes. By using the HSM service which migrate data from disk to tape, the usable file system space for the NAS clients becomes virtually unlimited. The ViNS Online Store (VOS) is then regarded as a cache with respect to the totally available file storage capacity. Even if files shall be kept in the VOS and never be migrated, the HSM service may still take care that additional copies are created on tape for data protection reasons.

Please note: VRB service and HSM service cannot be configured on the same file system, but for different file systems within the CS ViNS appliance subsystem.

VRB service

The purpose of file system replication is to generate a defined file system copy which can be used in disaster cases. The VRB service prevents data loss of non HSM based NAS file systems on the primary CS ViNS system. If the ETERNUS CS8000 RAID hardware failed partially (one or more file systems) or totally (disasters) the operation can be continued with the secondary CS ViNS system. After a manual failover, operation the

NAS clients are allowed to continue operation with the target system. Another recovery way is to preserve the data for being restored onto the primary system after that system has been repaired.

The base mechanism replicates NAS file systems (NASFS) of a CS ViNS to a replication target on another CS ViNS via LAN or WAN. The VRB services on both systems takes care that file systems from the VOS are periodically replicated in order to keep up-to-date copies to be used in disaster cases.

Using the CS ViNS subsystem with the VRB service offers various advantages:
- Recover from the loss of a complete file system by continuing operation on a recently replica of the file system on secondary CS ViNS system.
- To save bandwidth, only the deltas between master and replica are transferred. A file which has not changed since the last replication cycle must not be transferred in the current cycle.
- Replication to a central office: NAS file system (NASFS) of multiple ETERNUS CS8000 can be replicated to a shared ETERNUS CS8000 which provides replication targets for all replication sources via LAN or WAN; the shared ETERNUS CS8000 has to provide enough capacity to accommodate each replication target of all connected replication sources.

Please note:
File system replication between two different independent systems can be operated in parallel to the appliance internal cache mirror feature.

HSM service

Optionally CS ViNS controls a tight interlocked HSM (Hierarchical Storage Management) service which migrate files stored in the VOS to tape and manage the lifecycle of the files within the different storage media.

For all data which are completely migrated on tape, the CS ViNS subsystem keeps only a small amount of metadata online. When an application requires access to migrated data, the HSM service performs a recall of data from tape to the VOS. For the applications this recall is transparent, it will only recognize an increased latency for file access. For files that shall always be kept in the VOS without migration to tape, the HSM service can be configured to create additional copies on tape for data protection reasons.

As result data can be stored on the most appropriate media type for the given retention and restoration requirements of that data. This allows less time sensitive or infrequently accessed data to be stored on less expensive media, and eliminates the need for an administrator to manually migrate data between storage systems or storage tiers. Since tape libraries have very low power consumption, the operational expense of storing non active data on tape is greatly reduced.

Virtual unlimited capacity

With the capacity expansion by HSM services, the usable virtual NAS capacity results in the sum of the VOS capacity on the RAID system enhanced with a nearly unlimited tape storage capacity. This capacity expansion is totally transparent to applications because all files keep visible at the frontend interface even if they are stored on tape.

Using the CS ViNS subsystem with its scale-out cluster file system and tape storage behind one common interface offers various advantages:
- Easy integration with applications and services which are uncoupled from the underlying storage infrastructure.
- Extreme disk and tape scalability to accommodate capacity growth and store massive amounts of data.
- File system for up to 2 billion files with no file size limitation.
- Operational efficiency and administrative savings for managing massive numbers of files.
- Automated lifecycle management and migration to tape for archive data.
- Transparent usage of different service levels that make the supporting technology invisible.
- Central deployment and management of automated tiered storage consolidates separate file servers and eliminates complexity.
- Investment protection due to eliminating interference between archive software and disk or tape technology.
- Lower TCO with the right balance of access speed and storage cost.

Take advantage of the unified architecture

Concatenating the CS ViNS tape backend to the CS VTL frontend can simplify specialized disaster protection configurations. Just like the behavior with a physical tape library, CS ViNS/HSM can store tape volumes also to the CS VTL virtualized tape interface. Within an internal loopback all tape volumes of CS ViNS/HSM are routed to the CS VTL subsystem via an appliance internal connection. This means that the frontend of CS VTL is connected at the backend of CS ViNS/HSM. Such every tape volume generated by CS ViNS benefits from the extensively advanced CS VTL tape operation options which are described in the chapter 4.2. The mapping of objects is described in chapter 5.7.

Please note:
Due to different data structures in standard operation, the backend services of CS VTL and CS ViNS/HSM are operating independently from each other. This means that tape volumes from CS ViNS are basically not known by CS VTL and vice versa.

3.3 Hardware Architecture

3.3.1 System Overview

From a hardware point of view the appliance operates as a modular system built up with appliance processor nodes and internal disk storage, which consists of a number of RAID systems and acts as cache storage for data. The frontend processor nodes are operating the appliance virtualization and serve as interface to the host servers. The backend processor nodes are operating the HSM services with interfaces to the attached target storage systems. Each storage target with a tape interface can be attached, regardless the embedded physical storage such as tape, WORM tape, disk or disk with deduplication. Disk based backend targets can be secondary CS VTL systems. Currently the possible number of processor nodes with frontend/backend operation is 1 – 20. The possible number of RAID systems is 1 - 16. The scale-out cluster architecture allows seamless future extensions also beyond the current limits. Further hardware details and current upper limits are described in the Technical Overview (chapter 7).

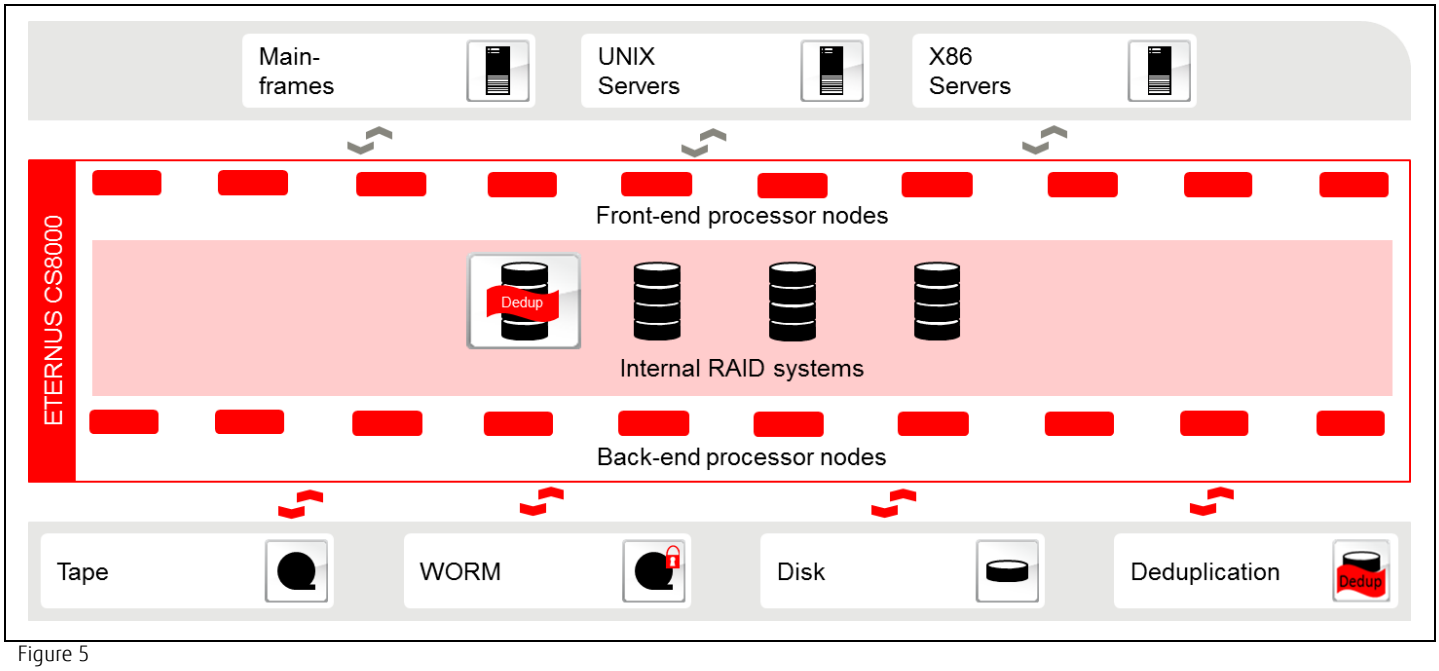


Figure 5

The software running on the appliance processor nodes provide a high available cluster framework. The internal RAID systems are formatted with a set of cluster file systems which allow all appliance processor nodes transparent data access in parallel. This cluster architecture permits dual-site configurations with automatic failover mechanism, easy expansion of disk capacity and seamless removal or addition of processor nodes for maintenance purposes or seamless enhancements.

Additional infrastructure components like internal SAN and LAN networks, a system console with keyboard, graphical user interface and KVM switch (Keyboard, Video, and Mouse) and administrative servers complete the appliance.

3.3.2 Processor Nodes

Integrated Service Processor (ISP):
The generic term of a single appliance processor node is Integrated Service Processor (ISP). Depending on the appliance size, number of installed processors, software and license keys, an ISP will run specific services and provide specific interfaces to the outside world. Related to the performing of specific services an ISP is called

- ICP (Integrated Channel Processor)
- IDP (Integrated Device Processor)
- VLP, SVLP (Virtual Library Processor, Standby Virtual Library Processor)
- TBP (Tie Breaker Processor)

Integrated Channel Processor (ICP):

An Integrated Channel Processor (ICP) runs frontend services which provide over the host interfaces the data transfers between hosts and cache storage. An ICP may run VTL or ViNS subsystem frontend services.

In order to provide VTL subsystem frontend services an ICP is equipped with Fibre Channel or FICON interface controllers. On top of the channel device drivers the ICP provides a set of virtual tape drive emulations with tape drive interfaces to handle data I/O from and to the hosts. Tape data (tape volumes) received from a host is stored as Logical Volumes (LV) in the Tape Volume Cache (TVC); tape data to be transferred to the host is read from the TVC and passed back to the host.

For providing ViNS frontend services an ICP is equipped with Ethernet interface controllers. On top of the Ethernet device drivers and the TCP/IP protocol stack the ICP provides a NAS interface with the NFS and CIFS protocols. Data received from hosts through the NAS interface is stored in the ViNS Online Store (VOS). The only backend service running on an ICP is the VRB (ViNS replication and backup) service which is responsible for replicating file systems between different CS ViNS systems.

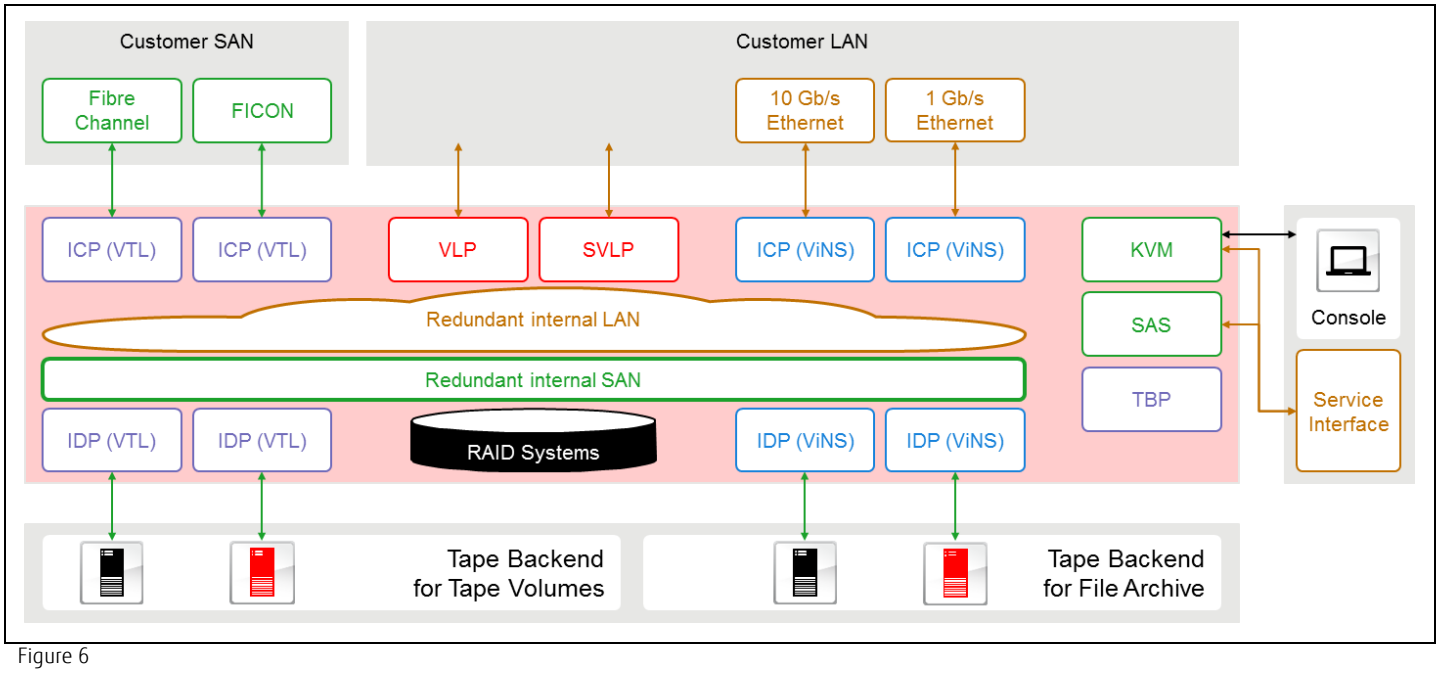


Figure 6

Integrated Device Processor (IDP):

An Integrated Device Processor (IDP) runs the appliance backend services. These are needed whenever data from the RAID systems (TVC or VOS cache storage) needs to be transferred to physical tapes or to a disk based library (with or without deduplication).

IDP backend services are not directly related to the ICP frontend services. However, when hosts require access to data that is currently not available in the cache storage, backend services are triggered by the appliance to recall the requested data from physical tape.

ETERNUS CS8000 models can also be operated without IDPs and physical tape at the backend and all data is kept in the RAID systems only (TVC or VOS cache storage).

The backend services of CS VTL and CS ViNS/HSM are completely separated from each other. This means that tape volumes from CS ViNS are basically not known by CS VTL and vice versa. In case a physical tape library control offers concurrent access from different systems, the CS VTL and CS ViNS subsystems can share the same tape library but have to use separated tape drives and tape cartridges.

Virtual Library Processors (VLP, SVLP):

The Virtual Library Processor (VLP) performs the appliance management services. Some of the services are used for configuring, operating and monitoring the overall system. Some other services are specific for the subsystems.

For the VTL subsystem the VLP performs in addition the inventory services for logical volumes, physical volumes and the TVC cache management. The VLP also provides TCP/IP based control services for the virtual library emulation at the frontend and the physical tape libraries at the VTL backend.

Since the VLP is crucial for the operation of the entire appliance, all models with full redundancy are equipped with a second VLP which is working as a hot standby - the Standby VLP (SVLP). The SVLP does not perform any active tasks except for monitoring the operating status of the active VLP (also referred to as the Master VLP) but is kept ready to take over the Master VLP tasks in the event of failure.

Tiebreaker Processor (TBP):
A Tiebreaker Processor (TBP) is specialized for ETERNUS CS8000 dual site configurations where the physical appliance components are distributed over two geographically distant locations. Despite the physical component separation over two sites, the appliance operates still as single logical entity. The TBP provides decision hints for the internal appliance framework in situations when the communication paths between the two locations break down.

3.3.3 Infrastructure Components

Redundant internal LAN:
All hardware components of an ETERNUS CS8000 system - including processor nodes, RAID systems and SAN switches - are connected with an internal LAN in a way that all components exchange control and management information with TCP/IP connections. The IP network is restricted to internal use; it is not possible to make the network available to the outside world by means of IP routing.

Since the internal LAN is crucial for maintaining the appliance infrastructure, the LAN is built around a redundant set of Ethernet switches where all processor nodes are connected with two interfaces, each cabled to a separate switch. The software on the processor nodes automatically detects link failures and switches the active link on demand.

By using deduplication within CS VTL, a dedicated independent internal storage LAN (SLAN) is used in parallel to the standard internal LAN (chapter 4.7.3)
. There is no connectivity between the two internal networks.

Redundant internal SAN:
The entire data flow between the processor nodes and the RAID systems is transferred via an internal SAN. Depending on the specific appliance model the SAN is built redundantly around two Fibre Channel switches to provide highly available data paths and scalability regarding data throughput. A processor node has two Fibre Channel controllers dedicated to the internal SAN connection, each being cabled to a separate Fibre Channel switch. The multipath software on the nodes takes care of link detection and distributes the data transfer between the available paths. In normal operation load balancing takes place on the communications paths between the processor nodes and the RAID systems.

Because of the high bandwidth of the internal SAN there is no possibility becoming the bottleneck of the whole system. Two FC technologies are available, *Multi Mode* and *Single Mode*. In *Multi Mode* operation, the devices which are connected via 8 Gbit/s Fibre can be up to 300 m apart, and in *Single Mode* the distance can be as much as 25 km with direct connection. With special convertors distances of more than 150 km can be realized.

System Console:
The console ports of all processor nodes are connected to a **KVM switch** (Keyboard, Video, and Mouse). A graphical console attached to the KVM switch allows access to the processor nodes directly at the rack. In addition the KVM switch contains an Ethernet port that may be connected to the enterprise LAN. This port is configured for remote access to the system via a secure TCP/IP connection.

Service Access System (SAS):
The Service Access System (SAS) is running on a dedicated PRIMERGY server to monitor over the internal LAN all processor nodes and the RAID systems. Additionally the SAS provides remote service access (for example Universal Tele Service) and functions such as CALL HOME.

3.3.4 Internal RAID System

The internal disk storage of ETERNUS CS8000 is set up with of one or more RAID (Redundant Array of Independent Disks) systems which form one logical entity within the internal appliance SAN infrastructure. The RAID systems serve as cache for CS VTL and CS ViNS to store and retrieve data from the host interfaces and policy-based migration of data by the internal HSM processes to tape. The file system for the CS VTL Tape Volume Cache (TVC) is organized as tape file system (TFS) for block based access over the SAN based VTL interfaces. Within CS VTL optionally the BLOCKPOOL files system serves for storing deduplicated tape data. The file system for the CS ViNS Online Store (VOS) is organized as NFS storage with file based access over the TCP/IP based NFS/CIFS interfaces. The file systems for CS VTL and CS ViNS are both based on the high performance high scalable ETERNUS CS Appliance File System (CAFS). The technical integration of the CAFS on base of the RAID system is described within the next chapter.

3.4 Shared Cluster File System

An ETERNUS CS8000 appliance forms a single shared disk cluster (CAFS Cluster) which consists of the appliance processor nodes (CAFS Nodes) which have shared access to the ETERNUS CS Appliance File System (CAFS). CAFS is designed as a shared cluster file system and responsible for arranging RAID system storage capacity, reliability, efficiency, and tuning with regard to the physical disk storage medium.

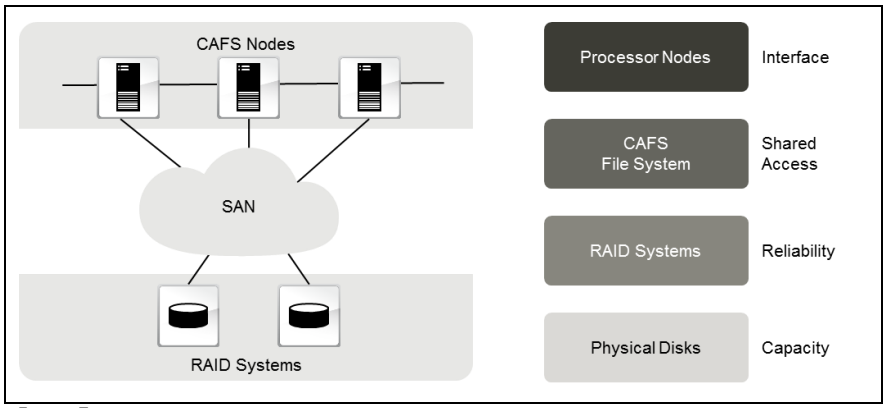


Figure 7

Within the CAFS cluster the RAID system is directly attached to all CAFS nodes. The nodes are physically connected to the RAID system using the internal SAN and to each other using the internal LAN. Cache data transferred by the nodes flows over the SAN and the cluster file system. Control information flows among the node instances in the cluster over the LAN. The direct connection of all nodes means that each shared block device is available concurrently to all of the nodes in the cluster.

Each RAID system is partitioned into Logical Units (LUNs) where LUNs are defined with the RAID configuration of an ETERNUS DX and thus provide the available entire RAID system capacity in logical units. Each LUN is presented across a Fibre Channel interface within the internal SAN and available for each CAFS node. On top of the available set of LUNs the CAFS is built on base of logical Network Shared Disks (NSD) which corresponds one-to-one to the LUNs of the RAID system. The Network Shared Disks provide a method for cluster-wide disk naming and high-speed node access to data. The disk space available for the CAFS is allocated by assigning Network Shared Disks.

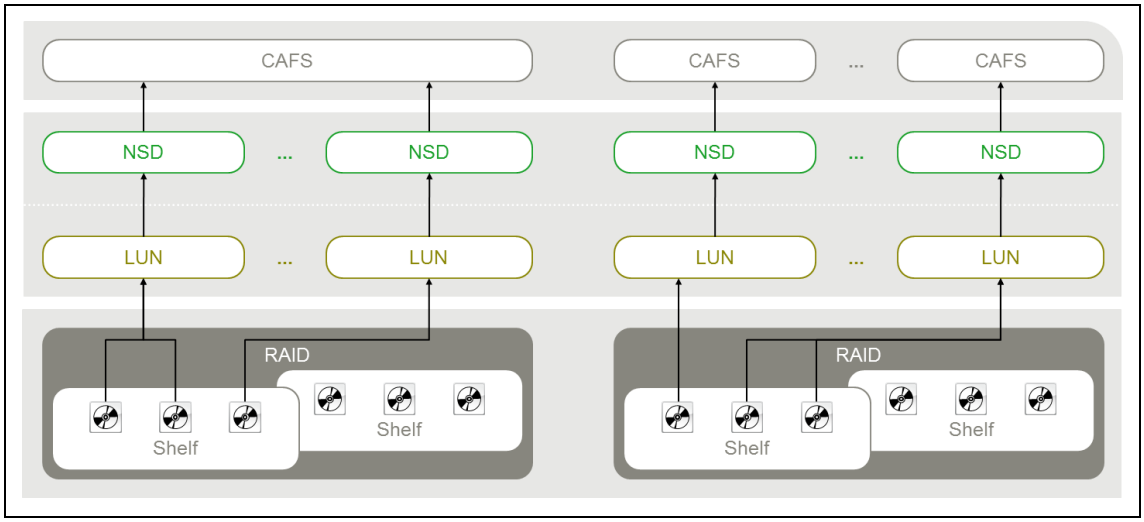


Figure 8

CAFS enables a powerful processing of very large volumes of data and presents highest data integrity and availability even when hardware components are affected by a fault or fail completely. CAFS is fault tolerant and enables uninterrupted appliance operation with continued access to data even when disk faults occur, a RAID system crashes, one or more nodes in the cluster crash, and faults occur in SAN and LAN communication, including complete failure of the SAN. This is accomplished though robust clustering features and support for data replication. CAFS continuously monitors the health of the file system components. In case of a SAN failure or inability the CAFS automatically manages the emergency disk accesses via the internal LAN and performs appropriate recovery tasks.

A single ETERNUS CS8000 system may contain multiple cluster file systems, each used for specific purposes by the two appliance subsystems. The cluster architecture within the appliance provides an enterprise level of data protection and business continuity capabilities. The high level of data availability in the event of data corruption or an unexpected disaster is ensured on several levels:

- Cluster architecture protects against individual node failures
- Cluster file system enables embedded data replication which protects against RAID subsystem failures
- Appropriate RAID level 5 or RAID level 6 protect against individual disk failures

The CAFS Cluster Components

- A **CAFS Node** is an appliance processor node within the **CAFS Cluster**. In ETERNUS CS8000 each Integrated Service Processor (ICP, IDP, VLP/SVLP, TBP) is a CAFS Node.
- The **CAFS File Systems** are managed and maintained by the CAFS File System Manager.
- The **CAFS File System Manager** implements status and configuration changes to the file system, deals with mount and unmount requests and checks the efficiency of disk utilization.
- The **CAFS Cluster Configuration Manager** manages the cluster configuration

When the CAFS Cluster is configured, the VLP/SVLP and TBP are specified for the roles of CAFS File System Manager and CAFS Configuration Manager. The CAFS Cluster Configuration Manager makes information of the cluster configuration available to all CAFS Nodes. CAFS Nodes can be dynamically removed from and later re-inserted into the cluster. Examples may include fatal error or an upgrade / maintenance of the node. When the CAFS Node is started, it contacts the VLP (the primary CAFS Cluster Configuration Manager or the secondary (SVLP) in case the primary does not respond) to obtain the current data of the cluster configuration. Without this data it cannot join the CAFS Cluster.

3.4.1 Internal File Systems

With the CS VTL subsystem several file systems on the assigned disk storage are used as Tape Volume Cache (TVC) for storing logical volumes and several BLOCKPOOL repositories for storing unique data blocks of deduplicated tape volumes.

The CS ViNS subsystem uses separately assigned disk storage with its own file systems for the ViNS Online Store (VOS). The VOS is for the file data that is directly available in the cache and also for the stubs of the files that are already migrated to tapes. Aside of the VOS the ViNS subsystem also needs some amount of disk storage for HSMS metadata.

CAFS is the technical base for all system configurations related to the Tape Volume Cache (TVC), BLOCKPOOL and to the ViNS Online Store (VOS). All data of both, CS VTL and CS NAS, are stored here in separate file systems. However, ETERNUS CS8000 does not expose the CAFS at the user interface. Instead the appliance file system management is adapted to the specific usage scenarios.

The following CAFS based file system types are used:

- **TFS = Tape File Systems** which serves for storing of tape volumes. The tape file system is especially designed to store sequential tape format files on disk in a self-describing form. Backup software on hosts uses these structures to store tape volumes over the SAN to tape drives. All tape file systems as a whole constitute the Tape Volume Cache (TVC) of the CS VTL subsystem.

- **BPFS = BLOCKPOOL File System** which serves as repository for deduplicated tape data on disk where the unique data chunks are stored together with the metadata needed to keep track of stored objects.

- **NASFS = NAS file systems** for data in the ViNS Online Store (VOS) of CS ViNS. NAS (Network Attached Storage) provides a file system on top of a block level based disk storage and appears to the operating systems as a file server. NAS storage enables sharing files over the TCP/IP network providing data access to a group of clients. Within CS ViNS a NAS file system (NASFS) corresponds to a CAFS file system.

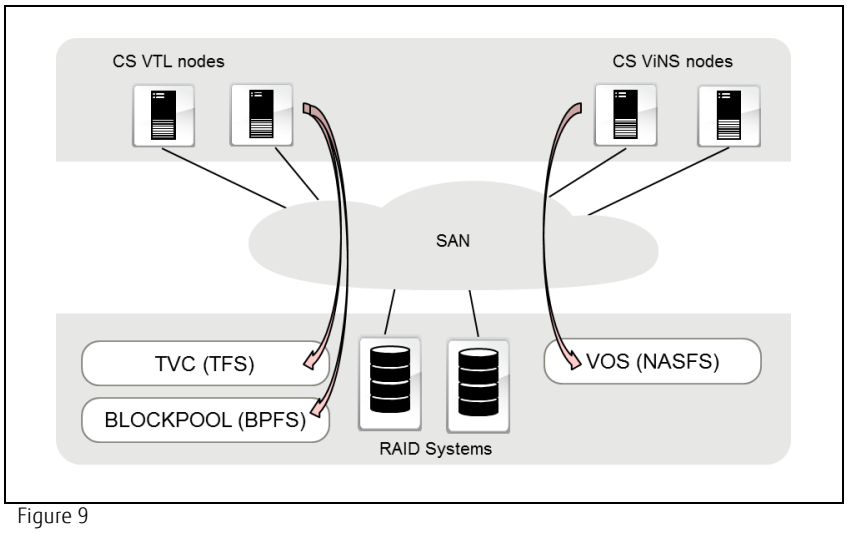


Figure 9

Beside several TFS, BPFS and NASFS file systems, some additional internal file systems are used for databases and administrative data of ETERNUS CS8000.

3.4.2 Cache Mirroring Feature

For an additional level of data availability and protection, a synchronous data replication between different RAID systems is available. This is particularly useful when the RAID systems are located at geographically distant sites. The Cache Mirroring Feature (CMF) ensures higher data protection against data loss as the result of disasters.

With the CMF dedicated assigned TFS and dedicated assigned NASFS can be configured for data mirroring. Such tape volumes stored in the Tape Volume Cache (TVC) and files stored in the ViNS Online Store (VOS) become replicated. With CMF the assigned TFS/NASFS stores its data synchronously on two RAID systems. Typically the two RAID systems with mirrored cache data are located at two different locations whereas both RAID systems belong still to one single cluster unit. Even if one location is completely destroyed, all the data which was stored on a mirrored file system is always available at the second location.

At least two RAID systems should be available for using the CMF, depending on the model, a maximum of 2 x 8 RAID systems (cf. chapter 7) within an entire cluster are supported. RAID systems which are used for CMF should be homogeneous in structure. The number of RAID systems used must be an even number and two RAID systems should be configured as an identical pair.

The following picture shows the assignment of several TFS located on two locations. TFS 101 is assigned for both locations which means that all data is replicated according the CMF.

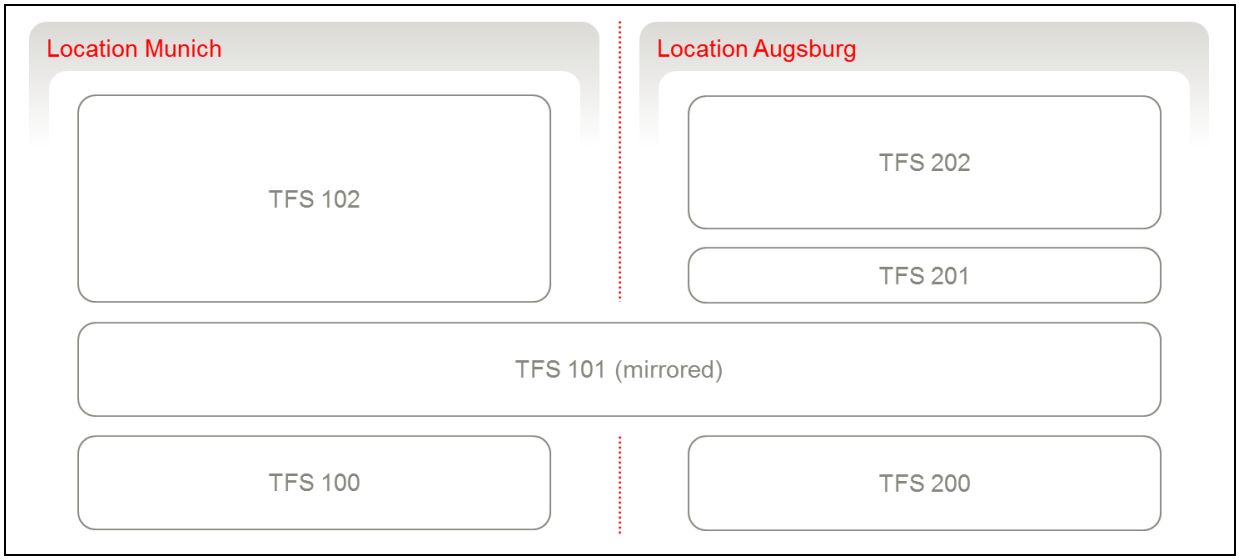


Figure 10

3.4.3 Dual Site Redundancy

ETERNUS CS8000 models for highest availability demands are built up with redundant processor nodes, redundant RAID systems and internal network redundancies. Therefore at least two frontend processing nodes, two backend processing nodes and two RAID systems as well as the appliance administration nodes are available twice. With such redundancies it is possible to distribute an ETERNUS CS8000 system configuration physically over two geographically distant locations A and B whereas the appliance cluster framework still operates as one single logical entity. The internal communication and data transfer infrastructure of the appliance is extended in a way such that the logical structure of the internal SAN and LAN is maintained. The distance between the two locations is high relying from the long distance network infrastructure which has to be provided to extend the internal SAN. The description of the SAN network extension is subject of the dedicated white paper: "ETERNUS CS8000 - Disaster-resilient Architectures" (cf. chapter 9).

With this architecture concept synchronous cache mirroring between locations A and B can easily be implemented for both virtual storage subsystems, CS VTL and CS ViNS. In the event of a network failure between the two system locations the cluster is split into two separated parts and thus has to be managed as a split-cluster with a failover mechanism. Even if the cluster is split into two parts the clustered nodes on each site are able to work with the cluster file system which resides on their side.

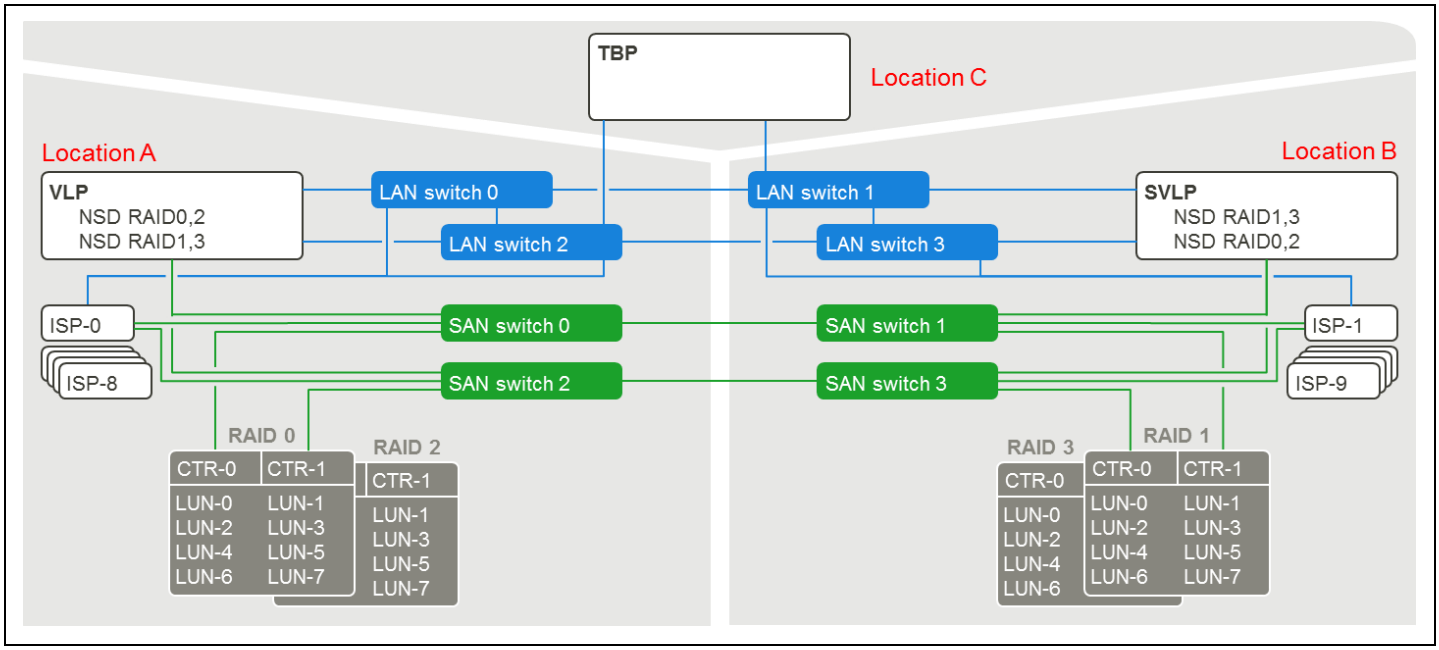


Figure 11

When the appliance CAFS Cluster is configured, a primary and a secondary Cluster Configuration Manager must be defined. By default they execute on the same processing nodes like the two Virtual Library Processors (VLP/SVLP). The Tiebreaker Processor (TBP) is required at a distinct third location C. The TBP functions as an arbitration entity in case the SAN and LAN connection between the locations A and B break down and the CAFS Cluster is split into two parts. The TBP helps to decide which side of the ETERNUS CS8000 system will continue to work and which side will be deactivated. The definition of such quorums is essential that a functioning and consistent file system remains available and that no split-brain situation occurs when a location fails or connections between locations are interrupted.

3.4.4 Automatic Failover

Though ETERNUS CS8000 is armed against the failure of single components by the cluster architecture, internal functions support further availability. The most important function is the VLP which controls the whole appliance. The VLP is prepared for general faults and to support recovery in case of failures or a disaster. Typically almost all internal ETERNUS CS8000 appliance control functions run on the VLP master. This processor is largely protected against disk errors by RAID system disks. If this processor were to fail nevertheless, the appliance would be inoperable. To prevent this situation ETERNUS CS8000 is armed with a hot standby *VLP* – the *SVLP* (Standby *VLP*). The SVLP is monitoring the VLP and with the "automatic *VLP* failover" function (Auto*VLP* failover) and takes over the VLP tasks in the event of failure.

If the VLP fails, the scenario is as follows:

1. The SVLP is active in the system and monitoring the VLP. If the VLP fails, the SVLP takes over control of ETERNUS CS8000.
2. The SVLP is activated automatically: With a switchover the VLP/SVLP pair swaps their tasks.
3. After the defective processor has been repaired, it is integrated once again into the overall system and takes over the role of the SVLP.

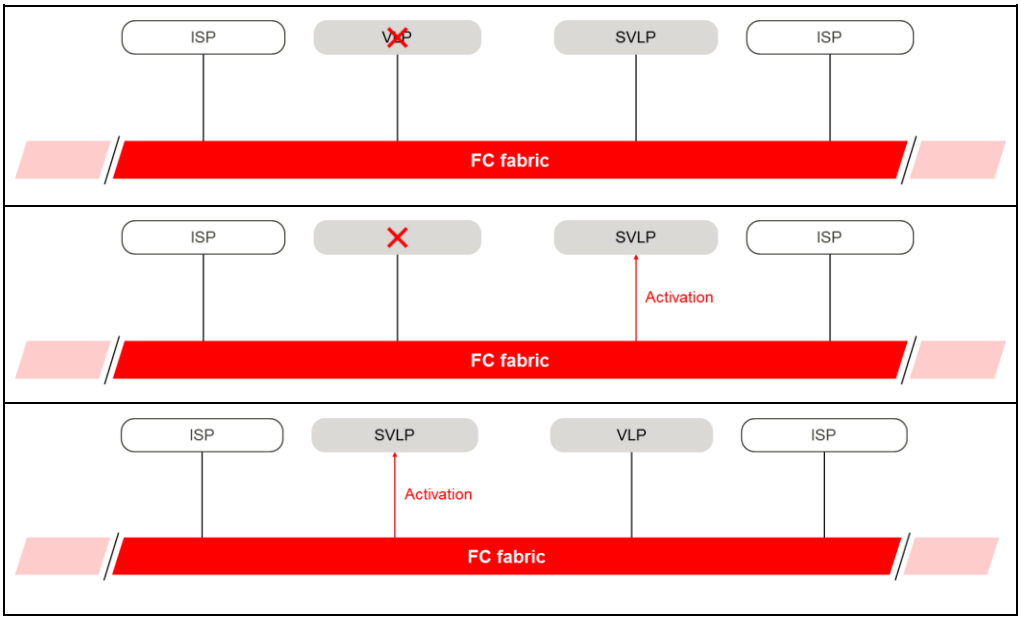


Figure 12

3.4.5 Recovery after Site Failures

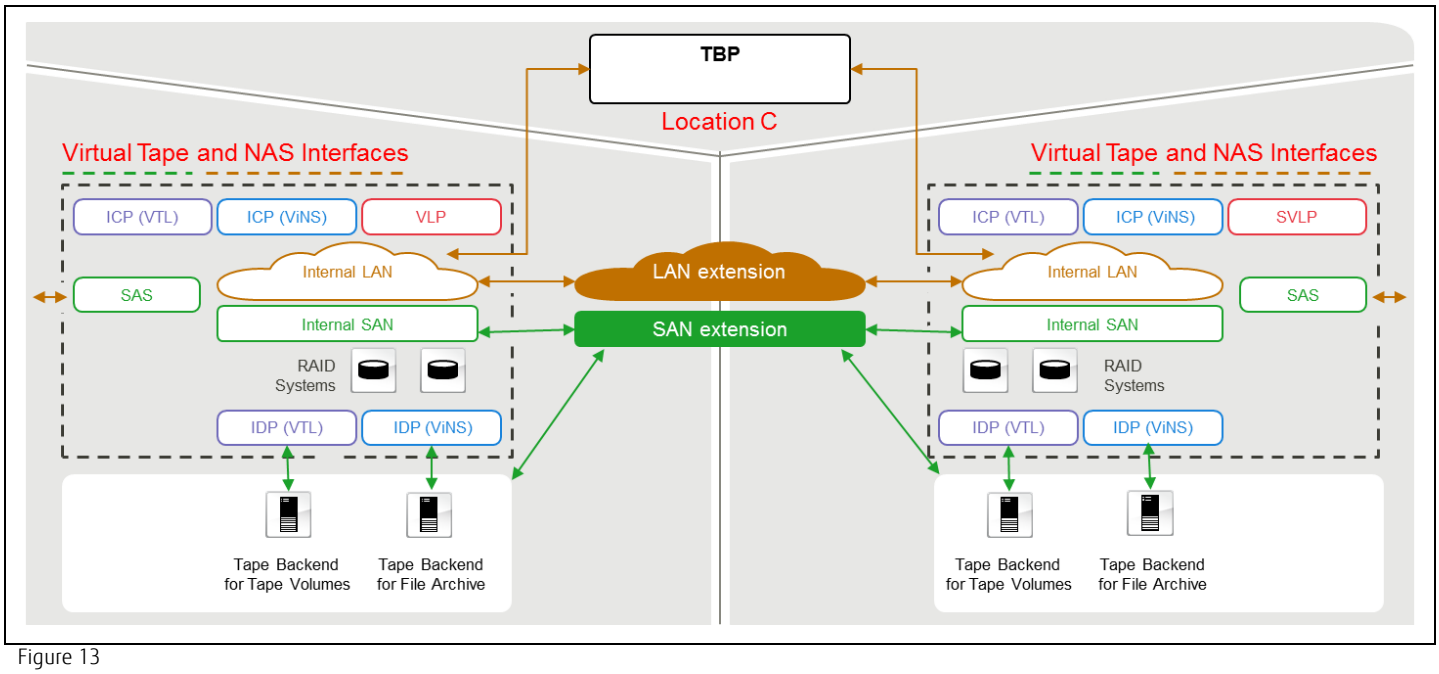


Figure 13

Recovery if only site A fails but site B and site C are not affected:

If the AutoVLP-Failover function is configured, no measures are necessary in order to use ETERNUS CS8000 resources at site B. The resources at site A cannot be used but operations can be continued at site B. Every file system that exists only at site A and has not been mirrored cannot be used; the newly started SVLP at site B takes over the role as VLP and marks the file system from site A as frozen.

Recovery if only site B fails but site A and site C are not affected:

The approach is identical to the one described above, except that site A has to be replaced by site B.

Recovery if only site C fails but site A and site B are not affected:

The ETERNUS CS operations continue at sites A and B. However, it is possible that another failure at site A or B causes. In such a case where the Tiebreaker NSDs are not accessible, operations on a mirrored file system terminates.

Recovery if both site A and site C fail but site B is not affected:

In case of a failure of this magnitude all appliance operations are terminated. Recovery measures will be carried out by Fujitsu Service personnel.

Recovery if site B and site C fail but site A is not affected:

The approach is identical to the one described above, except that site A has to be replaced by site B.

Approach if all FC connections between the sites fail:

Failure of FC connections between the sites does not cause termination of the appliance operations. The file systems managing disks at a remote site can be accessed via the LAN extension to the remote ISP, and also the SVLP at the remote site. However, this causes an extreme load on the appliance internal LAN, which may lead to a significant performance drop as well as internal and external timeouts. ETERNUS CS8000 automatically tries to restore access to the disks via SAN. After recovery of FC connections the disks are used again automatically.

Approach if all LAN connections between the sites fail:

Failure of the appliance LAN connections between the sites may lead to termination of all appliance operations or at least the operations at one site are terminated. What exactly happens depends on the physical connections that the connected network used at the time of the disturbance, as well as on the exact site where the failure happened. After recovery of the LAN connections the appliance will reintegrate the newly connected site or restart operations if necessary. This process runs automatically. However, it may be necessary to restart the ISPs in order to remove all error states or blocked processes.

4 CS VTL Subsystem Interfaces and Operations

The following chapters 4.1 – 4.6 describe the virtual tape operation without deduplication. The concluding chapter 4.7 introduces additional system components and their operation for deduplication of tape data based on the previously explained architecture and concepts.

4.1 CS VTL Interfaces and Logical Components

4.1.1 Introducing the CS VTL Operation

Emulation of physical tape:
The CS VTL subsystem presents by emulation a large number of virtual tape drives to the host and stores tape data on emulated tape volumes. These volumes are stored in the TFS of the RAID storage rather than on physical tape media. The disk space provided in the RAID storage is called Tape Volume Cache (TVC).

Virtual tape drives are also called Logical Drives (LD). Towards the host, logical drives look the same as physical tape drives. Emulation is not apparent to the host and applications. The host always writes to and reads from logical drives. It never accesses the CS VTL backend tape drives and has no knowledge about the existence of these drives. Every application that supports one of the emulated tape technologies can use the CS VTL frontend without any changes.

Because the host exclusively accesses the logical drives, all data must be written to or read from emulated volumes in the disk-based TVC. These emulated tape volumes in the TVC are called Logical Volumes (LV). When the host requests a logical volume that is still in cache, the volume is virtually mounted. After the virtual mount is complete, the host can access the data at disk speed. Mounting of scratch tapes is also a virtual operation.

Virtual tape is always available:
Without a VTL applications are contending for physical tape drives and jobs must wait because no physical tape drive is available. Tape virtualization efficiently addresses these issues by providing many virtual tape drives. Each host can request as many virtual tape drives as actually needed. CS VTL emulates various tape libraries and tape drive industry standards providing in sum a maximum of 1280 virtual tape drives.

High-speed access:
Physical tape requires linear motion to wind and unwind potentially very long reels of media. This tape motion may take several seconds to several minutes to move the read/write head from one end of the tape to the other. The TVC enables the hosts to write and read sequential tape volumes with the behavior of random disk I/O operations. Such the TVC guarantees high-speed access to the logical volumes for the hosts, and avoids bottlenecks which otherwise occur with direct host attached physical tape automation systems.

Efficient use of capacity:
A logical volume takes up just the space in cache that the data on the volume actually requires. A high efficient compression algorithm furthermore reduces capacity requirements. With CS VTL, the logical volumes are copied from the cache to the attached backend devices in physical tape format. CS VTL manages autonomously the tape drives and tape volumes in the tape library attached at the backend. It also controls the transfer of logical volumes between TVC and backend devices. As result the logical volumes need only the amount of capacity occupied by the data, making efficient use of disk and backend capacity.

This operation is identical for all possible attached backend devices which can be physical tape libraries or secondary CS VTL systems:

- **Physical tape backend:** Because many applications do not fill the high capacity media of modern tape technology, standard physical tape operation can end up with many underused cartridges. This wastes much space and requires an excessive number of tape cartridges. To utilize the entire physical tape capacity, CS VTL stacks the logical volumes and writes them sequentially on high capacity physical tape. The process of putting several logical volumes onto one physical tape is called stacking. A physical tape containing logical volumes is therefore referred to as a stacked volume. Tape virtualization with stacking of small logical volumes on one high capacity physical tape reduces the space required by volumes and fully uses the capacity of current tape technology. As an additional result the high data transfer rates of high-end tape drives are efficiently used in streaming mode.
- **Cascaded backend:** CS VTL can also generate un-stacked physical volumes which especially are used in cascaded configurations where one CS VTL is attached at the backend of another CS VTL for asynchronous replication of logical volumes. The use case is described in a later section.
- **Disk library system:** A CS VTL disk library is used to store logical volumes on disk only and such have no tape backend. Instead, logical volumes are maintained on disk until they expire. For this reason, cache capacity for a CS VTL disk library as final backup target should be larger than the capacity for a CS VTL with physical tape backend. Another option exists when one or more CS VTL disk library systems with inexpensive high capacity NL-SAS disk drives are attached as final backup target at the backend instead of physical tape libraries. In such a configuration the primary CS VTL system is used as speed matching buffer for the host attachments and manages data replication over sites. With the deduplication functionality, the disk capacity requirements maybe dramatically reduced (chapter 4.7).

