

PB Sized Storage Systems Are Not Unusual

Symposium Presentation
28th IEEE Conference on Massive Data Storage



Raymond L. Paden, Ph.D.
HPC Technical Architect
IBM Deep Computing
raypaden@us.ibm.com
512-286-7055

Version 1.0c
17 April 2012

Tutorial Outline

1. Storage Technology
Past, Present and Future Storage Technology
2. Using GPFS to Manage a Storage Hierarchy
3. Examples

Storage Capacity and Data Rates Are Growing Fast!

1965

- Capacity < 205 MB
- Streaming data rate < 2 MB/s
- 26 platters laterally mounted
- Rotational speed = 1200 RPM

CDC 6603

205 MB

2 MB/s

1965

1987

- Capacity < 1.2 GB
- Streaming data rate < 3 MB/s
- 2 spindles
- Rotational speed = 3600 RPM
- Average seek time = 12 ms

IBM SP Frame with SSA Disk

8 x Storage Trays (128 Disks)

1.2 TB @ 480 MB/s

1997

1997

- Storage Trays
- Capacity < 9 GB per disk
- Streaming data rate < 21 MB/s
- Rotational speed = 10 Krpm
- Average seek time = 7.7 ms



**IBM 3380
Model CJ2**
1.2 GB
3 MB/s
1987



IBM 3380 from IBM Archive

Storage Capacity and Data Rates Are Growing Fast!

2009

- SATA or Near Line SAS
 - ◆ Capacity < 1000 GB
 - ◆ Streaming data rate < 150 MB/s
 - ◆ Rotational speed = 7200 RPM
 - ◆ Average seek time = 9 ms
- Fibre Channel or Enterprise SAS
 - ◆ Capacity < 600 GB
 - ◆ Streaming data rate < 425 MB/s
 - ◆ Rotational speed = 15000 RPM
 - ◆ Average seek time = 3.6 ms



IBM DCS9900
10 x Storage Trays (600 Disks)
600 TB @ 5 GB/s
2009

**Performance
and Capacity
are Diverging!**

2012

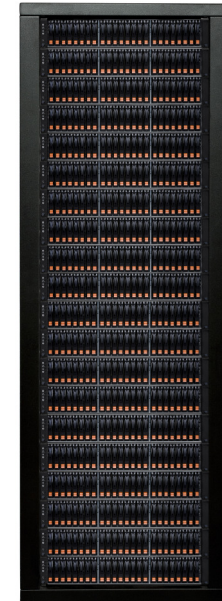
- Near Line SAS
 - ◆ Capacity < 3000 GB
 - ◆ Streaming data rate < 210 MB/s
 - ◆ Rotational speed = 7200 RPM
 - ◆ Average seek time = 9 ms
- Enterprise SAS
 - ◆ Capacity < 300 GB
 - ◆ Streaming data rate < 600 MB/s
 - ◆ Rotational speed = 15000 RPM
 - ◆ Average seek time = 2 ms

Capacity
Optimized



IBM DCS3700
10 x DCS3700
600 x 3 TB Disk
1.8 PB
Stream < 20 GB/s
Random < 70,000 IOP/s

Performance
Optimized



IBM DS3524
20 x DS3524
240 x 300 GB Disk
144 TB
Stream < 16 GB/s
Random < 90,000 IOP/s

Storage Capacity and Data Rates Are Growing Fast!

2018

- Disk Capacity 30 TB¹
- Disk speed 380 MB/s¹

Concern: Will the number of disks needed to meet capacity requirements fail to meet performance requirements?

Performance:Capacity Ratio

1965: 9.8 MB/s per GB

1987: 2.5 MB/s per GB

1997: 2.3 MB/s per GB

2012: capacity optimized = .050 MB/s per GB (3.5" x 7200 RPM NL-SAS)

performance optimized = 2.0 MB/s per GB (2.5" x 15000 RPM SAS)

2018: .013 MB/s per GB (Capacity Optimized)¹

2009: 1 PB (1200 x 1 TB/disk) < 10 GB/s

2012: 1 PB (340 x 3 TB/disk) < 12 GB/s

2018²: 1 PB (34 x 30 TB/disk) < 4 GB/s

Hope for 2018: Will customers use proportionally enough more disks in 2018 to offset the dropping performance:capacity ratio?

1. Gary Grider, Exa-Scale FSIO, 07/2010, LANL. Some storage architects believe Grider's prediction is over optimistic.

2. Assumptions: 30 TB disks can sustain Grider's projected 384 GB/s per disk and that RAID controller technology in 2018 can harvest data rates these disks as efficiently as we can today.

Storage Capacity and Data Rates Are Growing Fast!

2012 – Emerging Technology: SSD

Many flavors of SSD

- SSD Controllers; e.g., RamSan 820
 - ◆ eMLC Flash
 - ◆ Capacity < 1.008 PB
 - ◆ Streaming data rate < 168 GB/s
 - ◆ IOP Rate < 18.9 MIOP/s
 - ◆ Seek time: write \approx 100 us, read \approx 25us
- SSD Block Devices; e.g., DCS3700 with embedded SSD*
 - ◆ 2.5" eMLC Devices (200 GB and 400 GB)
 - ◆ Per block device statistics
 - Streaming data rate < 250 MB/s
 - IOP Rate < 26,000 IOP/s

* Can effectively use only up to 10 x SSD per DCS3700.

- SSD PCIe Card; e.g., Fusion-io ioDrive2+
 - ◆ MLC Flash (365 GB, 785 GB, 1.2 TB)
 - ◆ Streaming data rate < 1.2 GB/s
 - ◆ Latency: write < 15 ms, read < 68 ms

+ **Not** recommended for GPFS; it is a good product, but since it is embedded in a server, it becomes a single point of failure.



RamSan 820

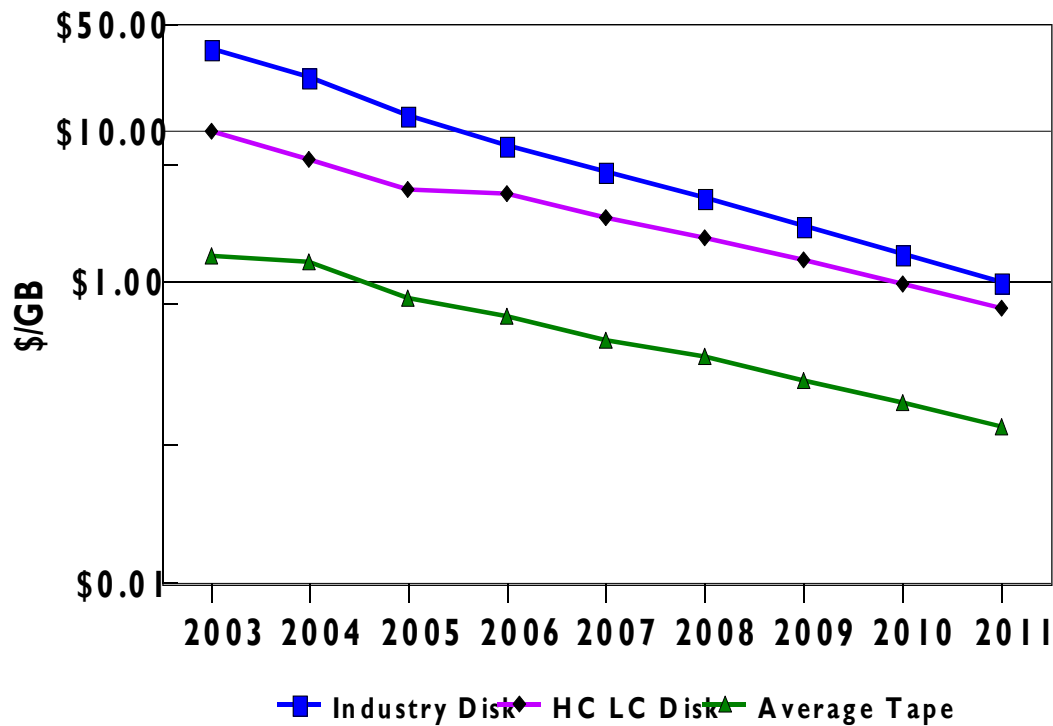


DCS3700
embedded SSD and Disk

Storage Capacity and Data Rates Are Growing Fast!

