IBM System Storage™

The IBM Virtualization Engine™ TS7700

Keijo Ekman
Consultant IT Specialist
kekman@fi.ibm.com

Supporting Business Continuity and Information Lifecycle Management

TS7700 Virtualization Engine Series

- Successor to the successful IBM TotalStorage Virtual Tape Server (VTS)
- The TS7700 Virtualization Engine Series is composed of:
  - One IBM Virtualization Engine TS7740 server (3957 Model V06)
  - One IBM Virtualization Engine TS7740 MODEL CC6 (3956 Model CC6) -- Cache Controller
  - Three IBM Virtualization Engine TS7740 MODEL CX6 (3956 Model CX6) -- Cache Drawer
- Introduces a new modular, scalable, high-performing architecture for mainframe tape virtualization.
- New TS7700 Grid Communication features provides peer-to-peer like copy capability between two TS7700's using IP network connections.
- General Availability
  - Sept. 29, 2006
- IBM TotalStorage Virtual Tape Server Withdrawal From Marketing:
  - 3494 Models B10, B20, CX1 and selected features
  - Last order date: Dec 1st

TS7700 Virtualization Engine Components

- A tape frame[1]
  - Frame provides up to 36u for mounting
    - A TS7740 Virtualization Engine
    - One TS7740 cache controller
    - Three TS7740 cache drawers
  - Supports High availability
    - Redundant power supplies
    - Two power feeds

[1] Machine Type 3952 Model F05

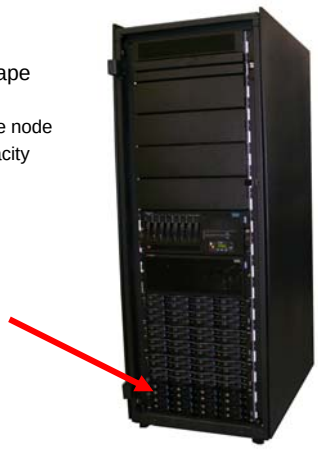
TS7700 Virtualization Engine Components (continued)

- One TS7740 node[1]
  - High performance IBM System p520
    - Two dual-core, 64-bit, 1.9-GHz processors
    - 8 GB RAM
    - Two or four 4 Gbps FICON host interfaces
    - Two 1 Gbps replication links
    - Additional adapters
      - Physical library and drive attachments (fiber)
      - Management interface (Ethernet)
      - Service interface (Ethernet)
  - Supports high availability
    - Dual power
    - Redundant hot-swap power supplies and fans

[1] Machine Type 3957 Model V06

TS7700 Virtualization Engine Components (continued)

- One TS7740 cache controller[1]
  - Provides high performance RAID 5 disk tape volume cache
    - Attach to one TS7740 Virtualization Engine node
    - Provide up to 1.5 TB of usable cache capacity
    - Includes 16 15k rpm 146GB FC HDDs
    - Includes four 4 Gbps FC interfaces
  - Supports high availability
    - Dual power
    - Automatic hot sparing/rebuild
    - Redundant hot-swap components
      - Raid Controllers
      - Power Supplies
      - Enclosure fans
      - Hard disks



[1] Machine Type 3956 Model CC6

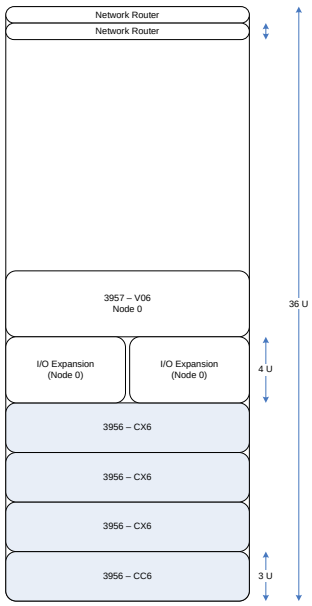
TS7700 Virtualization Engine Components (continued)

- Three TS7740 cache drawers[1]
  - Provide high performance RAID 5 disk arrays
- Each TS7740 cache drawer
  - High performance RAID 5 disk
    - Attaches to the TS7740 cache controller
    - Provide 1.5 TB of usable cache capacity
    - Includes 16 15k rpm 146GB FC HDDs
  - Supports high availability
    - Dual power
    - Automatic hot sparing/rebuild
    - Redundant hot-swap components
      - Power Supplies
      - Enclosure fans
      - Hard disks

