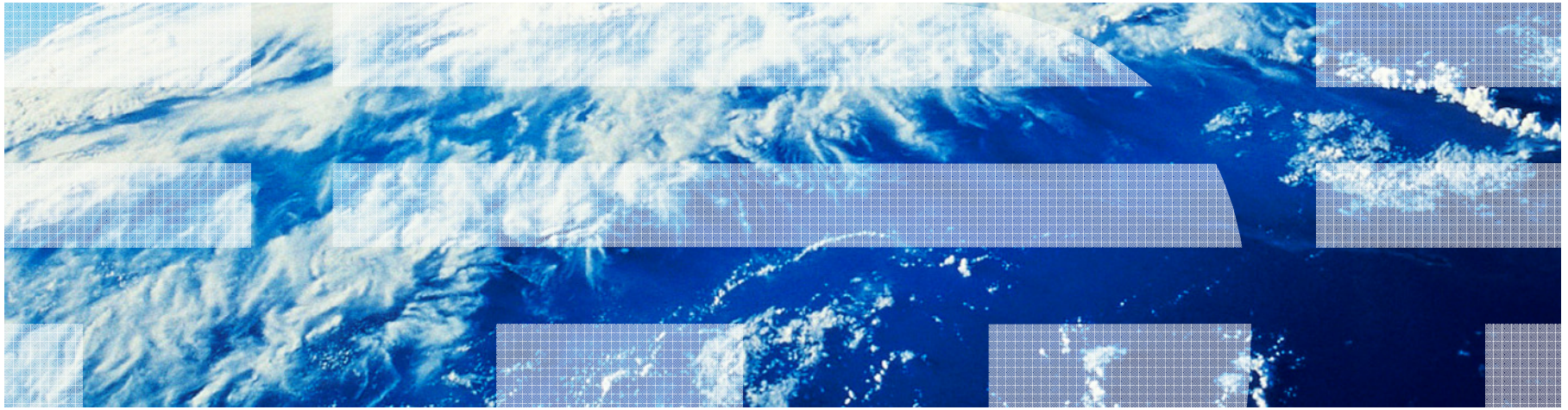
IBM Storage Solutions for High Performance Computing and Digital Media

Sandro De Santis *(sandro1_desantis@it.ibm.com)*



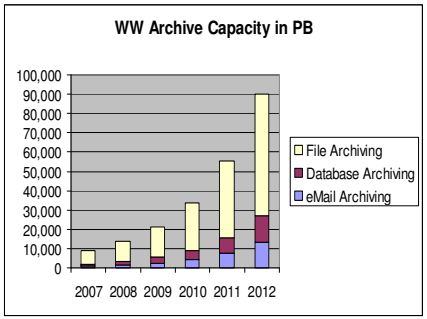
Agenda

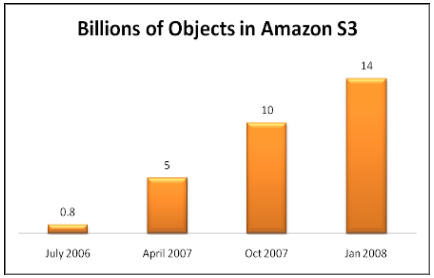
- Current scenario and the IT laws
  - Disk Technology Directions
- Virtualization: why & where
- IBM Solutions in action:
  - Highly virtualized storage system: IBM XIV
  - Virtualized storage infrastructure with IBM SVC
  - Streamline your back-up & restore process with IBM TS7650
- Virtualization, Optimization, Consolidation, ...
- Storage as a Service (for the Cloud)
- HPC
  - introduction
  - HPC requirement
  - IBM GPFS
  - DS5000
  - DCS9900
  - Future
    - SONAS
    - DCS3700
- Summary

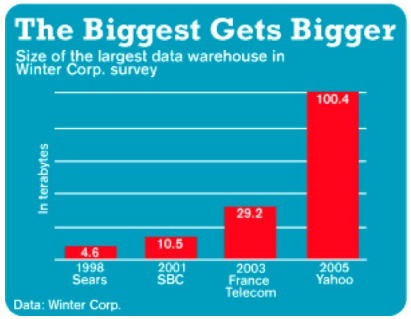
... the real mission

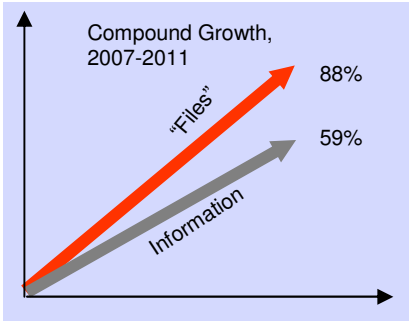
Drivers of Data Growth

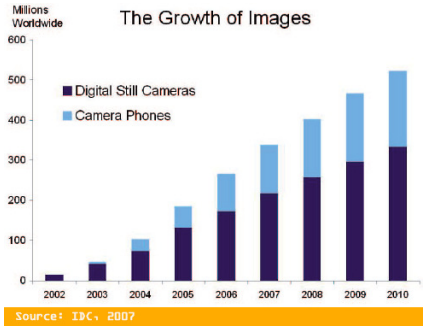
- An interconnected world creates a growing demand for data
  - Points of data create/capture
    - Video Cameras, Cell phones, Sensors, Satellites, Scanners, …
  - Points of access
    - Mobile, Laptop/desktop, Professional workstations, Home entertainment, …
  - Large object creation with file access
  - Long term retention
- Key Use Cases
  - Medical imaging for active data and archive
  - Digital simulations in Pharma, Automotive, Aerospace
  - Rich content records in Insurance, Construction, Realty
  - Video capture for security, process management, education
  - Content distribution in Media & Entertainment
  - Web 2.0 and Social Networking
  - Financial Services - Analytics

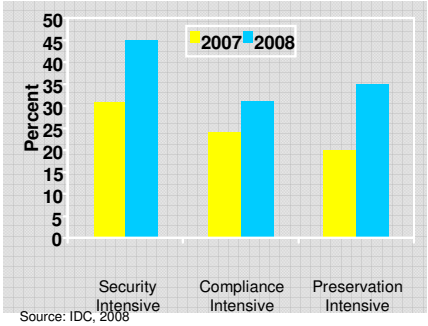












Source: IDC, 2008

Disk Technology Directions

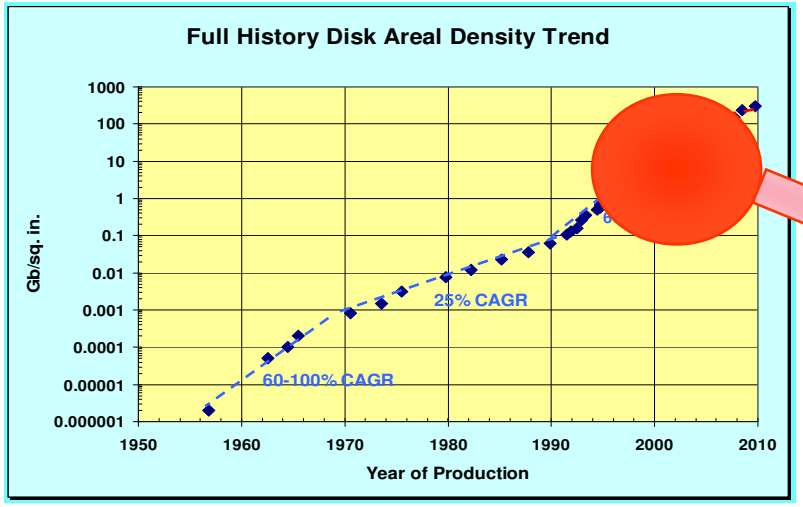
Change in Disk Areal Density Trend

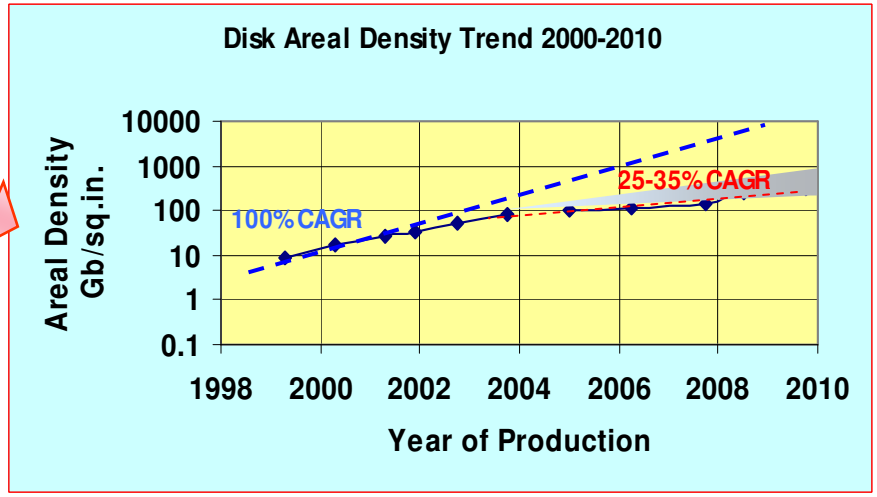
Historic trend

- From 1957 until today CAGR of Areal Density as averaged 30-35%/yr
- Between ~1990 and ~2004 it averaged 60-100%/yr

Current Decade

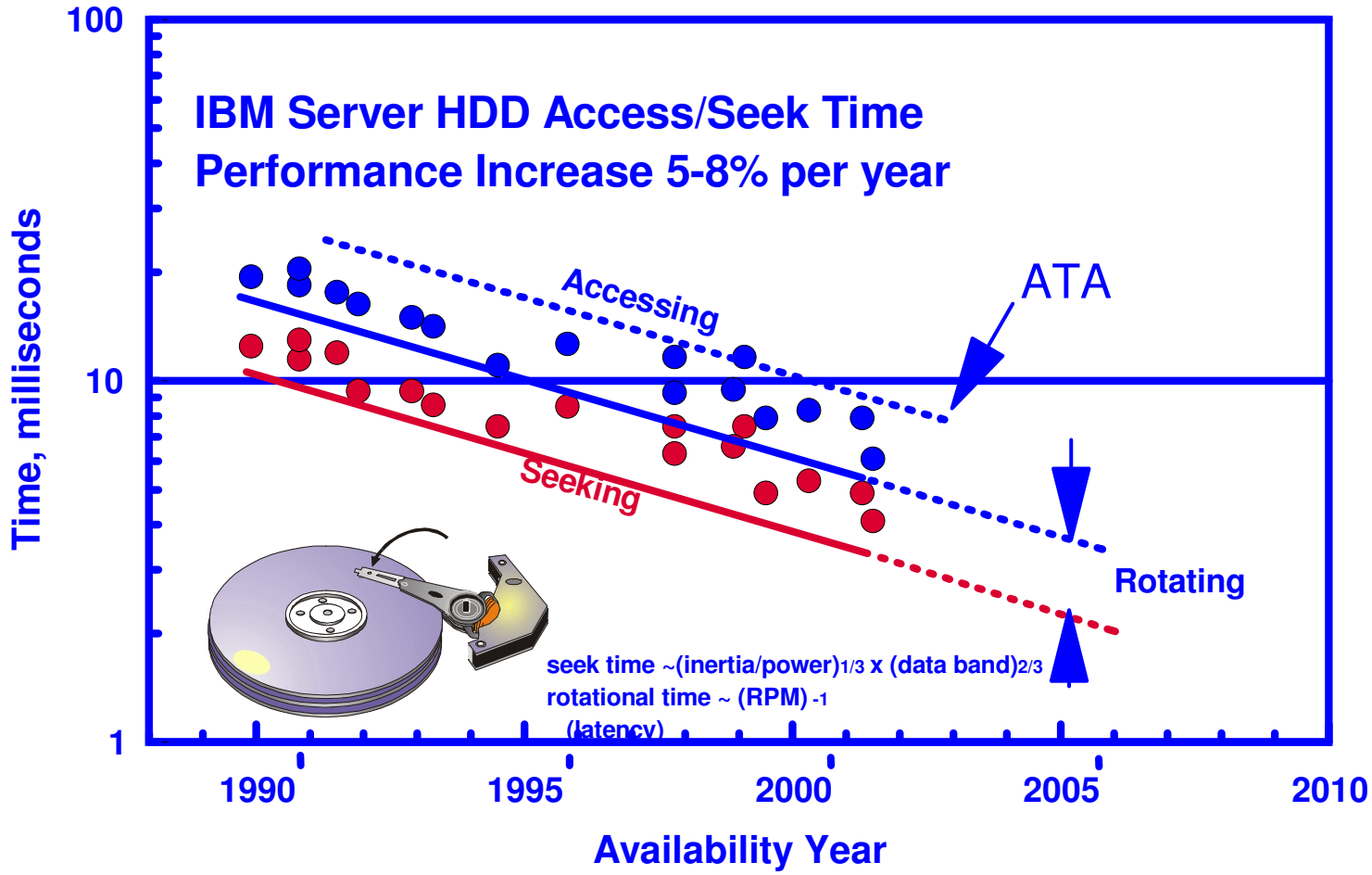
- Around 2004 the CAGR dropped to 25-35%/yr
- Disk vendors indicate that it will continue at 25-35% CAGR for the rest of the decade
- This will have a significant impact on the price and performance of disk drives through the end of the decade



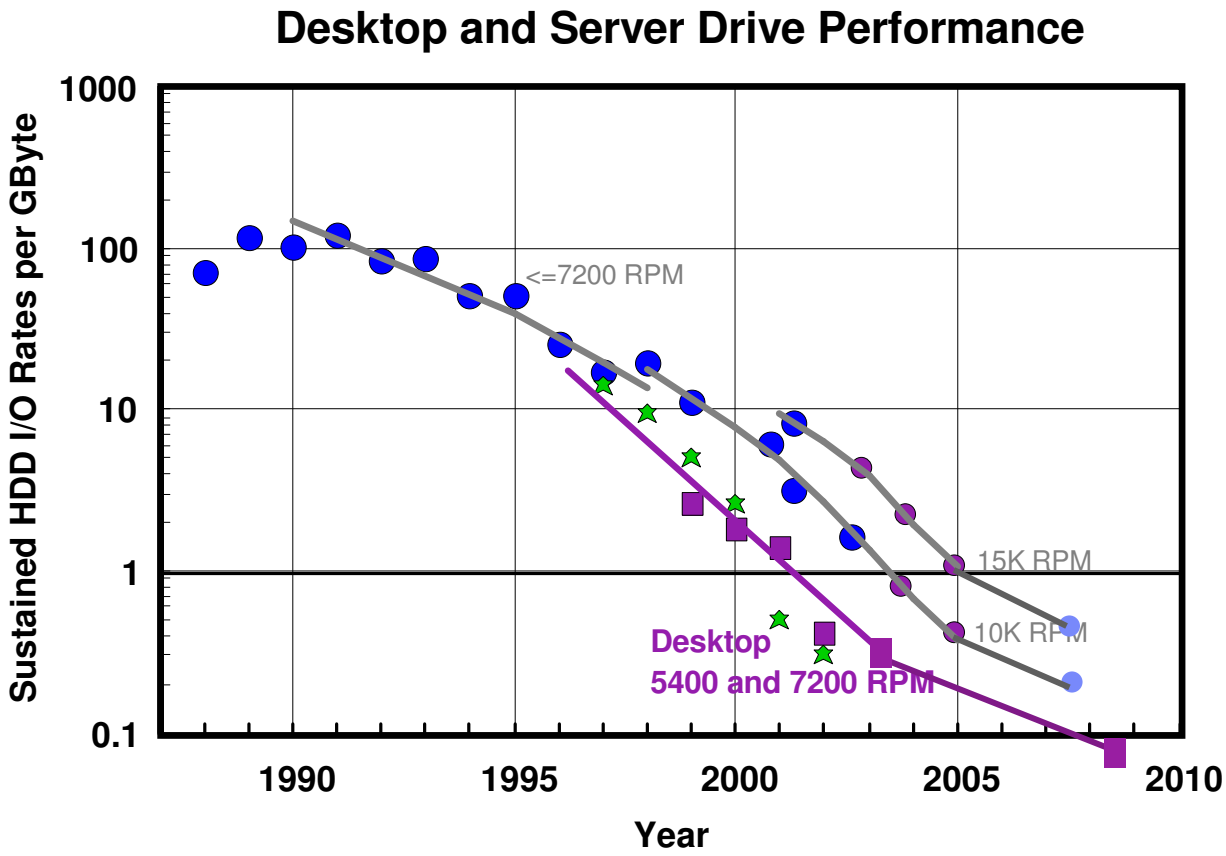


Sources—Historic: IBM & Hitachi, Current: disk vendors through IBM procurement

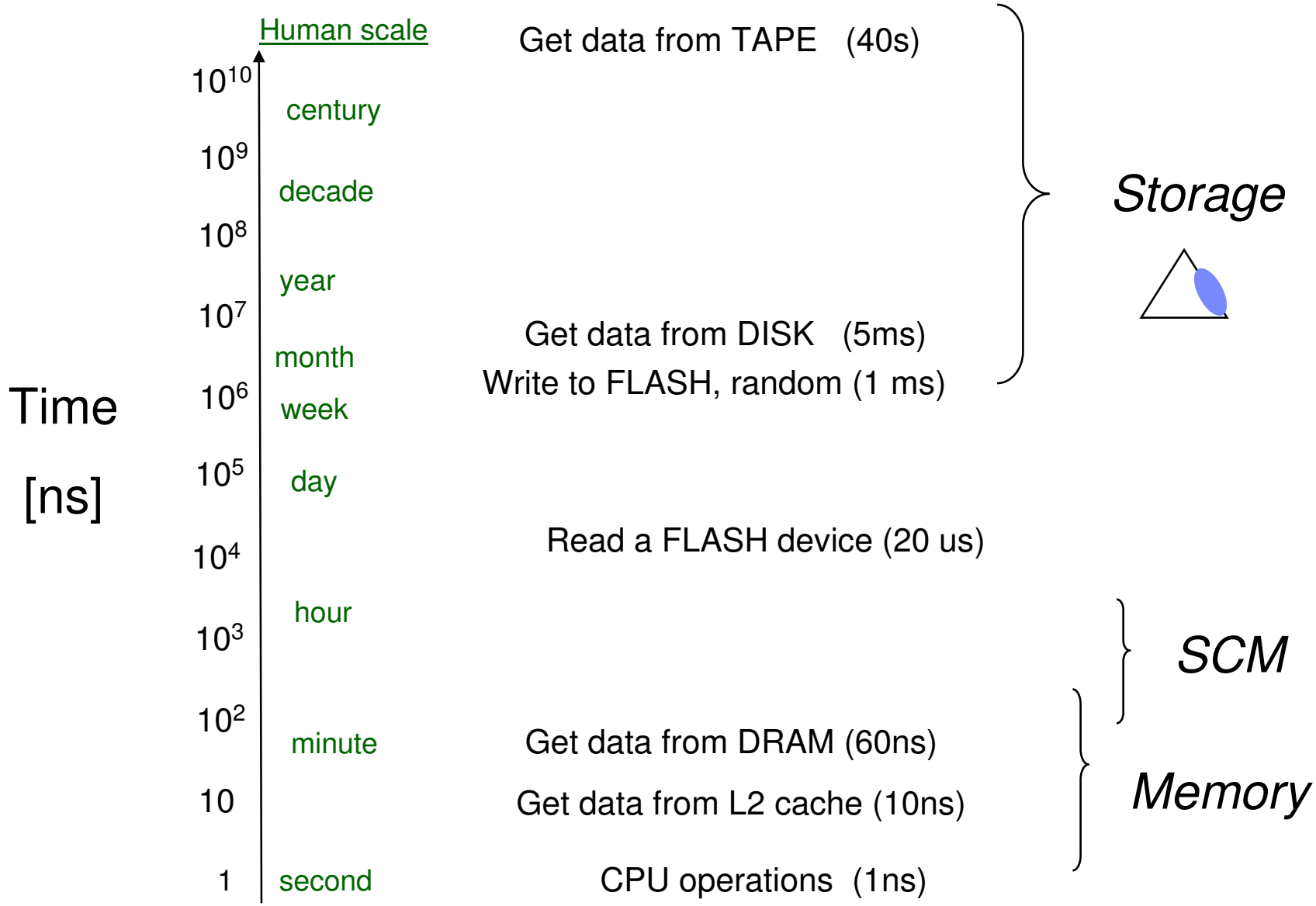
Drive Access Times Are Improving Slowly



Performance / Capacity losing ground



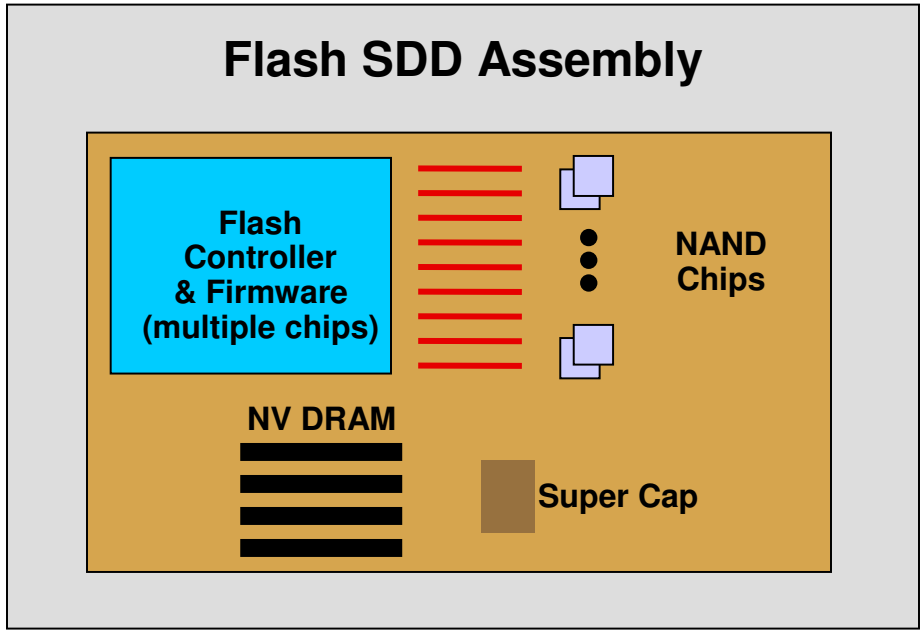
Typical System Time Scale



Flash Drives 101

Improve write endurance with wear leveling, over provisioning, and bit correction, and caching

Increase write performance via parallelism, caching and write coalescing

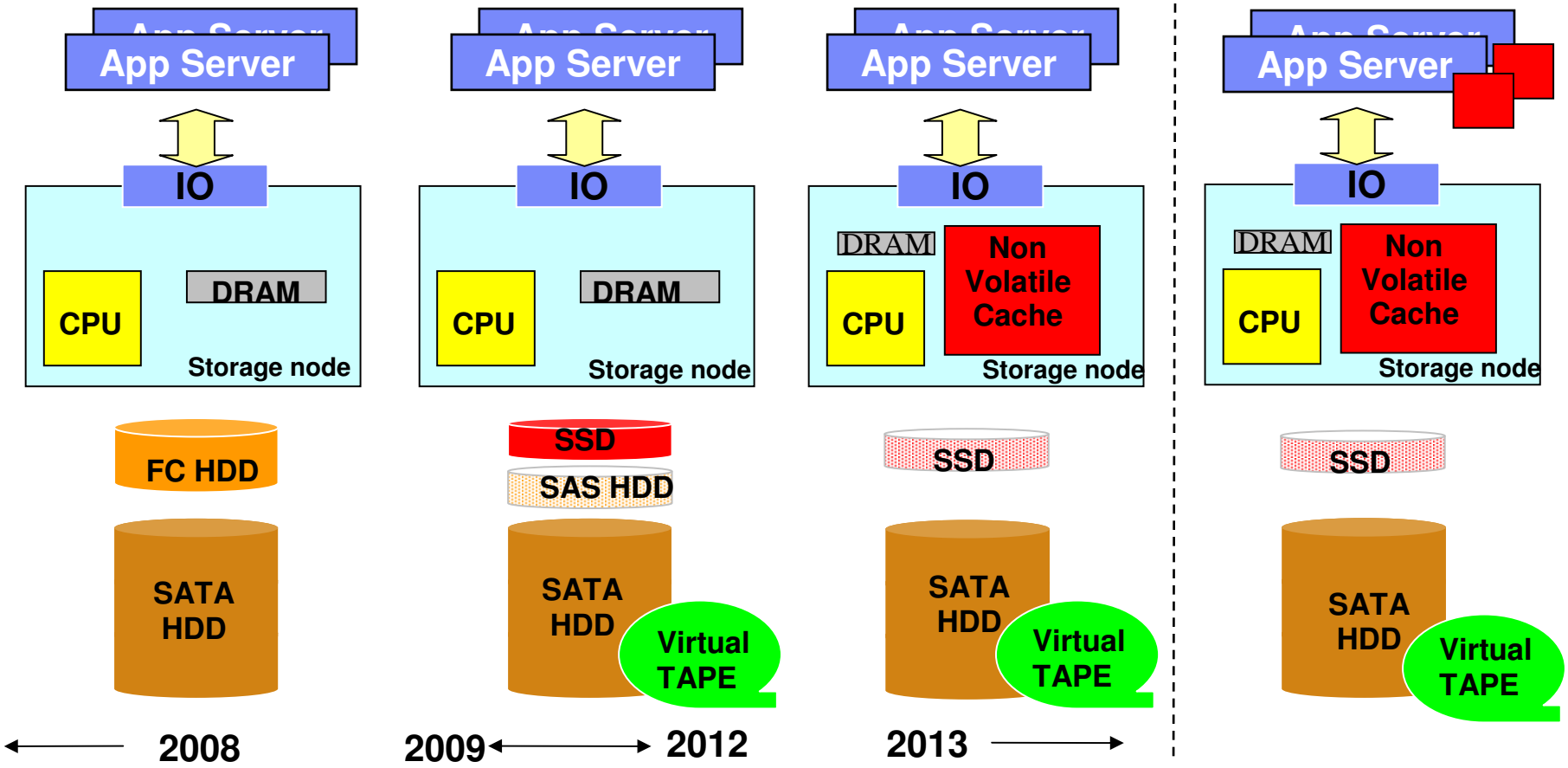


Provide reliability via over provisioning bad block detection and error correction

MLC: Multi-Level Cell Flash Technology Enables 2 or more bits per physical storage location on chip.

