IBM TotalStorage™

SAN Volume Controller Revealed

Bill Wiegand
IBM Advanced Technical Support

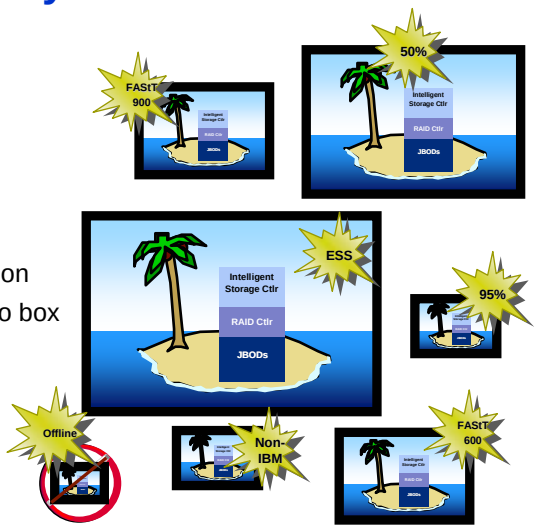
Agenda

- Current Environment
- SVC Architecture
  - Disk Management
  - Clustering
  - RAS
  - Master Console
  - Copy Services
- Zoning
- Data Migration

Intelligent Disk Array Limitation

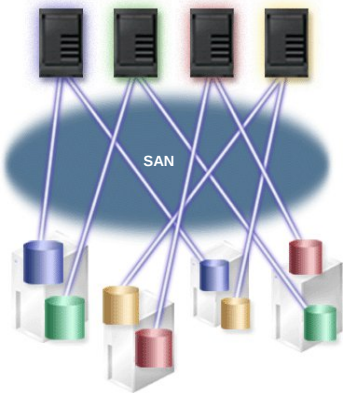
SAN Storage Islands

- Individually managed
- Stranded capacity
- Varied intelligence levels
- Lacks dynamic data migration
- Replication Service is box to box



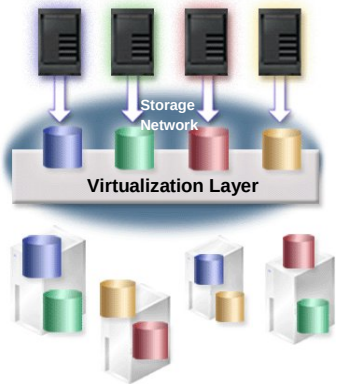
Virtualization Implementations

SANs Today



Servers are mapped to specific physical disks
i.e., "physical mapping"

SAN Volume Controller



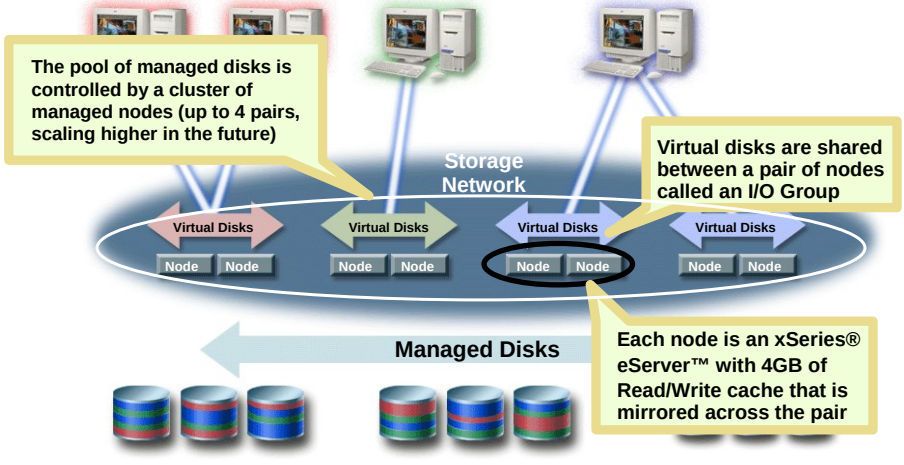
Servers are mapped to a virtual disk
i.e., "logical mapping"

Some Basic Terms

- **Managed Disk (mDisk)** – LUNs provided by backend storage carved up into equal-sized extents.
- **Managed Disk Group** – A group of **Managed Disks** using the same extent size, usually with similar performance characteristics.
- **Virtual Disk (vDisk)** – LUNs presented to hosts. A single **vDisk** actually resides within a single **Managed Disk Group. vDisk** extents are mapped to **mDisk** extents by the SVC.
- **Storage Engine** – a special xSeries Server that handles the mapping between **vDisks** and **mDisks** and provides virtualization, data migration and copy services. Also commonly known as an SVC node.
- **I/O Group** – a pair of SVC **Storage Engines** that back each other up for **vDisk** processing. A **vDisk** is serviced by exactly one **I/O Group**. Also commonly referred to as an SVC node pair.
- **Cluster** – one or more **I/O Groups** that share the same set of **Managed Disk Groups.**

IBM TotalStorage SAN Volume Controller

Designed to be a redundant, modular, scalable, solution



SAN Volume Controller - Hardware

Base Offering

- Dual Storage Engine Clustered System
  - Up To Two Engine Pairs Supported
- UPS (Required with the SAN Volume Controller)
  - 2 Per Cluster
  - 2U Form Factor
  - Supports 1-4 Engine Pairs
- Master Console
  - 1U 19" Rack Mounted xSeries Server
  - 2 Port, 2Gb FC HBA
  - Rack Mounted Monitor/Keyboard
- Each Engine Contains:
  - Modified xSeries Server
    - 1U 19" Rack Mounted Enclosure
    - Dual 2.4GHz Processor
    - 4GB of ECC Memory
    - Dual PCI-X 64 Bit 100 MHz Slots
    - Dual 10/100/1000 Cu Ethernet Ports
    - 18GB SCSI HDD
    - 2 x 2 Port, 2Gb FC HBA
  - Management Module
    - Heart Beat Timer
    - Control for VFD Display/Keypad
    - Power button intercept
    - Secondary Flashboot Device
  - Front Bezel
    - VFD Display
    - 5 Button Keypad
  - Pre-loaded Virtualization Software based on Version 2.4 of the Linux kernel

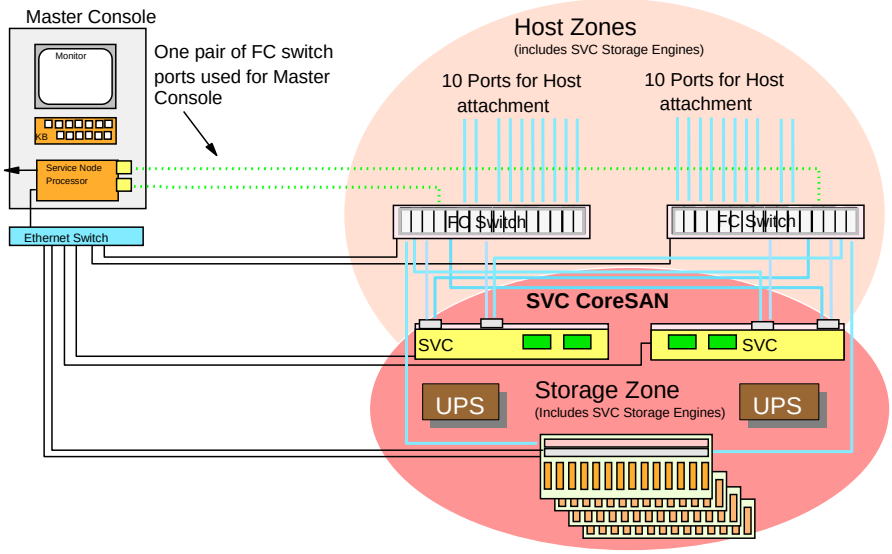
Why specify the engine?