[1] Machine Type 3956 Model CX6

TS7700 Frame Layout

- 3952 F05 Frame
- System P Server
  - Power5 processor (2 dual core CPUs)
- I/O Expansion drawers
- RAID Disk controller and 3 disk expansion drawers
- Redundant network routers
  - Provides internal network connection to Library Manager and Disk Controller configuration
  - Provides protected NAT interface for customer to access Management Interface services running on controller
- Expansion for 2nd 3957 controller and I/O drawers (SOD)



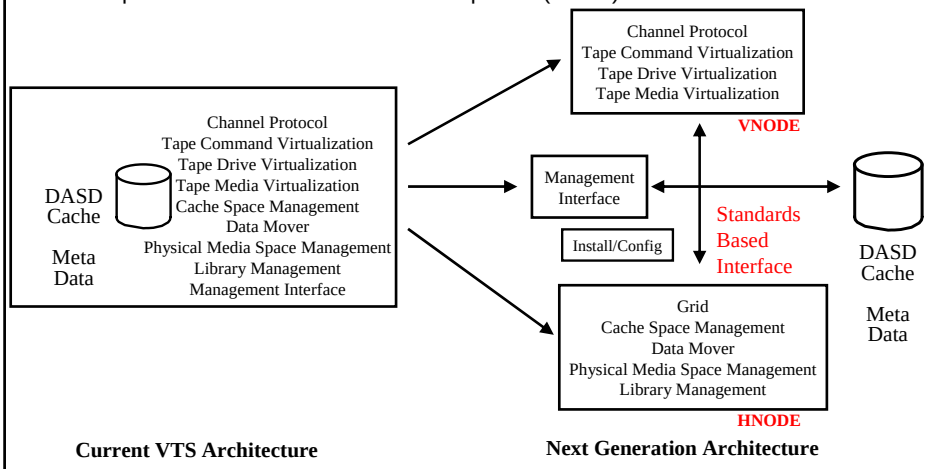
TS7700 Virtualization Engine Specifications

Specification	TS7740	Model B10	Model B20	Model B18
Number of Virtual Devices	128	64	128 / 256	64 / 128
Usable Cache Capacity	6 TB	216 – 432 GB	864 GB to 1.7 TB	72 GB to 1.7 TB
Compressed Cache Capacity (3:1)	18 TB	648 GB to 1.2 TB	2.4 TB to 5.2 TB	216 GB to 5.2 TB
FICON	4 (4Gbps)	2 / 4	4 / 8	
ESCON Channels		2 / 4 / 8	8 / 16	2 / 4 / 8
TS1120/3592 Tape Drive Attachment	4 - 16	4 - 12	4 - 12	
3590 Tape Drive Attachment		4 - 6	4 - 12	3 - 6
Number of Virtual Volumes	500,000	250,000	500,000	250,000
Supports upgrade path	planned		planned	

Statements of IBM future plans and directions are provided
for information purposes only. Plans and direction are subject to change without notice.

TS7700 - Architectural Partitioning

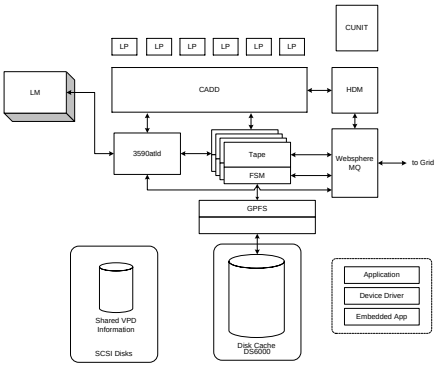
- Break up the monolithic VTS into scalable pieces (nodes)



Current VTS Architecture **Next Generation Architecture**

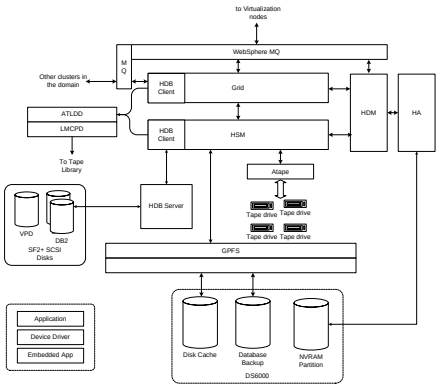
Virtualization Node (VNode)

- The “VNode” refers to a code stack which performs all of the actions needed to present a library image and drive images to a host
- The VNode code was designed to run along side of the HNode code in the same controller, or in a separate controller
- Uses standardized interfaces to talk with outside components (TCP/IP)



