

# GPFS: Building Blocks and Storage Tiers

Tutorial  
28th IEEE Conference on Massive Data Storage



"A supercomputer is a device for turning compute-bound problems into I/O-bound problems."

Ken Batcher

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

**Version 1.0c**  
**16 April 2012**

## Tutorial Outline

---

1. What is GPFS?
2. Building Block Architecture
3. Storage Tiers

# What Is GPFS?

## GPFS = General Parallel File System

GPFS GA date = 1998

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

Shared disk:

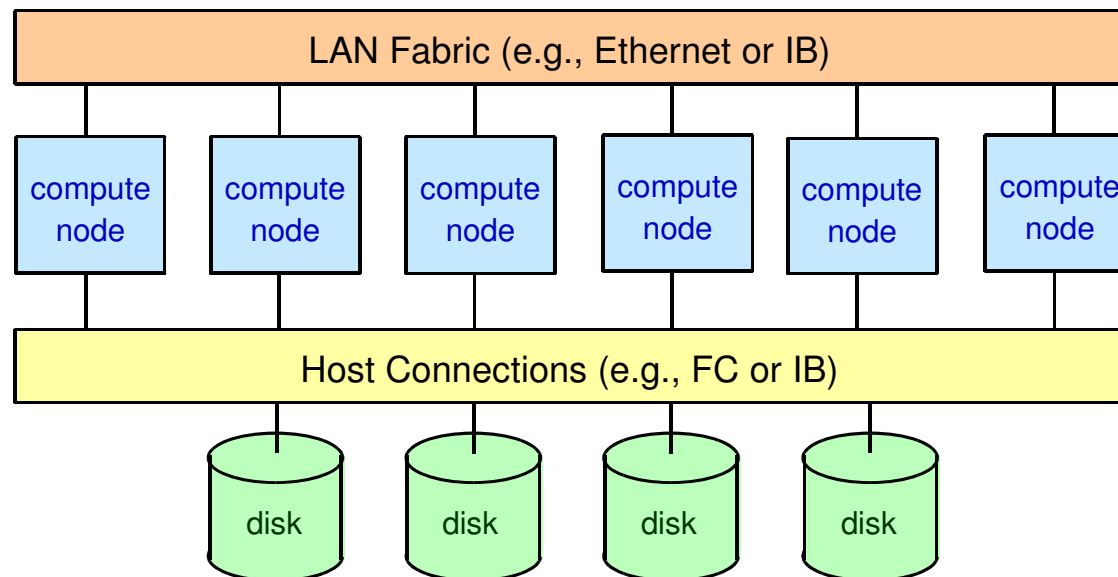
All userdata and metadata are accessible from any disk to any node

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

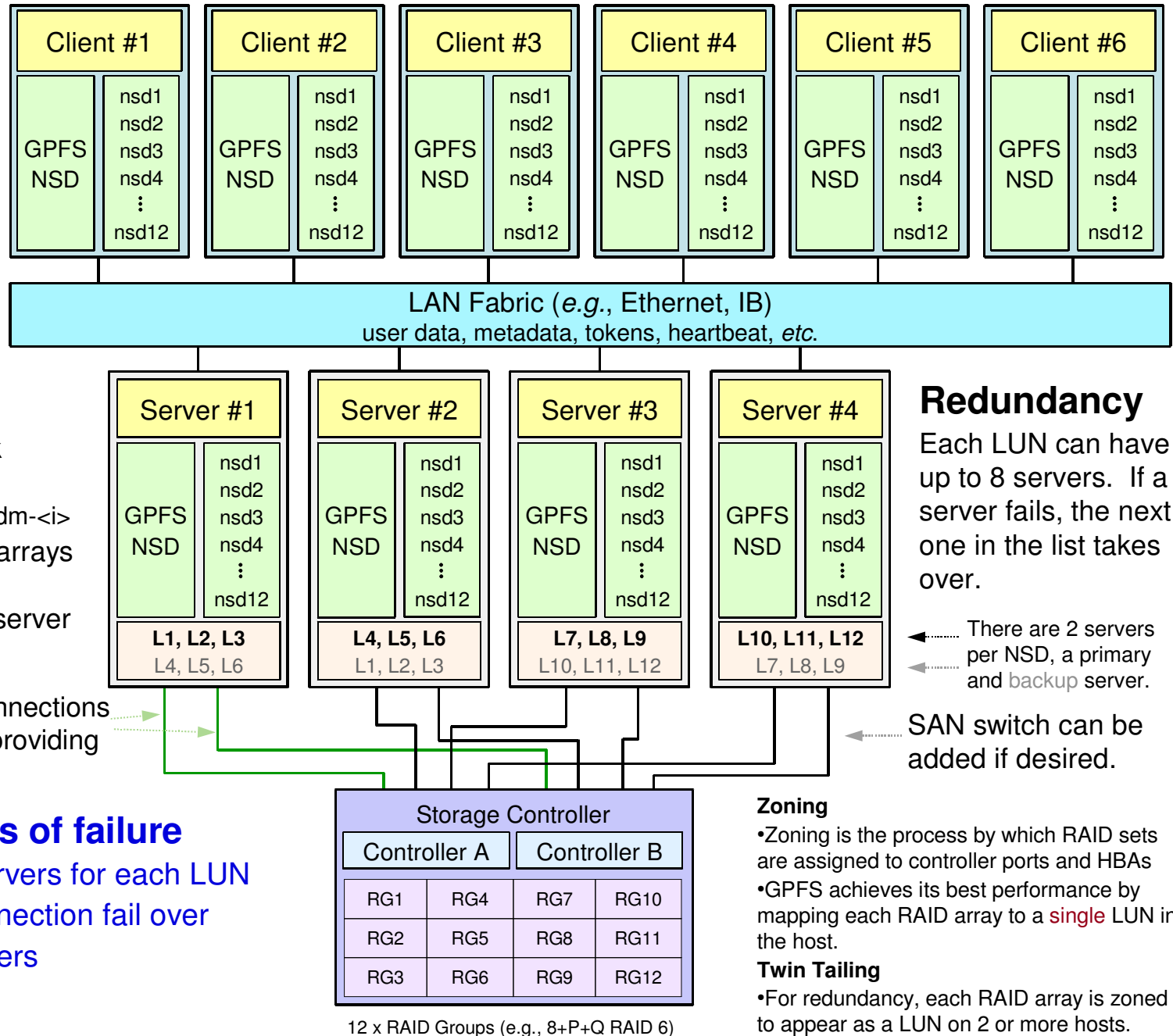
- Logical Unit
- Abstraction of a disk
  - AIX - hdisk<i></i>
  - Linux – sd<c> or dm-<i></i>
- LUNs map to RAID arrays in a disk controller or "physical disks" in a server

### 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

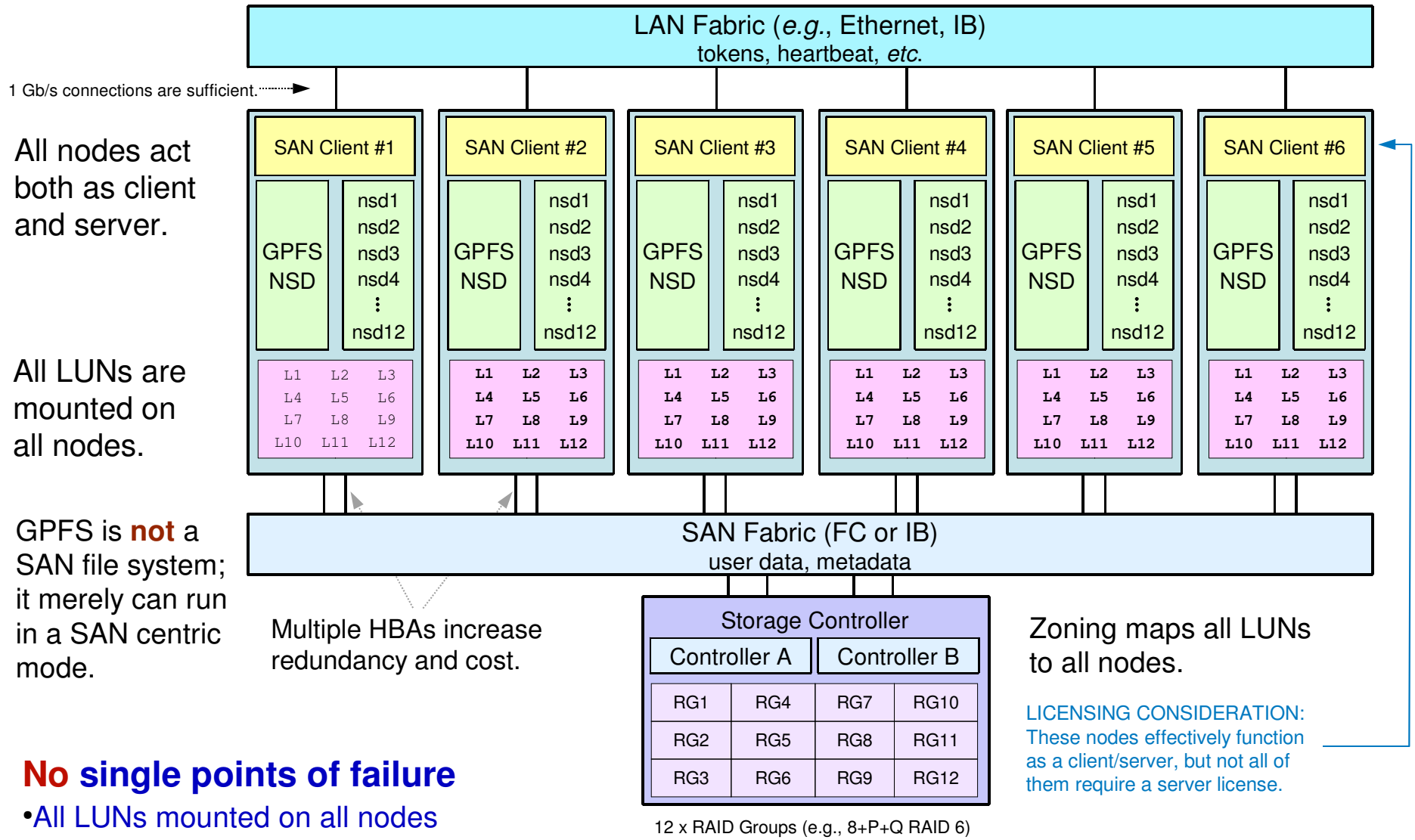
- For redundancy, each RAID array is zoned to appear as a LUN on 2 or more hosts.