- Advanced flash controllers enable performance, reliability and endurance
  - Reduce writes by ~1 order of magnitude (10^5 ➔ $2x10^6$)
- Next generation controller technology will allow use of MLC Flash chips
  - Eliminate writes by ~2 orders of magnitude, makes MLC technology affordable
  - Native MLC: ~10K writes, (< 50 days of use in a server 7X24 @ 100% utilization)
  - Increased integration will further increase density and lower cost

Storage System Evolution with Flash Technology



- **Step one - start with NAND flash SSDs, emulate HDD and embed within disk arrays**
- **Step two - use as an extension of subsystem Cache**
  - Keep Flash based SSDs as devices under the cache?
- **An additional model – server integration**

Summary – Storage Device Directions

- HDD still the leader in $/GB
- SSD already ahead in $/G/Sec
  - Flash based technology will grow in use over next 3-5 years
  - Other technologies in development
- Prices of SSDs will advantage smart data placement technologies
  - Tools and Internal automation
- Tape will remain the $/GB and GB/watt leader

- Return to historic disk drive price trends
  - **Use storage more efficiently – more intensive management**
  - **Compression / Deduplication**
  - **Hierarchical storage**
  - Fosters the use of lower cost drives in the Enterprise
  - Fosters the development of redundancy technology

IT Laws

First and Second Law of IT (Thermo) dynamics

1. ***The potential complexity of an IT infrastructure is unbounded.***
2. ***Without corrective action, the complexity of an IT infrastructure increases over time.***

IT Laws (cont.)

Newton's Three Laws of IT

1. An SMP server in motion tends to stay in motion;
   Data at rest tends to stay at rest;
   Installed software tends to stay installed.

2. Force needed to change an IT infrastructure =
   (magnitude of the impact) X (rate of change)

3. For every hardware architecture change there
   is an opposing software change of equal
   (or greater) magnitude.

Major Forces Are Driving IT Infrastructure Transformation

Costs & Service Delivery

- ⇨ Rising costs of operations
- ⇨ Explosion in volume of data and information
- ⇨ Difficulty in deploying new applications & services

Business Resilience & Security

- ⇨ Growing systems & applications availability needs
- ⇨ Security of your assets & your clients' information
- ⇨ Landslide of compliance requirements

Energy Requirements

- ⇨ Rising energy costs & rising energy demand
- ⇨ Power & thermal issues inhibit operations
- ⇨ Environmental compliance & governance mandates

Technology Advances

- ⇨ Service orientation
- ⇨ End-to-end service mgmt
- ⇨ **Comprehensive virtualization**
- ⇨ **Converged networks**
- ⇨ **Solid state storage**
- ⇨ IT appliances
- ⇨ **Storage de-duplication**
- ⇨ Many cores & threads per chip
- ⇨ Low-cost high-BW fiber optics
- ⇨ Petaflop supercomputers
- ⇨ **Cloud computing services**
- ⇨ Real-time data streams

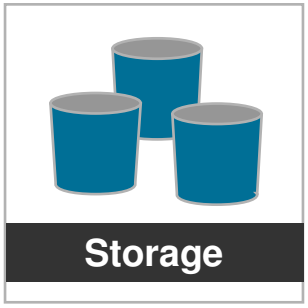
What is virtualization?

A logical representation of resources not constrained by physical limitations

Create many virtual resources within a single physical device

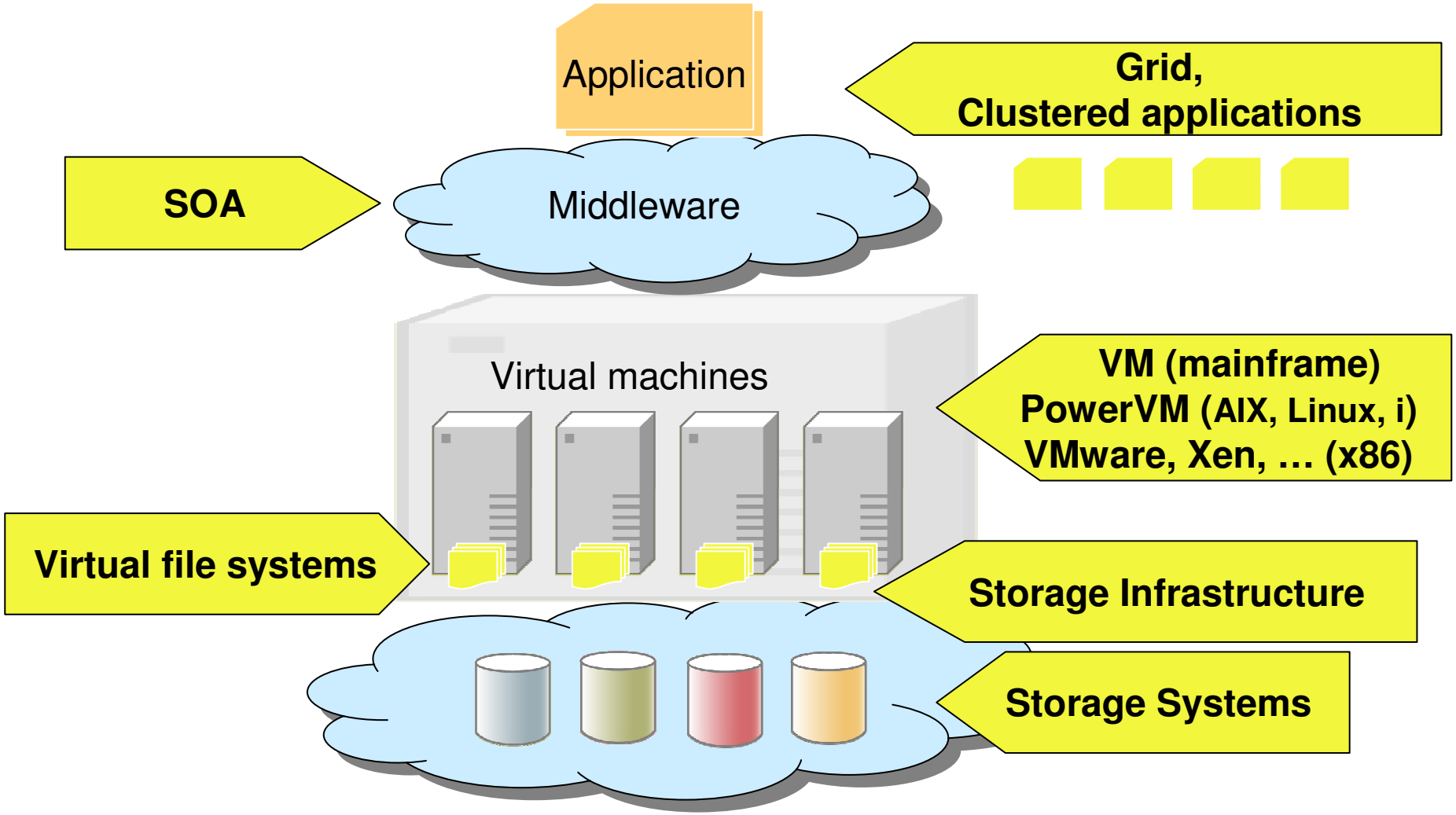
Reach beyond the box — see and manage many virtual resources as one

Dynamically change and adjust across the infrastructure



Virtualization creates an unprecedented freedom of choice

Virtualization is Everywhere



IBM Comprehensive Virtualization Offerings

Server virtualization

- **IBM System p, System i, System z LPARs, VMware ESX**
  - **Virtually consolidate workloads on servers**

File virtualization

- **Scale Out File Services. IBM General Parallel File System**
  - **Virtually consolidate files in one namespace across servers**

File system virtualization

- **IBM System Storage N series Virtual File Manager**
  - **Virtually consolidate different file *systems* into one namespace**

Disk and tape storage virtualization

- **IBM System Storage SAN Volume Controller, TS7500, TS7600, TS7700**
  - **Virtually consolidate storage into pools**

Storage Infrastructure Management

- **IBM Tivoli Storage Productivity Center**
  - **Consolidated management of virtual/physical storage resources**

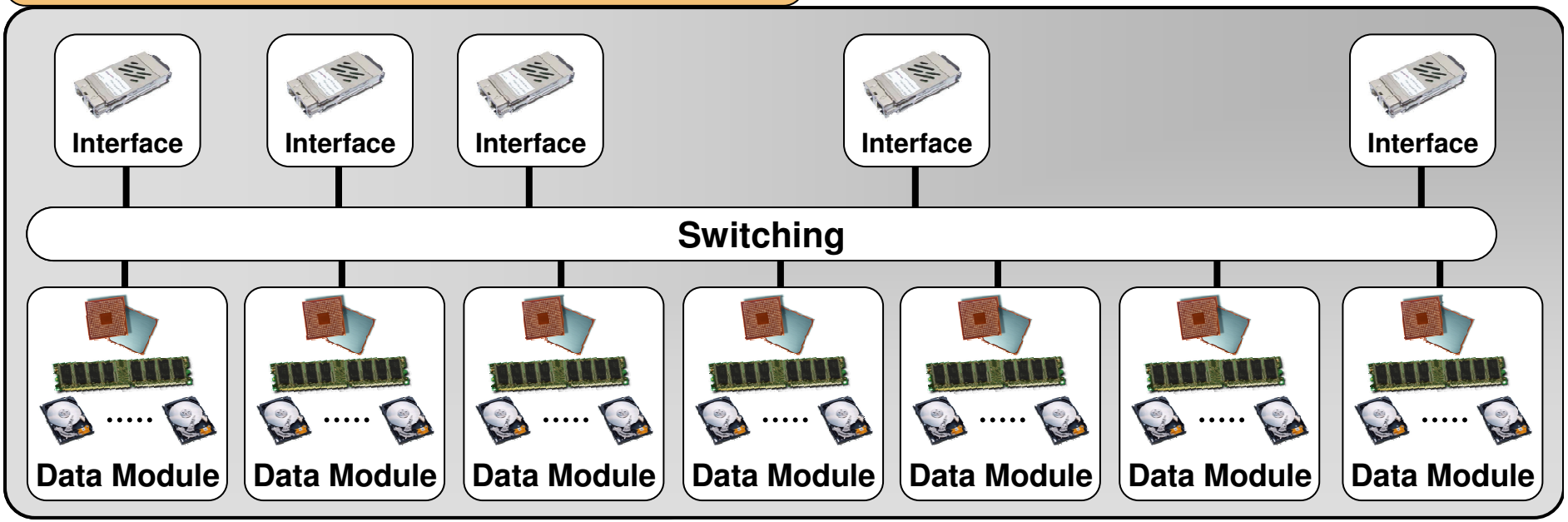
IBM Solutions in action: IBM XIV Storage System

IBM XIV Clustered Grid Architecture

Design principles:

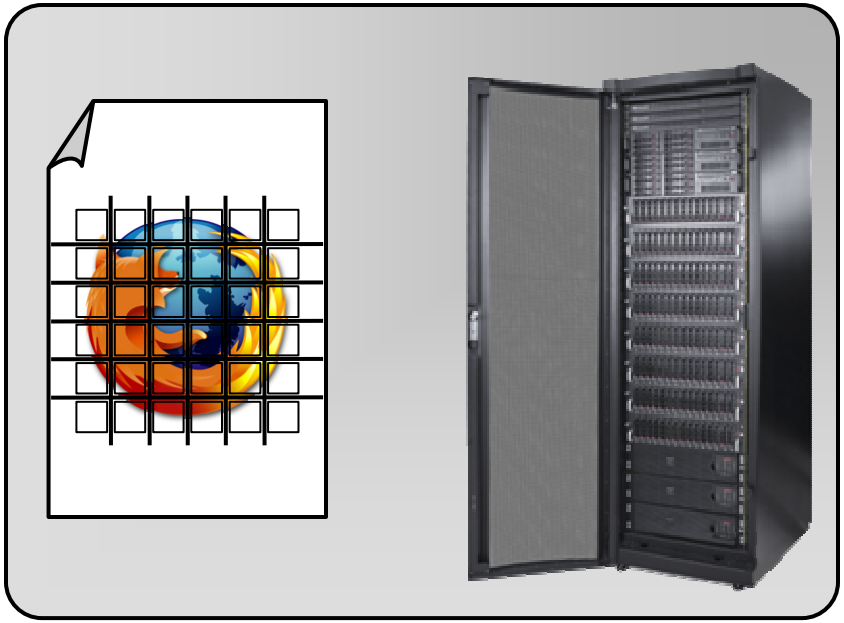
- Massive parallelism
- Granular distribution
- Off-the-shelf components
- Coupled disk, RAM and CPU
- User simplicity

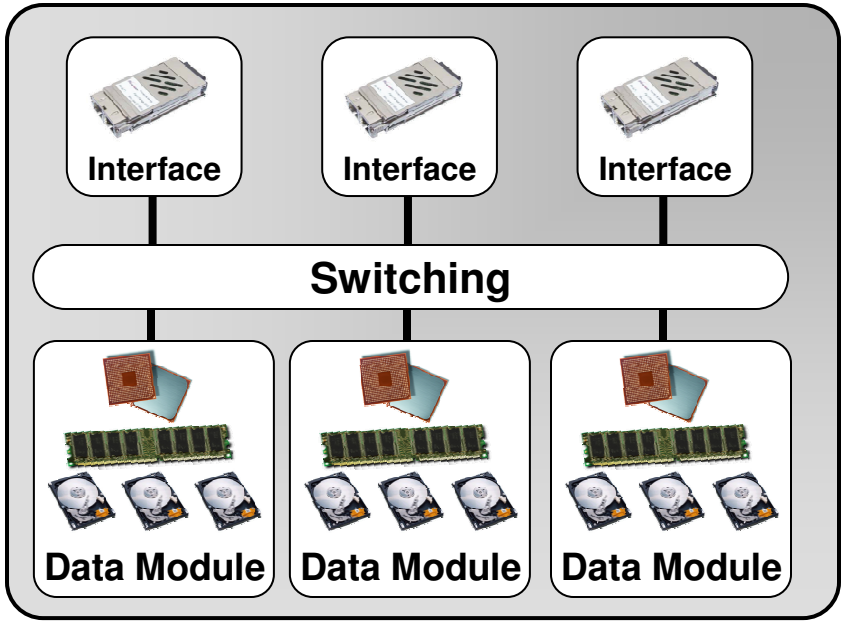
Scale Out



XIV Distribution Algorithm

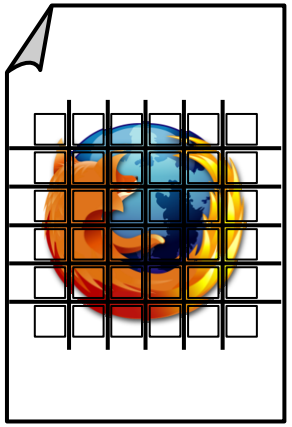
- Each volume is spread across all drives
- Data is “cut” into “partitions” and stored on the disks
- Distribution algorithm **automatically** distributes partitions across **all** disks in the system pseudo-randomly

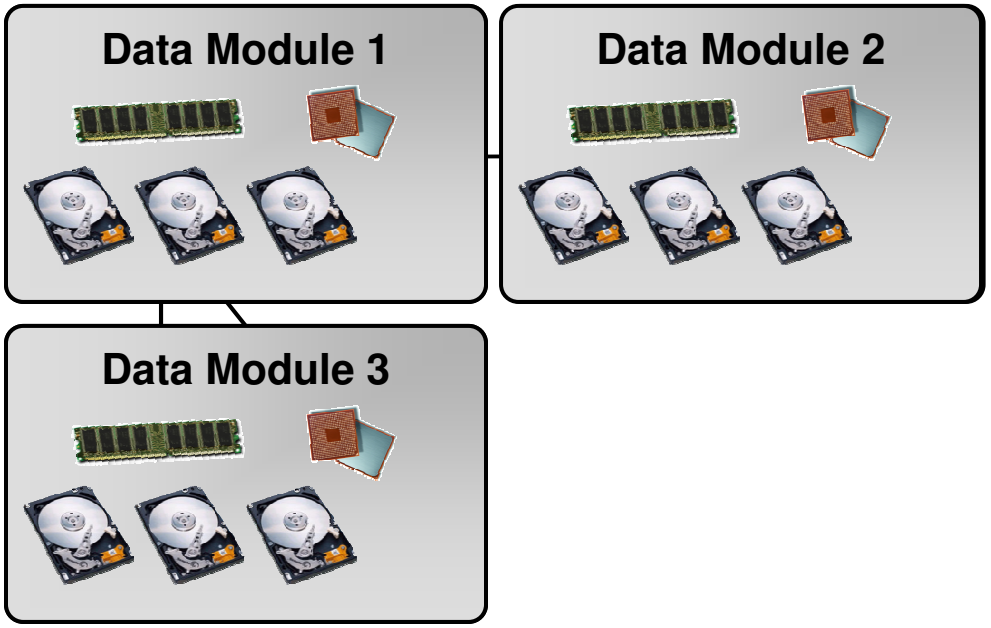




XIV Distribution Algorithm

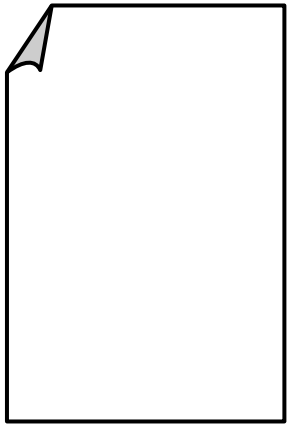
- Data distribution only changes when the system changes
  - Equilibrium is kept when new hardware is added
  - Equilibrium is kept when old hardware is removed
  - Equilibrium is kept after a hardware failure

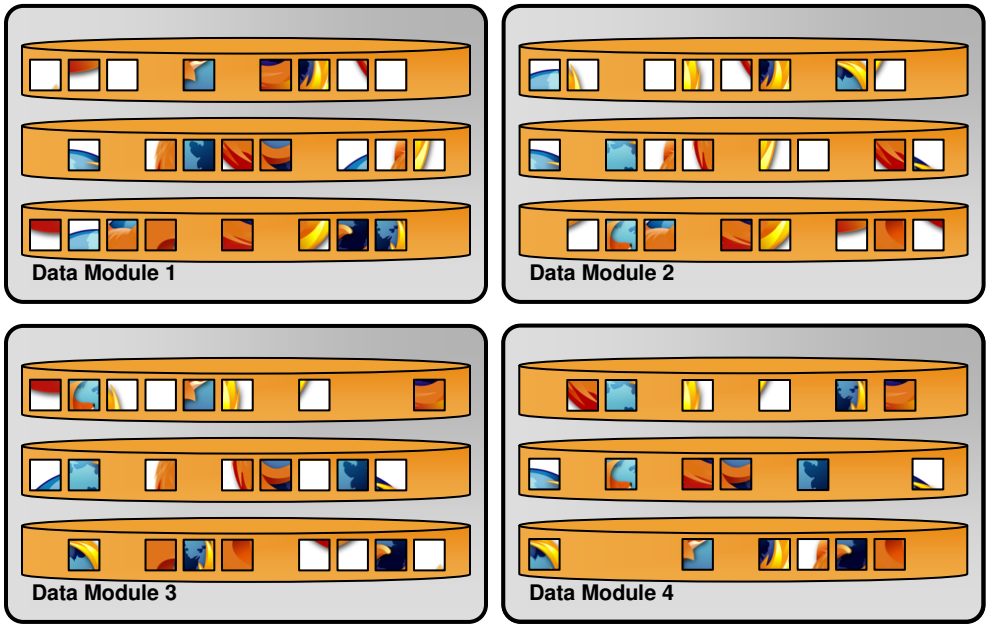




XIV Distribution Algorithm

- Data distribution only changes when the system changes
  - Equilibrium is kept when new hardware is added
  - Equilibrium is kept when old hardware is removed
  - Equilibrium is kept after a hardware failure



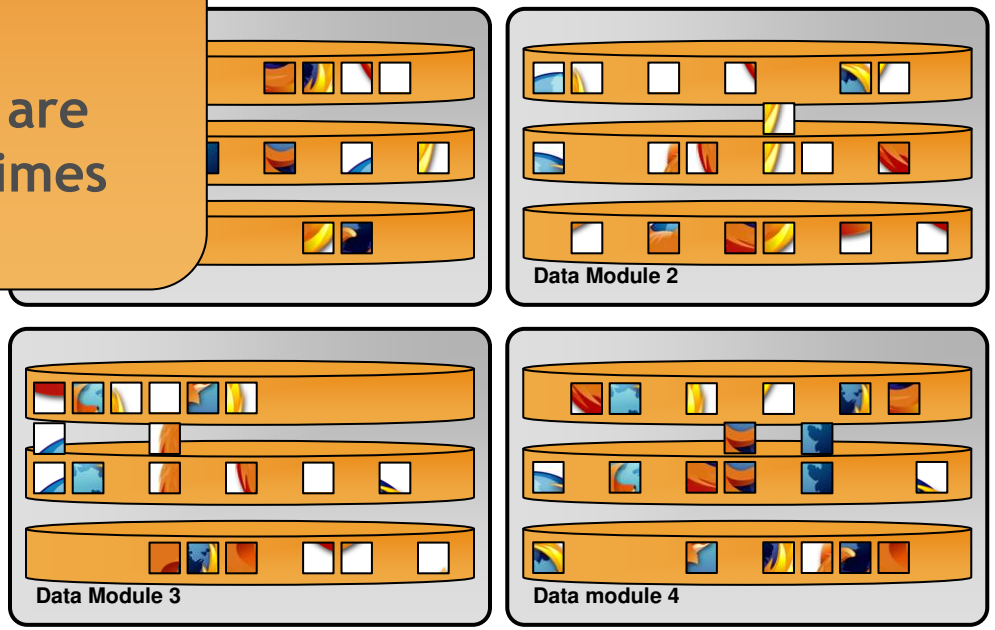


XIV Distribution Algorithm

- Data distribution only changes when the system changes
  - Equilibrium is kept when new hardware is added

The fact that distribution is full and automatic makes sure all spindles join the effort of data re-distribution after configuration change.