- Take complexity out
- Appliance mentality
- Increase availability
- Easily scalable

1 Model:

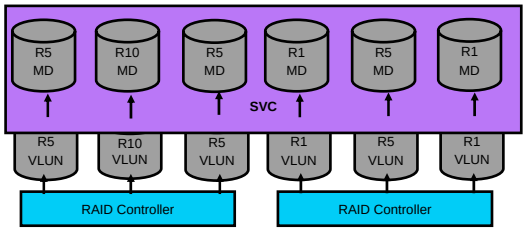
Model	Cache	FC Adapters
2145-4F2	4GB	2

SVC - Sample Configuration



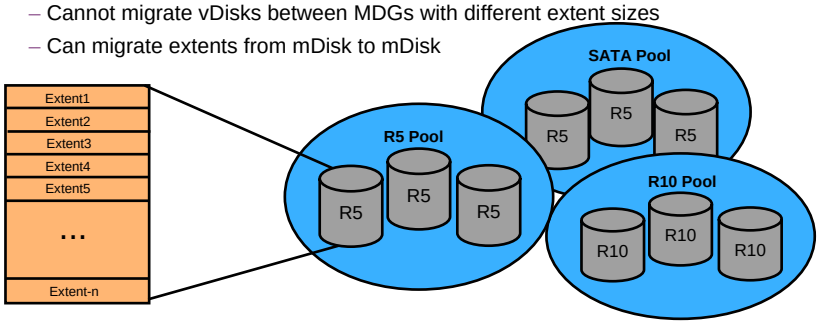
SVC - Managed Disks

- SVC does not perform RAID functions
  - Utilizes RAID capability of backend storage server
  - RAID-5, RAID-10, or RAID-1 recommended
- LUNs "surfaced" from RAID controllers are what hosts on the SAN see as physical disks
  - Disks surfaced by RAID controllers discovered by SVC as Managed Disks
  - Three modes for mDisks – unmanaged, managed, image
  - Spare capacity on mDisks can be reallocated transparently and dynamically



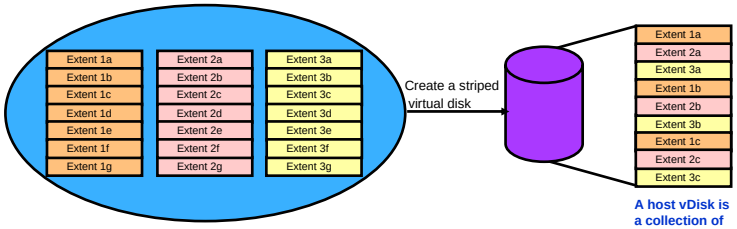
SVC - Managed Disk Groups

- SVC discovers mDisks, user assigns to pools called Managed Disk Groups
  - Sensible to pool like with like (i.e. RAID 5 with RAID 5)
  - Sets performance and availability characteristics of a MDG
  - Support for 128 managed disk groups per cluster
  - MDG can contain 128 managed disks
- These MDGs are addressed by the SVC in terms of extents
  - Extent size is determined at MDG creation time, default 16MB, max 512MB
    - Cannot migrate vDisks between MDGs with different extent sizes
    - Can migrate extents from mDisk to mDisk



SVC - Virtual Disks

- From these extents the user can build "virtual disks"
- Various policies can be used to build them
- Real physical capacity must be available to create a vDisk
  - No sparse allocation/over allocation functionality in V1.2
- Virtual disks can be expanded, reduced, or deleted
  - Some operating systems don't support dynamic vDisks
- I/O governing can be enabled to limit IO/s or MB/s



A host vDisk is a collection of Extents - each 16 MB - 512 MB

SVC - Virtual Disk Modes

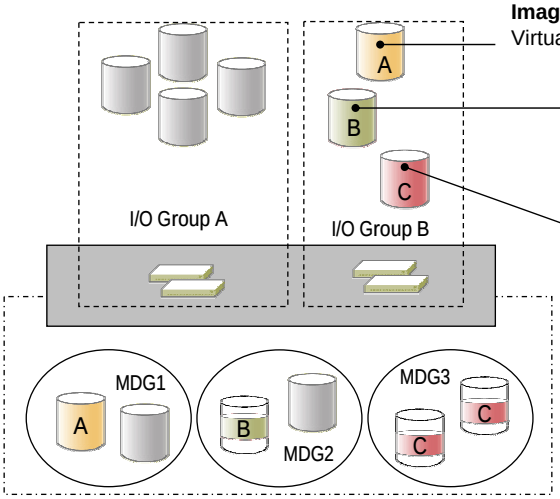


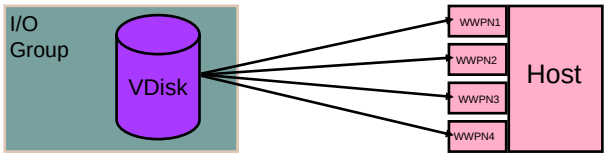
Image Mode:
Virtual Disk = Physical LUN

Sequential Mode:
Virtual Disk mapped sequentially to a portion of a managed disk

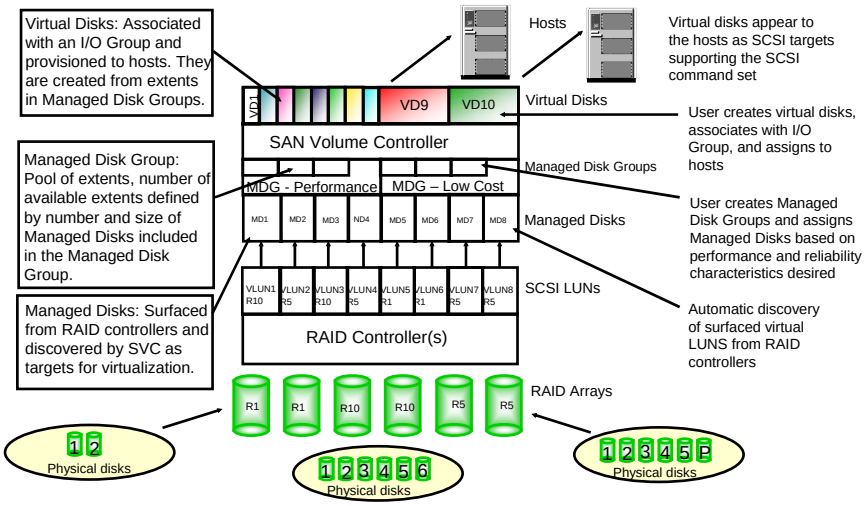
Striped Mode:
Virtual Disk striped across multiple managed disks

