GPFS Best Practices

"If you can do it, it ain't braggin."

Dizzy Dean

Raymond L. Paden, Ph.D.
HPC Technical Architect
Deep Computing

26 Oct 09
version 1.2.0

raypaden@us.ibm.com
877-669-1853
512-858-4261

Special Notices from IBM Legal

This presentation was produced in the United States. IBM may not offer the products, programs, services or features discussed herein in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the products, programs, services, and features available in your area. Any reference to an IBM product, program, service or feature is not intended to state or imply that only IBM's product, program, service or feature may be used. Any functionally equivalent product, program, service or feature that does not infringe on any of IBM's intellectual property rights may be used instead of the IBM product, program, service or feature.

Information in this presentation concerning non-IBM products was obtained from the suppliers of these products, published announcement material or other publicly available sources. Sources for non-IBM list prices and performance numbers are taken from publicly available information including D.H. Brown, vendor announcements, vendor WWW Home Pages, SPEC Home Page, GPC (Graphics Processing Council) Home Page and TPC (Transaction Processing Performance Council) Home Page. IBM has not tested these products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this presentation. The furnishing of this presentation does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local IBM office or IBM authorized reseller for the full text of a specific Statement of General Direction.

The information contained in this presentation has not been submitted to any formal IBM test and is distributed "AS IS". While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. The use of this information or the implementation of any techniques described herein is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. Customers attempting to adapt these techniques to their own environments do so at their own risk.

IBM is not responsible for printing errors in this presentation that result in pricing or information inaccuracies.

The information contained in this presentation represents the current views of IBM on the issues discussed as of the date of publication. IBM cannot guarantee the accuracy of any information presented after the date of publication.

IBM products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this presentation was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements quoted in this presentation may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this presentation may have been estimated through extrapolation. Actual results may vary. Users of this presentation should verify the applicable data for their specific environment.

Microsoft, Windows, Windows NT and the Windows logo are registered trademarks of Microsoft Corporation in the United States and/or other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

LINUX is a registered trademark of Linus Torvalds. Intel and Pentium are registered trademarks and MMX,Itanium, Pentium II Xeon and Pentium III Xeon are trademarks of Intel Corporation in the United States and/or other countries.

Other company, product and service names may be trademarks or service marks of others.

What is GPFS?

General Parallel File System

All of GPFS's rivals do some of these things, none of them do all of them!