12 x RAID Groups (e.g., 8+P+Q RAID 6)



# Storage Area Network (SAN) Topology

## Client/Servers Access Disk via the SAN



All nodes act both as client and server.

All LUNs are mounted on all nodes.

GPFS is **not** a SAN file system; it merely can run in a SAN centric mode.

Multiple HBAs increase redundancy and cost.

Zoning maps all LUNs to all nodes.

**LICENSING CONSIDERATION:** These nodes effectively function as a client/server, but not all of them require a server license.

### 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.,  $\geq 64$  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

---

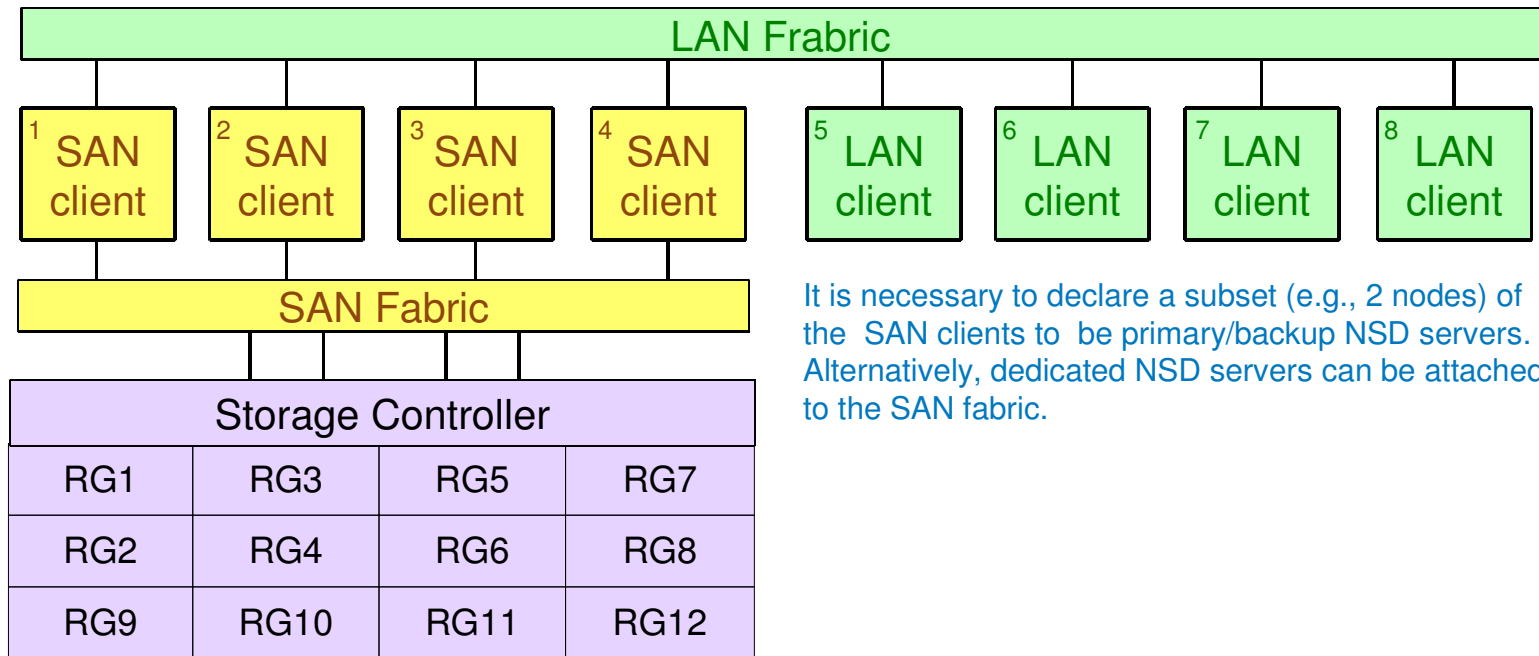
## • LAN Topology

- ◆ 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:

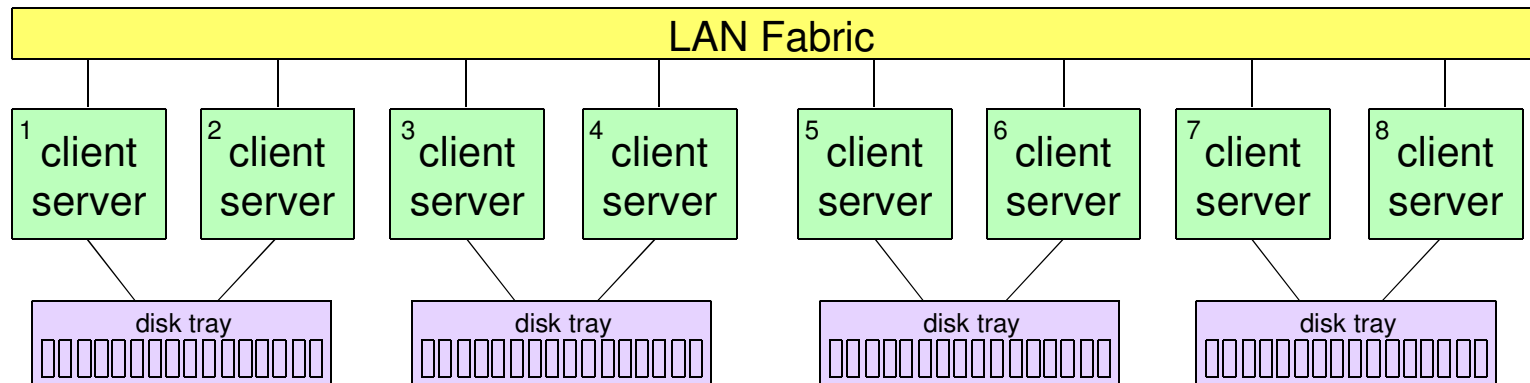
Nodes 1 - 4 (*i.e.*, SAN clients)

- 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

## **Which Organization is Best?**

---

**Its application/customer dependent!**

Each configuration has its limitations and its strong points.

# Designing a Storage System

You've got a problem!

You have requirements...

- Data rate
- Data capacity
- Disk technology
- LAN
- Servers
- Clients
- Cost

Now you need a storage strategy to put them into a solution!



## Strategy: Storage Building Block

---

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: Small vs. Large Building Blocks

## Small

2 x servers  
1 x DCS3700  
1 x Expansion tray  
180 x 2 TB disks  
Rate < 2.4 GB/s  
Capacity  $\approx$  360 TB



## Large

4 x servers  
1 x SFA10K  
20 x Expansion trays  
1200 x 2 TB disks  
Rate < 11 GB/s  
Capacity  $\approx$  2400 TB



Individual storage building blocks can be small or large offering varying degrees of

- cost of entry
- performance:capacity ratios
- flexible growth
- management complexity



## Strategy: Balance

---

***Ideally, an I/O subsystem should be balanced.*** There is no point in making one component of an I/O subsystem fast while another is slow. Moreover, overtaxing some components of the I/O subsystem may disproportionately degrade performance.

However, this goal cannot always be perfectly achieved. A common imbalance is when capacity takes precedence over bandwidth; then the aggregate bandwidth based on the number of disks may exceed the aggregate bandwidth supported by the controllers and/or the number of storage servers.

Performance is inversely  
proportional to capacity.  
- Todd Virnoche

# Strategy: Balance

---

***Ideally, an I/O subsystem should be balanced.*** There is no point in making one component of an I/O subsystem fast while another is slow. Moreover, overtaxing some components of the I/O subsystem may disproportionately degrade performance.

## Various Balance Strategies

1. Solutions *maximizing capacity* balance the number of servers and network adapters with the number of controllers, but use a large number of high capacity disks; the potential bandwidth of the disks exceeds the bandwidth of their managing controller.  
**Low** performance:capacity ratio
2. Solutions *maximizing performance* balance the number of servers and network adapters with the number of controllers, but use a smaller number of faster disks; the potential bandwidth of the disks matches the bandwidth of their managing controller.  
**High** performance:capacity ratio
3. Solutions providing *balanced performance/capacity* balance the number of servers and network adapters with the number of controllers, but use a smaller number of high capacity disks; the potential bandwidth of the disks matches the bandwidth of their managing controller, but the capacity is higher.  
**Moderate** performance:capacity ratio

# Measuring Performance: Storage Access

---

## Streaming

- ♦ records are accessed once and not needed again
- ♦ generally the file size is quite large (e.g., GB or more)
- ♦ good spatial locality occurs if records are adjacent
- ♦ performance is measured by BW (e.g., MB/s, GB/s)
- ♦ operation counts are low compared to BW
- ♦ most common in digital media, HPC, scientific/technical applications

## Comment:

Streaming and IOP access patterns are more common in HPC than transaction processing.