SVC - I/O Groups and Hosts

- Each virtual disk is assigned to a particular I/O Group (node pair)
  - Every node in cluster aware of vDisks, only owning I/O Group services requests
  - I/O targeted at either node in I/O Group for purposes of caching and load balancing
- It is these virtual disks that SVC presents to hosts on the SAN as targets of I/O
- The virtual disks are mapped to hosts
  - They are mapped to all ports on the host (SDD for mutli-path operation)
  - Can be mapped to multiple hosts for use with clustering software
  - 512 virtual disks per host or host cluster
- The hosts see these as physical disks (in terms of the OS)
  - e.g. SCSI targets, in AIX for example these are hdisks
- SVC knows hosts as groups of HBA World Wide Port Names (WWPNs)



SVC - Virtualization Summary

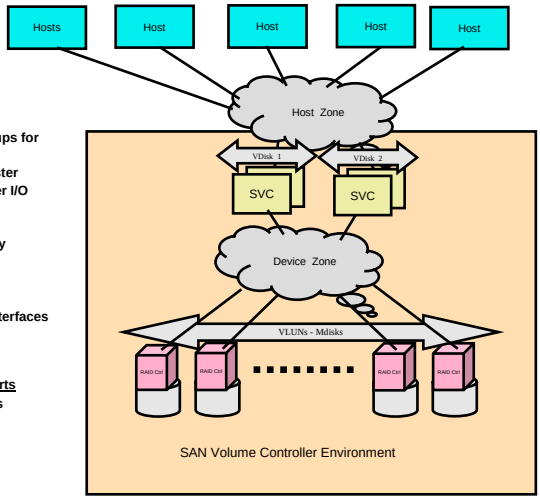


SVC - Supported Environment

FC Adapters
QLogic
Emulex
JNI
HP/Agilent

SVC
Nodes paired into I/O groups for high availability
Up to 4 I/O groups per cluster
8GB of read/write cache per I/O group
Point in Time Copy
Synchronous Remote Copy
Data Migration
1,024 vDisks per I/O Group
VDisk size 2TB Maximum
Command Line and GUI interfaces

Connect 128 controller ports
4,096 Managed Disks/LUNs
2PB physical storage



Connect 64 servers/LPARs
pSeries® (AIX) ®
xSeries/Intel (Windows 2000®, NT, Linux)
Sun (Solaris) ®
HP (HP-UX) ®
BladeCenter®
VMware®

FC 1Gb/2Gb Switches
Brocade
CNT/InRange
McData
Cisco

RAID Controllers
FAStT200, 500, 600, 700, 900
IBM ESS F20. 800, 750
HDS Thunder 92xx, 95xx
HDS Lightning 99xx
HPQ MA8000, 12000,16000
HPQ EVA 3000/5000
EMC/.Dell CX200, 400, 600
EMC Symm 8xxx
EMC Clarion FC4700

SVC - Clustering

- Cluster comprised of 2-8 storage engines or nodes but administered as single image
- One node automatically designated config/boss node for cluster
  - Assigned cluster IP address and responsible for coordination of node transitions
- Auto restart of a node on failure and re-admission to cluster via the management module
- Cluster requires majority of nodes remain operating to ensure quorum
  - Quorum disk used as tie-breaker
- Node stores writes in its cache and the write cache of its partner node - fast write mode
- On node failure, surviving node empties write cache and proceeds in write-through mode
- Utilizes 4K byte segments similar to ESS
- UPS/battery to destage and fail gracefully
- SDD manages multiple paths to ports on 2 SVC nodes for each virtual disk
  - Maximum of 8 paths from a host to a VDISK
- SDD performs failover, in case of host path or SVC node failure

RAS - Management Module

- Front bezel service panel displays error messages and used for initial configuration
- Provides WWPN for Agilent fibre channel HBAs
- Cluster/Node ID generation
- Watchdog timer/Deadman's handle
  - Deals with uncommunicative nodes
  - Monitors processor activity, if it detects code is hung, can warmstart code
  - I/O process can be restarted leaving memory intact
  - Data can be saved on power failure even with crashed I/O process
  - Allows a crashed node to be restarted by power cycling
- Interfaces with UPS to prevent loss of data during power failure

RAS - Power Loss

- Write cache, cluster configuration, and cluster metadata must be preserved in the event of a power loss
  - Assumes if SVC lost power, disk subsystem did as well
  - Must ensure no loss of data when power is restored
  - A UPS is used to provide power to SVC until DRAM contents is saved on the internal disk
  - Upon power restoration, SVC cluster restarted and cache rebuilt from disk image

RAS - Maintenance

- Concurrent Software & Hardware maintenance
  - Customer responsible for software upgrades
  - Upgrade all nodes or rollback to previous release
  - New node added to cluster, automatically upgraded or downgraded to running software version for that cluster -- Autonomic, self healing design
  - Add SVC nodes, and disk storage concurrently
- Concurrent test, repair and reconfiguration of nodes
  - Hardware repair to one node in pair - data access continues on other node
  - If HDD failure, CE replaces and SVC cluster rebuilds automatically
- Restoring repaired node to a cluster requires no knowledge of the cluster configuration or software levels - software and configuration data are automatically restored
- If cluster can't be started, service IP address used to access a node
  - Button combination on front panel of node enables Ethernet port with service IP address for access via web browser of error logs and dump information
- Call Home capability of SVC notifies IBM of hardware problem to dispatch CE

Master Console

Functionality
- Single platform for Configuration & Service
- Facilitates all install/upgrade and normal operations
- Provides Call Home capability
- Provides Remote Service capability with VPN
- SAN Topology rendering
- Access to all reference documentation

Components
- 1U Rack Mounted xServer (2 GHz/100MHz)
  - 1 GB of Memory
  - Dual 40 GB HDD
  - 2 Ethernet Ports
  - 2 Fibre Channel Ports
- 1U Rack Mounted LCD and Keyboard
- Windows 2000 Server
- CIM Agent and Console for SVC
- IBM Director V4.1
- PuTTY for Open SSH Support
- Java 1.4 plugin
- FAStT Storage Manager Client
- Tivoli SAN Manager V1.2 from Bonus Pack (64 Ports)
- Connection Manager for VPN
- Service Agent and e-Gate
- Adobe Acrobat for Publications