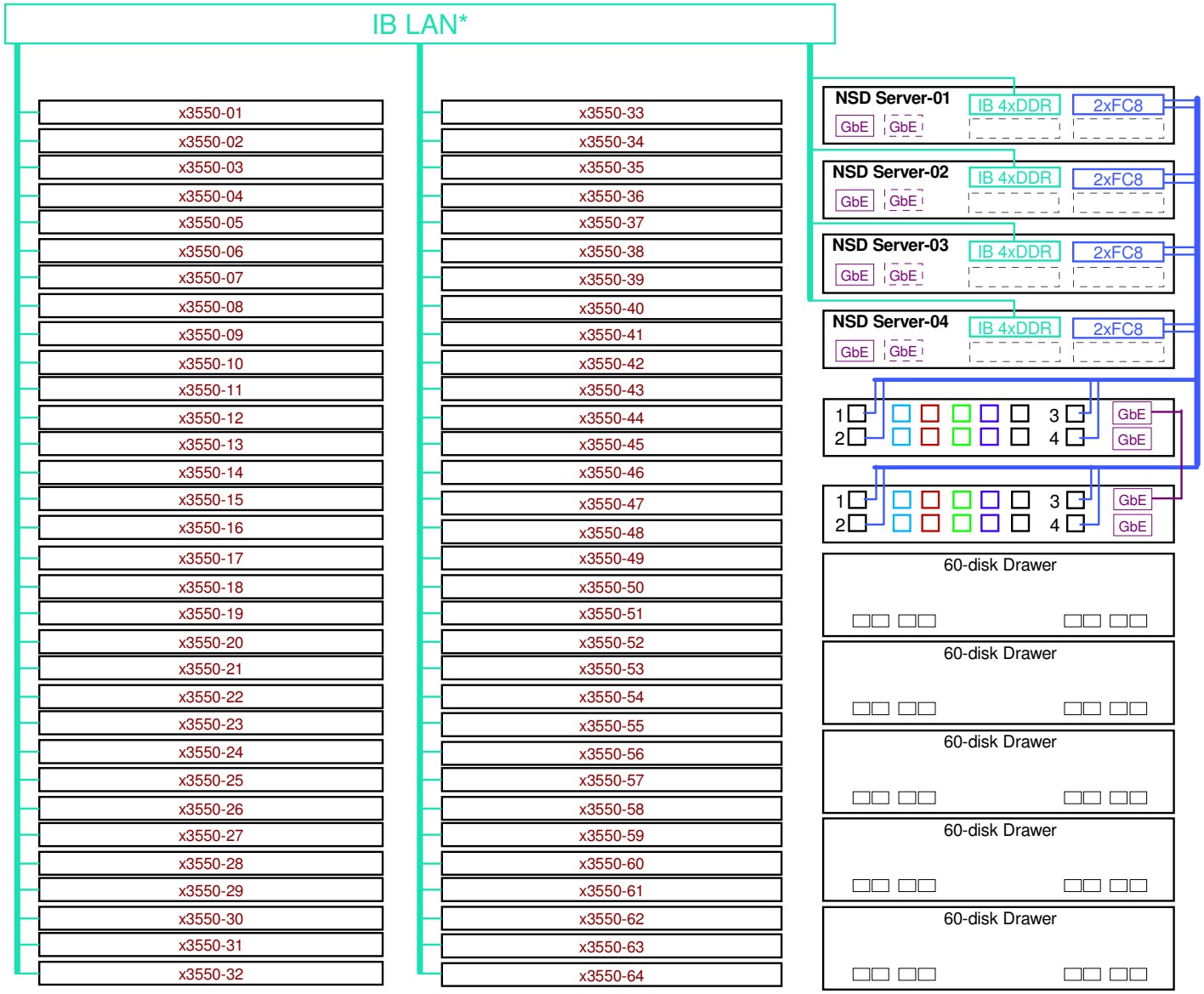
- *General*: supports wide range of applications and configurations
- *Cluster:* from large (4000+ in a multi-cluster) to small (only 1 node) clusters
- *Parallel:* user data and metadata flows between all nodes and all disks in parallel
- *HPC:* supports high performance applications
- Flexible: tuning parameters allow GPFS to be adapted to many environments
- *Capacity:* from high (4+ PB) to low capacity (only 1 disk)
- *Global*: Works across multiple nodes, clusters and labs *(i.e.,* LAN, SAN, WAN)
- *Heterogeneous:*
  - 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

What is GPFS?
Typical Example

Aggregate Performance and Capacity

Data rate: streaming rate < 5 GB/s, 4 KB transaction rate < 40,000 IOP/s

Usable capacity < 240 TB



LAN Configuration

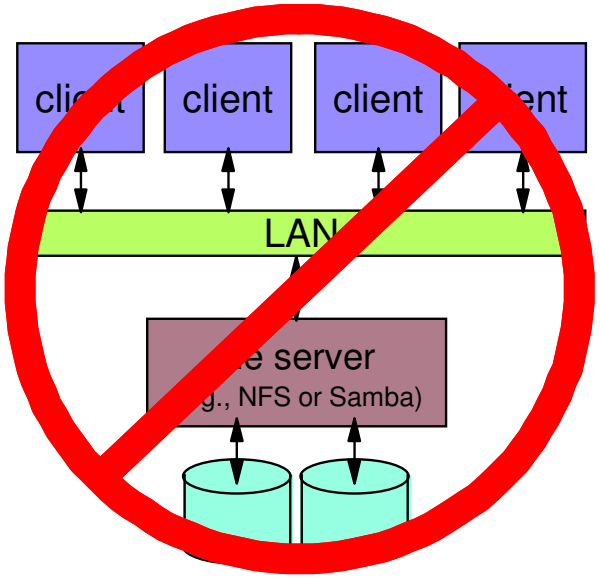
- Performance scales linearly in the number of storage servers
- Add capacity without increasing the number of servers
- Add performance by adding more servers and/or storage
- Inexpensively scale out the number of clients