## IOP Processing

- ♦ small transactions (e.g., 10's of KB or less)
  - small records irregularly distributed over the seek offset space
  - small files
- ♦ poor spatial locality and often poor temporal locality
- ♦ performance is measured in operation rates<sup>1</sup> (e.g., IOP/s, files/s)
- ♦ operation counts are high compared to BW
- ♦ common examples: bio-informatics, EDA, rendering, home directories

## Transaction Processing

- ♦ small transactions (e.g., 10's of KB or less), but often displaying good temporal locality
  - access efficiency can often be improved by database technology
- ♦ performance is measured in operation rates<sup>1</sup> (e.g., transactions/s)
- ♦ operation counts are high compared to BW
- ♦ common examples: commercial applications

## Footnote:

1. Correlating application transactions (e.g., POSIX calls) to IOPs (controller transactions) is difficult. POSIX calls result in 1 or more userdata and 0 or more metadata transactions scheduled by the file system to the controller. Controller caching semantics may then coalesce these transactions into single IOPs or distribute them across multiple IOPs.

# Measuring Performance and Capacity

---

## Performance Projections

Performance projections are based on HPC I/O benchmark codes<sup>1</sup>; the various systems are tuned according to standard best practice guidelines appropriate for a production configuration.

While these rates are reproducible in a production environment, they will typically be greater than the data rates observed using a mixture of actual application codes running on the same configuration.

## Units

Units for performance and capacity<sup>2</sup> are generally given in units of  $2^n$  with the following prefixes<sup>3</sup>:  
 $K = 2^{10}$ ,  $M = 2^{20}$ ,  $G = 2^{30}$ ,  $T = 2^{40}$ ,  $P = 2^{50}$

### Footnotes:

1. Stated data rates are least upper bounds, generally reproducible within 10% using standard HPC benchmark codes (e.g., gpfssperf, ibm.v4c, IOR, XDD).
2. The unit prefix for raw capacity is ambiguous; it is simply the value assigned by the OEM.
3. Alternatively, the International System of Units (SI) recommends the prefixes Ki, Mi, Gi, Ti, Pi. See <http://physics.nist.gov/cuu/Units/binary.html>

### Performance implications of high capacity disks:

Regarding 7200 RPM disks (n.b., SATA or NL-SAS), given the relatively recent availability of 3 TB disks, capacity calculations are based on the use of 2 TB disks. As 3 TB and larger disks are adopted, thereby lowering the performance:capacity ratio, the author is concerned that the number of disks needed to meet capacity requirements and satisfy cost constraints will make it difficult to meet performance requirements in HPC markets.

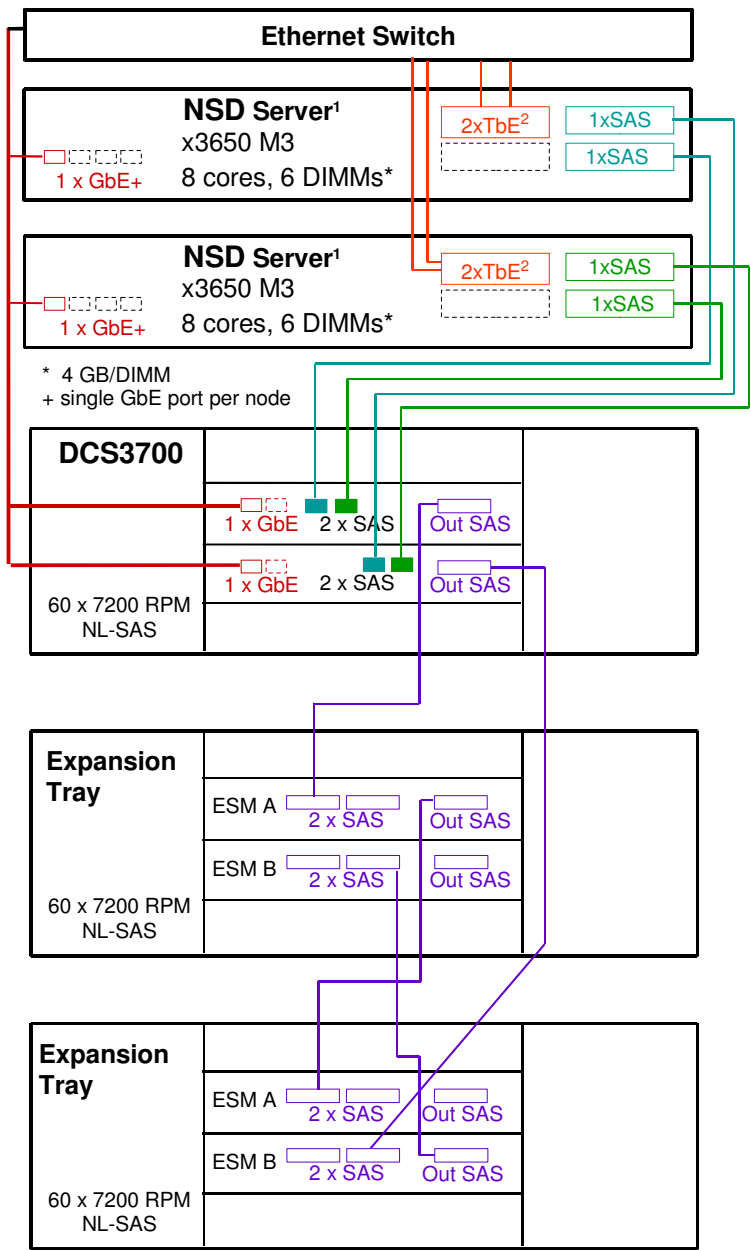
**WARNING:** Your mileage will vary depending on how you drive and maintain your vehicle.

## Maximum Capacity Solutions

---

The following slides demonstrate building block solutions using the supported **maximum** number of **high capacity** drives by the storage controller. The potential streaming performance of this number of drives generally exceeds what the controllers can sustain. This yields the lowest performance to capacity ratio.

# Building Block #1A: Logical View



## Analysis

### NSD Server

- Effective BW per NSD server < 1.4 GB/s
- x3650 M3 with 8 cores and 6 DIMMs (4 GB per DIMM)
- 1 x GbE < 80 MB/s
- 2 x TbE² < 1.4 GB/s
- 2 x single port 6 Gb/s SAS adapters

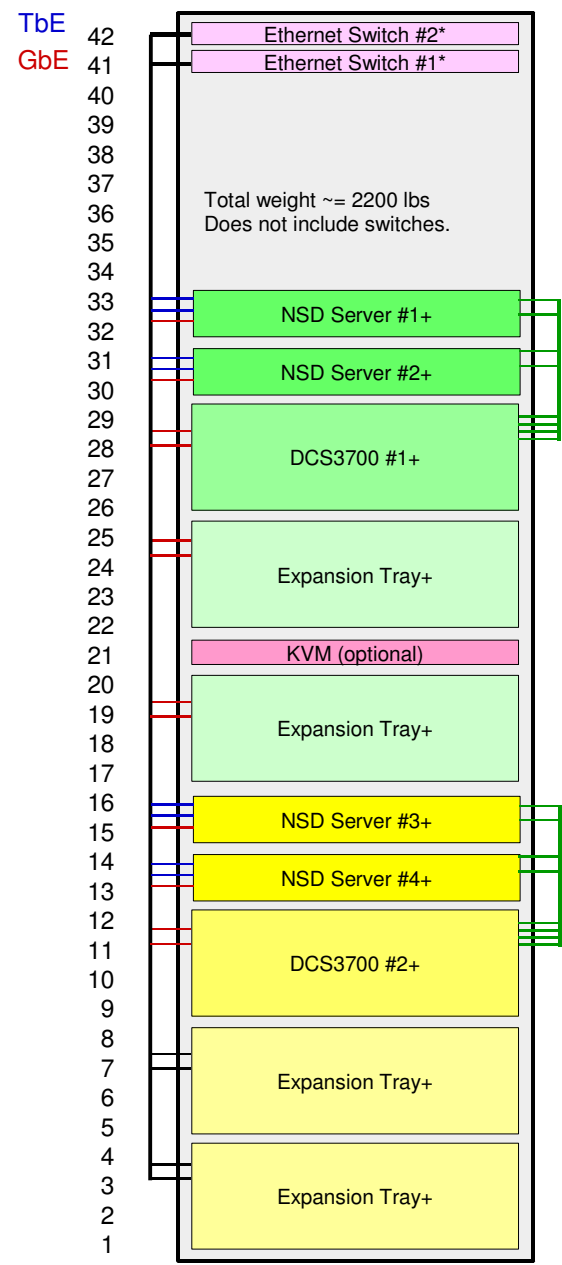
### 1 x DCS3700 Turbo with 2 x EXP3560 trays

- 180 x 2 TB near line SAS disks
- 18 x 8+2P RAID 6 arrays
- Capacity: raw = 360 TB, usable < 262 TB³
- Performance
- Streaming rate: write < 1.6 GB/s⁴, read < 2.0 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 can replace the dual port TbE adapter; performance will not increase.
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.

# Building Block #1A: 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)

2 x DCS3700, each with the following components  
 - 2 x Expansion Tray  
 - 180 x 2 TB, 7200 RPM Near Line SAS disks as 6 x 8+P+Q RAID 6 arrays  
 - Capacity: raw = 360 TB, usable < 262 TB  
 - 4 x SAS host ports @ 6 Gb/s; n.b., 2 SAS host ports per RAID controller

Switches: Provide Ethernet and IB switches as needed.

Comment: This configuration consists of 2 building blocks. Adding additional building blocks will scale performance and capacity linearly.