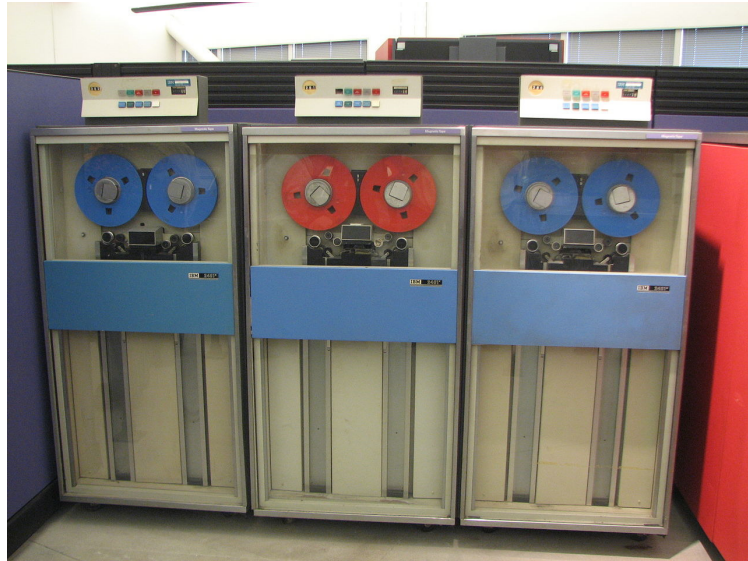
... And don't forget tape! It ain't dead!

Storage Prices



Source: Disk - Industry Analysts, Tape - IBM

Tape: Past and Today



From Wikimedia (Author: Erik Pitti from San Diego, CA, USA)
Use of this slide does not imply endorsement of this presentation.

Yesterday: 9 Track Tape
Introduced in 1964 with IBM System/360

IBM 2401
1600 BPI Cartridge
Capacity < 40 MB
Data rate < 320 KB/s (i.e., up to 200 in/s)
Library capacity: How big was your closet?

Today

IBM TS3500 Tape Library*

Tape drives:

- TS1050: 1.5 TB per cartridge @ 140 MB/s
- TS1140: 4.0 TB per cartridge @ 250 MB/s

Maximum Configuration (TS1140 with 3592 cartridges)

- Number of tape drives < 2880
- Capacity < 900 PB
- Data rate < 720 GB/s

* Statistics based on uncompressed data format (i.e., native).



From: IBM Asia Pacific Hardware Announcement AG11-0092, May 9, 2011

When customers buy storage, they do not want just a bucket of parts. They want to buy a solution.

Customers need a solution to manage storage in a integrated/hierarchical fashion.

What Is GPFS?

GPFS = General Parallel File System

GPFS GA date = 1998

GPFS is IBM's *shared disk, parallel clustered file system.*

Shared disk:

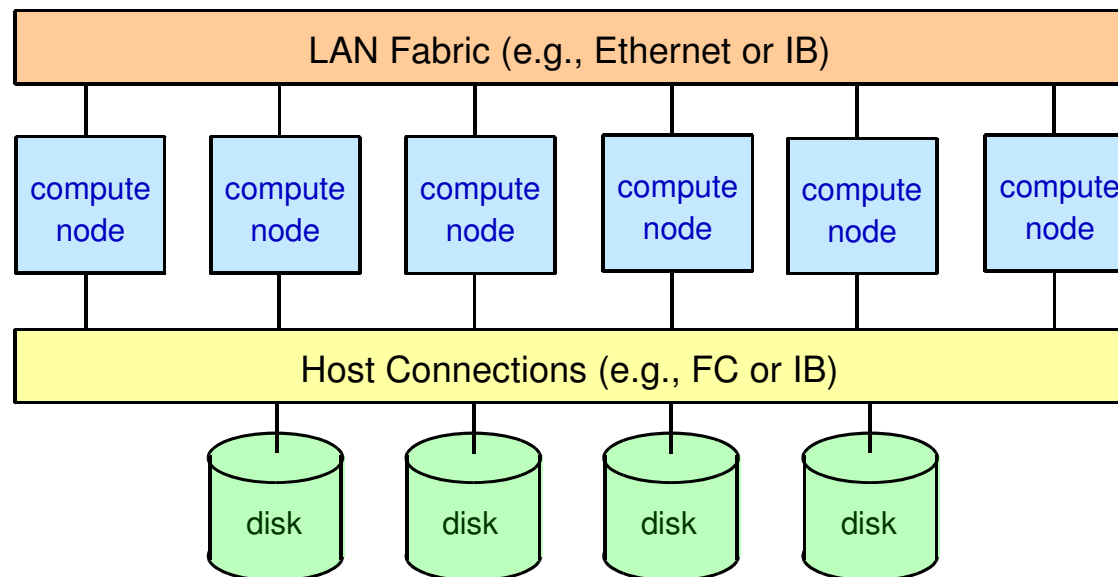
All userdata and metadata are accessible from any disk to any node

Parallel:

Userdata and metadata flows between all nodes and all disks in parallel

Clustered:

1 to 1000's of nodes under common rubric



GPFS Supports both direct and switched host connections.

Overview of GPFS Features

- *General*: supports wide range of applications and configurations
- *Cluster*: from large (5000+ nodes) to small (only 1 node) clusters
- *Parallel*: user data and metadata flows between all nodes and all disks in parallel
- *HPC*: supports high performance applications
- *Flexible*: tuning parameters allow GPFS to be adapted to many environments
- *Capacity*: from high (multi-PB PB) to low capacity (only 1 disk)
- *Global*: Works across multiple nodes, clusters and labs (*i.e.*, LAN, SAN, WAN)
- *Heterogenous*:
 - ◆ Native GPFS on AIX, Linux, Windows as well as NFS and CIFS
 - ◆ Works with almost any block storage device
- *Shared disk*: all user and meta data are accessible from any disk to any node
- *RAS*: reliability, accessibility, serviceability
- *Ease of use*: GPFS is not a black box, yet it is relatively easy to use and manage
- *Basic file system features*: POSIX API, journaling, both parallel and non-parallel access
- *Advanced features*: ILM, integrated with tape, disaster recovery, SNMP, snapshots, robust NFS support, hints

GPFS Architecture

1. Client vs. Server
2. LAN Model
3. SAN Model
4. Mixed SAN/LAN Model

Is GPFS a Client/Server Design?

Software Architecture Perspective: No

There is no single-server bottleneck, no protocol manager for data transfer. The mmfsd daemon runs symmetrically on all nodes. All nodes can and do access the file system via virtual disks (i.e., NSDs). All nodes can, if disks are physically attached to them, provide physical disk access for corresponding virtual disks.