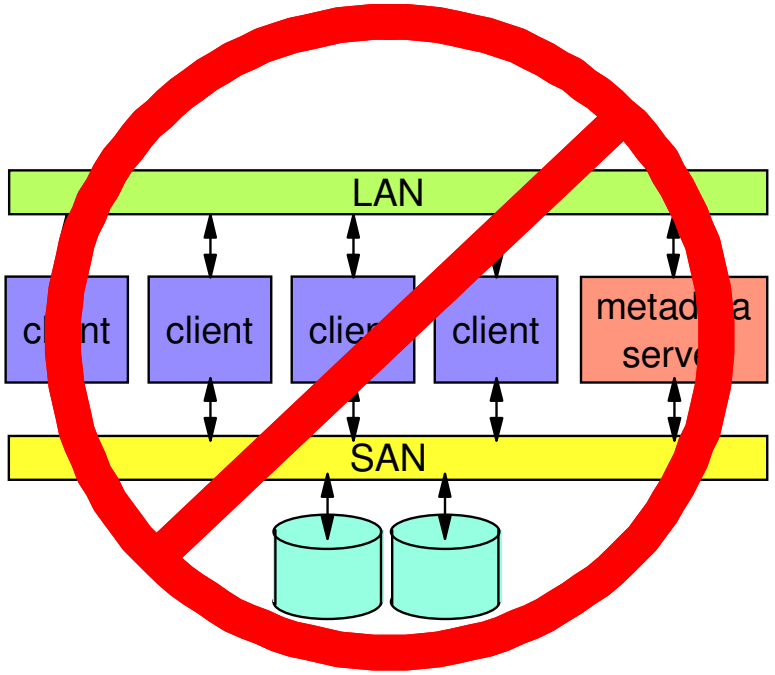
⋆Though not shown, a cluster like this will generally include an administrative GbE network.

What GPFS is *Not*

GPFS is ***not*** a client/server file system like NFS, CIFS (Samba) or AFS/DFS with a single file server.

- GPFS *nodes* can be an NFS or CIFS server, but GPFS treats them like any other application.





GPFS is ***not*** a SAN file system with dedicated metadata server.

- GPFS can run in a SAN file system like mode, but it does not have a dedicated metadata server.

GPFS avoids the bottlenecks introduced by centralized file and/or metadata servers.

What GPFS is *Not*

GPFS is not a niche file system for IBM system P products

- Yesterday
  - GPFS was a parallel file system for IBM SP systems
- Today
  - GPFS is a general purpose clustered parallel file system tunable for many workloads on many configurations.

Winterhawk

BlueGene/P

P6 p595

iDataPlex

BladeCenter/H

Where GPFS Is Used Today

GPFS is a mature product with established market presence. It has been generally available since 1998 with research development starting in 1991. Applications include...

- Aerospace and Automotive
- Banking and Finance
- Bio-informatics and Life Sciences
- Defense
- Digital Media
- EDA (Electronic Design Automation)
- General Business
- National Labs
- Petroleum
- SMB (Small and Medium sized Business)
- Universities
- Weather Modeling