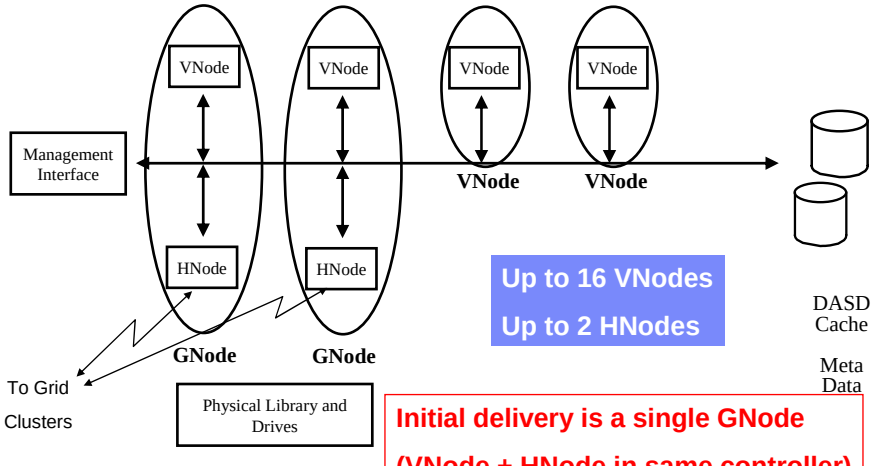
Hierarchical Storage Management Node (HNode)

- The “HNode” refers to a code stack which performs all of the actions needed coordinate the contents of the disk cache with the data on backend tape. It also includes the logic for managing changes and replication of the data across different sites.
- The HNode code was designed to run along side of the VNode code in the same controller, or in a separate controller
- Uses standardized interfaces to talk with outside components (TCP/IP)



TS7700 - Scalability

- Common DASD, Multiple Elements - redundancy, capacity, attachments



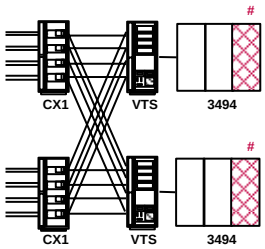
Up to 16 VNodes

Up to 2 HNodes

Initial delivery is a single GNode

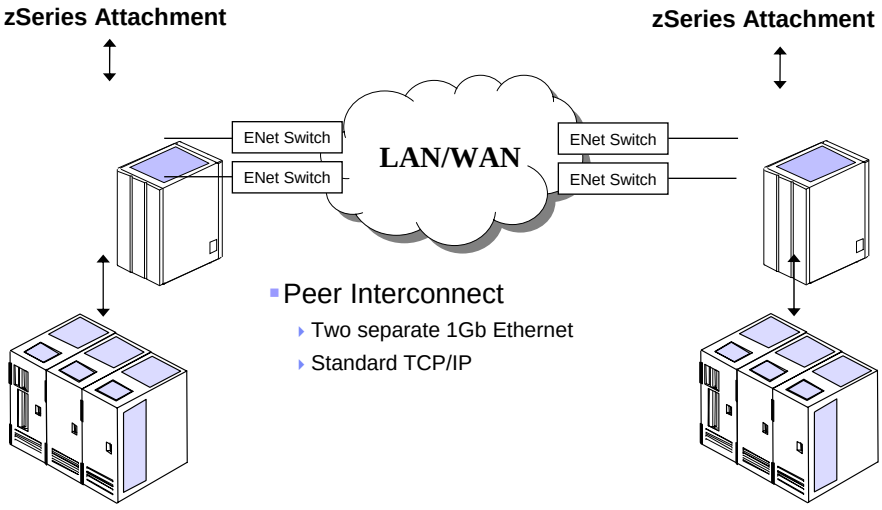
(VNode + HNode in same controller)

Current PtP VTS Implementation



- A PtP VTS Configuration
  - Appears as a single VTS 'image'
  - Provides 64, 128 or 256 Virtual Drives
  - Requires 1 or 2 CX1 frame(s) with
    - Required VTC feature codes
  - Requires two supported VTS Models
  - Two supported tape libraries
    - Two 3494 tape libraries with either
      - four to 12 3590 tape drives or four to 12 TS1120 or 3592 J1A tape drives or
      - four to six 3590 tape drives and one to 12 TS1120 or 3592 J1A tape drives installed
    - Two 3584 tape libraries with
      - four to 12 TS1120 or 3592 tape drives
    - One 3494 and one 3584 tape library
      - The 3494 can support 3590 and or TS1120 or 3592 tape drives
      - The 3584 only supports TS1120 and 3592 J1A tape drives