Tremendous performance gains are seen in recovery/optimization times thanks to this fact.



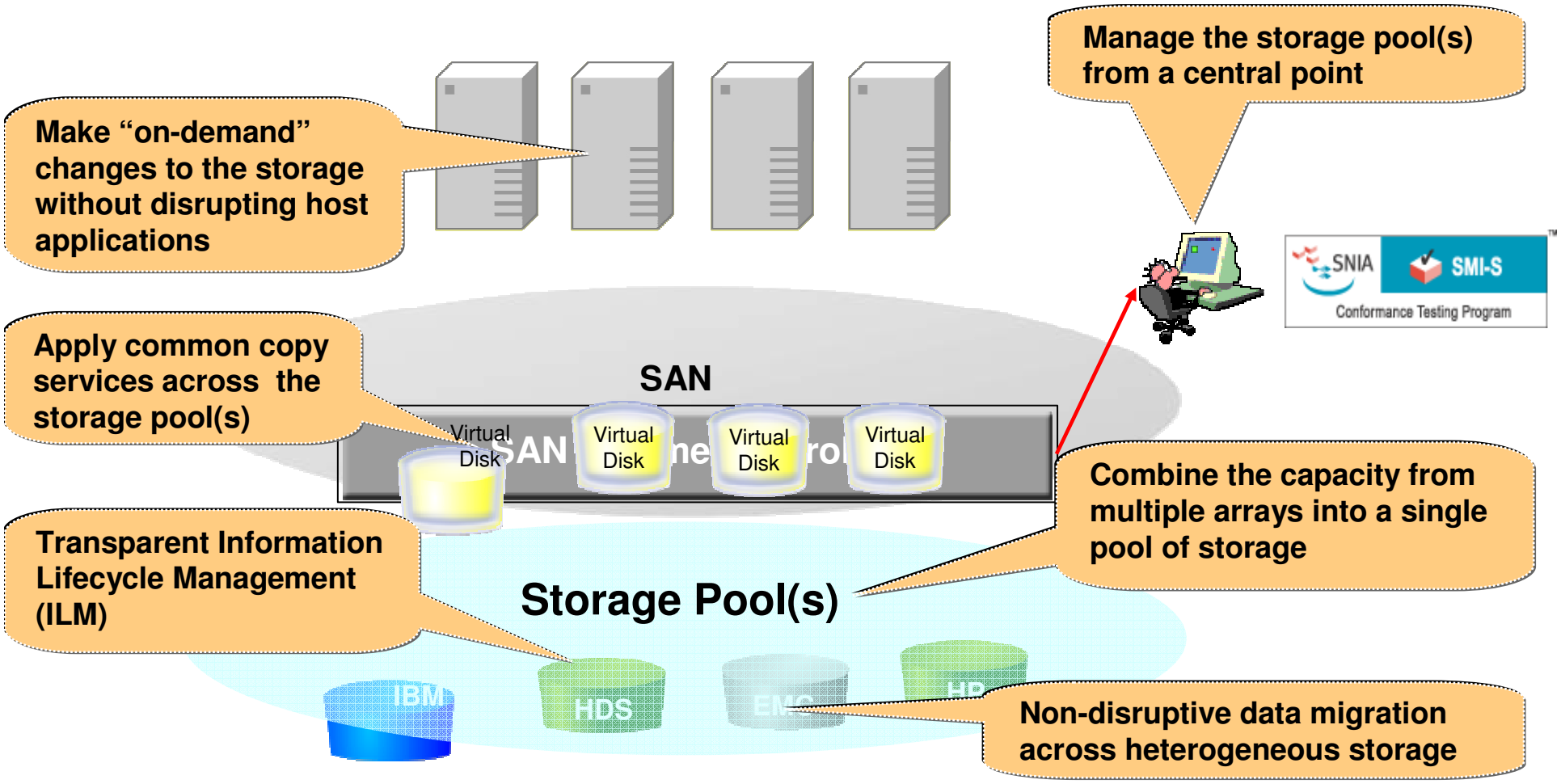
IBM XIV Storage System Proven Benefits

- Lower capital costs, no added charge for XIV software features – mirroring, snapshot, data migration, performance tuning, thin provisioning, …
- Savings in power, cooling, and space with large capacity drives
- Less storage and management needed, thanks to:
  - Automated distribution algorithm – eliminate hot spots – no performance management
  - Built-in thin provisioning – use real capacity when needed
  - Space Management efficiency
  - Efficient and high performance snapshots
  - No storage tiering management (“de-tiering”)
- Simple, intuitive management – more capacity with less staff
- Consistent performance
- High availability:
  - Revolutionary grid-based self-healing

xiv An IBM Company
Storage Reinvented

IBM Solutions in action: IBM SAN Volume Controller

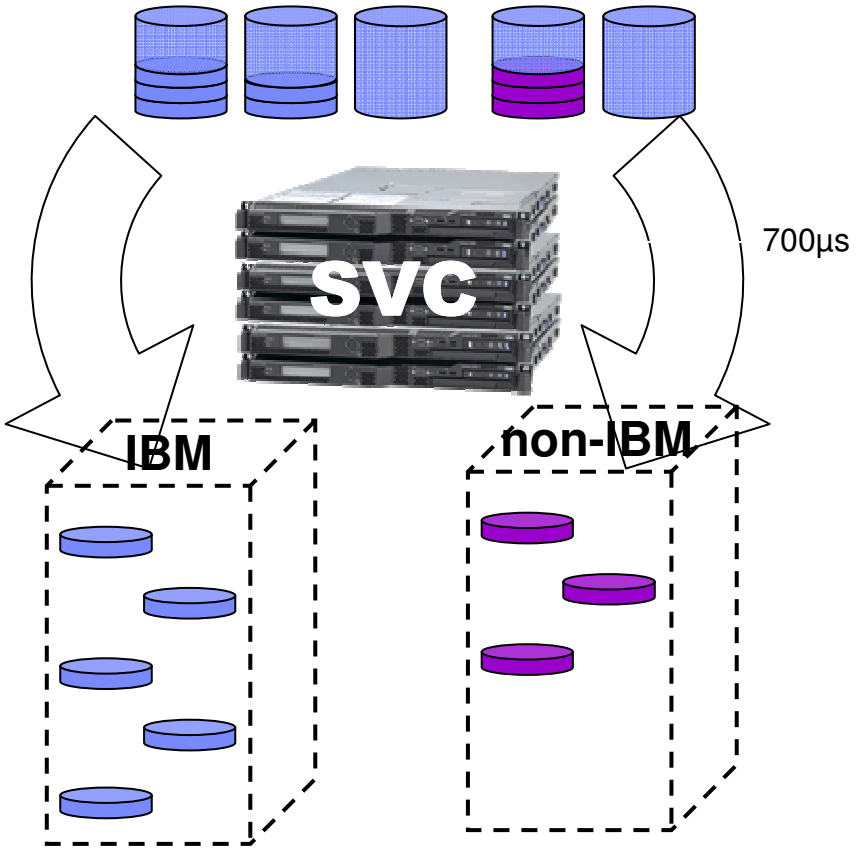
Flexible & Dynamic Storage Infrastructure with IBM SAN Volume Controller



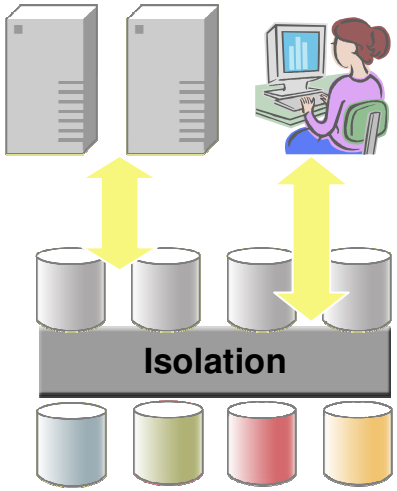
IBM Quicksilver Project

- Technology Demonstration: IBM SAN Volume Controller + Flash memory cards
  - Database workload (0% cache), running for 2 hours, delivered **1 Mio IOPS** at **700µs response time** (peak 1.1M) – August 2008
- SVC adds fine-grained "thin provisioning" to *any* storage

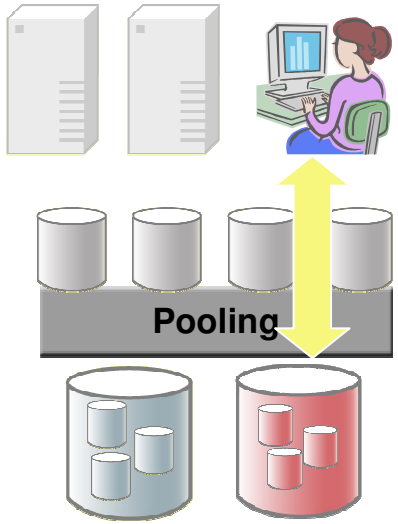
- Maximizes the use of existing storage systems
- Leverage SVC transparent data migration



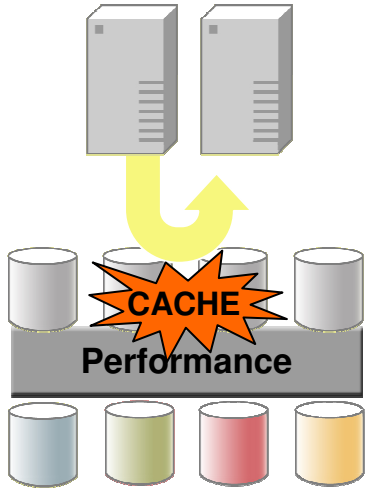
SVC Storage Infrastructure Virtualization : Compiled Benefits



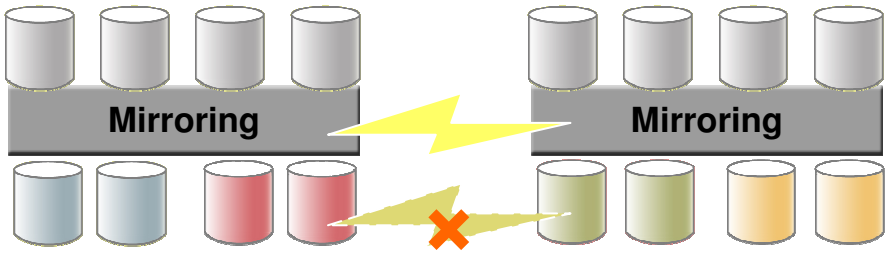
1. Flat interop. matrix
2. Single point administration
3. No-cost multipathing SW



1. Higher (pool) utilization
2. Cross-pool-striping: **IOPS**
3. Storage tier migration



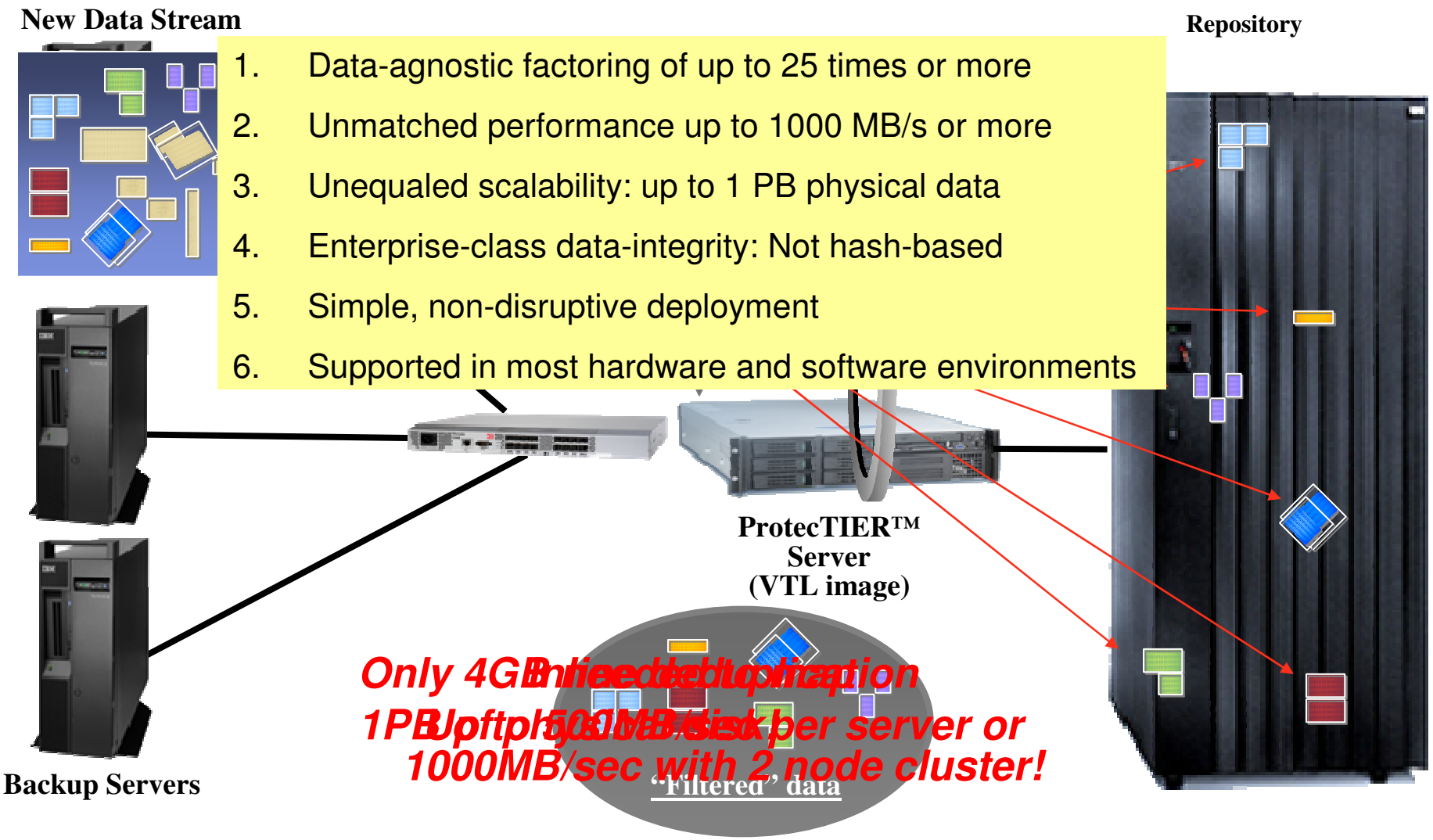
1. Performance increase
2. Hot-spot elimination
3. L1…L2 cache



1. License economies
2. Cross-vendor mirror
3. Favorable TCO

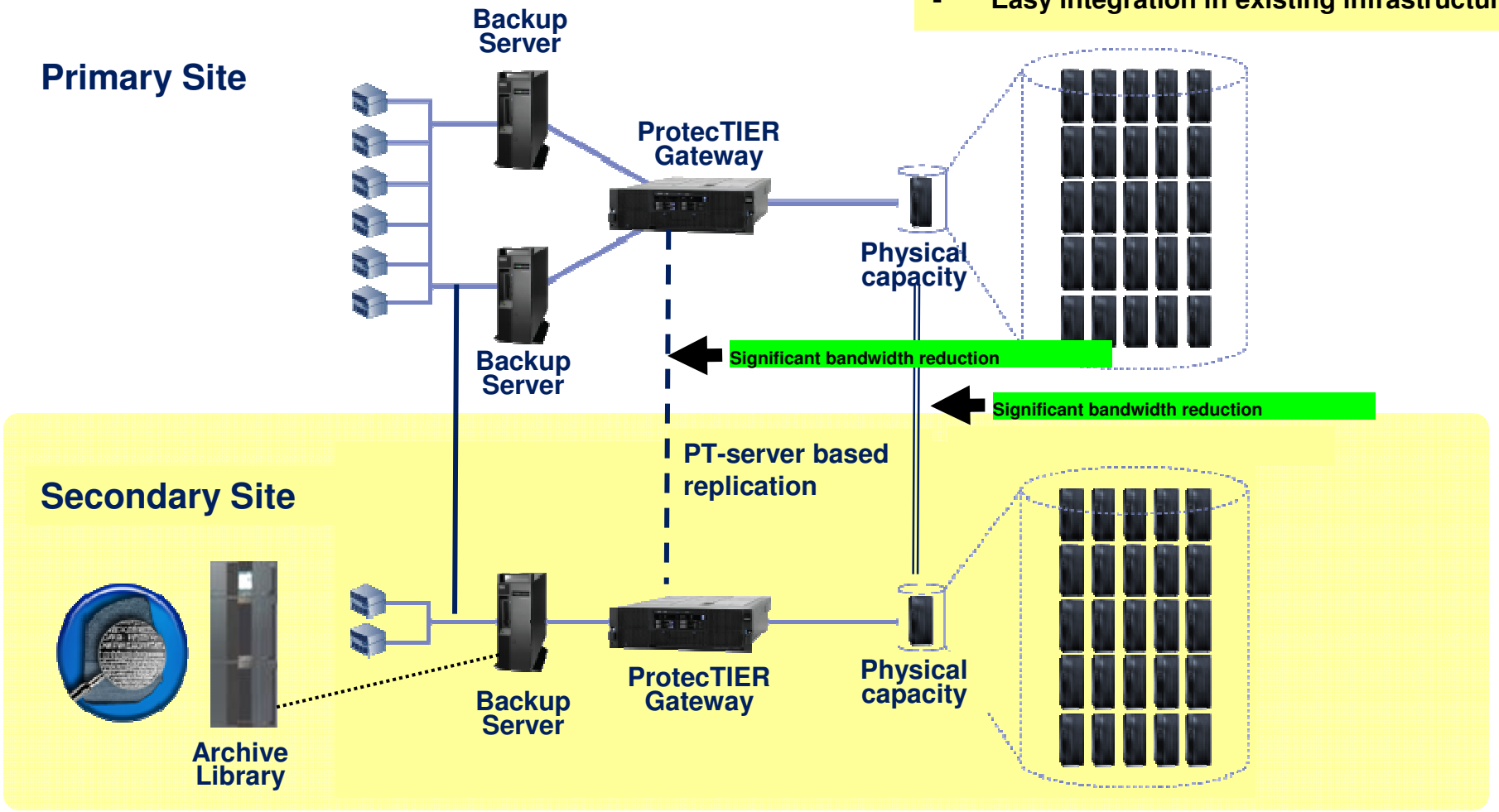
IBM Solutions in action: IBM TS7650 ProtectTier Deduplication

Enterprise Class Data Deduplication: The HyperFactor



Back-up and D/R Replication

- Reduced storage resources
- Faster data recovery
- More frequent checkpoints
- Meet back-up window timeframe
- Highly scalable solution
- 100% data integrity
- Easy integration in existing infrastructure



Deduplication provides significant data reduction
Allows longer retention periods with minimal expense

Actual IBM TS7650 customer examples:

Customer	Physical Disk	Nominal Disk	Data Reduction	Retention Period
Wireless carrier	13 TB	190 TB	16:1	30 days
Oil company	100 TB	900 TB	44:1	6 weeks
Hospital	15 TB	100 TB	15:1	30 days

Consolidation and Virtualization IBM Services Landscape

Drive Productivity in Storage Operations

Get More Out of Existing Infrastructure

Cut Costs

Flexible Services Options
- Out-tasking specific requirements
- Project-based consulting
- Managed Services

Transform & Streamline
- Implement optimal, proven processes throughout the storage management lifecycle
- Alleviate immediate storage problem areas
- Control demand for storage and decrease complexity

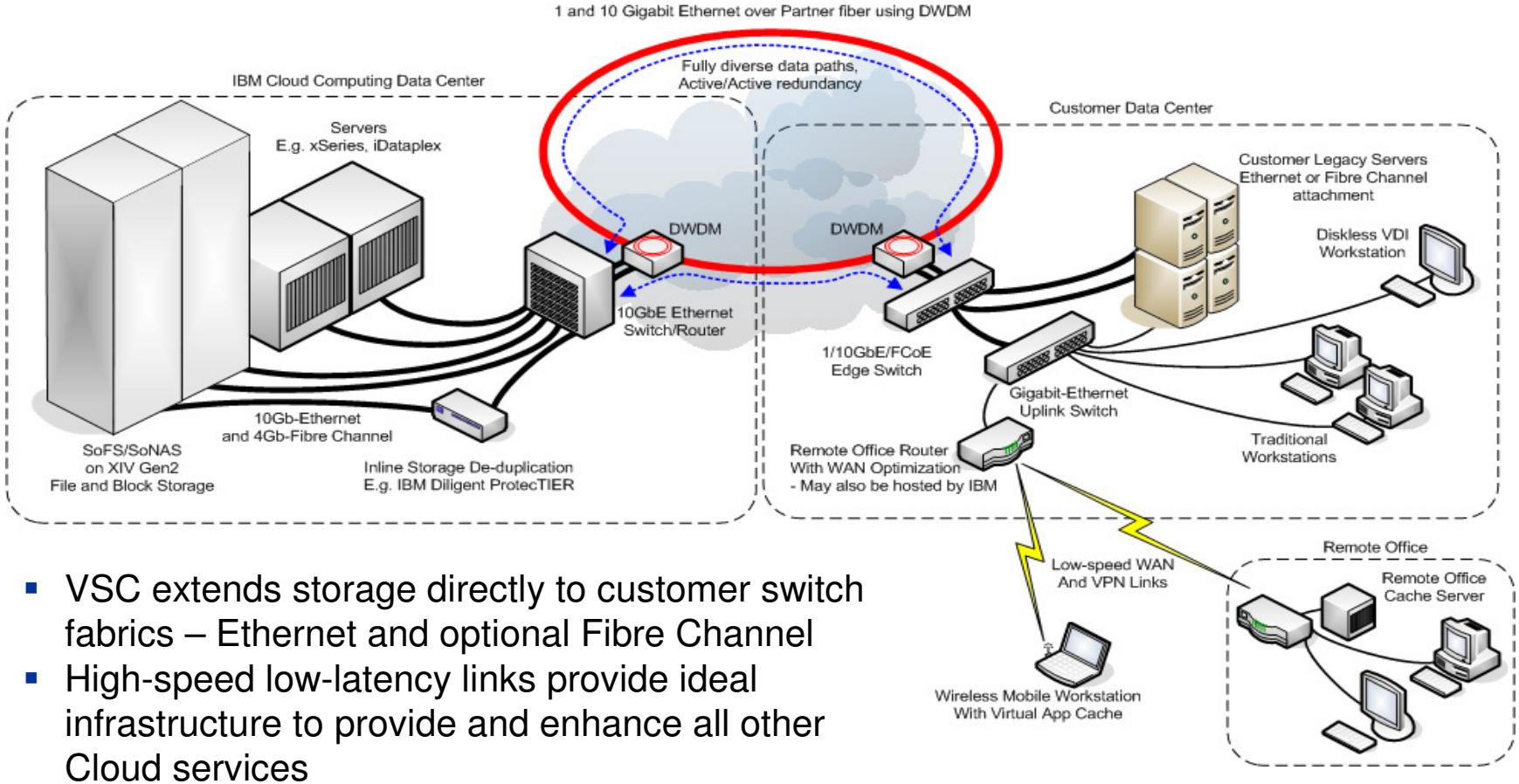
Consolidate & Virtualize
- Consolidate to save energy and floor space
- Virtualize to increase utilization, simplify operations, reduce administrative expense

Migrate your Data
- Non-disruptive data migrations across multi-vendor hardware
- Avoid costly downtime for migration windows
- Move data to more energy-efficient hardware
- Move data in support of consolidation and virtualization