Local Area Network (LAN) Architecture

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

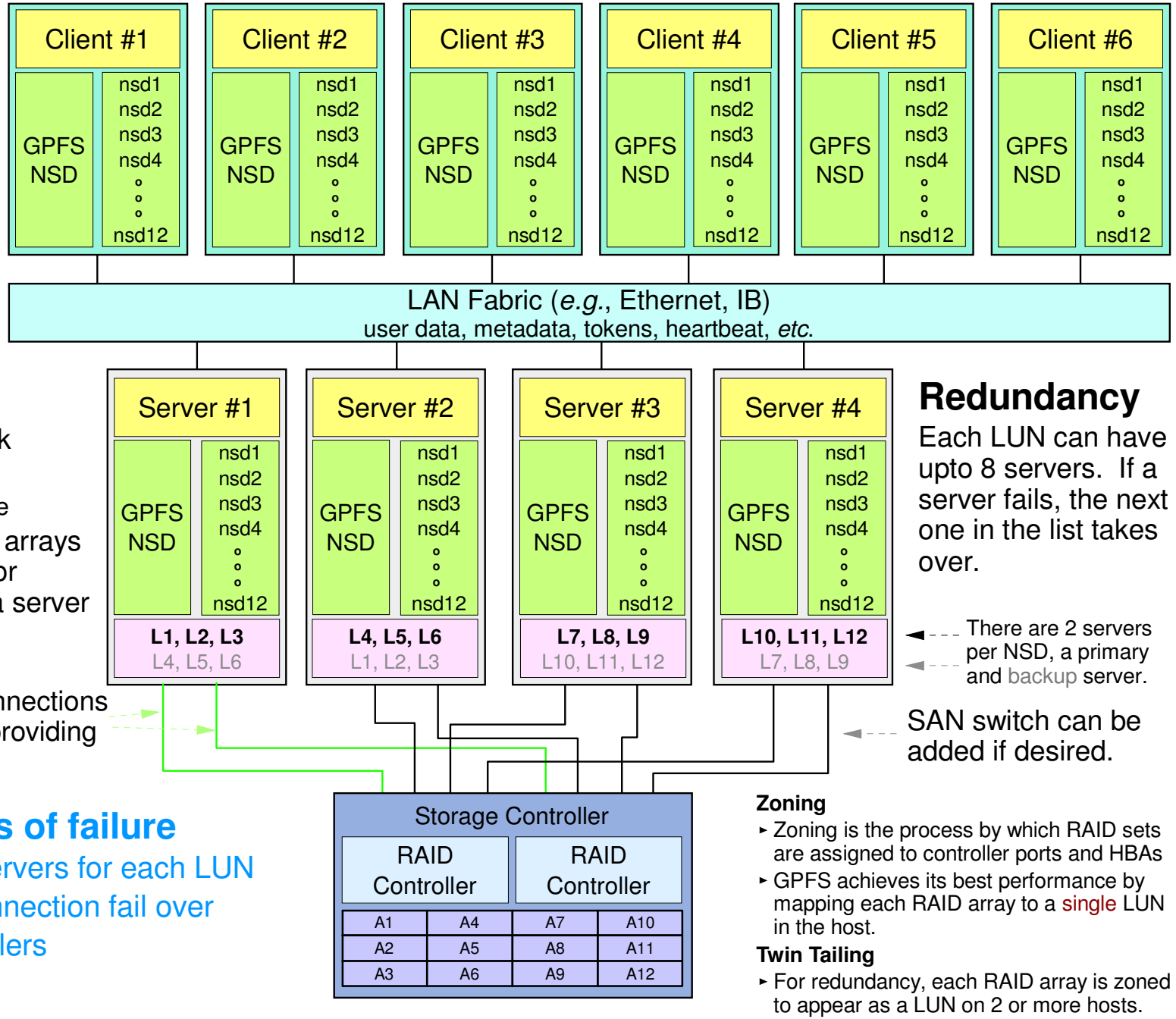
- Logical Unit
- Abstraction of a disk
  - AIX - hdisk
  - Linux - SCSI device
- LUNs map to RAID arrays in a disk controller or "physical disks" in a server

Redundancy

Each server has 2 connections to the disk controller providing redundancy

No single points of failure

- primary/backup servers for each LUN
- controller/host connection 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.