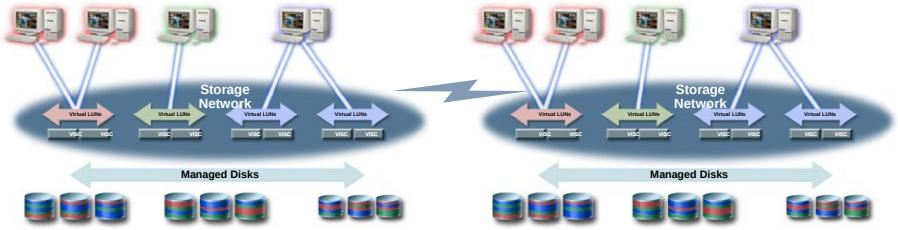
Common Platform for Advanced Functions

Single point for copy services

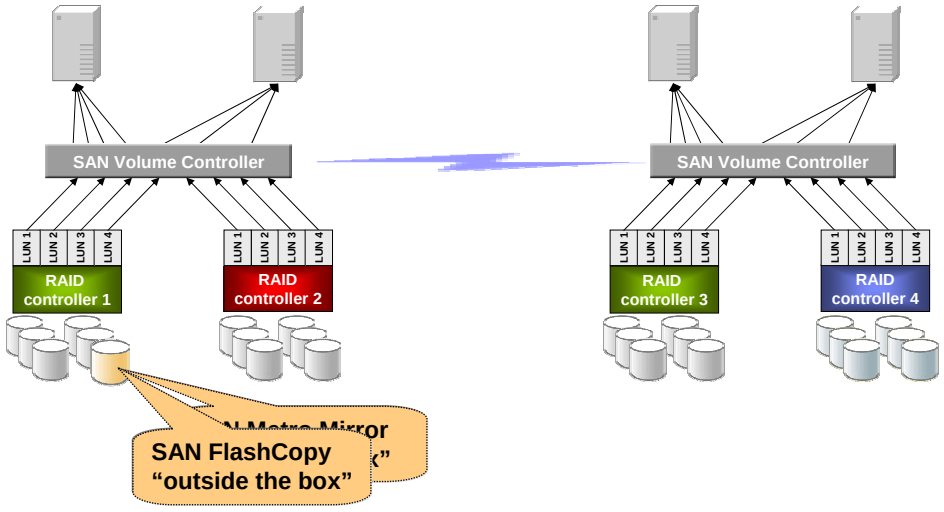
- Point-in-time copy/FlashCopy ®
- Synchronous remote copy/PPRC
- Data migration

Use to meet business needs

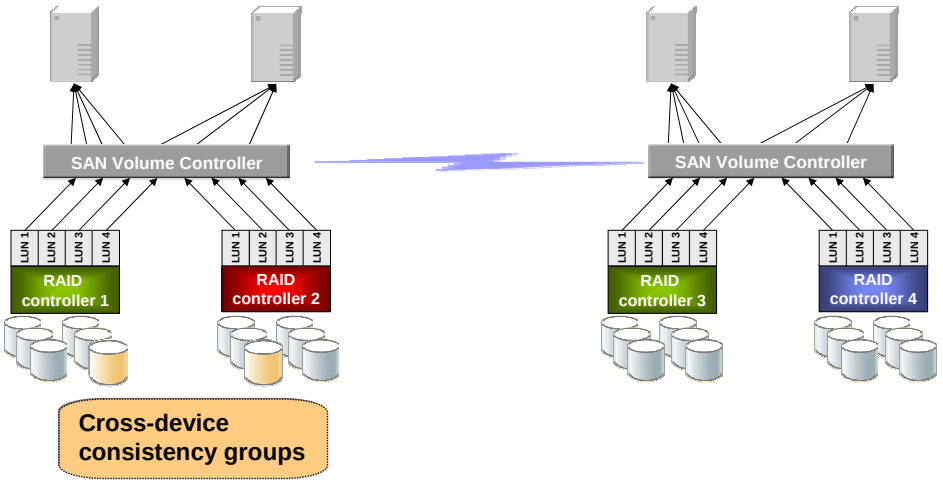
- Disaster recovery
- LAN free backup
- Server free backup
- Storage server replacement



SVC Copy Services



SVC Copy Services



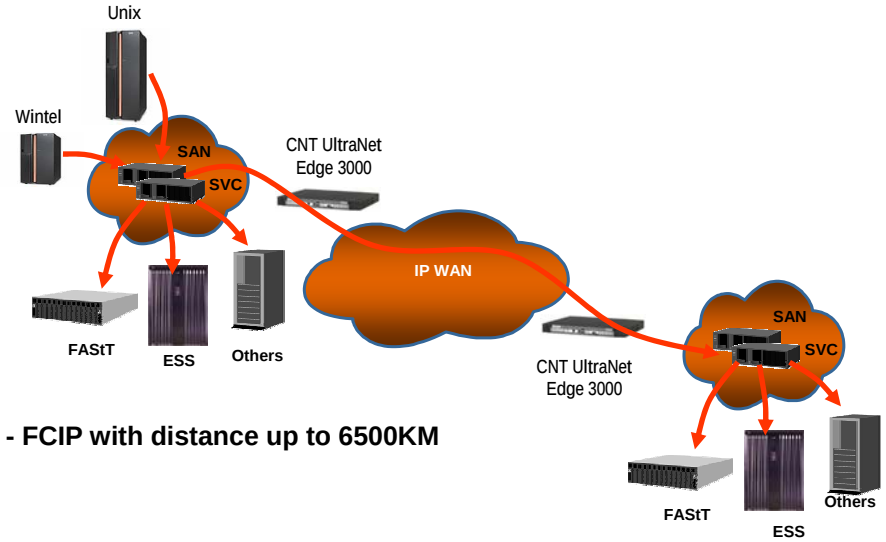
Copy Services - FlashCopy

- One to one mapping of Source to Target virtual disk
  - No multiple relationship support in V1.2
- Source and Target virtual disk must be the same size
  - No space efficient FlashCopy in V1.2
- No incremental support in V1.2
- Source/Target vDisks must be within same cluster but can be across I/O groups
- Source volume may be spread across multiple disk subsystems
- Target volume may be to one or more disk subsystems, different than the source
- Licensed independently of base virtualization software and PPRC

Copy Services - PPRC

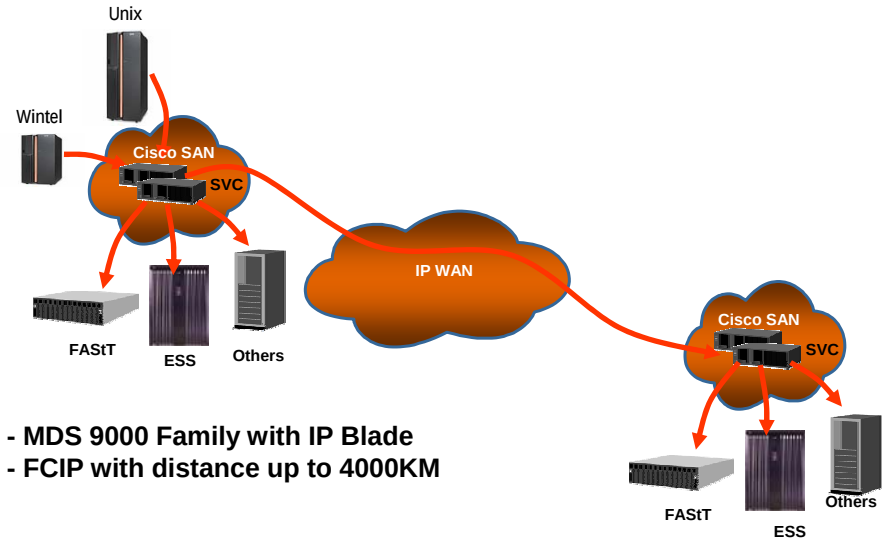
- Synchronous remote copy occurs between Source and Target virtual disks
  - Acknowledgment of write given to host when data has been written to secondary site
- Intra-cluster remote copy supported
  - Both virtual disks belong to the same cluster
- Inter-cluster remote copy supported
  - One virtual disk comes from each of two clusters
- Source volume may be spread across multiple disk subsystems
- Target volume may be to one or more disk subsystems, different than the source
- Licensed independently of base virtualization software and FlashCopy