Autonomous data migration:

When space is required in the TVC for new data, logical volumes that already have been copied to the backend are removed from the cache. By default, removal is based on a *least recently used* (LRU) algorithm. Using this algorithm ensures that no new data or recently accessed data is removed from cache. The process of copying logical volumes from cache to backend tape and then removing them is called *migration*. Logical volumes that have been deleted in the cache and exist only on tape backend are called *migrated volumes*.

Usually a previously migrated volume must be copied back from backend tape into the TVC to be accessed. It must be copied because the host has no direct access to the backend. When the complete logical volume is copied back into the cache, the host can access the data. The process of copying data back from backend tape to the TVC is called *recall*.

With this procedure the appliance software dynamically maps data written over virtual tape drives onto physical tape and transparently manages the usage of physical drives and media. Allocation and migration of the backup data to physical tape is managed automatically by means of policies. An intelligent cache management provides always sufficient space in the cache by migration of logical volumes to physical tapes in the attached tape libraries. The lifecycle of specific volumes in the TVC can furthermore be managed by specifying individual retention periods.

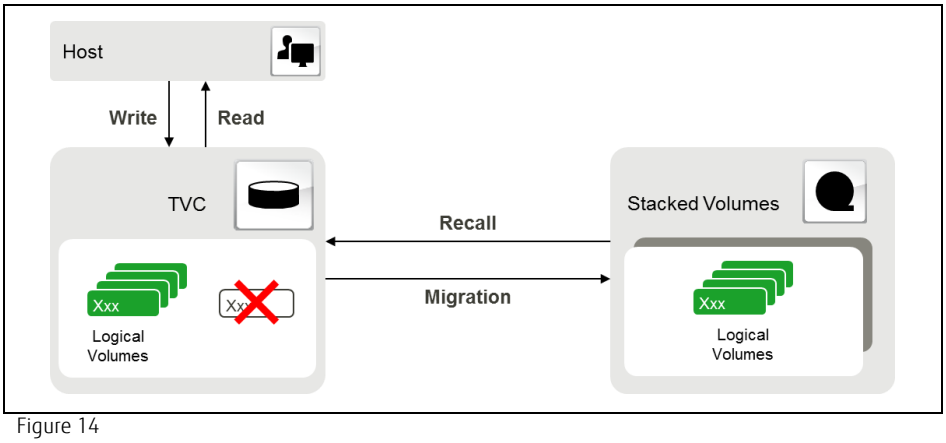


Figure 14

Highest data availability:

Depending on the configuration, CS VTL provides different types of integrated high availability and disaster recovery mechanisms:

- Synchronous replication of logical volumes on base of CMF (Cache Mirroring Feature) over short and medium distances, meaning that at each point in time the data on the source and target is the same. With CMF one CS VTL cluster configuration is distributed over two sites but seen as one logical entity.
- Asynchronous replication of logical volumes in cache over very long distances where CS VTL systems are cascaded behind each other. Asynchronous replication means that a certain delay exists for having data transferred from source to the target.
- Replication of physical volumes (Multiple Save) at the backend where CS VTL writes up to three copies of physical volumes at the same time to different tape drives (physical or/and virtual drives).

Any data replicated on physical volumes is accessible through any other CS VTL system. In worst case after a disaster when only offsite physical volumes are available, all stored data can be retrieved for recovery, as CS VTL stores also metadata on each physical tape. Through remote site access to the backend physical tape library, two separated CS VTL systems can also exchange tape volumes with each other. Setting policies defines where and when multiple copies of data are created.

4.1.2 CS VTL Frontend

The CS VTL subsystem operates as Virtual Tape Library and emulates as frontend service a set of physical tape libraries with all its interfaces. It provides similar interfaces for library control and tape I/O as defined by industry standards. In order to make use of a tape library, backup or archiving applications running on hosts use two kinds of interfaces:

- **Library control interface:**
  The library control interface is sometimes referred to as robotics control, since in real libraries a robot mechanism is in operation to perform tape cartridge movements between cartridge slots to a specific tape drive (mount) or from a tape drive back to their slot (dismount). For large tape libraries the library control interfaces are TCP/IP based and emulated within CS VTL by a dedicated library specific service on the VLP (Virtual Library Processor). Some backup applications address small libraries over Fibre Channel. Therefore the ICP (Integrated Channel Processor) supports the virtual SCSI library control emulation over Fibre Channel.
- **Tape I/O interface:**
  Once a tape cartridge is mounted in a tape drive, the host uses the tape I/O interface to read/write data from/to the magnetic tape cartridge or to perform other related operations such as seeking to specific positions. The frontend services with tape I/O interfaces are processed from CS VTL by the ICP (Integrated Channel Processor). Therefore the ICP appliance processor node is equipped with appropriate Fibre Channel hardware controllers. In a standard SAN configuration CS VTL can be connected with all state of the art FC controllers and SAN components up to 8 GBit.

Since the library interfaces and tape drives are virtualized by software running on the appliance processor nodes, different specific library and tape drive interfaces are possible in operating concurrently within the same CS VTL. The interfaces are configured depending on the needs of the related host systems. CS VTL can also be shared by several host backup applications and is seen by the hosts alike a partitioned physical tape library. The various emulations of CS VTL support different hosts with different behaviors and 100% requirements of their backup applications for tape storage.

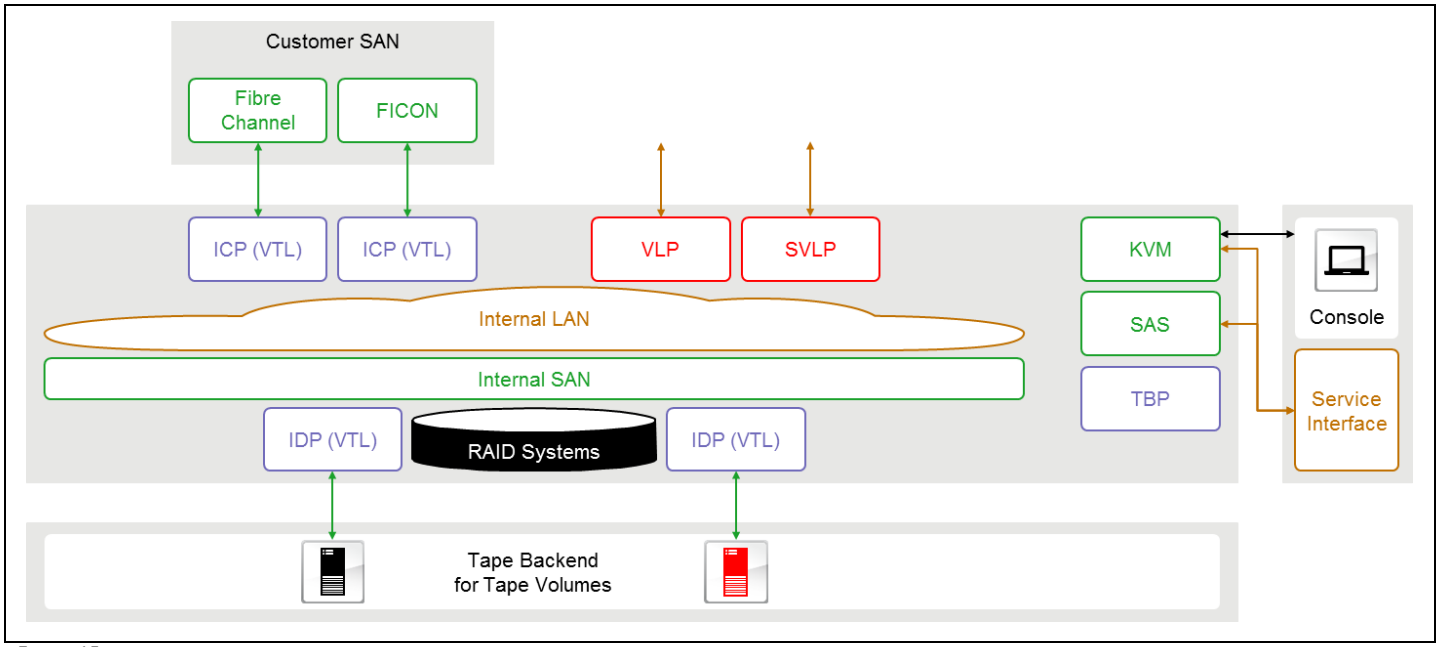


Figure 15

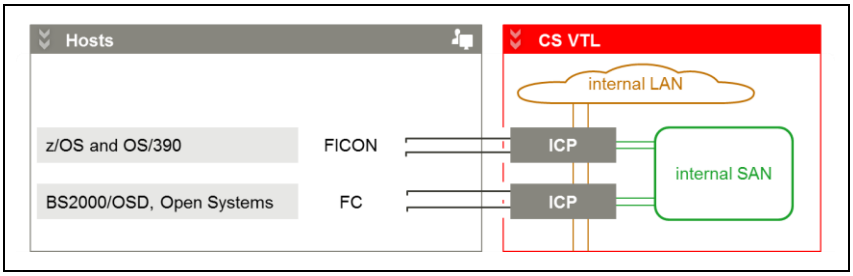


Figure 16

On top of the Fibre Channel device drivers the ICP implements with its logical drives (LD) a set of emulated tape drive interfaces to handle data transfer from and to the hosts. A logical drive is a tape drive emulated by software which presents the host applications the interface of a real physical drive. Two types of emulation programs are distinguished by the appliance; these are referred to as VTD (Virtual Tape Device) and 3490 Emu (or EMU for short). VTD emulations are used when the hosts are connected via FC SCSI. The emulated device types are IBM/IBM 3590E1A, IBM/LTO Ultrium 2, STK/9840B and a FSC specific type. 3490 emulations are used when the hosts are connected via FICON. The emulated device types are IBM 3490-E and Fujitsu F6473K.

The logical drive implements the protocol for communicating with hosts and transfer data to/from LVs (Logical Volumes) located in the TVC. They also communicate with the VLM (Virtual Library Manager) because they need to be informed about the LVs that are logically mounted in the logical drive.

Logical volumes show the same properties like physical tape cartridges normally written by hosts. Procedures established by users for allocating volume identifications can be used without changes. With CS VTL logical volumes may be of various types, formats and size, since they are mapped by software onto physical storage.

Like with physical tape drives compression can be switched on or off whereas the ICP internal processor performs a high efficient compression algorithm (3:1) which furthermore reduces needed disk space for storing LVs in the TVC and finally also on physical tape. Compressed data is transparently decompressed when requested from the host.

Virtualization of tape drives offers following benefits:
- The virtualization appliance offers a multitude of tape drives than a physical library
- The multitude enables enhanced job parallelism
- Virtual tape drives can be shared like physical tape drives
- Virtualization and buffering in the TVC requires fewer real tape drives
- Frontend compression decreases storage capacity
- Existing host processes can be served and operated unchanged

Multiple ICPs can be configured to scale up the system and to guarantee a high number of available logical drives, high performance as well as high availability. Simultaneous connection with Mainframes, Open System and NAS storage systems is possible.
A maximum of 10 ICPs is available for connection to the hosts, where each ICP can control a maximum of 128 logical drives. Mixed operation of the different device emulations in one ICP is fully supported.

Technical Data:

- Maximum number of logical drives: up to 1280
- Maximum number of logical volumes: up to 3,000,000
- Logical volume size (uncompressed): 900 MB and 2, 5, 10, 20, 50, 100, 200 GB.
  (additional capacity with compression factor usually 1.5 up to 3)

Following interfaces and tape drive emulations are available:

Interface	Tape Drive Emulation
FICON	IBM 3490-E, Fujitsu F6473K
Fibre Channel	Generic SCSI, EXABYTE Mammoth, FSC specific type, IBM 3590-E, STK T9840B, STK T9840D, LTO Ultrium 2, LTO Ultrium 4
ETHER	CS VTL device type for "Cascading Destination"

4.1.3 CS VTL Backend

The backend services of CS VTL are performed by the IDP (Integrated Device Processor) and process the communication with the physical drives (PD). The IDPs are equipped with Fibre Channel controllers for connection and data I/O to the PDs.

The IDP backend service handles the data transfer of logical volumes between the TVC to/from physical tape cartridges in the backend library converting the disk format into the technology of the physical drives. The IDP has the capability to write new files onto already written tape cartridges by appending another logical volume after the last one written. A cartridge filled with multiple logical volumes like this is also referred to as a "stacked volume". As the appliance generates such stacked volumes directly from disk, the physical drives at the backend are always operated with optimum speed. Optionally an additional compression of data can be switched on or off by control of these capabilities with the backend tape drives. WORM media which prevents overwriting can also be used. For special cases the appliance generates also un-stacked volumes in the same tape format like written from the host (native tape).

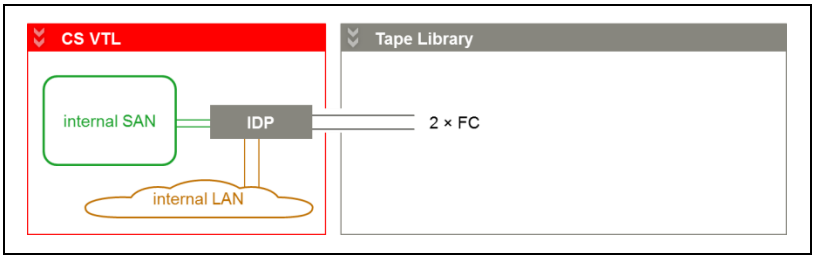


Figure 17

A maximum of 4 FC drives can be connected to one HBA (Host Bus Adapter) port of an IDP. This means: Being equipped with 2 FC boards, each with dual ports, an IDP supports up to 16 drives. Multiple IDPs can be configured to guarantee large number of attached physical tape drives, high performance as well as high availability. In sum up to 10 IDPs each with up to 16 external tape drives are supported. In case the tasks of an IDP fail, they can be taken over by another IDP.

If the backend library control is also done through the SCSI protocol, the backend service provides also the library logic control functions to a direct attached SCSI library which is connected via FC.

Following physical tape drives are supported:

Manufacturer	Tape Drives
LTO Ultrium	Gen 1, Gen 2, Gen 3, Gen 4, Gen 5
IBM	Magstar 3590, Jaguar 3592 / TS1120 / TS1130, TS1140
Oracle / StorageTek	T9840, T9940, T10000

Certain ETERNUS CS8000 models are operated in Disk Library mode without tape backend where all data is kept in the RAID storage. Such systems don't need backend data transfer services and thus don't need IDPs.

Alternatively to a physical tape library, a secondary CS VTL can be attached at the backend.

4.1.4 Virtual and Physical Library Control

The Virtual Library Processor (VLP) is the core of the ETERNUS CS8000 appliance for coordination of the internal processes of all components and to process commands from the administrator. As standard of dual site systems two separate Virtual Library Processors - VLP (VLP Master) and SVLP (Standby VLP) - are installed for Automatic Failover operation. The SVLP monitors the operational status of the VLP Master and is kept ready to take over the VLP Master's tasks if this is out of order (cf. chapter 3.4.4)

In addition, the VLP is responsible for the cache management and to maintain the database of descriptive data concerning logical volumes. The Meta data of the volumes to be read or written is also stored on the 1st RAID system (on two RAID systems for CS2000 and higher) to guarantee fast availability in case of a disaster. In the event of a total failure of the VLM/PLM (Virtual Library Manager/Physical Library Manager) system on one site, both components can be started on the other carrier system and can take over the tasks of the failed system. Failover in the event of a problem happens automatically.

The main logical components of the VLP are:
- Virtual Library Manager (VLM)
- Virtual Library Service (VLS)
- Physical Library Manager (PLM)
- Physical Library Service (PLS)

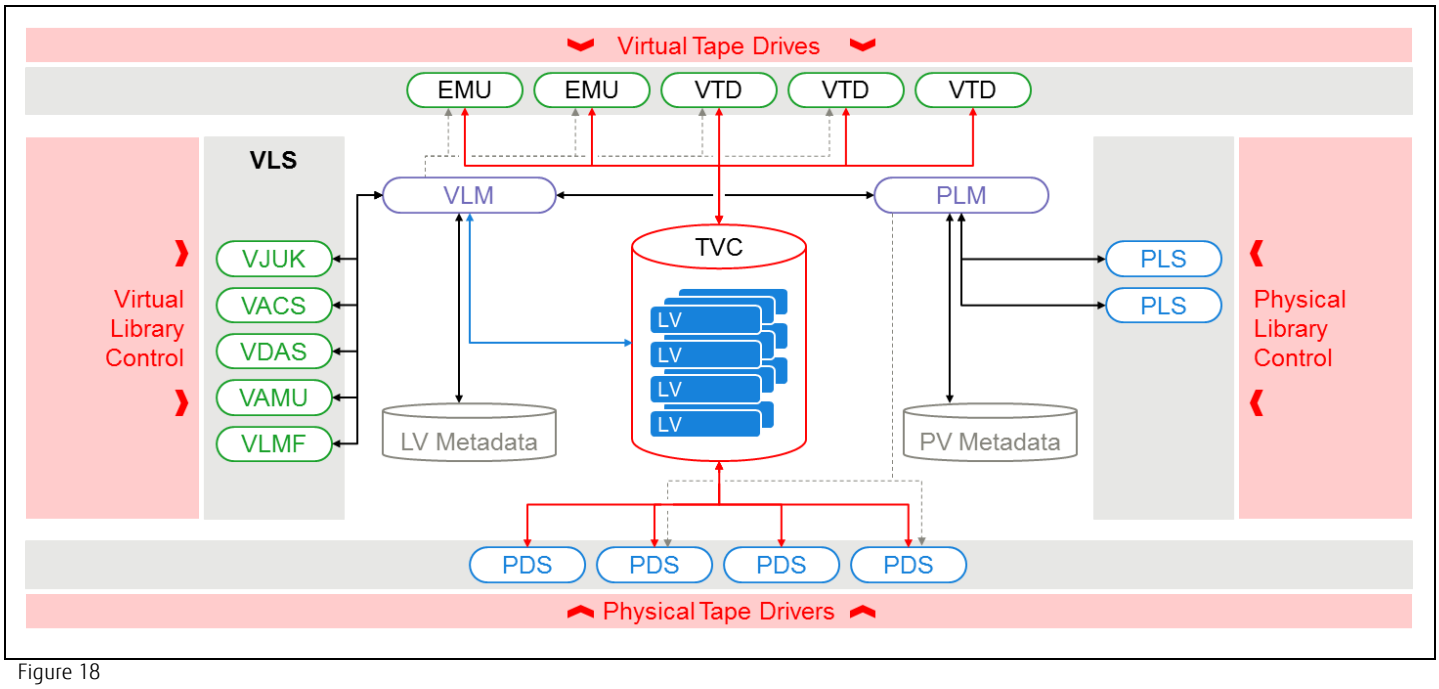


Figure 18

Virtual Library Manager (VLM)

The VLM functions as a central management instance for logical volumes in the TVC. It receives mount/dismount requests from hosts - indirectly through VLS - and assigns LVs to virtual drives. In a second function the VLM works as cache manager for the TVC. It keeps track of the state of LVs and the amount of available space in the TVC. Appropriate cache management algorithms are applied to select LVs to be displaced to physical tapes. The VLM keeps track of the logical volumes in an own database that contains the names, status and attributes of the LVs. Additionally the VLM manages the library control interface at the frontend with the VLS and at the backend with the PLM.

Virtual Library Service (VLS)

The VLS module implements the library control interface presented to hosts. The VLM communicates with the VLS which processes the virtual library control interfaces to exchange library control commands with the connected host systems at the frontend by emulating different industry standard physical library services. With this virtualization of different library types the connected hosts at the frontend are independent from type and manufacturer of different physical tape libraries attached at the backend.

Various emulations offer simultaneously different virtual library types for multiple parallel host connections which operate each with the corresponding software components in the host. The different instances of the VLS communicate with the host and forward requests to the VLM. Depending on the type, two groups of tape library control interfaces exist:

- SCSI library control - SCSI specification for medium changer devices.
- Library control via TCP/IP connections.

Virtual library services run on the VLP - except for VJUK which emulates SCSI tape libraries, which runs on ICPs. The SCSI-library system interface is used for the operation of SCSI devices (e.g. Autoloader in the low-price segment). SCSI tape library emulation is done via the Fibre Channel connections at the CS VTL interfaces. A SCSI library appears as a separate LUN (device) on an assigned FC target.

Depending on the emulated library type one or more sub modules on the VLP emulate logical libraries with TCP/IP control connection like StorageTek-ACSLS (VACS), Quantum-SDLC (VDAS or VAMU) or Fujitsu (VLMF) with the specific library control protocols. The originally proprietary standards from StorageTek and Quantum for the control of library systems, have found broad acceptance by several manufacturers.

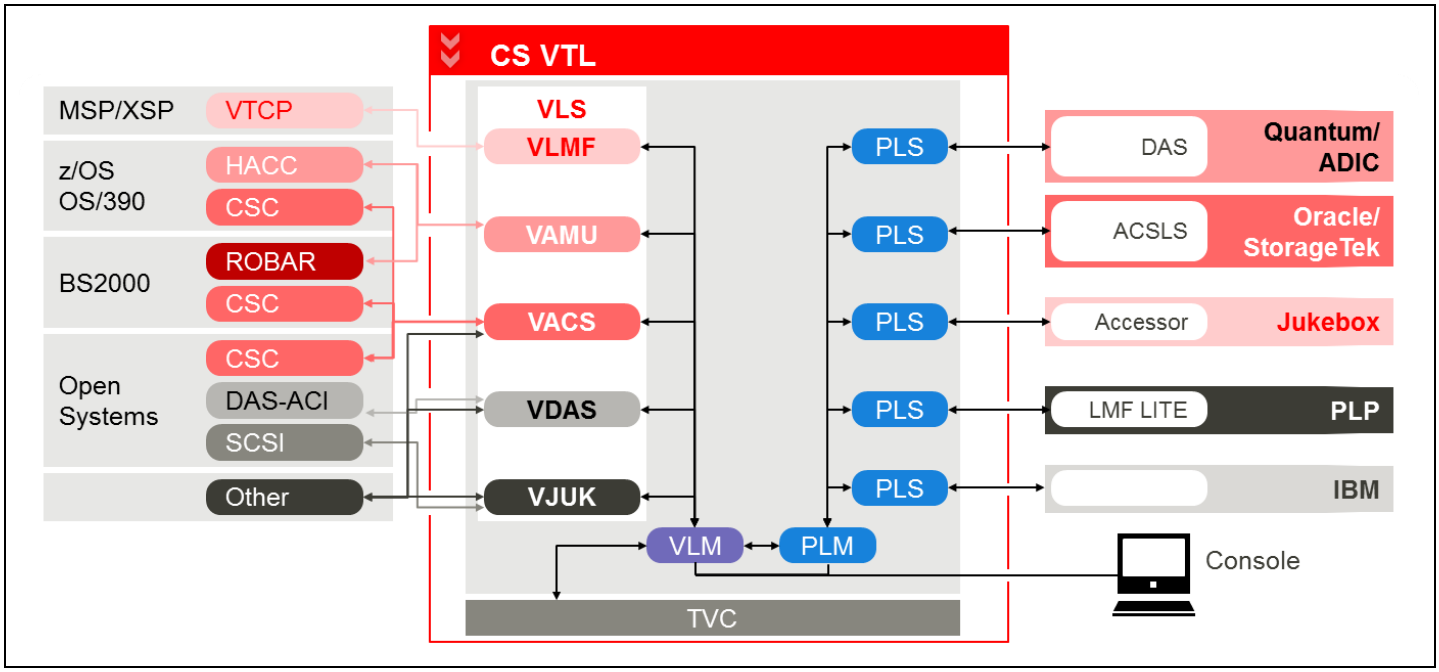


Figure 19

Physical Library Manager (PLM)

The VLM performs also the communication with the PLM which manages the physical tape backend. The PLM processes one or several PLS (Physical Library Service) control interfaces for physical tape libraries attached at the backend. Thus the PLM coordinates all jobs issued to the connected backend peripherals (tape libraries, tape drives). The PLM maintains an own database that contains information about where, on how many and on which physical volumes each logical volume resides.

Physical Library Service (PLS)

The PLS is the link between CS VTL and the backend tape library and maintains the control path of the connected physical tape library. One PLS runs for each physical library and implements the specific library control interface. Jobs to the tape library, e.g. mount/dismount of physical volumes (tape cartridges) to drives are executed on behalf of the PLM.

Physical Device Service (PDS)

Like the PLS a PDS is controlled by the PLM but maintains the connection to the physical drive. The PDS handles the data transfer of logical volumes in the TVC to/from a physical tape drive in a backend library. One PDS runs per physical drive.

Backend Support:
Following physical tape libraries are supported:

Manufacturer	Tape Library
Fujitsu	ETERNUS LT40, LT60, LT130, LT160, LT270 Legacy systems: 9084-221, 3560, TX24/TX48
IBM	3494, 3584 / TS3500
Oracle / StorageTek	SL500, SL1400, SL3000, SL8500, L180, L700/E, L5500, PowderHorn 9310, TimberWolf 9740
Quantum / ADIC	Scalar i500, Scalar i2000, Scalar i6k, Scalar 10k, Scalar 100, Scalar 1000, AML/J, AML/E, AML/2
Spectra Logic	Spectra T50, T120, T200, T380, T680, T950

Frontend Support:
The library control interfaces are attached to the respective software components running by the operating systems:

CS VTL control interface	Connection standards from	CS VTL service for operating systems
VJUK (emulation of SCSI library (jukebox))	SCSI specification for Media Changer Devices (Autoloader)	Open Systems
VACS (emulation of ACSLS Server)	Oracle/StorageTek	Open Systems (CSC) BS2000 (BS2000-CSC) OS/390 and z/OS (MVS-CSC) z/OS, i5/OS, OS/400, VM, VSE (LMS) GCOS 8
VDAS (emulation of DAS/ACI Server)	Quantum/ADIC	i5/OS, OS/400, VM, VSE (LMS)
VAMU	Quantum/ADIC	BS2000/OSD (ROBAR) z/OS and OS/390 (HACC-MVS)
VLMF	Fujitsu	Fujitsu MSP, XSP (VTCP Clients)

Following operating systems are certified (Mainframes and Open Systems simultaneous):

Manufacturer	Operating System
Bull	GCOS 8
Fujitsu	MSP, XSP, VME (ICL) BS2000/OSD
IBM	z/OS, OS/390, i5/OS, z/VM, z/VSE
Open Systems	AIX, HP-UX, Solaris, SUSE LINUX, Red Hat LINUX, z/Linux WINDOWS Server (2000, 2003 and 2008) NDMP Backup (NetApp, EMC)

Following backup software for open systems is certified (excerpt):

Manufacturer	Operating System
Atempo	Time Navigator
BakBone	NetVault
BridgeHead Software	Hypertape
CommVault	Simpana
Computer Associates	BrightStor ARCserve Backup
EMC	Networker, DiskXtender
Gresham	EDT
HP	DataProtector
IBM	Tivoli Storage Manager
Microsoft	Data Protection Manager
Seven Ten Storage	StorFirst Altus
Symantec	Backup Exec, NetBackup, EnterpriseVault

Following storage systems for NDMP backup are certified:
- NetApp Filer
- EMC Celerra

Following applications with Mainframes are certified
- Computer Associates – Dynam, VM:Tape, Dynam/T
- IBM – BRMS, Tivoli Storage Manager, DFSMShsm, DFSMSdss
- Fujitsu - ARCHIVE/HSMS

4.1.5 Tape Volume Cache (TVC)

The Tape Volume Cache (TVC) is a kind of speed matching buffer to which logical volumes (LV) are written in the same various formats as received from the ICP frontend written from host to the logical drives (LD). The Tape Volume Cache (TVC) on base of the internal RAID system enables high-speed tape storage processing and the attachment of high-capacity tape technology at the backend. Within the common cluster the TVC is accessible by all participating clusters components. The architecture allows for any logical volume in cache to be accessed by any cluster component. This capability results in the creation of an effective cache size and any logical volume stacked in physical tape can be recalled into tape volume cache making it available to any cluster component.

The following benefits are offered:
- Logical volumes are accessed at disk speed. Tape commands such as space, locate, rewind, and unload are mapped by the ICP into disk commands that are completed in tens of milliseconds rather than the tens of seconds required for traditional tape commands.
- Multiple logical volumes can be accessed in parallel because they physically reside in the tape volume cache. To ensure data integrity, a single logical volume cannot be shared by different jobs or systems at the same time.

4.2 Information Lifecycle Management with Virtual Tape

Basically the backup data from hosts is written in tape format into the TVC. CS VTL stores this tape backup data as logical volumes, this means logical volumes are stored as large files within the appliance tape file system. According predefined policies CS VTL manages the logical volumes in various manners and controls the storing on internal disk and on physical tape at the backend.

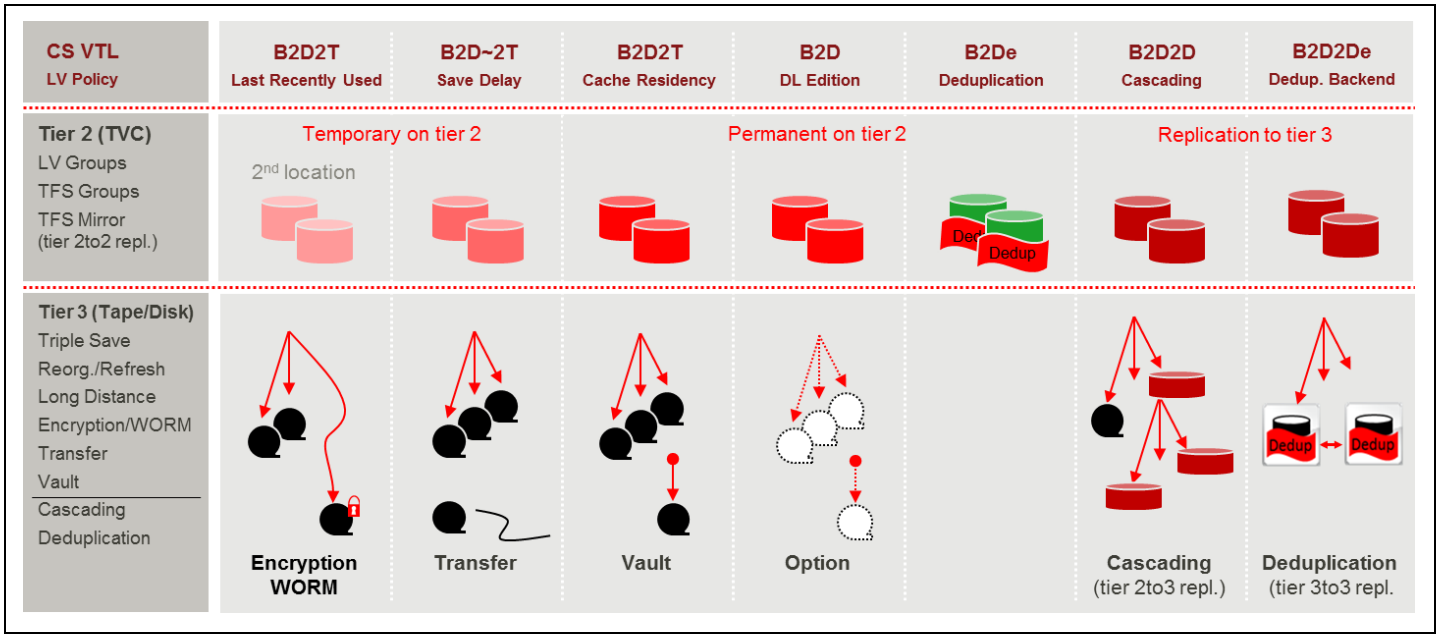


Figure 20

4.2.1 Overview of Policies for Tape Volume Processing

According the importance of tape backup data and their related service levels the logical and physical volume management allows various policy operations for the contents of the TVC and to assign Logical Volumes (LVs) and Physical Volumes (PVs) to separated groups. Sets of LVs are organized in Logical Volume Groups (LVGs). Certain attributes may be assigned to a LVG which then apply to all LVs that are member of the LVG. With a similar concept PVs managed by PLM are combined to Physical Volume Groups (PVGs). By defining rules that create logical associations between LVGs and PVGs, it is possible to assign the LVs belonging to a particular LVG exclusively to a dedicated PVG. The group assignment is managed by the VLM and stored into its own database.

This grouping allows determining which content in the TVC is preferred for various operations like:

- Storing of LVs within the TVC
  - Assign logical type of emulated tape media at the frontend
  - LV pooling with LVG for different policies
  - Storing two redundant copies of a LV in the TVC
  - Determine for removal of LVs in the TVC after the PV is written to physical tape
  - Improve mount response time by keeping often required LVs in TVC for longer periods
  - Keeping LVs always in the TVC even if they are already stored onto physical tape (Cache Residency)
  - Storing LVs only on disk in the TVC without writing on tape (Disk Library Mode)
- Writing LVs assigned to a PV on physical tape
  - Determine to write LVs immediate on tape (Sustained Mode) or after a predefined time (Save Delay)
  - Writing multiple redundant copies of a LVs simultaneously onto different physical tapes (Multiple Save)
- Target options for physical tape
  - Physical tape pooling with PVG for different policies, different locations, different type of media, kind of replication, type of backend storage (tape or disk)
  - Use tape encryption
  - Use WORM media
  - Determine for the PV to be transferred offline to another CS VTL system location (Transfer) or to be preserved at a vault location (Vault).
  - Determine cycles for Tape Refresh, Tape Reorganization, Tape Technology Migration
  - Use instead of a physical tape library a secondary CS VTL (with or without deduplication).

4.2.2 Correlation between Logical and Physical Volumes

LVs and PVs (related to physical tape cartridges) are named entities. Names of LVs have to be unique within one CS VTL. Data received through the virtual tape drive interface of the VTL is stored in LVs in the TVC.

A LV is a data container with a fixed maximum size (set by configuration when the LV is created). Data is written and read to/from LVs with sequential semantics (tape semantics). Each LV is assigned a unique identifier referred to as Volume Serial Number (VSN). This is similar to a barcode label used for physical tape cartridges. The VLM in its function of Volume Manager keeps track of the LVs, their VSN and their state.

LVs can be created with a size of 2 GB or more instead of with the standard size of 900 MB. Depending on the model, up to 3 million LVs can be defined. The LVs size can be selected in discrete steps for each LVG and can be enlarged (but not made smaller) for each LVG during running appliance operation:

- STANDARD: 900 MB
- EXTENDED: 2 GB, 5 GB, 10 GB, 20 GB, 50 GB, 100 GB, 200 GB

The data container of a LVs is a file which is normally stored only temporarily in a TFS of the TVC to be finally stored permanently in a PV. The VLM acting as TVC manager takes care of LVs data backup/displacement to physical tapes. A physical tape is internally called PV (Physical Volume). The VSN of a PV is defined when the PV is added to a PVG. The PLM is the instance that keeps track of the PV.

On the physical backend labeled tape cartridges are inserted into a tape library. Within a single tape library cartridges have to be uniquely labeled. They are identified by their location (in an attached physical library) and their barcode label (referred to as VSN). Within CS VTL the two names spaces of LVs and of PVs are independent, i.e. the cartridges used by CS VTL at the backend are not visible to the hosts.

Modification of LVs on tape and handling of old or deleted data:
Writing onto physical tapes starts always after the last valid tape mark. Thus, updating a LV which resides not at the end of a cartridge cannot be done in place. The updated version is rewritten in total, usually at a different place and a different physical tape cartridge. The space allocated to the LV before the update still contains the old data, but the data is marked in the appliance database as invalid, and CS VTL guarantees that invalid data cannot be accessed. Ultimate deletion of obsolete data takes place on technical decisions to reorganize the cartridge.

Example: Handling a volume mount request from a host

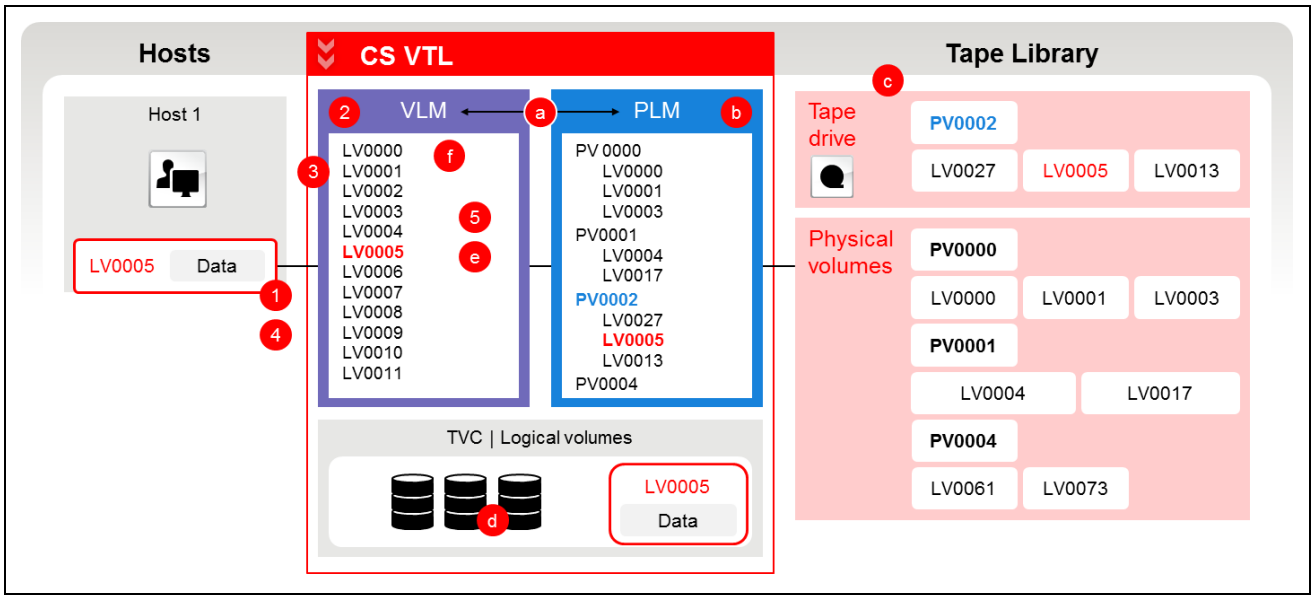


Figure 21

Initial situation: The logical volume LV0005 is already located on the physical volume PV0002. A mount job is executed as follows:

1. The host issues a mount job for LV0005, which is then accepted by the VLM. The VLM does not know at this point what task is involved:
- read the volume or a part thereof
- append a file to the end of the volume
- overwrite the entire volume

2. The VLM checks its database to establish whether LV0005 specified by the host is available and whether there is a corresponding free storage space in the TVC. If the TVC does not have enough free capacity at this point, the LRU (Least Recently Used) procedure is employed to delete the oldest data from the TVC. If a sufficient number of old files cannot be deleted, the mount job is suspended (Mount queued).

Depending on whether the LV is still in the TVC or is only on a PV, the following two situations arise:

Case 1: The volume is migrated to tape and is no longer located in the TVC.
a. The VLM issues a request to the PLM to read LV0005 into the TVC.
b. The PLM checks its database to determine the physical volume on which the requested LV0005 is located: PV0002.
c. The PLM requests the robot to mount the physical tape cartridge PV0002 onto a free tape drive.
d. The data of LV0005 is transferred from the tape drive into the TVC.
e. A flag is set in the VLM database to indicate that LV0005 is in the TVC.
f. Only at this point does the VLM grant the host access to the volume (mount acknowledged).

Case 2: The volume is present in the TVC. The VLM immediately grants the host access to the volume.

3. The host performs read and write accesses on the LV.

4. The host issues an unmount job. In contrast to a physical tape library system, the job will be confirmed immediately.

5. The VLM checks whether the LV in the TVC has been modified.

Case 1: The LV has not been modified: No further action is taken, since the copy of the LV on the PV is still valid.

Case 2: The LV has been modified.
a. The VLM informs the PLM that the LV is to be copied onto tape.
b. The PLM selects a suitable tape cartridge: a completely new tape, a scratch tape, or a tape onto which writing has not yet resulted in an overflow. If this cartridge is not yet mounted, the PLM checks whether at this point a physical tape drive is available in the tape library.
c. The PLM requests the selected tape cartridge to be mounted, if required, and begins data transfer from the TVC to the tape. The data of the modified LV is retained in the TVC until deleted by the VLM in accordance with the LRU procedure.

4.2.3 The Volume Group Concept

CS VTL allows the formation of volume groups, which are named entities. Thereby LVs or PVs which have identical attributes can be grouped and their attributes can be managed by set operations, i.e. for all elements of the group with a single operation, and have not to be managed individually for each single element. Management of LVs and PVs is simplified, which is important because of the high overall number of volumes in many installations.

Grouping is effective for LVs by LVGs (Logical Volume Groups) as well as for PVs by PVGs (Physical Volume Groups). Each LVG is associated to at least one PVG. Each PVG is uniquely associated to a tape library. In case of a tape library which has tape drives of different types, each PVG can only be associated to tape drives of one single type within the library. Therefore, in order to operate a tape library with two different tape technologies at least two PVGs are needed.

It is possible to configure:
- Up to 3 million LVs
- Up to 513 LVGs
- Up to 101 PVGs

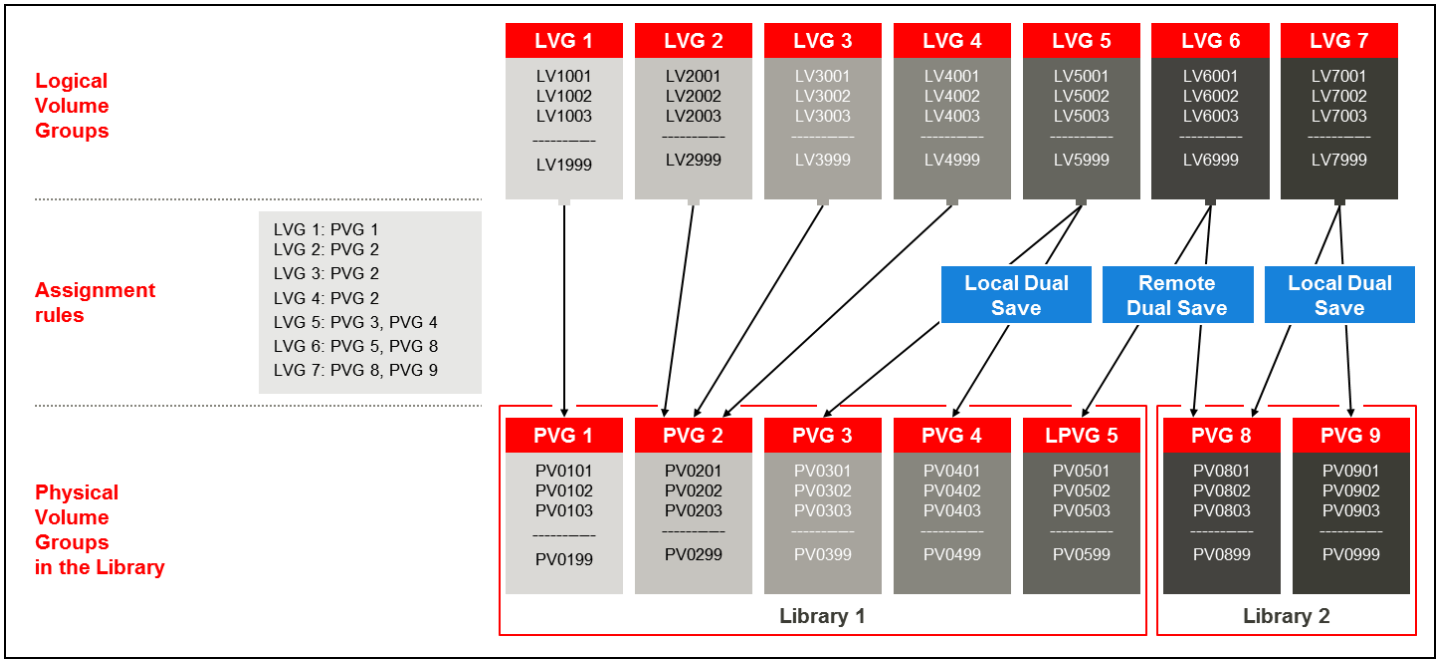


Figure 22

Because hosts do not know the concepts of volume groups, and because they address volumes preferably by "name" (dependent on backup software: Volser, VSN, Label), sets of LVs is most conveniently formed by grouping LVs by their name. Grouping of cartridges by a PVG is also accomplished by the administrator in consideration of the LVGs assigned to the PVG. PVGs are one of several means for the administrator to influence the operation of CS VTL. Thus practically, virtual tape libraries of different type and capacity can be provided for the hosts, based on a single, centrally managed physical data storage.

For each LV, as seen and addressed by the host, the LVG determines uniformly for all its LVs:
- Properties of the LVs as format, type of media and maximum size.
- On which PV the data of LVs are permanently stored, by the assignment of the LVG to one or more PVG. E.g. in case of a LVG being assigned to two/three PVG, all the LVs of the LVG are saved redundantly on both/three PVGs (Dual Save/Triple Save).

The assignment of a LVG to a PVG is limited by the size and capacity of LVs from this LVG and the capacity of cartridges from the PVG. A LV of maximum size must be stored as a whole on one single physical cartridge and cannot be separated into parts and continued on a second physical cartridge.

Separating of backup data with the concept of volume groups is an appropriate means to establish different operational behavior for separate usage requirements:

- Rule-based management of tape data including redundancy in the TVC (Cache Mirror)
- Write throughput (dependent from TVC disk type SAS or NL-SAS)
- Recovery time including maximum delays by mount wait times
- Storing LVs of various host systems or applications on different PVs.
- Data separation from different organizations on separate PVs to fulfill certain outsourcing contracts. Organizations can "see, feel, and touch" their data by having only their data on dedicated media. Easy charging for tape by the number of volumes in a tape library.
- Multiple Save with redundant copies of a LV on different PVs. This offers an extra degree of data security for situations where a tape becomes unreadable
- Manage different media types (large capacity media with long recall time or fast recall media)
- Manage different data locations (for LVs stored in the TVC and for PVs stored in the tape library)
- Data protection on media including encryption
- Data retention and statutory archiving regulations including archiving on WORM tapes

Example:
Two different systems (a mainframe host and a UNIX system) using CS VTL in conjunction with a common type library system. By grouping volumes, mainframe data and UNIX data can be stored on different PVs. The LVs of the mainframe are assigned to the LVG1, while those of the UNIX system are assigned to the LVG2. These LVGs are assigned to different PVGs. As a result of these assignments, mainframe data will now be stored on the PV PV0001 through PV0300, while UNIX data will be stored on the PV PV0501 through PV0600.

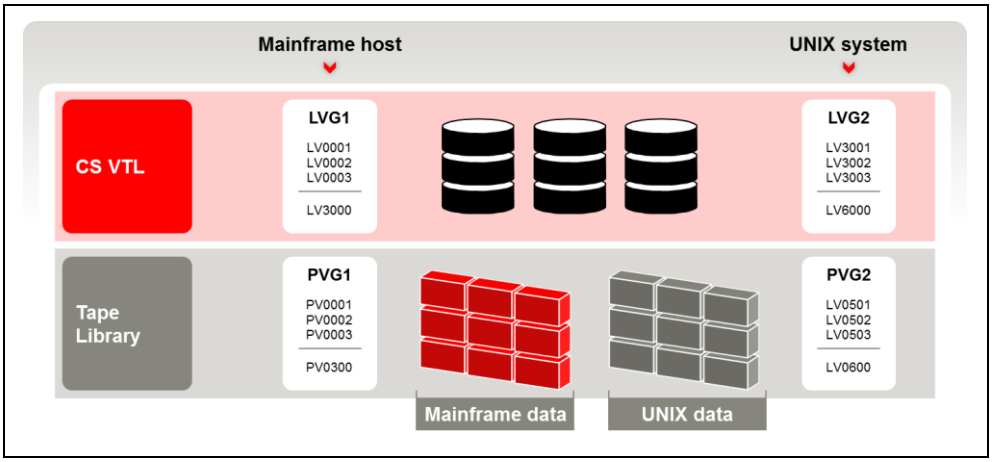


Figure 23

Before defining LVs with CS VTL, the total number of LVs required has to be considered, the volume serial ranges to be defined, and the number of volumes within each range. The VSN for LVs and PVs must be unique. The number of LVs that are required depends from the planed CS VTL workload. The numbers of PVs that are required depends from the workload and accommodate capacity at the backend of CS VTL. The following information needs to be considered:

- The number of LVs/LVGs
- The average amount of data on a volume
- The average compression ratio achieved for the data
- If the Multiple Save function is to be used
- Whether the VAULT function is to be used
- The number of PVs/scratch PVs
- The number of PVGs

Partitioning on the basis of volume groups comprises the following system administrator activities:

- Adding a LVG
- Removing LVG
- Adding LVs to a LVG
- Moving LVs to another LVG
- Removing LV
- Adding a PVG
- Removing PVG
- Adding PVs to a PVG
- Removing PVs from a PVG
- Assigning an LVG to a PVG
- Removing an assignment between a LVG and a PVG

4.2.4 The TVC within the Data Path of Tape Backup

The TVC is the data hub of the entire system and each data transfer between the host and the physical drives is handled exclusively via this TVC. The CS VTL acts as a **speed matching buffer** between host and physical tape, for writing as well as for reading. Speed matching supports frontend demands (fast mount, wind, rewind, mount, unmounts, write, recalls, start/stop) and therefrom different backend technology possibilities (performance, capacity, start stop elimination). A detailed look into the way how speed matching with tape volume caching works shows that this is multiple:

- All hosts reads and writes are directed to the TVC with disk speed. New or modified data is always stored first within the TVC. Hosts are only aware of logical virtual tape drives from the frontend ICP as the access point to all LVs.
- Fast mount, positioning, load, demount: Main memory used by the emulation software in the ICP serves for a speed matching effect on the level of single data blocks transferred between host and virtual tape drive.
- Removes common physical tape delays for the hosts because all data access is to TVC disk: If the data requested by the host is already (or still) in the TVC, the data does not have to be retrieved back from the physical tape to the TVC, and there is no need to mount the physical tape cartridge. The data transfer from and to the host can therefore begin immediately.
- LVs require large, physically contiguous and aligned storage blocks, allowing transfers without additional internal copying. The TFS residing on the TVC supports this file structure for LVs and high speed data transfers with physical tape drives at the backend IDP which avoids start-stop operation of physical drives and guarantees high performance streaming mode. Those measures are a prerequisite in order to exploit the high transfer rates offered by tape drives of the latest technology.