## AGGREGATE STATISTICS

**Disks**  
 - 360 x 2 TB, 7200 RPM Near Line SAS disks  
 - 36 x 8+P+Q RAID 6 arrays, 1 LUN per array

**Capacity**  
 - raw = 720, usable < 524 TB<sup>1</sup>

**Performance**  
 - streaming rate : write < 3.2 GB/s, read < 4 GB/s  
 - IOP rate (random, 4K transactions): write < 7200 IOP/s, read < 12,000 IOP/s<sup>2</sup>

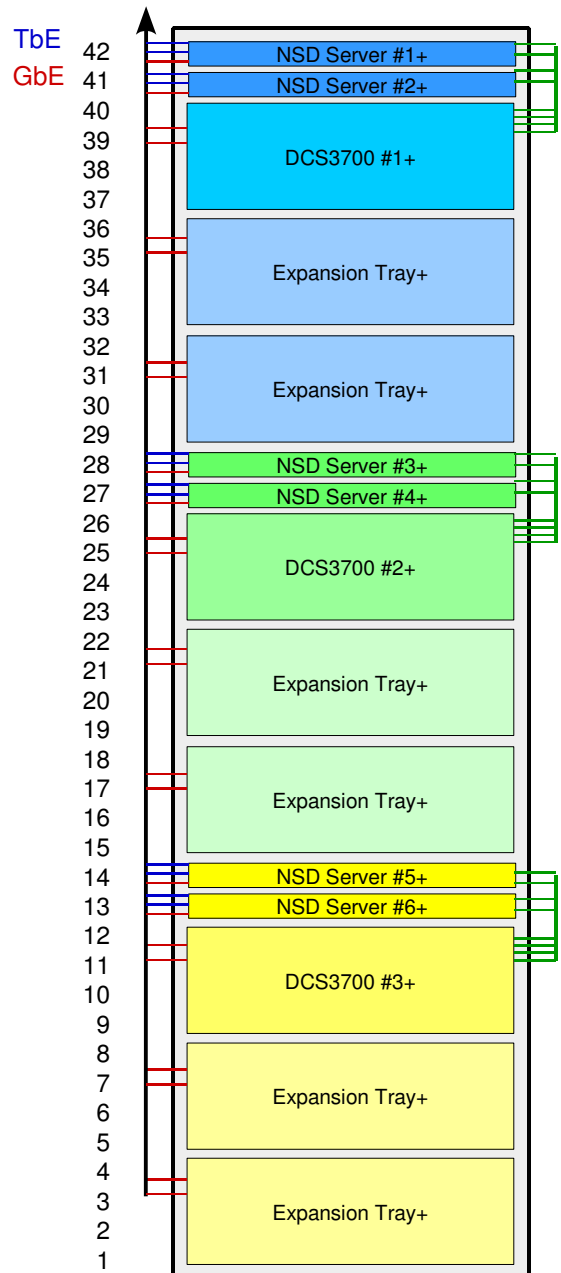
**COMMENT:** Maintaining good streaming performance requires careful attention being given to balance. Alterations disrupting balance (e.g., inconsistent number of disks or expansion trays per DCS3700) will compromise performance.

## FOOTNOTES:

1. Usable capacity is defined as the storage capacity delivered by the controller to the file system. Additionally, file system metadata requires a typically small fraction of the usable capacity. For the GPFS file system, this is *typically* < 1.5% (~= 8 TB in this case). With a very large number (e.g., billions) of very small files (e.g., 4 KB), the metadata capacity *may* be much larger (e.g., > 10%). Metadata overhead in this case is application environment specific and difficult to project.  
 2. These numbers are based purely 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. (n.b., These rates are based on actual tests using 15000 RPM disk assuming seek rates on 7200 RPM disk < 33% of 15000 RPM disk.)

\* These switches are included in this diagram for completeness. If the customer has adequate switch ports, then these switches may not be needed.  
 + Due to SAS cable lengths (3M is recommended) it is necessary to place the NSD servers in the same rack as the controllers.

# Variation on Building Block #1A: Physical View



## COMPONENTS

4 x NSD servers (**x3550 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, **1 x dual port SAS (6 Gb/s)**

3 x DCS3700, each with the following components  
 - 2 x Expansion Tray  
 - 180 x 2 TB, 7200 RPM Near Line SAS disks as 6 x 8+P+Q RAID 6 arrays  
 - Capacity: raw = 360 TB, usable = 262 TB  
 - 4 x SAS host ports @ 6 Gb/s; n.b., 2 SAS host ports per RAID controller

Switches: Provide Ethernet switches as needed.

Comment: This configuration consists of 3 building blocks. Adding additional building blocks will scale performance and capacity linearly.

## AGGREGATE STATISTICS

**Disks**  
 - 540 x 2 TB, 7200 RPM Near Line SAS disks  
 - 54 x 8+P+Q RAID 6 arrays, 1 LUN per array

**Capacity**  
 - raw = 1080, usable = 786 TB<sup>1</sup>

**Performance**  
 - streaming rate : write < 4.8 GB/s, read < 6 GB/s  
 - IOP rate (random, 4K transactions): write < 10,800 IOP/s, read < 18,000 IOP/s<sup>2</sup>

## FOOTNOTES:

1. Usable capacity is defined as the storage capacity delivered by the controller to the file system. Additionally, file system metadata requires a typically small fraction of the usable capacity. For the GPFS file system, this is *typically* < 1.5% (~= 12 TB in this case). With a very large number (e.g., billions) of very small files (e.g., 4 KB), the metadata capacity *may* be much larger (e.g., > 10%). Metadata overhead in this case is application environment specific and difficult to project.
2. These numbers are based purely 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. (n.b., These rates are based on actual tests using 15000 RPM disk assuming seek rates on 7200 RPM disk < 33% of 15000 RPM disk.)

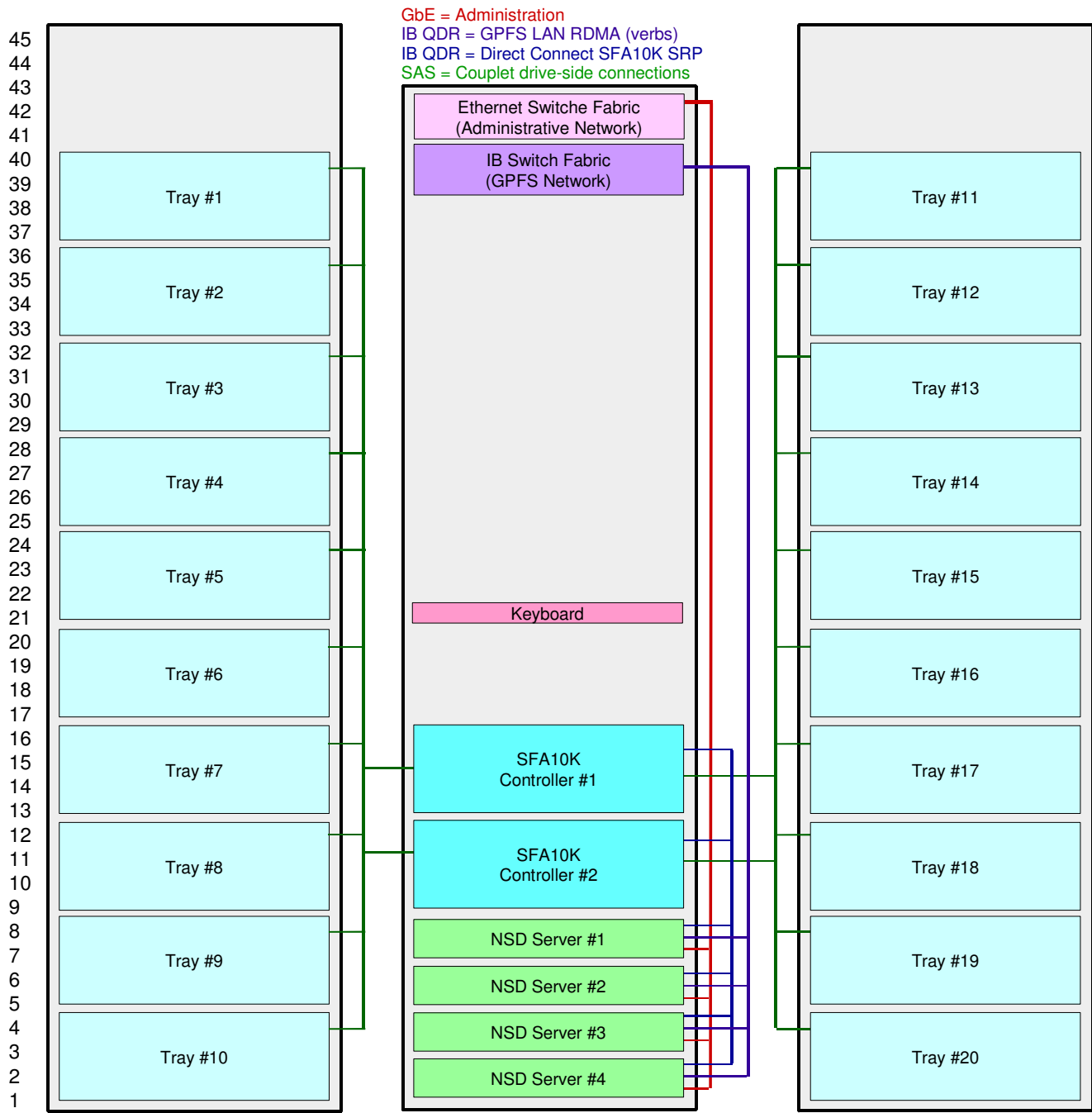
**COMMENT: This solution is similar to the previous one, but uses slightly different components achieve greater rack density.**

**Comment:**  
 Denser servers with fewer PCI-E slots are used in order to increase rack density.

+ Due to SAS cable lengths (3M is recommended) it is necessary to place the NSD servers in the same rack as the controllers.