SVC and Long Distance Sync PPRC with CNT



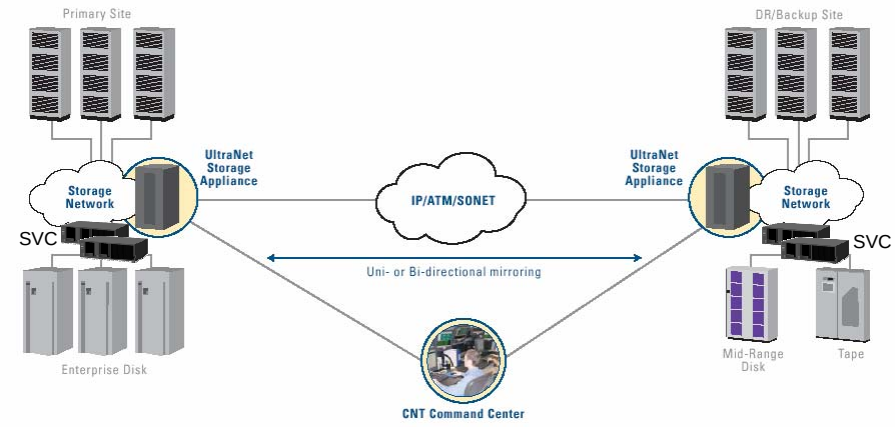
- FCIP with distance up to 6500KM

SVC and Long Distance Sync PPRC with Cisco



- MDS 9000 Family with IP Blade
- FCIP with distance up to 4000KM

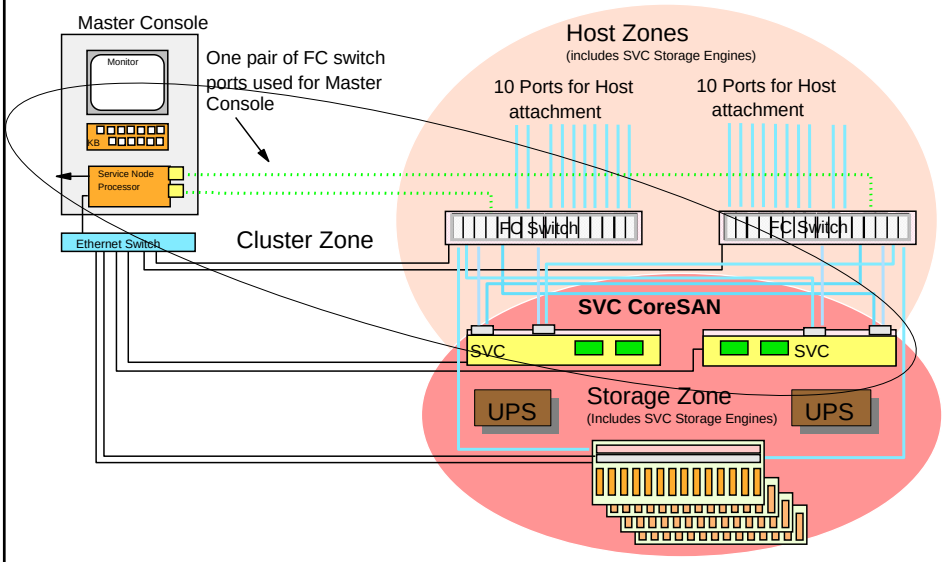
SVC and Async PPRC with CNT
UltraNet Storage Appliance



Zoning for the SAN Volume Controller

- **Zoning is a way to carve up a SAN fabric**
  - Entities in a zone can only access other entities in the same zone
  - Zones may overlap (Entities can be in more than one zone and see everything in all zones for which it is a member)
  - Data traffic only. Fabric-based traffic not affected
- **SVC requires three kinds of zones (SVC sees everything)**
  - Zone for SVC ports and Master Console HBAs
  - Zone(s) for SVC ports and Storage Device ports
  - Zone(s) for SVC ports and Host HBAs
- **Extra zone when doing Remote Copy between two clusters**
  - Zone with all SVC ports from both clusters
  - Do NOT allow one cluster to see other cluster's storage

SVC - Sample Configuration



SVC - Virtual Disk Modes

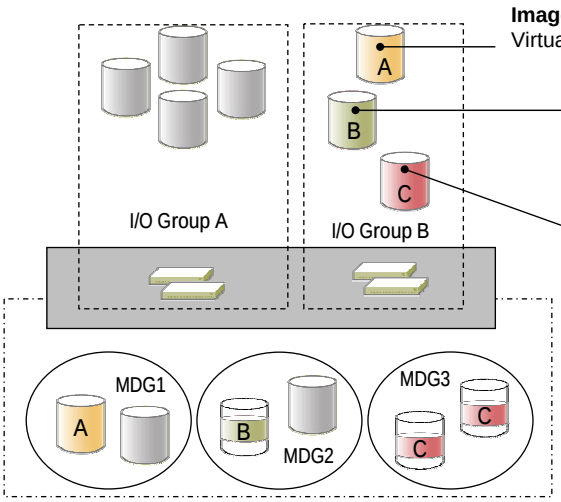
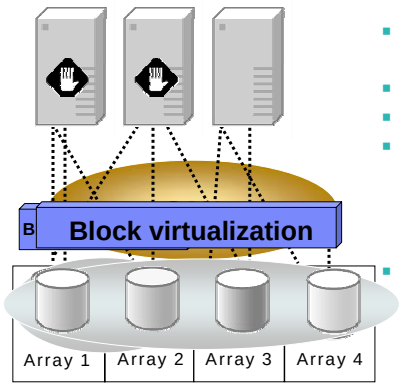