Align Information
- Move low-value data to less expensive storage and free up expensive Tier 1 hardware
- Archive structured and unstructured data for quick payback, improved access, and better performance

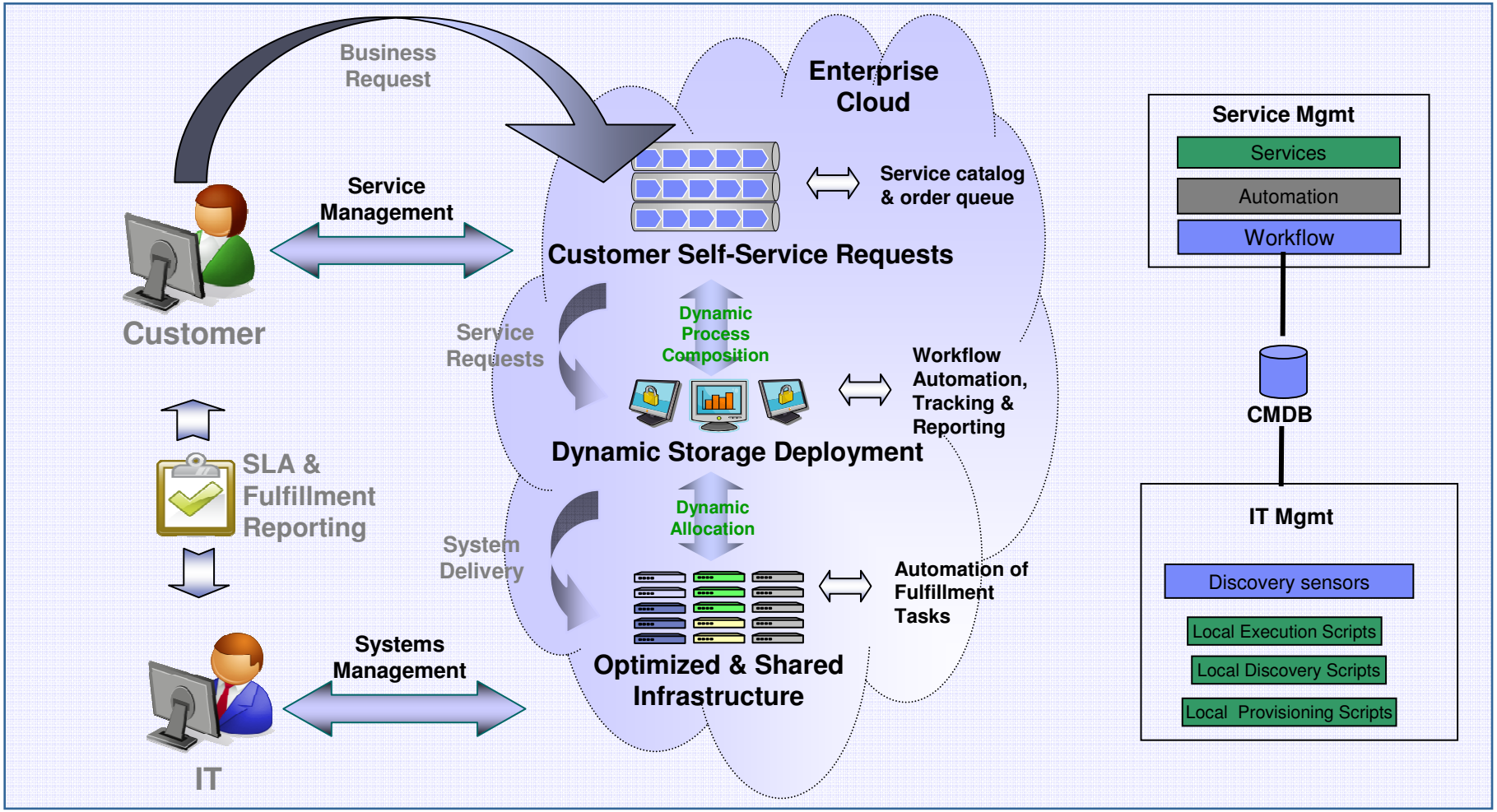
Increase Storage Utilization
- Gain insight into capacity utilization across vendor platforms
- Align usage to line of business – enabling effective chargeback
- Avoid capital expenditure by reclaiming storage

Platform Architecture Overview for Virtual Storage Clouds



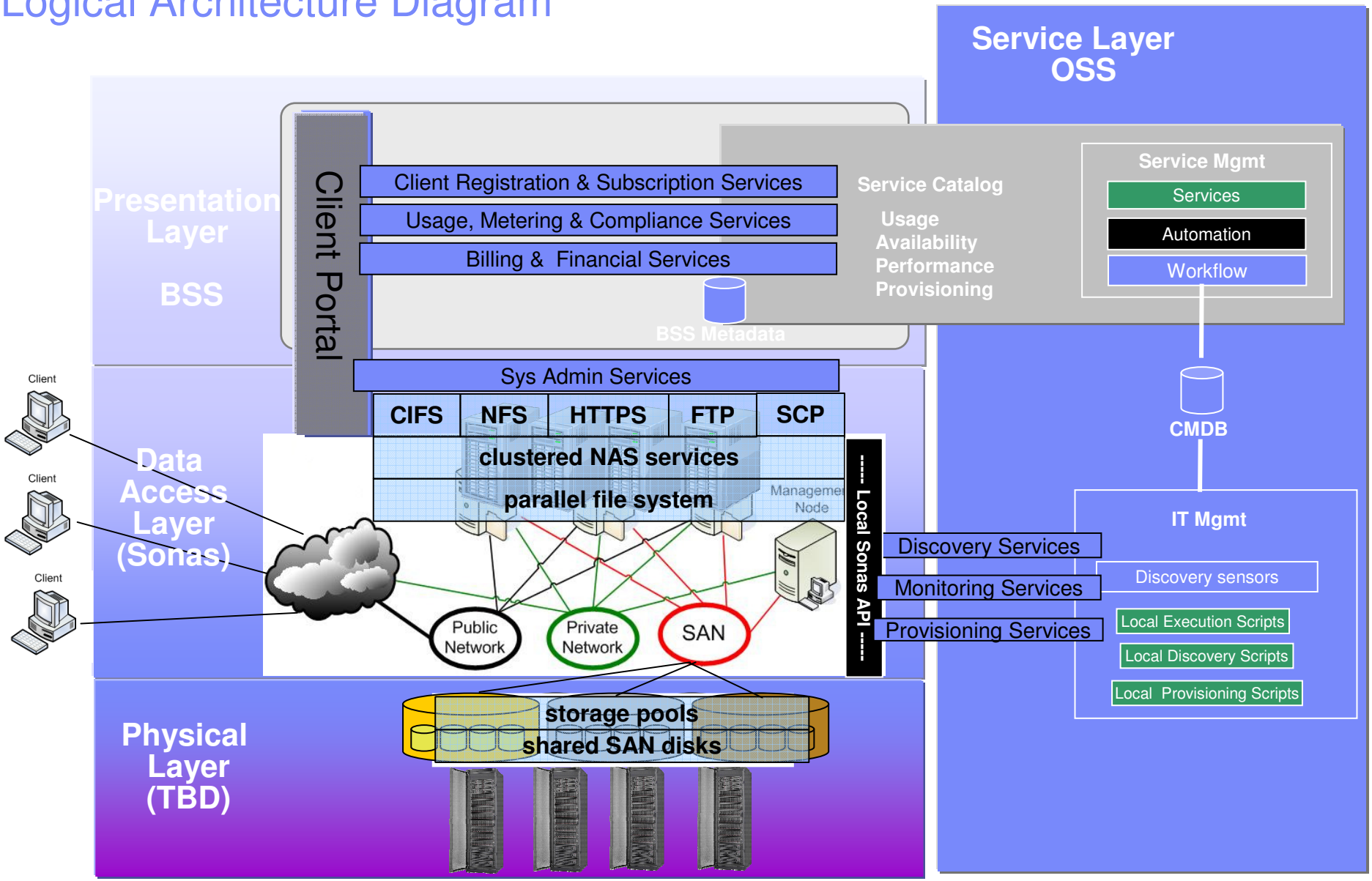
- VSC extends storage directly to customer switch fabrics – Ethernet and optional Fibre Channel
- High-speed low-latency links provide ideal infrastructure to provide and enhance all other Cloud services
- New Cloud services are made possible, including Live VM Migration, Network Boot, et al.

Requirement Overview of Storage Cloud Services

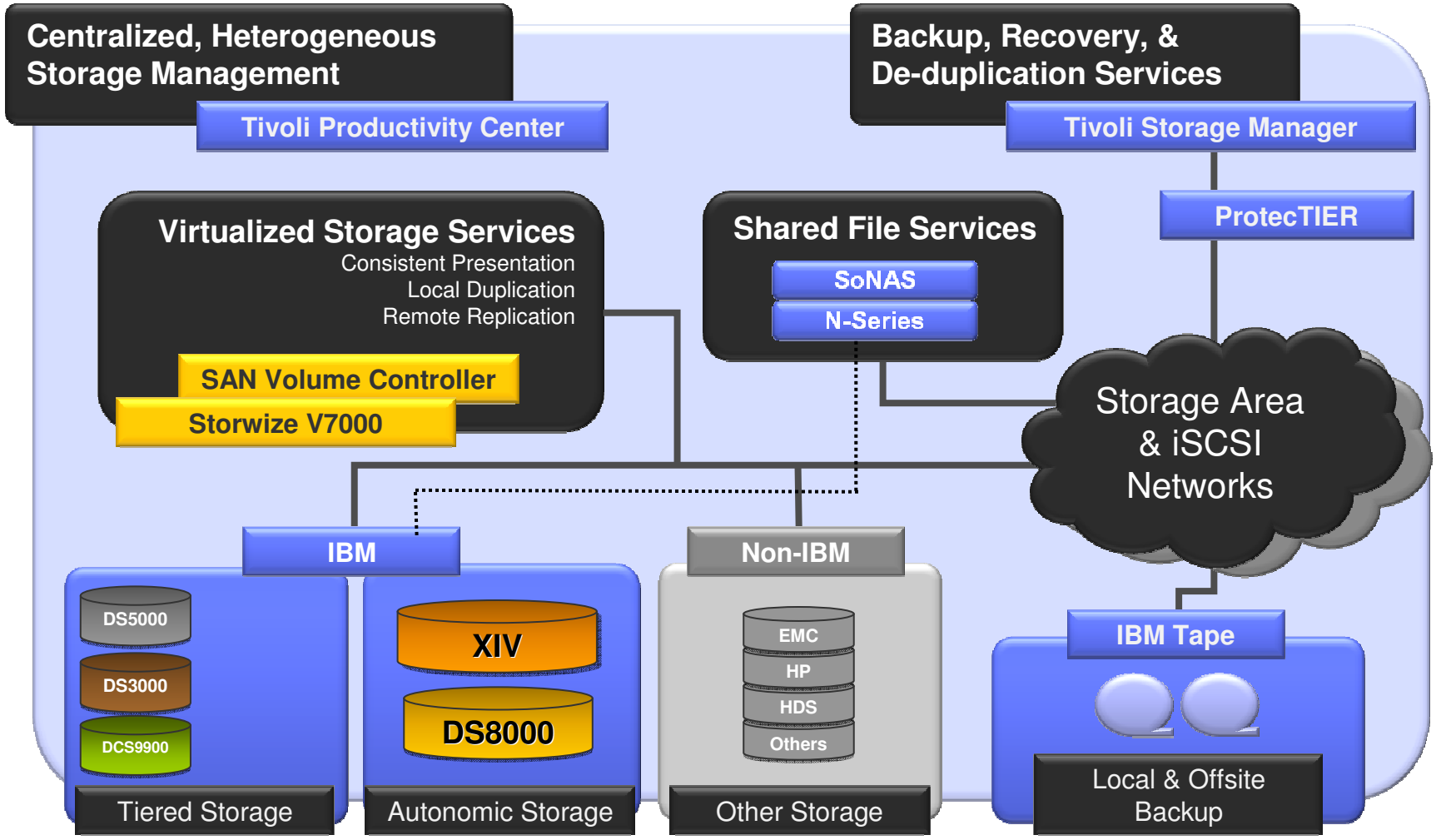


Storage Cloud services dynamically ordered, deployed and managed in response to user demand

Logical Architecture Diagram



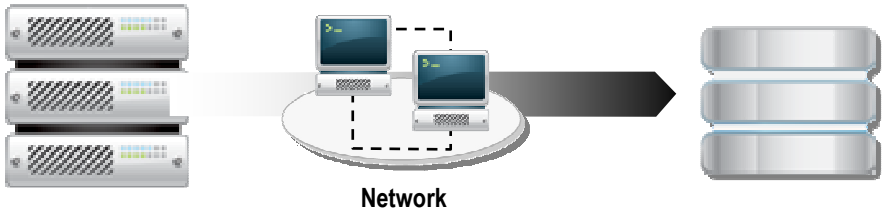
The IBM Storage Solution/Platform Portfolio



Added value Features: ***Manageability***

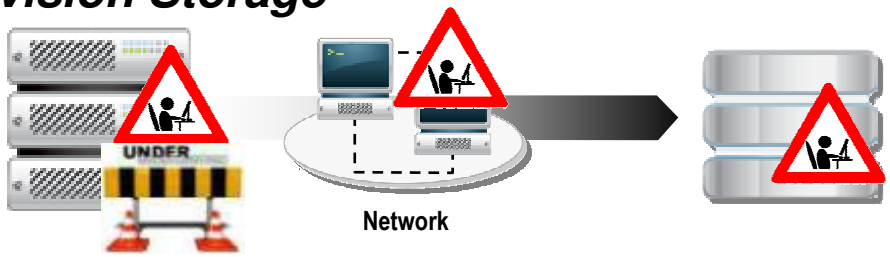
MANAGEABILITY

Visualize the SAN



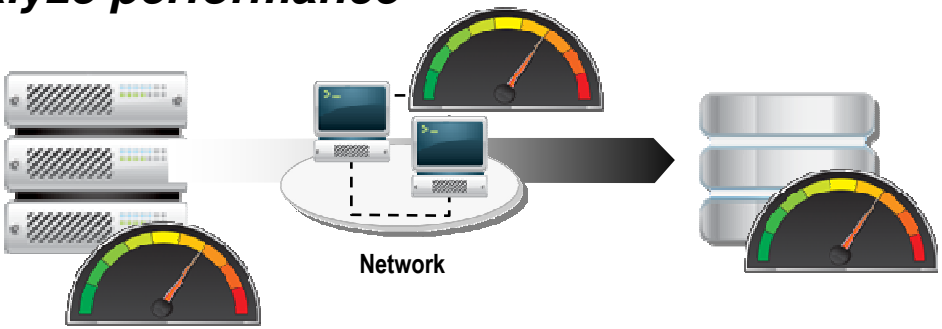
- ✓ From virtual machines to physical servers, the SAN, to disk systems …
- ✓ Physical topology and logical data path.
- ✓ Health/Status Monitoring.
- ✓ Event Management.
- ✓ Device Capacity Management.
- ✓ Policy-based Alerting.

Provision Storage



- ✓ For new or expanding applications, apply application-centric provisioning policies
- ✓ Provision host multi-pathing, SAN zones, and array capacity in a single action

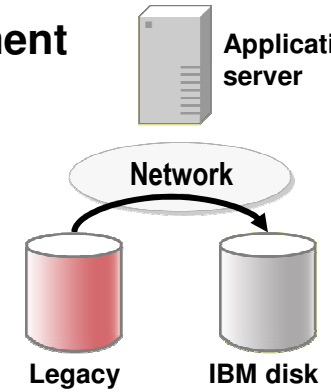
Analyze performance



- ✓ Isolate application I/O performance problems.
- ✓ Report on performance history.

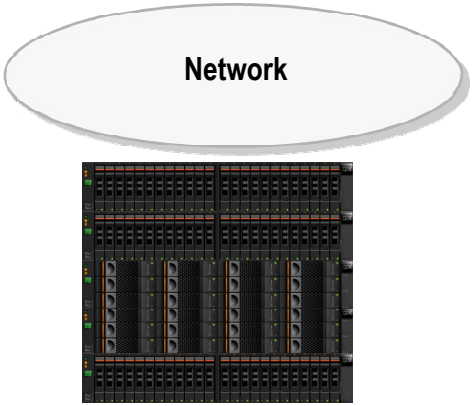
SAN VC and V7000 External Virtualization Features

Transparent data movement



✓**Efficiently manage technology upgrades and lease terminations by transparently moving application data from legacy disk arrays to new IBM Virtualized Disk Solutions**

Legacy disk attach



Hitachi	EMC	NetApp	IBM	Sun	HP	NEC	Bull	Fujitsu	Pillar
USP Lightning, Thunder TagmaStore	CLARiiON, Symmetrix, VMAX	FAS	ESS, FAStT DS3/4/5/6/8 XIV	StorageTek	MA, EMA MSA, EVA XP	iStorage	StoreWay	Eternus	Axiom

For the most current, and more detailed, information please visit **ibm.com**/storage/svc and click on “Interoperability”.

SAN VC and V7000 Host System Attach

VMware
SRM

Microsoft
Windows
MSCS
MPIO, VSS, GDS

Novell
NetWare
Clustering

IBM
AIX
i6 VIOS
HACMP /XD
GPFS / VIO

Sun
Solaris
VCS/SUN
clustering

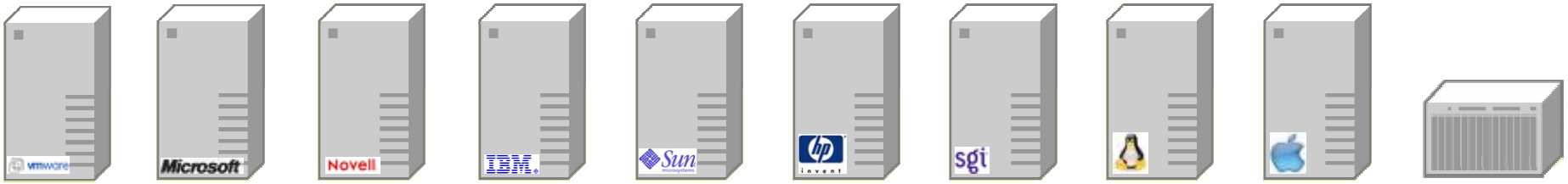
HP
HP-UX
Tru64
OpenVMS
ServiceGuard with SDD

SGI
IRIX

Linux
(Intel/Power/zLinux)
RHEL/SUSE
RHEL 5 ia32, x64
RHEL 3 Power
SLES 9 ia64

Apple
Mac OS X

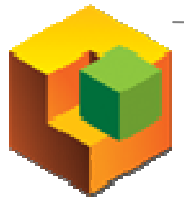
IBM
BladeCenter
Win/Linux/VMWare/AIX
OPM/FCS/IBS



Storwize V7000

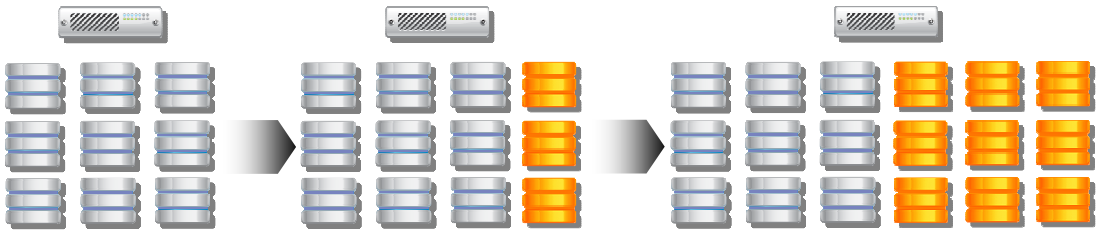
For the most current, and more detailed, information please visit **ibm.com**/storage/svc and click on “Interoperability”.

SAN VC and V7000 Features: ***Scalability***



Dynamically scale …

… capacity



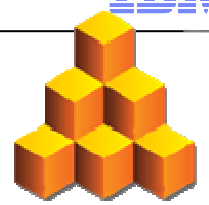
- ✓ For high capacity applications such as archive, dynamically add capacity by adding disk enclosures.
- ✓ …or mix with additional performance…

… features

FlashCopy	Practice DR recovery
SAN Visualization	Automated failover / fail-back
Performance management	Thin Provisioning
Metro Mirror	Performance optimization
Global Mirror	Virtualization

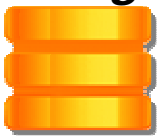
- ✓ Many features are included.
- ✓ Software is preinstalled in the system for easy setup.
- ✓ Premium features are already installed.

SAN VC and V7000 Features: ***Efficiency***



EFFICIENCY

Thin provisioning

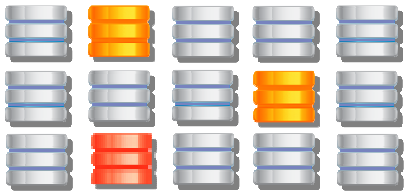


Without thin provisioning, pre-allocated space is reserved whether the application uses it or not.

With thin provisioning, applications can grow dynamically, but only consume space they are actually using.

- ✓ More productive use of available storage.
- ✓ Across all supported host platforms.
- ✓ Improve storage utilization by up to 30%

Performance optimization

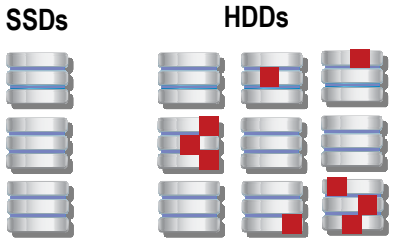


Hot-spots due to poor data layout.

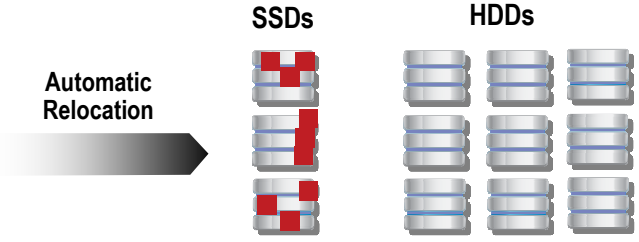
Optimized performance and throughput.

- ✓ Analyze system performance and throughput.
- ✓ Transparently rearrange the data to eliminate hot-spots and balance utilization of all components.

IBM Easy Tier SSD Management



Hot-spots due to poor data layout.



Optimized performance and throughput.

- ✓ Busiest data extents are identified and automatically relocated to highest performing Solid-state Disks
- ✓ Remaining data extents can take advantage of higher capacity, price optimized disks

Protection


Standalone


Cluster


Hot swap


RAID 0


RAID 1


RAID 5


RAID 0+1

HPC intro

What is HPC?

- High performance computing (HPC) is a branch of computer science that concentrates on developing supercomputers and software to run on supercomputers. A main area of this discipline is developing parallel processing algorithms and software: programs that can be divided into little pieces so that each piece can be executed simultaneously by separate processors.
- High performance computing (HPC) refers to the use of supercomputers and computer clusters, that is, computing systems comprised of multiple processors linked together in a single system with commercially available interconnects.
- Because of their flexibility, power, and relatively low cost, HPC systems increasingly dominate the world of supercomputing.

Where HPC?

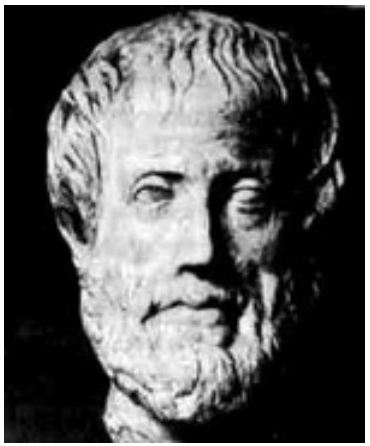
- HPC is most commonly associated with computing used for scientific research. A related term, High performance technical computing (HPTC), generally refers to the engineering applications of cluster-based computing (such as computational fluid dynamics and the building and testing of virtual prototypes).
- Recently, HPC has come to be applied to business uses of cluster-based supercomputers, such as data warehouses, line-of-business (LOB) application and transaction processing.