Zoning

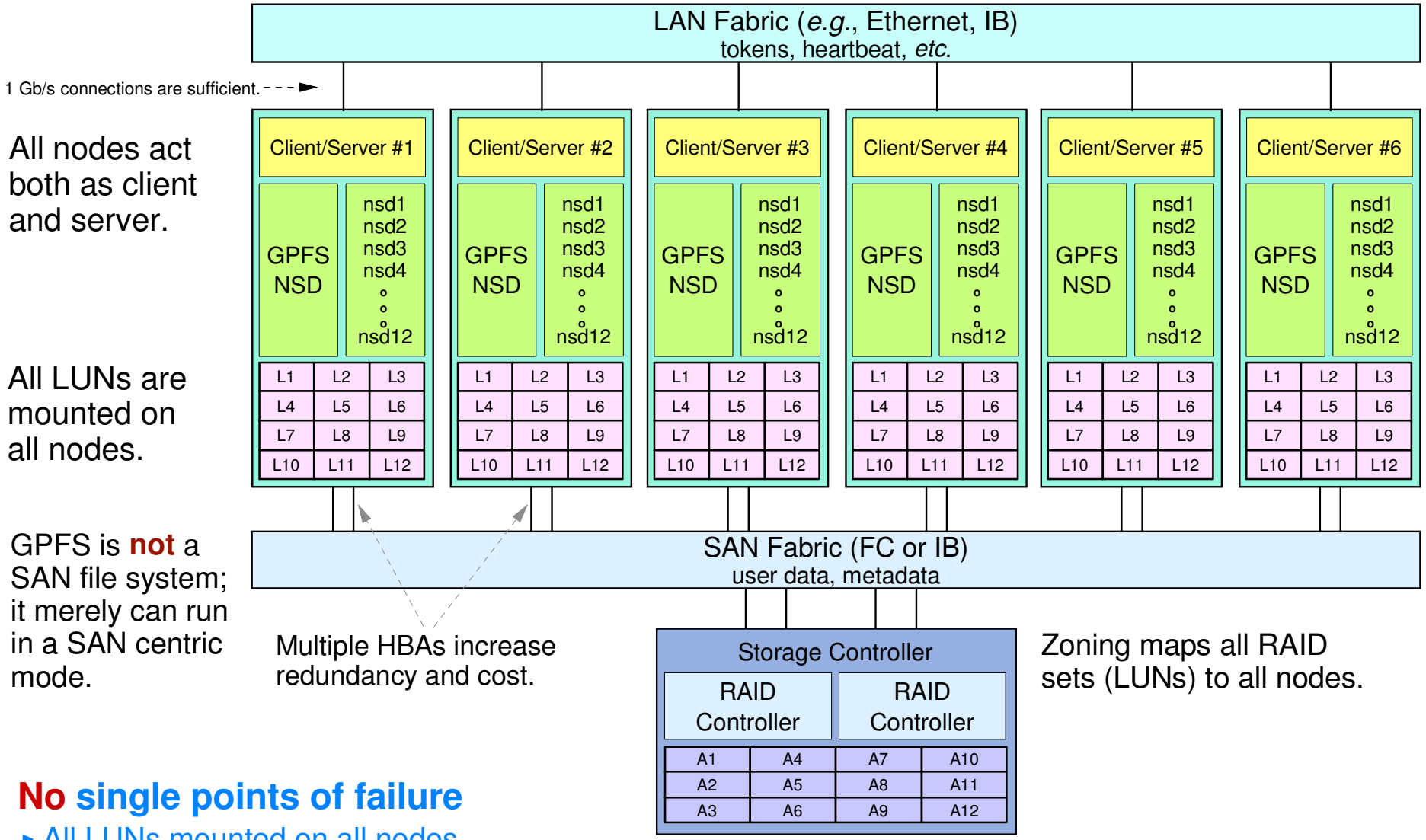
- Zoning is the process by which RAID sets are assigned to controller ports and HBAs
- GPFS achieves its best performance by mapping each RAID array to a single LUN in the host.

Twin Tailing

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

Storage Area Network (SAN) Architecture

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 RAID sets (LUNs) to all nodes.

No single points of failure

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

CAUTION:

Not recommended for SANs with > 32 host ports. Scaling beyond this requires special tuning (*e.g.*, set queue depth very small).

GPFS Best Practices

The following pages provide examples of what are and what are not GPFS best practices.

The examples are based on iDataPlex solutions using currently available storage solutions.

The principles motivating these examples can be extended to other server technologies (*e.g.*, rack optimized, blade, System X, System P, and so forth).

Recommended Storage Solutions

For reasons of best practices, the DS5300 and DCS9900 are generally the preferred HPC storage solutions.

For customers with smaller capacity and/or lower performance requirements that unlikely to grow over time to larger/faster solutions, the DS3200, DS3400, DS5020 and DS5100 are acceptable solutions.

The DS8300, XIV or the 3U iDX storage servers should not in general be used for HPC solutions.

COMMENT: The distinction between large and small capacity or high and low performance is not exact. However, as a guideline, if greater than 240 SATA disks or 128 FC or SAS disks are needed, or if greater than 3 GB/s is needed, then a DS5300 or DCS9900 is recommended.