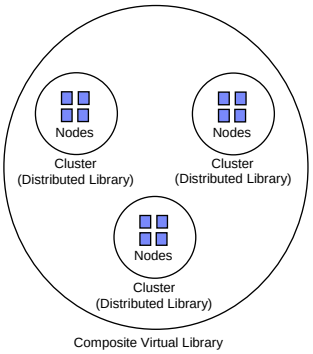
TS7700 - Grid Interconnect



- Peer Interconnect
  - Two separate 1Gb Ethernet
  - Standard TCP/IP

TS7700 Grid - Scalability

- Architecture for high scalability
  - Partition function into Nodes
  - Group nodes together in a Cluster
- Clusters will scale horizontally
  - 1-16 Virtualization Nodes (VNodes) – 1 at GA
  - 1 or 2 Hierarchical Storage Management Nodes (HNodes) – 1 at GA
  - Tape Volume Cache (disk)
  - 1-4 Physical libraries – 1 at GA
- Clusters interconnect via network
  - Interconnect clusters to form a composite virtual library
  - Up to 8 clusters in a composite virtual library – 2 at GA
  - PTP is integrated into the architecture



Retain Host OS view of Composite and Distributed Libraries

VTCs Versus TS7700

- VTCs in the PTP VTS performed the following key functions
  - Management of what and how to replicate data between the VTSs
  - Determination of which VTS has a valid copy of a logical volume
  - Selection of a VTS to handle the I/O operations for a tape mount and the routing of host I/O operations to that VTS.
  - Directed all volume management commands to the Master VTS and determined which VTS was the Master (volume ownership)
  - Each represented 16 or 32 virtual tape drive addresses
  - Each attached to the VTSs through a single ESCON or FICON interface
  - The VTSs had only partial awareness of being in a PTP configuration
- In the TS7700
  - First two functions (management of replication and volume validity) have been integrated into the base product
  - Hosts are directly attached to each TS7700 and have to manage the use of the virtual drive addresses
  - *There is no Master, volume ownership is dynamic across the TS7700s*
  - Each TS7700 provides 128 virtual tape drive addresses
  - Inter-TS7700 interface is dual Gb ethernet

TS7700 - Grid Architecture – Critical Concepts

- **All configurations have a Composite and Distributed Library**
  - *All configurations are considered part of a Grid (even a grid of only 1 cluster)*
- **Any logical volume can be accessed by any virtual device in the subsystem**
  - *Mounts of logical volumes to a cluster without a consistent copy will access the logical volume remotely across the WAN*
  - Replication policies and policy overrides assist the subsystem in choosing the location to access the volume, and whether a copy needs to be made
- **The microcode tracks changes to logical volumes, and ensures consistencies across the subsystem**

TS7700 - Grid Architecture – Critical Concepts

- **Each cluster has its own logical device range to the host**
  - Logical control units and associated devices are on VNode boundaries
  - *Similar to current PtP, each VNode presents a group of logical control units with 16 devices per LCU*
  - Paths to each VNode are configured separately to the host (like each VTC had its own definition)
- **All logical volumes within subsystem are known by every cluster**
  - Like the current PtP, each cluster knows about every logical volume, and the microcode works to keep all information about the logical volumes in sync
  - Maximum number of logical volumes is enforced per TS7700
- **The placement of logical volumes across the clusters, and replication are performed according to customer policies & defaults**

TS7700 - Grid Architecture – Critical Concepts

- **A logical volume is serially 'owned' by a cluster**
  - *Initial owning cluster is the one the insert was performed at*
- **The owning cluster handles changes to a volume's state and status**
  - Category change, construct changes
  - Then synchronizes with the other clusters
- **A cluster must own the volume to process a mount request**
  - *If a cluster that does not currently own the volume receives a mount request through its vnode, it first obtains ownership from the current owning cluster before it can proceed. If ownership cannot be obtained, the mount will fail.*
  - Added steps in the mount process may impact the number of mount per hour
- **Customer settable ownership takeover modes are provided**
  - *Used when a cluster has failed or is otherwise unavailable and access to volumes it had owned is required.*
  - Read Only and Read/Write modes
  - When original owning cluster is restored, ownership takeover relationships are resolved.