Tape mounts by the host:
Tape mount requests by hosts for writing of data can immediately be fulfilled if the LVs to be mounted are empty (so called scratch volume mount) or already stored in the TVC. Subsequently written data are stored within the TVC, and as a standard LVs are almost immediately written by the backend onto physical tape after the REWIND/UNLOAD of the LVs at the frontend.

Tape mount requests by the host for reading of filled volumes require the complete LVs to be stored within the TVC, before reading or writing of data by the host may take place. To optimize restore access time the TVC provides instantaneous access to the last recently used LVs with the according data still kept stored in the TVC after having been written to a cartridge.

LRU Algorithm:
As default the "period of residence" of the LVs in TVC is indefinite and is essentially determined by its capacity and the manner in which the hosts access this data. The algorithm for replacement within the TVC used is basically the “Least Recently Used (LRU)” algorithm. When the TVC is full, the oldest logical volumes are automatically overwritten by new data.

This LRU algorithm is used by default, (when no other Cache Policy Parameter is specified) and the LVs are managed with the LRU procedure. Several LRU stacks are maintained for each TFS. Each LVG is managed by one stack and each of these stacks is responsible for LVs in a particular size range. The LVs are entered accordingly in these LRU stacks as soon as the conditions for a possible release in the TVC are satisfied. On all the stacks a weight is calculated from the size and the age of the topmost volume for the release. The LVs with the highest weight is released.

Cache Policy Parameter:
The TVC management controls the release of LVs from the TVC, as soon as the amount of stored data exceeds the TVC capacity. The algorithm (cache policy) to release LVs from the TVC can be influenced by different parameters. These parameters, like the cache policy itself, can be configured individually per LVG. A simple configuration parameter enables the specification that particular LVs are never displaced from the TVC, as they normally would be in accordance with the LRU algorithm. The same parameter defines whether or not these LVs should automatically be backed up to physical tape.

This configuration is implemented by means of the LVG attribute Cache Policy with the values:

- **STANDARD:**
  LV displacement of this LVG from the TVC according to the LRU principle, automatic backup of the LVs to PV of the assigned PVG. The status of these LVs is then referred to as migrated.
- **TVC_RESIDENCE:**
  LVs of this LVG are not displaced from the TVC but are automatically backed up to PV of the assigned PVG.
- **NO_BACKUP_TO_TAPE:**
  LVs of this LVG are not displaced from the TVC and under no circumstances are they backed up to PV.

Cache Retention:
A “retention period” (protected against displacement) can be configured for the LVs of a STANDARD LVG. LVs in an LVG for which a “retention period” has been configured are kept in TVC for the specified period of time. This LVs remain usable for a particular period of time without accessing the backend (i.e. restore). To limit the data volume actually stored in the TVC, possibly necessary earlier release of LVs due to TVC shortage can be initiated, configuration dependent either automatically or by means of manual administrator’s interaction.

4.2.5 Tape Volume Stacking

The following cases for storing LVs on PVs must be distinguished: The VSN of a PV is defined when the PV is added to a PVG. The other configurable attributes of a PV correspond to the attributes of the assigned PVG. One Type attribute has the value STACKED which means that multiple smaller LVs are stacked onto one physical cartridge. The other type attribute has the value NATIVE for PVs in native tape format.

As default CS VTL writes multiple LVs of different sizes onto one PV (tape cartridge) according to the stacked volume principle where a series of LVs is stored consecutively on the PV together with additional Meta information in a CS VTL internal tape format. In this way, tapes are filled almost to capacity. There will be only a small section of unused tape, since a LV will always be written in full onto a physical tape cartridge. After each write operation a directory is created at the end of the tape. This prevents the entire tape from being scanned when disaster recovery might be required and consequently permits high-speed data access.

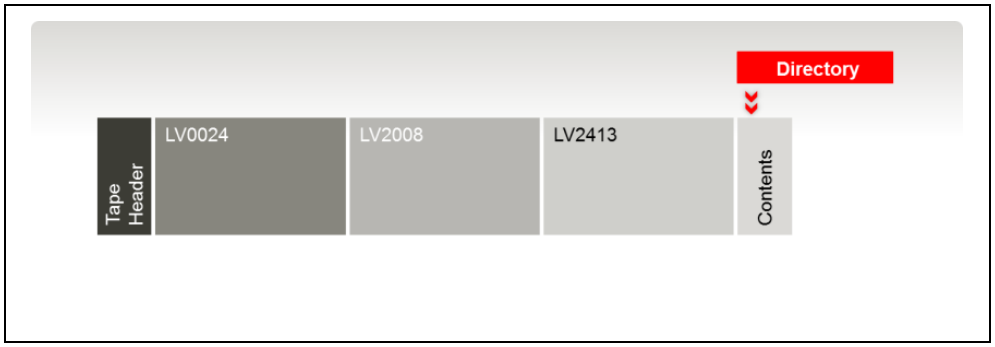


Figure 24

If a LV which has already been saved onto tape is written to tape a second time following an update, the first backup will be declared invalid. The current volume is appended after the last volume of this tape or another tape with sufficient storage space.

Benefits of Tape Volume Stacking:
- Fully utilize state of the art high tape cartridge capacity and tape library capacity
- Radically reduces amount of cartridge requirement
- Reduces footprint requirement for tape storage

4.2.6 Un-stacked Volumes as Native Tape

In some situations it is desirable for hosts to be able to read physical tapes written by a CS VTL directly. This is not possible with the stacked data format, since the stacked format of the data on tape is not the one the host can work with. For this purpose an alternative data format can be configured per PVG, namely the native tape format. A LV written to a PV from such a PVG corresponds exactly to the original format of the LV when it is transferred from the host applications (note that the maximum LVs capacity is 200 GB).

The PV with native tape format will contain exactly one LV and contains no CS VTL Meta information. If a PVG is of the type "native" the VSN of the contained PVs have to be identical with the VSN of the LVs in the linked LVG. Only with identical VSN a correct identification by the host is possible.

There are two scenarios to use native tape:
- PVs with native tape format can also be read by backup applications directly from the physical backend devices.
- Native tape format is mandatory for a PV when the CS VTL system is functioning as a cascading source and the volume is to be saved at a CS VTL cascading destination (cf. chapter 4.4.3).

4.2.7 Data Compression

When writing a PV with the tape drives connected at the backend onto physical tape the IDP can switch the compression function of the tape drives on or off. By analogy to data compression with physical drives the tape device emulations at the ICP frontend are also able to perform data compression. The IPC compression can be set to ON, OFF or HOST for each logical drive. In standard operation the frontend compression is ON.

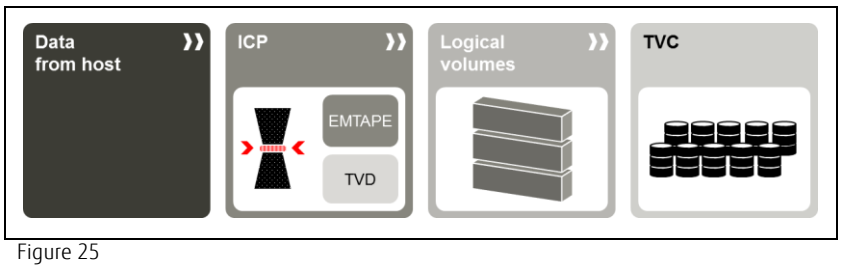


Figure 25

In HOST mode compression is controlled by the host writing tape volumes to the tape emulation (the VLM and PLM do not have any information about this). If the data is already compressed on the host, e.g. if backup data is supplied in compressed format, then ICP compression should be switched OFF.

Compression results in a whole range of advantages:
- TVC utilization is significantly improved depending on the compression level, i.e. without changing the TVC size; it is possible to keep considerably more LVs online in the TVC than without compression (the effective capacity for caching of LVs in the TVC is enlarged), frequently resulting in a very high-performance response time vis-à-vis the host system.
- The amount of data to be transferred between TVC and physical cartridges is reduced by a factor up to 4 (the actual compression factor depends strongly on the nature of the data; e.g. encrypted data do not offer the possibility of effective compression).
- The performance of the overall system is improved due to the fact that the load on the FC network is reduced by the compression factor.

4.2.8 Tape Volume Cache Management

The following describes the logical management of a TVC. The entire TVC must be planned and configured in detail before its usage; the structure options in CS VTL and the RAID system requirements must be observed. The structure options for the TVC configuration are at:
- Sizing the physical level
  - Disk types and capacity
  - Number of RAID systems with according RAID levels
- Logical level
  - Logical Unit Numbers (LUN)
  - Network Shared Disks (NSD)
  - Tape File Systems (TFS)
  - Tape File System Groups (TFG)
  - Logical Volume Groups (LVG)

Example: Logical Partitioning Options
A TVC consists at logical levels from one or more TFS (Tape File System/s) which are combined to one or more TFG (Tape File System Group/s). A TFG is defined in order to be able to configure the distribution of the LVs to the TFS clearly and in a flexible manner.
- A TFG consists of one or more TFSs.
- A TFS can only be assigned to one TFG.

The configuration function TFG and cache policy attributes of a LVG enable the following to be specified for the LVs of an LVG:
- how they are distributed to the TFSs
- whether they are to be stored at a particular location (Location A or B)
- whether they can be displaced from the TVC or are always to remain resident
- whether or not the permanently resident LVs should automatically be saved on tape cartridges

Before one or more TFS are configured, the following questions must be clarified:
- How many LVs are required?
- Which LVs sizes are required?
- How many LVGs are used for the LVs?
- How many LVs should always be completely available in the TVC?
- How many LVs should not be saved to a physical tape?
- Should the TFSs be replicated (Cache Mirroring)? If yes: only for some or all TFSs?

A LV is always assigned to a LVG. Using the TFG attribute, a LVG is always assigned to a particular TFG - either to an explicitly defined and named TFG or to the implicitly defined TFG FLOATING.

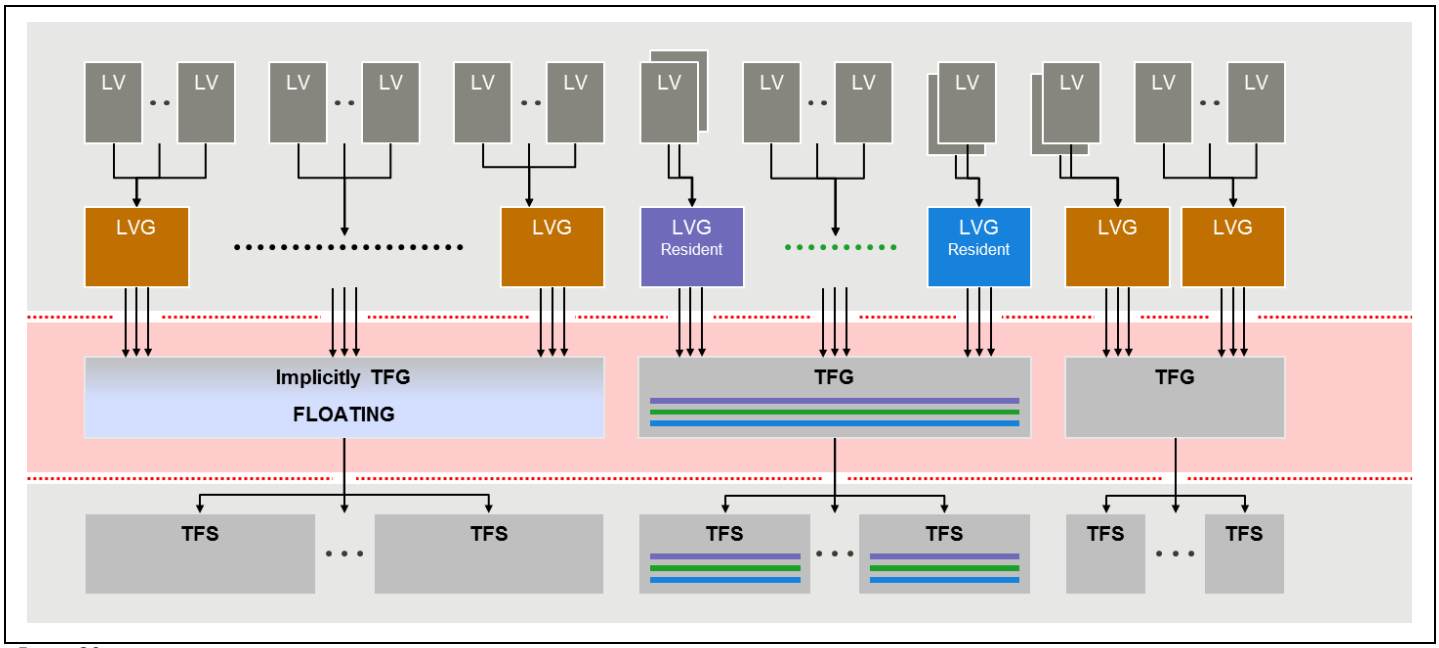


Figure 26

TFGs are configured to determine the distribution of the LVs of an LVG over the TFSs. A TFS can - but need not - be assigned to a TFG. The TFG attribute of a not explicit assigned LVG is represented by the value FLOATING. The TFG FLOATING is always available and collects all TFS that are not explicitly assigned to a named TFG.

The FLOATING value means that the LVs of this LVG are distributed evenly over those TFS which belong either to no TFG or to a TFG without assigned LVG. When an LVG is created, the TFG assignment is defined at the same time (presetting is FLOATING) and thus the storage location of the LV. With the specification FLOATING, the storage location in the TVC is left completely to CS VTL. In this case, CS VTL uses all TFSs that are not explicitly allocated to a TFG for the LVs of the LVG.

For each LVG with the Cache Policy values TVC_RESIDENCE or NO_BACKUP_TO_TAPE a defined storage space is reserved.

The distribution of the LVs on the TFSs is performed by the component VLM. VLM assigns a TFS to each LV according to the LVG and TFG configuration. This assignment is not persistent, but can also change dynamically.

Another possible application of the TFG feature consists of defining the location (e.g. Location A) for the LVs of an LVG. This is done by assigning the LVG to a TFG which only contains TFS for which only NSDs of the required location are used.

The compilation of several TFS to a TFG makes an even capacity utilization possible for these TFS, if they are homogenous regarding the capacity. Within a TFG, the LVs of an LVG are distributed evenly across the TFS.

Automatic TVC load balancing, i.e. moving LVs between TVC parts with the purpose of establishing an equal distribution of LVs mounts (correlating with I/O load at the frontend) among the individual TFS is possible.

The definition of TFGs also provides an additional option of logical TVC partitioning according to different requirements in regard to:
- storage location of dedicated LVs
- reserved storage capacity for dedicated LVs
- permanent or volatile residence of the LVs in the TVC
- not saving on tape for dedicated LVs

Example 1: The entire TVC is divided into four TFGs related to the different requirements. Each TFG can consist of one or more TFS.

The TFG for D2D-LVs (disk to disk) is intended for LVs
- which have to be permanently resident in the TVC
- and which are not saved to tape

The TFG for TVC-resident D2D2T-LVs (disk to disk to tape) is intended for LVs
- which have to be permanently resident in the TVC
- and which are also saved to tape

The TFG for a dedicated LV range is intended for LVs
- which need not be permanently resident in the TVC
- and which are saved to tape
- and which belong to a dedicated LVs range for which decided requirements exist in regard to storage location and storage availability within the TVC

The TFG FLOATING is intended for LVs
- which need not be permanently resident in the TVC
- and which are saved to tape
- and for which no decided requirements exist in regard to storage location and storage availability within the TVC

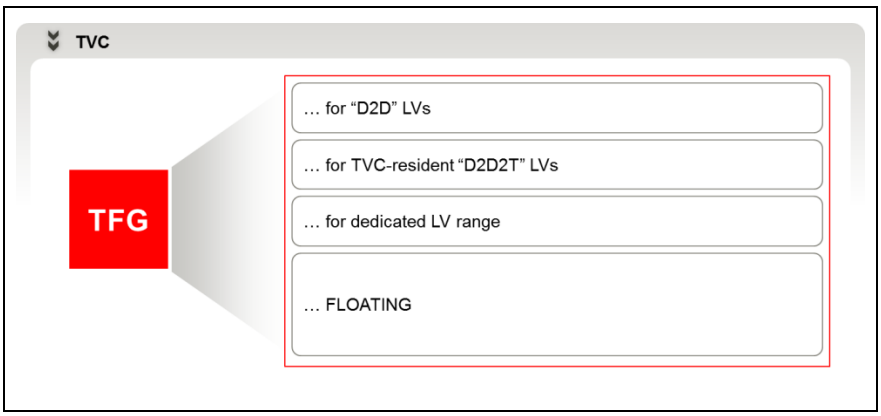


Figure 27

Example 2: TVC partitioning with mirrored and non-mirrored TFGs:

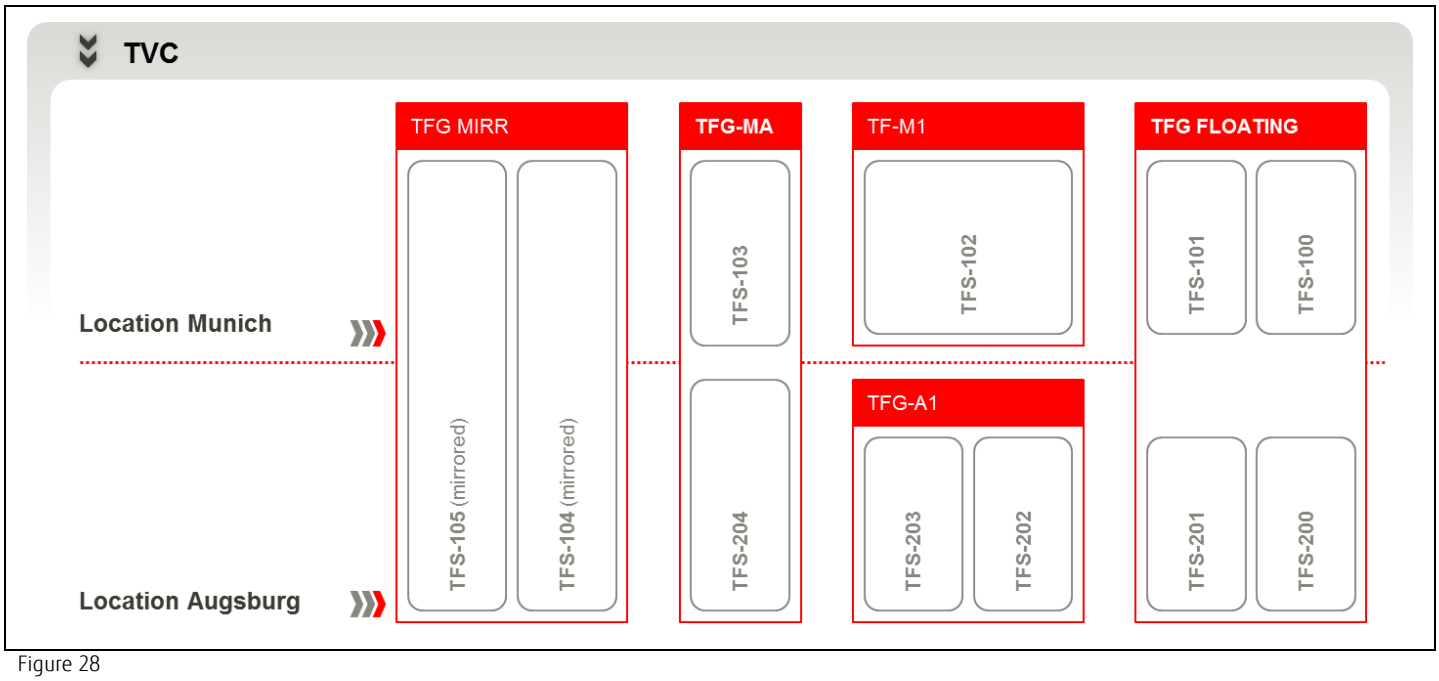


Figure 28

- The entire TVC is divided into 5 TFG. The TFG FLOATING consists of four TFS (100, 101, 200, 201, non-mirrored); 2 are in the location Munich and two in Augsburg
- TFG M1 consists of the TFS 102, location Munich
- TFG A1 consists of the TFS 202 and 203, location Augsburg
- TFG MA consists of the TFS 103 in location Munich and TFS 204 in location Augsburg
- TFG MIRR consists of the TFSs 104 and 105; both are mirrored TFS. The effective cache capacity of a TFS is provided in both locations

4.3 Administration of Physical Tape Storage

Each physical tape cartridge belongs uniquely to a single PVG. A tape library may contain cartridges of different PVGs. Each cartridge has to be named uniquely for all of the attached tape libraries, i.e. there are no two cartridges with the same name. All cartridges of a single PVG have the following identical properties:

- Type of the physical drive to be used for reading and writing.
- Tape library with the physical drives used.

4.3.1 Save and Restore to and from Cartridges

Transferring for save and restore of LVs between the TVC and physical tape is accomplished independently from writing or reading by the hosts. Storing LVs on cartridges is completely under the control of CS VTL. Also the format (stacked/native volumes) on cartridges is specific to CS VTL. In order to read LVs from a cartridge which contains stacked volumes, a CS VTL is always needed. CS VTL operates autonomously several physical drives in parallel, writing or reading.

Transferring data of LVs between TVC and cartridges only takes place, if no concurrent transfer of data concerning the same LVs takes place between TVC and hosts. Thus, writing the LVs at the backend takes place after REWIND/UNLOAD at the frontend. After writing of LVs into the TVC by the host being finished, an internal request for transferring the data onto cartridge is triggered. As default several LVs are stored sequentially on a single cartridge (Stacked Volume).

Reading of the LVs by hosts starts only after the LVs is stored entirely in the TVC. Host mounting of a LV which is not yet stored within the TVC requires to access the cartridge, on which it is stored and to restore the LVs over the TVC from the cartridge. If this cartridge is already mounted restoring of the LVs starts without any delay from physical mounting. Otherwise, the cartridge has to be mounted.

4.3.2 Administration and Secure Handling of Meta Data

Descriptive data concerning LVs (header information, status, size, storage media etc.) is kept permanently available by CS VTL within its volume catalog. For each LV the first bytes (about 256 kB), i.e. including the complete volume label information, is permanently stored within the catalog. This catalog is stored within the TVC, and an additional copy of the catalog is kept within the local file storage of the VLP, thus ensuring additional redundancy.

Creating and releasing of LVs are like bookkeeping operations concerning the volume catalog. Creating LVs does not require accessing a physical cartridge or the tape library. In order to release a LVs only indicators have to be set concerning data within the TVC and the occupied storage space on physical cartridges. These indicators guarantee that the data can no longer be accessed. Also operations as Mount, Positioning and Rewind are implemented as mere bookkeeping operations.

Descriptive data concerning LVs are essential for operation of CS VTL. In order to continue operation after a complete failure of CS VTL they have to be recovered. Disaster recovery, i.e. complete restoration of descriptive data within the volume catalog, is accomplished by reading in all tape directories which are stored after the last written LV at the end of each cartridge. The information of the localization of the directory is stored in the cartridge memory chip.

4.3.3 Tape Reorganization and Tape Refresh

CS VTL takes care of data on cartridges in order to magnetically refresh the media. It also takes care of available cartridge capacity. Data on sparsely filled cartridges - because of gaps due to deleted data – are copied onto new cartridges which are written in a compact way. Time and amount of data concerned by such reorganization processes is determined automatically dependent on threshold values specified by the administrator. During these processes also such cartridges are detected and discarded which are likely to become defective.

With the procedure Tape Reorganization and Tape Refresh the LVs saved on the PV are automatically copied onto other PV of the same physical volume group (PVG). This is done for the following reasons:

- Automatic reclamation eliminates unused space from deleted volumes
- Effective use of the physical tape
- Prevention of physical and chemical aging processes
- Detection of faulty or inaccessible tape cartridges
- Technology migration with changing from one tape technology to another new one

4.3.4 Secure Vaults

One of the key reasons to use tape is for recovery of critical operations in the event of a disaster. The CS VTL with Cache Mirror provides for automatic replication of LVs in the TVC also at different locations, which supports recovery time and recovery point objectives measured in seconds.

Additional the replication of tape data can be obtained by redundant copies of PVs. Redundancy of physical tape data is achieved by storing LVs redundantly on multiple PVs locally, remote or at a vault location.

Data Redundancy based on "Multiple Save":
CS VTL allows for storing LVs on two or three separate cartridges (PVs). These cartridges are written simultaneously by two/three independent physical tape drives, which may also reside in distantly located tape libraries. This feature offers protection against failures or loss of a single physical cartridge, undetected failures of a tape drive during writing, and against local disasters concerning the loss of one of the tape libraries.

The administrator has only to use the Dual/Triple Save option and the redundant copies are produced by CS VTL in one step without any additional load for the host. CS VTL also takes autonomous actions in case of detected failures. A LV stored on two cartridges at one location allows CS VTL to read in case of media error from the local redundant copy. CS VTL automatically replaces the spoilt copy of a concerned volume by a new one. With the "Triple Save", redundancy at one location is a means to compensate locally failures of a cartridge, whereas the third copy at a second different location compensates the case of a site disaster.

Export and Import of physical tape cartridges from the tape library:
In general, concepts relating to the export/import of magnetic tape data are closely linked to the problem of protection against emergencies. With export and import facilities and transferring the data to another site, backup data can be made redundant accessible at different independent CS VTL systems. In case of one CS VTL going down, all the data are available after a time lag at the other system.

As with the function for volume export/import is completely superseded from the control of individual hosts, CS VTL is the facility that offers centralized export and import of data. It is not controlled by the tape management tools on the hosts. CS VTL exporting/importing is designed for various data center scenarios for storing tapes outside a tape library on different purposes:

- **Special storage location:**
  Secure storage in special premises to avoid the consequences of a local emergency e.g. fire-proof basement, strongroom or at a greater physical distance.
- **Manual storage:**
  Export of cartridges that are extremely unlikely to be accessed through daily operations are managed manually e.g. because they are only stored for contingencies of a disaster.
- **External copies for data exchanged between systems:**
  Magnetic tape data is independent of an individual system and can be accessed and used by other CS VTL systems in the case of local emergencies at separate locations.
- **Physical bulk transfer with selective distribution:**
  Reduction in costs as a result of manual transfer of bundled bulk data to possibly extreme distant locations.

Two CS VTL functions are available for export/import:

- The **Vault attribute** of a PVG, which extends the management of LVs to cartridges that are temporarily exported ("VAULT" state).
- The **Transfer PVG**, whose cartridges can be exchanged between different CS VTL systems.

VAULT Attribute:
Cartridges are considered to usually reside permanently within a tape library. However, the so-called "Vault-Attribute" may have been assigned to a PVG. Then the cartridges of this PVG are allowed to be removed out of the tape library.

The link between a LVG and the PVG with the VAULT attribute is established in the usual way. All the LVs of the corresponding LVG are then transferred to cartridges of this PVG, in duplicate in the case of Dual Save. In this respect, this is no different to operation without the VAULT attribute. But with the "VAULT" state for a cartridge to be set, CS VTL prevents all further online accesses. The cartridges are then marked in the catalog of CS VTL as in Vault-state. Though the cartridges and the LVs stored on it are still known in the catalog of CS VTL, no access to the LVs on the cartridges is possible (even if the cartridges are still in the tape library), but a warning is triggered to the administrator, to make the cartridges accessible again, i.e. to reset the Vault- state to accessible.

Only LVs should be simultaneously assigned to a PVG with a VAULT attribute if they are also to be exported together (i.e. the LVs have e.g. the same protection period or the same storage location).

Cartridges in the VAULT state are handled by CS VTL as though they were not accessible, and can be removed from the tape library. Since read accesses are rejected with an error message by CS VTL, the VAULT state can be used to protect a cartridge that must not be used for a specific time. If it is necessary to access the cartridge again, the VAULT state must first be reset again, and obviously the respective tape cartridge must also be available in the tape library.

The user should determine the protection period if necessary, so that a LVs cannot be overwritten while the corresponding cartridge is still in the VAULT state. Even Refresh and Reorganization activities are impossible for cartridges while they are in the VAULT state. In order to enable Refresh or Reorganization of a respective cartridge, the cartridge has to be imported again into CS VTL.

Transfer PVG:
With Transfer PVG the data to be exported (LVs of an LVG) is copied by CS VTL onto additional cartridges (a special additional PVG known as the Transfer PVG). The administrator is in charge to issue commands for triggering the copying of LVs concerned and synchronize with usage. Removal of the generated Transfer PVG cartridges from the tape library takes place via the tape library interface.

CS VTL does not keep a record of which LVs were copied and exported to the Transfer PVG. These details must be held outside of CS VTL using recorded log information. The existing copies of these LVs on other PVs when Multiple Save is used remain unaffected by the transfer: These copies are retained in the physical library unchanged.

The copies on the Transfer PVG can be imported on another CS VTL system. The original LVs remain in the tape library under the control of the first CS VTL. When importing a cartridge, all the LVs that are stored on it are copied onto identically named LV, which must already exist, on the target system. If exported LVs are exported again, the previously exported copies are an earlier version and should no longer be imported; otherwise the versions of the LVs in the target CS VTL system will be overwritten by the imported versions that are out of date.

4.3.5 Encryption of Magnetic Tape Data

The importance of data protection has become increasingly apparent with news reports of security breaches, loss and theft of personal and financial information, and government regulation. Tape data have to be encrypted in many cases in order to guarantee security against loss or stealing during transportation to another location. Encrypting the tape volumes minimizes the risk of unauthorized data access without excessive security management burdens or subsystem performance issues.

Possibilities for seamless integration of encryption solutions with CS VTL in all environments:
- Encryption with tape drive
  - within physical tape library
- Encryption in each Fiber Channel path
  - between host and CS VTL interface
  - between backend and physical tape library
  - encryption of the data path for cache mirror
- Encryption in hosts
  - host writes already encrypted data to CS VTL
- Encryption for tape data on transit
  - accomplished by Tape Pipelining Components

For the encryption between the CS VTL backend (ICP) and physical tape library special devices can be employed in combination with FC tape drives in order achieve strong protection. With CS VTL encryption for example takes place before data is stored onto cartridges by intercepting the transferred data between an IDP and the tape drive through an encryption device. Because only selected data have to be encrypted, these data can be directed to a distinguished IDP, which is armed with an encryption device. CS VTL with an attached encryption device delivers concentrated service for all hosts at a single place, which is economical, since encryption devices are quite expensive. In that way also data delivered by hosts systems connected over mainframe connections to CS VTL can be encrypted by encryption devices connected over FC at the backend.

4.3.6 Backend Support of WORM Tapes

Tape devices with WORM semantic for tape volumes offer features for document archive systems with revision security requirements. Tape based WORM offer the following functionality:

- It is not possible to overwrite data blocks that have been written to a WORM tape.
- It is not possible to overwrite tape marks other than those found between the last data record and the end of the recorded media (EOT).
- It is possible to append data to a WORM tape.
- LTO WORM tapes have a unique unchangeable PV identification which can be used to check if a volume has been replaced.

The WORM functionality is supported on the following types of physical tape device:

- IBM-LTO4; HP-LTO4
- IBM-LTO5; HP-LTO5
- IBM-3592E05
- STK-T9840D
- STK-T10000C

CS VTL use cases for physical WORM tapes:

Stub-save:
During a version migration or migration between different RAID systems, the LVs stubs are saved to tape from the old system and then restored into the new system. The use of a PV type WORM would ensure that this data cannot be overwritten using the migration process.

Transfer in/out:
During the transfer of LVs between two different CS VTL systems it would be possible to make use of WORM PVs to ensure that the data cannot be modified or manipulated in transit. The WORM PV identifier would ensure that the PV cannot be replaced without detection. This requires an alternative secured transfer path for the WORM PV identifier data. It should be noted that the 'WORM' data is not secure as it can still be stolen unless it is also encrypted for transfer.

Vault:
If LVs are to be transferred to a vault location it would be possible to make use of WORM PVs to ensure that the data cannot be modified or manipulated while stored externally from the CS VTL system. The WORM PV identifier would ensure that the PV cannot be replaced without detection. But this requires additional secured storage for the WORM PV identifier data. It should be noted that the 'WORM' data is not per se secure; although it cannot be modified it can still be stolen unless it is also encrypted for vaulting.

Special Archiving (WORM specific LVG):
The administrator defines a set of LVs which are to be used for this special functionality and assigns them to a LVG which is linked to a WORM only PVG. This set of LVs must be known in the application and used for the data that should be archived in this special way.

4.4 Disaster Resiliency with Virtual Tape

The following describes concepts for disaster resilient configurations with CS VTL and point out replication scenarios for LVs and PVs beyond the appliance virtual tape interface. All mechanism for high data availability and disaster resiliency are completely hidden to the host applications and are operated transparently only by CS VTL internal functions.

To protect against any kind of disaster and loss of important enterprise data, a company's disaster scenario plan has to consider following details:

Measures to ensure business continuity
- Protection against data center site outages
- Availability and redundancy of data center environments and important enterprise applications

Protection against loss of enterprise data
- Redundant data protection applications
- Backup and recovery requirements regarding RPO and RTO
- Availability of the backup and recovery environment
- Redundancy of storage devices keeping the backup data

Protection against loss of tape data and measures for fast recovery
- Prevent outages of the virtual tape environment
- Replication of backup data
- Vault location for backup data
- Device allocation of virtual and physical environment
- Ensure high availability of backup data for recovery
- Ensure and take care for long term high capacity data store

The ETERNUS CS8000 architecture supports disaster resiliency with features that guarantee the highest possible availability level for backup data:
- Various configurations to protect against site outages including loss of the entire appliance
- General availability of the appliance with automatic failover
- Protection of the appliance internal configuration and tape data base
- Redundant component configuration
- Redundancy of each data path
- Appliance internal RAID protection
- Redundancy of disk and tape storage
- Tier 2 data replication (synchronous/asynchronous) and recovery of LVs in each scenario
- Tier 3 data replication and recovery of PVs in each scenario
- Backend storage media options (tape, disk, deduplication, long distance vaults)
- Outbound replication between different locations (cascading)
- Support of long distance replication mechanism
- Support for branch offices, centralization and consolidation

4.4.1 Multiple Library Support and Multiple Save

One of the important characteristics of CS VTL is the parallel connection of multiple physical tape libraries of different types. A maximum of ten tape libraries can be operated at the backend (extendable with ACSLS). However, since at least one PVG is required per library, it is only possible to support as many libraries as there are corresponding PVGs. Within a tape library different physical tape drive technologies are also possible. However, a separate PVG must be configured also for each drive type.

With the Multiple Save functionality, based on the volume group functionality, a replication of one LV on two (Dual Save) or three (Triple Save) PVs is possible. This second or third PV can reside in either the same tape library (Multiple Local Save) or in a remote tape library (Multiple Remote Save). If a PV which usually contains a large number of LVs is in some way corrupted (e.g. due to a tape error), the appliance marks this PV as defect, automatically excludes it from the PVG and a new PV is generated. Due to Multiple Save, CS VTL can access the copy of a LV created on a different PV. If the copy is located in a second remote tape library, then even the complete destruction of the first tape library would not cause any irrevocable loss of data.

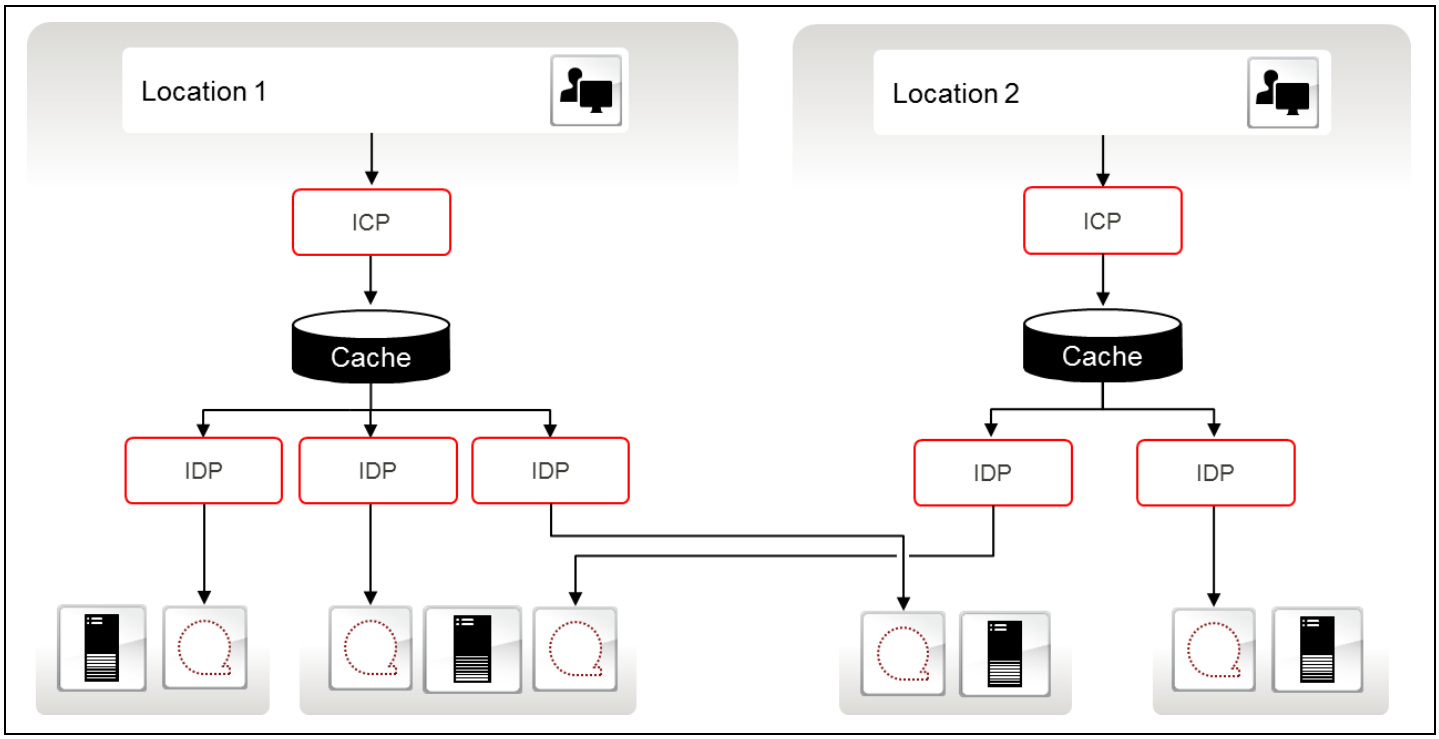


Figure 29

In many data centers the procedure used today is to export the physical volumes which have been written, e.g. the second or third copy, immediately after data backup, for example, so that they can then be stored at a particularly secure location. The Multiple Remote Save functionality combined with exporting the tape cartridges using the Vault attribute permits this procedure to be automated elegantly: not only does it relieve the host application of any copy or move operations, it also eliminates the need to transport the cartridges to a second archive (and back again). The associated risk of data manipulation is thus excluded.

Replication of PVs over long distance FCIP connections with Remote Multiple Save:
In some cases by operating Remote Multiple Save it is desired that a second tape library is located far away from the CS VTL system with the first tape library attached. Therefor the FC data paths between the IDP and to the physical tape drives within the far away tape library have to be extended with FCIP (Fibre Channel over TCP/IP). Network components from different vendors support Tape Pipelining and Tape Acceleration over FCIP in a way that the target tape drive can be operated in streaming mode.

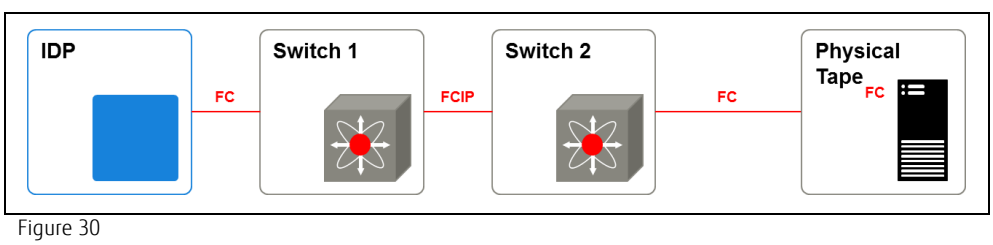


Figure 30

4.4.2 Dual Site Configuration

In a Dual Site configuration (cf. 3.4.3) comprising the Cache Mirroring and Multiple Save feature, a single CS VTL appliance entity is spanned over two sites. Redundant components (*ICP*, *IDP*, *VLP/SVLP and TVC*) are placed at the two sites. A third site operates the TBP. Data delivered from the host are then redundantly stored at the two sites from the beginning.

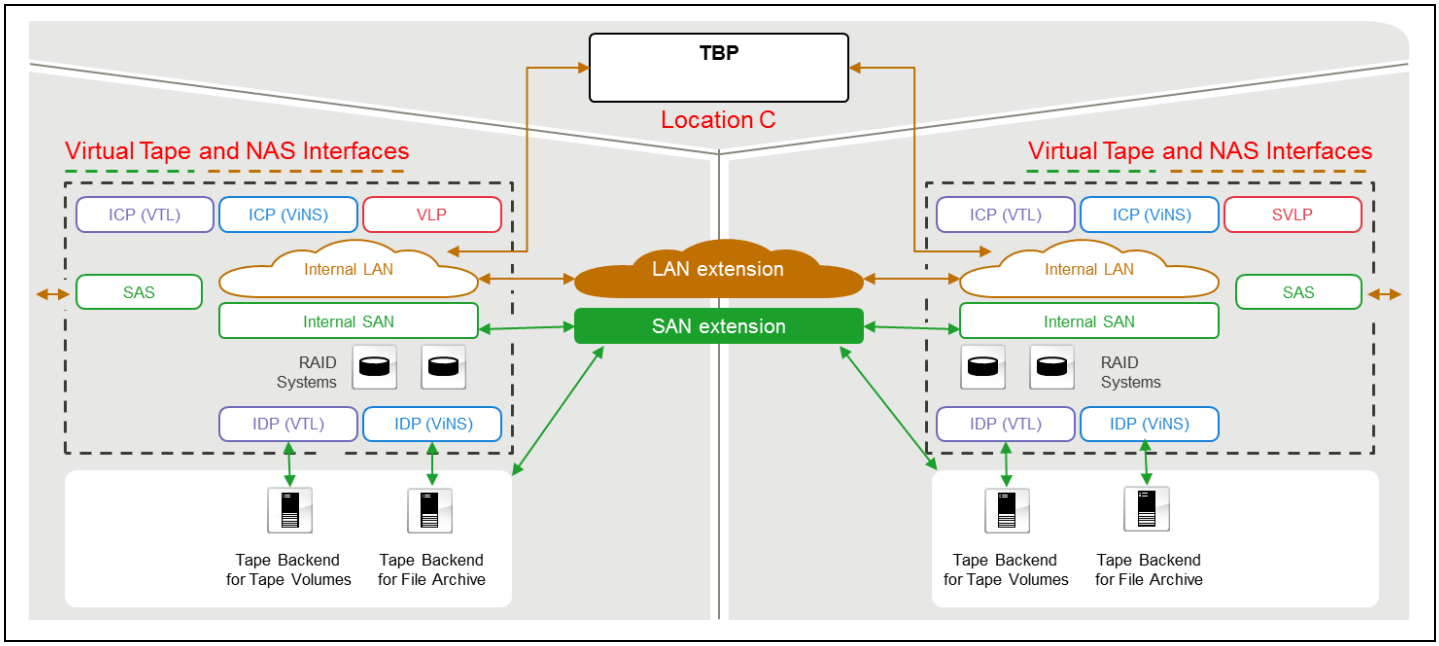


Figure 31

If on one location a RAID system has failed, for example through a local disaster, the remaining CS VTL part is able to continue working (cf. chapter 3.4.2 - 3.4.5). As the status of the data is at all times identical on both sites, a restore is significantly quicker and simpler. No modifications to applications or data backup processes are required. This configuration offers high availability of data over the distantly located sites and fast access in case of immediate restores. The maximum distance between two sites with CMF depends of the network configuration and can be extended by using additional extender components, which may cause, however, limitations of the overall throughput.

Replication of LVs over long distance FCIP connections with CMF: Synchronous long distance LV replication (CMF cf. 3.4.2) enables the transmission of LVs between two parts of a CS VTL system via an FCIP connection. FCIP is a storage networking technology (RFC 3821) based on the Internet protocol which enables FC data to be tunneled via IP networks. The FCIP connections can be established between the internal FC Switches (FCS). For disaster resiliency two redundant connections in the WAN ensure the redundancy of the data path.

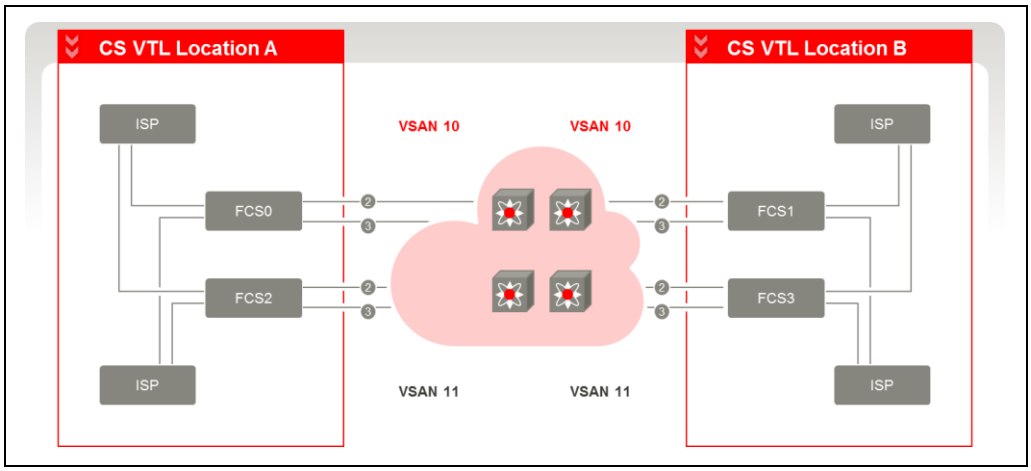


Figure 32

The Dual Site configuration represents the best way to add robustness to backup environments including disaster resilience while still meeting very aggressive RTO targets. However it will reach its limits with growing distance between the two locations. For distances very well beyond 100 km the latencies induced by growing round trip times are becoming too large and the configuration reaches its limits.

4.4.3 Cascaded Configurations

In cases where synchronous cache mirroring is impossible because of the long distance, the cascading concept offers an asynchronous volume replication mechanism over long distances. In contrast to CMF, independent CS VTL systems are involved in cascading.