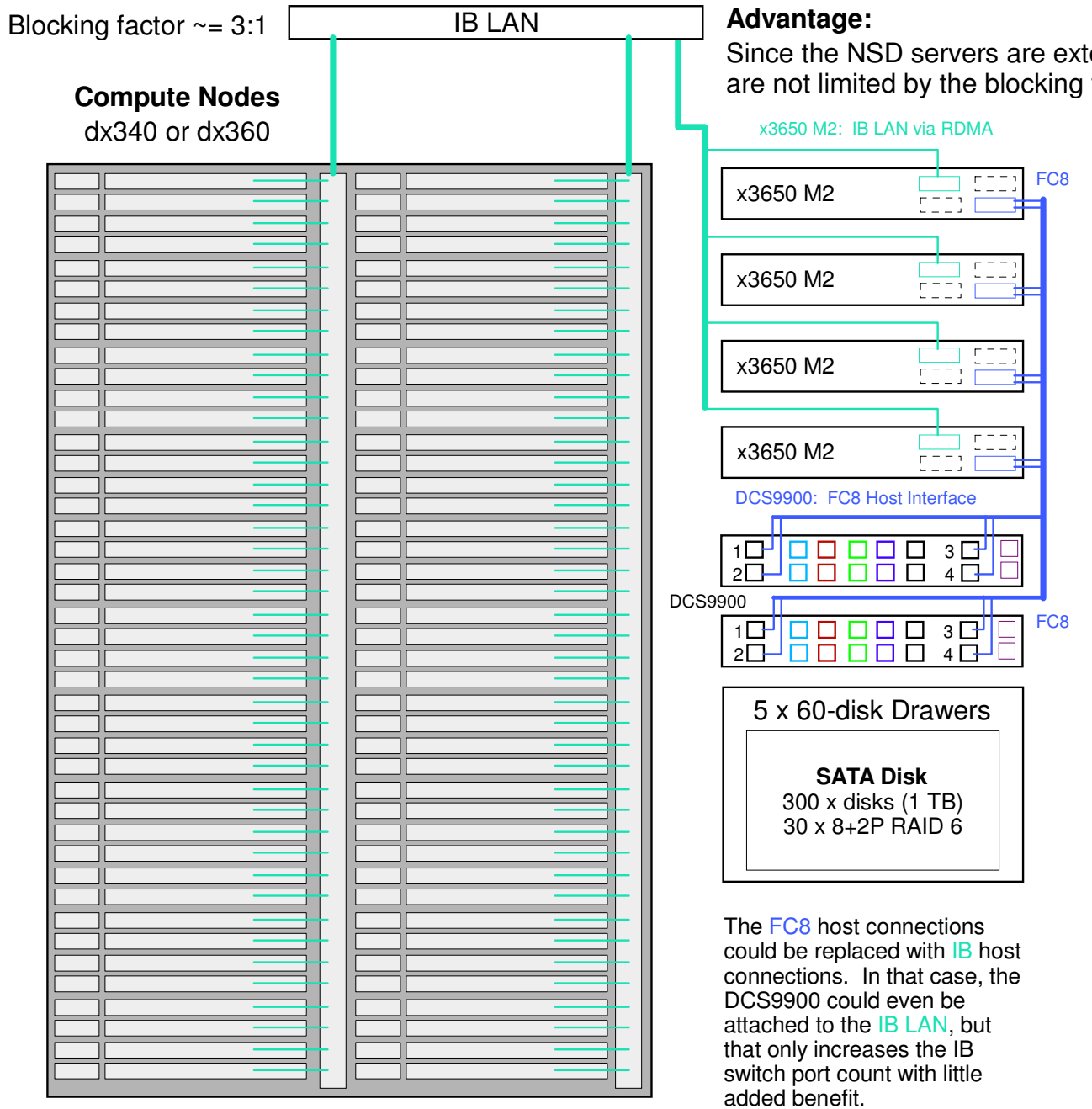
GPFS Building Blocks

A convenient design strategy for GPFS solutions is to define a "storage building block", which is the "smallest" increment of storage and servers by which a storage system can grow.

Therefore, 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)

This solution is made feasible since GPFS scales linearly in the number of disks, storage controllers, NSD servers, GPFS clients, and so forth.