# Building Block #1B: Physical View



## NSD Servers

- 4 x NSD servers (x3650 M3):
- 2 x quad core westmere sockets
  - 6 x DIMMs (4 GB per DIMM)
  - 1 x IB QDR: GPFS LAN using RDMA (Verbs)
  - 1 x IB QDR: Direct Attached Storage SAN (SRP)
  - 1 x GbE: Administrative LAN

## Storage

- 1 x SFA10K + 10 x Expansion Trays
- 1200 x 2 TB, 7200 RPM SATA disks
- 120 x 8+P+Q RAID 6 pools, 1 LUN per pool
- Capacity: raw 2400 TB, usable < 1800 TB
- Streaming Performance:
  - write < 11 GB/s
  - read < 10 GB/s
- IOP rate: Varies significantly based on locality

# Balanced Performance/Capacity Solutions

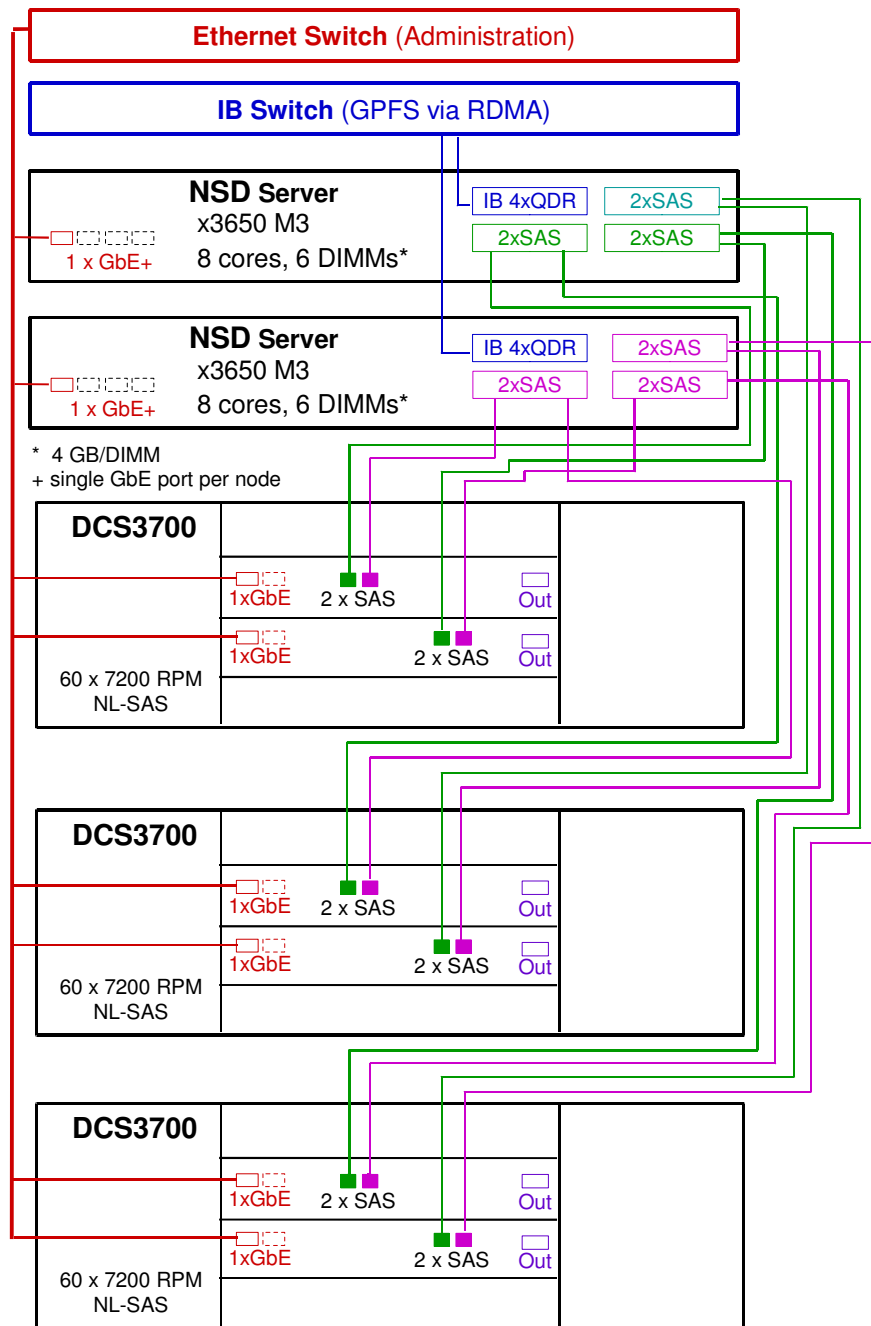
---

The following slides demonstrate building block solutions using the **minimum** number of **high capacity** drives necessary to saturate controller streaming\* performance. This improves the performance to capacity ratio.

**Footnote:**

\* Most storage controllers can sustain higher IOP rates than spinning disk can produce doing small random IOP transactions.

# Building Block #2A: Logical View



## Analysis

NSD = Network Storage Device  
These are the storage servers for GPFS.

### NSD Server

- Effective BW per NSD server < 3 GB/s
- x3650 M3 with 8 cores and 6 DIMMs (4 GB per DIMM)
- 1 x GbE < 80 MB/s
- 1 x IB QDR < 3 GB/s (n.b., using RDMA)
- 2 x dual port 6 Gb/s SAS adapters<sup>1</sup>

### DCS3700

- 60 x 2 TB near line SAS disks
- 6 x 8+P+Q RAID 6 arrays
- Capacity: raw = 120 TB, usable = 87.3 TB<sup>2</sup>
- Performance
- Streaming rate: write < 1.6 GB/s, read < 2.0 GB/s
- IOP rate (random 4K transactions): write < 1200 IOP/s, read < 2000 IOP/s<sup>4</sup>
- IOP rate (mdtest): find mdtest results below<sup>5</sup>

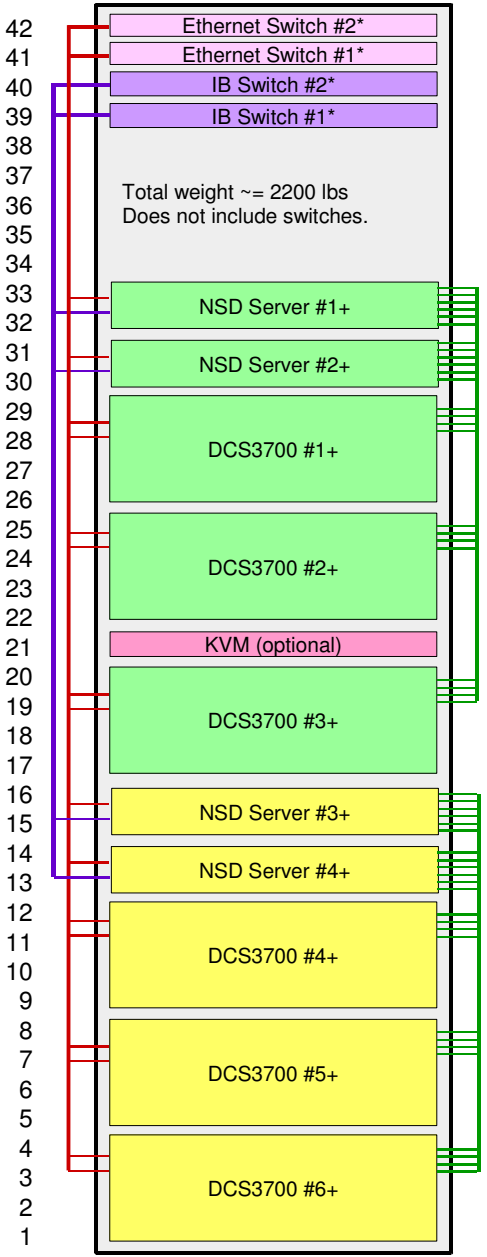
## Aggregate Building Block Statistics

- 2 x NSD servers
- 3 x DCS3700
- 180 x 2 TB near line SAS disks as 18 x 8+P+Q RAID 6 arrays
- Capacity: raw = 360 TB, usable = 262 TB<sup>2</sup>
- Performance
- Streaming rate: write < 4.8 GB/s, read < 5.5 GB/s<sup>3</sup>
- IOP rate (random 4K transactions): write < 3600 IOP/s, read < 6000 IOP/s<sup>4</sup>
- IOP rate (mdtest): scaling tests remain to be completed<sup>5</sup>

### FOOTNOTES:

1. Wire speed for a 1 x 6 Gb/s port < 3 GB/s (n.b., 4 lanes @ 6 Gb/s per lane); with 6 ports per node, the potential SAS aggregate BW is 18 GB/s! The 6 ports are needed for redundancy, not performance.
2. The DCS3700 provides a capacity of 14.55 TB per RAID 6 array for the file system to use. Theoretically, this solution should be able to deliver 6 GB/s; however, this requires pushing performance to IB QDR limit. While this may be feasible, performance expectations are being lowered as a precaution.
3. 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.
4. These limited scale tests are included to show the impact that file system optimization can have on small transaction rates; i.e., the random 4K random transaction test is a worst possible case.

# Building Block #2A: 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, 1 x IB QDR, 2 x dual port SAS (6 Gb/s)

6 x DCS3700, each with the following components  
 - 60 x 2 TB, 7200 RPM Near Line SAS disks as 6 x 8+P+Q RAID 6 arrays  
 - Capacity: raw = 120 TB, usable = 87.3 TB  
 - 4 x SAS host ports @ 6 Gb/s; n.b., 2 SAS host ports per RAID controller

Switches: Provide IB and Ethernet switches as needed.

Comment: This configuration consists of 2 building blocks. Adding additional building blocks will scale performance and capacity linearly.