Is GPFS a Client/Server Design?

Practical Perspective: Yes

1. GPFS is commonly *deployed* having dedicated storage servers ("NSD servers") and distinct compute clients ("NSD clients") running applications that access virtual disks (*i.e.*, "NSD devices" or "NSDs") via the file system.
 - This is based on economics (*its generally* too expensive to have 1 storage controller for every 2 nodes)
2. Nodes are designated as clients or servers for licensing.
 - Client nodes only consume data
 - Server nodes produce data for other nodes or provide GPFS management functions
 - producers: NSD servers, application servers (e.g., CIFS, NFS, FTP, HTTP)
 - management function: quorum nodes, manager nodes, cluster manager, configuration manager
 - Server functions are commonly overlapped ← This reduces cost, but use caution!
 - example: use NSD servers as quorum and manager nodes
 - Client licenses cost less than server licenses ← The new licensing model is much cheaper!
 - Server nodes **can** perform client actions, but client nodes can **not** perform server actions

Local Area Network (LAN) Topology

Clients Access Disks Through the Servers via the LAN

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

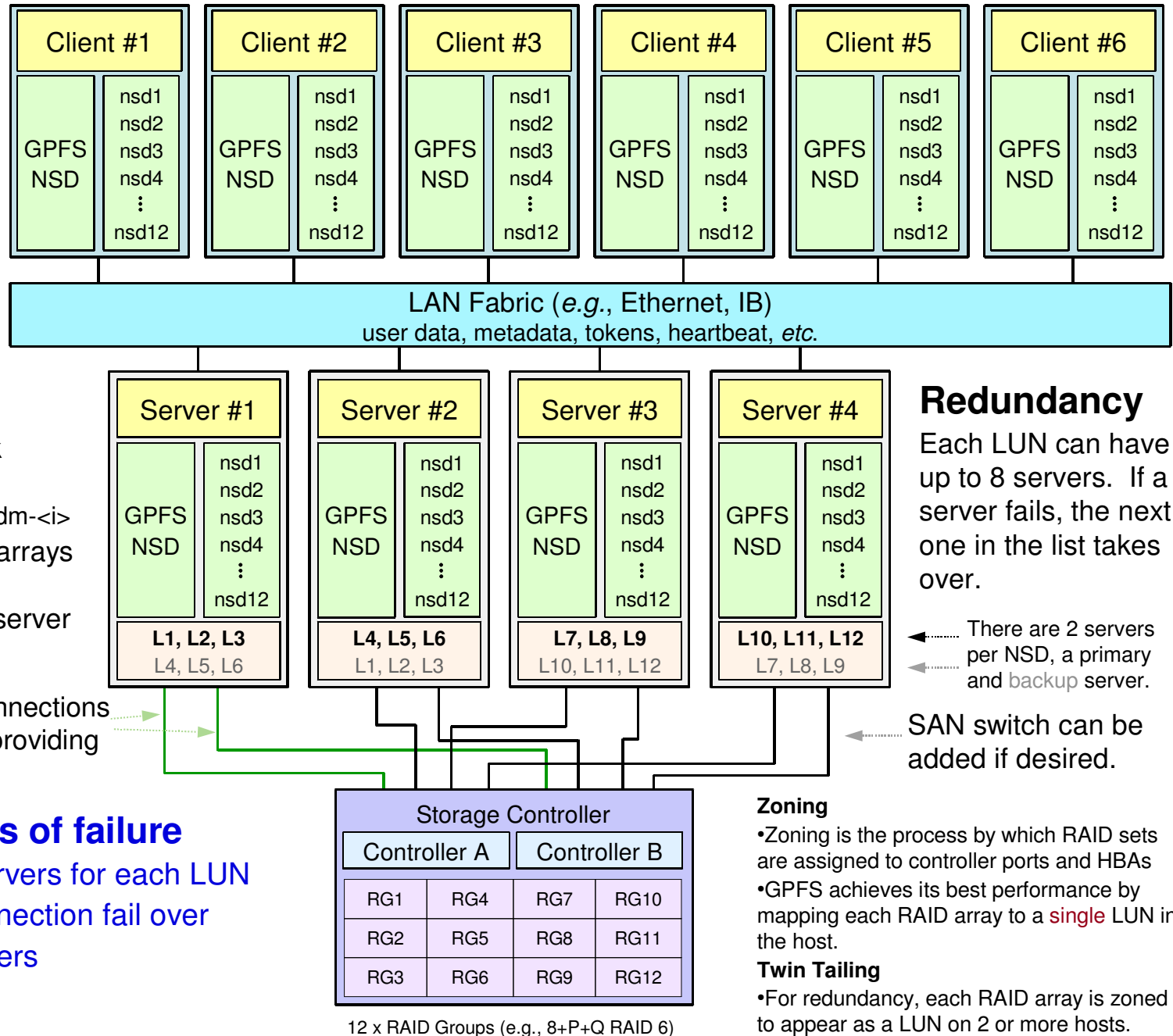
- Logical Unit
- Abstraction of a disk
 - AIX - hdisk<i></i>
 - Linux – sd<c> or dm-<i></i>
- LUNs map to RAID arrays in a disk controller or "physical disks" in a server

Redundancy

Each server has 2 connections to the disk controller providing redundancy

No single points of failure

- primary/backup servers for each LUN
- controller/host connection fail over
- Dual RAID controllers