TS7700 - Grid Architecture – Critical Concepts

- **Outages to a cluster affect access to volumes owned by that cluster from other clusters**
  - Other clusters can be given permission to access or access/modify volumes owned by a cluster with outage
  - *Permission granted by Operator via Management Interface*
- **Clusters "gracefully" removing themselves from the Grid (Service Mode), automatically enable other clusters to access their owned volumes**

Grid Host View

- Logical control units and physical paths are defined on a VNode/GNode boundary (similar to VTCs in the current PtP)
  - All of them are part of the same composite library image to the host
- The subsystem ID information returned in the RDC command must uniquely identify the logical control unit from others in the composite library image
  - The first release defines the possible subsystem IDs as follows:

Distributed Library	Logical CUs	Subsystem IDs
0	0-7	0x1-0x8
1	0-7	0x41-0x48

Cluster to Cluster Replication Policy

- New definition for copy as ***Copy Consistency Point***
  - Defines when a target volume is consistent on a TS7700 with the source volume
- The Copy Consistency Point Policies determine:
  - The cluster which should have the initial location of data
  - Which clusters get copies of logical volumes, and at what point the cluster should have copies
- **Copy Consistency Point Policies are configured using Management Class**
  - Policies are assigned to a management class name
  - Policies are set through Library Manager

Cluster-Cluster Replication Policy (con't)

- Consistency Policy Options:
  - **No copy (N)** – this cluster does not receive a copy for volumes in this management class
  - **RUN (R)** – this cluster will have a valid replication of the logical volume before we provide Device End to the Rewind Unload (RUN) command from the host (this is a direct parallel to current PtP Immediate mode copy setting)
    - Corresponds to VTS Immediate Copy Mode
  - **Deferred (D)** – this cluster will get a valid replication of the logical volume at some point in time after the job completes (same as the deferred mode in the PTP)
    - Corresponds to VTS Deferred Copy Mode

Default Management Class (-------):

Cluster 0 – (R) RUN

Cluster 1 - (D) Deferred

Any previously undefined MC name received by the LM for a scratch mount will use these defaults.

You can change the (-------) MC Consistency Point to provide different defaults.

Cluster to Cluster Replication

- Setting of the Management Class policy at the LM defines the consistency sync point for every defined cluster in the subsystem
  - The policies become an array of sync values, with each element of the array representing the policy for a given cluster.
    - ***At GA, the array has only 2 values, the first for Cluster 0, and the second for Cluster 1.***
  - e.g. If it is desired to have a copy of a volume at unload time at Cluster 0, and a deferred copy at Cluster 1, the array would be:

R	D

Cluster 0 – RUN (Immediate Copy made here)

Cluster 1 – Deferred (Deferred Copy made here)

Cluster to Cluster Replication

The replication policies can be different at each LM

- The setting of the policies at each LM dictates what the resulting actions would be if the volume is mounted on a virtual device address associated with the clusters associated with that LM.

Example:

Site A (Cluster 0): | R | D | Site B (Cluster 1): | D | R |

If volume mounted at Cluster 0 (Site A):

Cluster 0 – RUN (Immediate Copy made here)

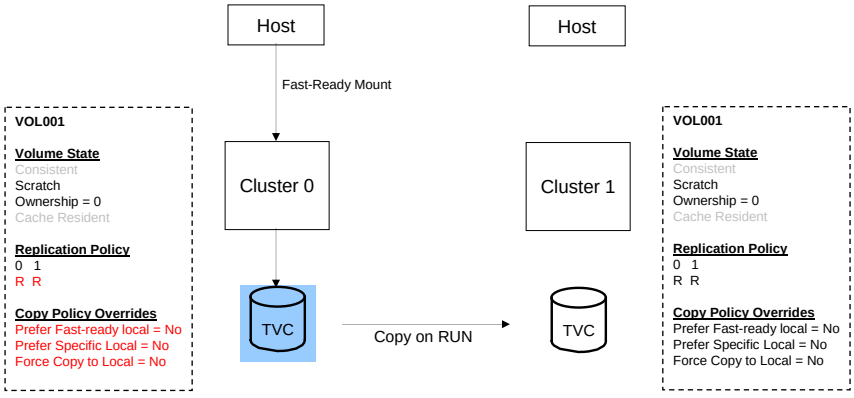
Cluster 1 – Deferred (Deferred Copy made here)

If volume mounted at Cluster 1 (Site B):

Cluster 0 – Deferred (Deferred Copy made here)

Cluster 1 – RUN (Immediate Copy made here)