A Cluster Is Described As:

"A cluster is a collection of interconnected computers used as a unified computing resource"

(*Gregory Pfister - In Search of Clusters*)

- Clusters are comprised of standard components that could be used separately in other types of computing configurations
  - Compute nodes (servers/Blades)
  - Networking adapters and switches
  - Local and/or external storage
  - Systems management software



Aristotle

"...the true object of architecture is not bricks, mortar, or timber, but the house; and so the principal object of natural philosophy is not the material elements, but their composition, and the totality of the substance, independently of which they have no existence…"
-- Aristotle

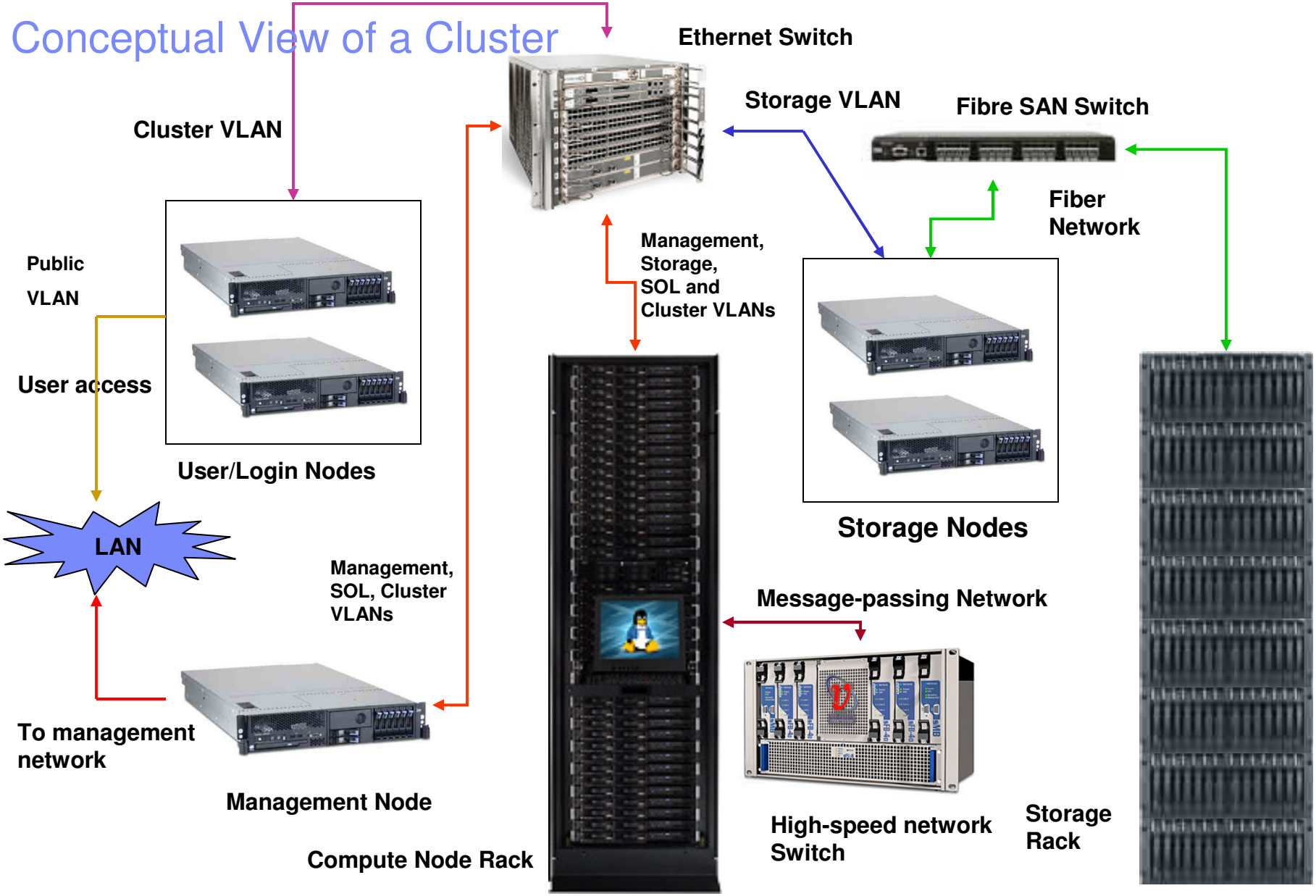
The whole is greater than the sum of the parts

HPC Clusters – Applications

HPC clusters typically run:

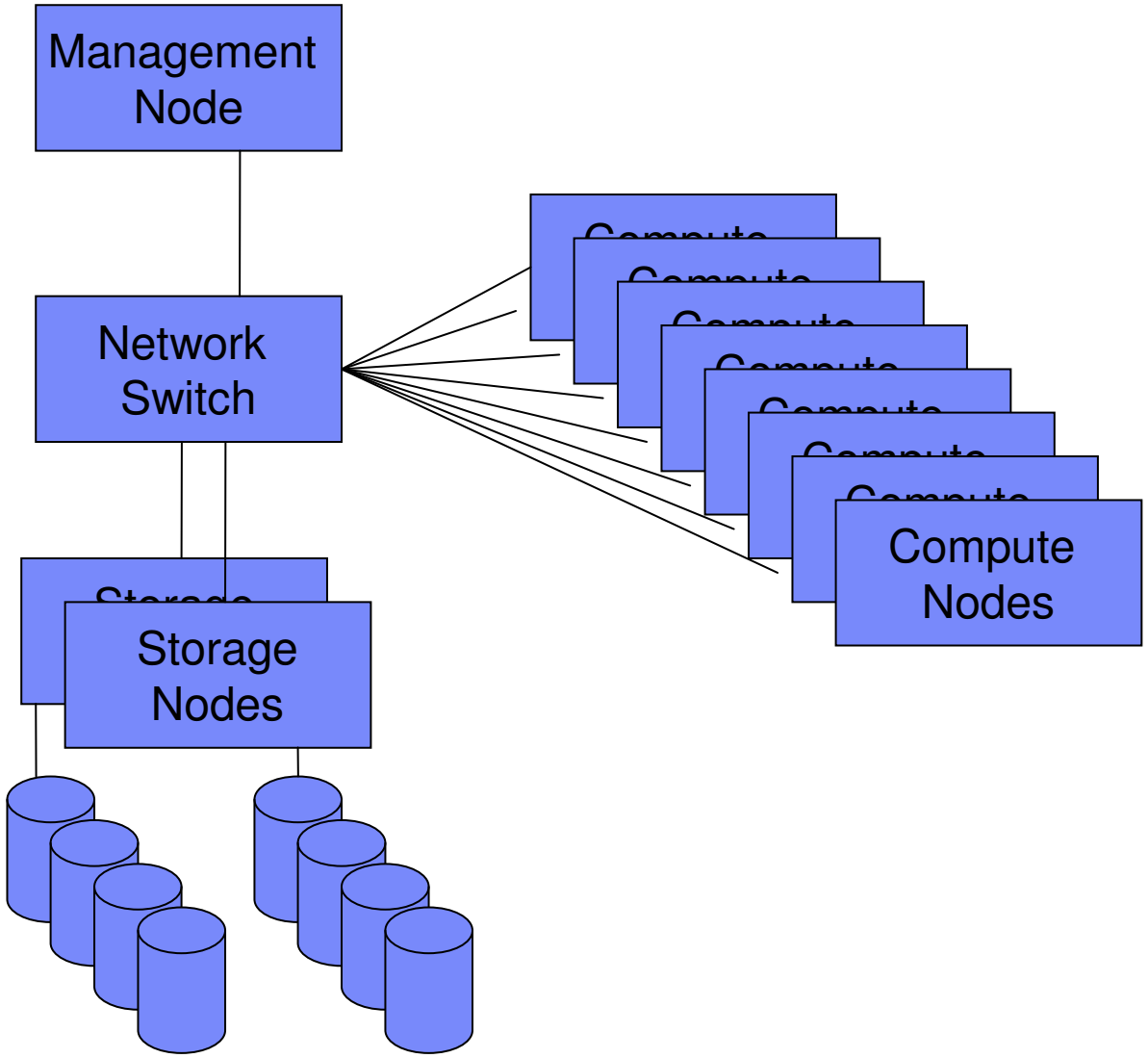
- Parallel applications with one to many tasks on one to many nodes computing in parallel and exchanging messages during their run time (e.g. big CFD simulation of a Formula 1 racing car).
- Lots of single jobs with different input data on lots of individual nodes (or their CPU cores) without communication during their run time (e.g. matching of gene or protein sequences vs. a common database).
- Mix of both (multiple/different parallel apps + single jobs).

Conceptual View of a Cluster

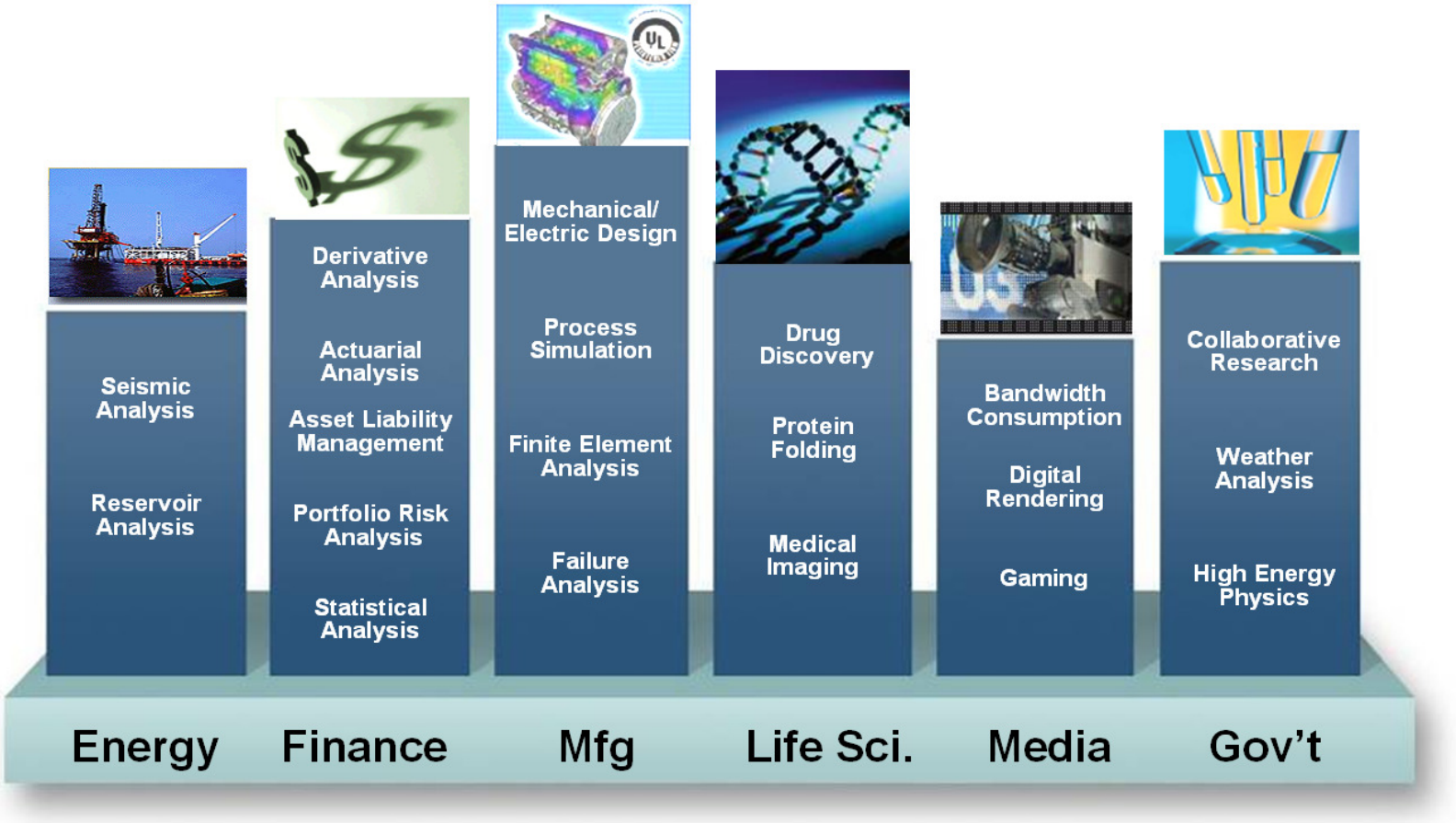


Very Basic Cluster Design

A Cluster that is designed like this is ready to perform a variety of compute functions.

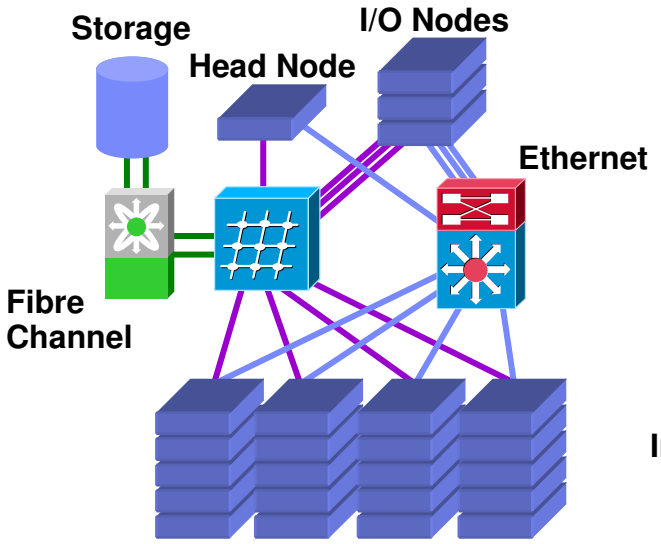


HPC marketplace



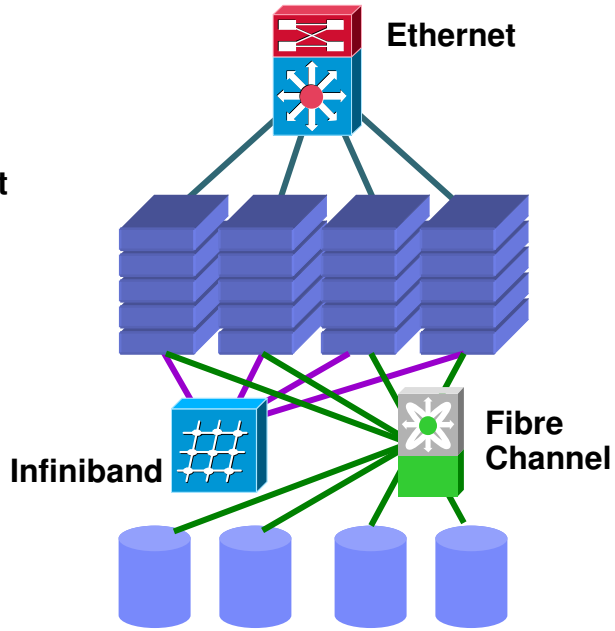
High Performance Clusters need Networks

HIgh-Performance Computing



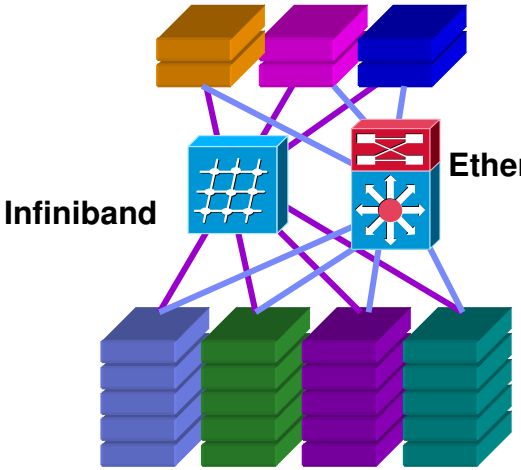
Fluent, Ansys, Charmm, Amber, LS-Dyna

Clustered Database and Storage



Oracle, IBM DB2
Lustre, IBM GPFS, Panasas

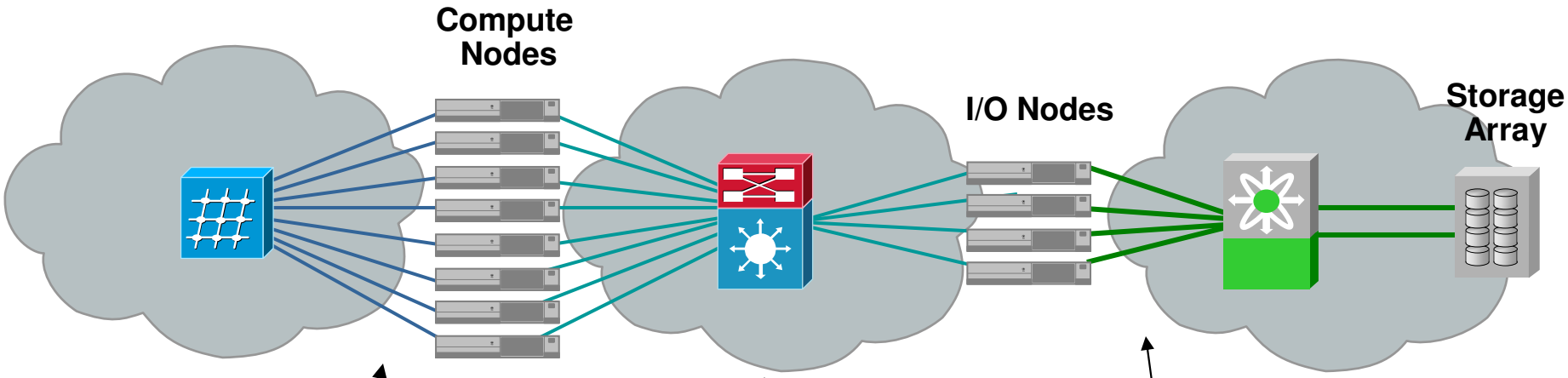
Distributed Systems



TIBco, Wombat, 29West

The right network technology is key to efficiency and ROI

The Role of the Network in HPC clusters



Inter-Process Communication (IPC)
Network for Message Passing (usually MPI)
- latency and bandwidth sensitive

Management and I/O
- bandwidth sensitive for large I/O

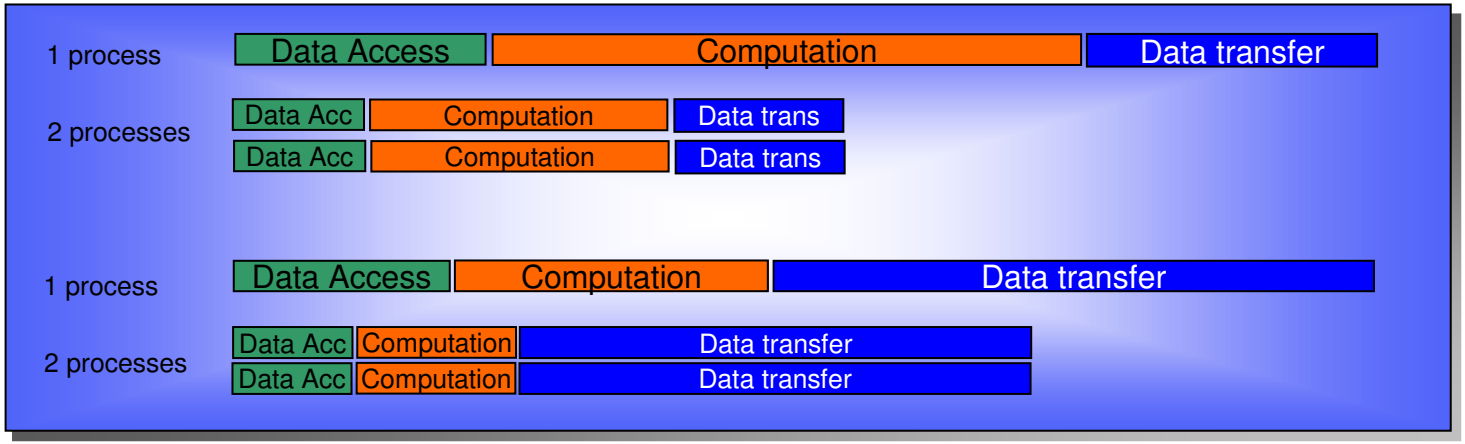
Storage
- scalable and fast access to block storage

Infiniband/Myrinet **Gigabit Ethernet** **Fibre Channel**

Interconnect – Why is it important for a Cluster?