Redundancy

Each LUN can have up to 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

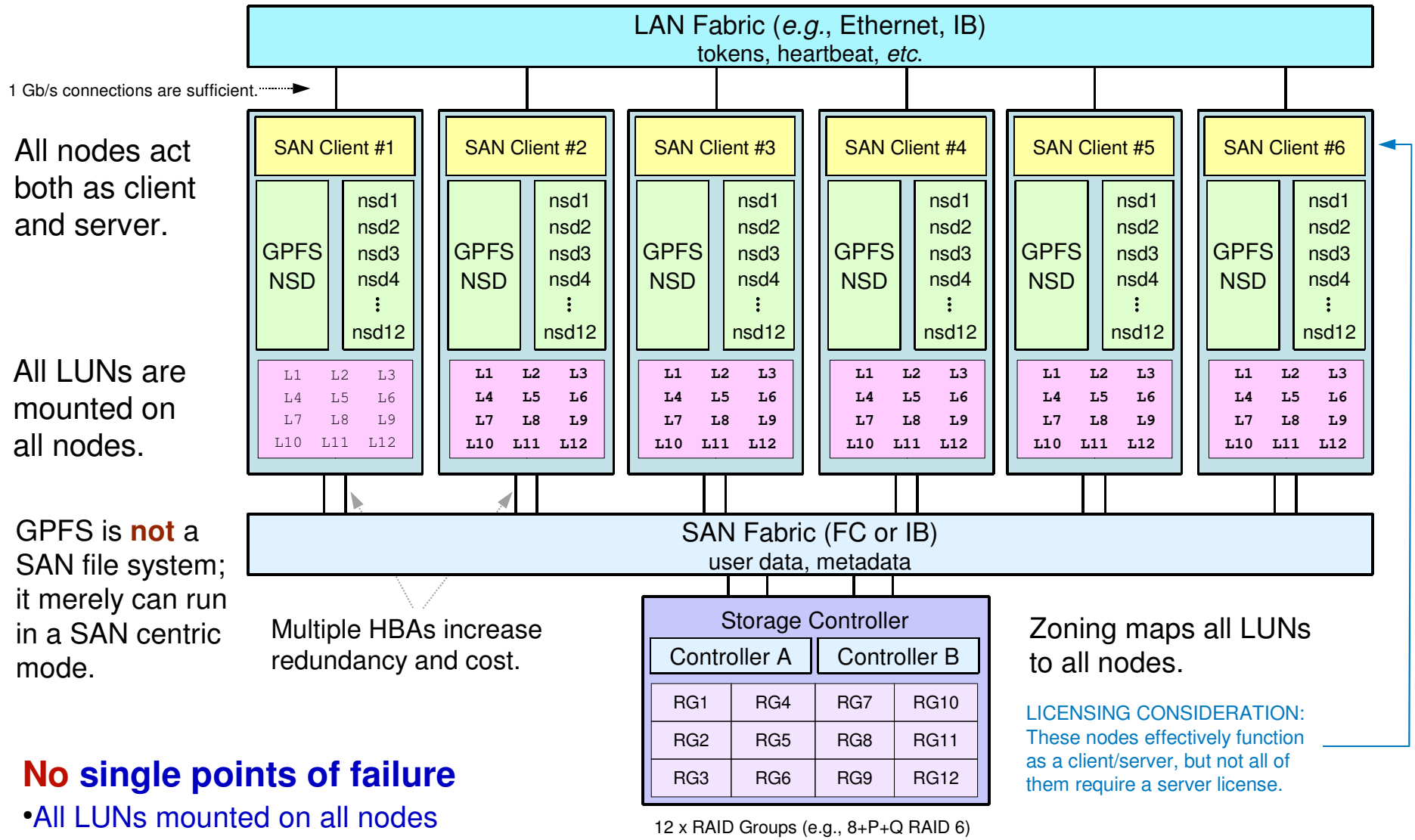
- Zoning is the process by which RAID sets are assigned to controller ports and HBAs
- GPFS achieves its best performance by mapping each RAID array to a **single** LUN in the host.

Twin Tailing

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

Storage Area Network (SAN) Topology

Client/Servers Access Disk via the SAN



No single points of failure

- All LUNs mounted on all nodes
- SAN connection (FC or IB) fail over
- Dual RAID controllers

CAUTION:

A SAN configuration is **not** recommended for larger clusters (e.g., ≥ 64 since queue depth must be set small (e.g., 1))

The largest SAN topologies in production today are 256 nodes, but require special tuning.

Comparing LAN and SAN Topologies

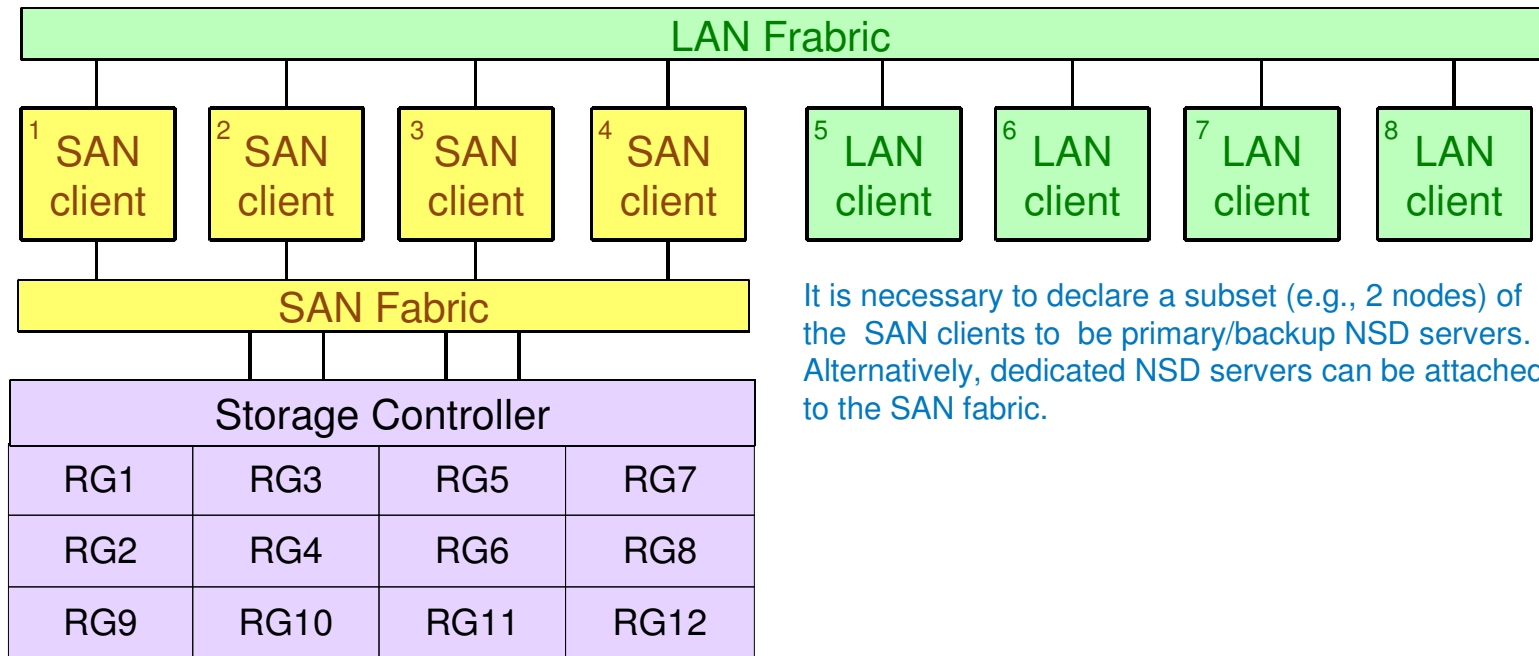
• LAN Topology

- ◆ All GPFS traffic (user data, metadata, overhead) traverses LAN fabric
- ◆ Disks attach only to servers (also called NSD servers)
- ◆ Applications generally run only on the clients (also called GPFS clients); however, applications can also run on servers
 - cycle stealing on the server can adversely affect synchronous applications
- ◆ Economically scales out to large clusters
 - ideal for an "army of ants" configuration (*i.e.*, large number of small systems)
- ◆ Potential bottleneck: LAN adapters
 - *e.g.*, GbE adapter limits peak BW per node to 80 MB/s; "channel aggregation" improves BW

• SAN Topology

- ◆ User data and metadata only traverse SAN; only overhead data traverses the LAN
- ◆ Disks attach to all nodes in the cluster
- ◆ Applications run on all nodes in the cluster
- ◆ Works well for small clusters
 - too expensive to scale out to large clusters (*e.g.*, largest production SAN cluster is 250+ nodes)
 - ideal for a "herd of elephants" configuration (*i.e.*, small number of large systems)
- ◆ Potential bottleneck: HBA (Host Bus Adapters)
 - *e.g.*, assume 180 MB/s effect BW per 4 Gb/s HBA; multiple HBAs improves BW

Mixed LAN/SAN Topology



It is necessary to declare a subset (e.g., 2 nodes) of the SAN clients to be primary/backup NSD servers. Alternatively, dedicated NSD servers can be attached to the SAN fabric.