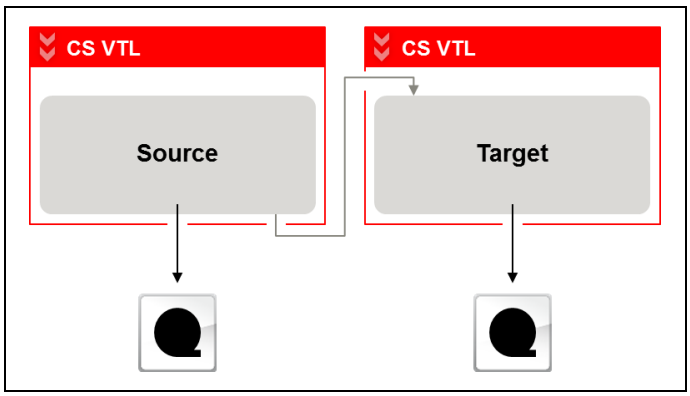


Figure 33

In a cascaded configuration the CS VTL which acts as cascading destination is attached at the backend of another CS VTL which serves as cascading source. With this cascading the backend tape operation of a CS VTL system is the source for an attached frontend tape operation of the target CS VTL. The target CS VTL system is seen by the source system like a physical tape library - with tape drives and library control - and such can be used by the source system as an additional backend target that is able to store physical volume copies within the standard Remote Multiple Save operation.

The benefit of CS VTL cascading over long distances results from a number of logical tape drives on the target systems which are connected to the source system as if they were physical tape drives. But in contrast to physical tape drives the logical tape drive don't need to be operated in streaming mode but may also run at low speed.

4.4.4 Native Tape Format for Cascaded Configurations

The idea of disaster protection with cascading follows the idea that a host or possibly also a second host has always access to the backed-up volumes of the cascading source and can thus continue to operate with the volume copies stored at the cascading destination.

A prerequisite here is that the second host can run the relevant backup applications, knows the volumes, and has the relevant connections to the cascading destination. Another prerequisite is that the CS VTL cascading source generates PVs in native tape format (cf. chapter 4.2.6). Native tape format corresponds to the original format generated through the backup applications from the first host and stored as LVs in the CS VTL cascading source.

If a PVG is of the type "native" the VSN of the including PVs has to be identical with the VSN of the LVs in a linked LVG. The number of possible PVs in native mode is the same like the number of LVs. With identical VSNs, the LVs at the cascading destination are available in such a manner as if the (second) host had backed them up to the cascading destination itself. Only with identical VSNs a correct connection to the host is possible so that data can be restored.

This means that the following VSNs have to be identical:
- VSN in Host \(Cascading Source)
- VSN of LV in LVG \(Cascading Source)
- VSN of PV in Native PVG \(Cascading Source)
- optionally: VSN of PV in Physical Library \(attached to Cascading Source e.g. for Dual Save)
- VSN of LV in LVG \(Cascading Destination)

When larger distances are involved, frequently a smaller bandwidth is available, which results in longer transmission times. To adjust the wait time to the transmission time, a minimal bandwidth can be specified for tapes in native format. The wait times for data- intensive operations (e.g. reading/writing) are derived from the specified bandwidth.

4.4.5 Cascading Connections

With CS VTL two different cascading connections are possible

- Long distance transmission of complete PVs in native tape format (corresponding the LVs) over a FC connection from the cascading source to the cascading destination.
- Long distance replication of LV files over a TCP/IP connection with delta transmission and delta error recovery.

Long distance transmission of native tape PVs over FC

With transmission of PVs in native tape format over FC the CS VTL systems are connected to each other in such a way that the cascading source performs the role of a backup host and the cascading destination the role of a virtual tape library. The cascade at the external interfaces of the cascading source and cascading destination is transparent here. The cascading destination cannot distinguish between a "real" backup hosts and CS VTL cascading sources.

The tunneling protocol FCIP (Fibre Channel over TCP/IP) is used as the connection technology between the cascading source and the cascading destination. The drive emulations of the cascading destination being visible on the cascading source as SCSI drives connected via FC, while the connecting WAN is transparent for the connected CS VTL systems.

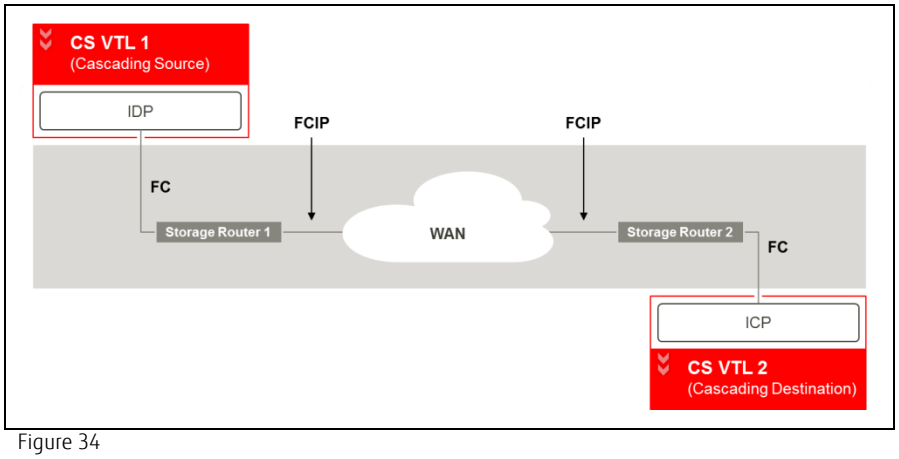


Figure 34

The three core components for cascading - the cascading source, cascading destination and connecting network - are administered and operated separately from each other. CS VTL offers no integrating, global management for this architecture. The FCIP connecting network must be implemented outside CS VTL and is therefore outside the sphere of influence of CS VTL. The quality and performance of the network are reflected in the CS VTL systems as quality and performance features of the FC connections and are handled accordingly.

Long distance replication of LV files over TCP/IP

With this cascading connection it is possible to cascade CS VTL systems using direct TCP/IP connections. No expensive FCIP routers and dedicated FC connections between the locations are needed. The LV is transferred between the source/backend and destination/frontend over existing, low-cost LAN/WAN connections. No tape-based procedure is involved here, but replication/synchronization of LVs as files between the systems. Transmission of files using TCP enables that transmitted data mainly comprises the new data, and if the connection fails, data transmission can be reinitiated and continued. The Meta information contained in the LV file is used to restart the transmission.

Benefits of long distance LV file replication over TCP/IP

- Low cost LAN/WAN connection instead of expensive FC environment
- Reduced network load with delta transmission and delta error recovery
- Compress/decompress for transmission possible but not necessary
- Secure data transfer with SSH authentication and encryption

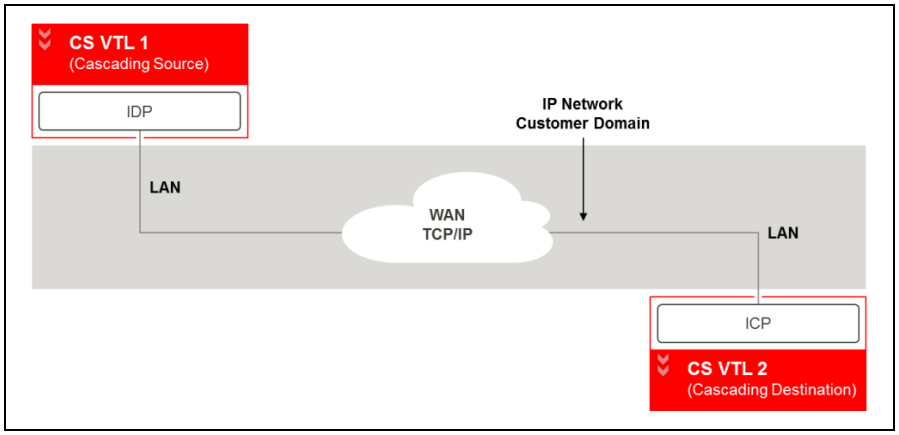


Figure 35

Technology for replication of LV files over TCP/IP

This cascading architecture involves a CS VTL specific LAN-based type of tape drive. It is realized as a form of cascading of two CS VTL systems with a special type of tape drive emulation called TVTD (TCP Virtual Tape Device) on the cascading destination and the corresponding special type of TPDS (TCP Physical Device Service) on the cascading source. The cascading source sees the cascading destination as a special tape library to which LVs are replicated. Communication between the TPDS on the IDP and the TVTD on the ICP takes place using TCP/IP. To permit this, LAN boards with special configured IP addresses must be installed at the cascading source and at the cascading destination. A TPDS always communicates with the same TVTD.

A combined control/data connection is used to exchange information and to transfer the LV data. The write/read operation between the cascading source and destination is mapped to writing to/reading over TCP/IP connections from a TPDS to a TVTD. The TPDS implements all the interfaces which are required by other processes at the cascading source (PLM, PLS). The TVTD implements the required interfaces accordingly at the cascading destination (VLS, VLM). ACSLS is used as the library controller, i.e. STKCSC on the cascading source side and VACS on the cascading destination side. The connection between cascading source and cascading destination can be encrypted. In this case ssh with a dedicated ssh daemon on the cascading destination is used.

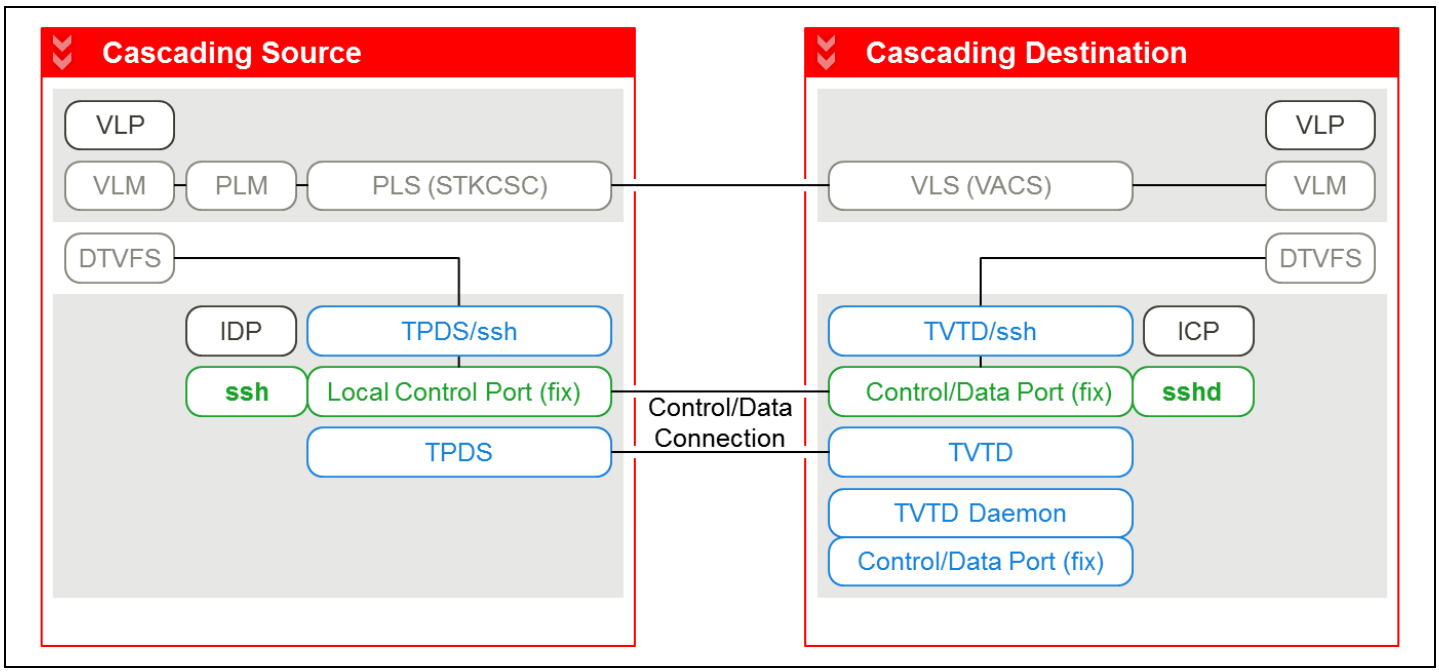


Figure 36

Benefits of replication with LV file copies:

- LVs on the source side will be copied to the target side as data files using TCP/IP based logical drives. Due to the 1:1 copy mechanism that preserves the native tape format on the target side the LVs become immediately available for the backup application in case of requested fast recoveries.
- Seamless integration in the Multiple Save concept maintains multiple copies of the LVs written only once by the backup application at the source side. For example it is possible to maintain one cache resident copy on disk and one on physical tape at the source side for fast restores while another disaster recovery copy is transparently replicated to the far away target side.
- LV files can be compressed for transmission and a delta transmission algorithm is used on top to minimize the amount of network traffic. This delta transmission technique also ensures that a broken I/O process will not require data transmission to be started for a whole LV over and over again but will instead continue with smaller data entities based on the last synchronization point.

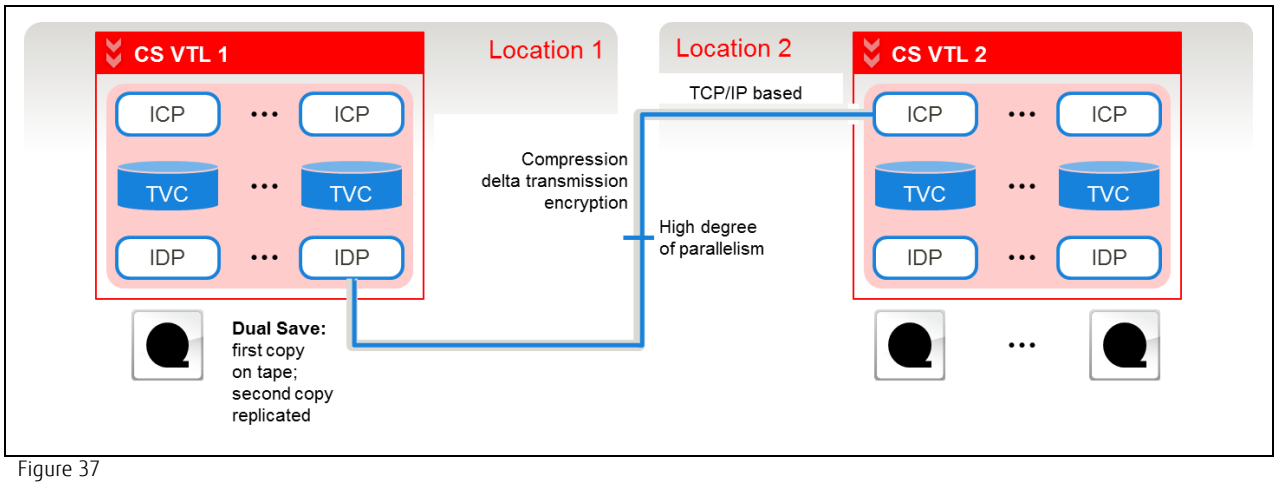


Figure 37

4.4.6 Cascading Usage Scenarios

Long distance volume replication with mono site source system:
A CS VTL with a non-split site configuration operates as source by storing the first volume copy on physical tape close to the source system. Another CS VTL is attached as second target receiving a Dual Save copy of the source volumes.

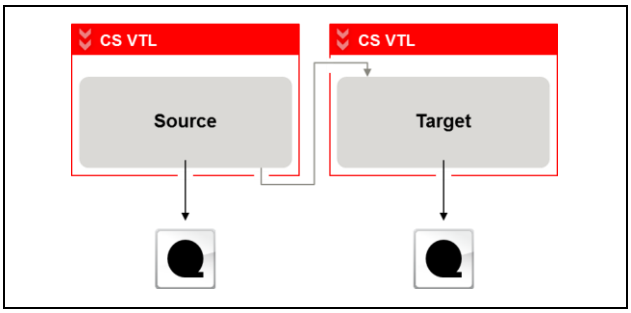


Figure 38

Long distance volume replication with split site source system:
A split site CS VTL configuration operates as source already maintaining two volume copies (like Dual Save). A second CS VTL system at the remote site is connected over a long distance network. This target system can also be a split site system or just a mono site system. Via Triple Save as seen by the source system, the target system receives the third volume copy. The second CS VTL system that can be very far away from the split site locations is suited for disaster recovery in case a catastrophe would affect the two locations covered by the split site configuration.

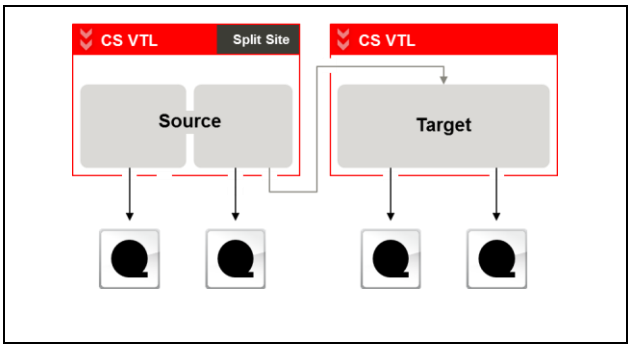


Figure 39

Long distance volume replication and vice versa:

It is also permissible to use cascading in both directions, which means a cascading source can also be operated as cascading destination and vice versa. The first system replicates its volumes to a second one and the second one replicates volumes to the first one

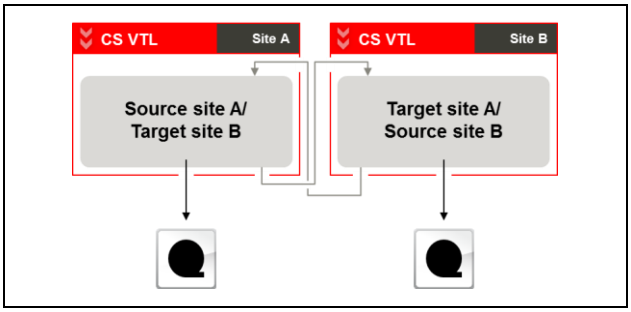


Figure 40

Multiple cascaded systems:

It is also imaginably that multiple CS VTL systems are cascaded behind each other.

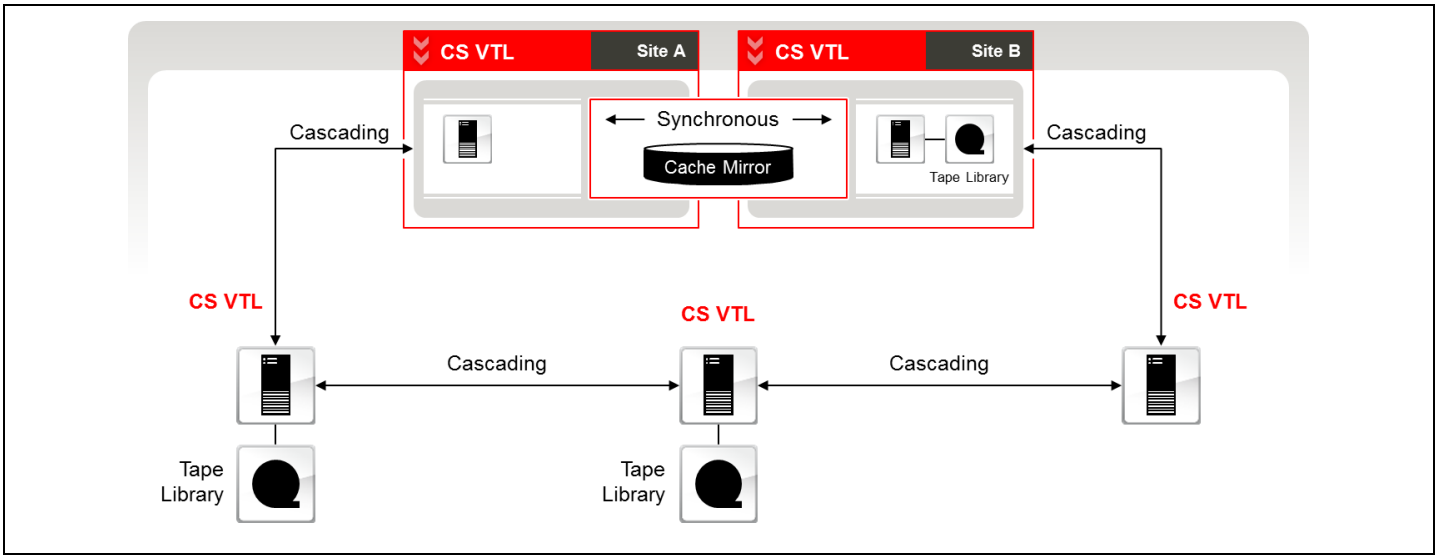


Figure 41

n:1 - branch office concept:

Some companies which operate branch offices with relatively small data volumes need a local backup to physical tape but also a second copy at a data center which is a long distance away. For this case the cascading relationships supported between the cascading source and cascading destination can be n:1. For example, ETERNUS CS8200 is used in the branches with local attached small auto changers and is also connected via WAN to an ETERNUS CS8800 in the central data center.

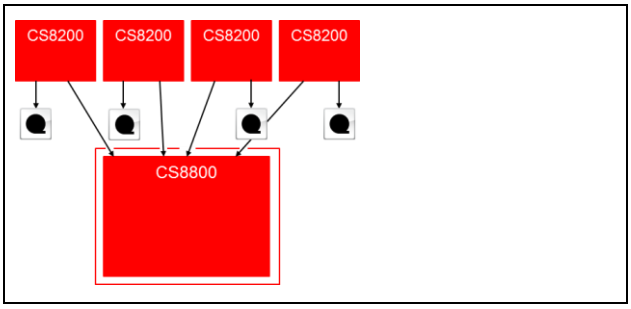


Figure 42

n:m - multiple source and target configuration:

With cascading several CS VTL systems disaster protection concepts can be realized where one source system replicates to multiple target systems (i.e. generates multiple volume copies) and where multiple source systems replicate into one central target system (i.e. branch office concept). Finally also concepts are possible that allow "n" source systems to replicate to "m" target systems.

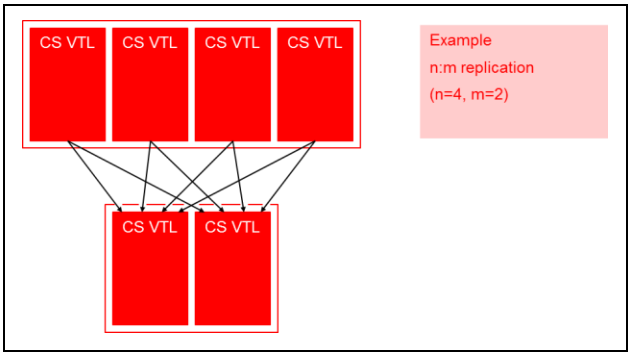


Figure 43

4.5 Calculating System Performance

In each data center environment the performance requirements depend on the overall data center configuration and the kind of backup data. CS VTL offers many possibilities to meet each specific requirement. However customer specific sized configurations of CS VTL systems are not homogeneous. Different configurations of the ICP host adapters with FICON and FC channels, different RAID system disk types (SAS, NL-SAS) and different backend tape drive technologies are possible. For the calculation of an efficient CS VTL system performance various thoughts have to be considered.

Data transfers from hosts are buffered by the TVC, which operates as a speed matching buffer. Thus, there is no need for hosts to comply with streaming mode requirements as it is with physical drives. CS VTL allows arbitrary transfer speeds with the virtual tape drives. Even with small transfer rates an overall huge amount of data can be transferred in total thanks to the high number of available virtual tape drives. But also virtual tape drives with very high transfer rates are available due to the maximum transfer rate determined by the RAID system of the TVC. Writing of data from the TVC onto physical cartridges takes place by one or several continuous streams of data with operating the attached physical tape drives in streaming mode.

The ETERNUS CS8000 cluster architecture avoids bottlenecks and optimizes the use of the TVC. When writing a LV into the TVC the RAID system with the least number of concurrent accesses is selected - out of the plurality of all RAID systems. Moreover, by having different ISP functions like ICP and IDP, peak transfer rates do not depend strongly from migration or clean-up processes. Peak transfer rates and sustained transfer rates are almost equally high. Variations in performance to be experienced in different situations are based on rather minor effects of the shared TVC on the RAID storage systems.

Exploitation of media capacity and guaranteed transfer rates are no complete separate issues. In order to exploit media capacity CS VTL runs clean-up processes internally scheduled which may conflict with processes running in parallel responsible for backup or recovery. Conflicts of this kind are avoided by the distributed cluster architecture. Such internal "housekeeping" processes of CS VTL do not reduce transfer rates from the hosts.

The workload for a single ICP is triggered by the hosts using the virtual drives. The workload for the IDPs is distributed by the VLP by internal orders concerning data transfer of LVs. The IDPs have to be synchronized on accessing physical volumes on cartridges and LVs in the TVC. The IDPs compete on the internal transport channels of the TVC with the ICPs, which transfer data to and from the hosts. The TVC's transfer capacity is exploited concurrently by ICPs and by IDPs.

There are limitations in distributing the workload between components. A given ICP serves a given set of virtual drives. Workload from these virtual drives cannot be accomplished by another ICP. In order to comply with throughput demands, which cannot be processed by this ICP, additional virtual drives served by another ICP have to be added.

Likewise a given IDP serves a defined set of physical drives within a tape library. In order to comply with throughput demands, which cannot be processed by this IDP, additional physical drives have to be added, served by an additional IDP.

There is also gained flexibility. By adding RAID systems of a different kind, different storage qualities become available within a single CS VTL. This can be utilized to comply with different SLAs concerning different LVGs. SAS disks are very fast and support high transfer rates; NL-SAS disks offer large inexpensive capacity. Thus, one partition of the TVC may support high transfer rates, whereas another partition offers large caching capacity.

4.5.1 Avoiding Bottlenecks

Host data is written by the ICP into the TVC, from which the IDP reads it for writing onto physical media. This means, that the TVC has to allow transfer for the double amount of data than received from the host. In case of Dual Save data is read by two IDPs, because data is written simultaneous onto two cartridges. Thus, the amount of transferred data to and from the TVC is triple the amount of data received from the hosts. In case of CMF, data is written into two partitions of the TVC, from which it is read by two independent IDPs. This means that the amount of transferred data to and from the TVC is fourfold the amount of data received from the hosts. The components TVC, IDP (plus physical drives) and ICP have to cope with these load requirements. With these considerations the TVC has to be sized carefully to achieve the overall performance requirements of a CS VTL system.

The example shows the internal flow of data with a single stream, with Dual Save and with CMF:

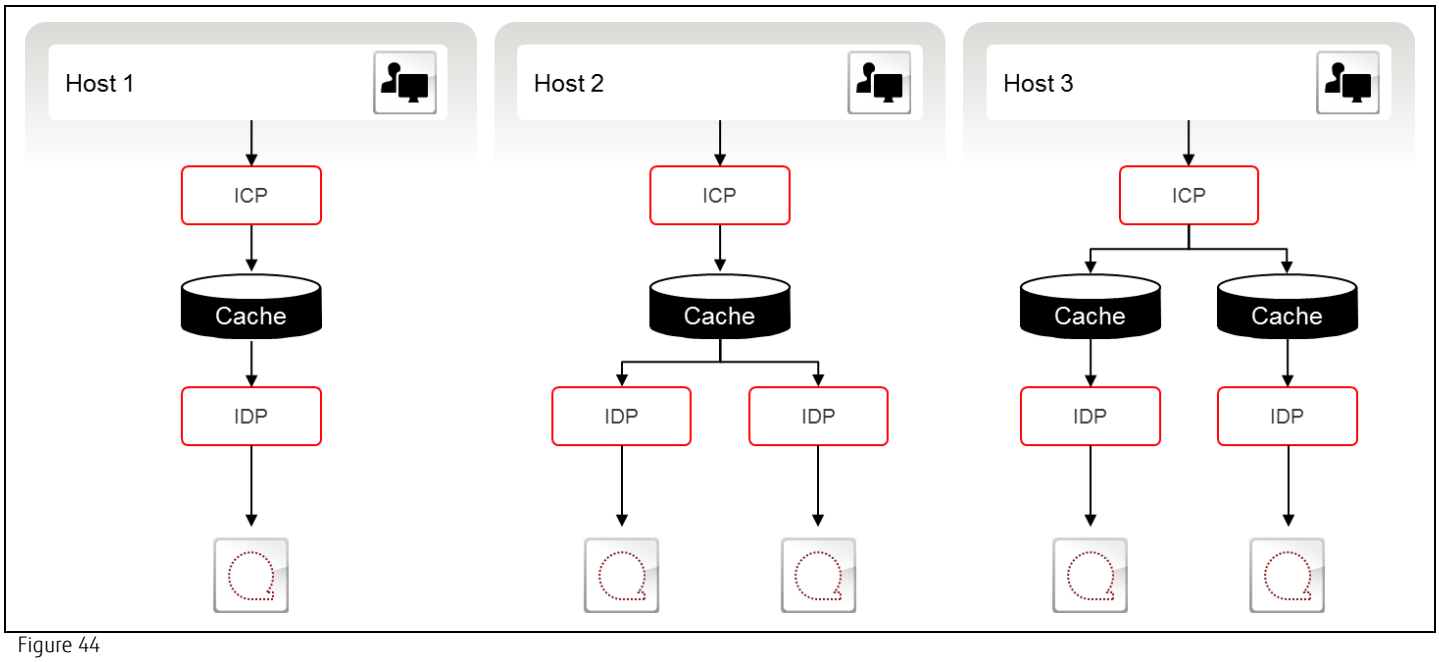


Figure 44

Comparable performance considerations as with Dual Save and CMF apply to Triple Save and to frequent Export/Import processes.

Rules: Dual Save and CMF impact performance in case of writing by a factor of 1.5 and 2. Performance of restore by reading of LVs is not affected, since data are fetched only from one copy. Because reading is less frequent than writing, the number of components needed is mainly determined by the requirements of writing during backup.

The number of physical drives required depends on the overall amount of data to be transferred and should be larger than the number of concurrently used PVG.

4.5.2 Host Data Transfer Rates

Writing and Backup

Writing from hosts may concern empty and non-empty volumes. The majority of writing concerns empty LVs. Because an empty volume is considered as already residing in the TVC, waiting for accessing a physical cartridge including physical mount process is not necessary. If a non-empty LV is affected, this LV has to reside in total within the TVC, and if not present there, it has to be restored from physical cartridge.

As a standard, LVs are stored in the TVC already compressed by the ICP. LVs are written onto a cartridge almost immediately after REWIND/UNLOAD. In the long run the same amount of data has to be migrated as there are data coming in from hosts. There has to be equilibrium, and the transfer capacity of TVC is exploited equally by ICPs writing and by IDPs reading.

Rules: The short term write transfer rate depends mainly on the number of virtual drives and the overall transfer rate of the TVC multiplied by the compression ratio of the data achieved.

The long term, so called sustained write transfer rate depends mainly on the number of concurrent RAID subsystems installed for the TVC and their type. With CS VTL the upper bound of sustained write rates is roughly given by the half (in case of Dual Save: one third) of the overall transfer rate of the TVC multiplied by the compression ratio of the data achieved.

Reading and Recovery

In case of reading there are also situations which are in some analogy to those of writing. Peak transfer rates are achieved if data can be fetched directly out of the TVC, so called "TVC hit", without the need to access cartridges and without concurrent accesses to the TVC. Accesses to LVs which do not reside within the TVC require the access to physical tape storage, so called "TVC miss" or "Read with Restore". In case of long periods of recovery the transfer rate for hosts is bound by the transfer rate of the physical equipment (in streaming mode and concerning compressed data!).

A major impact on transfer rates in case of recovery is waiting for mounting cartridges in the physical library, which depends on the number of physical drives and cartridges to be accessed. The number of physical drives should be at least equal to the number of cartridges to be addressed in parallel. Recovery usually concerns only a single host or a single application. Therefore, deliberations concerning transfer rates are different.

Rules: Speed up for bulk recovery by a single host can be achieved by

- the last backup data being fully kept within the TVC
- the host using virtual tape drives supported by several ICPs
- the cartridges of the last backup are prefetched before the host accesses the data (wait for mounts do not delay reading). That means that the data from the physical tape cartridges are restored to the TVC before or in parallel to the recovery process (parallel processing decreases the time of the recovery significantly)
- the TVC is sufficiently sized

Effect of Compression

Compression of data is a means for

- using the TVC more efficiently, because more LVs can be stored within the TVC
- using the capacity of physical cartridges more efficiently, because less space is needed by each LV
- the throughput of the CS VTL is increasing, because less data have to be moved between the components

The combination of compression by CS VTL and in addition by the physical drives provides a very high compression ratio. As a rule CS VTL achieves a higher compression ratio than host applications. Moreover, the host becomes free from an unnecessary task. Therefore compression by the host becomes obsolete. However, encrypted data or already very highly compressed data like pictures or videos cannot be compressed effectively by CS VTL.

Rule: In order to avoid performance degradation with LVs which cannot be compressed effectively, the host should opt out from compression for those LVs by operating virtual drives without compression.

Size of Transfer Blocks

Larger transfer blocks lead to improved throughput of the ICP and improve also the maximum transfer rates of FICON and FC channels. Standard applications on open systems support choosing large transfer blocks, as 256 KB (i.e. 256*1024 Bytes or 256 KiB). Applications on mainframes are usually restricted to 32 KB or even smaller blocks. A block size of 256 KB improves the throughput by almost 50% compared to 32 KB, independent of compression switched on or off.

Rule: In order to increase performance, large buffers should be chosen as large as possible.

4.6 Tape Processing with CS VTL in the Data Center

4.6.1 Introducing a CS VTL System

A preconfigured CS VTL becomes installed with a customer individual process by Fujitsu Technology Solutions service or another service provider. The introduction and installation of CS VTL causes no serious problems and there is no requirement for complicated adaptation of host software procedures. Prior to starting operation, settings in the host, the application, CS VTL itself and the physical robot archive must be checked and modified if necessary.

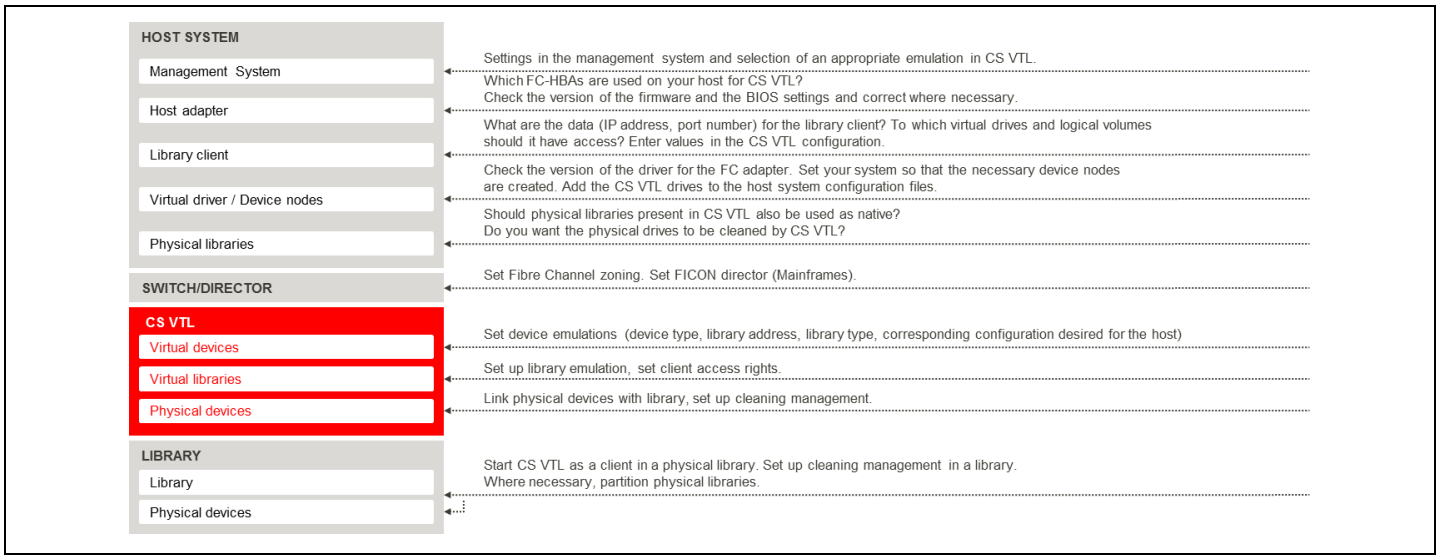


Figure 45

Introducing CS VTL is accomplished in several steps with deriving concepts for:

- Configuration of CS VTL (sizing, components, data and control paths, floor space)
- Virtual tape operation based on the envisaged demands (amount of data, concurrency, etc.)
- Disaster recovery (location of tape libraries, storage vaults, disaster recovery in case of CS VTL destruction)
- Volume group concept with LVG and PVG (according to service level requirements)
- Migration from traditional operation to virtual operation (coexistence with native operation if necessary)
- Test operation without affecting native operation (plus revising the above concepts according to lessons learned)
- Only CS VTL operation or a mix with CS VTL virtualization and direct physical tape operation permanently in parallel. Parallel operation of physical and virtual tape operation is easy, and host operation is not severely interrupted. Only host parameters describing the drive configuration have to take into account for additional available virtual tape drives and virtual libraries. With this dualism, the transformation of data from a physical tape library to CS VTL can be done on demand as soon as the volume is accessed the first time.

4.6.2 Benefits for Mainframes and Open Systems

So far the main benefits from tape virtualization with CS VTL are:

- Brings efficiency to the tape operation environment.
- Reduces batch window.
- Provides high availability and disaster recovery configurations.
- Provides fast access to data through caching on disk.
- Provides utilization of advanced tape drive, tape media, and tape automation technology.
- Automated tape data migration to new tape technology without influences at the frontend
- Provides the capability of filling high capacity media 100%.
- Provides a large number of tape drives or concurrent use.
- Provides data consolidation, protection, and sharing.
- Requires no additional software.
- Reduces Total Cost of Ownership (TCO)

Regarding to mainframes and open systems environments more specific benefits can be identified:

Mainframes
Applications on mainframes are functionally mature, however limited in the use of advanced tape technology with respect to block size, high capacity cartridges, and transfer rates.

Rules of thumb: LVs used by mainframes should preferably be chosen of small size, dependent on the backup software in operation. The increase in the overall number of volumes in the host catalogs is no problem.

Main benefits for mainframe operation	
Without Virtualization	**With CS VTL**
Cartridges not full, many slots needed in the physical tape library	Cartridges full, few slots needed in the physical tape library
Small backup window requires many physical tape drives	Fewer physical tape drives needed, since data transfer with virtual tape drives is exploited optimally
Start/Stop operation of physical tape drives	Physical tape drives are operated in streaming mode
Expensive high-end devices, which have to tolerate Start/Stop operation	Cost-effective LTO tape drives, which do not need to operate with Start/Stop
Allocation and re-allocation of drives requires applications to synchronize, i.e. not to block one another	With many available virtual tape drives concurrent data streams co-exist without mutual blocking

Open Systems
Applications on open systems with backup and storage functionality exploit high capacity cartridges and optimize transfer rates; however they have to cope with operational complexity, high administration efforts and performance loss in case of recovery. Read access can become even blocked for periods in case of data which reside on a cartridge to be written until the capacity of the cartridge is exploited. Organizational requirements concerning the usage of separate volumes for certain backup runs, in order to avoid possible overwrites by a follow-up backup run, restrict the exploitation of high capacity cartridges. Non-technical or product-based restrictions on the amount of data on specific volumes (e.g. isolated, extra volumes for backing up Meta data, which are of small amount) have a similar effect.

Rules of thumb: Backup software for open systems is sometimes restricted in the maximum number of volumes to be managed within their catalogs. Thus, the size of LVs dedicated to open systems has to be larger than with mainframes in order to keep the overall number low. But care should be taken, to choose the maximum size of LVs as small as possible (e.g. 5-10 GB is a pragmatic number), because large LVs restrict the degree of freedom provided by virtualization, in that:

- The complete LV needs to be copied into the TVC before reading or writing; this is time consuming and consumes space within the TVC. Unused capacity increases.
- Recovery will be faster (complete LV has to be read into the TVC)

Main benefits for open systems operation	
Without Virtualization	**With CS VTL**
Temporarily unused physical tape drives because of single host connections	Virtual tape drives are available in large numbers and sharing of physical resources is automatically done by CS VTL
Decentralized administration and operation for peripherals of different types	Central administration with consolidated physical environment due to emulation of various device types
Multiple device drivers have to be adapted within each host and/or client	Device drivers have to be adapted only within the virtual tape system
Multiplexing of data streams in case of writing via intermediate servers slows down read and recovery runs by a factor of 2 to 10	Concurrent data transfers to virtual tape drives have not to be serialized by the hosts , thus read and recovery runs are not slowed down
Conflicts in accessing different data which are filed on the same cartridge, i.e. mutual exclusion of backup and recovery runs on the same cartridge	Small LVs allow for arbitrary degrees of concurrency of applications, without limitations resulting from storing data on shared media
Some cartridges are only partially used because of non-technical, e.g. product-specific requirements for data separation	LVs are stored according to internal optimization strategies independent of external requirements for data separation

4.6.3 Availability of the Appliance and Tape Processing

Failures of physical tape devices are the most prominent problem in traditional backup procedures. Backup software is often blocked by drive failures and unable to recover automatically and is not able to switch between tape drives. Thus manual interactions by human operators are required and backups cannot run unattended. Because CS VTL isolates physical drives from the hosts, also failures at the backend are completely invisible for the host. CS VTL is able to recover from transient device errors, and to overcome permanent failures of a device by switching to any working device not in failure. If the TVC has adequate capacity even the failure of a complete physical tape library can be bridged for a quite long period.

An error message returned to the host during operation with a virtual tape drive looks for the host like errors with physical tape operation and allows him to avoid data losses. In case of the failure of a virtual tape drive caused by the failure of a single ICP, the host can set back and repeat the current work with a new requested virtual tape drive from another ICP.

CS VTL offers stable virtual tape interfaces to the hosts. Thus, no host software updates are required in case of technological upgrades at the physical backend of CS VTL.

Benefits for the host in case of failures	
Features of CS VTL	**Benefits for the Host**
VLP failover and mutual monitoring	Appliance resistance against disasters
Redundancy of CS VTL hardware components	Component failure causes only reduced performance
Internal error control, prevention and correction	Operation without interruption
Automation and centralization of administrative tasks	Eliminates human errors and downtimes
Automatic switch in case of physical device failures	Protection against failure of physical tape drives
Multiple Save of PVs	Protection against failure of tape media
Grouping and storing LVs according service levels via the volume group concept	Defined recovery time for each class of data
Stable host interfaces with virtual tape drives and virtual tape library interface	Due to full virtualization no host interruptions by technology upgrades of the physical devices

Measures to ensure the availability of tape data	
Features of CS VTL	**Benefits to Availability**
Data received from host are replicated by cache mirroring onto two independent RAID systems	No loss of data even of LVs not yet saved from TVC onto tapes in case of a complete RAID system failure
Data received from host are replicated by cascading of two independent CS VTL systems	LVs are available over a faraway emergency system in case of a complete CS VTL break down
Multiple Save as an attribute of LVGs, and reading from copy of PVs in case of failures	Failures of cartridges are automatically compensated by reading and restoring redundancy
Multiple Save on Tape Libraries on separated sites	Site failures do not impair data availability
Vault attribute of Volume Groups: Cartridges can be removed out of a physical tape library keep marked within the catalogs of CS VTL.	Removal of physical volumes secure against disaster or malevolent actions, by storing them at secure vaults
Transfer-PVG with specific properties: Removed copies are no longer in the catalog of CS VTL	Cartridges can be removed for being stored manually and can be re-installed in any other CS VTL system

Measures to increase availability of physical tape processing	
Features of CS VTL	**Benefits to Availability**
Operation of tape drives in streaming mode; avoided Start/Stop operation	Stress-free usage of tape drives and magnetic tape media
Automated refresh and reorganization of cartridges	Prevention against failures of media and avoidance of human errors
Internal catalog data are kept redundantly distributed over several computing components, i.e. failure resistant	Decoupling of internal computing components from the failure of a single RAID subsystem
Copy of the volume catalog at the end of each cartridge	Recovery in case of a total system crash
SNMP extensions for integration into network management centers	Inclusion into a user-specific emergency concepts

4.6.4 Operational Advantages

Control by the Host:
The main functions of operation and TVC usage are implicit controlled by the emulation software for LVs. The mechanisms of load balancing, leveraging the variations of the load by buffering data within the TVC, are effective without any interactions of the user.

CS VTL can compress the LV through the administrator either setting compression "on" or "off" or "host". In host mode the host may choose via the emulation interface of a virtual drive whether CS VTL should compress or not the data of a given LV. Host mode has the same behavior like using the traditional interfaces of a physical drive.

Variation in Traffic Demands:
During operation the load on the equipment changes typically over time. There are traffic gaps on single connections or hosts. They may be compensated by higher traffic from other hosts. Many data streams which may only require low transfer rates on the host side are automatically mapped onto fewer data streams of high transfer rates at the backend, exploiting the streaming mode of physical drives.

Options for the Administrator:
The administrator can control the storing of LVs on cartridges and tape library systems. A concept of volume groups allows him to specify which kind of LVs should be stored on which group of cartridges residing in which physical tape library. The physical locations and the number of copies can be specified.

Monitoring:
As a rule, no administrative interactions are required during operation. But there are extensive monitoring facilities allowing for the visualization of various kinds of operational aspects in an intuitive graphical representation. At any time the administrator can check the current status of archiving and recovery processes. The overall status window allows the user to zoom-in, which also works for single components. History functions support to recognize trends.

Administration of Media:
CS VTL takes care of available cartridge capacity, tape reorganization and refresh for physical tape media or recovering data of defective cartridges and removing them. Refreshing media, typically to be done manually by administrators and involving host operation, is automatically included thereby, in that migration dates of LVs are additionally taken into consideration. The administrator becomes informed if during operation the danger of a bottleneck (shortage of physical cartridges, TVC filling up, etc.) is detected.

Reorganizing Storage by Backup Products:
Backup products usually come with mechanisms for storage reorganization of various types, which operate independently of the usage of a virtual or a physical tape library. Reorganizing the data within a LV requires the volume to be transferred into the TVC, modified and migrated back onto physical cartridges. Therefore, some manufacturers give some guidance for avoiding unnecessary traffic with virtual tape systems. Reorganization of cartridges by CS VTL does not be in direct conflict with storage reorganization within single LVs. By choosing small LVs and inserting at most a single "save set" the demand and amount for volume-internal reorganization can even be reduced substantially.

Scratch Volumes:
The interfaces of physical tape libraries as ACSLS or HACC, offer the functionality of releasing a volume ("RECYCLE"), i.e. to allocate the volume to a "scratch pool" of volumes. Before writing new data onto the volume the released data are still recoverable. By issuing an "Unscratch" command for the volume, the data become accessible again. This is used traditionally with mainframes, and therefore has to be possible also with LVs. CS VTL supports the option of an "Unscratch" only within a given time period after releasing the LV. Therefore, there is a time period of grace, which can be specified, within which the released data are still secured. Thereby, procedures may remain unchanged which run a sequence of "Scratch" commands, and resort after some time to an "Unscratch" for some volumes. With that CS VTL allows for compatibility, but also ensures to get rid of stale data after this period.

With mainframes usually volumes are released periodically and transformed into scratch volumes. This affects as well bookkeeping data in the host as those in CS VTL. But, between bookkeeping by hosts and by CS VTL may raise a synchronization gap of some seconds. To overcome situations of conflicts CS VTL offers several solutions specific to the virtual tape library interface and the requirements of the IT environment.

4.7 Data Deduplication with CS VTL

Data deduplication is a technology that is used to reduce the amount of space that is required to store data on disk and is transparent to users and applications. Data deduplication eliminates the need to store multiple copies of the same data over time. As data is written to the appliance in the native format of the backup application, it is deduplicated. Backup applications are not requiring any additional software installation.

The space reduction of deduplication is achieved by storing only a single instance of data that is backed up repetitively. The effectiveness of data deduplication depends on many variables, including the rate of data change, the number of backups, the type of data, and the data retention period. The benefit is determined by data and backup policies. Data deduplication can provide great data reduction but it does not always make sense to use data deduplication because not all types of data can be deduplicated with identical efficiency. Data deduplication might interfere with other technologies, such as compression, encryption, or data security requirements.

Since Version 6 the ETERNUS CS8000 VTL subsystem leverages data deduplication technology for virtual tape to dramatically increase the role of the internal RAID system for the protection of critical data. With the CS VTL deduplication technology users can retain much more backup data on the fast RAID system than with the conventional Tape Volume Cache (TVC) configuration. CS VTL combines leading deduplication technology with an optimal combination of high scalability, total system performance, and easy integration in heterogeneous environments. The CS VTL tape emulations in conjunction with deduplication support all major Mainframes and Open Systems backup software solutions. By deduplicating data and then compressing it, CS VTL makes optimal use of disk capacity, further reducing associated disk costs.

4.7.1 Deduplication Technology

With CS VTL data deduplication is processed "in-line". CS VTL does not wait for the backup job to be completed before it begins the deduplication process. The incoming data stream is read and analyzed by the data deduplication engine on the fly. An advantage of inline data deduplication is that the data is only processed once and there is no other processing later. Inline data deduplication requires less disk storage because the native data is not stored before data deduplication.