### AGGREGATE STATISTICS

**Disks**  
 - 360 x 2 TB, 7200 RPM Near Line SAS disks  
 - 36 x 8+P+Q RAID 6 arrays, 1 LUN per array

**Capacity**  
 - raw = 720, usable = 522 TB<sup>1</sup>

**Performance**  
 - streaming rate : write < 9.6 GB/s, read < 11 GB/s<sup>2</sup>  
 - IOP rate (random, 4K transactions): write < 7200 IOP/s, read < 12,000 IOP/s<sup>3</sup>  
 - IOP rate (mdtest): scaling tests remain to be completed<sup>4</sup>

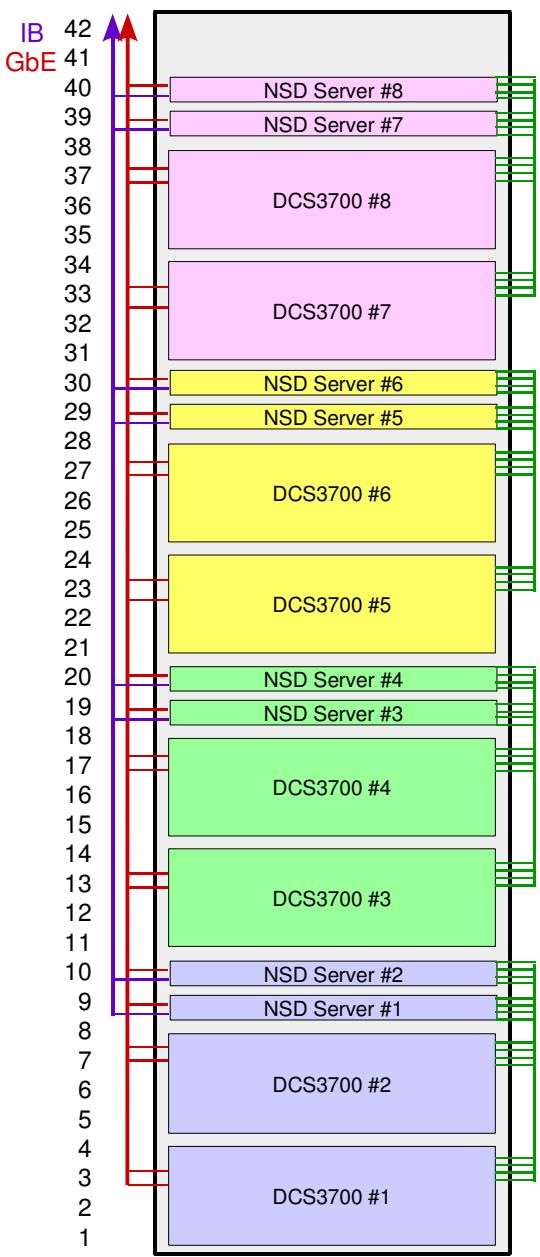
**COMMENT:** Maintaining good streaming performance requires careful attention being given to balance. Alterations disrupting balance (e.g., replacing controllers with expansion trays, or indiscriminately adding expansion trays) will compromise performance.

### FOOTNOTES:

1. Usable capacity is defined as the storage capacity delivered by the controller to the file system. Additionally, file system metadata requires a typically small fraction of the usable capacity. For the GPFS file system, this is typically < 1.5% (~= 8 TB in this case). With a very large number (e.g., billions) of very small files (e.g., 4 KB), the metadata capacity may be much larger (e.g., > 10%). Metadata overhead in this case is application environment specific and difficult to project.
2. Theoretically, this solution should be able to deliver 12 GB/s; however, this requires pushing performance to IB QDR limit. While this may be feasible, performance expectations are being lowered as a precaution.
3. These numbers are based purely 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. (n.b., These rates are based on actual tests using 15000 RPM disk assuming seek rates on 7200 RPM disk < 33% of 15000 RPM disk.)
4. Limited scale mdtest results are included on the previous page to show the impact that file system optimization can have on small transaction rates; i.e., the random 4K random transaction test is a worst possible case.

\* These switches are included in this diagram for completeness. If the customer has adequate switch ports, then these switches may not be needed.  
 + Due to SAS cable lengths (3M is recommended) it is necessary to place the NSD servers in the same rack as the controllers.

# Variation on Building Block #2A: Physical View



## COMPONENTS

8 x NSD servers (x3550 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, 1 x IB QDR or 1 x dual port TbE, 1 x quad port SAS (6 Gb/s)

8 x DCS3700, each with the following components  
 - 60 x 2 TB, 7200 RPM Near Line SAS disks as 6 x 8+P+Q RAID 6 arrays  
 - Capacity: raw = 120 TB, usable = 87.3 TB  
 - 4 x SAS host ports @ 6 Gb/s; n.b., 2 SAS host ports per RAID controller

Switches: Provided externally

Comment: This configuration consists of 4 building blocks. Adding additional building blocks will scale performance and capacity linearly.

## AGGREGATE STATISTICS

**Disks**  
 - 480 x 2 TB, 7200 RPM Near Line SAS disks  
 - 48 x 8+P+Q RAID 6 arrays, 1 LUN per array

**Capacity**  
 - raw = 960, usable = 698.4 TB<sup>1</sup>

**Performance using IB QDR<sup>2</sup>**  
 - streaming rate : write < 12.8 GB/s<sup>2</sup>, read < 16 GB/s<sup>2</sup>  
 - IOP rate (random, 4K transactions): write < 9600 IOP/s, read < 16,000 IOP/s<sup>3</sup>  
 - IOP rate (mdtest): scaling tests remain to be completed<sup>4</sup>

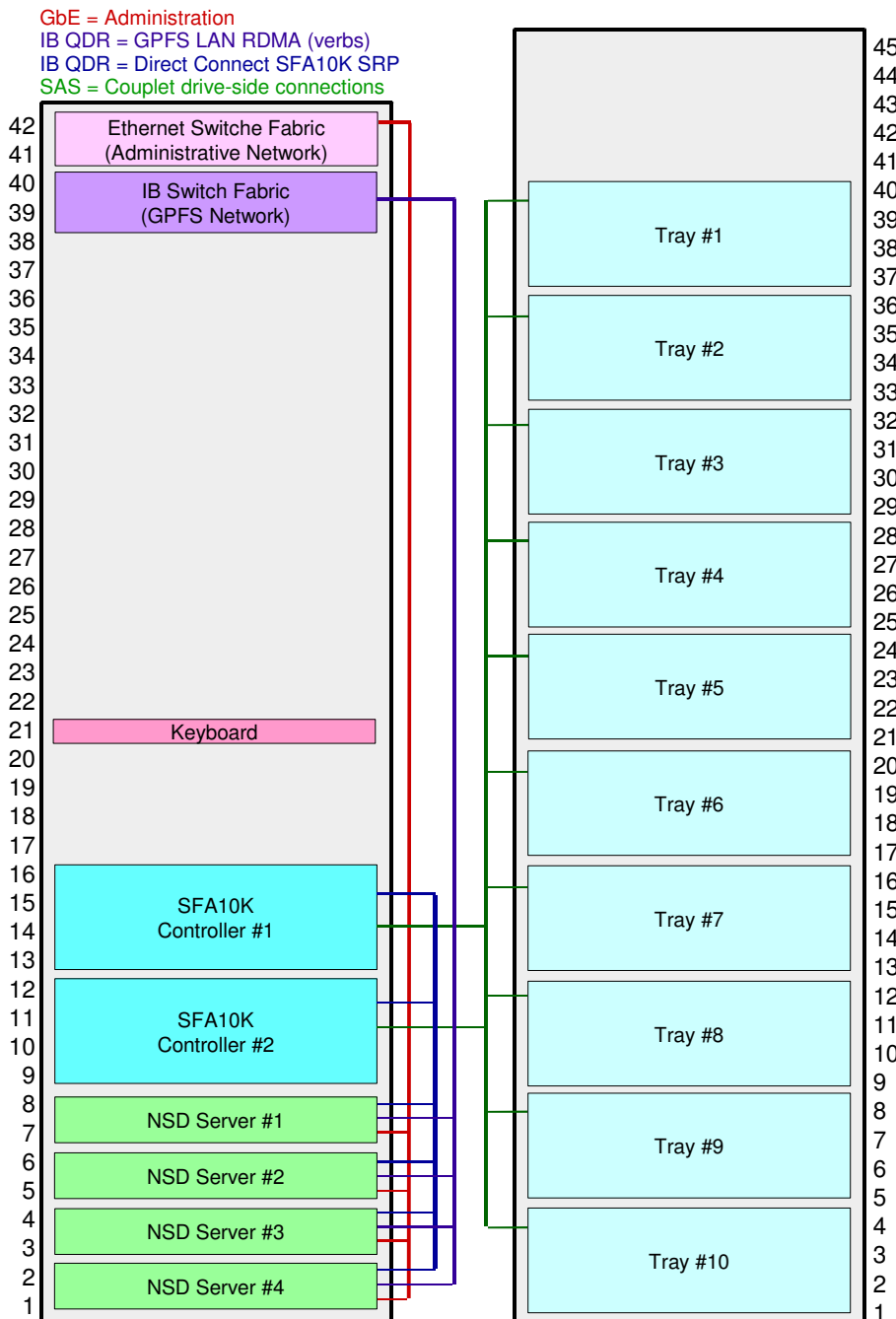
## FOOTNOTES:

1. Usable capacity is defined as the storage capacity delivered by the controller to the file system. Additionally, file system metadata requires a typically small fraction of the usable capacity. For the GPFS file system, this is *typically* < 1.5% (~= 10.5 TB in this case). With a very large number (e.g., billions) of very small files (e.g., 4 KB), the metadata capacity *may* be much larger (e.g., > 10%). Metadata overhead in this case is application environment specific and difficult to project.
2. If the IB QDR HCAs are replaced with 2xTbE adapters, aggregate streaming are: write < 11 GB/s, read < 11 GB/s. IOP rates should not be impacted by the choice of LAN adapter.
3. These numbers are based purely 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. (n.b., These rates are based on actual tests using 15000 RPM disk assuming seek rates on 7200 RPM disk < 33% of 15000 RPM disk.)
4. Limited scale mdtest results are included below to show the impact that file system optimization can have on small transaction rates; i.e., the random 4K random transaction test is a worst possible case.