COMMENTS:

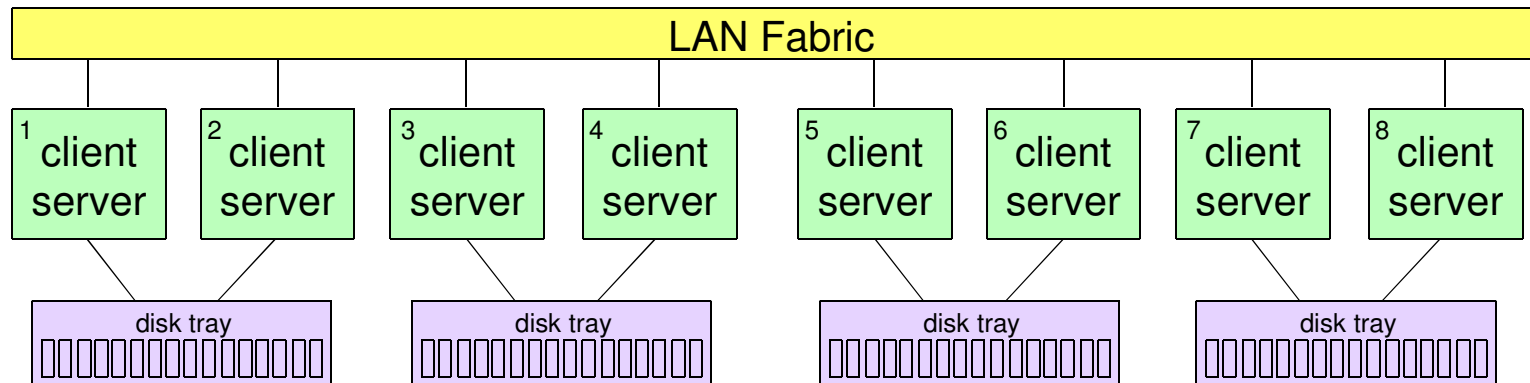
Nodes 1 - 4 (*i.e.*, SAN clients)

- GPFS operates in SAN mode
- User and meta data traverse the SAN
- Tokens and heartbeat traverse the LAN

Nodes 5 - 8 (*i.e.*, LAN clients)

- GPFS operates in LAN mode
- User data, meta data, tokens, heartbeat traverse the LAN

Symmetric Clusters



COMMENTS

No distinction between NSD clients and NSD servers ← Requires special bid pricing under new licensing model

- not well suited for synchronous applications

Provides excellent scaling and performance

Not common today given the cost associated with disk controllers

Use "twin tailed disk" to avoid single point of failure risks

- does not necessarily work with any disk drawer ← New products may make this popular again.
 - do validation test first
 - example: DS3512 - yes, EXP3512 - no

Can be done using internal SCSI

- Problem: exposed to single point of failure risk
- Solution: use GPFS mirroring

GPFS Facilitates a Building Block Design Strategy

A storage building block is the smallest increment of storage, servers and networking by which a storage system can grow.

It provides a versatile storage design strategy, especially conducive to clusters.

Using this strategy, a storage solution consists of 1 or more storage building blocks. This allows customers to conveniently expand their storage solution in increments of storage building blocks (i.e., "build as you grow" strategy).

COMMENT: This solution strategy is facilitated by external storage controllers and file systems that work well within a LAN (e.g., GPFS).

Strategy: Storage Tiers

Storage Building Blocks can be used under the GPFS Information Life-cycle Management (ILM) feature to configure multi-tiered storage solutions.

GOALS

- ▶ Manage data over its life cycle ("cradle to grave")
- ▶ Keep active data on highest performing media and inactive data on tape of low cost, high capacity disk
- ▶ Migration of data is automatic and transparent to the client
- ▶ Lower levels can serve as backup for higher levels

Tier-1

- ▶ Performance Optimized Disk
 - e.g., FC, SAS disk
- ▶ Scratch Space

Tier-2


- ▶ Capacity Optimized
 - e.g., SATA
- ▶ Infrequently used files

Tier-3

- ▶ Local tape libraries

Tier-4

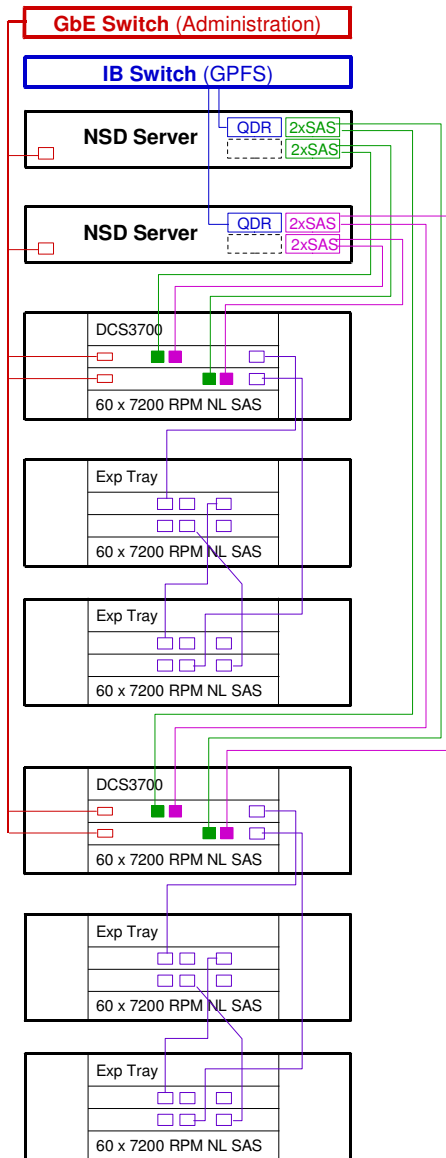
- ▶ Remote tape libraries



frequent use
smaller capacity
high BW/low latency
more expensive

infrequent use
larger capacity
lower BW
higher latency
less expensive

Capacity Optimized Building Block – Logical View



Analysis

NSD Server

- Effective BW per NSD server < 3.0 GB/s
- x3650 M3 with 8 cores and 6 DIMMs (4 GB per DIMM)
- 1 x GbE < 80 MB/s
- 1 x IB QDR² < 3.0 GB/s
- 2 x single port 6 Gb/s SAS adapters

2 x DCS3700 Turbo each with 2 expansion trays

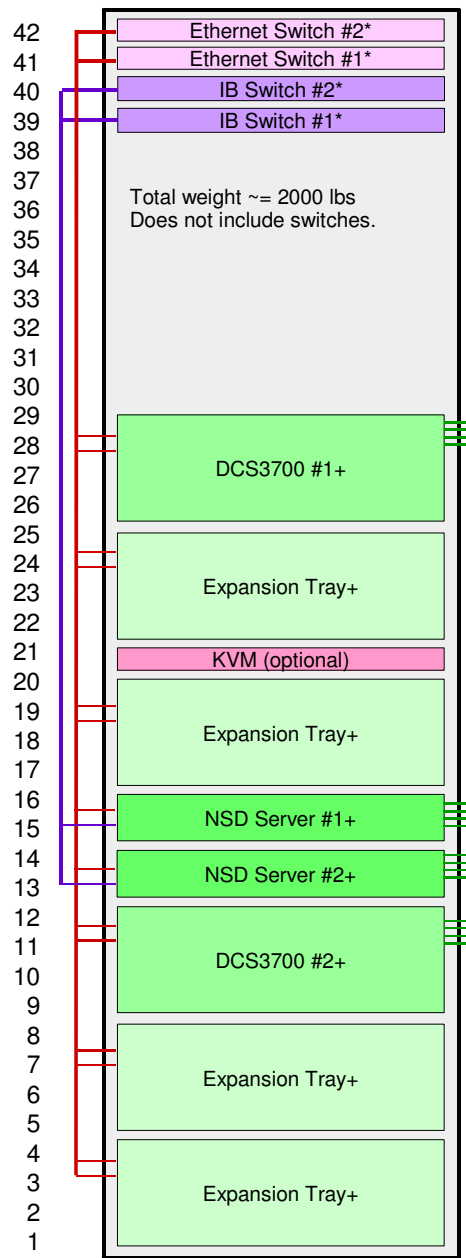
- 360 x 2 TB near line SAS disks
- 36 x 8+2P RAID 6 arrays
- Capacity: raw = 720 TB, usable = 524 TB³
- Performance
- Streaming rate: write < 3.2 GB/s⁴, read < 4.0 GB/s⁴
- IOP rate (random 4K transactions): write < 3600 IOP/s⁵, read < 6000 IOP/s⁵