The procedure is breaking up the data into elements and creates a signature or identifier for each data element. The data element signature values are compared to identify duplicate data. After the duplicate data is identified, one instance of each element is stored, pointers are created for the duplicate items, and the duplicate items are not stored but just referenced.

CS VTL deduplication technology uses a sub-file, variable-length approach to identify redundant blocks in a data stream - blocks that have appeared before in the same dataset or in datasets processed at an earlier time. When a block appears that has already been stored, the CS VTL inserts a reference pointer to the earlier instance of the data segment instead of storing another copy. The result is a reduction in the storage capacity needed to store the data set. Block-level deduplication is not dependent on the type of data, it is generally data-agnostic in the sense that it will compare blocks and such it is more granular.

CS VTL uses compression technology after duplicated blocks have been identified and replaced as part of the deduplication process. With compression, unique data that has been through the data deduplication process can be compressed at a typical ratio of approximately 2:1.

Advantages of using Deduplicated Logical Volumes: Deduplicated logical volumes are kept in the deduplication data pool on disk and may be accessed randomly, without the need to read the contents from backend storage. Handling is similar to traditional disk resident volumes that don't get migrated to backend storage. This may significantly reduce volume mount times and data seek operations. Another main advantage is that disk space is saved because redundancy is eliminated from the ingest data streams.

4.7.2 Deduplication Process

When processing an ingest data stream with deduplication (while data is written to a volume) redundancy is eliminated from the data and only unique parts are stored to disk. However, enough additional information is stored about the data stream so it can be restored to its original state.

The repository on disk holding the deduplicated data - a pool of all unique data blocks and associated metadata - is called a BLOCKPOOL. When subsequent backup jobs occur, the data deduplication engine searches for new data entering the CS VTL and uses the variable length compression algorithm to compare new data to existing data in the BLOCKPOOL. Unique blocks are added to the BLOCKPOOL and known blocks are indexed.

Processing when data is stored: The data stream is broken into segments; the size of a segment is in range of several tenths of megabytes. Each such segment is processed for deduplication and stored as an object we call a BLOB. Each such BLOB has a unique identifier – a BLOB tag.

The following outlines the steps of deduplication processing:

- Chunking: The segment data is broken into variable length chunks (blocklets) using special algorithms that try to detect boundaries adequate for deduplication. Blocklet are rather small, their size is in the range of tenths of kilobytes.
- Blocklet hashing: For each blocklet a hash value is built using state-of-the-art hash algorithms. The resulting hash value is considered a distinct identifier for the blocklet.
- Blocklet hash value lookup: The hash value of a blocklet is looked up in the global index of data blocklets already stored. The blocklet is considered unique if the hash value is not found, else it is redundant.
- Blocklet store (with optional compression): If the blocklet is unique then it is stored with a reference count of 1. The blocklet data may be compressed with conventional algorithms in order to save disk space. The blocklet hash is stored in the global index. If the blocklet is redundant the reference count of the blocklet with the same hash value is incremented by one. Information about the blocklets that are part of a BLOB is stored in the per BLOB index so it can be reconstructed correctly when data is retrieved.
- Processing when data is retrieved: Based on the BLOB index that was created at store time the blocklets of the data stream are located in the data pool and delivered in the correct sequence.

The sum of all data stored with the deduplication engine makes up the BLOCKPOOL. The more unique blocklets are stored in the pool, the higher is the chance to find redundant blocklets when new data is stored.

4.7.3 Deduplication System Architecture

The deduplication (dedup) feature is implemented in a CS VTL system as a set of one or more deduplication services. Each deduplication service is associated with a set of resources that are distributed among the components of the CS VTL system. The traditional components of a CS VTL system (library services, virtual tape drives) are not part of the deduplication services themselves but are aware of them and may get linked to a deduplication service when needed.

The following picture gives an overview of the components of a dedup service and outlines how traditional components are involved.

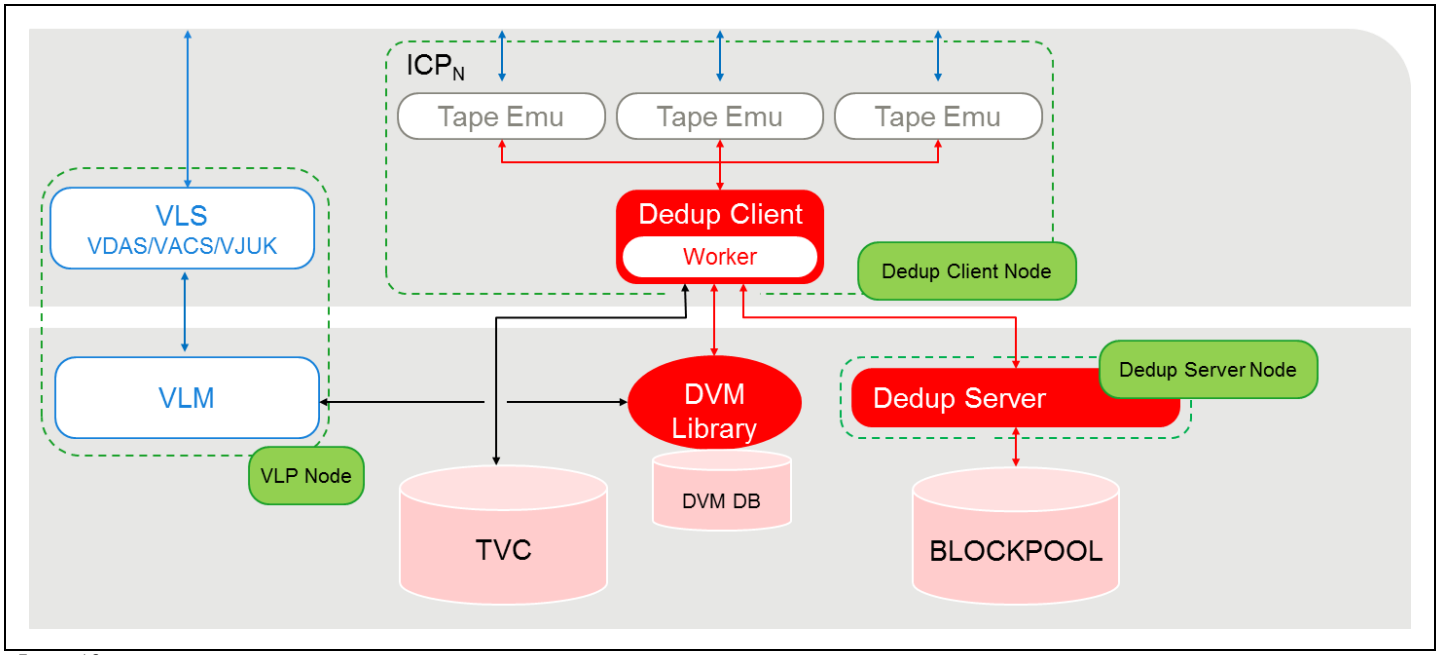


Figure 46

Traditional virtual library components:

- **Virtual Library Services (VLS: VDAS, VACS, VJUK, ...):** These components are used by hosts to mount and unmount/unload logical volumes to/from logical drives. "Dedup LVs" are handled in the same way as traditional volumes.
- **Virtual Library Manager (VLM):** The VLM maintains a database of all LVs defined in the system. It knows about membership in volume groups and location of logical drives on ICPs. In case of deduplication it keeps also track and controls the deduplication processing of LVs.
- **Logical drives:** Logical drives with specific tape emulations become aware of Dedup LVs whenever such a volume is mounted. They then connect to the dedup client on the ICP they are running on and handle all volume I/O through that client.

Resources on RAID:

- **Tape Volume Cache (TVC):** The TVC traditionally contains LVs that are mounted. For Dedup LVs only a stub file is maintained in the TVC. The volume data is never stored there; it is directly written to or read from the dedup data pool. Since Dedup LVs are represented in the TVC with a stub file, a small amount of space is required in the TVC.
- **Dedup Service Home File System (Home Directory):** A dedup service is assigned a home directory located on a dedicated CAFS file system. All persistent data maintained by the dedup service is kept below the home directory:
  - The BLOCKPOOL of the dedup service: this is the resource consuming almost all of the required disk space.
  - The Dedup Volume Management (DVM) database

Specific components for the Deduplication Service:

Dedup Client: Data is deduplicated at the frontend following the tape emulations running on ICPs. A dedup client operates on each ICP which performs deduplication processing and keeps track of all dedup volumes mounted on logical drives on its ICP. The I/O processing is then performed by worker processes assigned to volumes at mount time. The worker processes implement the client side data deduplication (chunking, hashing, compression, chunk data transfer) and the protocol with the dedup server. Data processing is done inline through shared memory buffers (in memory, without temporary storage on disk) and is synchronous in the context of the tape emulations. The dedup client is configured to store unique LV blocks to the BLOCKPOOL over a dedup server. The communication with the dedup server uses a dedicated internal storage LAN.

Dedup Server: The dedup server is a multithreaded process that handles parallel connections from several dedup client worker processes. The dedup server is running exclusively on a dedicated IDP processor node (not used for tape backend processing) and transfers metadata and chunk data in/from the related BLOCKPOOL.

BLOCKPOOL: The dedup data pool (BLOCKPOOL) is the repository on disk where the unique data chunks are stored together with the metadata needed to keep track of stored objects. The BLOCKPOOL uses a dedicated file system type BPFS (BLOCKPOOL-FS) adapted to the I/O profile of the dedup service. The BPFS is based on the Shared Cluster File System (CAFS Cluster) which is provided by the appliance file system architecture. It does not interfere with the logical volumes (and stubs) stored in the TVC but inherits all advantages of the CAFS file system (e.g. the option of mirroring).

Dedup Volume Management (DVM), DVM Database (DVM DB): Dedup volumes are broken into smaller pieces we call segments. Each such segment is stored as a separate BLOB in the BLOCKPOOL. The metadata and mapping information for LV elements is stored in the dedup data pool DVM DB (Dedup Volume Management Data Base). Access to the database takes place through the DVM library. As representation of LVs stored in a dedup pool metadata and administrative data is stored resident in the 'normal' TVC as well.

Internal Storage LAN: Deduplication within the appliance produces the exchange of huge amount of data between all components - such as dedup clients and servers. Therefor a dedicated independent internal storage LAN (SLAN) is used in parallel to the standard internal LAN. There is no connectivity between the two internal networks. The components which perform deduplication processing are using distinct interfaces in parallel.

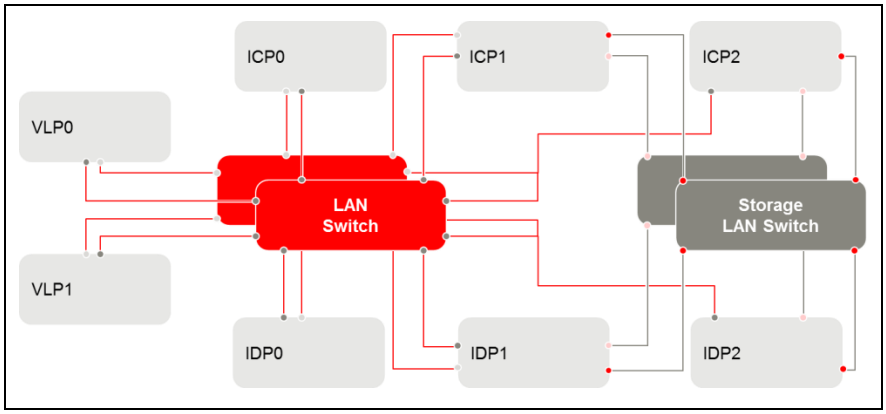


Figure 47

4.7.4 Deduplication and Storage of Logical Volumes

Client/Server deduplication processing: An important part of the deduplication processing is the handling of LVs. The following outlines the single steps:

1. Client side: A single LV is split into data chunks. The dedup client parses the data stream to be deduplicated into chunks (blocklets).
2. Client side: Compute hash values for chunks.
3. Client-server communication: Check for uniqueness of chunks (using the hash values).
4. Client: Compress unique chunks before transfer.
5. Client-server communication: Data transfer. Unique chunks are transferred (compressed), for non-unique chunks reference count increments signaled.
6. Server: Store unique chunks in the BLOCKPOOL; increment the reference counts for non-unique chunks.

Storage of Dedup LVs in the BLOCKPOOL:
All Dedup LVs will be stored in deduplicated form in the following way:

- A Dedup-LVG is assigned a TFG/TFS (part of the TVC) and a dedup service.
- Dedup-LV data is directly written or read in the context of the tape emulation (on the ICP) using the dedup service.
- Only the Dedup-LV metadata is stored in the TVC

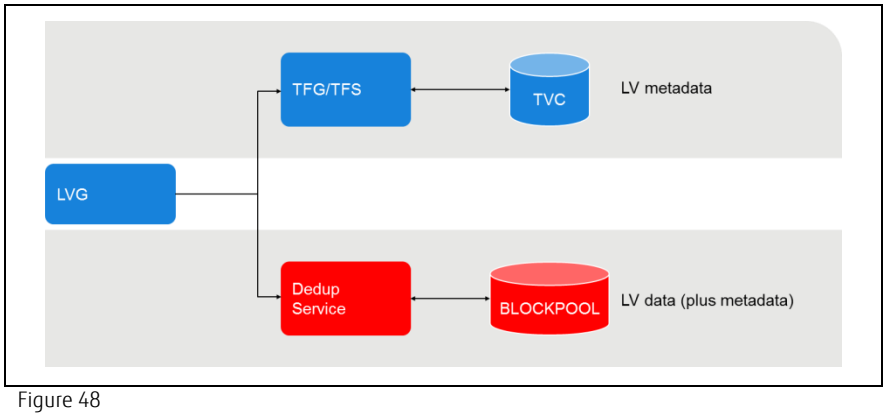


Figure 48

Processing of Logical Volumes:

Generally in a CS VTL system all logical volumes (LVs) are grouped in volume groups (LVGs). All LVs within a LVG have properties assigned to the LVG. In order to store LVs with a dedup service the LVG configuration is extended by parameters for deduplication. That means storage resources (LVGs, LVs) are linked to a deduplication service when they are created (LVG configuration, volume creation) and become part of that service.

Once a dedup service is available, at least one new logical volume group (LVG) must be created. As a configuration detail the LVG type is set to DEDUP-MANAGED and the dedup service to link-to is elected. Then a set of LVs is created within the LVG. Because the LVG is linked to a dedup service all volumes of the LVG are implicitly also linked. We call such volumes "Dedup logical volumes (Dedup LV)".

The tape emulations and the dedup client run on one or more ICPs. Each tape emulation handling a Dedup LV is associated with a dedup client and is able to handle "new style" LVs (dedup type LVs, stored in the BLOCKPOOL) and legacy type LVs (handled through the TVC). The knowledge about the LV type is available in the VLM, based on LVG configuration information. At LV mount time the VLM inform the ICP about the type of LV that is currently loaded in a logical drive. If it is a dedup type LV the dedup client will prepare to handle I/O to/from the BLOCKPOOL using the dedup service.

The type of tape emulation selected for the logical drive is not relevant for handling Dedup LVs; all emulation types provided by CS VTL support deduplication. Such the handling of Dedup LVs is not restricted to specific tape emulations. Every emulation on every ICP may mount a Dedup LV and perform I/O - assuming that a client for a specific server is running on the ICP. There is however a restriction: Dedup LVs will only mount successfully on logical drives located on ICPs that have been selected as clients for the given dedup service.

In order to access a Dedup LV, the volume is mounted by the application to a logical drive. This is done by using one of the traditional VTL library services such as VACS, VDAS, VJUK and so forth. Once a Dedup LV is mounted on a logical drive it is used in the same way as traditional LVs. For the duration of the mount the logical drive is linked to the dedup service and all I/O to the volume is done together with that service. After I/O is finished the volume is unmounted/unloaded by using the usual virtual library services. Once the volume is unloaded the logical drive is unlinked from the dedup service. The same logical drive may subsequently be used to handle traditional logical volumes as well.

Dedup Housekeeping: Efficient dedup processing of logical volumes need periodical housekeeping actions that are performed asynchronously to the I/O processing and are scheduled according to configuration parameters.

Currently there are two housekeeping actions related to dedup services:

- Cleanup of volume segments that have been marked for delayed deletion (delete segments): This action results from the way logical volumes are processed when they get overwritten. Old data needs to be deleted; however, since real deletion would be too expensive to be processed during the I/O phase the data (segments) are only marked for deletion and the real delete processing is deferred to the housekeeping action.
- Reclamation of space from the dedup data pool (compact): Remember that unique chunks of data get stored in the data pool together with a reference count. When data gets deleted from the dedup pool, the reference count of such unique chunks may drop to zero, which indicates that they are not needed any more. Much of the space occupied by unreferenced data is reclaimed automatically when new unique chunks get stored; however, an explicit reclamation action is required from time to time in order to avoid wastage of disk space. The space reclamation is done by the compact housekeeping action.

Restrictions and Limitations: The current implementation of dedup services has some restrictions and limitations regarding handling of logical volumes:

Restrictions:

- Only 'empty' LVGs – not containing any LVs – will be able to get the dedup service attribute assigned.
- The move of LVs stored in a dedup pool from its current LVG to another LVG is only allowed if these LVGs are linked to the same dedup service and use the same dedup pool and the same tape file systems.

It is not possible

- To link a Dedup LVG to PVGs (Physical Volume Groups). As a consequence it is not possible to migrate or copy dedup volumes to physical tapes or other backend storage.
- A 'conversion' between traditionally stored LVs to LVs stored in a dedup pool and vice versa.
- Migration of existing LVs into a dedup data pool.
- Import/export of LVs into/from the dedup data pool.
- For dedup volumes to move the LV to a different LVG - not even to LVGs linked to the same dedup service.
- Cascading over TCP/IP because it is not compatible in conjunction with dedup services; that is, dedup volumes can neither be source nor target in a TCP/IP based cascade between two ETERNUS CS8000 systems. Note that this restriction only applies to dedup volumes. Cascading with traditional volumes is not affected.

4.7.5 Operation of Dedup Components in the Cluster

Tape emulations run on ICPs distributed in the cluster. They are bound at LV mount time to a dedup client – running on the same ICP. A dedup service is operated on an IDP and has an assigned BLOCKPOOL stored in a CAFS file system. A dedup service is able to handle multiple clients in parallel. The central operation control by the VLM and the central DVM DB (Dedup Volume Management Data Base) ensures the storage of unique data blocks in the BLOCKPOOL regardless wherefrom the incoming LV are written by hosts.

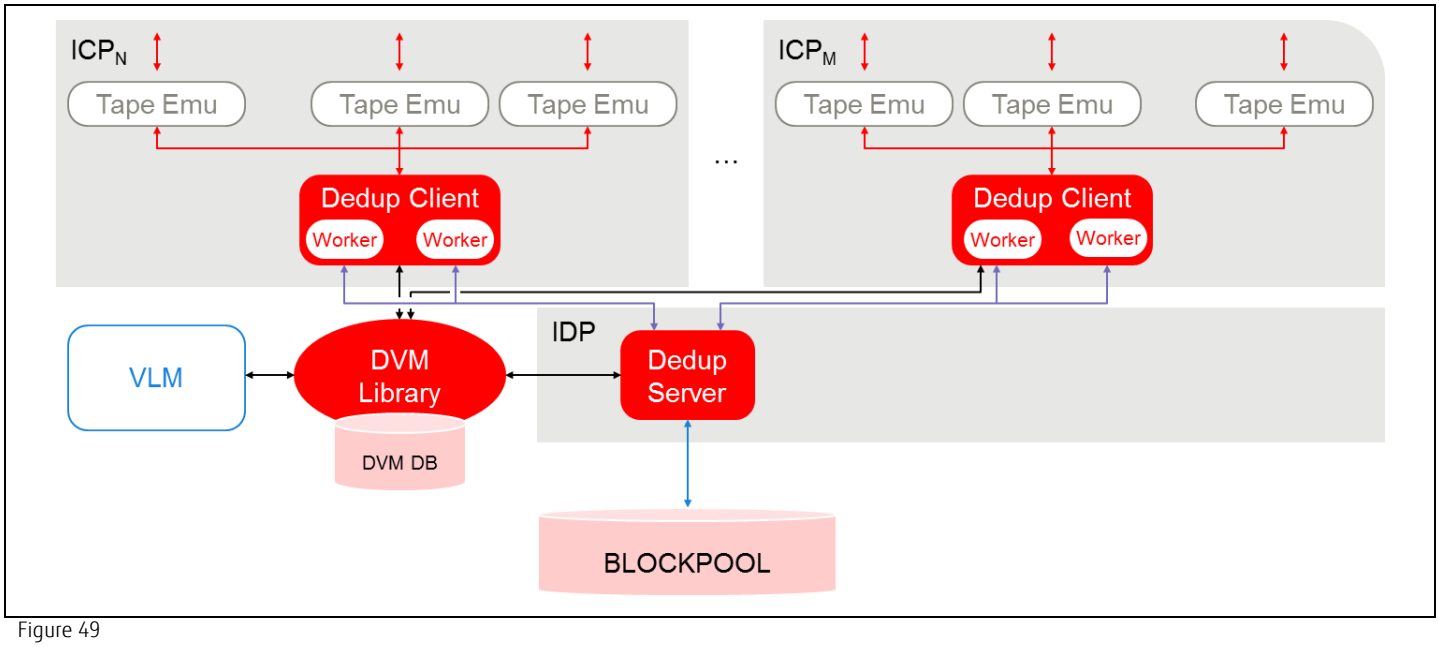


Figure 49

Example: A typical CS VTL system example with a dedup service may look as follows:

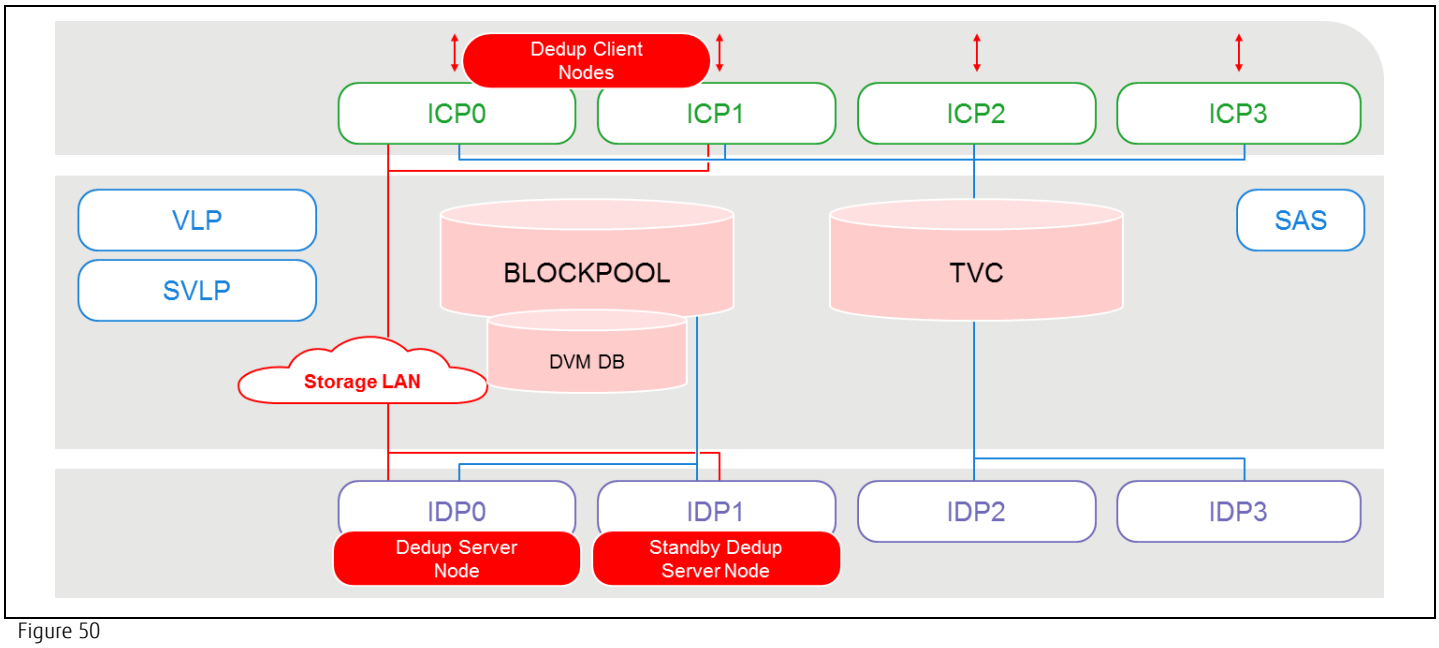


Figure 50

The example CS VTL system consists of the following components:

- VLP and SVLP (standby VLP) nodes.
- 4 ICPs: ICP0 and ICP1 with dedup functionality are able to handle dedup and non-dedup volumes, ICP2 and ICP3 are able to handle non-dedup volumes only.
- 4 IDPs: IDP0 as exclusive dedup server node, IDP1 as standby server node, IDP2 and IDP3 as tape backend IDPs.
- One or more RAID systems for the TVC.
- A RAID system for the dedup service (BLOCKPOOL and metadata)
- A 10Gb/s Storage LAN where nodes ICP0, ICP1, IDP0 and IDP1 are connected.

4.7.6 Multiple Deduplication Services in one System

CS VTL supports multiple dedup services in one system. Each service maintains its own pool of deduplicated data (BLOCKPOOL); as a consequence there is no deduplication effect (additional elimination of redundancy) across the dedup services.

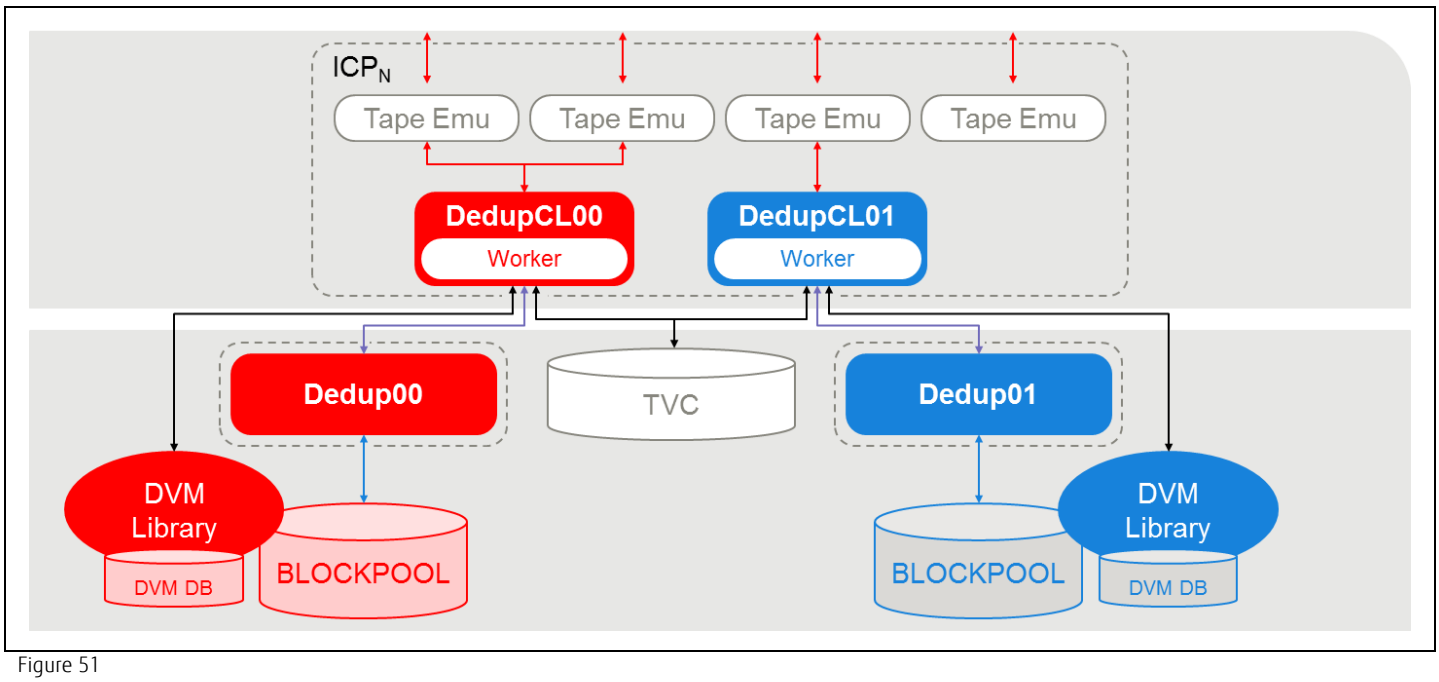


Figure 51

It may however be of advantage to operate more than one service:
- For multi-tenancy services the data can be cleanly separated into disjoint BLOCKPOOLs.
- Better I/O performance if BLOCKPOOLs reside on different RAID systems.

In some cases it may be required to operate more than one dedup service:
- Currently the disk space available to a dedup service is limited to one CAFS file system; file systems spanning multiple RAID systems are not supported.
- In order to store huge amount of deduplicated data that needs to be distributed to multiple RAID systems multiple dedup services are needed.
- Up to 10 Dedup Services, each with 480 TB Capacity for its BLOCKPOOL (10 x 480 TB native) are possible.

While each dedup service requires its own storage resources for the BLOCKPOOLs other resources may or may not be shared:
- Tape Volume Cache (TVC): TVC file systems are not exclusively assigned to a dedup service. The same TVC file system may hold volume stubs for arbitrary dedup volumes and also serve as a cache for non-dedup volumes.
- Dedup Server Nodes: For performance reasons each dedup server should operate on a separated IDP server node. However, it is technically possible to run more than one server on a node, in which case the CPU and memory resources of that node are shared among the server processes.
- Dedup Client Nodes: Dedup clients are bound to ICP nodes where logical drives are operated. The same logical drive may handle dedup volumes from different dedup services. Therefore one ICP must be able to operate multiple dedup clients.
- The resources of an ICP client node (CPU, memory) are shared among all dedup client processes on that node.
- The "special LAN" that serves as a transport LAN between dedup clients and server is shared among all dedup services.

4.7.7 Failover and Disaster Resiliency for Deduplication

For high availability purposes a failover concept for the Dedup Service is provided. The dedup data pool (BLOCKPOOL) is stored in a CAFS file system and is considered highly available (accessible from arbitrary nodes in the cluster). Since the Dedup Service works node bound a Dedup Server failover concept with a standby IDP is provided.

The Dedup Server failover is using following failover mechanism:
- Two IDPs are designated to operate a Dedup Server.
  - An active dedup IDP.
  - A standby dedup IDP.
- The BLOCKPOOL and any volatile data needed by the service are stored on a CAFS file system that is accessible by both IDPs.

During normal operation, the active dedup IDP is running the service, while the standby dedup IDP runs a monitoring service. The monitoring service actively executes tests to determine the status of the active IDP and of the service from the client's point of view. If a failure of the Dedup Server is detected, the dedup monitoring service initiates a failover of the service to the standby IDP. Finally the service will be available again on the formerly standby IDP. Currently CS VTL does not support automatic failover but a manual takeover of the server process to a standby is possible.

Dedup Clients which have noticed a service breakdown/failover because the connection of the dedup worker processes is closed will reconnect to the service on the new active IDP.

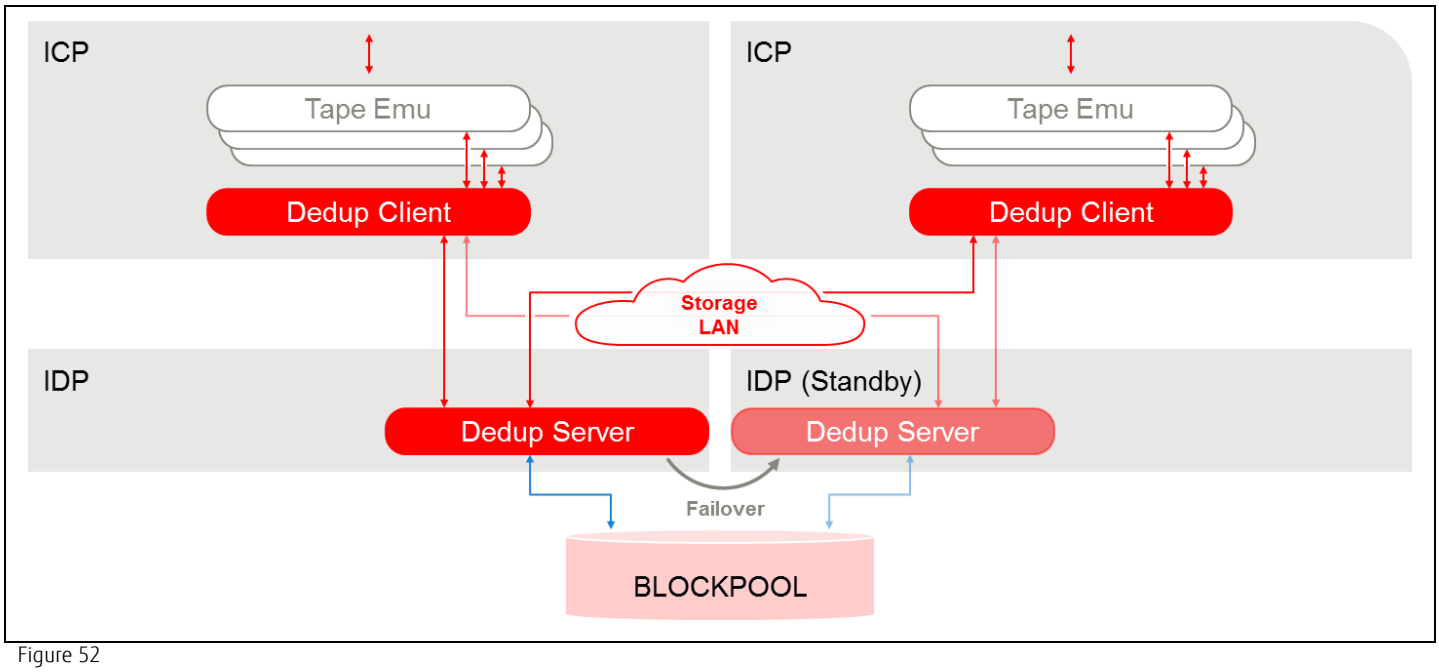


Figure 52

When operating more than one Dedup Server in a system each service may be assigned a dedup monitoring service. Configurations where one standby dedup IDP monitors multiple active Dedup Servers (each on its own IDP) are also supported. In such a configuration a standby dedup IDP runs multiple dedup monitoring services. However, once a failover takes place and the standby dedup IDP becomes active for one service the monitors will stop working. That is, only one failover is possible.

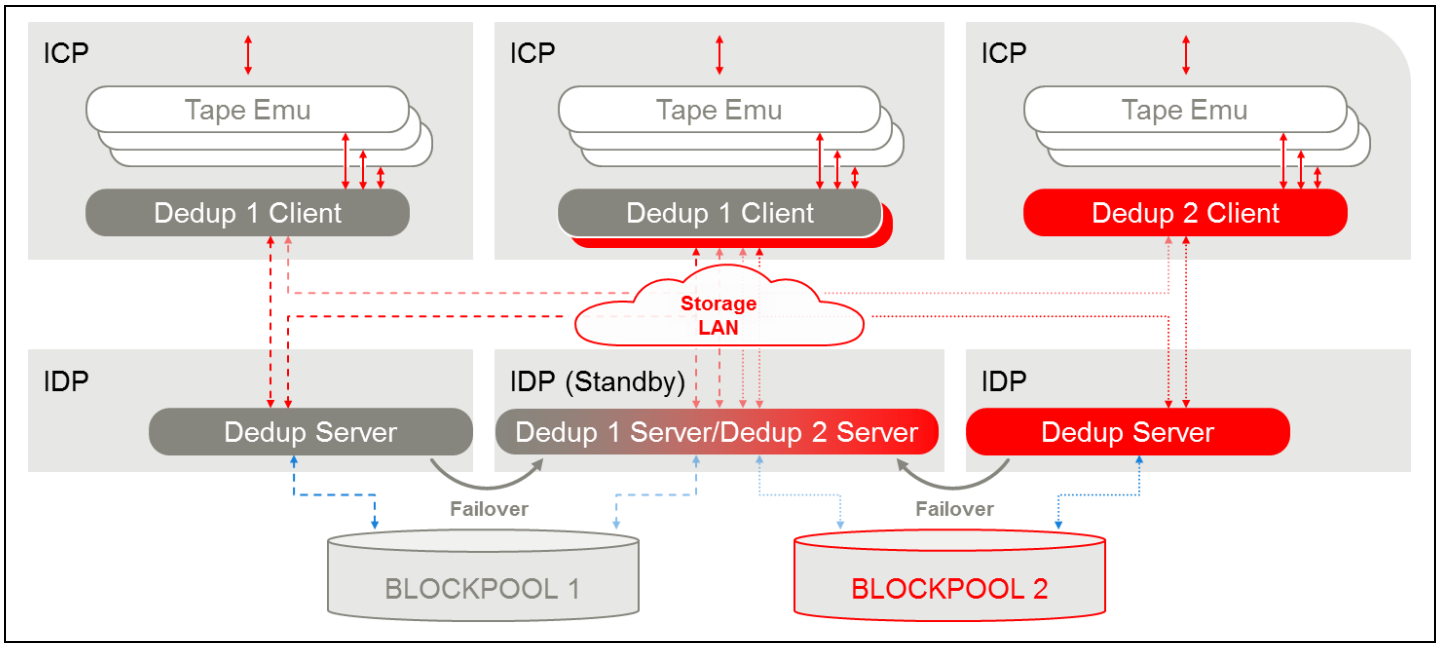


Figure 53

Rules and hints for configuration:
- Only configured ICPs will run a Dedup Client to a Dedup Server, ICPs without running a Dedup Client cannot access Dedup-LVs (mount of LV will fail).
- More than one Dedup Client per ICP is possible, that means that each ICP need to run at least one Client per Dedup Server in order to access LVs in the respective BLOCKPOOL.
- All tape emulations on the ICP have access to the Dedup Client and thus may handle Dedup LVs, it is not possible to restrict access at tape emulation granularity. Tape emulations on an ICP may access all Dedup Clients on that ICP; binding is done at LV mount time, depending on the Dedup Server the LV is stored with.
- Each Dedup Server instance owns a unique address, runs on a dedicated IDP and has its own BLOCKPOOL; data sharing between the BLOCKPOOLs is not possible. There is maximal one Dedup Server working on a single File System.

Availability of a Deduplication Service
- **Protection against RAID system failures:** Currently it is not possible to create tape copies of the logical volumes stored with a dedup service. Therefore it is crucial to protect the file system holding the BLOCKPOOL of the dedup service against disk and RAID systems failures. The general recommendation is to use a mirrored file system and to place the corresponding mirror LUNs on different RAID systems.
- **Availability of the Dedup Server:** A dedup service is based on a client/server architecture where the clients may be distributed among several cluster nodes (ICPs) while the server runs on one node only. In order to protect against failures of the dedup server node a server process failover is provided. The prerequisite for a server failover is the availability of a standby node in the CS VTL that provides similar resources and connectivity as the dedup server node does.
- **Availability of the Dedup Transport LAN (SLAN):** Another critical point in the client/server infrastructure is the SLAN as transport between clients and server. To protect against failure of single components (LAN switches, Ethernet controllers) all such components are doubled and redundantly cabled.

4.7.8 Use Cases

The following describes use cases where use of data deduplication may give significant advantage over using traditional storage methods. The combination of using disk space efficient deduplication together with traditional long term storage on physical tapes gives best results for the customer.

Use Case: Backup using two data pools - the customer defines its backup policy:

Type1 backup - daily (whatever level): The retention period of such backup data is typically short, let's assume 4 weeks. After this period of time the backup application automatically deletes the backed up data and reuses the volumes.
Type2 backup - every N week (normally at level FULL): The retention period of such backup data is typically much longer than that for the type1 backups, let's assume 12 months. After this period of time the backup application automatically deletes the backed up data.

Experience shows that backed up data are accessed with much higher probability during the first 4 weeks. Therefore it should be fast accessible during this period. After this period the backup data that is still valid (Type2) should be migrated to inexpensive media in order to free disk space - at the cost of higher delays in case the data is accessed after migration.

On a CS VTL system with a dedup data pool this scenario would be implemented as follows:

Configure two LVGs: LVG1 for backup data of Type1 and LVG2 for backup data of Type2.
- LVG1 is configured to store LVs in the dedup data pool.
- LVG2 is configured to store LVs in the traditional way and is linked to one or more PVGs to migrate data to physical tape.

Resulting data flow:
- Type1 backup to LVs of LVG1:
  - Data is stored in the dedup data pool only. The user may choose to always perform full backups to avoid fragmentation and longer restore times. Due to deduplication the space required for full backups is about the same as required for incrementals.
  - LVs expire after 4 weeks and are recycled by the application (deleted from the dedup data pool).
- Type2 backup to LVs of LVG2:
  - Data is stored in volumes in the TVC and migrated to physical tape so space in the TVC is freed. Multiple copies are created according to the configuration.
  - Restore from such volumes is only possible after they are loaded into the TVC again. However, as long as other copies of the data are available in the dedup volumes the probability for need of accessing this data becomes very low.
  - After 12 month the LVs are recycled (overwritten) by the application.

Use Case: Linking the CS ViNS Subsystem to a Dedup Service

The ViNS subsystem in an ETERNUS CS8000 appliance uses tape technology to migrate file data from NAS file systems to backend storage. The loopback configuration, where the CS ViNS subsystem is connected to the CS VTL subsystem of the same ETERNUS CS8000, is one possibility to build an integrated HSM system with multiple copies of the data.

By defining the primary storage group of a CS ViNS backend to be a Dedup LVG it is possible to take advantage of deduplication technology. This provides fast access to migrated data at lower disk space requirements. It is even possible to create additional tape copies of the data by configuring the CS ViNS copy storage pool to use a traditional LVG that gets migrated to tape.

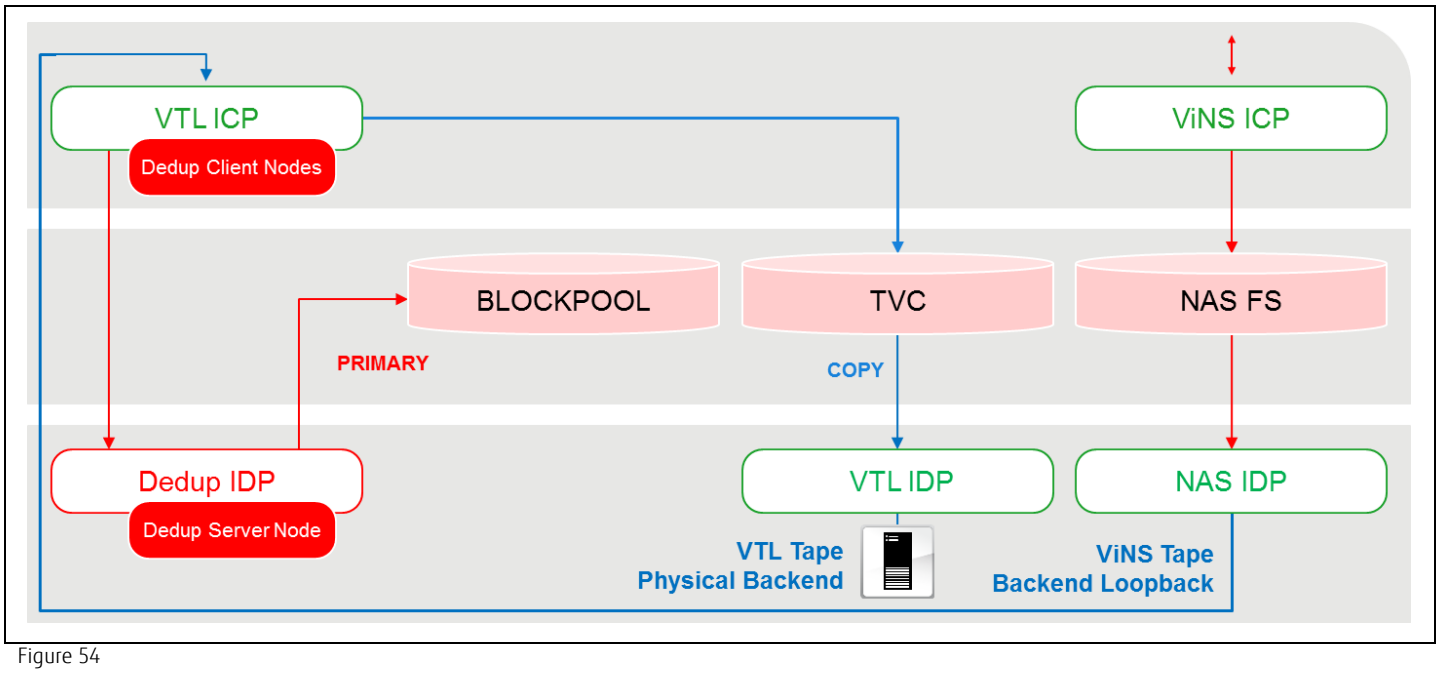


Figure 54

4.7.9 Benefits of Deduplication

When appropriately deployed, data deduplication can provide benefits to users that need to store their traditional backups on disk only. It can significantly reduce disk capacity requirements and enables longer retention periods for data that are required to keep online. With deduplication reducing disk storage requirements, the environmental costs for running and cooling the disk storage also are reduced. In addition performing restores from disk can be faster than restoring from tape.

The CS VTL deduplication offers a couple of advantages:

- Highly scalable deduplication appliance
- Unified deduplication for Open Systems and Mainframes
- Backup application writes data as it would to tape
- Flexible policy driven administration by choosing which LV to store deduplicated
- By using multiple deduplication services the system can be partitioned into multiple storage areas that are clearly isolated from each other.
- Store up to 10- to 30- times more tape data on given physical storage capacity
- Additional disk space is saved as unique data chunks stored on disk are compressed with standard data compression techniques.
- Enables more data to be protected with low RTO (Recovery Time Objective)
- Disaster resiliency by way of cache mirroring,
- Significantly reduces cache mirror bandwidth and enables large amount of data for one-to-one synchronous replication
- Failover of the deduplication engine and failover of the appliance

5 CS ViNS Subsystem Interfaces and Operations

5.1 CS ViNS Interfaces and Logical Components

The CS ViNS subsystem provides a virtual very large file storage that is accessed through NAS Interfaces, currently using the NFS (V2/V3/V4) and/or CIFS (V1) protocol. The simplest view of CS ViNS is that of a NAS filer providing access to a file store over a number of LAN interfaces. Since NAS clients use the file store through CIFS or NFS mounts, it appears to them just like local storage.

However, CS ViNS is especially designed as a storage server for NAS clients in data protection environments. It is also an ideal secondary storage target to relieve a primary host storage system by moving less used or outdated data from the primary storage to ETERNUS CS8000. To perform such special requirements it comprises a variety of scalability, availability and other features which are important in such environments, such as:

- Highly available IP access to the file store
- Replication of file systems within the ViNS Online Store to a second CS ViNS system
- Long-term file protection against modification or deletion
- A hierarchical storage management for copying and migration files from the ViNS Online Store to tape storage

File system replication and HSM cannot be configured on the same file system. The main reason is that replication cycles may cause unwanted HSM recall operations. This is especially true when replication is configured for a HSM managed file system which already contains a large set of data with a high percentage of migrated files.

CS ViNS is divided into two main functions: the ViNS frontend and optionally two different ViNS backend services.