Application performance

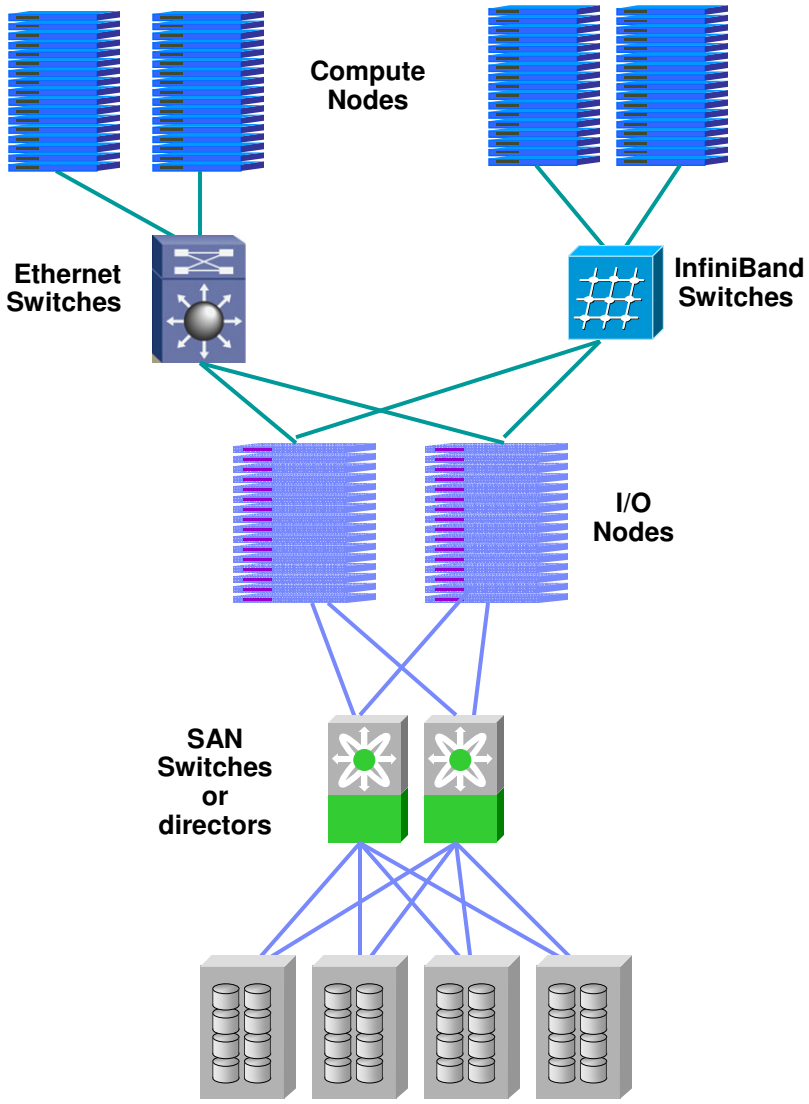
- Compute time (CPU + Memory access)
- Communication (between processes)
- Data access



Interconnect

HPC Storage

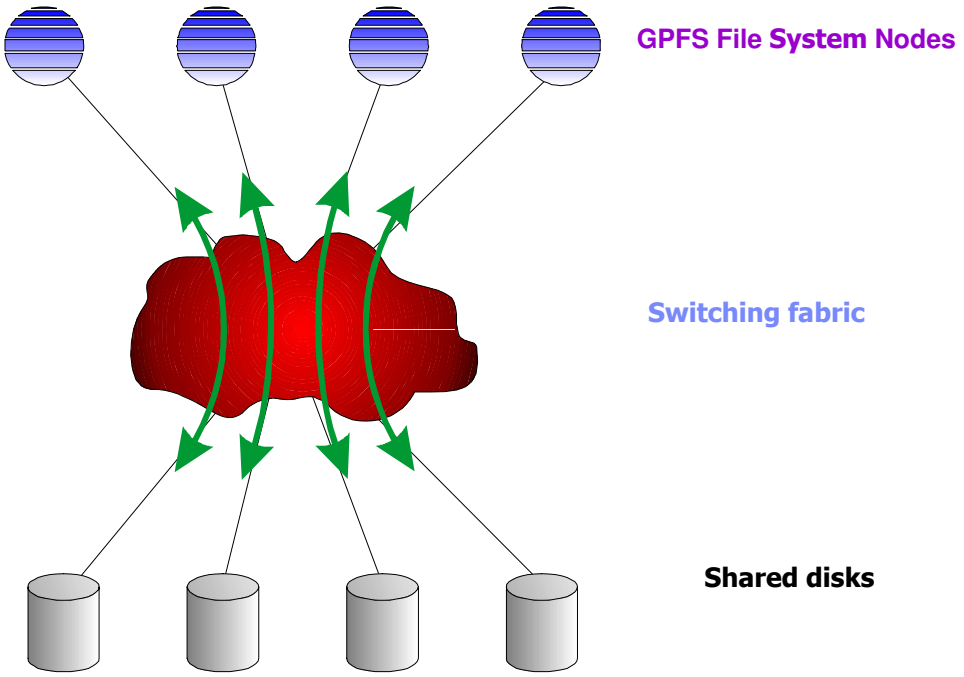
- HPC has large storage requirements
  - Often in Peta-Bytes of Data
  - Retrieving and saving data sets is problematic
  - Data integrity and coherency is a key concern
- NAS typically used for HPC storage
- High-performance parallel file systems increase I/O performance
  - iBrix, Panasas, GPFS, PVFS, Lustre, etc
- InfiniBand-attached storage provides high-bandwidth access to storage using native InfiniBand RDMA protocols
  - Ideal Performance characteristics for High-performance Database clustering
- Key Products:
  - IB attached storage and FibreChannel Gateway
  - FibreChannel Switching
  - Catalyst 6500, etc – NAS



GPFS and storage

GPFS File System

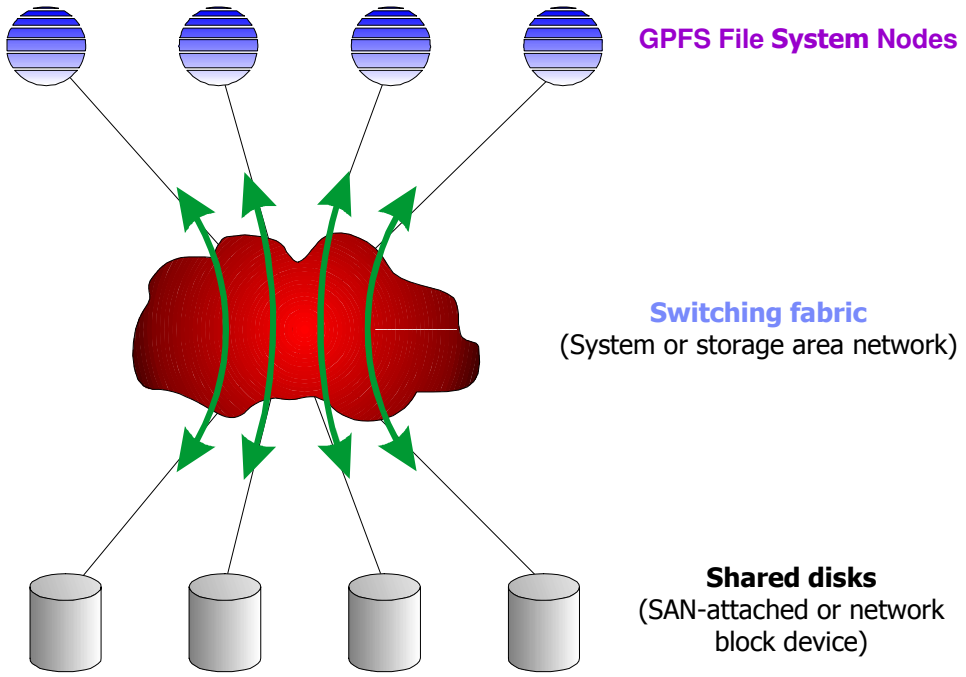
- Available continuously since 1998.
- Product available on pSeries and xSeries (IA32, IA64, Opteron), on AIX and Linux, and on Blue Gene.
- Also runs on compatible non-IBM servers and storage.
- Thousands of installs, including many Top 500 supercomputers
- Customers use GPFS in many applications
  - High-performance computing
  - Scalable file and Web servers
  - Database and digital libraries
  - Digital media
  - Analytics, financial data management, engineering design, …



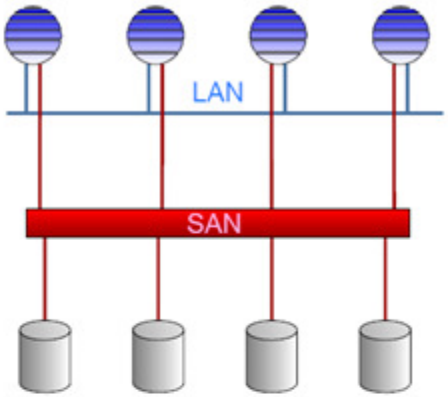
GPFS: parallel file access

Parallel Cluster File System Based on Shared Disk (SAN) Model

- *Cluster –* fabric-interconnected nodes (IP, SAN, …)
- *Shared disk -* all data and metadata on fabric-attached disk
- *Parallel -* data and metadata flows from all of the nodes to all of the disks in parallel under control of distributed lock manager.

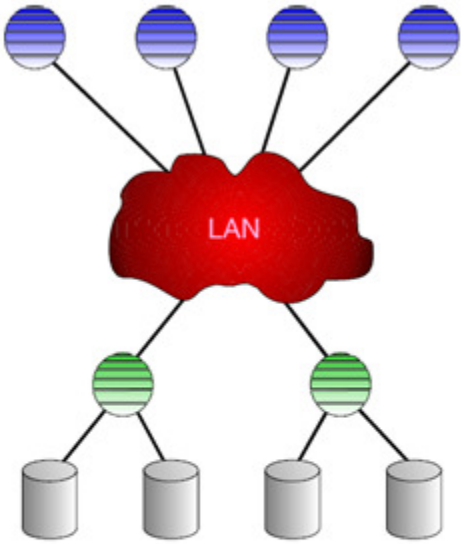


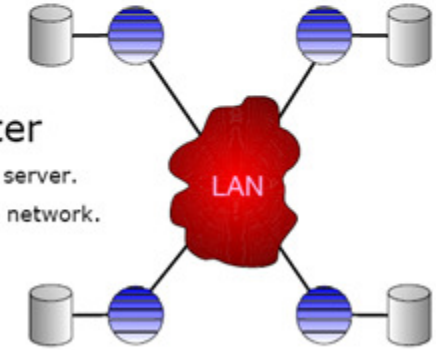
GPFS configuration examples



Storage Area Network

All disk data traverses the SAN.





Symmetric Cluster

All nodes are both client and server.

Disk data traverses the node network.

Cluster with Dedicated Storage Nodes

Disk data traverses the node network.

These are some of the basic possible configurations... there are many others!

GPFS performance features

- Striping
- Large blocks (with support of sub-blocks)
- Byte range locking (rather than file or extent locking)
- Access pattern optimizations
- File caching (i.e. pagepool) that optimize streaming access
- Prefetch, write behind
- Multi-therading
- Distributed/parallel overhead functions (e.g. metadata, tokens)
- Multi-pathing (i.e. multiple, independent paths to the same file data from anywhere in the cluster)

Network Storage Device Organzation

NSD

- SW layer in GPFS providing a "virtual" view of a disk
- virtual disks which correspond to LUNs in the NSD servers with a bijective mapping

LUN

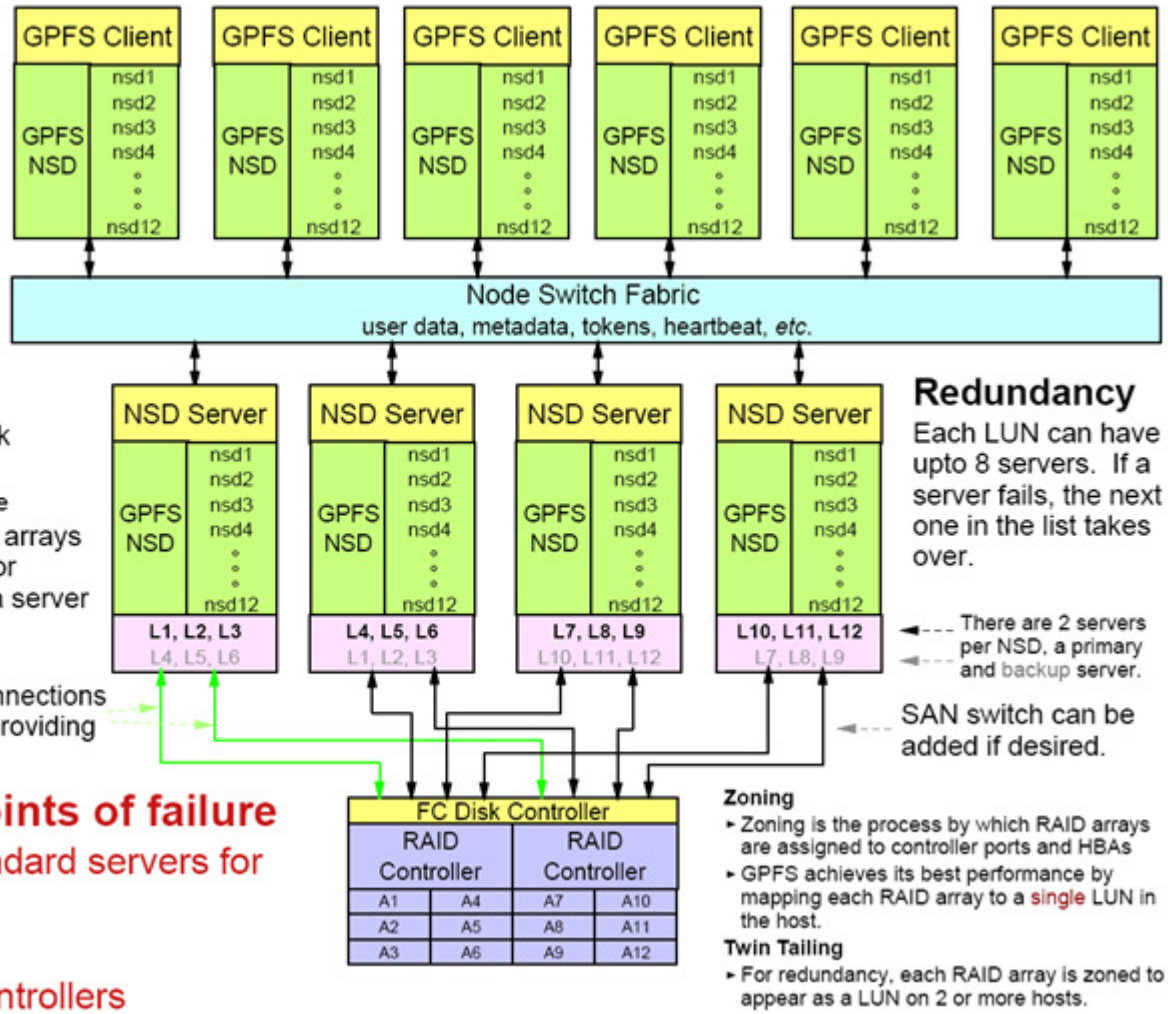
- Logical Unit
- Abstraction of a disk
  - AIX - hdisk
  - Linux - SCSI device
- LUNs map to RAID arrays in a disk controller or "physical disks" in a server

Redundancy

Each server has 2 connections to the disk controller providing redundancy

No single points of failure

- primary/secondard servers for each LUN
- FC fail over
- Dual RAID controllers



Redundancy

Each LUN can have upto 8 servers. If a server fails, the next one in the list takes over.

There are 2 servers per NSD, a primary and backup server.

SAN switch can be added if desired.

Zoning

- Zoning is the process by which RAID arrays are assigned to controller ports and HBAs
- GPFS achieves its best performance by mapping each RAID array to a single LUN in the host.

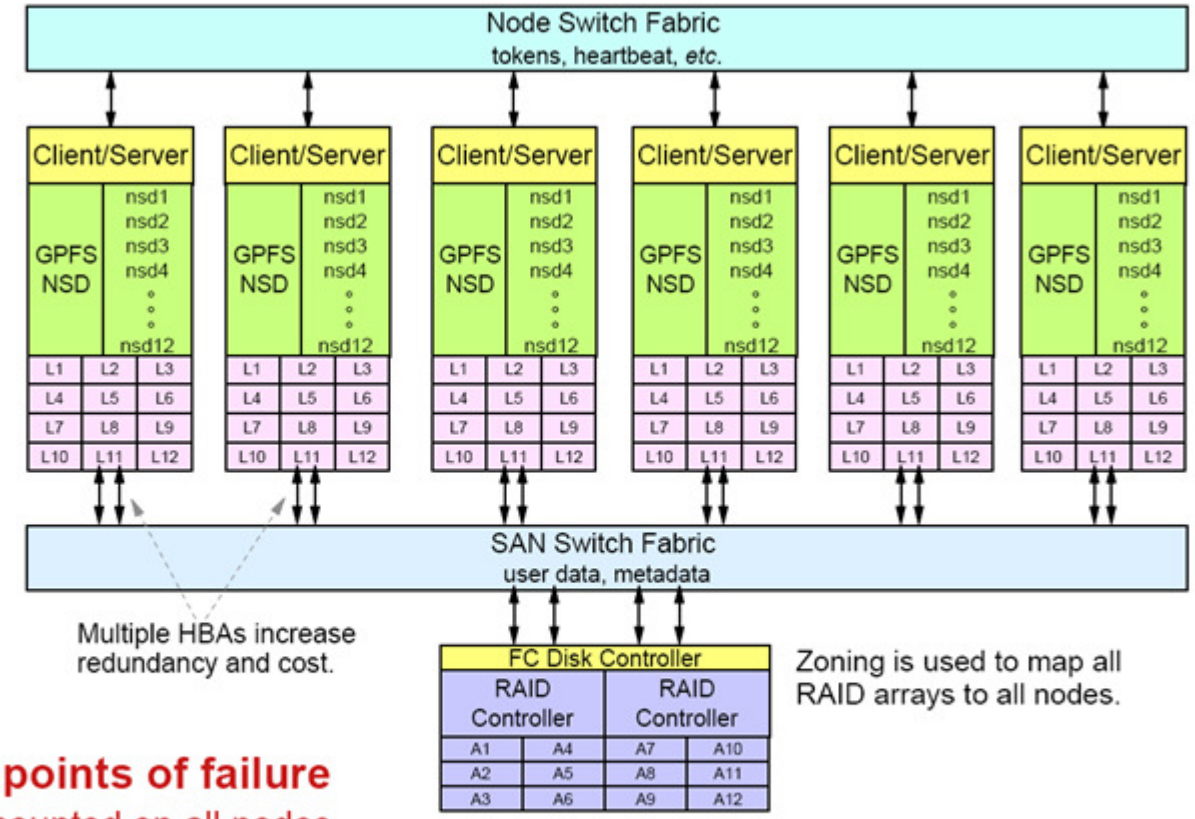
Twin Tailing

- For redundancy, each RAID array is zoned to appear as a LUN on 2 or more hosts.

Storage Area Network organization

All nodes act both as client and server.

All LUNs are mounted on all nodes.

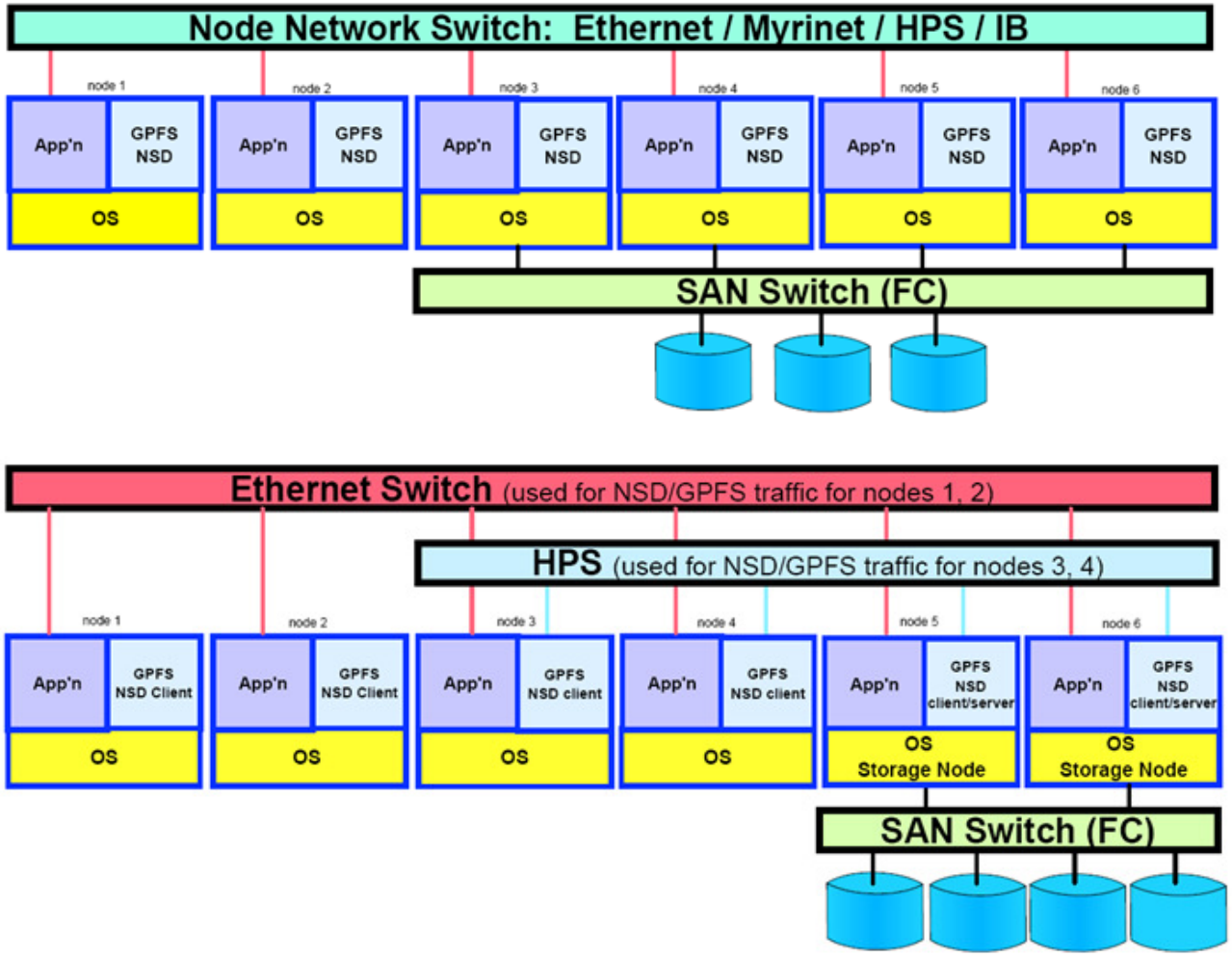


No single points of failure

- All LUNs mounted on all nodes
- FC fail over
- Dual RAID controllers

Due to the cost of SAN/FC networks, these solutions do not scale out very large.

Different models



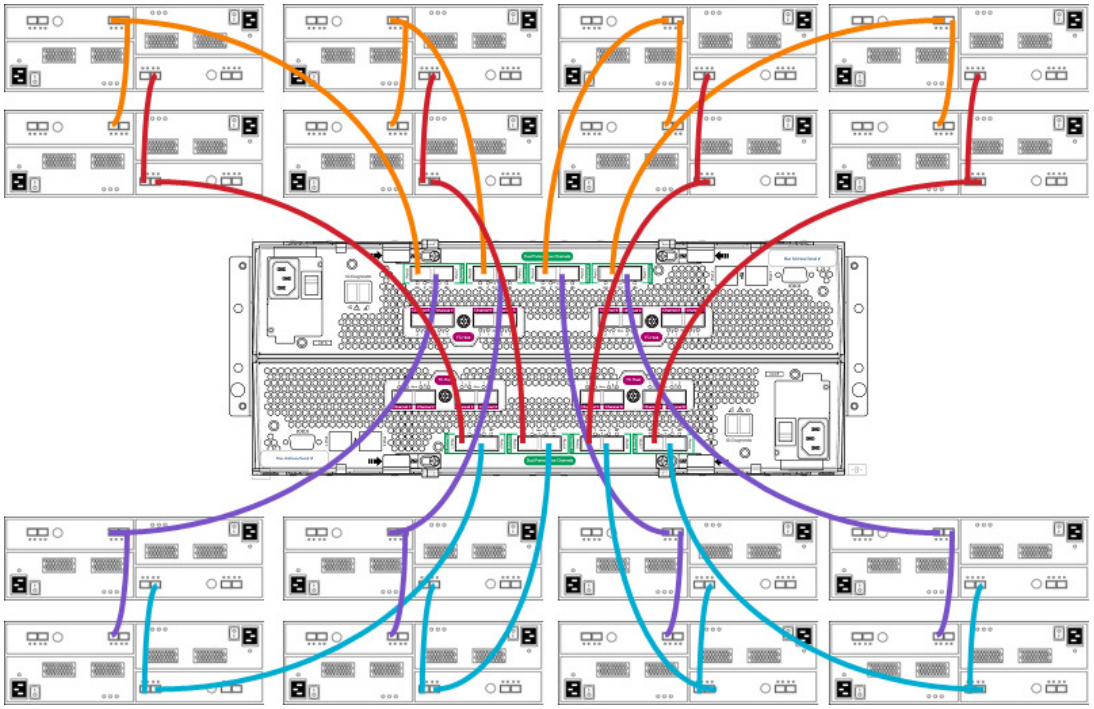
IBM disk solutions for HPC

- DS5000
- DCS9900
- future

DS5000

DS5000 Back-end Designed For Low Latency

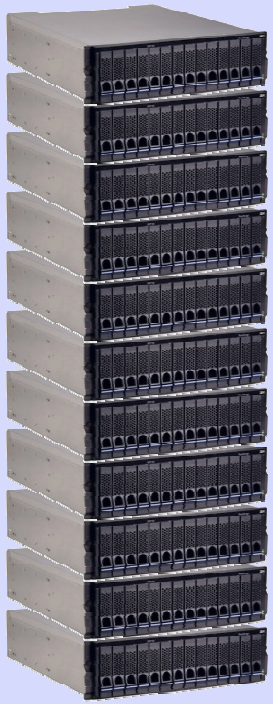
- Short, quick drive loops reduce latency and create more responsive applications
- Sixteen switched back-end drive ports
- 256-drive configuration has only 32 drives per dual-loop
- 448-drive configuration has a max of 64 drives per dual-loop



EXP5060 Expansion Unit

- Support 1 TB and 2 TB 3.5” SATA II Drives
- 5 horizontal drawers with 12 drives each
  - Eliminates excessive front heavy weight in the rack
- All drives remain online when drawer is extended for service
  - Allows for single drive replacement without affecting all other drives in the enclosure
- 4U and fits in a standard 19” rack
- Supports two cabling options
  - Traditional
  - Trunking
- Can be intermixed with EXP5000 expansion units