FOOTNOTES:

1. The x3650 M3 can be replaced with an x3550 M3 if a single dual port SAS HBA in place of 2 single port SAS HBAs.
2. An IB QDR HCA is essential to this solution in order to harvest the full BW potential of the DCS3700s.
3. The DCS3700 provides a capacity of 14.55 TB per RAID 6 array for the file system to use.
4. The stated streaming rates are least upper bounds (LUB); these rates are based on GPFS/DCS3700 benchmarks using 60 x 7200 RPM Near Line SAS disks. Extrapolating from other tests, greater LUB rates may be expected (e.g., write < 1.7 GB/s and read < 2.4 GB/s using at least 80 of these disks).
5. These rates are extrapolated from actual tests using 15000 RPM disk assuming seek rates on 7200 RPM disk < 33% of 15000 RPM disk. These tests assume completely random 4K transactions (n.b., no locality) to raw devices (n.b., no file system). Instrumented code accessing random 4K files will measure a lower IOP rate since they can not measure the necessary metadata transactions. Favorable locality will increase these rates significantly.

COMMENT: This building block is similar to #1a except that by using IB QDR for the LAN, it can double the capacity and performance of the building block while reducing the number of NSD servers. However, when scaling out, building block increments are quite large.

Capacity Optimized Building Block – Physical View



COMPONENTS

4 x NSD servers (x3650 M3) each with the following components:
 - 2 x quad core westmere sockets, 6 x DIMMs (2 GB or 4 GB per DIMM)
 - 1 x GbE, 2 x TbE or 1 x IB QDR, 2 x single port SAS (6 Gb/s)

Switches: Provide Ethernet and IB switches as needed.

Comment: This configuration consists of 1 building blocks. Adding additional building blocks will scale performance and capacity linearly.

AGGREGATE STATISTICS

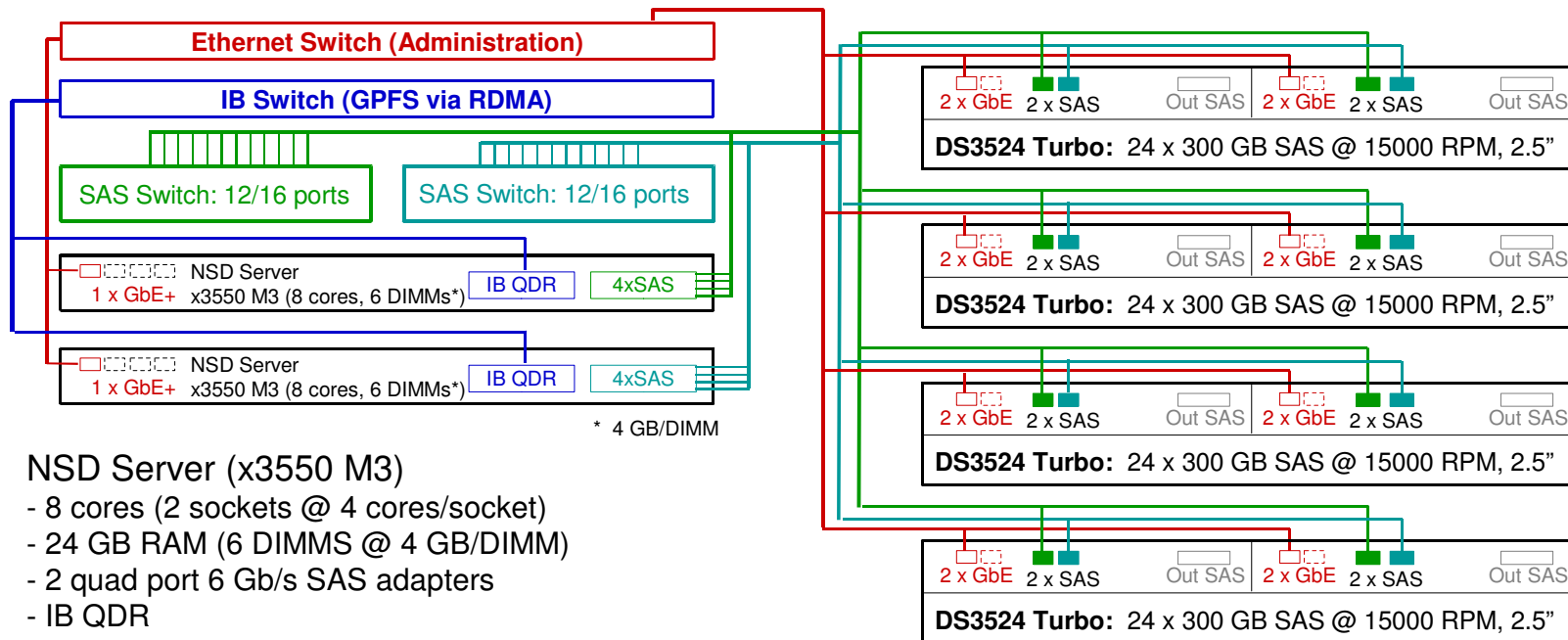
2 x DCS3700 Turbo each with 2 x EXP3560 trays
 - 360 x 2 TB near line SAS disks
 - 36 x 8+2P RAID 6 arrays
 - Capacity: raw = 720 TB, usable = 524 TB³
 - Performance
 Streaming rate: write < 3.2 GB/s⁴, read < 4.0 GB/s⁴
 IOP rate (random 4K transactions): write < 3600 IOP/s⁵, read < 6000 IOP/s⁵

FOOTNOTES:

1. The x3650 M3 can be replaced with an x3550 M3 if a single dual port SAS HBA in place of 2 single port SAS HBAs.
2. An IB QDR HCA is essential to this solution in order to harvest the full BW potential of the DCS3700s.
3. The DCS3700 provides a capacity of 14.55 TB per RAID 6 array for the file system to use.
4. The stated streaming rates are least upper bounds (LUB); these rates are based on GPFS/DCS3700 benchmarks using 60 x 7200 RPM Near Line SAS disks. Extrapolating from other tests, greater LUB rates may be expected (e.g., write < 1.7 GB/s and read < 2.4 GB/s using at least 80 of these disks).
5. These rates are extrapolated from actual tests using 15000 RPM disk assuming seek rates on 7200 RPM disk < 33% of 15000 RPM disk. These tests assume completely random 4K transactions (n.b., no locality) to raw devices (n.b., no file system). Instrumented code accessing random 4K files will measure a lower IOP rate since they can not measure the necessary metadata transactions. Favorable locality will increase these rates significantly.