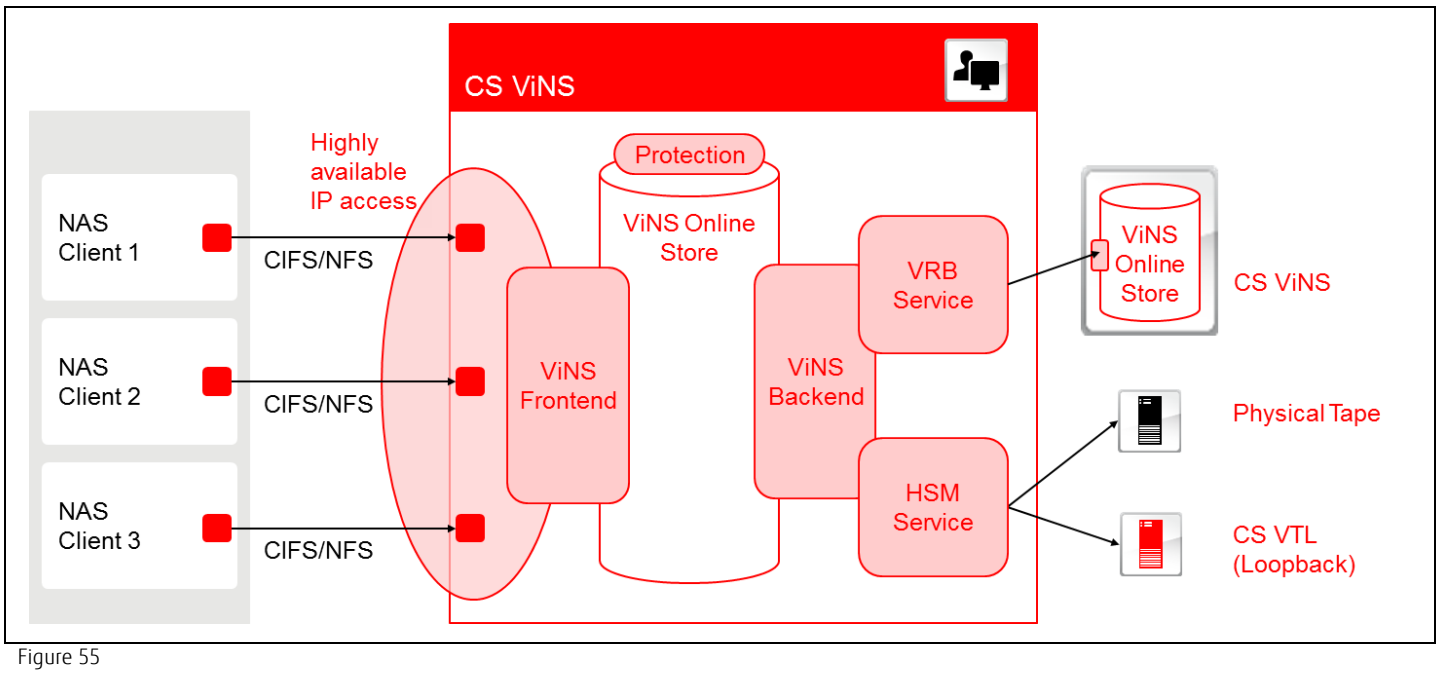


Figure 55

ViNS frontend

The ViNS frontend provides the ViNS Online store access to the NAS clients. The hardware and software components are:

- RAID systems for the ViNS Online Store (VOS)
- NAS ICPs and their external network interfaces
- IP pools
- NAS services
- NAS shares
- File systems and file groups (ViNS frontend view)

ViNS backend for file system replication

The VRB (ViNS replication and backup) service offers a data protection method by CS ViNS which performs the replication of file systems kept in the VOS to another CS ViNS system. The components are:

- A second CS ViNS system to store the replicated file systems
- Access paths to this system:
  - NAS ICPs and external network interfaces
  - IP pools
- File systems to be replicated
- Further replication attributes

The VRB service (ViNS backend) is running on an ICP such using the same network interfaces like the ViNS frontend.

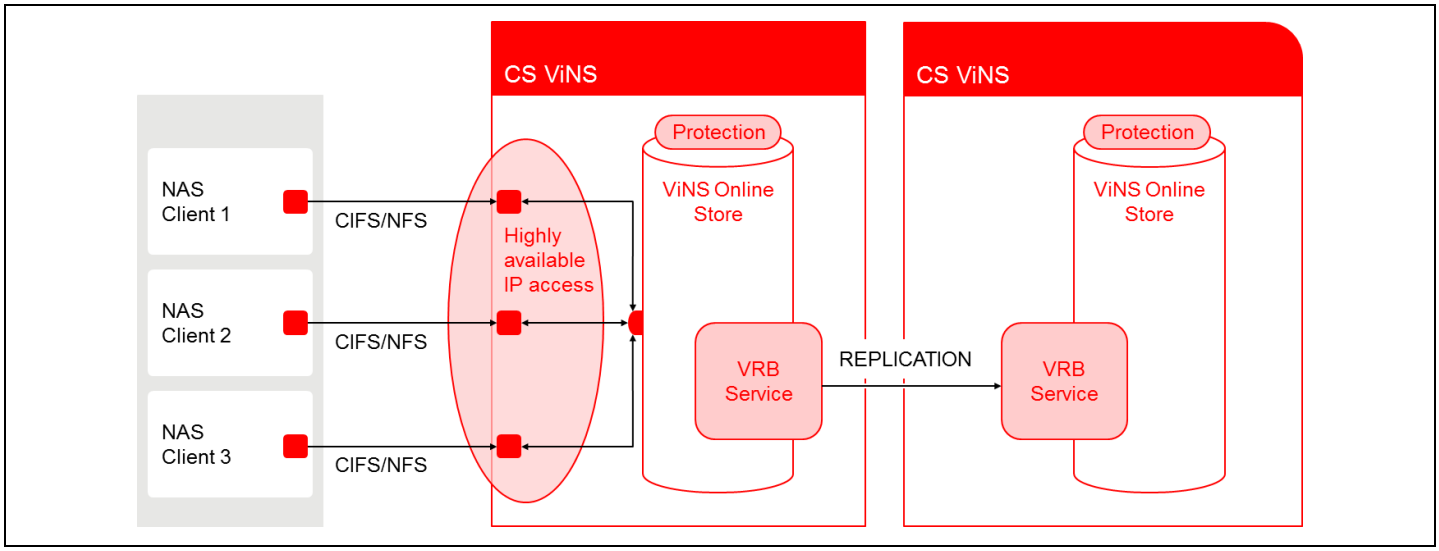


Figure 56

ViNS backend for hierarchical storage management

CS ViNS can be equipped with a tight interlocked HSM (Hierarchical Storage Management) service which migrate files stored in the VOS to tape and manage the lifecycle of the files within the different storage media. The HSM service comprises the following resources:

- File systems and file groups (ViNS backend view)
- NAS IDPs with access to tape device hardware
- Tape libraries
- Tape drives
- Tape volume pools
- Tape volume pool groups

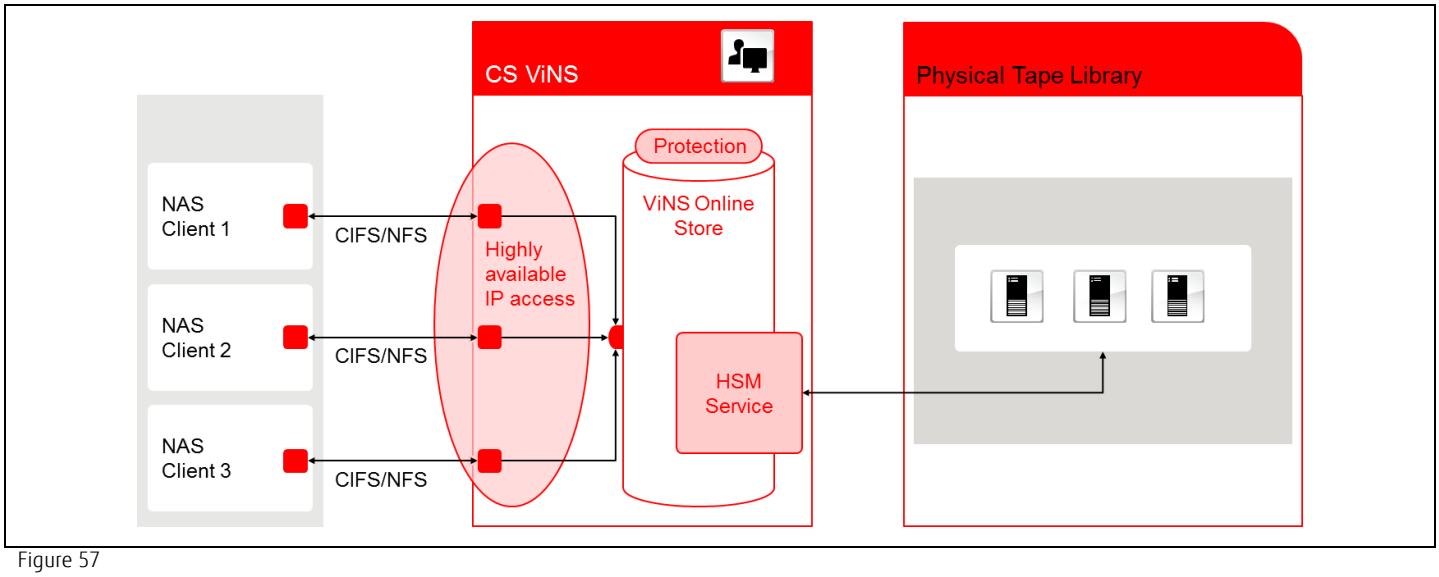


Figure 57

Since the HSM service uses tape libraries to migrate and recall HSM file copies, it may particular use a CS VTL for this purpose. The connection of a CS ViNS system to a CS VTL may be implemented in two ways:

- Cascaded access: The tape libraries used by the CS ViNS are implemented as CS VTL in a separate ETERNUS CS8000 system.
- Loopback access: The tape libraries used by the CS ViNS are implemented as CS VTL in the same ETERNUS CS8000 system where the CS ViNS is running.

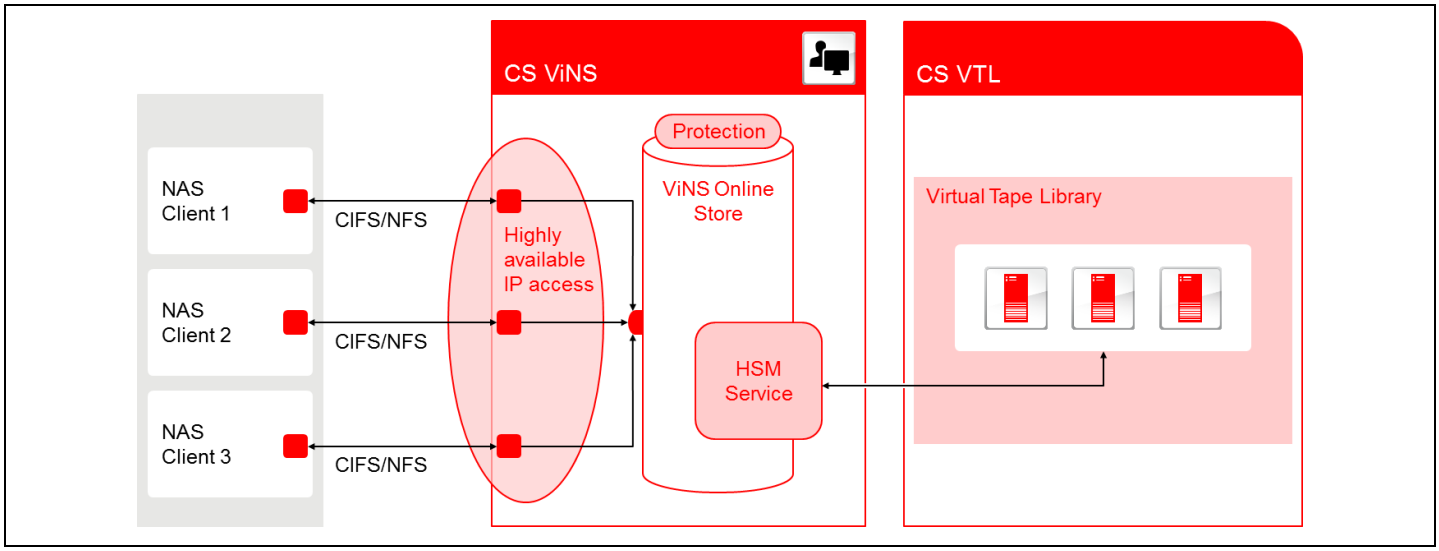


Figure 58

5.2 CS ViNS Frontend

The CS ViNS frontend provides an interface for direct access to the RAID systems configured for ViNS. At the CS ViNS frontend so called NAS shares, based on file groups (i.e. subdirectories of NAS file systems), are exported for being accessed by NAS clients. NAS share access is possible via the network file services NFS and/or CIFS.
Both services can be used for mounting the NAS shares exported by CS ViNS on the customer's system. The NAS shares reside in the cluster file system and are exported by several NAS ICPs. This means that different NAS clients can access the same NAS share and file system via more than one NAS ICP.

NAS services: Each NAS ICP is set up to provide NFS and/or CIFS services, i.e. to run the respective nfsd or smbd service processes. The supported NAS protocols are NFSv2/v3/v4 and SMBv1.

In addition to the NFS and CIFS daemon processes, a ctdbd (Cluster Trivial Database Daemon) service is running on each NAS ICP. This service is responsible for:

- Synchronizing the node-specific data of the CIFS servers throughout the cluster to enable the CIFS service to be used in the overall CS ViNS cluster.
- Switching over the IP addresses within an IP pool to standby interfaces when a failure occurs.

File systems and file groups: NAS file systems (NASFS) configured as being of type NASFS are not exported to NAS clients in their entirety. Instead, they are divided into file groups. A file group is a directory underneath the root node of a NAS file system. At the ViNS frontend, file groups are the entities which are exported to NAS clients via shares. The diagram shows a NAS file system /nas/500 with three file groups fg1, fg2 and fg3.

If a NASFS is not HSM managed its file groups must be set up in the NAS Frontend/File Groups configuration dialog. Otherwise, it is recommended to create them in the NAS Backend/File Groups dialog because at the ViNS backend file groups have a special meaning as well.

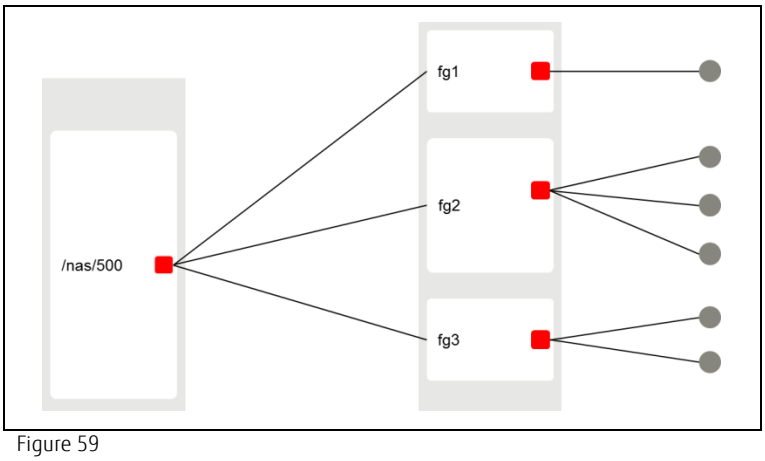
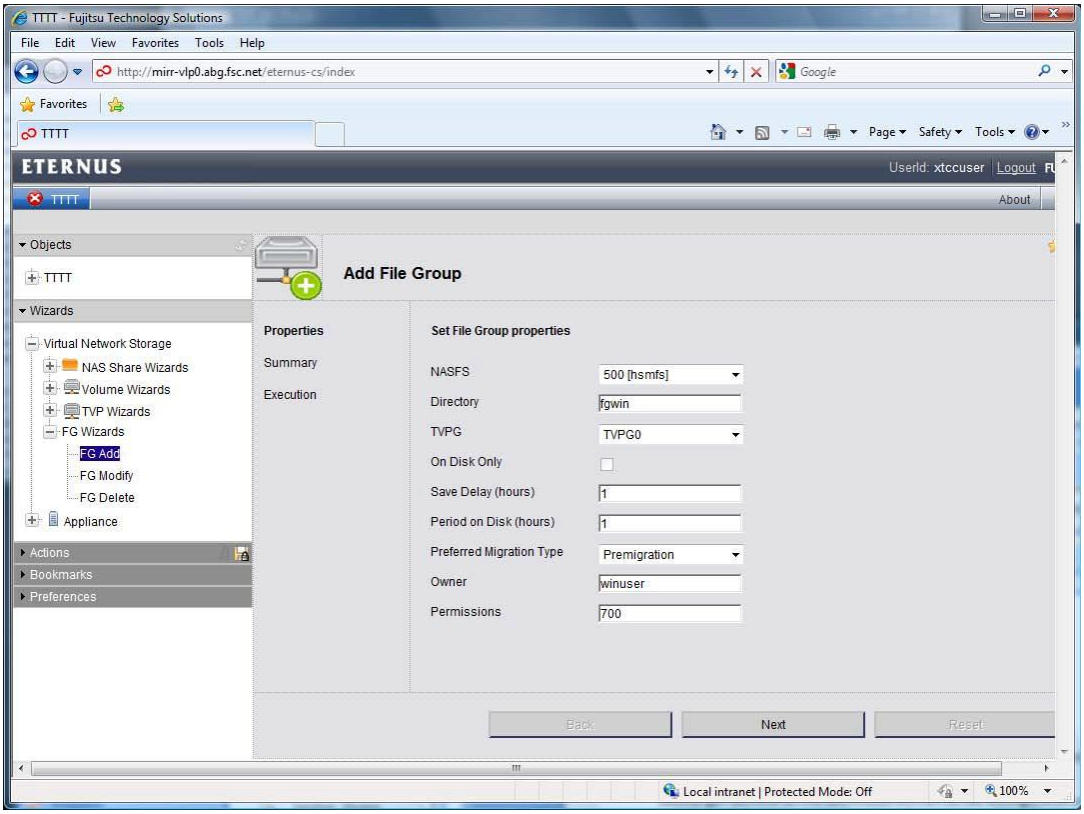


Figure 59



File protection period: There is only a single ViNS frontend parameter for NASFSs. It controls the protect ability of a NAS file system. NAS clients can establish a protection period for a file by simple file operations, provided that the corresponding file system was configured as "protectable". To do so, it is only required to set the access time of the file to a time stamp in the future and to remove the writing authorization on the file. The file remains protected against modifications until the time stamp specified in the access time is reached. The protection period can also be set by the protection command.

Shares: A share is a set of attributes controlling the accessibility of a file group by NAS clients. For example, it defines which NAS client hosts may access the file group at all. Based on the information contained in shares, file groups are exported to NAS clients so that they may access the configured NAS services on all NAS ICPs. There are separate shares for NFS and CIFS, e.g. any given file group may be exported only via CIFS, even though the CS ViNS cluster contains NAS ICPs which also run NFS services. Shares can be created, deleted or modified at any time by means of the WebGUI or the administration CLI.

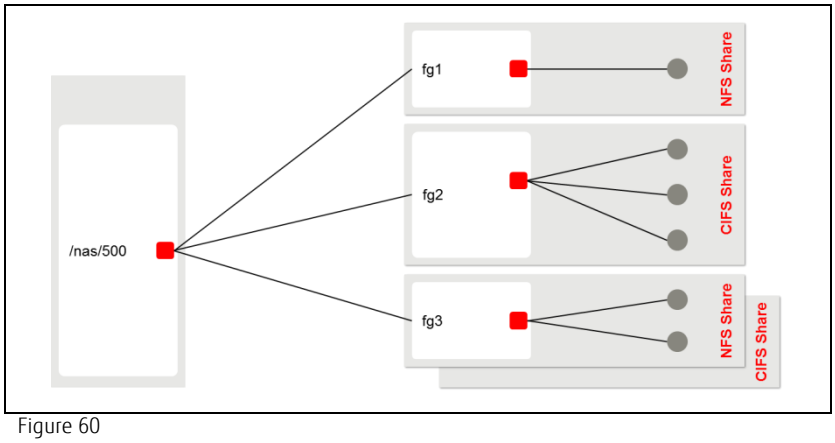


Figure 60

IP pools and failover of IP addresses within IP pools: An IP pool consists of a set of IP addresses in the same sub-network. In the CS ViNS configuration each IP pool is assigned a set of external LAN interfaces of the NAS ICPs to which the addresses of the pool can be distributed.

The data traffic on the incoming side is distributed over the available interfaces of a NAS ICP by the NAS clients making equal use of the IP addresses of the IP pools with their assigned interfaces. Routing settings on the NAS ICP are responsible for the distribution of the outgoing data to the interfaces. Effective load distribution is achieved on the incoming side and also on the outgoing side

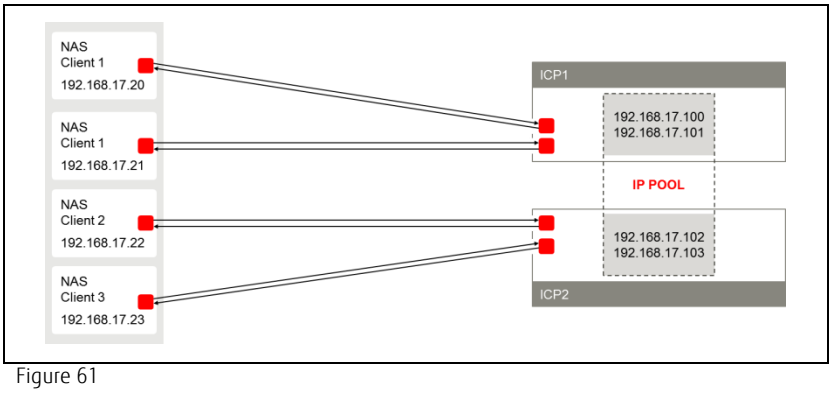


Figure 61

For reasons of high availability IP addresses are evenly distributed over NAS ICPs. If a NAS ICP (global failover) or a single network interface (local failover) fails, the IP addresses of the interface are transferred to an operable interface that is assigned to the same IP pool. This ensures that the IP addresses assigned to a service are always available.

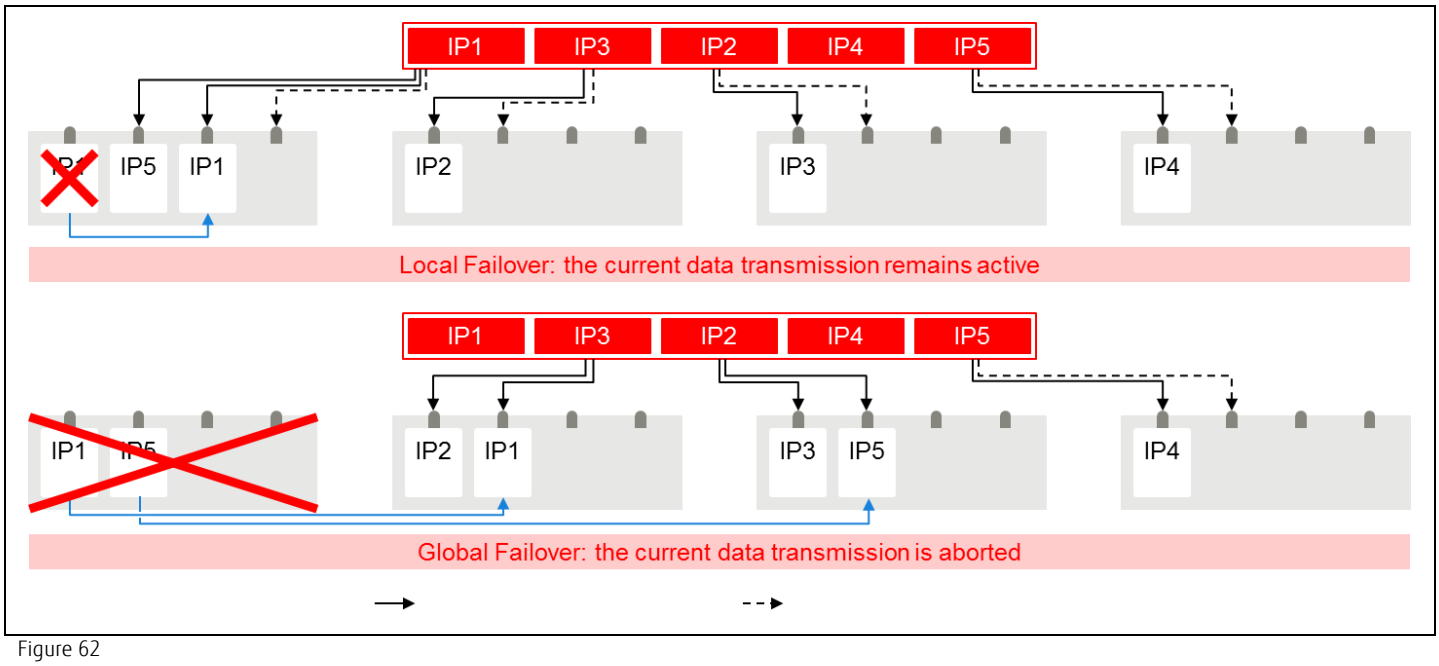


Figure 62

If possible and configured correspondingly a local failover is performed first. That means the same NAS ICP takes over the IP address. If a local failover is not possible (failure of a NAS ICP) another NAS ICP takes over the IP address/es.

To speed up the failover process the following is performed automatically:

- The NAS clients are informed about the changed MAC address for this IP.
- In case of a local failover the TCP connection remains active and retransmission of lost packets happens automatically.
- In case of a global failover the TCP connection is reset and the NAS client reconnects itself.

5.3 CS ViNS Data Replication

ViNS data are replicated asynchronously from a source CS ViNS system to a target CS ViNS system. The replica of a file system is activated and used for productive operation in disaster cases, e.g. when the original copy gets lost. It is also possible to switch operation to the replica in maintenance scenarios.

The asynchronous replication of NAS file systems is running on the NAS ICPs of a CS ViNS cluster. It periodically selects data to be replicated and copies the data to a second, remote CS ViNS system.

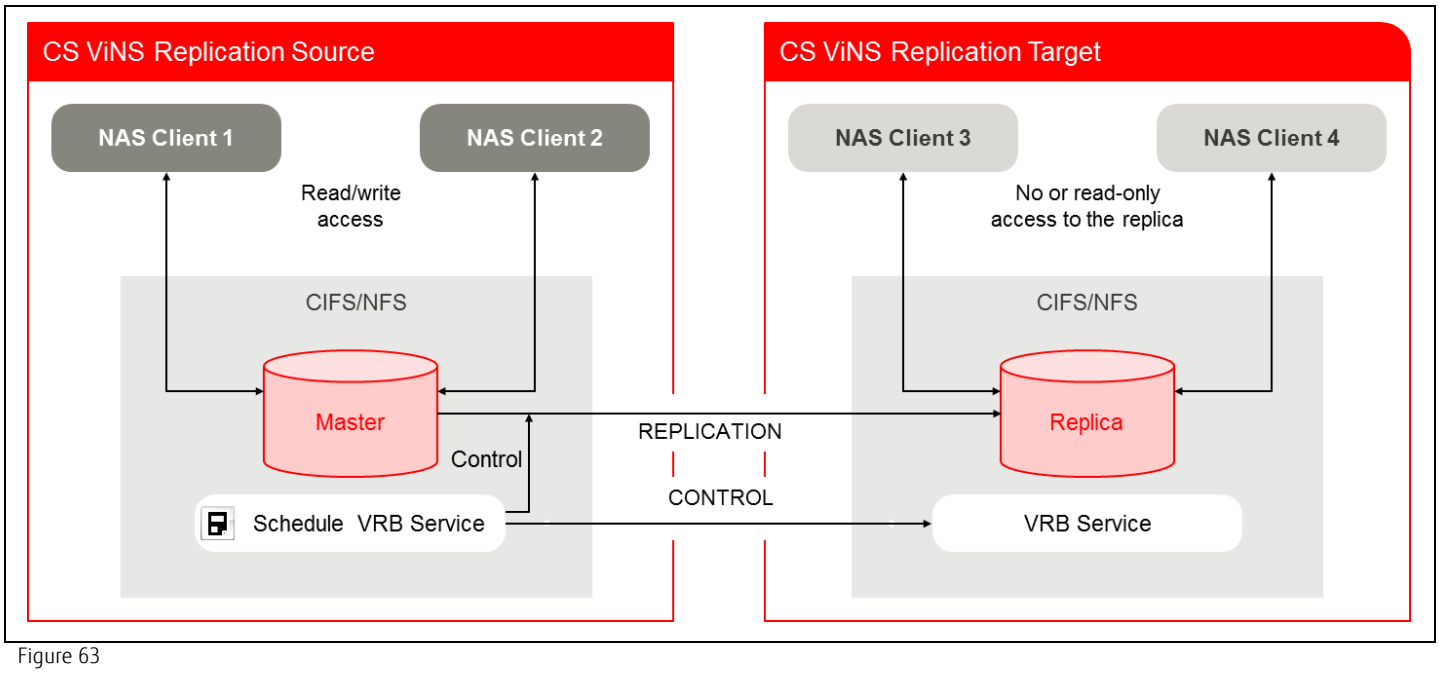


Figure 63

The ViNS replication and backend service (VRB service) creates and updates copies of CS ViNS file systems on a remote CS ViNS system. A file system is replicated from a source system to a target system. The terms "source system" and "target system" (or "source" and "target" for short) are meant per replicated file system. The source system is usually the productive one, i.e. it is the system whose NAS file systems are used by NAS clients to store data from the user's backup and archive services. A given file system cannot be replicated within the same CS ViNS. However, a CS ViNS can be source for one set of file systems and target for another set.

Hints

- Master and replica of a NAS file system will be on two different CS ViNS systems. Within a single ETERNUS CS8000, the CMF feature already offers protection against loss of data.
- The replication will be done asynchronously to the usage of the file system by applications running on NAS clients.
- The replication target can be placed off-site (long distance).
- The data at the replication target is used to quickly recover from data loss at the replication source. The RPO ("recovery point objective") depends on the snapshot period + transport time of snapshot differences to the replication target.
- The same CS ViNS system may act as replication source and target at the same time for any two different NAS file systems.

Replication takes care that all modifications of a file system master copy at the source system are propagated to a replica copy at the target system within a time interval to be configured. Both the master and the replica are frozen into snapshots at the source resp. at the target. A snapshot is a file system image from a given point in time. While the file system contents may be changed from then on, the contents of the snapshot does not change.

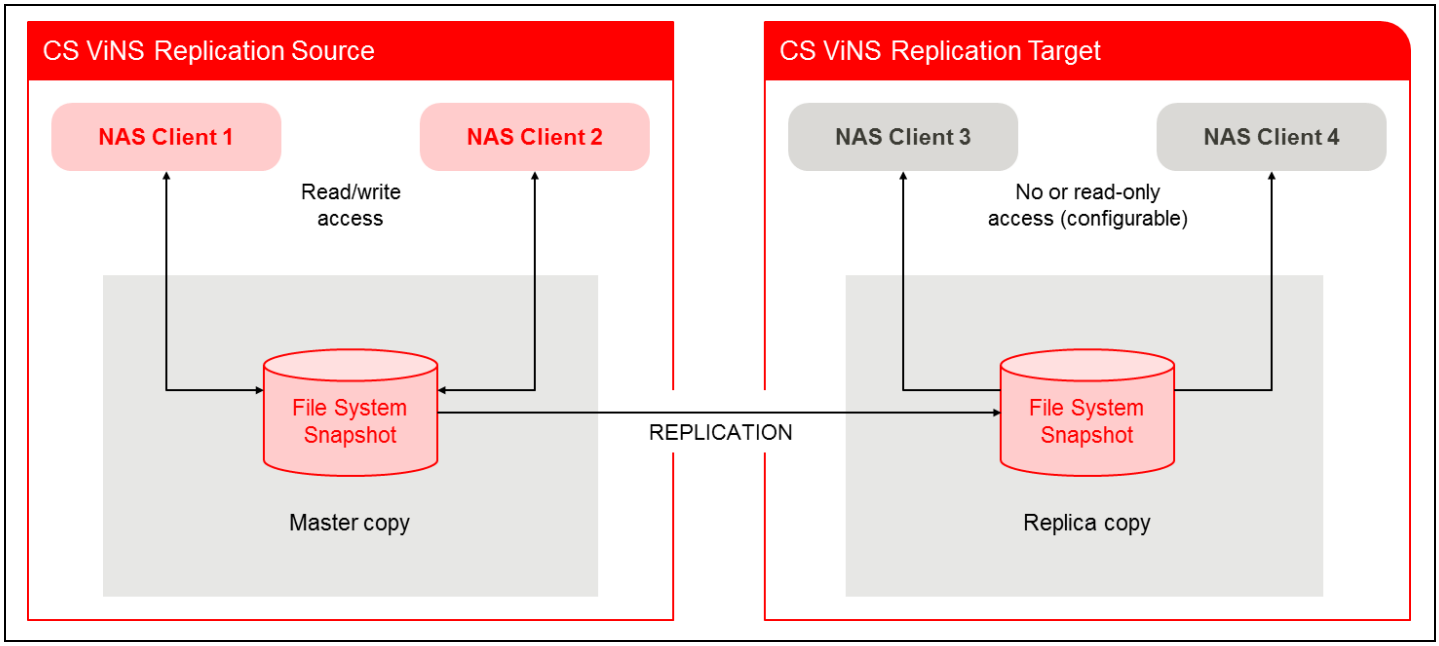


Figure 64

5.3.1 Setting up the Replication of a NAS File System

A NAS file system can be configured as a file system which is replicated between two CS ViNS systems. The replication must be separately configured on both systems. For each master file system, there must be exactly one replica file system and vice versa. The replica should not be smaller than the master.

For a given NAS file system, the user must specify on both systems at least
- the identity of the replicated file system
- whether the local CS ViNS system is source or target for the file system
- the access path to the replication peer (local and remote IP address to be used for the replication)

At the source system, the NAS access to the master is not restricted by the VRB service. At the target, the administrator decides whether to export the file groups of the replica for read-only access or not at all. However, NAS clients cannot be given write access to a replica.

The VRB service is automatically started on each ICP node which takes part in an IP pool associated with a master or replica file system. On each such node, a pair of processes is running – an active VRB service process and a passive VRB service process. The active VRB service replicates file systems for which the CS ViNS system has the source role. The passive VRB service acts on the target system. Its activities are mostly controlled by the active counterpart.

The VRB service configured on ICP nodes uses IP pools for highly available communication to the peer. Depending on the expected NAS and replication load, it may be useful to separate the IP pools used by the NAS services from the IP pools used by the VRB services or even to run NAS services and VRB services on different sets of ICP nodes.

5.3.2 Replication Operation

The replication of file systems happens asynchronously. At configurable points in time, a scheduled replication cycle is started for a given file system: The replication procedure determines the differences between the current file system contents and the most recent contents that have been successfully replicated. The result of this procedure is used to create a new replica. Only files and directories which have been modified between two replication cycles are selected for being replicated in the current cycle.

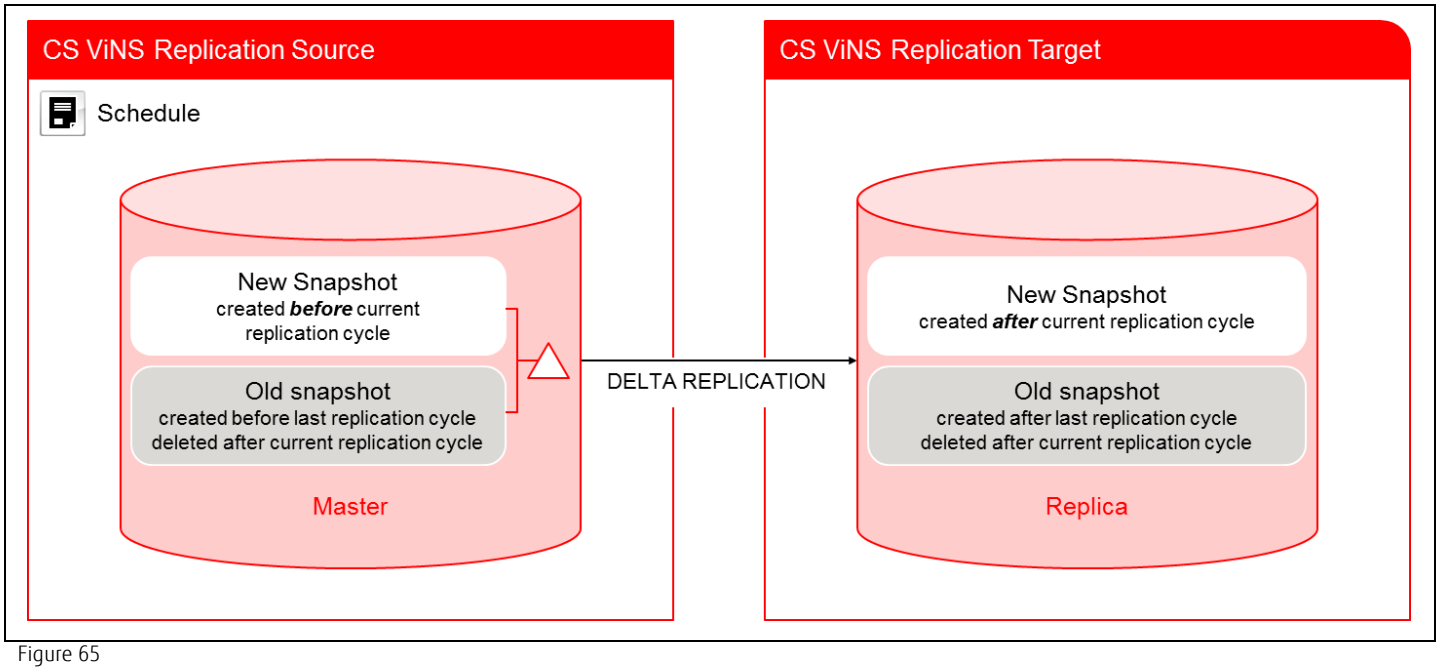


Figure 65

The picture above illustrates how a replication is executed. At the start of a replication cycle, a snapshot of the file system is created at the source system. The creation of a snapshot preserves the current contents of the file system in a read only image within the file system. File system data are copied "by reference", i.e. real data copies are only made when data referenced by the snapshot are modified or deleted in the file system.

The files to be replicated during a cycle are determined as follows:

Assuming that there is a snapshot from the previous replication cycle, the new snapshot is compared to the old one resulting in a list of differing files and directories. Only these different files and directories are replicated into the target's file system. When the replication is terminated successfully, a new snapshot is created at the target system as well. The old snapshots at both sides are then deleted.

The set of replicated data is limited to a minimum. The VRB service does not replicate entire files but only the portions which have been changed. If file metadata are modified, but the data remain the same, only the metadata are replicated.

Replications are scheduled and started automatically with a configurable file system specific interval. Each VRB service allows a maximum number of replications to be running in parallel. A scheduled replication job may be rejected because that maximum number is exhausted. If this happens repeatedly for the same file system, then the file system may not replicated for a possibly long time. Therefore the next replication job for the file system is scheduled with a special rescheduling interval if a replication attempt for the given object results in a "too many active jobs" error. Should the same error occur during the retry, then the waiting period is extended by another rescheduling interval, and so on.

Apart from the automatically scheduled operation, replication jobs can also be started manually through the User Interface (CLI and GUI).

The file system image which was replicated is available in a snapshot at both the source and the target.

A replication job may fail for various reasons, e.g. because a replication peer is shut down or the connection between the peers is broken. At this point, the replica may have only partially replicated contents. The user may set up the configuration such that the replica is automatically restored from the last valid snapshot in case of an error. However, the next replication job will usually put back the replica into a consistent state again without such extra measures.

5.3.3 Data Synchronization

The VRB service takes care that any file modifications are replicated from the source to the target. The modifications may cover file data or metadata, including extended file attributes and access control lists. If a file is deleted or renamed at the replication source, the operation is propagated to the target in the next replication cycle. If a file must be protected against modification and deletion, the containing file system needs to be configured as protectable at both the replication source and the target.

5.3.4 Encrypted Replication

Depending on the data path between the source and the target system, it may be necessary to encrypt the data to be replicated. Encryption is accomplished by sending the data through a SSH tunnel.

The picture below shows configured from a source system to two different target systems. The replication of the "green" file system is configured without encryption because the link between the source system and the first target system is secure. Since this is not true for the link to the second target system, the "red" file system is configured with encryption. The replication data stream between the source system and the second target system is passed through a SSH tunnel (blue).

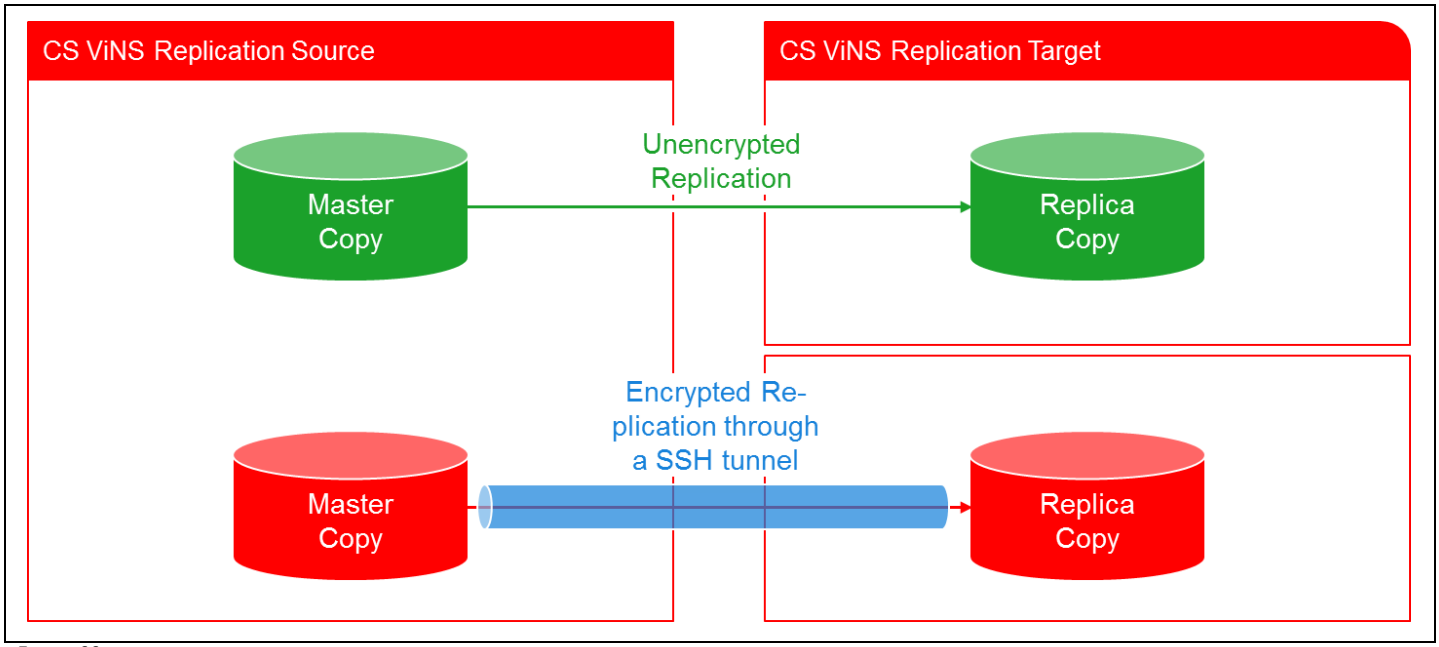


Figure 66

For setting up an SSH tunnel, the SSH public and private SSH keys must be created and distributed to the nodes on the source and target systems. Then the file system for which an encrypted replication is required can be configured accordingly.

5.3.5 Manual Failover and Failback

It is assumed that there is a preferred primary CS ViNS system which normally serves as the replication source, while the file system is replicated to a secondary CS ViNS which serves as the replication target. The primary/secondary roles are static, while the source/target and the active/inactive roles may change.

The replica serves to protect the file system against the loss of the file system on the primary CS ViNS. In case of such a failover, the administrator can execute a manual failover procedure. The objective of this procedure is to make the replica available for productive operation.

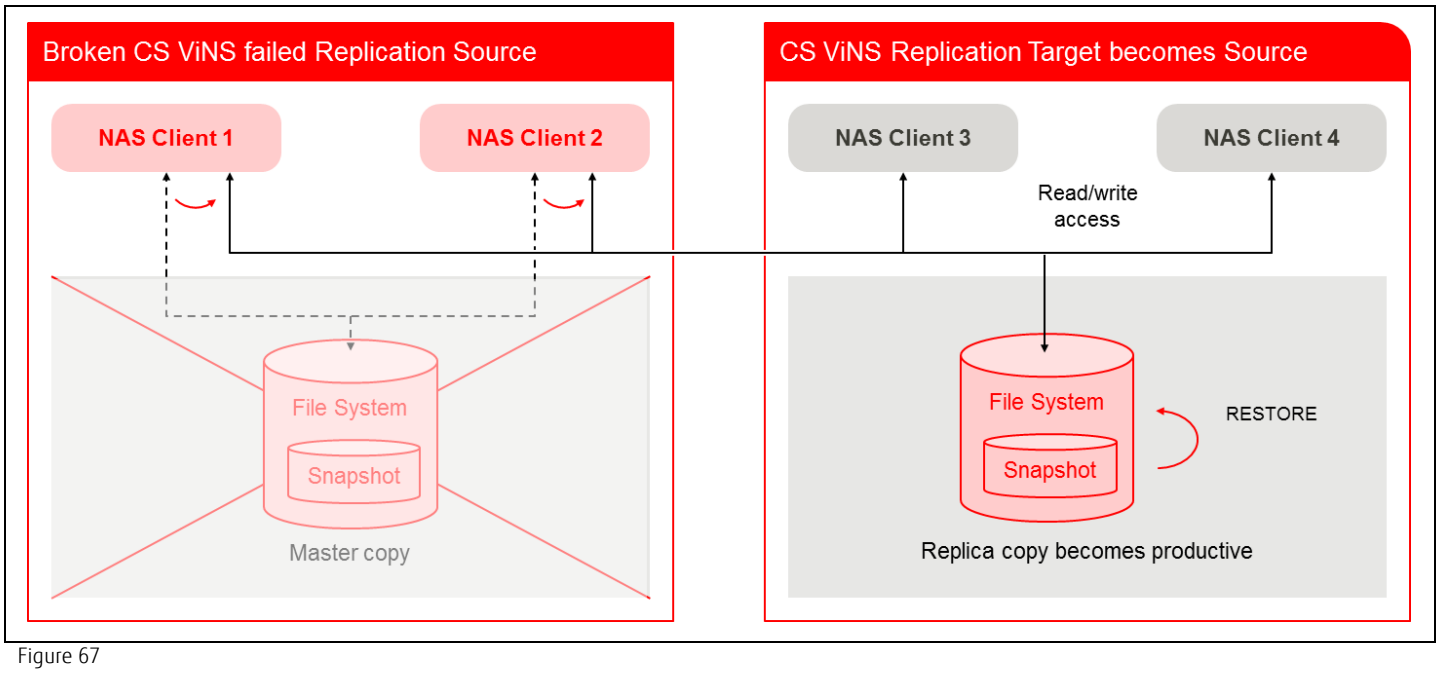


Figure 67

Essentially, the CS ViNS system which used to be the replication target of the file system ceases to be the target. It is reconfigured to host the master copy of the file system, potentially restored from the last valid snapshot. The NAS clients which had mounted the file system from the failed primary system need to remount it from the secondary system.

After the primary system has been repaired, the administrator can execute a manual failback procedure to shift the productive operation back.

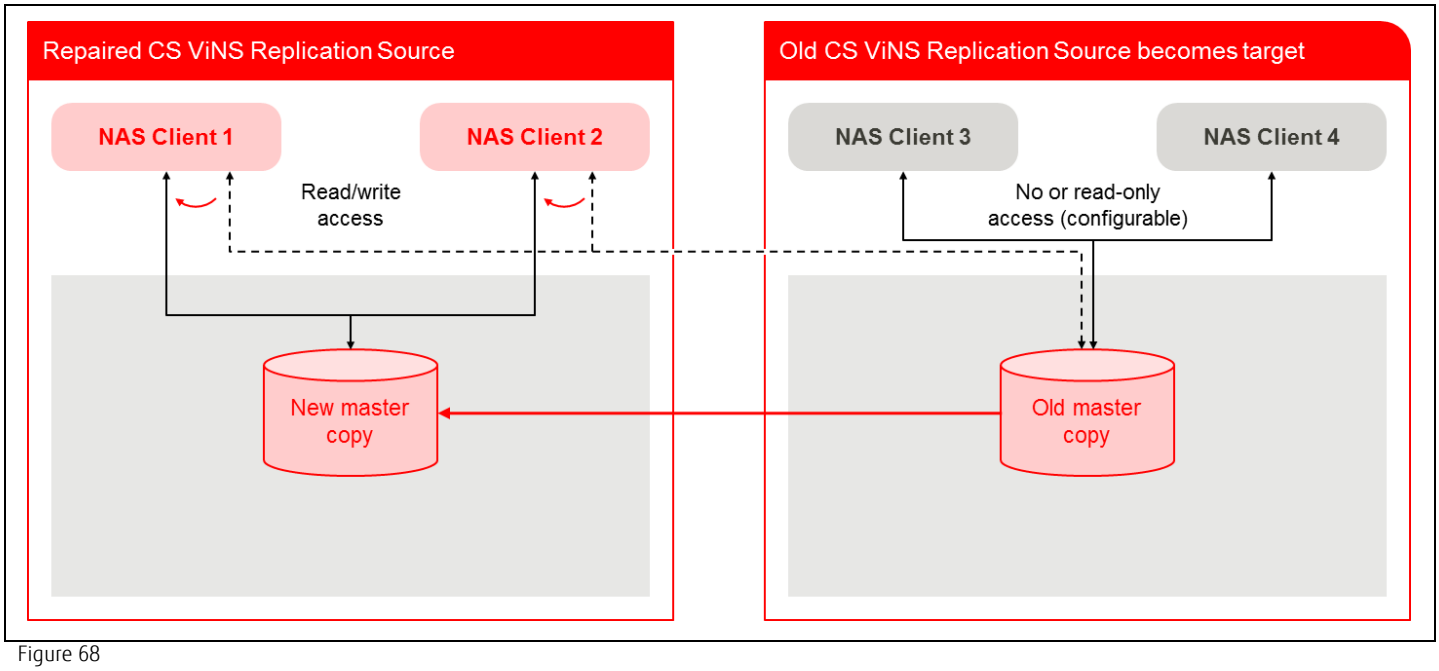


Figure 68

5.3.6 Automatic Recovery from Replication Failures

The following table lists the possible VRB resource failures and their automatic recovery measures:

Failure	Measure
VRB service process failure (unexpected termination)	Automatic restart
IP link failure to the peer	IP address reassignment to another network interface on the local node (if possible) or to a network interface on another ICP node with interfaces in the same IP pool
ICP node failure	IP address reassignment to a network interface on another ICP node with interfaces in the same IP pool
ETERNUS CS8000 site failure	IP address reassignments to network interfaces on ICP nodes at the surviving site which have interfaces in the same IP pool
File system / RAID system failure or total ETERNUS CS8000 failure	No automatic recovery – but a manual failover to the replication target system(s) is possible

Link, node or site failures have an impact only on those replication jobs which send/receive data using the failed IP addresses. Other replication jobs remain unaffected.