Image Mode:
Virtual Disk = Physical LUN

Sequential Mode:
Virtual Disk mapped sequentially to a portion of a managed disk

Striped Mode:
Virtual Disk striped across multiple managed disks

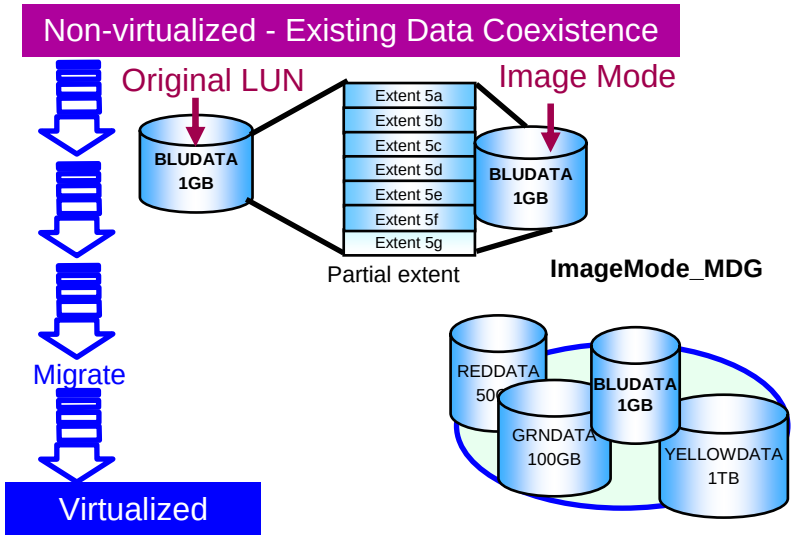
Migration from Existing Environment



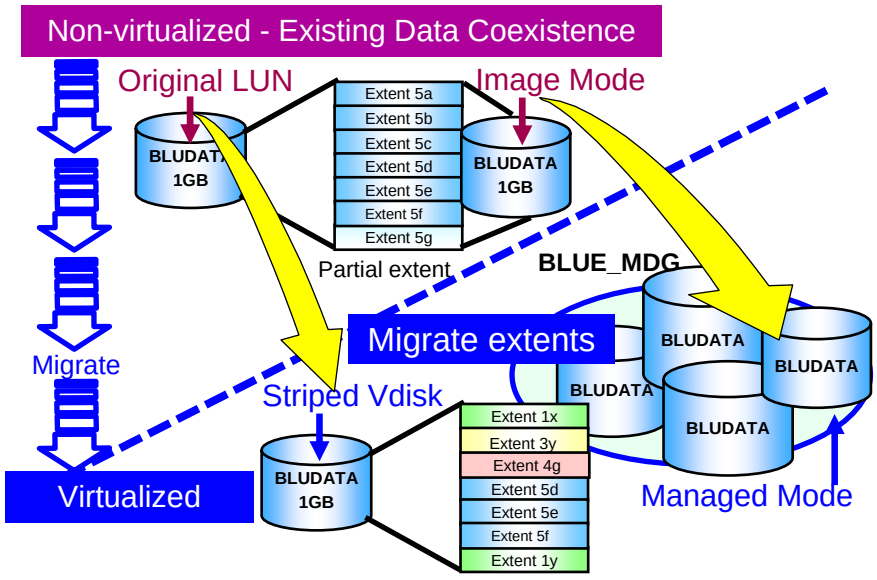
Evolutionary steps

- **Install** SAN Volume Controller
- **Pause I/O to storage chosen for migration**
- **Add existing LUNs to SAN Volume Controller in image mode**
- **Reconfigure host LUNs to SAN Volume Controller**
- **Restart applications**
- No data movement required
- But…arrays may now be managed as a virtualized pool
  - **Data moved, striped, rebalanced**
  - **Application servers unaware of physical changes**
- Evolve the rest of the SAN in the same manner
  - **At already planned downtime**
  - **As fast or slow as you need**

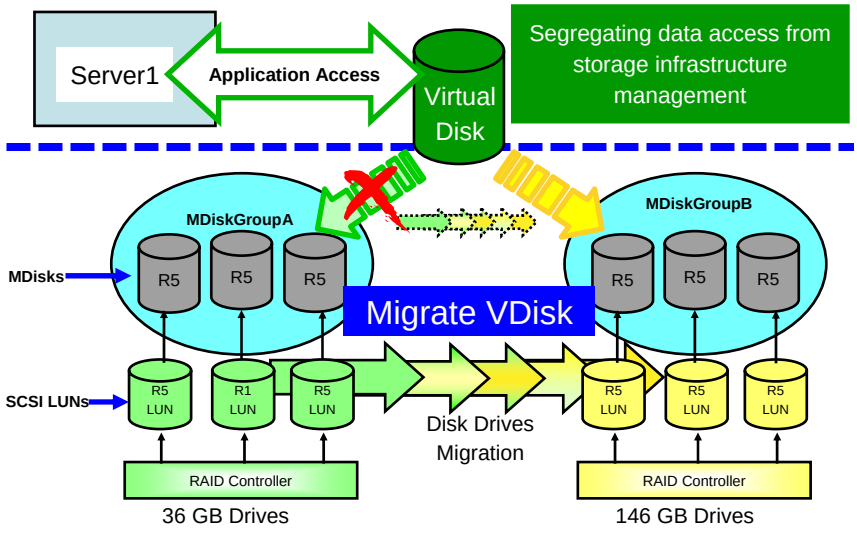
Data Migration: Existing Data



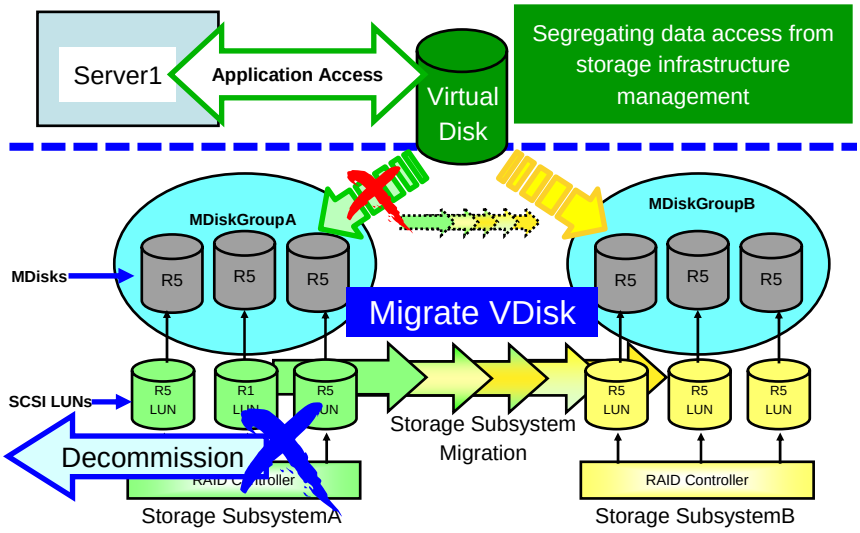
Data Migration: Virtualize Existing Data



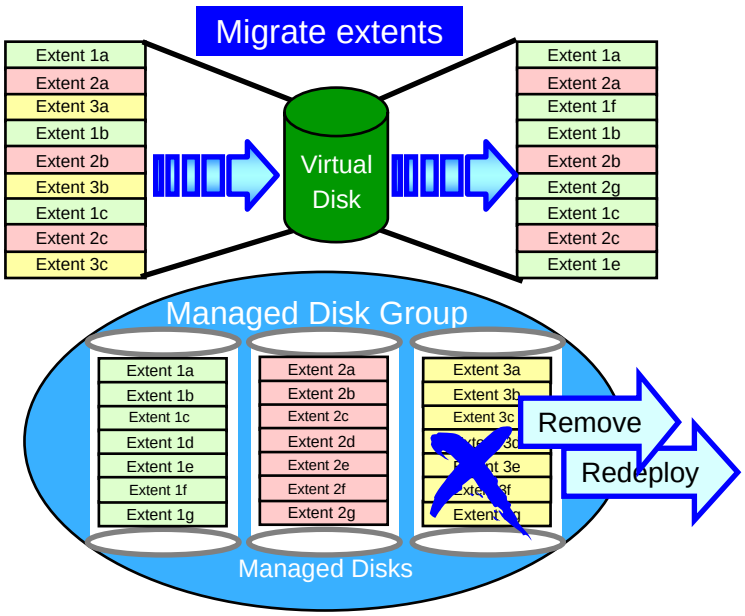
Data Migration: Replacing Disk Drives within a Storage Subsystem