COMMENT: Maintaining good streaming performance requires careful attention being given to balance. Alterations disrupting balance (e.g., inconsistent number of disks or expansion trays per DCS3700) will compromise performance.

Building Block #3A: Logical View



NSD Server (x3550 M3)

- 8 cores (2 sockets @ 4 cores/socket)
- 24 GB RAM (6 DIMMs @ 4 GB/DIMM)
- 2 quad port 6 Gb/s SAS adapters
- IB QDR

DS3524 Turbo (dual controller)

- 2 SAS ports per controller

Disk per DS3524

- 24 x 300 GB SAS disks @ 15000 RPM
- 6 x 2+2 RAID 10 Arrays
- Capacity: raw \approx 7.2 TB, usable < 3.3 TB

Expected Disk Performance per DS3524

- Streaming write rate¹ < 500 MB/s
- Streaming read rate¹ < 800 MB/s
- IOP write rate:² 3000 to 4500 IOP/s
- IOP read rate:² 4500 to 10,000 IOP/s

FOOTNOTES: Data rates are based on theoretical calculations for a GPFS file system spanning 24 disks in a single DS3524 configured as described using -j scatter. **Validation testing is recommended.**

1. Assumes sequential access pattern measured by well written instrumented code.

2. Assumes 4K "to media" transactions measured by the controller. The lower bound assumes random 4K transactions while the upper bound assumes good locality. These rates include both GPFS data and metadata transactions. Instrumented code not measuring metadata transactions will measure lower IOP rates.

Other supported disk choices:

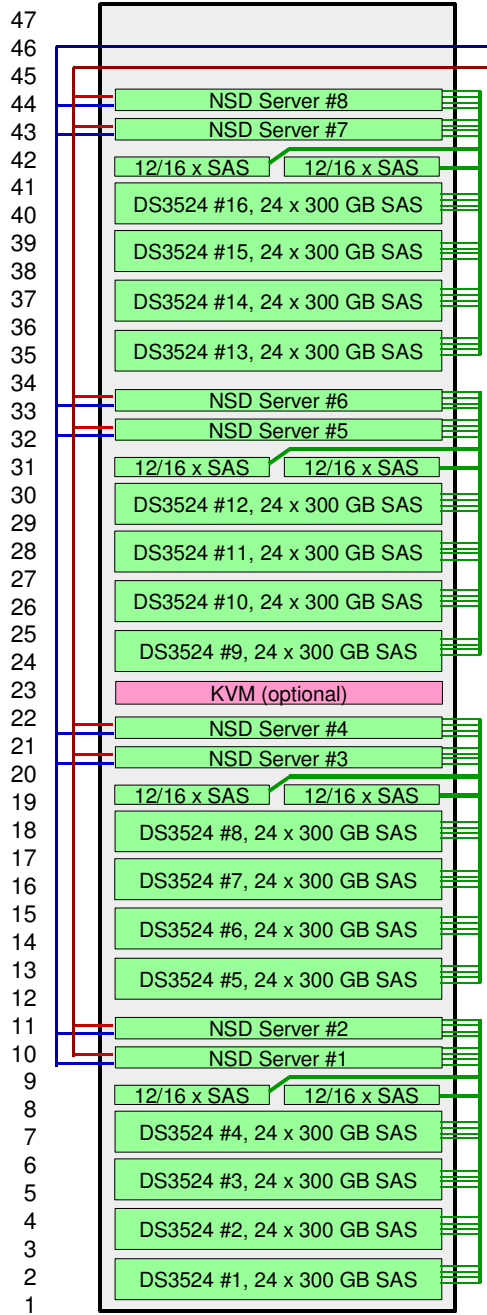
600 GB x 2.5" 10,000 RPM SAS

- IOP rate may be slightly less than for 15,000 RPM disks since its average seek time is slightly greater (n.b., 3 milliseconds vs. 2 milliseconds)

400 GB x 2.5" SSD

- While its seek time is much less, its robustness is not as good as spinning media, and its much more expensive.

Building Block #3A: Physical View

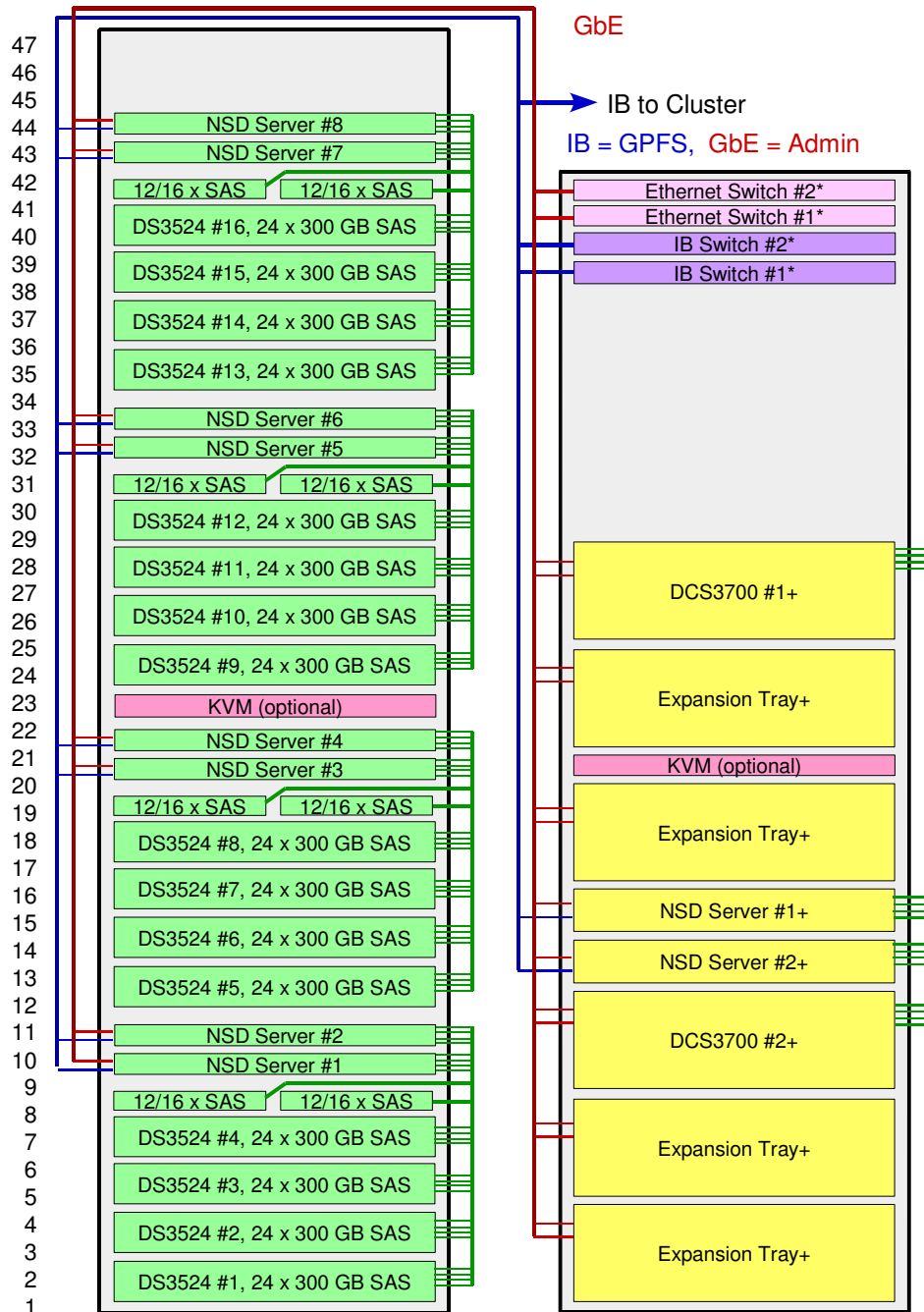


IB QDR = GPFS LAN RDMA (verbs)
GbE = Administration
SAS = Couplet drive-side connections