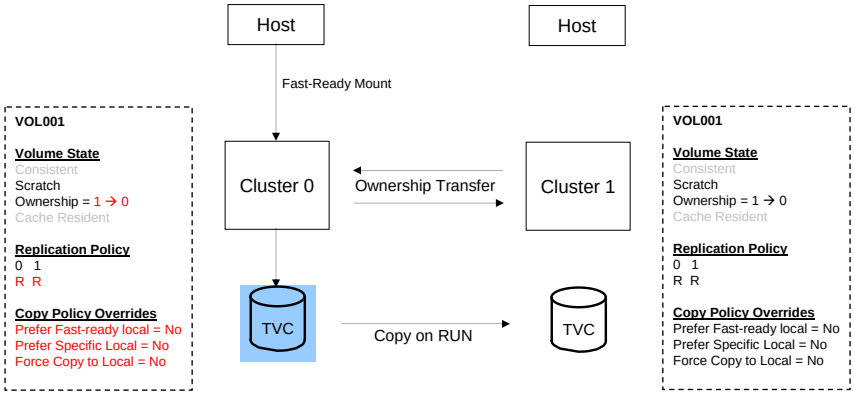
TVC Selection – Fast Ready
Mounting Cluster 0 is Owner – RUN on Cluster 0 – RUN on Cluster 1



For a Fast-Ready (scratch) mount, replication policy is the primary factor for picking the TVC

Typical ‘synchronous’ mode setting

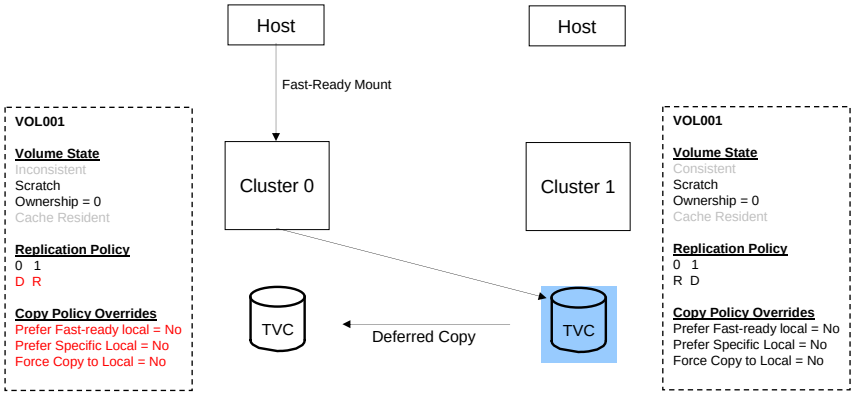
TVC Selection -- Fast Ready
Mounting Cluster 0 is NOT Owner – RUN on Cluster 0 – RUN on Cluster 1



For a Fast-Ready (scratch) mount, replication policy is the primary factor for picking the TVC

Typical 'synchronous' mode setting

TVC Selection – Fast Ready
Mounting Cluster 0 is Owner – Deferred on Cluster 0 - RUN on Cluster 1



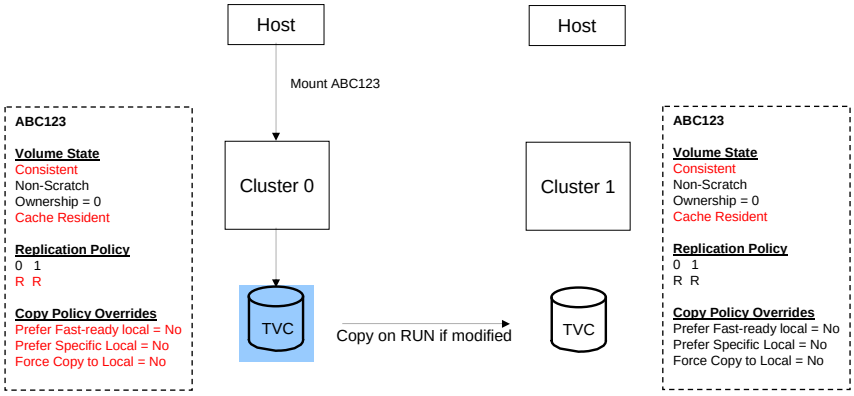
For a Fast-Ready (scratch) mount, replication policy is the primary factor for picking the TVC