Pooling of interfaces and IP addresses protects the communication path uses for replication in the same way as for the NAS services (see 5.2 ViNS frontend). If a pooled interface fails, its IP address is transferred to another interface - preferably on the same ICP. If an ICP hosting a pooled interface fails, its IP address is transferred to another interface on another ICP.

5.4 CS ViNS HSM Backend

5.4.1 HSM Service

The ViNS backend comprises the services of the hierarchical storage management for ViNS Online Store (VOS) within CS ViNS. These are optional services and referred to under the generic term HSMS (Hierarchical Storage Management Service). A HSMS runs on a ViNS IDP. It ensures that files are premigrated or migrated from the NAS file systems to the tape storage connected to the ViNS IDP according to configurable criteria.

The HSMS consists of four main components:

- **Information Lifecycle Management Component (ILMC):** The ILMC is requesting the HSMC to perform premigration and migration actions based on HSM-specific configuration settings for NAS file systems and file groups.
- **Hierarchical Storage Management Component (HSMC):** The HSMC is requesting the BSMC to create file copies on the tape storage in the context of premigrations, migrations and recalls. The HSMC manages the states of premigrated and migrated files in the VOS and executes the automatic recalls.
- **Backend Storage Management Component (BSMC):** The BSMC copies the data between the VOS and the background storage and manages the resources and the content of the background storage.
- **HSM Service Monitor (NASMON):** The task of the monitoring component (NASMON) depends on the role of the HSMS in which it is running. If the NASMON is configured together with the other components, its task is the monitoring of resources needed by the HSMS on the local NAS IDP.

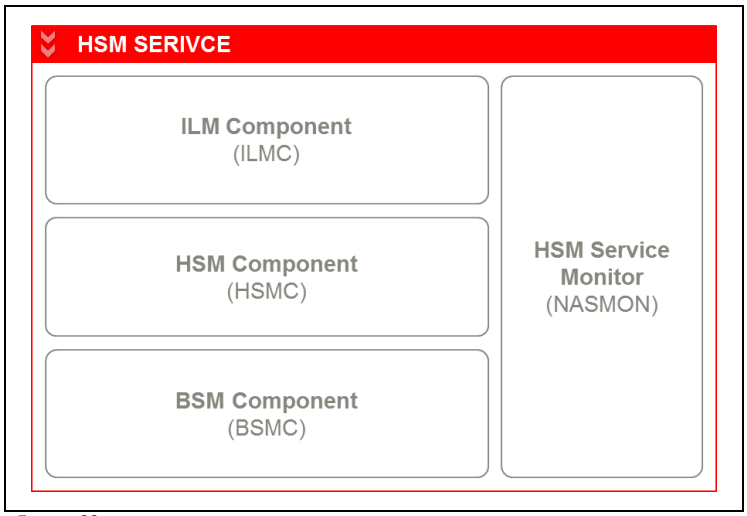


Figure 69

Information Lifecycle Management Component: The ILMC selects files within a CAFS according to a number of selection criteria in order to premigrate or migrate them to backend storage. The selection criteria are defined in the configuration of the NAS file systems and file groups. The ILMC converts these criteria into ILM policies which can be applied directly to the CAFS.

Hierarchical Storage Management Component: The HSMC manages the possible HSM states of the files in the NAS file systems. From the HSM viewpoint a file can be in one of the following states:

- **resident:** The file exists in the VOS. There is either no copy of the data in the background storage or the background copy is obsolete.
- **premigrated:** The file exists in the VOS, and there is a valid copy of the data in the background storage.
- **migrated:** The file occupies no data blocks in the VOS. The data exists only in the background storage.

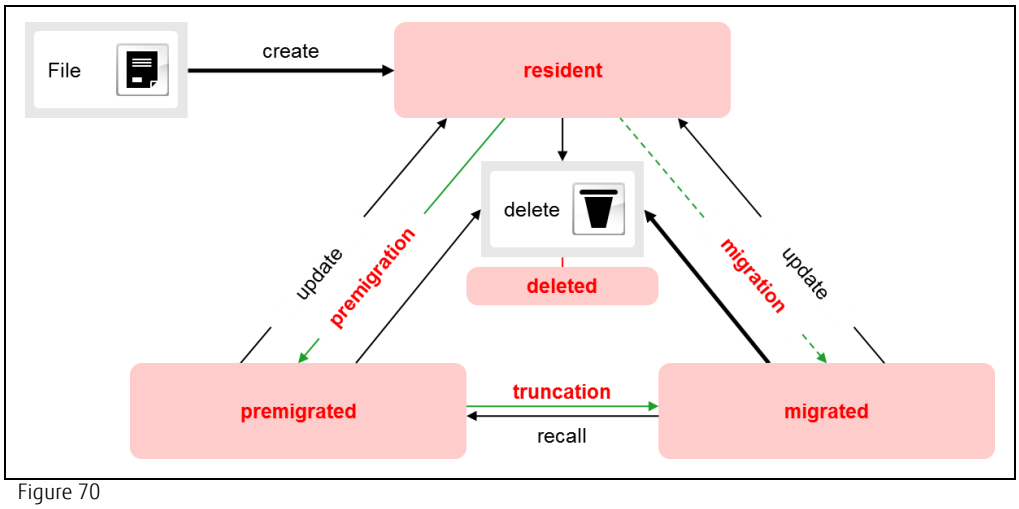


Figure 70

The HSMC receives requests from the ILMC to premigrate, truncate, or migrate files which the ILMC has selected in accordance with the configured selection criteria.

The HSMC is also informed by the CAFS if a process
- accesses a premigrated file in write mode or
- accesses a migrated file in read or write mode.

In both cases the HSM state of the file changes. In the latter case the HSMC executes a recall of the file. If a change of state requires data to be transferred between the VOS and the background storage, the HSMC requests the BSMC to perform this transfer.

Backend Storage Management Component: The BSMC is responsible for reading and writing data from/to the background storage and keeps a record of which files are stored on which tape volumes. It manages and monitors all the resources of the background storage. These include hardware resources such as tape libraries, drives and volumes, but also logical storage objects such as tape volume pools.

From the BSMC viewpoint a tape library must be homogeneous, i.e. it may only contain drives which can read and write media of the same type. An existing library may consequently need to be partitioned in such a manner that this requirement is satisfied. A tape volume pool is a set of volumes within a tape library. All volumes of a pool are of the same type, e.g. T10000. A tape volume pool group is a set of tape volume pools. The media on which HSM copies of files are to be stored (media type, WORM attribute, number of file copies) can be defined specifically for each file group.

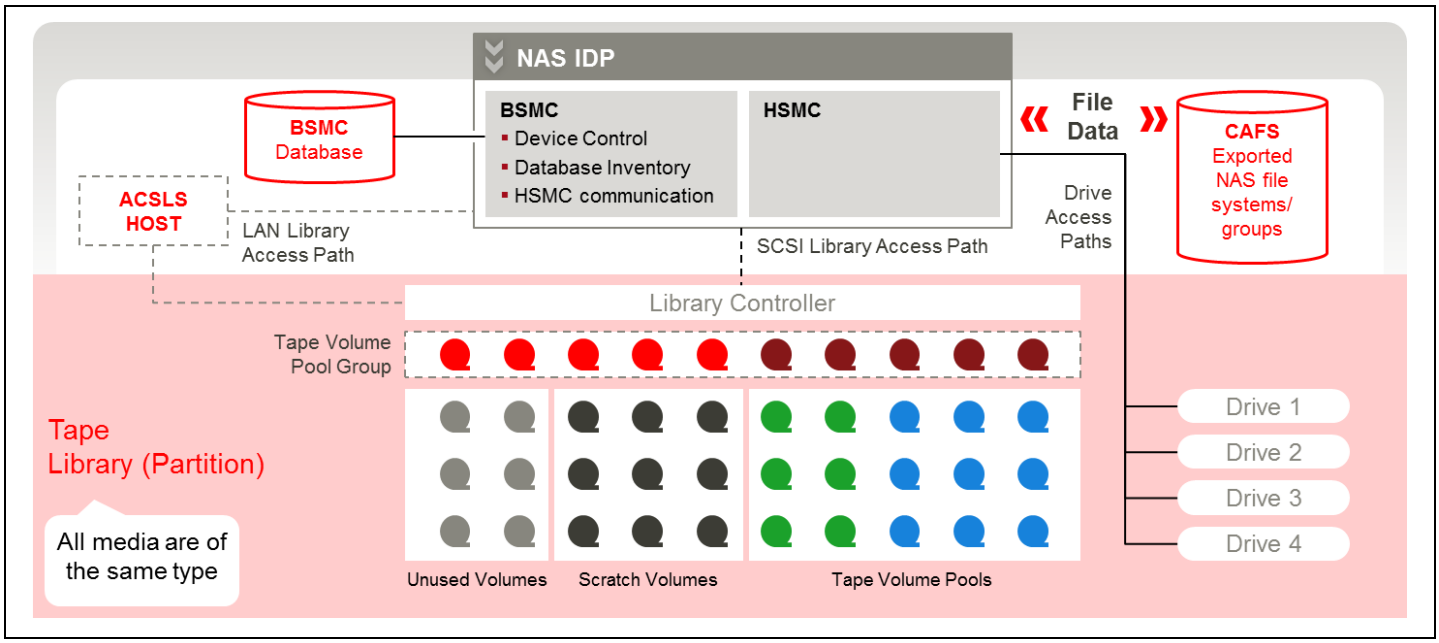


Figure 71

HSM Service Monitor: The ViNS backend provides the NASMON component to ensure the availability of the HSM service in case of hardware or software problems. The NASMON is responsible for monitoring either the node-local resources of an active HSMS or the availability of a node running an active HSMS. In the latter case, the NASMON is running on a standby NAS IDP as the only component of a "standby HSM service". If the NASMON detects a failure of the observed NAS IDP, it triggers a failover of that HSMS to the local NAS IDP.

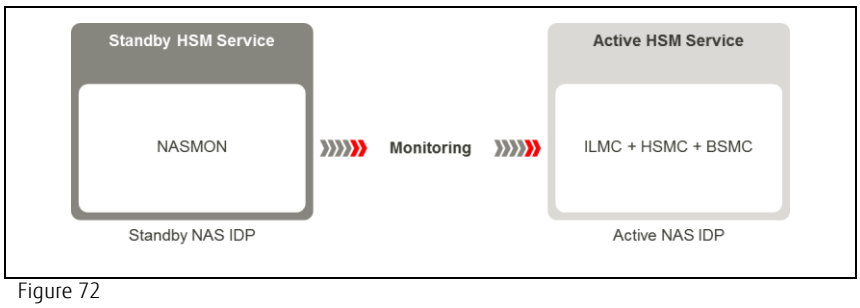


Figure 72

Types of HSM services:
Each HSM service has a role which determines the components to be run within the service. The available roles are:

- Active:
  An active HSM service consists of the components ILMC, HSMC, BSMC and optionally NASMON. It runs on an active NAS IDP and provides hierarchical storage management for one of more file systems.
- Active_BSM:
  "Active_BSM" is to be read: "active HSM service with only a BSM component". This is a special case of an active HSM service. A HSM service in role Active_BSM consists of a BSMC component and optionally a NASMON component. It runs on an active NAS IDP. It does not provide hierarchical storage management for any file systems. Instead, it acts as a library manager for tape libraries used by active HSM services on other nodes.
- Standby:
  A standby HSM service is an auxiliary service which provides no HSM on its own. It contains only a NASMON component and it runs on a standby NAS IDP. It monitors a set of active HSM services and triggers a failover when a problem with the monitored service or its NAS IDP is detected.
- Fallback:
  A fallback HSM service is an auxiliary service on the VLP nodes which is only used in situations where no NAS IDP is available. In these special cases, it provides read and write access to premigrated files. Fallback services are bound to the VLP nodes and do not take part in any failover scenarios.

How the HSMS components work together:
The interplay of the components ILMC, HSMC and BSMC within the HSM service are internal defined operational sequence for the following basic operations:

- Periodic premigration or migration:
  ILMC periodically runs jobs which, depending on the preferred migration parameter of a given file group, either premigrate or migrate the files of that file group.
- Truncation:
  Truncation of the files in a NAS file system is started when the file system disk space reaches the Threshold High value and ILMC knows about the existence of premigrated files (information collected during previous periodic scans). The purpose of truncation is to free storage space in the NAS file system as rapidly as possible.
- NoSpace Migration:
  NoSpace migration of the files of a NAS file system is started when there is a shortage of space and no more premigrated files are known to ILMC. The purpose of migration is to create free storage space in the NAS file system as rapidly as possible.
- Recall:
  A file is recalled when an application accesses the file in read or write mode.

Validity of file copies in the background storage:
The data of a file which is stored in the background storage is always valid if the state of the file is premigrated or migrated.

It becomes invalid when a file in the premigrated or migrated state is modified. In such cases the file state is set to resident. However, this only applies when the user data is changed. If only the metadata (e.g. permissions) is modified, the state is not changed. A file copy in the background storage also becomes invalid when the file is deleted. Invalid file copies are removed by the HSMC during what is known as reconciliation. In this operation the data administered by the file system is compared with the one in the background storage. Detected invalid copies are not discarded by the BSMC immediately but rather only after a certain expiration date.

5.4.2 File Systems and File Groups

If a NAS file system is associated with a HSM service in the ViNS Backend/File System Management configuration, it becomes a HSM file system (HSMFS). The basic definition of a file group at the ViNS backend is the same as at the frontend: A file group in a HSMFS is a directory one level below the HSMFS root directory. However, the ViNS backend allows setting additional HSM relevant attributes for file groups. The main ViNS backend parameters of HSM file systems and file groups are shown in the following diagram:

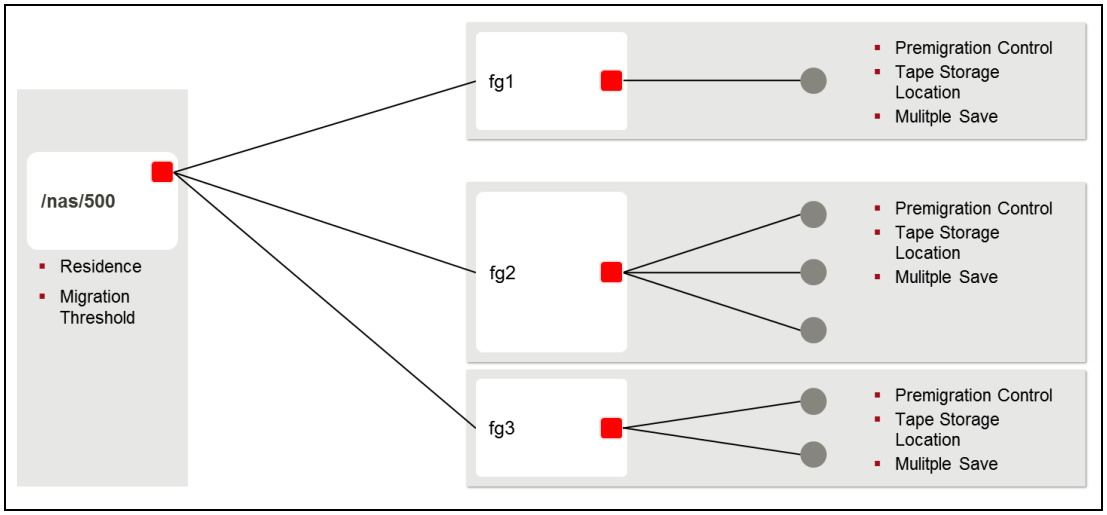


Figure 73

The HSM relevant attributes of a file system control if and when files of the HSMFS are migrated. The residence attribute determines whether files are migrated at all, and the migration thresholds tell when it is time to start or stop the NoSpace Migration of the file system. More HSM attributes exist at the file group level. These determine when and where files within the group get periodically premigrated or migrated. The tape storage location of a file group implicitly defines whether files are premigrated on a single copy or on dual copies.

5.4.3 Physical Tape Storage

Tape Libraries:
The ViNS backend stores HSM copies of files onto tape volumes contained in tape libraries. A HSM physical tape library must be homogenous. This means that it must only contain volumes of a single type - e.g. LTO-5 - which can be read and written by each drive in the library. No two different HSM services may use the same logical tape library, except for the special case of library sharing. Tape libraries used by the ViNS backend and the tape libraries used by the VTL must also be strictly separated.

Tape drives:
Tape drives are automatically recognized and preconfigured by the ETERNUS CS8000 hardware scan. The scan distinguishes between tape drives connected to VTL IDPs and NAS IDPs and preconfigures the devices as either VTL or HSM physical drives. Therefore the same drive cannot be connected to nodes of both types.

Tape volume pools:

A tape volume pool (TVP) is a set of volumes within a tape library. Since a tape library is required to be homogeneous, all volumes of a TVP are of the same type. A TVP may be of type PRIMARY, COPY or BACKUP. A TVP of type PRIMARY is used for storing HSM copies of files. A TVP of type COPY serves to hold a backup copy of the data stored in a primary TVP. A TVP of type BACKUP does not contain any HSM copies but only contains metadata stored by the HSM service backup job. The diagram contains an example layout of TVPs in two tape libraries:

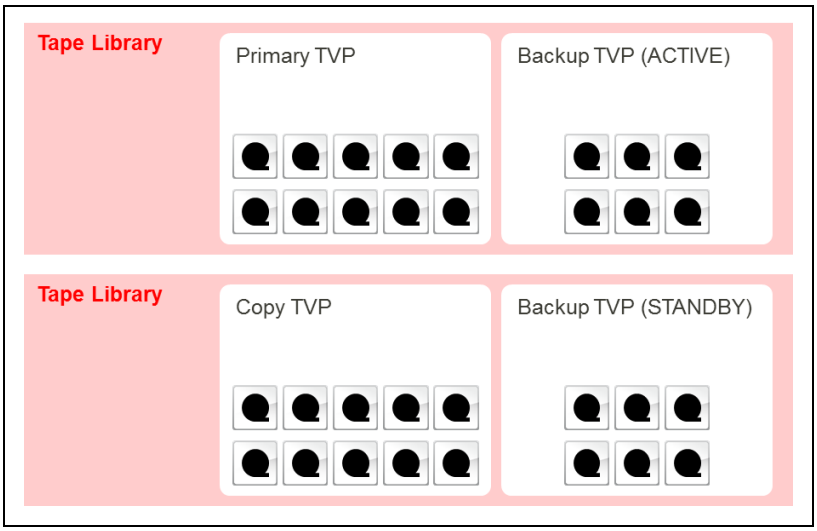


Figure 74

For availability reasons, primary and copy TVPs should be distributed over different tape libraries. In case of a library failure, a second copy of the data remains accessible. It may also be useful to define an additional "standby" backup TVP as shown above. This TVP is only used for backup purposes if the "active" backup TVP becomes unavailable.

There is an important difference between primary and copy TVPs. While HSM copies of files can be written to and read from the primary TVP, the copy TVP can only be used for reading.

Tape volume pool groups:

A tape volume pool group (TVPG) is an ordered set of one or two TVPs. If the group consists of only one TVP, this must be a primary TVP. If the group has two TVPs, they must be a pair of a primary and a copy TVP. A TVPG can thus be understood as a "single-save" or as a "dual-save" TVP group.

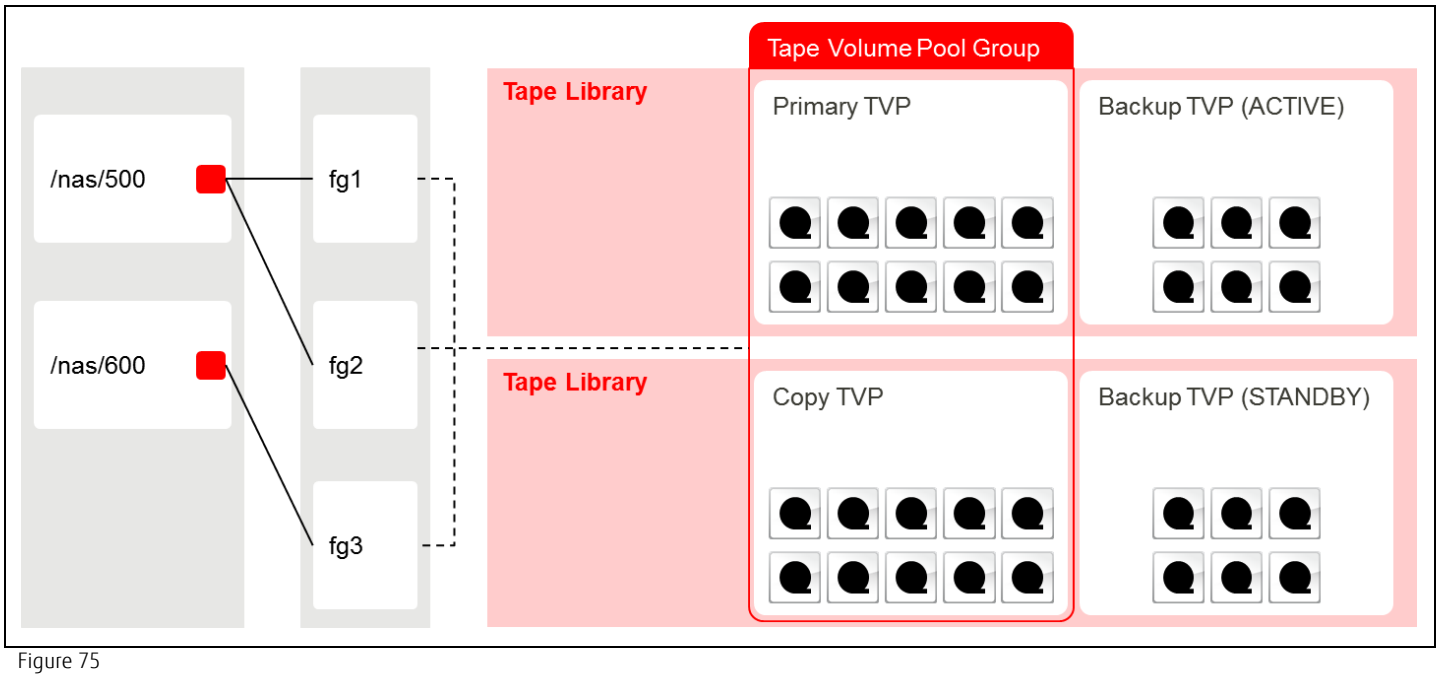


Figure 75

As shown in the example above, a TVPG serves as the HSM storage location for one or more file groups. The figure illustrates a dual-save TVPG which stores the HSM copies of three file groups contained in two HSM file systems.

5.5 HSM Data Lifecycle

CS ViNS primarily aims at supporting backup and archiving environments. These are characterized by high storage capacity needs and a dominance of write accesses over read accesses. In the case of archiving, long-term storage and special protection requirements are important.

The data stored in a ViNS system go through a lifecycle which can be described as follows:
1. Data are stored in file groups by backup or archiving applications running on NAS clients.
2. Data may be designated as write- and delete-protected while being stored.
3. If stored in a HSM file system, data are premigrated or migrated according to the configured rules.
4. If a backup is restored or if data are retrieved from an archive, transparent recalls are triggered if the requested data have been migrated to tape.
5. If any stored data are no longer needed, they are deleted from the file system.

5.5.1 Storing and Protecting Files

Files are written to CS ViNS file groups over standard NAS interfaces. If files need to be protected, they must be stored in "protectable" file systems, and their access times must be set to a protection end date in the future. Thus they are made immutable until the protection period is over.

The premigration and migration of files are tasks executed by the ILM component. ILMC functionality comprises the following:
- Periodic premigration and migration
- Truncation due to disk space shortage
- Migration due to disk space shortage (NoSpace Migration)
- Protection against system overload

The periodic migration and the truncation are the usual actions to free space in the NAS file system.

5.5.1.1 Periodic premigration and migration

The following can be specified for each file group:
- whether premigration should take place periodically (default), or
- whether migration should take place periodically, or
- whether no periodic premigration or migration should take place.

To permit the periodic premigration and migration jobs to be started, each file group is scanned periodically in order to select resident data with a change date which is older than "Save Delay" (in the case of premigration) or "Period on Disk" (in the case of migration). If the periodic scan detects new or changed data, the file names are entered into a list with which a premigration or a migration job is started. The file group scan is started when:
- the previous HSM job has been terminated, and
- no restrictions exist because of resource consumption, and
- the "Save Delay" or "Period on Disk" time has elapsed since the last scan was started. The time to wait between scans self-adjusts automatically based on whether the previous scan detected new or changed data.

The periodic scans over the file groups also save a list of all files which have already been premigrated in the Scan List Depot (SLD).

5.5.1.2 Truncation due to disk space shortage

When the file system's occupancy level reaches the configured value _THRESHOLD_HIGH, truncation is initiated for the file system. The file lists of the premigrated files which are available in the SLD are used for this purpose. Lists of files that are older (M1) or younger (M2) than the "period on disk" are compiled for each file group. First the M1 lists, and then the M2 lists, are used for truncation until as many files have been selected as necessary to fall below the _THRESHOLD_LOW occupancy level.

The truncation procedure does not affect resident files; therefore the _THRESHOLD_LOW value cannot be reached if the size of the premigrated files in the SLD is sufficient.

5.5.1.3 Migration due to disk space shortage (NoSpace Migration)
When the occupancy level reaches the _THRESHOLD_HIGH value but the truncation procedure finds no premigrated files in the SLD, or when the file system's occupancy level exceeds 95%, NoSpace Migration is initiated for the file system. This event should not occur under error-free backend operation and correct frontend dimensioning (Frontend Backend Sizing).

Under these circumstances ongoing periodic premigration, migration and truncation jobs are canceled. A NoSpace Migration scan is started for each file group to determine its current status. All files are categorized in 4 classes:
- **M1:** premigrated files which are older than the period on disk,
- **M2:** premigrated files which are younger than the period on disk,
- **M3:** resident files which are older than the period on disk,
- **M4:** resident files which are younger than the period on disk.

The M1 through M4 file lists are used for migration in the following way (migrating resident files or stubbing of premigrated files):
- The target for NoSpace Migration is determined. Files from M1 and M2, in that order, are selected for truncation. If this is not enough to fall below the target, files from M3 and M4, in that order, are selected for migration.
- Migration jobs of classes M3 and M4 are given priority on the backend. Any ongoing HSM jobs of all other file systems which are currently not performing NoSpace Migration are cancelled if they are using the same HPLS as would be needed by a started M3 or M4 job.
- After the NoSpace Migration operation has been completed, the periodic premigration / migration and truncation operations are resumed.

5.5.1.4 Protection against system overload
Because resources are required both to compile file lists (scanning) and for the HSM jobs, certain limits must be set on ongoing operations to prevent system overload.

The configuration parameters _SCAN_FS and _NUM_PARALLEL_HSMCALLS allow limiting the maximum number of scan and HSM jobs which run in parallel per file system.

The number of file group scan jobs per node (_SCAN_NODE) which are running at the same time is also taken into account. By default a maximum of 6 scan jobs are allowed to run in parallel.

The truncation jobs running in parallel are also limited for each file system (_TRUNC_FS) and for each node (_TRUNC_NODE).

The order of file groups is determined at ILMC startup and they are administered as having equal rights without a priority. When the ILMC is started, all file group scans within the first hour are started with a time delay. This time delay is at most 60 seconds in the case of a few file groups (< 10) and one second in the case of a large number of file groups (> 1000).

5.5.2 Reading Files

If a NAS client reads data from resident or premigrated files, the requests are immediately satisfied from the RAID storage.

If a file has been migrated, the file system contains merely the file's inode, but not its data. The HSM component intercepts the read request and obtains the data from background storage. The NAS client is not aware of this procedure, but its read access may take longer compared to the case where the data would be read from disk.

From the HSM component's viewpoint, the state of the file changes from "migrated" to "premigrated". The file remains in this state until it is either modified or selected again for migration by the ILM component.

5.5.3 Cleaning up Obsolete HSM Copies

HSM copies become obsolete in two cases:
- A premigrated or a migrated file is written. In both cases, the file becomes resident. The HSM copy contains obsolete data and can be no longer used.
- A premigrated or a migrated file is deleted.

Obsolete HSM copies are cleaned up by reconciliation jobs within the HSM service as a periodic ViNS housekeeping job. Reconciliation identifies obsolete tape copies and takes care of freeing the corresponding space on the tape volumes.

5.6 Multiple HSM services

More than one HSMS may exist in a CS ViNS cluster. The available HSM services are preconfigured based on the installed licenses.

5.6.1 Distributing HSM Service Resources

The HSM responsibility for a given NAS file system is assigned to a HSMS at configuration time. Each HSMS thus manages a dedicated subset of the configured NAS file systems. The association between a HSM service and a NAS file system cannot be modified once it has been set up. In the same way, each HSM service is assigned a subset of the available tape libraries for HSM usage. The partitioning of resources in an example configuration is depicted in the following diagram:

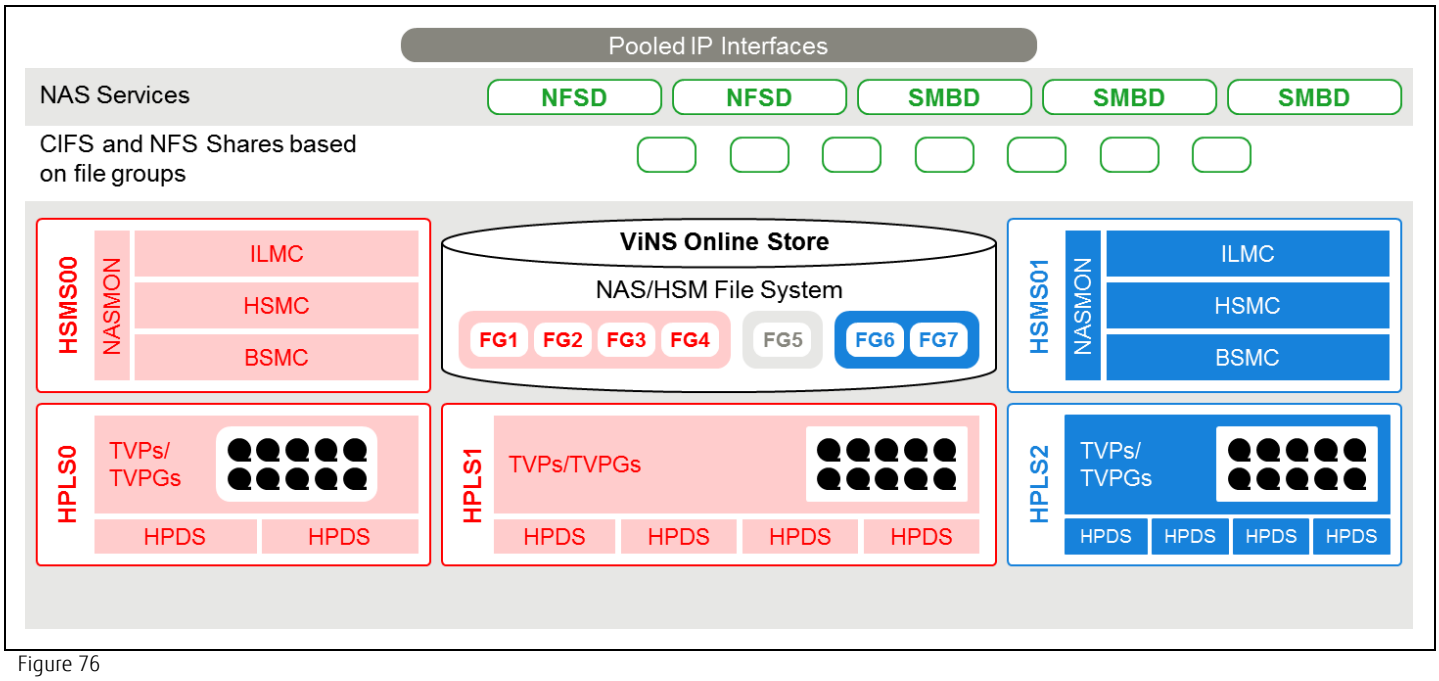


Figure 76

The diagram shows two HSM services HSMS00 and HSMS01. The resources owned by either service are shown in corresponding colors (red or blue). The tape libraries HPLS<i> consist of tape drives (HPDS) and TVPs. TVPGs may span libraries, but they belong to a single HSM service.

The file system with file groups FG1-4 belongs to HSMS00, while the one with FG6-7 belongs to HSMS01. The file group FG5 is part of a file system which is not HSM managed. To complete the picture, the ViNS frontend components (shares, NAS services and pooled IP interfaces) are shown as well.

5.6.2 Library Client/Manager Configurations

A BSMC may only access a single ACS host. Hence it is not possible to set up a TVPG with a primary and a copy TVP in two different ACS systems. This case can be handled by creating a HSM service configuration with shared tape libraries.

If a HSM service has been configured e.g. with an associated library of type STKCSC, it may use another STKCSC library which is owned by a different HSM service. The HSM service which "exports" the library is called the library manager, while the "importing" HSM service is the library client. Note that the library manager/library client role of a HSM service applies to a given library, not to the HSM service as a whole. In the ViNS Backend/HSM Physical Tape Libraries configuration dialog, a tape library of type LINK must be defined as owned by the library client. This library must have a reference to the exported library.

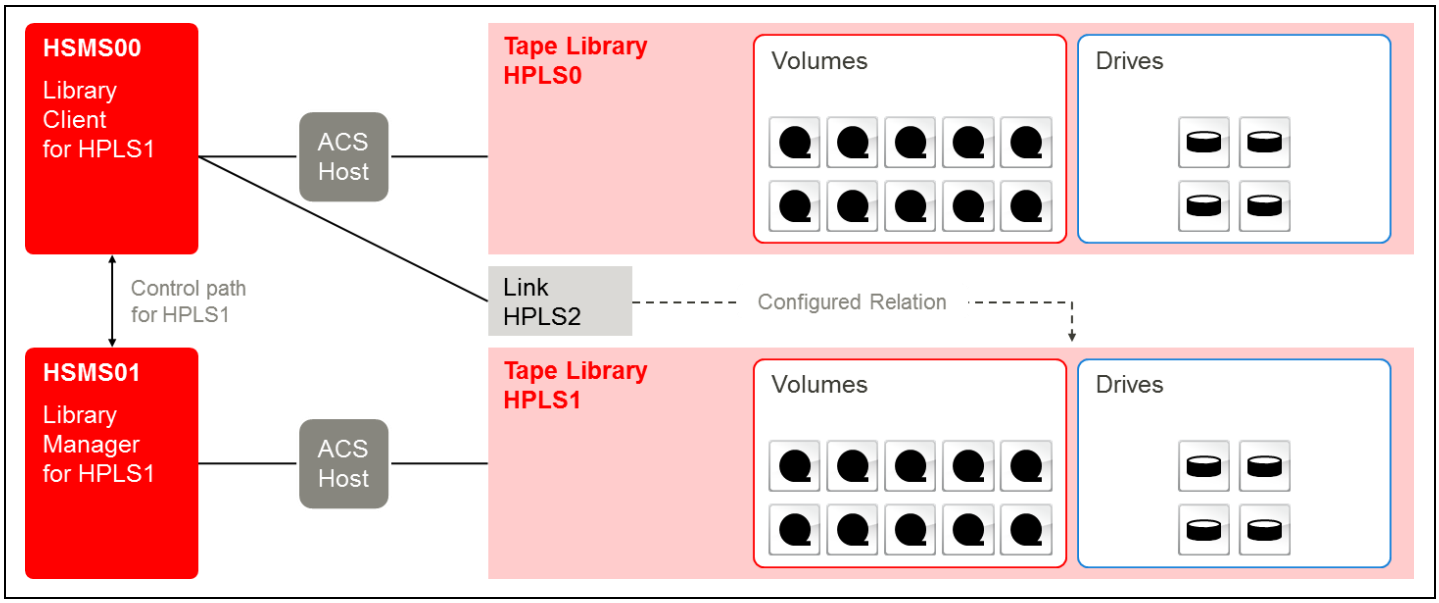


Figure 77

The diagram illustrates an example of a library client/manager configuration. The HSM service HSMS01 (on the right side) exports the library HPLS1 for usage by a library client. This library client is the HSM service HSMS00 (on the left side). HPLS2 is configured as a tape library of type LINK, and as such, it has a reference to HPLS1. This setup takes care that HSMS00 can use HPLS1 by forwarding the control commands for HPLS1 to HSMS01. Logically, HSMS00 can now access two STKCSC libraries over different ACS hosts.

Library client/library manager configurations follow a set of rules:

- A library of type LINK has access to all drives of the referenced logical library.
- If a logical library is exported to a library client, this export is "exclusive", i.e.
  - A given logical library may only have one referencing library of type LINK.
  - A HSM service which exports a library must not use the same library for its own HSM.
- A HSM service may be designated as "library manager only". Its role is then "Active_BSM". Such a service exports all of its libraries to library clients, and does not itself perform HSM for any file systems.

5.7 HSM Services with Loopback to CS VTL

Just like the behavior with physical tape, a HSM service can be attached also to the CS VTL virtualized tape interface. The CS VTL subsystem resources such can be used within the unified architecture as resources from a HSM service in the same cluster. Such a setup is called a "loopback configuration".

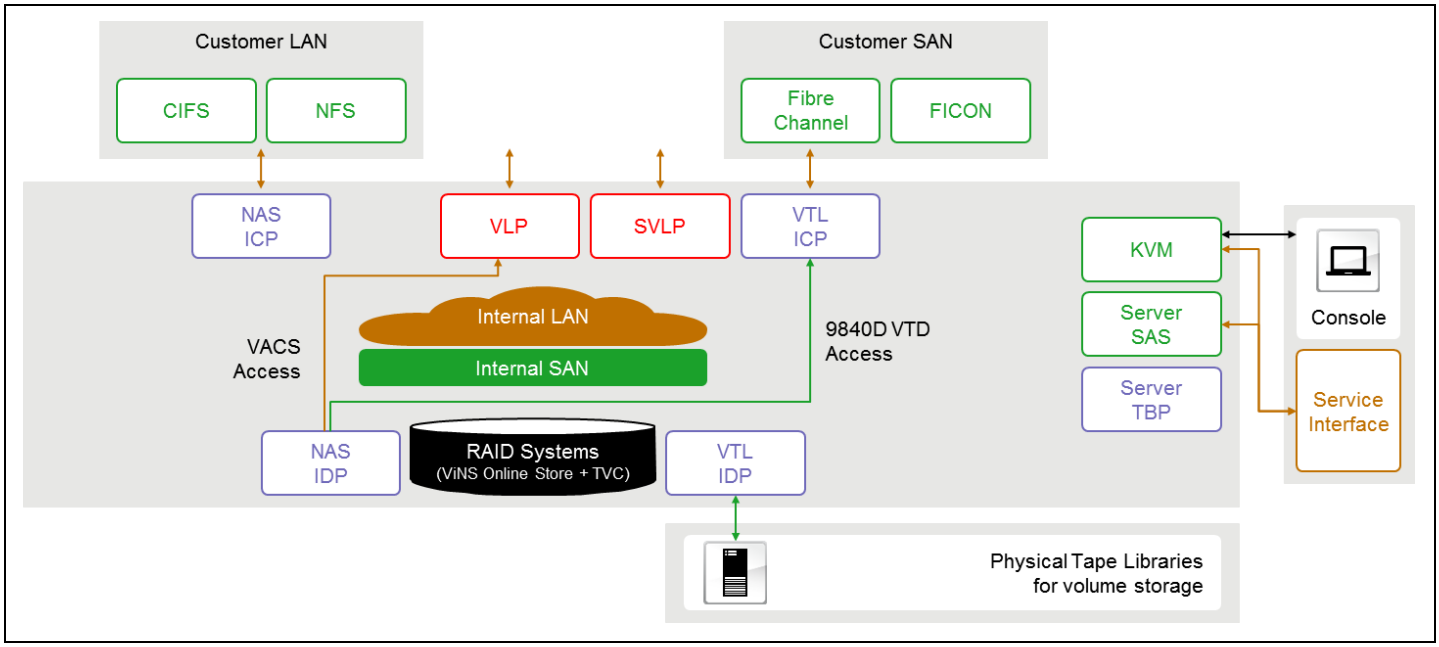


Figure 78

The external FC connections of the NAS IDP are connected back into the external FC connections of the CS VTL ICP. Likewise, the VACS on the master VLP is accessed by the NAS IDP over the appliance internal LAN. The available types of loopback devices are restricted to virtual ACS libraries and virtual drives of type T9840D. In other words, the HSM service uses a VACS as one or more tape libraries ("loopback HPLSs" or "loopback VTLs") and T9840D VTDs as tape drives ("loopback HPDSs" or "loopback VTDs").

5.7.1 Mapping of HSM Service Objects to CS VTL Objects

Mapping of the objects of the HSM service to objects of the VTL is illustrated in the following diagram:

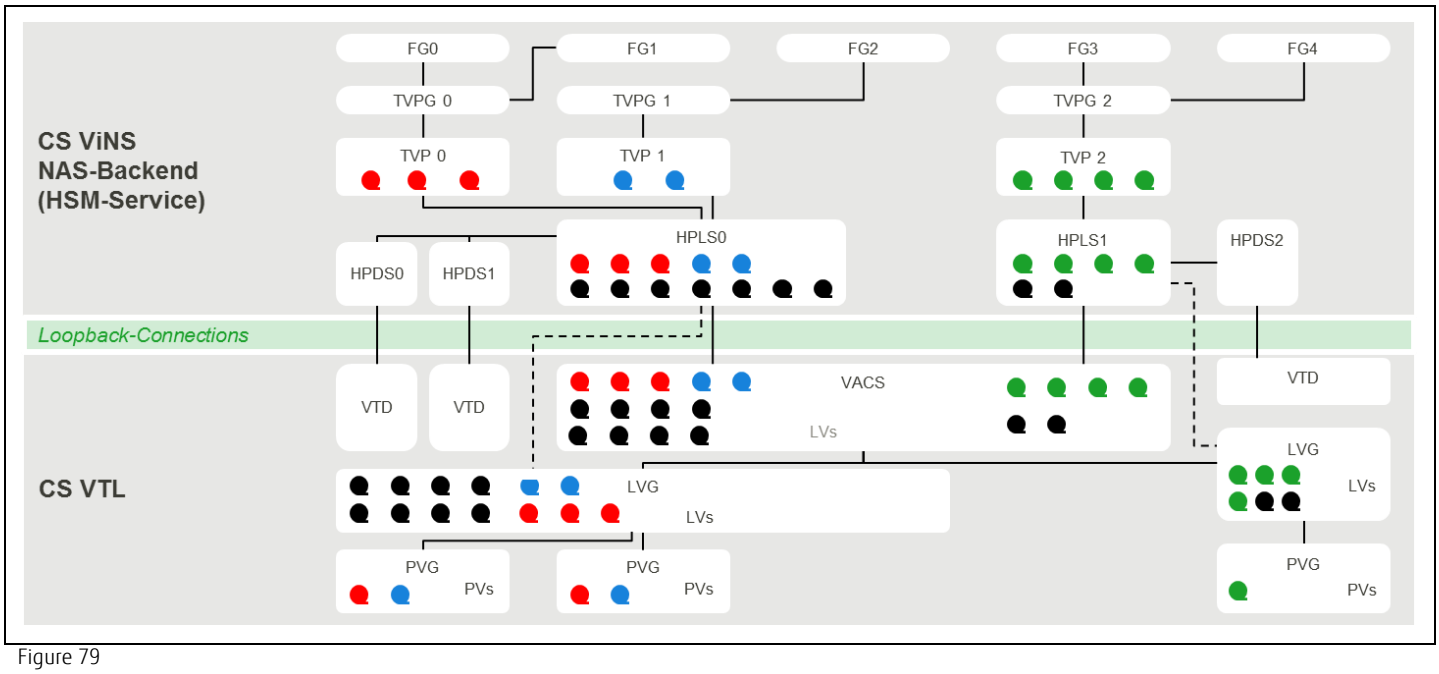


Figure 79

HSM managed files are organized in file groups (FG0-FG4). A file group is associated with a storage location that holds the HSM copies of files in the group. The storage location is a tape volume pool group. In the above picture, FG0 and FG1 files are stored in TVPG0; FG2 files are stored in TVPG1 etc. TVPGs for libraries connected via loopback must consist of only a single tape volume pool (TVP). Copy TVPs are not needed because the replication of HSM copies is handled at the CS VTL layer. This is a major advantage of loopback configurations.

In the example, TVP0 (red) has the HSM copies of FG0 and FG1 TVP1 (blue) has the copies of FG2 and TVP2 (green) holds the ones from FG3 and FG4. Tape volume pools belong to tape libraries. TVP0 and TVP1 are both contained in library HPLS0 while HPLS1 hosts the volumes of TVP2. The gray volumes indicate the scratch pool of the respective library.

The libraries and drives used by the HSMS are mapped to CS VTL objects of the same ETERNUS CS8000 cluster. HPLS0 and HPLS1 use separate sets of drives in the same VACS service. Each drive (HPDS object) on the HSMS side is associated with exactly one logical drive (VTD object) of CS VTL. Within CS VTL, the drives must be configured with type STK T9840D.

The logical volume groups of the VTL reflect the grouping of volumes on the HSMS side: The LVG on the left hand contains the volumes of HPLS0. It is associated with two PVGs so that all HSM data of files stored in HPLS0 have two tape copies. The LVG on the right- hand side is a single-save LVG, and it contains the volumes from HPLS1.

The key point in the diagram is the relation between HPLSs and LVGs: A HPLS object which is implemented by a virtual library accessed via loopback has a 1:1 relation with a logical volume group on the CS VTL side. The LVG represents the set of volumes available to the HPLS.

5.7.2 Availability Considerations

The usage of a loopback CS VTL has several advantages with respect to high availability because the HSM service profits from the HA properties of the CS VTL which offers transparent single-save or multiple-save capabilities for its users. This means that primary tape volume pools within a loopback CS VTL do not need associated copy pools to hold multiple HSM copies of the same data.

The absence of copy pools has important advantages:

- In a multi-site configuration, primary and copy TVPs for HSM data would normally be distributed over the two sites. Therefore the data center SAN needs to span both sites. If there is no need for copy TVPs, the requirement for a site-spanning SAN does not exist.
- There is no need build library client/manager configurations to allow an HSM service to access more than one ACSLS library.
- No HSMS housekeeping jobs for synchronizing the contents of primary and copy TVPs are required.

5.8 HSM Service Availability

CS ViNS implements a two-stage monitoring of HSM services:
- Node-local monitoring of a HSMS
- Monitoring of an active HSMS by a standby HSMS on a different node

Monitoring is based on periodic executions of check commands for resources used by a HSM service. If a check fails, there is an associated action to report and/or fix the problem.

5.8.1 Node-local-monitoring

The node-local monitoring functions executed by the NASMON cover state checks of:
- HSM-managed NAS file systems
- HSM metadata file systems
- Tape libraries
- Tape drives
- Tape volume pools
- Scratch volume pools
- Tape volumes
- HSM service components (ILMC, HSMC, BSMC)
- ViNS backend housekeeping jobs

5.8.2 Monitoring of an Active HSMS by a Standby HSMS

For automatic failover of a HSMS from a failed NAS IDP, a CS ViNS cluster may be equipped with a standby NAS IDP. The standby node provides HSMS availability in case of node failure or severe HSMS failures. The standby NAS IDP runs a NASMON component configured for remote monitoring, which periodically checks the availability of the observed NAS IDPs. A standby NAS IDP needs access to the same hardware resources (tape libraries, tape drives) as the monitored active NAS IDP.