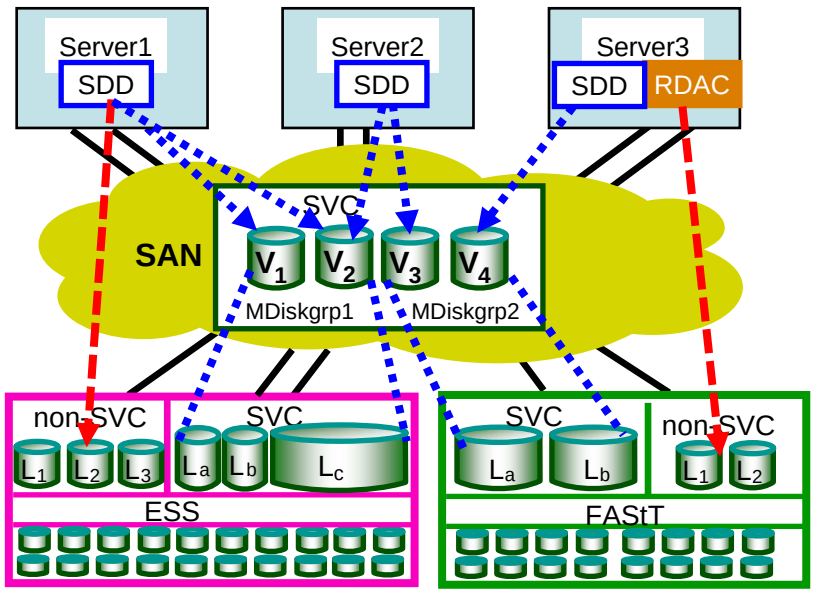
Data Migration: Replace Storage Subsystem

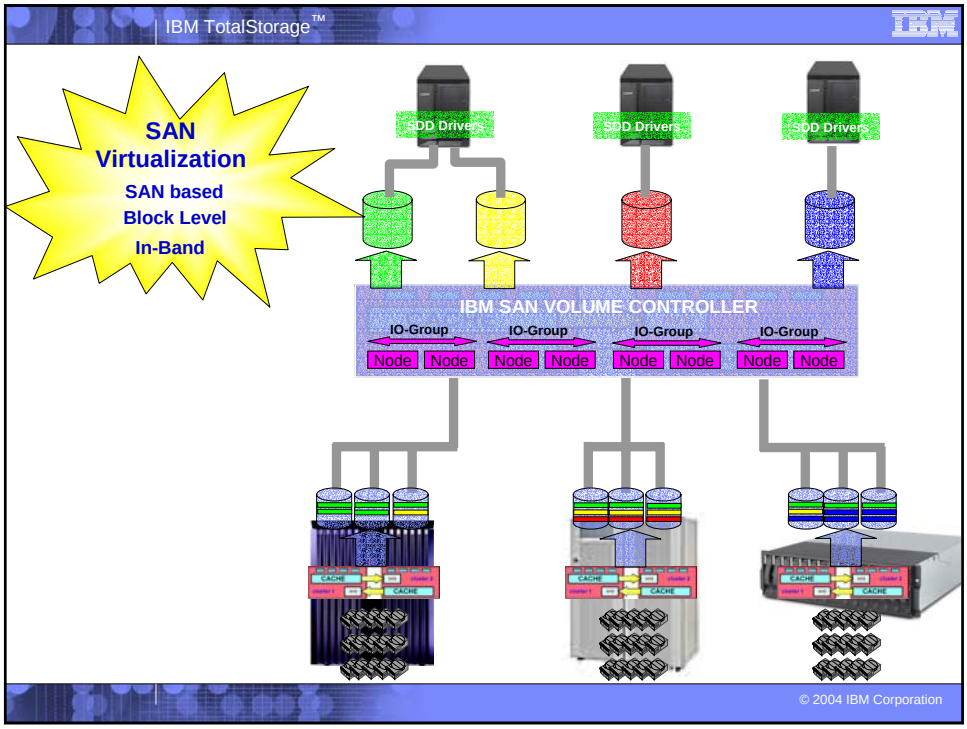


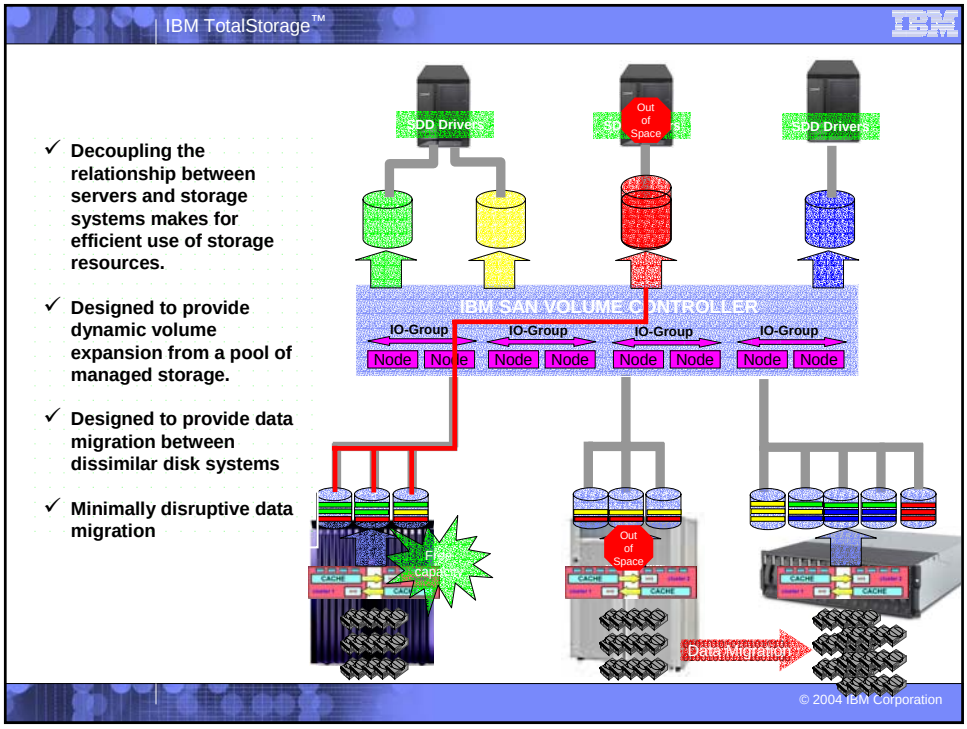
Data Migration: Redistribute MDisk Usage



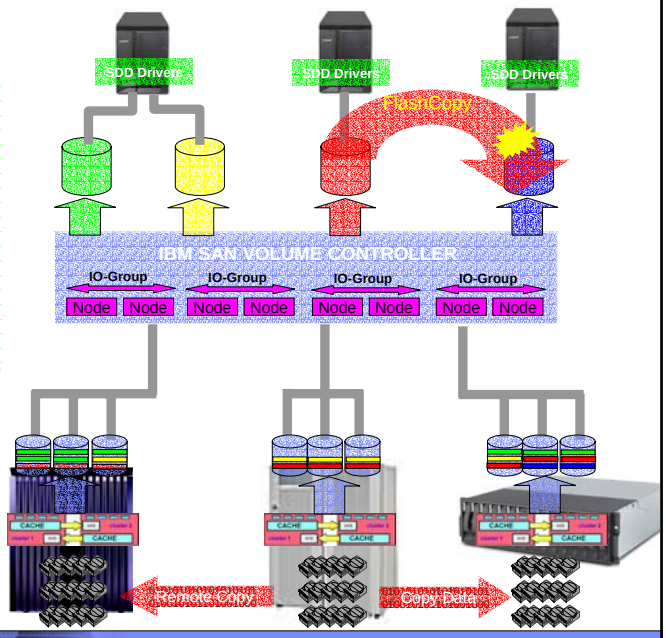
SVC and non-SVC Storage Subsystem Sharing




IBM TotalStorage™
SAN
Virtualization
SAN based
Block Level
In-Band
SDD Drivers
SDD Drivers
SDD Drivers
IBM SAN VOLUME CONTROLLER
IO-Group
IO-Group
IO-Group
IO-Group
Node
Node
Node
Node
Node
Node
Node
Node
© 2004 IBM Corporation