**Comment:**  
 Denser servers with fewer PCI-E slots are used in order to increase rack density.

**COMMENT:** Maintaining good streaming performance requires careful attention being given to balance. Alterations disrupting balance (e.g., replacing controllers with expansion trays, or indiscriminately adding expansion trays) will compromise performance.

# Building Block #2B: Physical View



GbE = Administration  
 IB QDR = GPFS LAN RDMA (verbs)  
 IB QDR = Direct Connect SFA10K SRP  
 SAS = Couplet drive-side connections

## NSD Servers

- 4 x NSD servers (x3650 M3):
- 2 x quad core westmere sockets
  - 6 x DIMMs (4 GB per DIMM)
  - 1 x IB QDR: GPFS LAN using RDMA (Verbs)
  - 1 x IB QDR: Direct Attached Storage SAN (SRP)
  - 1 x GbE: Administrative LAN

## Storage

- 1 x SFA10K and 10 x Expansion Trays
- 520 x 2 TB SATA, 7200 RPM disks
- 52 x 8+P+Q RAID 6 pools, 1 LUN per pool (dataOnly LUNs)
- 80 x 400 GB SSD
- 40 x 1+1 RAID 1 pools, 1 LUN per pool (metadataOnly LUNs)
- Capacity: raw 1040 TB, usable < 780 TB
- Performance **Estimate**<sup>1</sup>
- Streaming rate: write < 11 GB/s, read < 10 GB/s
- IOP rate: Varies significantly based on locality.

## FOOTNOTES:

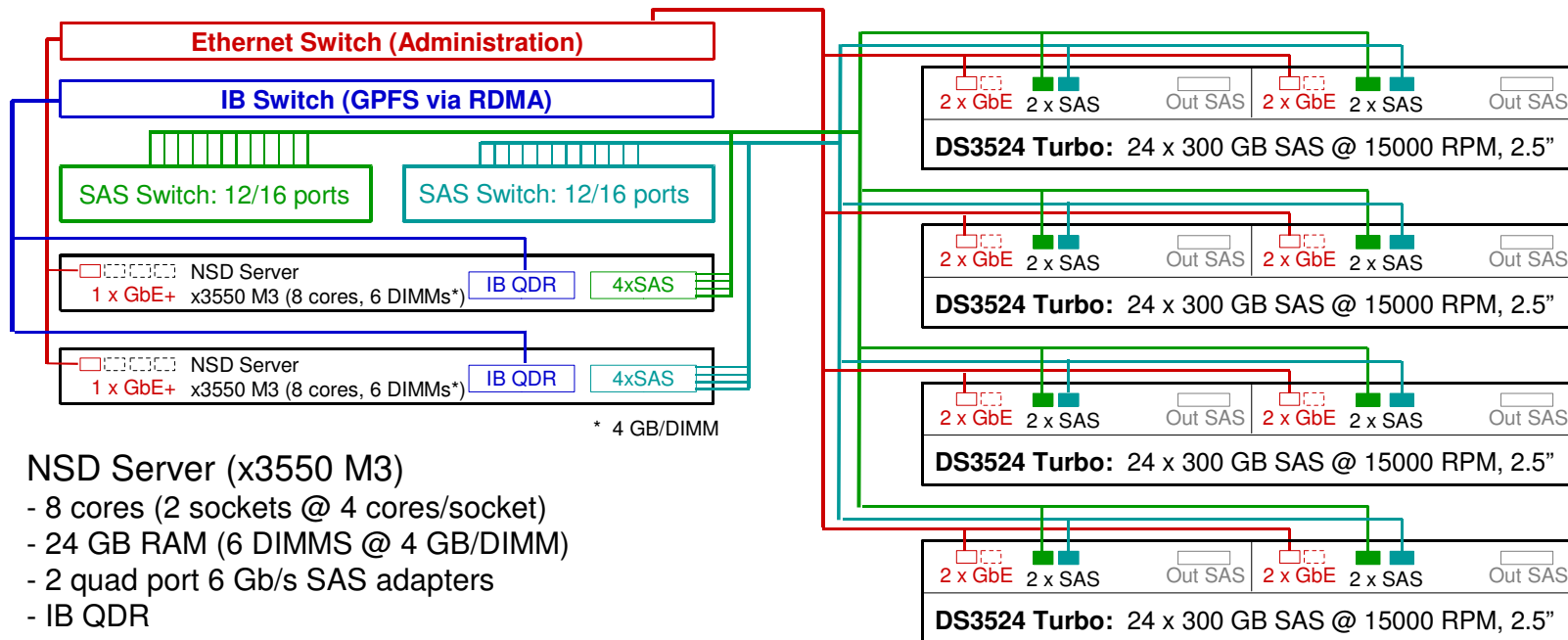
1. Benchmarks based on 240 x SATA (7200 RPM), but no SSD:  
 write < 4600 MB/s, read < 6000 MB/s.  
 Benchmarks based on 290 x SAS (15000 RPM) + 10 x SSD:  
 write < 11000 MB/s, read < 10,000 MB/s.  
 Without the use of SSD, the SAS write performance was 40% less.

## Maximum Performance Solutions

---

The following slides demonstrate building block solutions using a **minimum** number of **high performance** drives. Since 7200 RPM drives generally yield adequate streaming performance for the HPC market, the goal with these solutions is to improve IOP performance (though streaming performance is generally optimized as well). This yields the highest performance to capacity ratio for both streaming and IOP 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 rate<sup>1</sup> < 500 MB/s
- Streaming read rate<sup>1</sup> < 800 MB/s
- IOP write rate:<sup>2</sup> 3000 to 4500 IOP/s
- IOP read rate:<sup>2</sup> 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

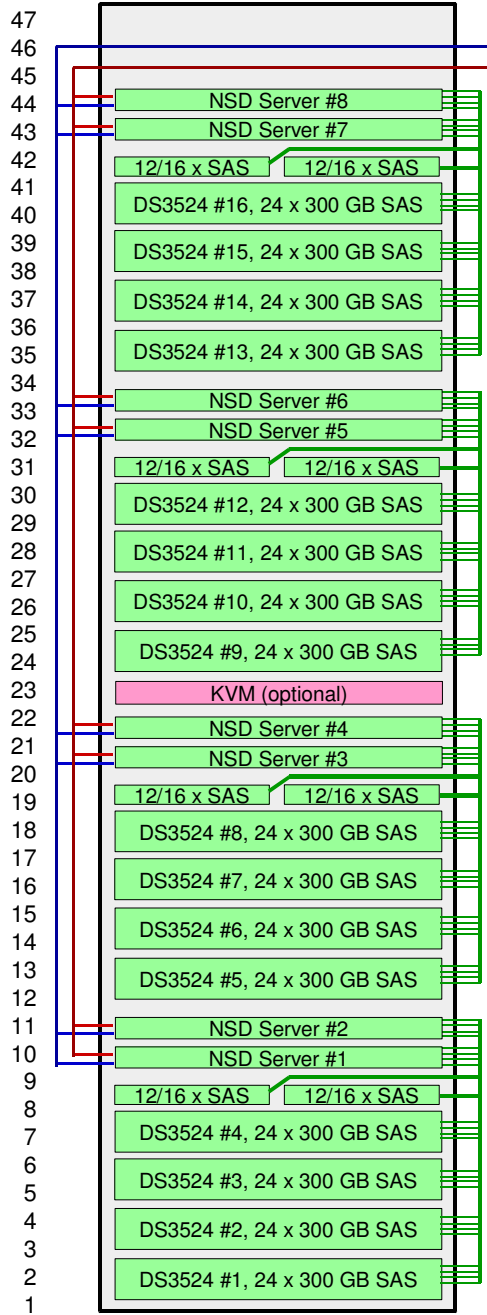
- IOP rate may be slightly less than for 15,000 RPM disks since its average seek time is slightly greater (n.b., 3 milliseconds vs. 2 milliseconds)

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

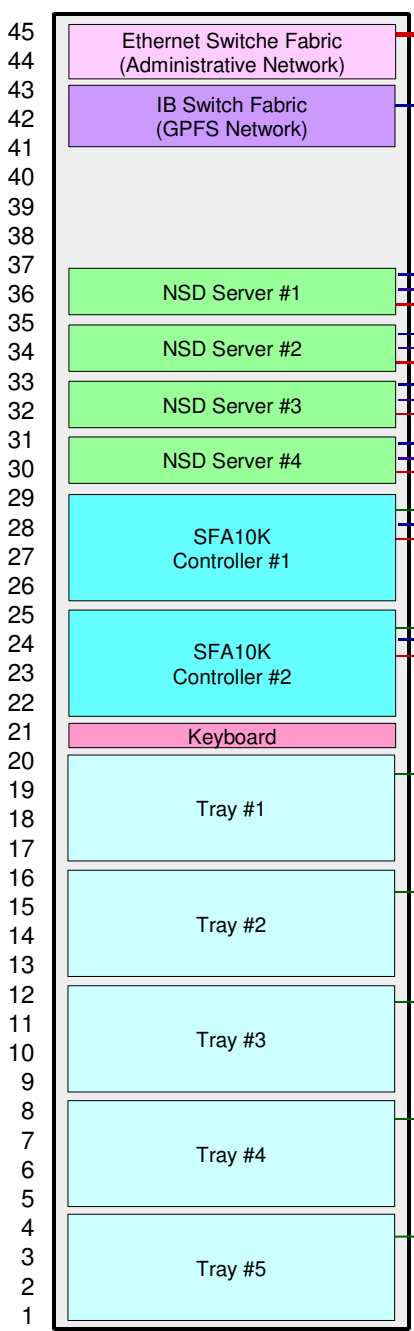


IB QDR = GPFS LAN RDMA (verbs)  
GbE = Administration  
SAS = Couplet drive-side connections

## 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

# Building Block #3B: Physical View



GbE = Administration  
 IB QDR = GPFS LAN RDMA (verbs)  
 IB QDR = Direct Connect SFA10K SRP  
 SAS = Couplet drive-side connections

### NSD Servers

- 4 x NSD servers (x3650 M3):
- 2 x quad core westmere sockets
  - 6 x DIMMs (4 GB per DIMM)
  - 1 x IB QDR: GPFS LAN using RDMA (Verbs)
  - 1 x IB QDR: Direct Attached Storage SAN (SRP)
  - 1 x GbE: Administrative LAN

### Storage

- 1 x SFA10K and 5 x Expansion Trays
- 280 x 600 GB, 15000 RPM SAS disks
- 28 x 8+P+Q RAID 6 pools, 1 LUN per pool (dataOnly LUNs)
- 20<sup>1</sup> x 400 GB SSD (metadataOnly LUNs)
- 10 x 1+1 RAID 1 pools
- Capacity: raw 168 TB, usable < 126 TB
- Performance
- Streaming rate: write < 11 GB/s, read < 10 GB/s
- IOP rate: Varies significantly based on locality.