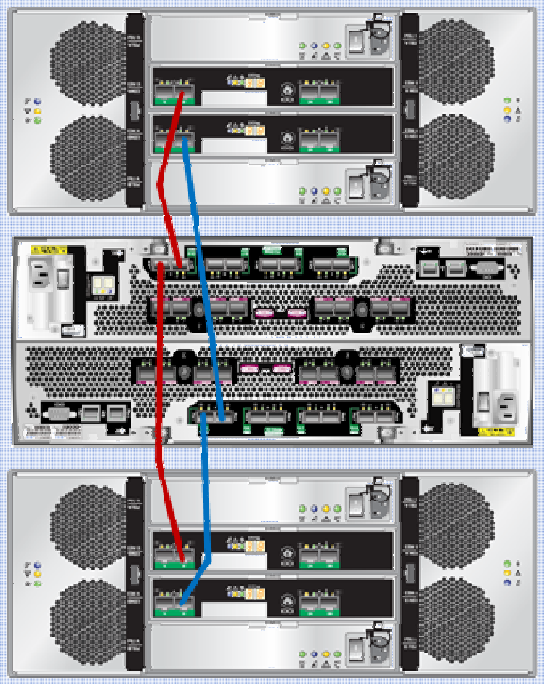
The Evolution of SATA Drive Expansion Units
Raw Capacity in 33U (32U for EXP5060)

2003 – EXP100	2006 – EXP810	2010 – EXP5060
		
27.5 TB in 33U	**88 TB in 33U**	***960 TB in 32U***
110 Drives (250GB)	176 Drives (500GB)	480 Drives (2TB)

EXP5060 – Cabling Options

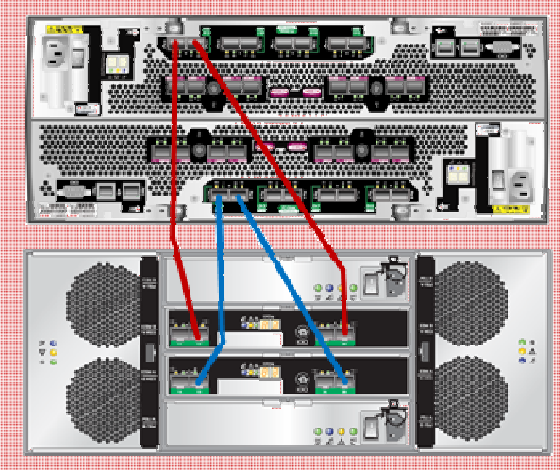
Traditional

- Two loops per enclosure
- Used when intermixing with EXP5000 on loop pair



Trunking

- Four loops per enclosure
- Used to provide maximum throughput per enclosure



DS5x00 performances with GPFS 3.3

Parameter	Value
Read Cache	Enabled
Write Cache mirroring	Disabled
Write Cache without batterie	Disabled
Dynamic prefetch	Enabled
Volume segment Size	512kB
Controller Cache bloc size	32k
Start stop flush	80%

Table 1: Raid volume and controller parameters

GPFS Parameters	Value
maxblocksize	4194304
pagepool	4294967296
maxFilesToCache	10000
maxMBpS	3200
verbsRdma	enable
verbsPorts	mlx4_0/1
nsdbufspace	70
prefetchThreads	96
nsdThreadsPerDisk	4

Table 2: GPFS parameters for 4MB blocksize

DS5300 with 16 x 8 Gbps FC host ports
16 GB cache, 240 SATA drives
24 arrays with RAID6

	GPFS Client Sequential read file creation (GB/s)	GPFS Client Sequential write file creation (GB/s)	GPFS Client Sequential read (GB/s)	GPFS client Sequential write (GB/s)
	5.2	5.0	5.0	5.2

DCS9900

IBM System Storage DCS9900 Features

- Massive storage for highly scalable data streaming applications
- Designed to support managed Quality of Service to provide uninterrupted data delivery
- Hardware enabled RAID 6, which protects data in the event of double disk failure in the same redundancy group
- LUN mapping/LUN masking by host WWN
- Support for SATA and/or SAS Drives
- Dual controller with 5GB of RAID protected cache (2.5GB cache per controller)
- Eight FC8 or IB 4X DDR host ports (4 ports per controller)
- Twenty SAS 4-lane (3Gb/s)connections (10 per controller)
- 1024 LUNS, 512 concurrent logins (IB) and 1024 concurrent logins (FC) per DCS9900 couplet
- Full duplex host transfer operation, sustained performance up to 5.7 GB/s in both reads/writes
- No performance penalties in degraded mode operation, very fast rebuild rate
- Tier journaling with partial rebuild
- Block level virtualization, to virtualize storage deployment and system management
- SNMP, GUI, Telnet and API support

DCS9900 Features for High Performance

- Bandwidth or throughput requirements
- Low latency and QoS features
- FC8 or IB 4x DDR host interface
- Shared or parallel filesystems
- Large capacities
- SAS or SATA drives
- SAN or NAS environments

Key Technical Differentiators

Performance

- *Up to 5.7 GB/sec bandwidth with large sequential I/O*
- **Writes as fast as reads**
- **No loss of performance during drive rebuilds**

Density

- *Up to 600 drives/600 TB per rack*

Scalability

- *Up to 2.4 PB per system in just two racks*
- **Scaling to multiple PB with multiple systems under parallel file system**

Energy efficiency

- **Fewer controllers, power supplies and fans per TB**

Availability and reliability

- **Hardware based RAID 6…8+2…with no performance penalty**
- **Parity computed on every read – no SATA silent corruption errors**
- **All data remains visible to all clients even in the case of a disk, enclosure or controller failure – parallel architecture at every level**
- **Redundant power supplies and fans**

Simplicity

- **One array where others require many**
- **All managed from a single controller pair**
- **Storage can be managed from a single client on the SAN**
- **Fewer servers, switches, power supplies, fans and cables**

Twice the performance and 25% more capacity than the DCS9550

Key Benefits

HPC
- Less time lost waiting for checkpoints
- More capacity to capture results
- Faster computations

→
- More oil reserves found, quicker
- Better weather predictions, sooner
- Better plane & auto designs, faster

Digital Media
- QoS: zero frame loss due to disk, enclosure or controller errors
- Real-time collaboration through access to shared data
- 4 uncompressed 4K HD streams (1 GB/second each)
- 640+ uncompressed HD streams (25-50 Mbps each)
- 8000+ MPEG2 video streams (3.75 Mbps each)

→
- Quicker monetization of film assets
- Better utilization of creative talent

All
- Less waiting - high performance read and write
- Management simplification through storage consolidation
- Cost savings through use of SATA drives

→
- Reduced management costs
- Energy savings
- Less data center build-out

DCS9900 RAID 6 Advantages

- Always 8 data drives (A-H), plus 2 dedicated Parity drive(s)
- Allows drives to reorder commands
- No performance impact in degraded modes
- RAIDed cache – Both read and write operations
- Parity calculated on both read and write operations
- One stripe unit per disk channel (Byte striping)

DCS9900 Data Flow

•Serial Data Stream received from host into front-end interface (FC8 or IB 4x DDR)

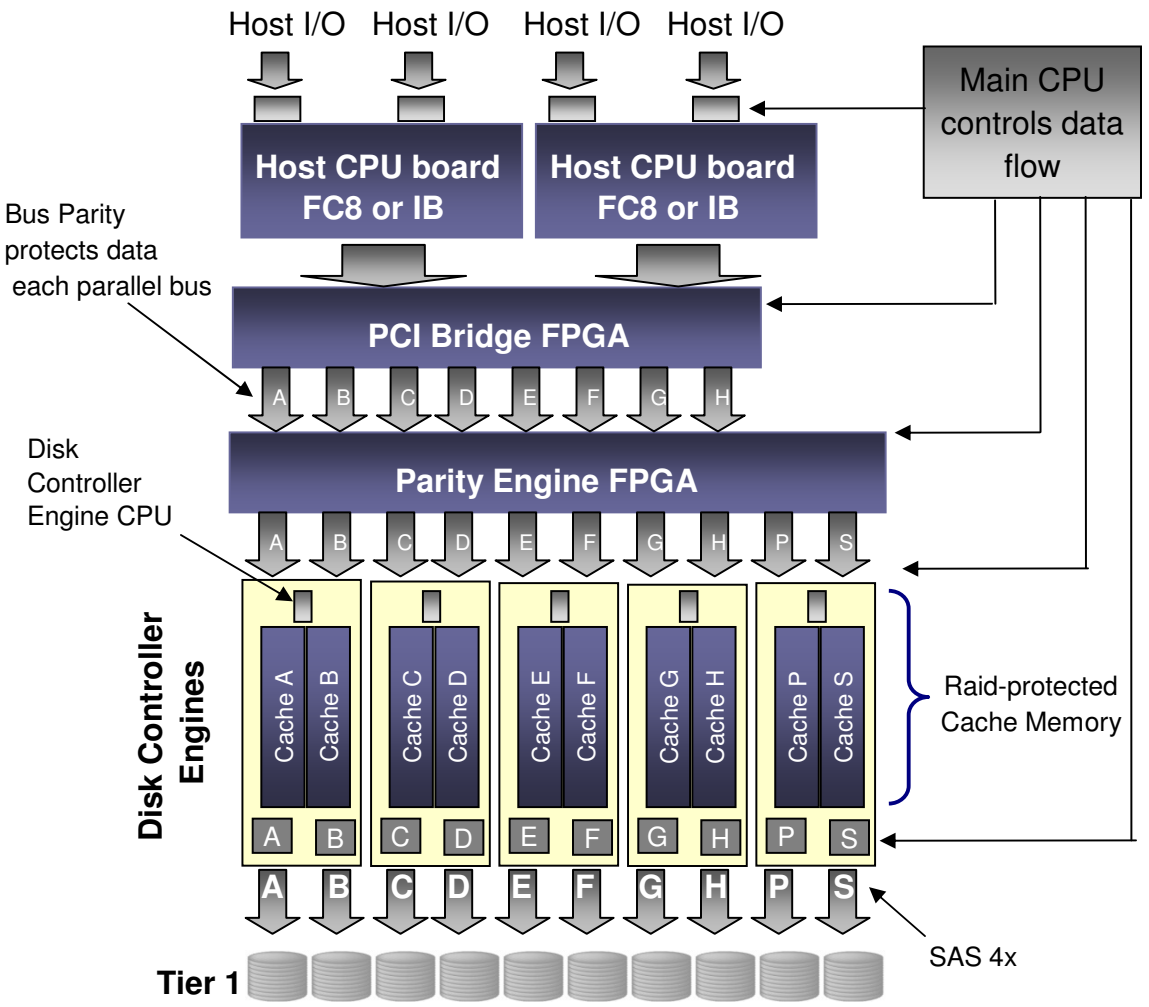
•Serial data stream enters PCI Bridge where the serial stream is divided into eight 512 byte parallel streams

•Parallel stream enters Parity Engine where one or optionally two parities are calculated from the data synchronously

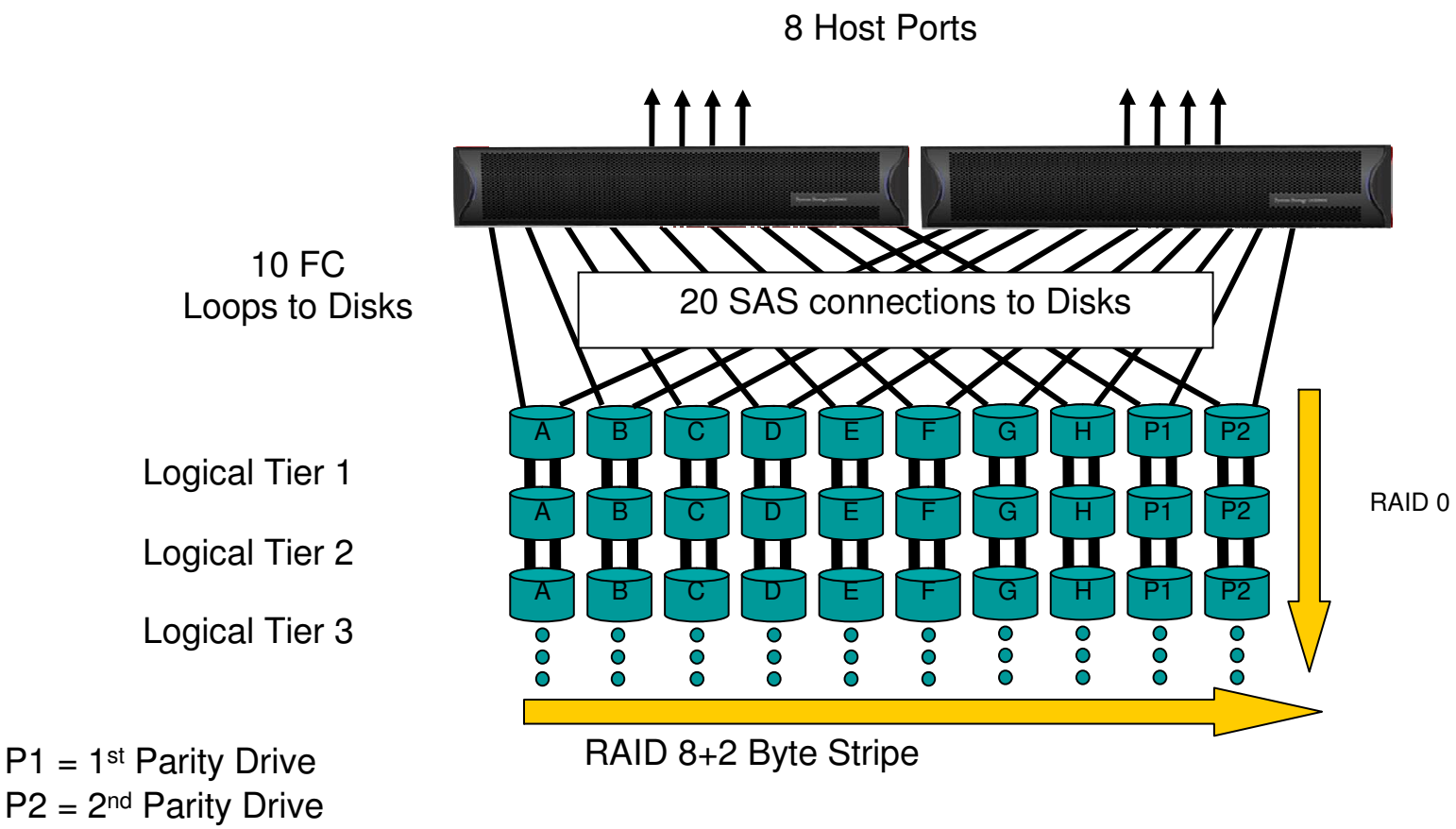
•Work orders from the central CPU received by Disk Controller Engines CPU's Queued command reordering in queued Cache.

•Simultaneous parallel data stream sent out the back-end disk interfaces

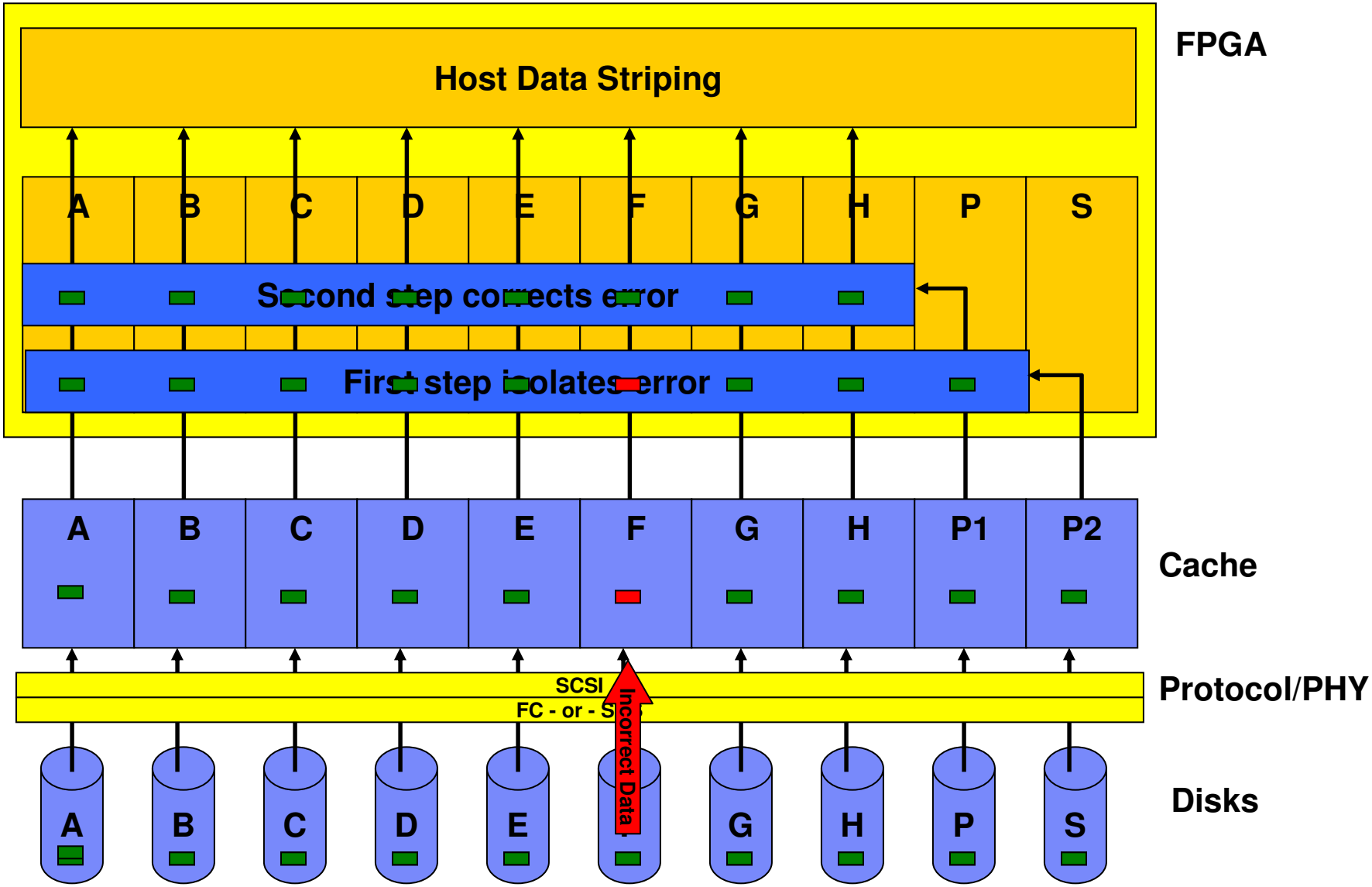
•Disks are arranged in a vertical stripe across back-end channels in a "tier"



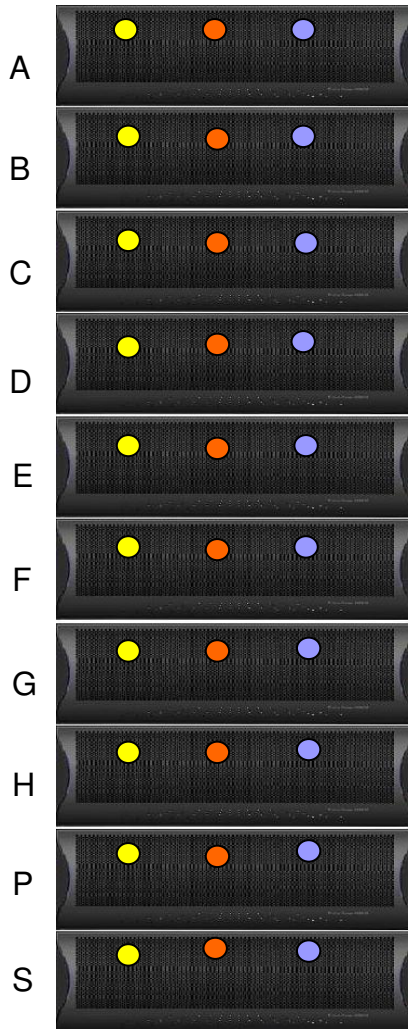
DCS9900 Backend Parallelism



Data Corruption Error Handling



Tiers of Disks, continued



A tier is a physical grouping of disks

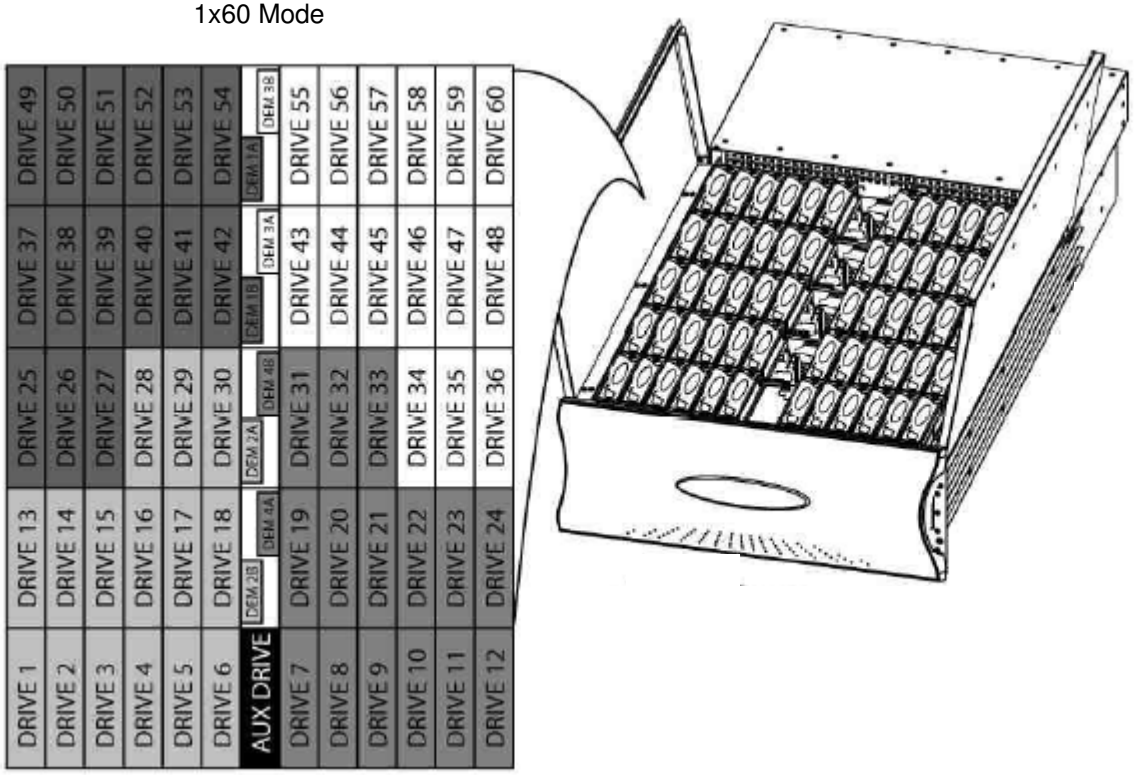
Data is byte striped down the drives in the tier.

The byte stripe consists of 8 data drives formatted to 512 bytes/block each.