Send primary copy offsite

TVC Selection – Specific

Mounting Cluster 0 is Owner – RUN on Cluster 0 - RUN on Cluster 1

Cache Resident Both Clusters – Consistent Both Clusters

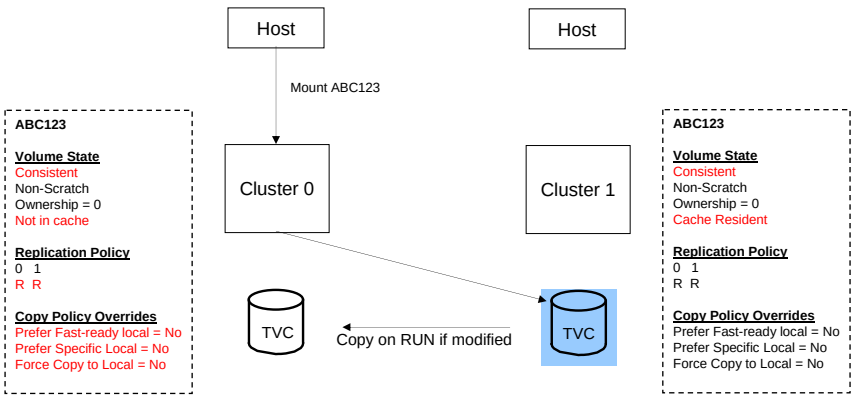


For a specific mount, volume consistency and replication policy are the primary factors for picking the TVC. Cache residency is a secondary factor if cache state is not equal.

TVC Selection – Specific

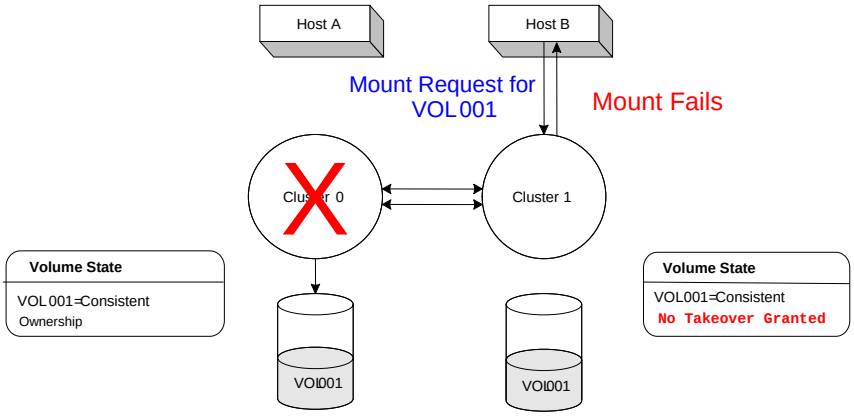
Mounting Cluster 0 is Owner – RUN on Cluster 0 - RUN on Cluster 1

Not in Cache Cluster 0 – Cache Resident Cluster 1 - Consistent Both Clusters



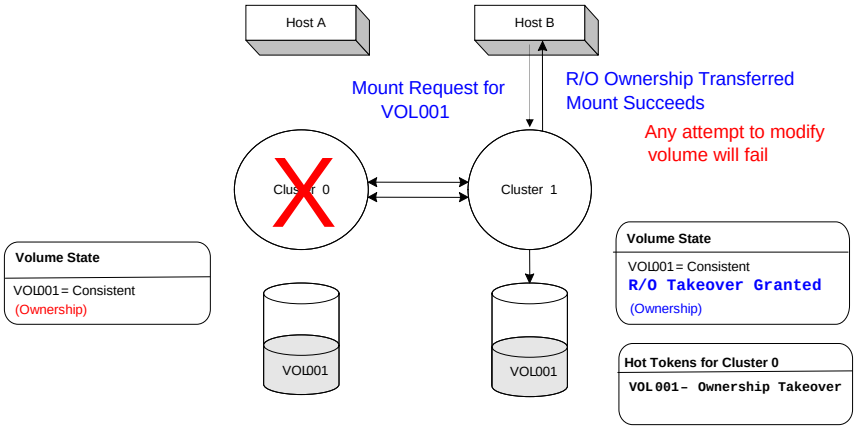
For a specific mount, volume consistency and replication policy are the primary factors for picking the TVC. Cache residency is a secondary factor if cache state is not equal.