IOP Optimized Storage

- ◆ 4 Building Blocks
- ◆ Aggregate Statistics
Capacity: raw = 28 TB, usable < 13 TB
Streaming
write < 8 GB/s
read < 13 GB/s
- IOP rate
write: 48,000 to 72,000 IOP/s,
read < 72,000 to 160,000 IOP/s

Two-Tier Solution: Fast Disk, Capacity Disk



Tier #1 – IOP Optimized Storage

- ◆ Building Block #3A
- ◆ 4 x Building Blocks
- ◆ Aggregate Statistics
- Capacity: raw = 28 TB, usable < 13 TB
- Streaming
 - write < 8 GB/s
 - read < 13 GB/s
- IOP rate
 - write: 48,000 to 72,000 IOP/s,
 - read < 72,000 to 160,000 IOP/s

Tier #2 – Capacity Optimized Storage

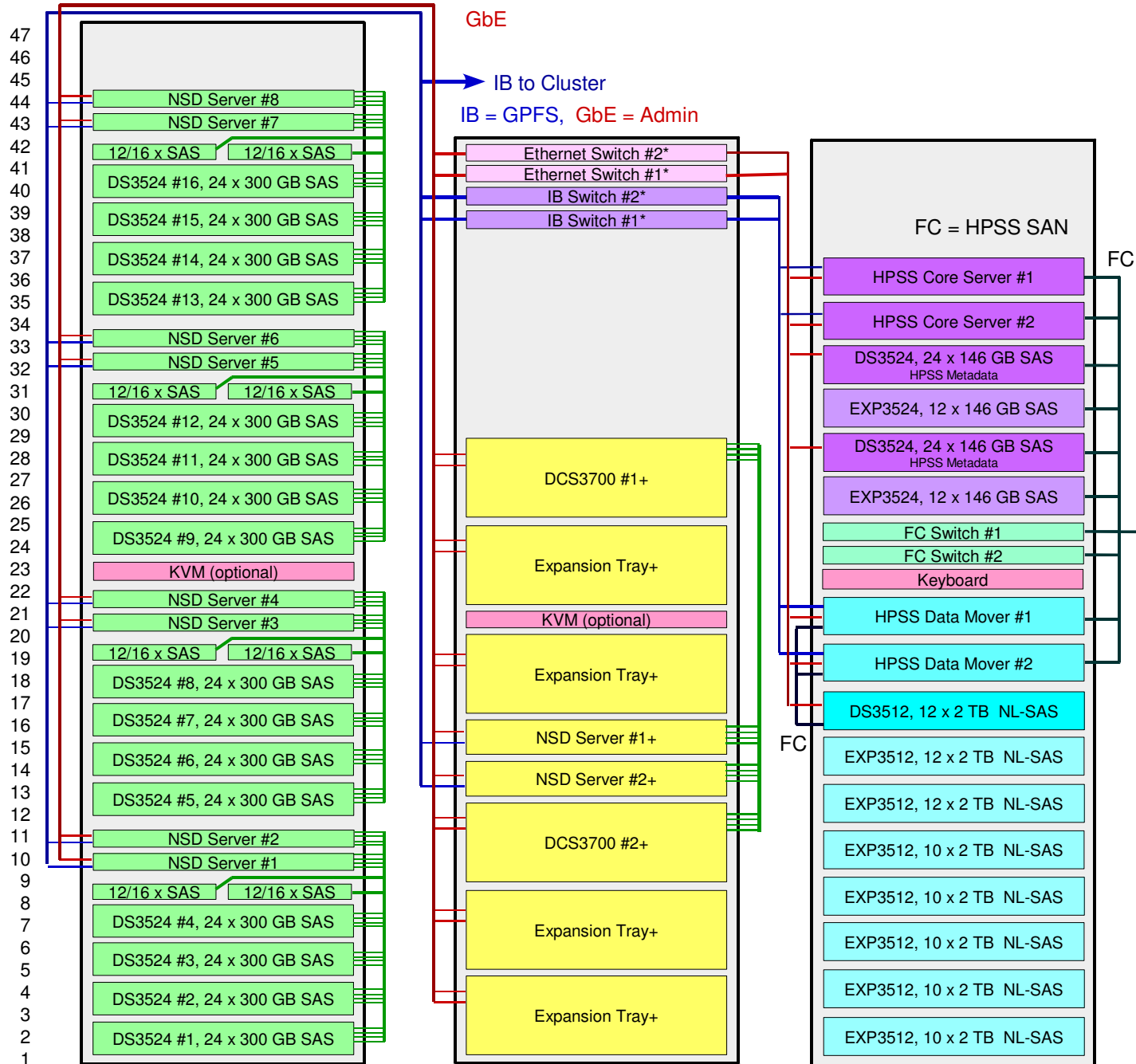
- ◆ Variation of Building Block #1A
- ◆ 1 x Building Block
- ◆ Aggregate Statistics
- Capacity: raw = 720 TB, usable < 524 TB
- Streaming
 - write < 3.2 GB/s
 - read < 4 GB/s

COMMENTS:

The general idea behind this solution is to provide a tier of storage supporting high transaction rates combined with a second tier of cost effective storage. The GPFS file system provides a “policy engine” that manages these 2 tiers of storage.

A 47u rack is recommended for Tier #1 as it can hold 4 building blocks. But if this frame is infeasible, a 42u frame easily be used instead holding 3 building blocks. This solution also requires SAS switches, but these are not available from IBM. If this solution is adopted, the LSI SAS6160 is recommended.

Three-Tier Solution: Fast Disk, Capacity Disk, Tape



Tier 1 – 15000 RPM Disk
 Build Block #3A
 Usable capacity < 13 TB
 Streaming write < 8 GB/s
 Streaming read < 13 GB/s
 IOP write: 48,000 to 72,000 IOP/s
 IOP read < 72,000 to 160,000 IOP/s

Tier 2 – 7200 RPM Disk
 Variation of Building Block #1A
 Usable capacity < 0.5 PB
 Streaming write < 3.2 GB/s
 Streaming read < 4.0 GB/s

Tier 3 – LTO5
 Usable capacity < 1.5 PB
 - 1000 cartridges
 Write < 2.0 GB/s
 Read: TBD

FC cables to tape drives.
 2 Options
 a. 3 x LTO5 < 336 MB/s
 b. 5 x LTO5 < 560 MB/s
 Assumes **un**compressed rates.

HPSS Manages the tape tier and integrates with GPFS ILM

HPSS Data Movers manage the HPSS disk cache and tape drives. The DS3512 storage is used for tape caching and storing small files while tape is used to store large files.

HPSS Core servers manage HPSS Metadata is stored on a DS3524