The total byte stripe is 4kb

1269-3S1 Storage Expansion Unit
60 Bay Drive Mapping

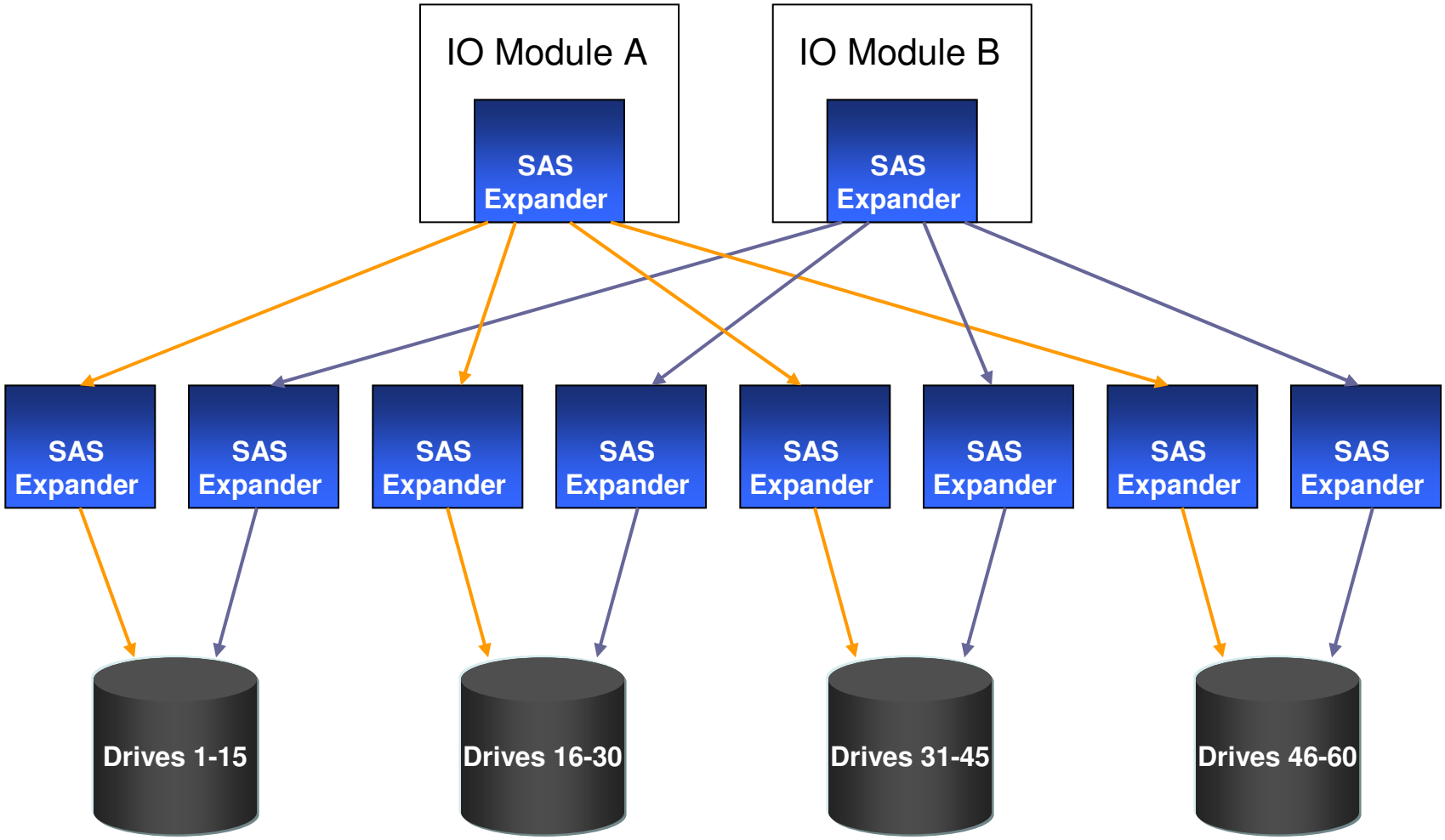
1x60 Mode



2x30 Mode

DRIVE 1	DRIVE 7	DRIVE 13	DRIVE 19	DRIVE 25
DRIVE 2	DRIVE 8	DRIVE 14	DRIVE 20	DRIVE 26
DRIVE 3	DRIVE 9	DRIVE 15	DRIVE 21	DRIVE 27
DRIVE 4	DRIVE 10	DRIVE 16	DRIVE 22	DRIVE 28
DRIVE 5	DRIVE 11	DRIVE 17	DRIVE 23	DRIVE 29
DRIVE 6	DRIVE 12	DRIVE 18	DRIVE 24	DRIVE 30
DEM 2B	DEM 4A	DEM 2A	DEM 4B	DEM 1B	DEM 3A	DEM 1A	DEM 3B
AUX DRIVE	DRIVE 1	DRIVE 7	DRIVE 13	DRIVE 19	DRIVE 25
	DRIVE 2	DRIVE 8	DRIVE 14	DRIVE 20	DRIVE 26
	DRIVE 3	DRIVE 9	DRIVE 15	DRIVE 21	DRIVE 27
	DRIVE 4	DRIVE 10	DRIVE 16	DRIVE 22	DRIVE 28
	DRIVE 5	DRIVE 11	DRIVE 17	DRIVE 23	DRIVE 29
	DRIVE 6	DRIVE 12	DRIVE 18	DRIVE 24	DRIVE 30

1269-3S1 Storage Expansion Unit
SAS Expanders



Enabling Technology – Sleep Mode

Up to 85% of data **is** dormant **and** infrequently accessed

For many customers, **it is desirable to keep this data** readily accessible **while finding** more economical **ways to store it**

Sleep Mode reduces power and cooling costs by idling designated drive tiers

- **Drive controllers remain active to reduce delays on awakening**
- **System allows specific drives to be idled – keeping important data readily accessible**
- **Industry-leading spin-up time**
- **No impact on performance when not in sleep mode**

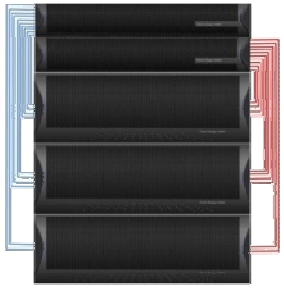
Sleep Mode ***supports IBM's Big Green Initiative***

A single 960TB DCS9550 system can realize savings up to $28,740 per year in power and associated cooling costs in expensive metro areas

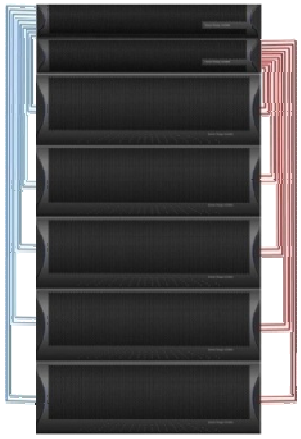
Configuration	Fully Active	80% Sleep	Power Savings	Annual Cost Savings @ $0.20/KWHr *
240TB (240 x 1TB SATA)	**4.89 kW**	**2.84 kW**	**2.05 kW**	$7.185
480TB (480 x 1TB SATA)	**9.29 kW**	**5.19 kW**	**4.10 kW**	$14,370
960TB (960 x 1TB SATA)	**18.1 kW**	**9.88 kW**	**8.20 kW**	$28,740

* Includes savings from power required for cooling, estimated to be equal to savings from direct power.

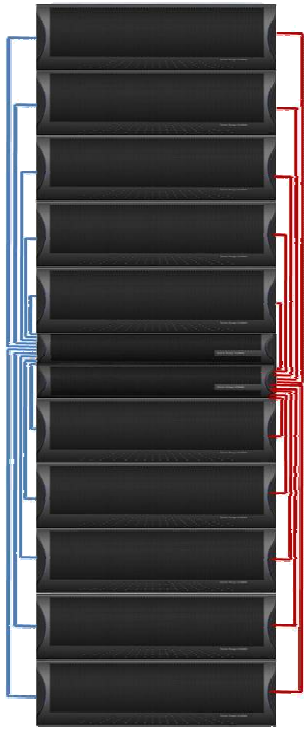
DCS9900 Configurations



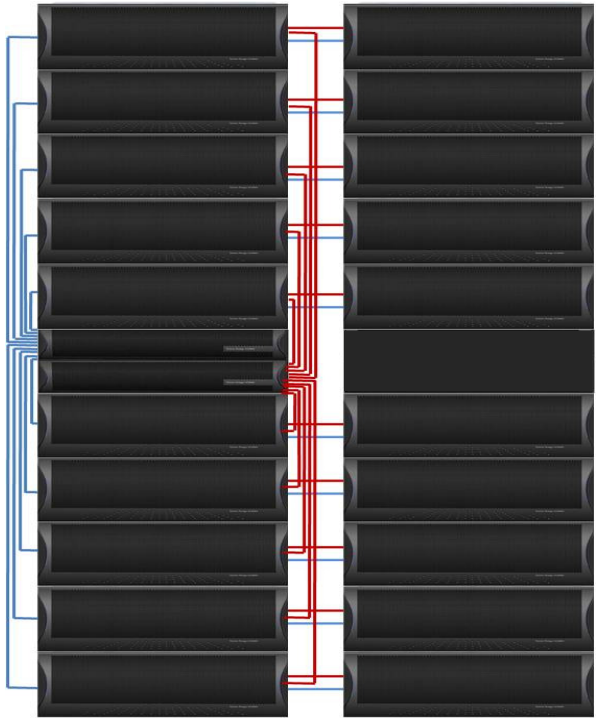
Three 60-Slot Enclosures
150 Drives
Up to:
300 TB raw w/2TB Drives
229 TB usable*
8 FC8 or IB 4x DDR ports
1 @ 42U or 45U Rack



Five 60-Slot Enclosures
150 - 300 Drives
Up to:
600TB raw w/2TB Drives
458 TB usable*
8 FC8 or IB 4x DDR ports
1 @ 42U or 45U Rack



Ten 60-Slot Enclosures
150 - 600 Drives
Up to:
1.2 PB raw w/2TB Drives
896 TB usable*
8 FC8 or IB 4x DDR ports
1 @45U Rack

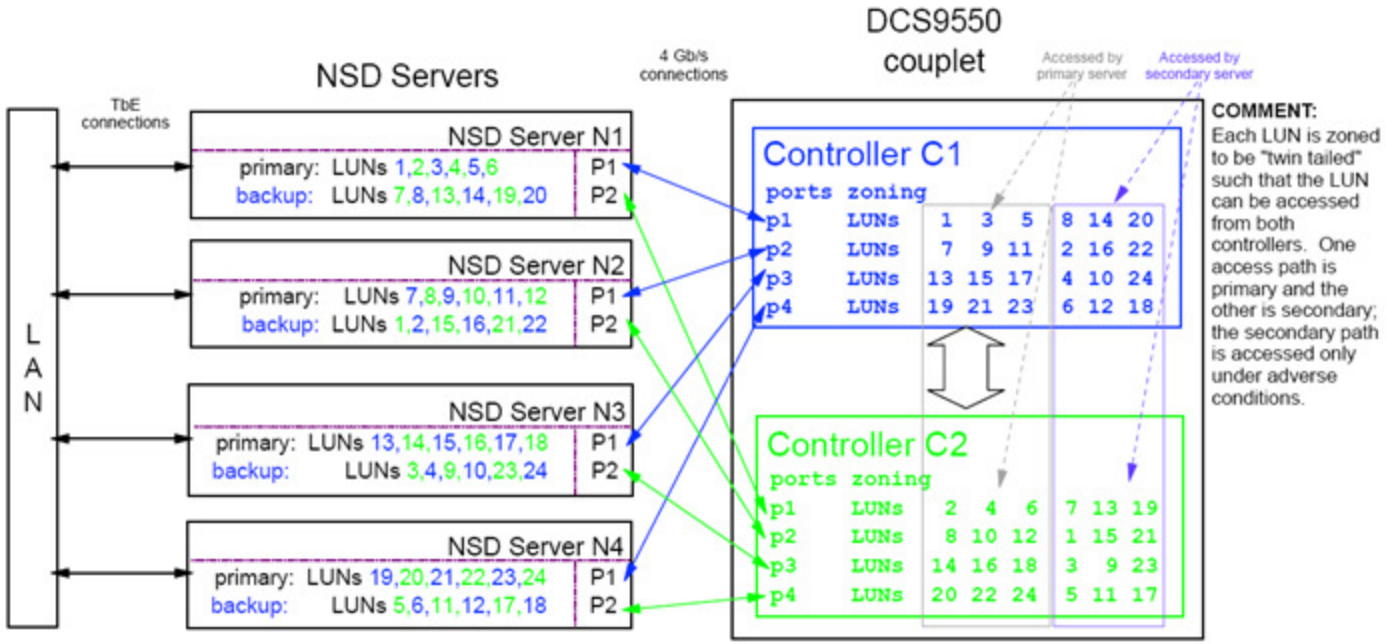


Twenty 60-Slot Enclosures
150 - 1200 Disks
Up to:
2.4 PB raw w/2TB Drives
1.832 PB usable*
8 FC8 or IB 4x DDR ports
2 @ 45U Rack

←========== SAS or SATA ==========→

←=== SAS or SATA or SAS/SATA Intermix ==→

Disk Controller DCS9xx0 logical configuration



COMMENTS:

- Requires TbE or other high speed LAN
- This configuration completely satisfies all 4 design criteria.
- Topologically equivalent configurations can be achieved using different cabling and zoning schema.

COMMENTS:

- C1 "owns" LUNs 1,3,5,7,9,11,13,15,17,19,21,23
- C2 "owns" LUNs 2,4,6,8,10,12,14,16,18,20,22,24
- Requries cache coherence to be enabled
  - this is a DCS9550 parameter
- Controllers and drivers support "active:active" protocol

Future (and now)

New solutions

- SONAS
- Next DCS xxxx
- DS8800

SONAS Architecture

Solves the storage problem

3 basic components

✓**Interface Nodes = *how fast***
✓**Storage Pods = *how big***
✓**Management Node**

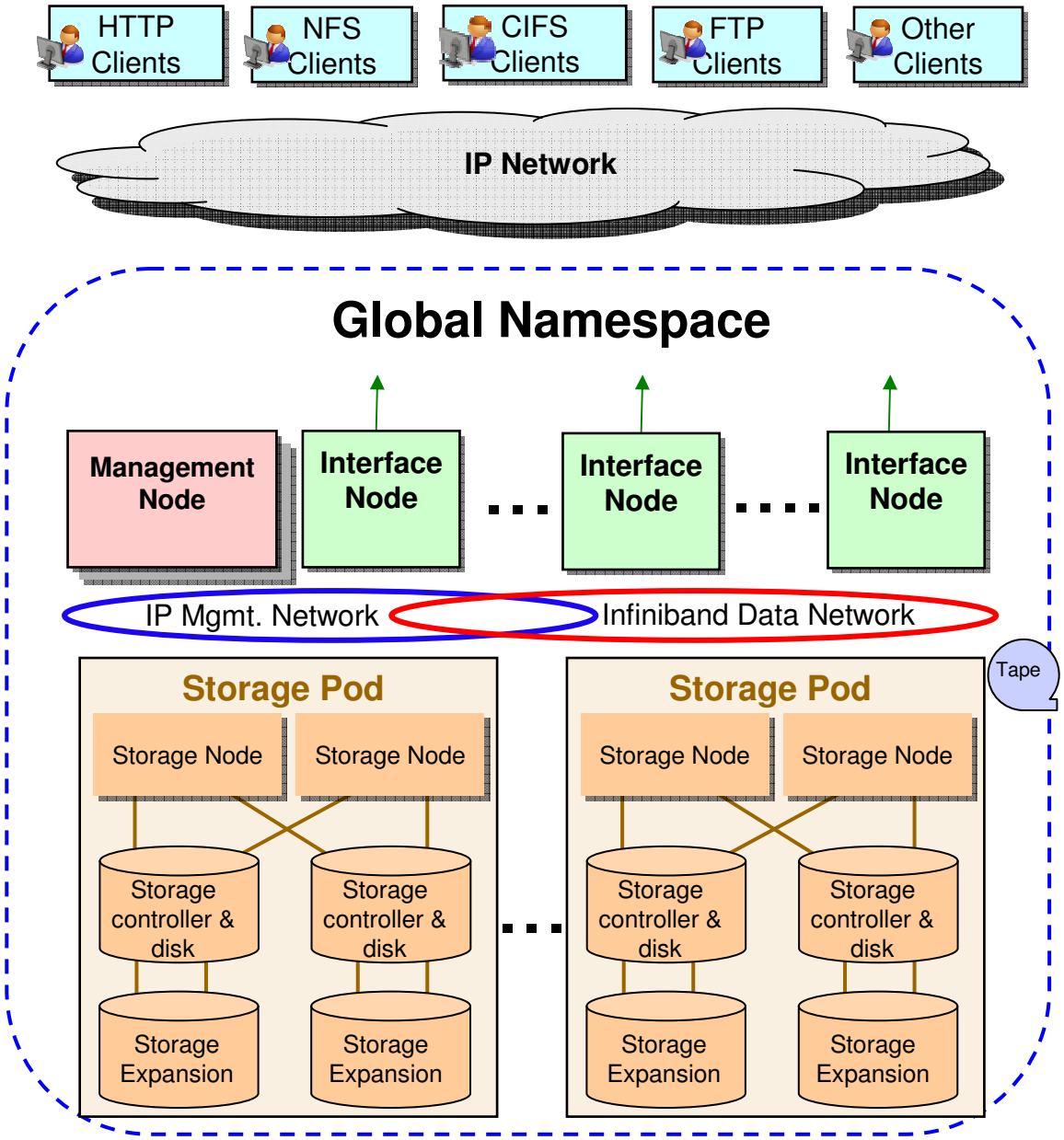
All nodes are clustered for availability
•Users connected through 1GbE or 10GbE
•All nodes are connected through private Infiniband network

Parallel Grid Architecture

•Massive linear scalability
•High performance
•High availability & redundancy

SONAS Software runs on all nodes

•Policy automation
•Global file system
•GUI and operating system



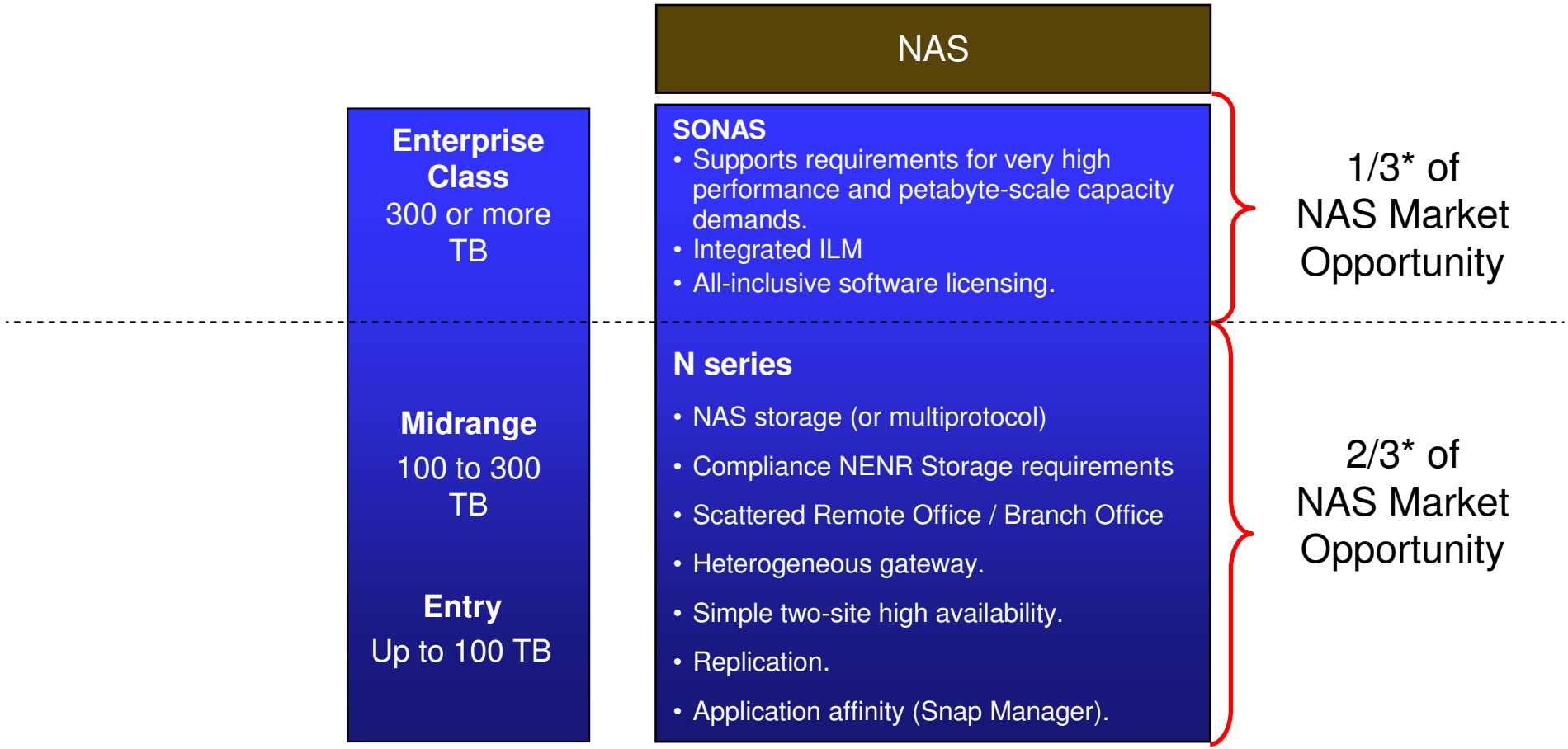
SONAS Performance Learning Points: Parallelism, Scale

- SONAS performance is a clustered, scale-out approach
- Single stream performance from SONAS to a single client is pretty good, but *that by itself isn't hugely different than an unconstrained traditional NAS filer with a single stream*
- The performance problem with traditional filers is that:
  - as you start to add users the traditional system degrades at some point
  - but at that point, the SONAS continues to run with linear scale
  - and SONAS will continue to scale just by adding nodes or storage
- **The real power of SONAS performance** is ***parallelism*** and ***massive scale***: multiple concurrent users, multiple concurrent streams, across different attaching platforms, with massive scalability, including multiple writers to the same file

Watch for **1H2011** announcements in **SONAS performance benchmarking**

IBM N Series and IBM Scale Out NAS Positioning

N series is a very important component of the IBM Portfolio
SONAS is positioned for specific opportunities



New „HPC“ Disk Systems

DCS3700 for the HPC Market

planned to be announced in May

- Dual-active intelligent array controllers enabled with Turbo Performance
- 4 GB mirrored Cache (upgradeable to 8 GB)
- 4x 6 Gb SAS host ports standard with optional 8x 8 Gb FC host ports
- Sixty SAS drive bays, with support for up to 180 drives with the attachment of two DCS3700 Expansion Units
- Data replication options with FlashCopy, Volume Copy, and Remote Mirroring over Fibre Channel
- IBM DS Storage Manager for administrative and management activities
- 128 Partitions

DCS3700 Expansion

- Dual-active environmental services modules (ESMs)
- Sixty SAS drive bays
- 6 Gb SAS attachment to the DCS3700 Storage System or to another DCS3700 Expansion Unit

Both models feature:

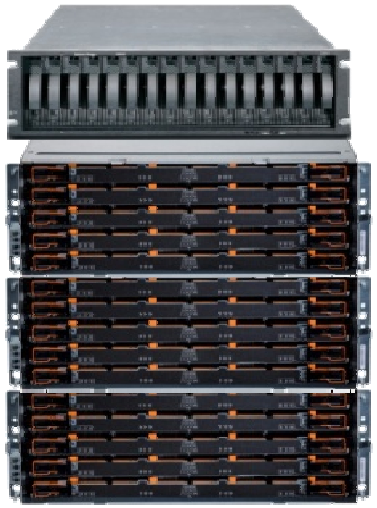
- Slim 4U, 19-inch rack mount enclosure
- Dual port, hot-swappable 2 TB SAS nearline disk drives
- Redundant, hot-swappable hardware components
- IBM installation and a one year warranty with 24 x 7 on-site repair, 6 hour response

HPC-Entry Disk Systems

DCS3700 for the HPC Market

Outlook

It is IBM's current plan and intent to offer
- additional drive options
for the DCS3700 in second half 2011,
including high-performance disk drives and solid state drives.



HPC-Storage installations in Italy

- Cineca – Bologna
  - 6 x DCS9900 x 1,5 PB
- CMCC Euromed – Lecce
  - 2 x DCS9900 x 240 TB
  - 1 x DS4800 x 20 TB
- Enea Cresco – Napoli
  - 1 x DCS9550 x 120 TB
- Banca Intesa – Parma
  - 2 x DS5300 x 10 TB
- Cineca x ENI – Bologna
  - 3 x DCS9900 x 1 PB
  - 3 x DDN SFA10000 x 3 PB (on going)
  - 2 x DS5100 (metadata) (on going)

Questions ?

That's all folks !
Thanks !