### FOOTNOTES:

1. If this configuration is adopted for a "many small files" workload where the SSD is used as a metadata store, then there is an inadequate amount of SSD to hold it all. One way to manage this would be to replace spinning disk with SSD (e.g., 200 x 600 GB, 15000 RPM disks + 100 x 400 GB SSD). This may lower streaming performance. Validation testing is required.

# Strategy: Storage Tiers

---

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

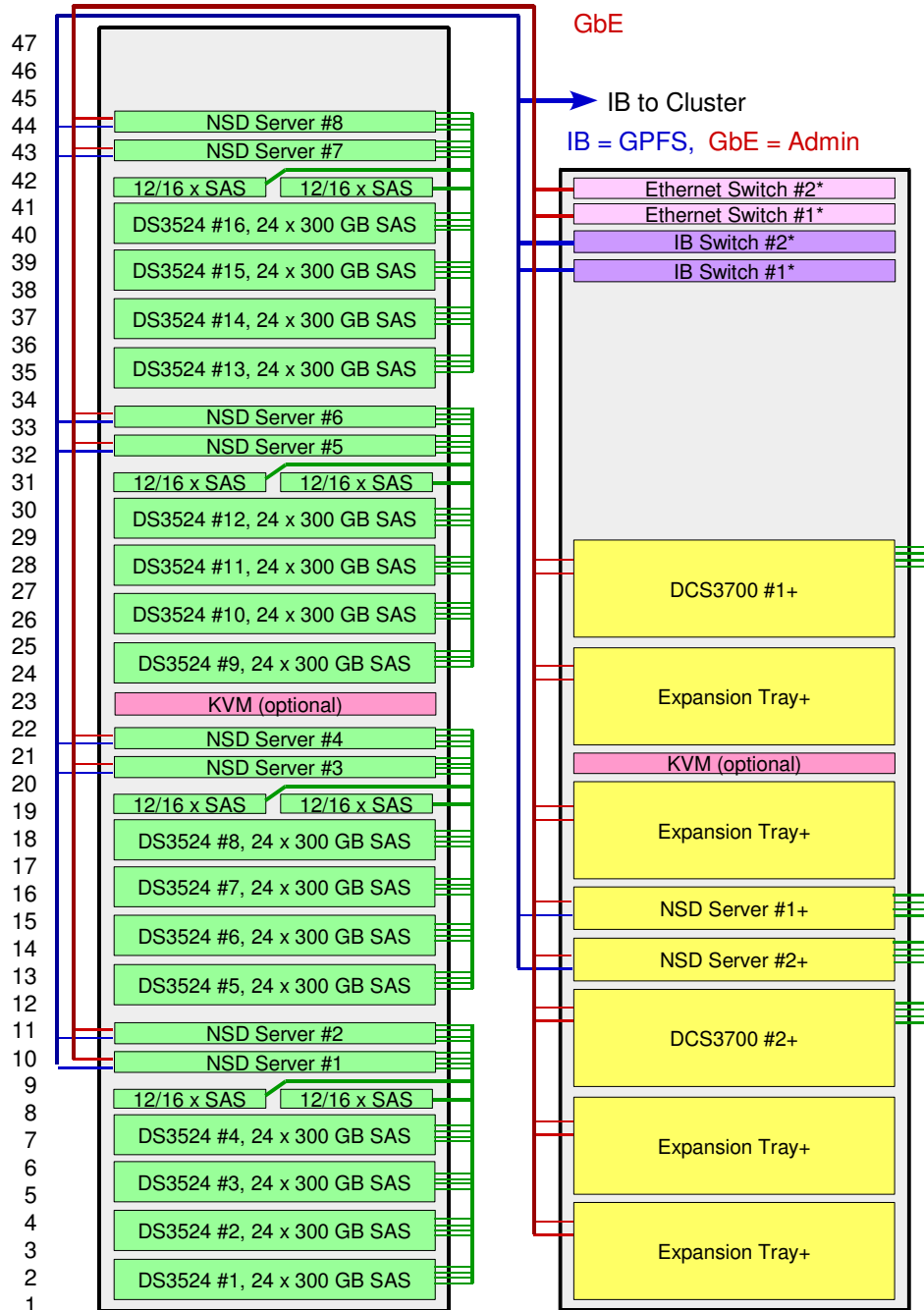
- ▶ Remote tape libraries



frequent use  
smaller capacity  
high BW/low latency  
more expensive

infrequent use  
larger capacity  
lower BW  
higher latency  
less expensive

# 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

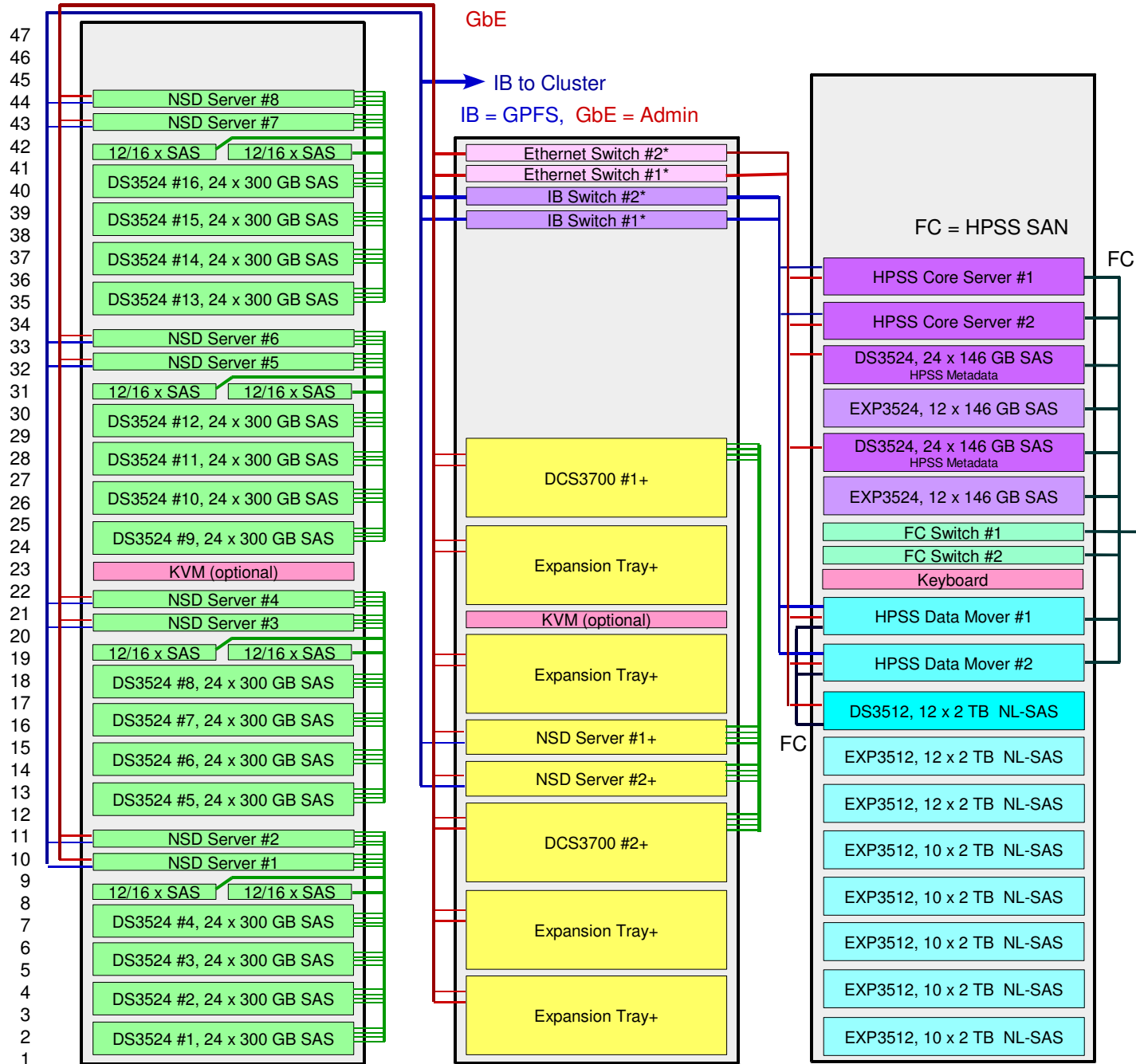
- ◆ Variation of Building Block #1A
- ◆ 1 x Building Block
- ◆ Aggregate Statistics
- Capacity: raw = 720 TB, usable < 524 TB
- Streaming
  - write < 3.2 GB/s
  - read < 4 GB/s

## 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