The following example assumes that the NASMON of HSMS00 monitors the active HSM service HSMS01.

The NASMON checks for the following situations:
- node failure of the active NAS IDP1
- external LAN failure of the active NAS IDP1 (in case the monitored node has access to an ACSLS library)
- failure of the HSMS management on the active NAS IDP1 (while the node itself remains up and running)

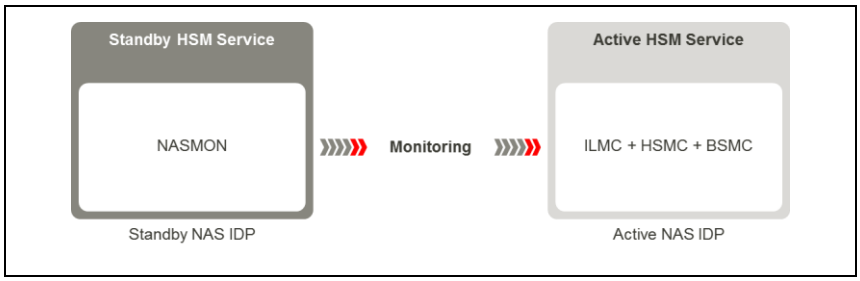


Figure 80

If any of these conditions occur, the standby HSMS executes a HSMS failover:

The standby HSMS initiates a role change between standby NAS IDP0 and the monitored NAS IDP1. The active HSMS01 is placed on the standby NAS IDP (which becomes an active NAS IDP), and the former active NAS IDP takes on the standby role.

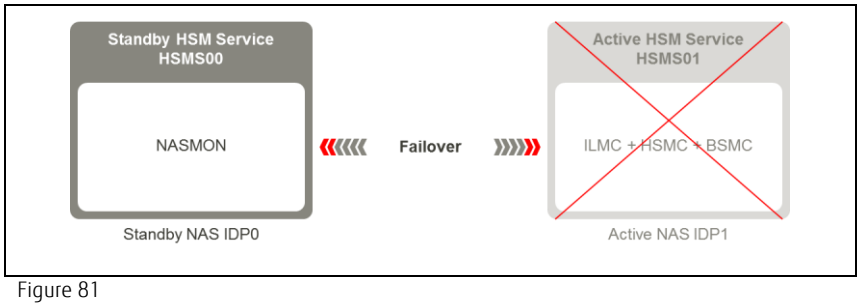


Figure 81

The failed node NAS IDP1 (now a standby NAS IDP) enters the "maintenance state". In maintenance state, the node does not immediately run a NASMON, so it does not monitor its new active counterpart. The maintenance state is expected to be lifted by the administrator when the failed node or HSMS has been repaired.

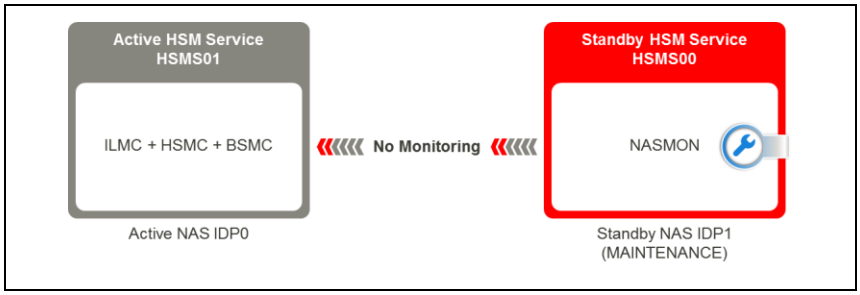


Figure 82

There is a difference between active/standby HSM services and active/standby NAS IDPs. The active or standby role of a HSM service is fixed in the CS ViNS configuration. In contrast, the active or standby role of a NAS IDP changes whenever a failover occurs which concerns the NAS IDP. In other words, a NAS IDP becomes active resp. standby when an active resp. standby HSM service is assigned to it as a result of a HSM service start or a failover.

A standby HSMS (i.e. a NASMON) can monitor more than one active HSMS. When any of the monitored HSM services fails, it is taken over by the standby node. Each standby HSMS monitors its own group of active HSMSs. However, two different standby HSMSs cannot monitor the same active HSMS.

CS ViNS supports the following redundancy configurations for a given number N of active NAS IDPs (N<= 4):

- **N+1 redundancy:** One standby HSM service monitors one or more active HSM services.
- **N * (1+1) redundancy:** There are one or more pairs (up to 4) of active and standby NAS IDPs.

In multi-site configurations, it is preferable to set up monitoring between standby and active HSM services such that a standby HSM service on the local site monitors an active HSM service on the remote site. Consider the following example for 2*(1+1) redundancy:

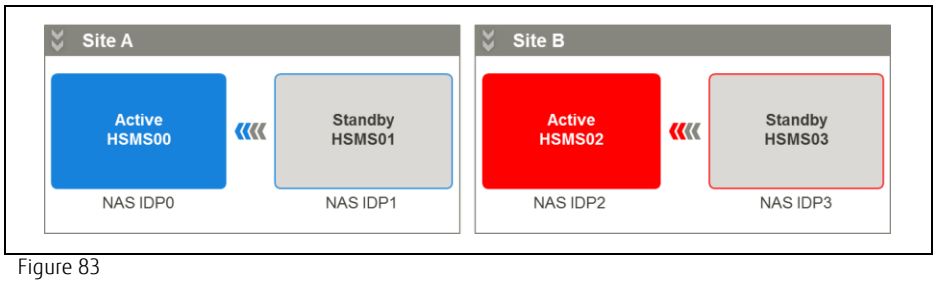


Figure 83

HSMS01 monitors HSMS00, and HSMS03 monitors HSMS02. This configuration is safe against a NAS IDP failure, but it is not safe against a site failure. For example, if site A becomes unavailable, there is no HSM service which can perform the hierarchical storage management for file systems managed by HSMS00.

It is therefore better to arrange the monitoring as in the next diagram:

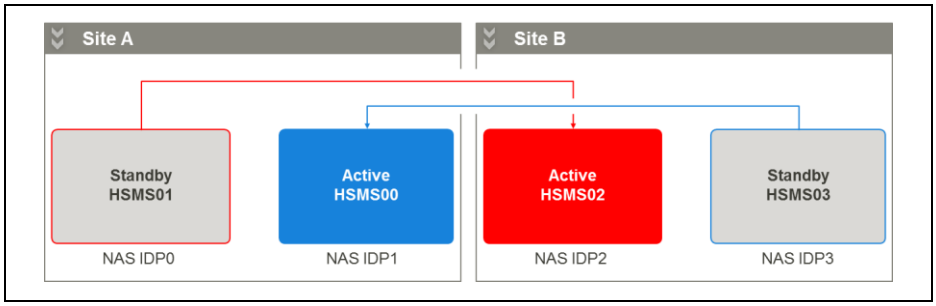


Figure 84

This setup ensures that two HSM services remain active even in the case of site failure. If site A fails, then the active HSM service HSMS00 is taken over to NAS IDP3 on site B.

5.8.3 Failure Scenarios

The following list describes the necessary measures to protect the HSM service against certain failures to access the tape devices used by the ViNS backend. Some scenarios differ depending on whether the service is used with a physical tape library or with a loopback library.

- **Failures of connections to tape libraries or drives:**

  A connection to physical devices fails: If an active NAS IDP is monitored from a standby NAS IDP, a failure of the network connection from a NAS IDP to an ACS host triggers a switch of the active HSM service to the standby NAS IDP. A failure of an FC/SCSI link to a tape drive or a SCSI library is not automatically handled.

  A connection to loopback device fails: Network links to the VACS server are redundant because they use the ETERNUS CS internal LAN. The failure of a single link is handled by the operating system. If the VTL configuration of a loopback library distributes the virtual drives of the library onto different controllers and/or ICP nodes, then the library remains usable after failure of a single FC/SCSI link between the NAS IDP and a VTL ICP.

- **Failures of tape volumes:**

  A physical tape volume fails: Dual-save tape volume pool groups must be used for protection against volume failures. If the primary copy fails, data is still available from the secondary copy.

  A virtual tape volume in a loopback library fails: Dual-save tape volume pool groups are not used. Volumes are kept in the TVC. Redundancy of volumes is achieved by mirroring TFSs and by defining dual-save LVGs to store the logical volumes containing HSM data.

- **Failures of tape drives**

  Tape drive failures are handled by the BSM component. The containing HSM service does not use a failed drive any longer.

- **Failures of tape libraries**

  A physical tape library fails: For protecting against library failures, it is recommended to use dual-save tape volume pool groups with primary and copy TVPs in different libraries. If a library fails and some primary TVPs are no longer available, it is possible to switch the primary TVPs in other libraries.

  A loopback library fails: A library failure is equivalent to a VACS or VLP failure. The HSM service should be run in a 2-site configuration with VLP failover. A VACS failure leads to an automatic VACS restart. A VLP failure is handled by the VLP failover function. The VACS library is switched to the standby VLP.

- **Failure of a HSM service or a NAS IDP**

  The HSM service should be run in a 2-site configuration with an active service on one site and a standby service on the other site. In case of a failure, the active HSMS is switched to the standby node.

- **Failure of an ETERNUS CS8000 site**

  A site failure is a combination of the failure cases described in the preceding scenarios. Consequently, the precautions against site failures consist of the superset of the described measures.

5.9 HSM Backend Housekeeping

Housekeeping is the generic term for actions which are performed at regular intervals in order to support smooth execution of hierarchical storage management. Such actions include:

- reorganizing resources
- initiating backup operations

The actions described in this section should be performed once a day. Certain activities are interdependent, which means that actions must be executed in a predefined order. The following actions should be executed at least once every 24 hours:

- **Reconciliation:**
  Comparison of the content of the NAS file systems with the content of the volumes in the background storage.
- **Expiration:**
  Release of the storage space on the tapes which is no longer required after its expiration time has elapsed.
- **Reclamation:**
  Reorganization of the fragmented storage space on the tapes depending on the percentage of user data.
- **Backup of the primary tape volume pool:**
  Comparison of the primary tape volume pool with the assigned copy pool.
- **Disaster backup:**
  Backup of the metadata of the NAS file systems and the data management system of the BSMC to tape.

The housekeeping actions are executed automatically on a periodic basis. A number of commands permit administrative intervention in the procedures or in the status and the history of various actions to be displayed.

5.10 Archiving Applications

Archiving is about assuring the retention and authenticity of information that will be needed at some indefinite point in the future. Organizations retain information for two reasons: because the information has some inherent business value or because they have to do due to legal requirements.

The process of retaining information, more commonly referred to as "archiving," is challenging for several reasons:

- Companies continue to generate new information, at least a subset of which the business wants or has to save.
- Information is generated by multiple applications in multiple formats. E-mail, instant messages, productivity files, database records, SharePoint sites, wikis, and videos are just a few examples of information types that may need to be archived.
- Because of the importance of the information being archived, organizations have to treat it as "mission critical" and put proper data protection and business continuity processes in place to prevent downtime or data loss.
- Companies have to secure archives in accordance with industry- or government-specific privacy regulations.

CS ViNS is a high scalable, flexible archiving solution designed to help organizations of all sizes address their complete information retention needs, whether business, legal or regulatory. CS ViNS stops the cost explosion which is caused by keeping all archived data in near-line storage for the duration of its retention period or for long-term archiving. CS ViNS is designed to support data migration to lower storage tiers (from disk to tape) with policy-based and hierarchical storage management (HSM) capabilities. The integrated support of mixed media storage for both disk and tape helps to lower the total cost of ownership. A direct match can be achieved between the value of archived information and the cost of the storage media on which it is stored.

5.10.1 Certified Applications

CS ViNS supports generally all applications, which are able to write their data to a NAS file system. As a result, regardless of the archiving application, CS ViNS can be used to archive, manage and secure critical business information from a single point of control.

The following Archiving applications have been certified for CS ViNS:

- Heilig & Schubert - H&S PAM-HSM
- Metalogix HSM
- CommVault Simpana
- Symantec Enterprise Vault
- Fujitsu SecDocs

5.10.2 File Protection Period

CS ViNS allows specifying a protection period for regular files in a NAS file system whose attribute "Protectable" was set in the configuration. The file remains protected against modification and deletion attempts until the protection period expires. Protection periods can be set through both NAS interfaces (CIFS, NFS) as well as cluster-internally by respective administration commands.

If a file has been protected, operations on the file behave as described in the table.

Operation	Is the operation allowed on a protected file?
Modify the contents of the file	No
Rename the file	No
Delete the file	No
Add, modify or delete extended file attributes	No
Extend the contents of the file (*append operation*)	No
Modify atime, mtime or ctime	Yes
Modify the file mode bits	No
Modify the file owner and/or ACL	No
Set a hard link to the file	Yes

A NAS client may set the protection period through "ordinary" file system calls. Furthermore, there are simple administration commands for setting, deleting and querying file protection attributes. The functions of the protection period implementation in CS ViNS are summarized in the following diagram:

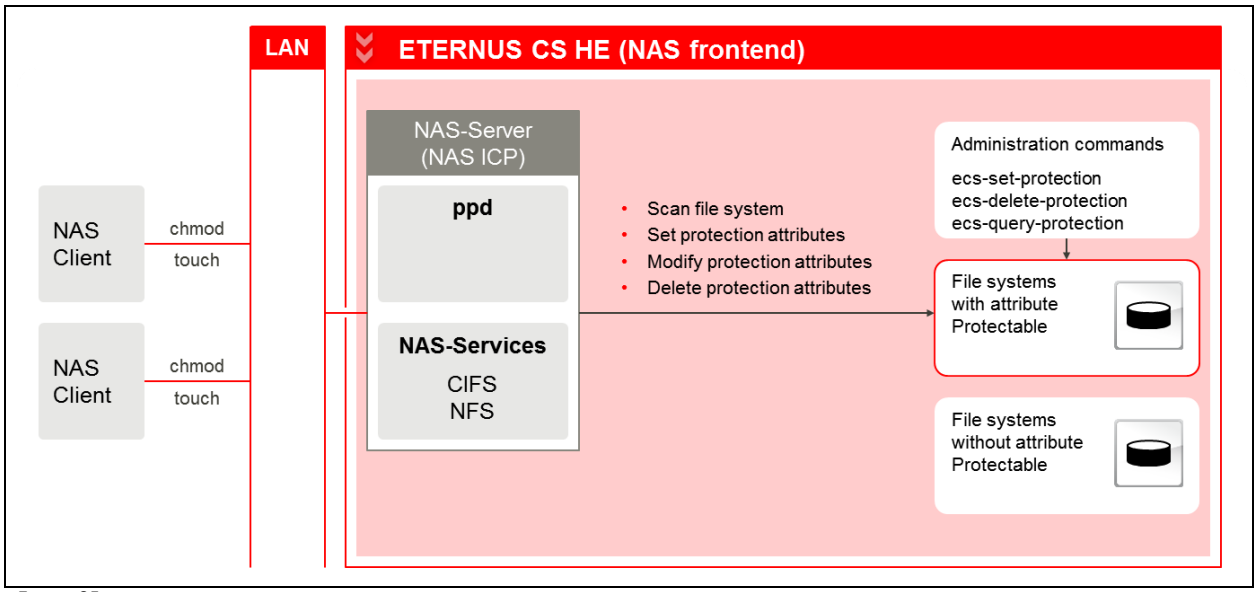


Figure 85

In order to set the protection attributes, these file systems are periodically scanned by the ppd (protection period daemon process). ppd executes periodic scans on all NAS file systems with the Protectable attribute. The scans identify files for which a protection period is to be established, modified or deleted.

5.10.3 File Protection and HSM

There are no HSM specific restrictions for protected files. In other words, protected files can be premigrated, migrated and recalled from/to tape volumes just like unprotected files.

However, protected files should be (pre)migrated to WORM media. Otherwise there is a chance to manipulate the file contents by directly writing to the tape volume. However, the protection period implementation does not enforce that HSM copies of protected files can only be created on WORM volumes. ppd merely logs a warning if the CS ViNS configuration permits files from file systems with the Protectable attribute to be (pre)migrated to non-WORM tape volumes.

In case of a disaster e.g. the loss of data from the RAID system, the HSM copies of files on tape volumes can be used to recover the lost data. The data recovery from the HSM copies take care that files which were protected before the disaster will again be protected after the disaster with the same protection expiration date/time as before.

6 Managing the Data Protection Appliance

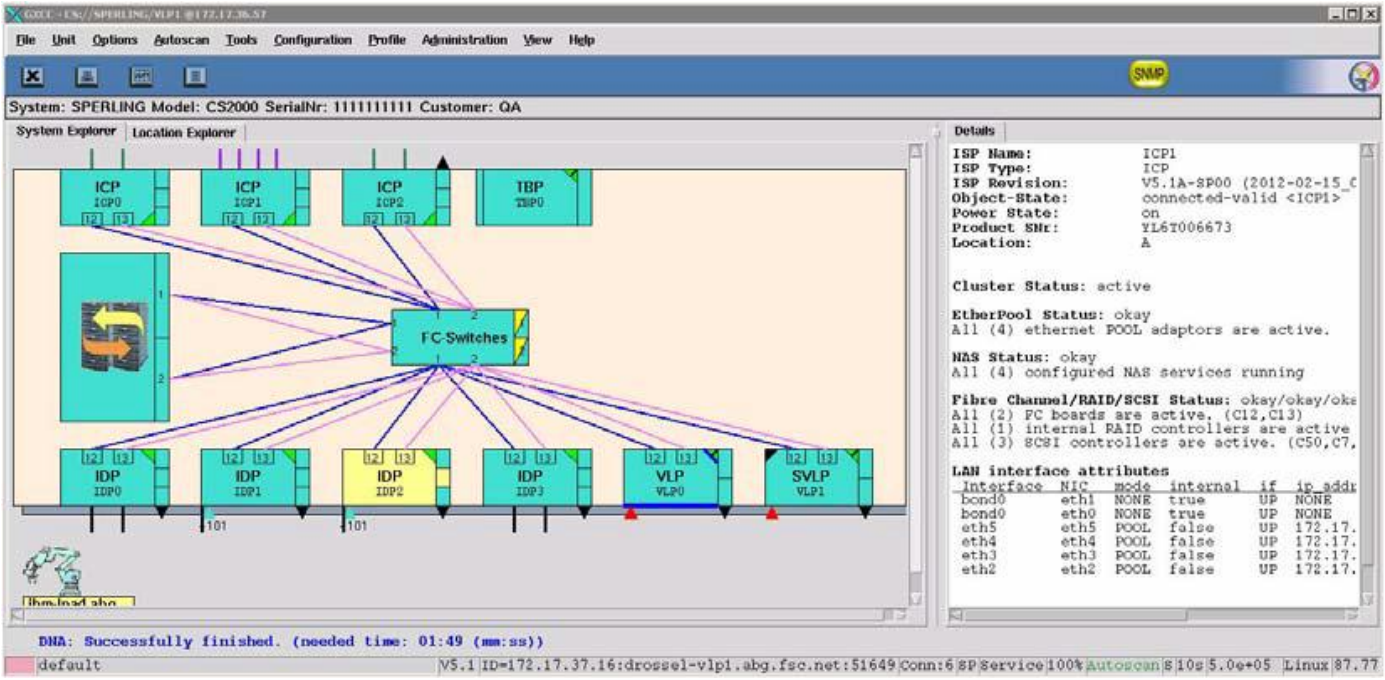
6.1 User Interfaces

For operating and monitoring different user interfaces are provided:
- Console graphical user interfaces - Tcl/Tk GUI (GXCC, XTCC)
- Web based graphical user interface (Web GUI)
- Command Line Interface (CLI)

An ETERNUS CS8000 can be configured not only via the Tcl/Tk GUI, but also via the Web GUI and the CLI. When more than one user is operating simultaneously over different user interfaces, only one of these users can perform a configuration. While the different user interfaces can be used for change of the configuration data, a synchronized access is guaranteed and which user interface is effective at any given time.

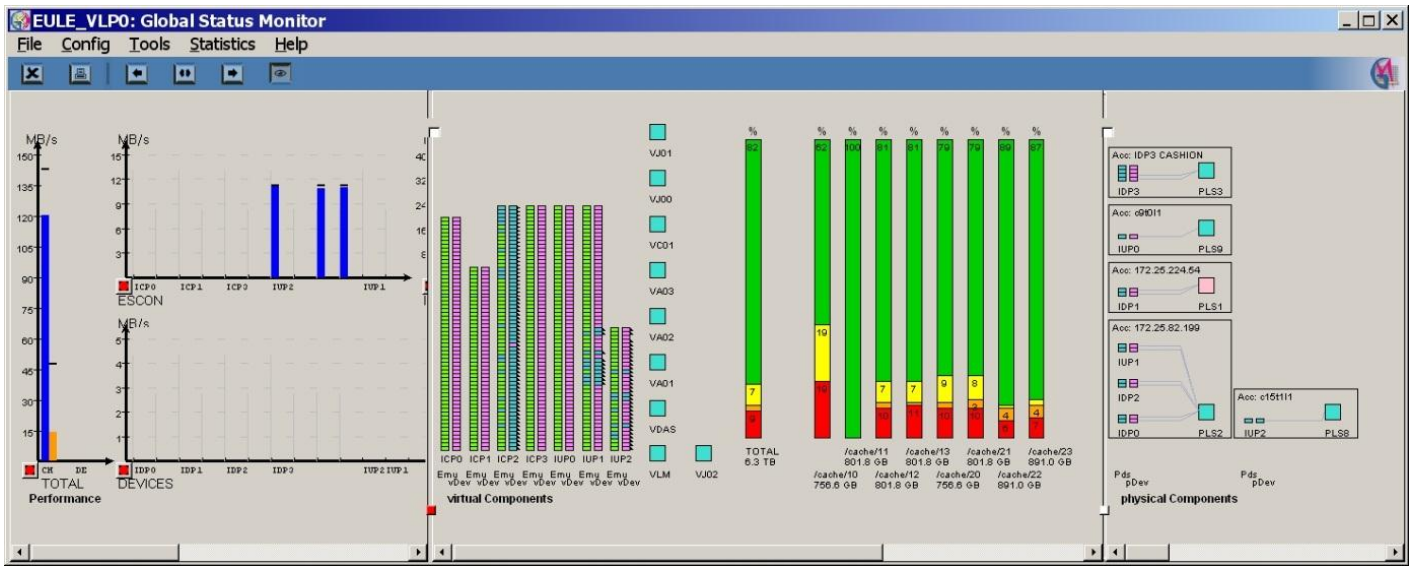
6.1.1 Console Graphical User Interfaces

The GXCC (Global Extended Control Center), Global Status and XTCC (Extended Tape Control Center) are graphical user interfaces running on the integrated console and are based on the Tcl/Tk (http://www.tcl.tk) scripting language and toolkit. The Tcl/Tk GUI supports functions relevant for the appliance administration as well for operating system specific activities. The console GUI provides a complete graphical representation of an ETERNUS CS8000 system, and covers all connected devices and ISPs such as ICPs, IDPs and VLPs. GXCC processes each ISP and other components of the cluster as if they were a single unit.



ETERNUS CS8000 includes a Performance Monitor, which constantly records and saves component operating data at intervals of a few seconds. The **Global Status Monitor** analyzes the performance data and displays all important operating and system data in graphical format over configurable periods. For CAFS file systems also some maintenance functions are offered depending on the respective state of these file systems.

The Global Status Monitor window consists of three vertically aligned sub windows which show (from left to right):
- the throughput through the host interfaces and the device controllers
- the status of the virtual components (drives, CAFS file systems)
- the status of the physical devices and physical control processes as well as data on the PVGs

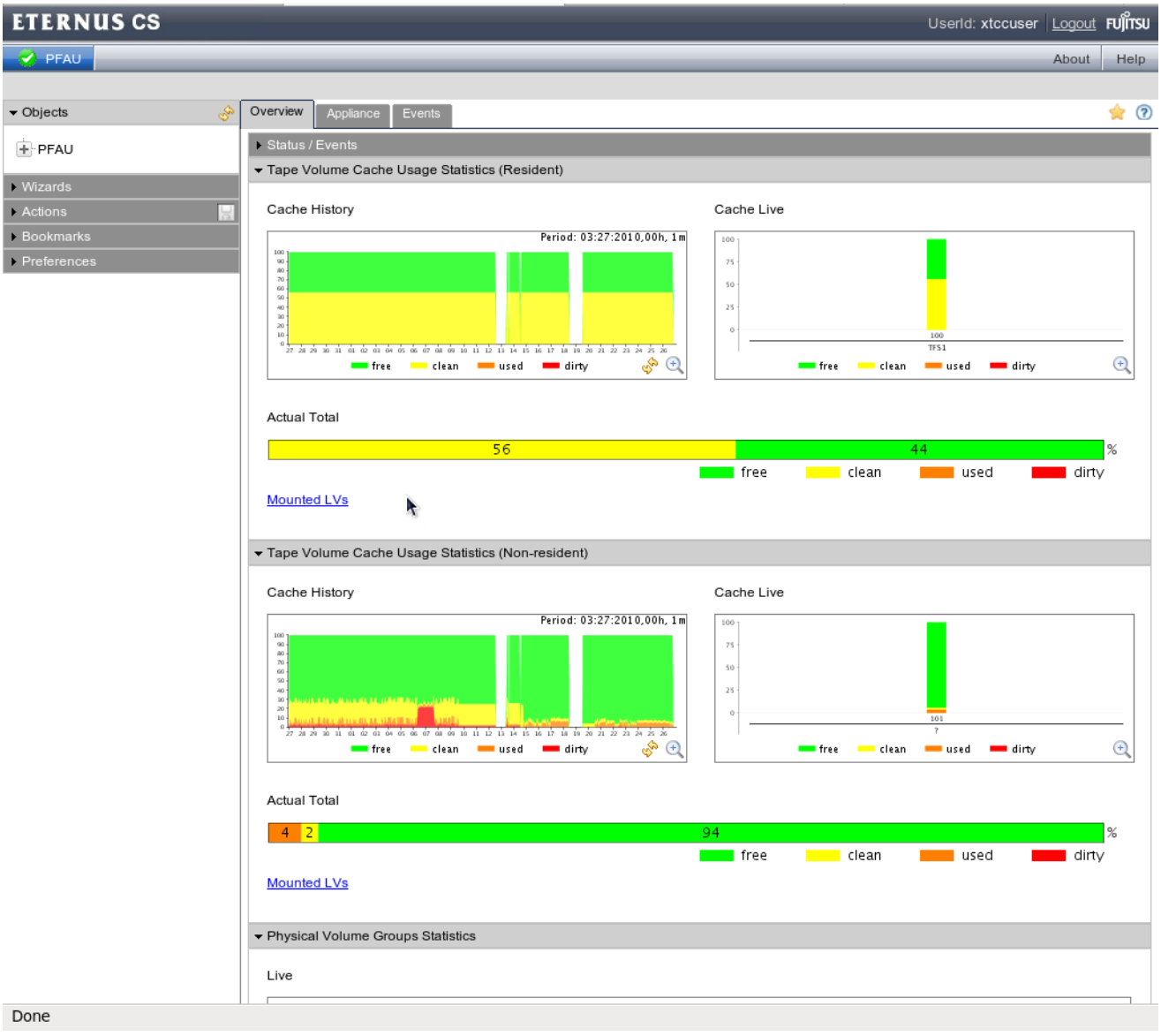


XTCC is an application for displays and operations within an ISP. While GXCC regards an ETERNUS CS8000 system as a whole, XTCC is limited to one ISP together with the connected peripherals. The XTCC is started for an ISP by positioning the mouse pointer in GXCC over the symbol of the ISP and right-clicking to select the Show Details (XTCC) option from the function menu.

GXCC, Global Status and XTCC are standard components of the appliance software package, and are installed on all the ISPs. Different to the Web GUI all ISPs (not only the VLP) of an ETERNUS CS8000 system can be accessed with the Tcl/Tk GUI from „outside“. Provided the ISPs have an external LAN connection GXCC/XTCC installed on SOLARIS, LINUX and WINDOWS workstations or on an SNMP system management station get access to these ISPs. The “Remote Management” function enables in this way one or more ETERNUS CS8000 system to be administered via the GUI from a remote installed console.

6.1.2 Web-based Graphical User Interface

The web based graphical user interface (Web GUI) offers all functions necessary for media management (i.e. LV, LVG, PV, and PVG) and some configuration functionality. The Web GUI can be used with a standard Web browser (Mozilla Firefox or Microsoft Internet Explorer) and does not need any further software. The web server components of the Web GUI are only started on the VLP. If a failover occurs, the Web GUI is reconfigured to the SVLP. Web GUI access is possible over all ISPs of an ETERNUS CS8000 but also over a PC independent and outside of the appliance. For communication between the client (browser) and the Web server, the customary HTTP connection and also the encrypted HTTPS (SSL) connection for secure communication is supported.



6.1.3 Command Line Interface

The Command Line Interface (CLI) is an integral part of the ETERNUS CS8000 software package and is installed on all ISPs. It can, also be operated on a PC independent and outside of the appliance via ssh command. Return codes also provide automatic access (scripting).

The CLI commands cover:
- Querying and statistics commands
- User relevant administrative commands
- Administrator specific commands
- VLM and PLM specific administration (VTL subsystem) especially for media management:
  - VLM administration: Add and delete logical volumes, set logical volumes into scratch state, fetch logical volumes into the cache (prefetch), list logical volumes.
  - PLM administration: Configure physical volumes, perform physical transfers, influence the priorities of PLM jobs, and query PLM information.
- ViNS HSMS management: Query information of the entire cluster about the HSM Service (HSMS) and about administrative actions in HSMS in an ETERNUS CS8000 with NAS interface and HSMS.

6.2 User Management Service

In ETERNUS CS8000 two types of user need to be managed:
- Administrative users
  - may log in to the system (via CLI or GUI),
  - need to be authenticated at the Linux operating system level,
  - are assigned some administrative role.
- NAS client users
  - may connect to NAS shares,
  - are authenticated by the NAS service (CIFS),
  - are mapped to local Linux users, which are then used to create/check ownership of file system objects (files, directories) and to check access to file system objects.

The User Management Service provides a common access method to user (and group) accounts throughout the cluster. User accounts are either taken from a locally managed repository, or - when explicitly configured - from an external directory.

The account management has knowledge of the ETERNUS CS8000 administrative roles and how these roles map to group memberships. It provides commands enabling system administrators to create, manage and delete internal user accounts.

For integration with external user account directories, the following services are supported:
- Active Directory:
  The User Management Service and Samba have access to the Active Directory. As a consequence, NAS client user accounts and eventually administrative user accounts may come from the Active Directory.
- LDAP based directory:
  The User Management Service has access to the LDAP (Lightweight Directory Access Protocol) directory; Samba does not use accounts from the directory. As a consequence only administrative user accounts may be kept in the LDAP directory.

The role concept is based on the UNIX/Linux group concept and permits ETERNUS CS8000 specific activities to be supported by person-related user accounts.

The following roles are defined:
- Operating System Administrator - Performance of administrative activities which relate to the underlying operating system
- ETERNUS CS8000 Administrator - Performance of appliance specific administrative activities which are relevant for appliance users
- Service - Performance of appliance specific service activities
- Observer - Observation of the appliance activities
- User Administrator - Management of user accounts
- NAS Client - Connection to NAS Services

6.3 Serviceability

Service Access System (SAS):

Monitoring of ETERNUS CS8000 (except CS50) takes place on the Service Access System (SAS). The SAS is running on a dedicated PRIMERGY server and monitors over the internal LAN all processor nodes and the RAID systems. SAS collects and consolidates important system messages and provides:

- System Logging
- Monitoring and alerts on different levels
- Call home
- Remote service access for maintenance and diagnostics

System Logging:

To support information system audits related to the ETERNUS CS8000 system operations, several system logging and reporting features provide the ability for examination of important system activities:

- Logging system processes and events
- Logging activities which are relevant to security
- Logging su (change the current user to the superuser the system)
- Logging authorization using the sudo command (certain actions which can only execute with root authorization)
- Logging ETERNUS CS8000 specific commands
- Logging activities of GXCC/Web-GUI sessions
- Logging the firewall activities

Alerts:

In the event of serious errors in ETERNUS CS8000 operation, the following measures for reporting are initiated automatically via various alarm paths:

- The "call home" feature is a communication link that is established between a product and a service provider. Fujitsu AIS Connect provides this feature so that reports can be automatically sent to the Fujitsu Support Center when severe problems are detected. The communication link is encrypted.
- "Hot Messages" to BS2000 hosts with ROBAR connection (ROBAR-SV as of V6.0 can be installed on the SAS computer. A LAN connection is necessary for communication between ROBAR-SV and ROBAR-CL which runs in BS2000.)
- Via a customer specific script in which it is possible to define specific reactions.
- Alerting by E-mail: For certain messages the administrator can configure the message management to execute the alerting action forwarding as e-mail.
- Alerts via SMS

The responses to the individual error events are preset for various service provider profiles. One of these can be selected. In addition, the selected default can be adjusted via the GUI on an individual basis.

Furthermore ETERNUS CS8000 can be integrated into SNMP remote monitoring. An integration package for Windows is available to activate the application launch connection especially for the "Unicenter" SNMP remote monitoring software from CA.

Remote service access, maintenance and diagnostics:

ETERNUS CS8000 can be logged into from another system by means of ssh without implementing any extra configuration measures. The following can be used instead of the telnet or rlogin command:

- Using openssh for GXCC operation
- Setting up the ssh tunnel for GXCC
- Secure access using https for operating the Web GUI

Security against attacks from the outside:

OS hardening embraces the measures of securing ETERNUS CS8000 against attacks from the outside network. For intrusion prevention and intrusion detection various protective tools are available in order that no serious security gaps are left. The system administrator is responsible for modifying following preset network services settings after installation:

- Closing open ports and setting the required TCP and UDP ports
- SSH configuration
- TCP Wrapper configuration to specifically block or restrict TCP-based accesses to system on the external LAN
- Firewall configuration to specifically lock accesses to the ISP via the external LAN.

Remote power-on/Remote power-off:

The function Remote Power ON/Off enables ISPs to be tested after they have crashed, ISPs to be powered off remotely, and the voltage status to be queried. The function also offers a reboot mechanism (power down/up) when ISPs in the network no longer reply. It is also possible the VLPs in failover status to be powered off if the active VLP is hung up as a result of being overloaded.
The three main connection scenarios are as follows:

- Remote connection from a Windows PC to an ETERNUS CS8000 system via GUI
- Remote power-on of ISPs in a configuration over two distributed locations
- Remote power-on of all ISPs from a SAS via Tele Service

Diagnostic tools: Discover system configuration changes with SnapIt:

To discover suffered problems with a changed system configuration a tool called "SnapIt" is available. Configuration changes in the system are found by means of a snapshot method in which two snapshots are created at different times. The first snapshot is "frozen" and is used as a reference by the following snapshot when the subsequent comparison takes place. The result of this comparison is marked with Severity Levels and written to the output interface (CLI / GUI). The check is performed for these configuration files to identify the changes:

- ETERNUS CS8000 configuration files
- LINUX operating system configuration files
- SAS operating system configuration files

Health Check:

Health Check is a service which is offered by the Fujitsu maintenance department, if certain service level is agreed. The health check is used for rapid pre-analysis after an ETERNUS CS8000 problem has occurred and, in normal operation, permits proactive maintenance of the ETERNUS CS8000 system. It enables the current status to be ascertained and checked.

Extended health monitoring:

Extended Health Monitoring provides additional features to inform the system administrator about potential problems regarding the overall function of the system. Extended Health Monitoring is an extension of the standard monitoring and addresses Hardware and Software components and resources. The following are the main functions addressed:

- Node health check for all cluster nodes
- Service processes health check
- LAN switches: remote logging
- SAN switches: usage of Fabric Watch
- NAS file systems: space and inode usage monitoring

Accounting:

The accounting subsystem provides a source of well-defined resource usage data about the operation of the appliance. The accounting subsystem is not in itself a billing system, but it is intended that the collected data is suitable as input to an external billing scheme or product.

The data provided aims to cover all the different aspects of resource usage in an ETERNUS CS8000 system. It enables flexible accounting of the resources used without imposing any particular schema. It offers resource usage classified by LVG, virtual tape drive and PVG. It includes I/O throughput and different views of the volume of stored data, including usage of the TVC.

The accounting subsystem also makes the data suitable for evaluation of the existing resource usage and to make capacity predictions for the future.

6.4 Total Cost of Ownership

Combining disk and tape increases efficiency by introducing ILM into the backup process. The ability to manage with a single appliance two storage tiers and store data automatically for short term on disk and for long-term retention on tape, as well as to manage the entire lifecycle through the appliance itself, reduces the overall disk capacity expense. The cost of tape cartridges is much less than disk and on a MB for MB basis, disks are still much more expensive than tape, even if deduplication is in use. And spinning disks require power and cooling, so moving long-term data onto tape reduces also power, cooling, and floor space requirements. Just in time provisioning of virtual tape resources with ETERNUS CS8000 further reduces overall disk cache requirements by allowing users to only provision virtual tape volumes on disk as it is needed.

Considerations for TCO calculations:

- TCO reductions derive from
  - Consolidation of multiple storage tiers
  - Single point of administration
  - Single target for backup and archiving for all major operating systems
  - Only one tape and disk technology in the data center
  - Optimized disk usage with compression and deduplication
  - Automated and optimized operations using policy based management
- Efficient use of tape storage
  - Lowest cost for storage
  - Full usage of media capacity
  - Smaller or less tape libraries
- Consolidated storage
  - Less physical tape drives, less media
  - Reduced costs, power, maintenance, floor space, operations and support staff

7 Models of the ETERNUS CS8000 Series

The core components of the ETERNUS CS8000 systems are based on the rack mountable Fujitsu PRIMERGY RX Industry Standard Server and Fujitsu ETERNUS DX RAID storage system. One ETERNUS DX RAID system exists of a base unit and RAID expansion shelves. The appliance supports up to 16 RAID systems for use as Tape Volume Cache (TVC), ViNS Online Store (VOS), and BLOCKPOOL. The RAID systems are configured with RAID 5 for SAS disk based configurations and RAID 6 for NL SAS disks. The components are installed in Fujitsu standard racks. All racks are 42HUs height and 1100mm deep with an asymmetric construction for optimized cable management and enhanced service access.

These standard building blocks enable that each system can be configured and seamless scaled up according individual needs. Each configured system is manufactured in Fujitsu's own factories and has to pass strongest quality assurance processes before delivery.

Models at a glance:
Each model configuration can be tailored to almost any individual needs with the possibility for further specific extensions. The storage subsystems CS VTL and CS ViNS can coexist within a single appliance, but depending on the specific customer requirements, some functions may be limited or excluded. Tailored configurations differ basically in the degree of component redundancies and data throughput. Extensive throughput requirements are possible with a flexible scalability regarding number of interfaces and RAID storage capacity. The basic functions of the embedded appliance operating software can easily expanded with optional functions by activating respective license keys.

Models	CS8200	CS8400	CS8800
Frontend	VTL	VTL / ViNS	
Data Reduction	VTL compression / VTL deduplication		
Scalability	RAID system scale-up	Full scale-out	
Scalability Nodes	2 nodes (standard)	Up to 20 nodes	
Scalability RAID System	1 RAID system	Up to 16 RAID systems	
Site Location	Single – site		Split - site
System Failover	Single - site failover		Split – site failover
Synchronous Replication	None	Single - site cache mirror	Split – site cache mirror
Asynchronous Replication	VTL cascading	VTL cascading ViNS replication	
Backend Tape Replication	VTL dual save	VTL multiple save ViNS dual save	

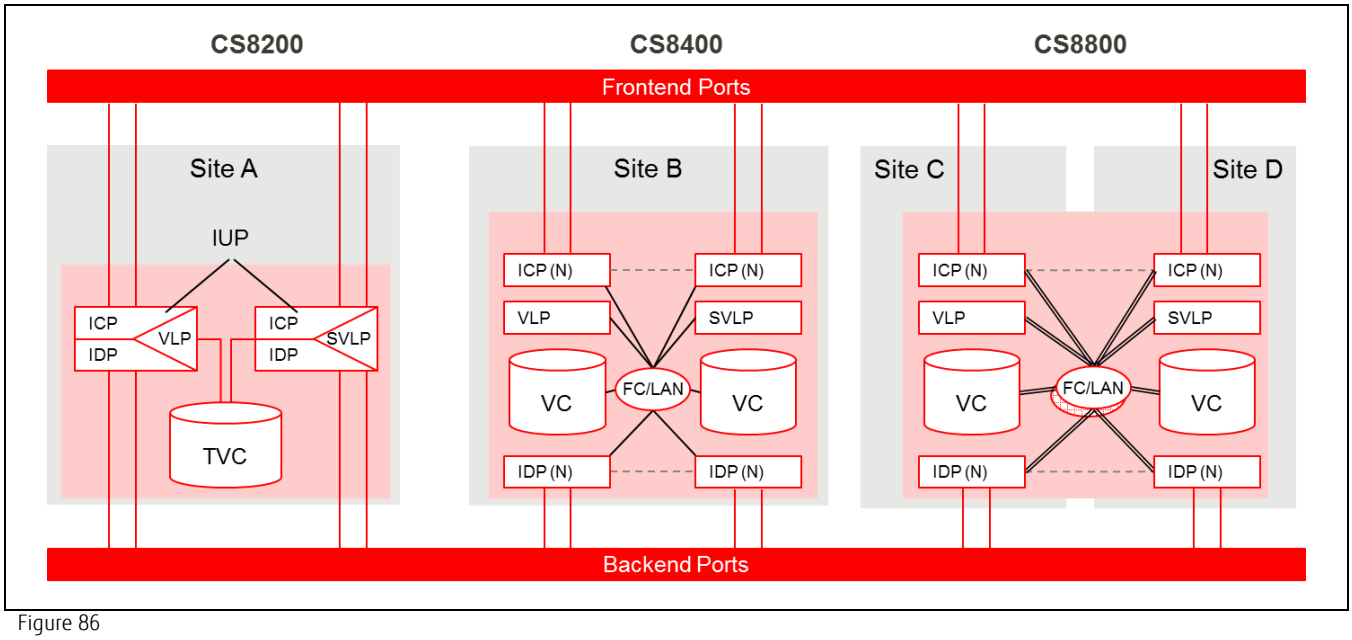


Figure 86

ETERNUS CS8200:

The ICP, IDP and VLP processors are designed to run also together on a single physical server, referred as Integrated Universal Processor (IUP). The model CS8200 is built with two IUPs and one dedicated RAID system. The two IUP nodes with type ICP_IDP combine the attributes of ICP, IDP and VLP in a single ISP. An ICP_IDP has both, a FC host adapter and a SCSI or FC controller to operate physical tape drives and tape libraries. This model delivers the entire CS VTL functionality and can be deployed excellently in cascaded environments.

Due to two VLPs, the two ICP_IDP nodes can provide automatic failover between each other. Together with the internal RAID system the ETERNUS CS8200 model provide basic redundancy and high availability for tape backup.

ETERNUS CS8400 / ETERNUS CS8800:

Within the models ETERNUS CS8400 and ETERNUS CS8800, the processors ICP, IDP and VLP are distributed among several appliance nodes, each running dedicated ICP, IDP or VLP services. This provides a reliable and scalable operation at the frontend and backend. With the dual site model ETERNUS CS8800 the appliance can be spanned over two sites. A dual site configuration providing cache mirroring and automatic failover is the ideal choice for fastest disaster recovery.

General Limits:

- Max 16 RAID storage base units per system
- Max 22 ISPs (10 ICP, 10 IDPs, 2 VLP) and one TBP
- Max 10 dedicated ICPNs for ViNS at frontend
- Max 8 (4 active + 4 hot stand-by) dedicated IDPN for ViNS backend
- No mixture of VTL and ViNS in one ICP or IDP
- Max 4 per ICP / 6 per IDP interface ports

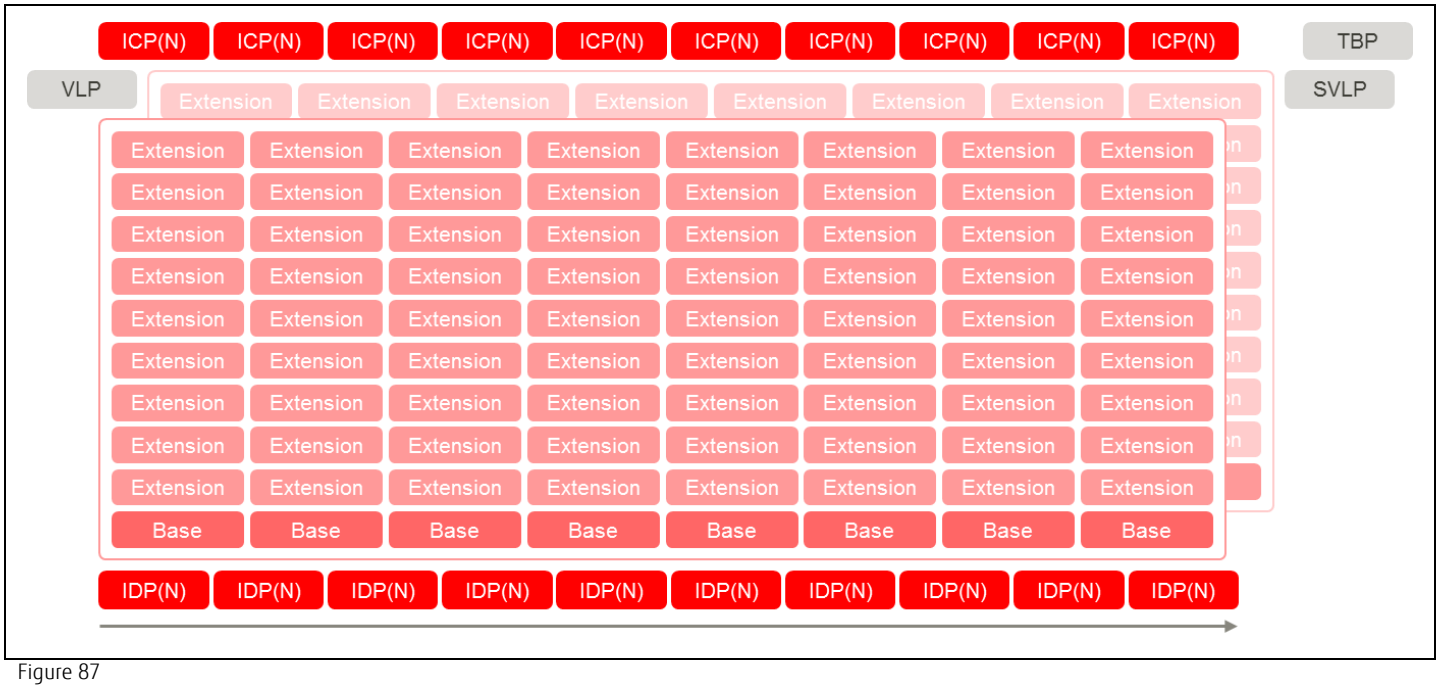


Figure 87

Logical Scalability:

Overview of maximum Logical Volumes and Volume Groups for ViNS, VTL / ViNS mixed mode
([1)] the number of PV in native mode is the same like the number of LV).

Model	Max RAID Systems	LV	LVG	PV [1)]	PVG	TVPs/TVPGs
CS8200	1	300,000	340	50,000	100	-
CS8400 – CS8800	16	3,000,000	512	50,000	100	100

8 Additional Information

- Website: ETERNUS CS:
  http://www.fujitsu.com/fts/eternus_cs

- Data Sheet ETERNUS CS8000:
  http://docs.ts.fujitsu.com/dl.aspx?id=2e0b9508-9e6a-4197-915e-e378d794fb3c

- Supported Environments:
  http://docs.ts.fujitsu.com/dl.aspx?id=02c0faa4-9961-49e0-a099-cfd01100482d

- Brochure:
  ETERNUS CS8000 – Improve the economics of your backup and archive infrastructure
  http://docs.ts.fujitsu.com/dl.aspx?id=b696b187-8d4a-4a19-8587-f90ef32e2e55

- White Paper:
  ETERNUS CS8000 - Disaster-resilient Architectures
  http://docs.ts.fujitsu.com/dl.aspx?id=cceb3772-7c7d-4797-b0f4-3f355649ed93

- Meeting Today's Backup and Archive Challenges
  IDC Whitepaper #IDCWP36U October 2012
  http://docs.ts.fujitsu.com/dl.aspx?id=1fb2c41d-fcf0-4a67-b3cd-16a61d343263

Published by
FUJITSU
E-mail: storage-pm@ts.fujitsu.com
Website: www.fujitsu.com/fts/eternus_cs
2014-06-30 WW EN