IBM TotalStorage™
✓ Decoupling the relationship between servers and storage systems makes for efficient use of storage resources.
✓ Designed to provide dynamic volume expansion from a pool of managed storage.
✓ Designed to provide data migration between dissimilar disk systems
✓ Minimally disruptive data migration
SDD Drivers
Out of Space
SDD Drivers
IBM SAN VOLUME CONTROLLER
IO-Group
IO-Group
IO-Group
IO-Group
Node
Node
Node
Node
Node
Node
Node
Node
Out of Space
Data Migration
© 2004 IBM Corporation

- ✓ **Decoupling the relationship between servers and storage systems makes for efficient use of storage resources.**
- ✓ **FlashCopy and PPRC can be between dissimilar disk systems.**
- ✓ **Designed to offer flexible, SAN wide copy services.**



Performance Planning Guidelines

- **In general, configure disk systems as you would without SVC**
  - Disk drives
    - 73 GB disks are recommended for most environments
    - For very demanding environments, consider 36 GB, 15K RPM disks
    - 146 GB drives offer lower cost for less active data and as FlashCopy targets
  - RAID types
    - RAID-5 suggested in most cases
    - SVC does not provide any RAID capability
  - Array sizes
    - 8+P or 4+P suggested for FAStT disk family
    - For ESS and FAStT create LUN size equal to array
    - Create minimum of one LUN per active fibre port on disk server used with SVC
    - For ESS present LUNs to SVC from multiple loops/LSSs
    - Use FAStT segment size of 128KB, helps sequential performance

Performance Planning Guidelines

- **Latency is delay added to response time for an I/O operation**
- **In-band solutions add latency to cache read miss I/Os**
  - Not unique to SAN Volume Controller
- **SAN Volume Controller latency is very low**
  - Minimal impact, roughly 50-60 microseconds on read misses
- **"Real world" impact of latency will usually be minimal**
  - All writes are cache hits and add no latency
  - Some reads will be cache hits with no extra latency
- **SVC or SVC4MDS caching can potentially improve performance with older or uncached disk systems**
- **"Generally performance neutral" for cache insensitive workloads**

IBM TotalStorage ™

Performance Planning Guidelines

- **Existing disk systems**
  - No need to change LUNs
  - Keep same set of paths into disk system
  - Keep same number of host ports
  - Deploy virtualization as a "middle layer" between hosts and disk systems
- **Quorum disks require some extents on mDisks**
  - May wish to spread quorum disks onto multiple backend disk systems
- **SVC Managed Disk Group extent size**
  - Generally not a significant performance factor
    - Smaller extents may distribute load I/O load across managed disks better
  - Maximum cluster capacity is related to extent size
  - Smaller extents may help reduce wasted space

© 2004 IBM Corporation

Performance Planning Guidelines

- **All mDisks in an MDG should have similar performance**
  - Same drive size, speed, RAID type
  - Otherwise, may get “lumpy” performance within vDisks
- **Preferred path for vDisks in an I/O Group can be tailored**
  - Default algorithm will usually provide good results
    - SVC alternates vDisks across nodes in the order created
  - “Unusual” configurations could cause concerns
    - Very different vDisk sizes, wide variations in I/O load per vDisk
- **Consider striped vDisk layout**
  - Default choice
  - Balances load across physical disks

Performance Planning Guidelines

- **Performance scales very well as I/O Groups added to SVC cluster**
- **Configuration different from traditional disk systems**
  - SVC is not a disk system itself; it needs “back-end” disk systems
- **But ... configuration also similar to traditional disk systems**
  - Need to ensure “back-end” storage can deliver “front-end” requirements
  - Can aggregate together performance from “back-end” storage to deliver overall system throughput
- **Review Part 7 in the SVC Configuration Guide for assistance with sizing of backend storage pool to accommodate host workload. This chapter includes information on calculating physical disk requirements based on I/O rates expected. There must be adequate backend capacity in terms of physical disk spindles for SVC to perform as expected.**

SVC Reference Materials

- Websites for marketing information and SVC supported environments
  - **http://www.ibm.com/storage/software/virtualization**
  - **http://www.ibm.com/storage/support/2145**
- Publications
  - **http://www.ibm.com/shop/publications/order**
  - Planning Guide – GA22-1052
  - Installation and Hardware Reference Guide – SC26-7541
  - Service Guide – SC26-7542
  - Configuration Guide – SC26-7543
  - Command-Line Interface User's Guide – SC26-7544
  - CIM Agent Developer's Reference – SC26-7545
  - Host Attachment Guide – SC26-7563
- Redbooks
  - **http://www.ibm.com/redbooks**
  - IBM TotalStorage SVC and SIS – SG24-6423

SAN Volume Controller - Value Proposition

- Increase Storage Administrator Productivity
- Enable Advanced Copy Services across a Virtual SAN
- Improve Capacity Utilization
- Increased Data Availability and Protection
- Enhanced Modular Scalability
- Supports Heterogeneous storage
- Facilitates Migration of Data from Outmoded Storage Assets
- Architected to Open Standards
  - IBM adds value where it counts most, and avoids proprietary technology where there are open standards

Designed to reduce the complexity and costs involved in managing SAN-based storage

Trademarks

The following terms are trademarks or registered trademarks of the IBM Corporation in either the United States, other countries or both.

IBM, Enterprise Storage Server, SSA, ESCON, FICON, OS/390, S/390, RS/6000, AIX, Netfinity, z/OS, zSeries, pSeries, iSeries, xSeries and TotalStorage

Windows 2000, Windows NT, Windows 95 and Windows 98 are registered trademarks of Microsoft Corporation.

UNIX is a registered trademark in the United States and other countries, licensed exclusively through X/Open Company Limited.

Tivoli and Tivoli Storage Manager are registered trademarks of Tivoli.

Intel in a registered trademark of Intel Corporation.

HP and HP-UX are trademarks of Hewlett-Packard Company.

Sun and Solaris are trademarks of Sun Microsystems, Inc.

Brocade, SilkWorm, Extended Fabrics, Remote Switch, Fabric OS, Fabric Watch, QuickLoop, Zoning, Inter-Switch Link Trunking are trademarks or registered trademarks of Brocade Communications System, Inc.

INRANGE is a registered trademark of INRANGE Technologies Corporation. FC/9000 and IN-VSN are trademarks of INRANGE Technologies Corporation.

McDATA and Fibre Channel Director are a trademark or registered trademark of McDATA Corporation.

Nortel is a registered trademark of Nortel Networks.

Cisco, MDS is a registered trademark of Cisco Systems.

Other company, product, and service names may be trademarks or registered trademarks of their respective companies.

Disclaimers

Product data is accurate as of initial publication and is subject to change without notice.

No part of this presentation may be reproduced or transmitted in any form without written permission from IBM Corporation.

References in this document to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM program product in this document is not intended to state or imply that only IBM's program product may be used. Any functionally equivalent program may be used instead.

The information provided in this document has not been submitted to any formal IBM test and is distributed "As Is" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into their operating environment.

While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

This statement contains information providing general direction on IBM's product plans. Such plans are subject to change without notice and IBM may not make such products available.

Thank You!!!

Questions???