Failure – Owning Cluster Unavailable (No Takeover Granted)

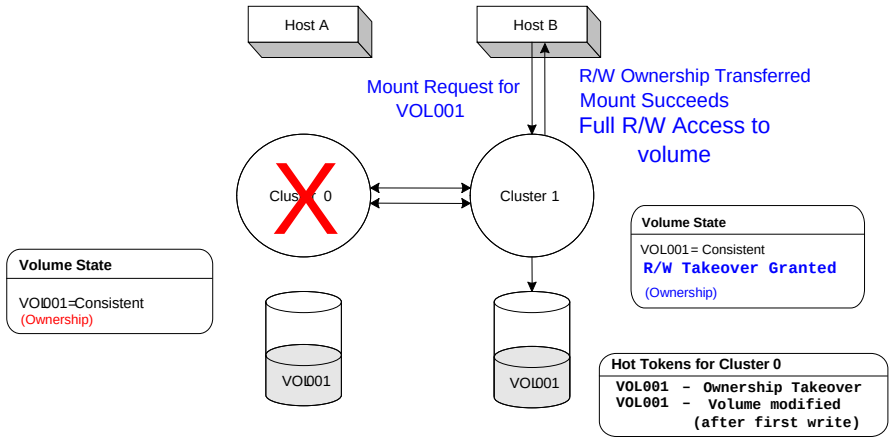


In R1, Ownership Takeover always requires customer involvement

Failure – Owning Cluster Unavailable (R/O Takeover Granted)



Failure – Owning Cluster Unavailable (R/W Takeover Granted)



Operating Systems Supported

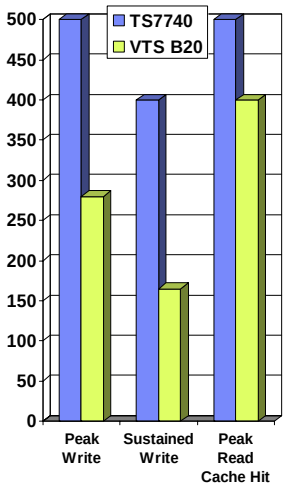
- z/OS V1R4 and higher
- z/VM 4.4.0 and higher
- z/VSE.3.1.2
- TPF 4.1 and z/TPF V1.1

No software changes required!

However, performance monitoring using the old VTSSTATS tool is no longer Possible, and SMF94 records are not produced. Use the BVIR tool in stead

TS7700 Preliminary Performance[1]

- Single TS7740 Cluster
  - Laboratory measurements indicate a potential peak write data rate of >500MB/s
  - Lab measurements indicate a potential sustained write data rate of >400MB/s
  - Lab measurements indicate a potential peak read cache hit data rate of >500MB/s
- Customer performance may vary
  - Block size, compression ratio and batch window characteristics
  - Processor and channel configuration
- Laboratory measurements
  - Four FICON channels
  - 128 Concurrent jobs
  - 800 MB volumes
  - 32 KB blocks
  - 3:1 compression ratio
  - 20 buffers



[1]Lab measurements, customer results will vary

Product Roadmap

A short preview of what's in STORE for you

TS7700 Design Objectives

- Current offering
  - Exploitation of the latest IBM technologies
    - ✔ IBM System p5 servers to increase performance
    - ✔ IBM modular disk subsystems to significantly increase cache capacity
  - Extending our Virtual Tape leadership position
    - ✔ Re-engineer the software architecture
      - Reuse a substantial percentage of the existing code
      - Rewrite components that will substantially improve performance or function
    - ✔ Address emerging customer requirements
      - Cost effectively manage Information lifecycles
      - Address Business Continuity Challenges
- Future Plan Strategy
  - Extending the hardware architecture to
    - Support a high availability subsystem
    - Reduce the disruption of planned or unplanned outages
  - Protecting customer investment
    - Provide upgrade path for current B20 VTS installations
    - Supporting 3494 Tape Libraries

Statements of IBM future plans and directions are provided for information purposes only. Plans and direction are subject to change without notice.

Previous Statements of Direction - Continued Focus

- Enhanced Import/Export capability for VTS products that will allow 'sets' of cartridges to be interchanged
  - Can enhance operational efficiencies for customers who choose to move cartridges offsite for manual vaulting
- Support for full-duplex communication between three sites for enhanced electronic vaulting with PtP VTS

TS7700 Statements of Direction

- IBM plans to :
  - Enhance the TS7700 Virtualization Engine by supporting the installation of a second TS7740 Server (Machine Type 3957, Model V06) within the 3952 Model F05 Tape Frame
  - Support attachment of TS7700 to the 3494 Tape Library
  - Provide an upgrade conversion of VTS Model B20s to a TS7740 server
    - Upgraded Model B20s will operate as a single server TS7700, including ability to participate in a TS7700 grid configuration
  - Utilize the encryption capability of the TS1120 Tape Drive