Using GPFS in iDataPlex as an IB/LAN File System
External NSD Servers



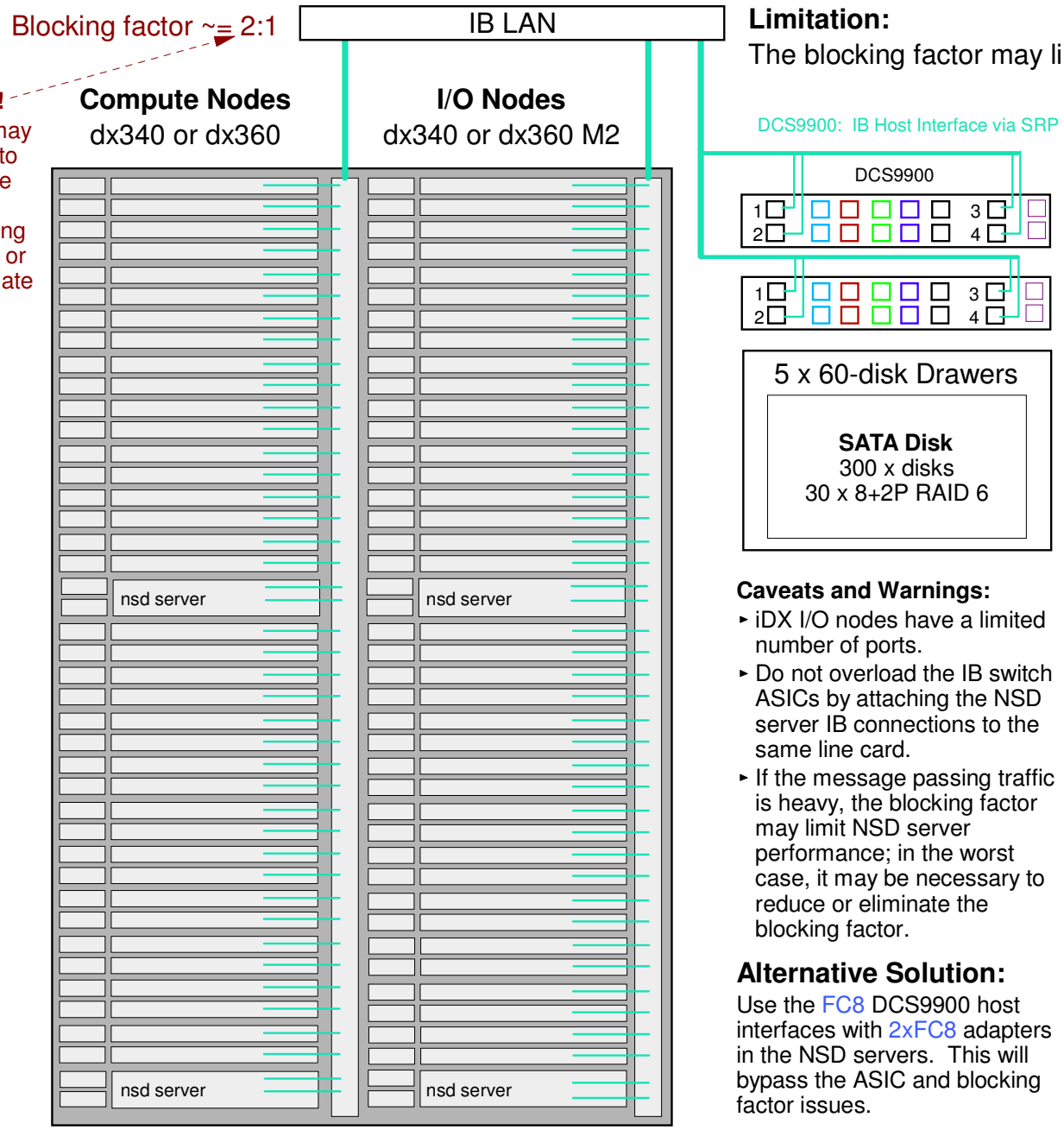
Advantage:
Since the NSD servers are external to the iDX frames, they are not limited by the blocking factor.

The FC8 host connections could be replaced with IB host connections. In that case, the DCS9900 could even be attached to the IB LAN, but that only increases the IB switch port count with little added benefit.

Analysis

- 84 nodes (1 frame)
  - aggregate rates
    write < 5.4 GB/s
    read < 3.6 GB/s
  - peak rate per node
    500 to 1500 MB/s
    depending on blocking
  - average rate per node
    write < 64 MB/s
    read < 40 MB/s
- 336 nodes (4 frames)
  - aggregate rates
    write < 5.4 GB/s
    read < 3.6 GB/s
  - peak rate for per node
    500 to 1500 MB/s
    depending on blocking
  - average rate per node
    write < 16 MB/s
    read < 10 MB/s
- Capacity
  - raw < 300 TB
  - usable < 240 TB

Using GPFS in iDataPlex as an IB/LAN File System
Internal NSD Servers



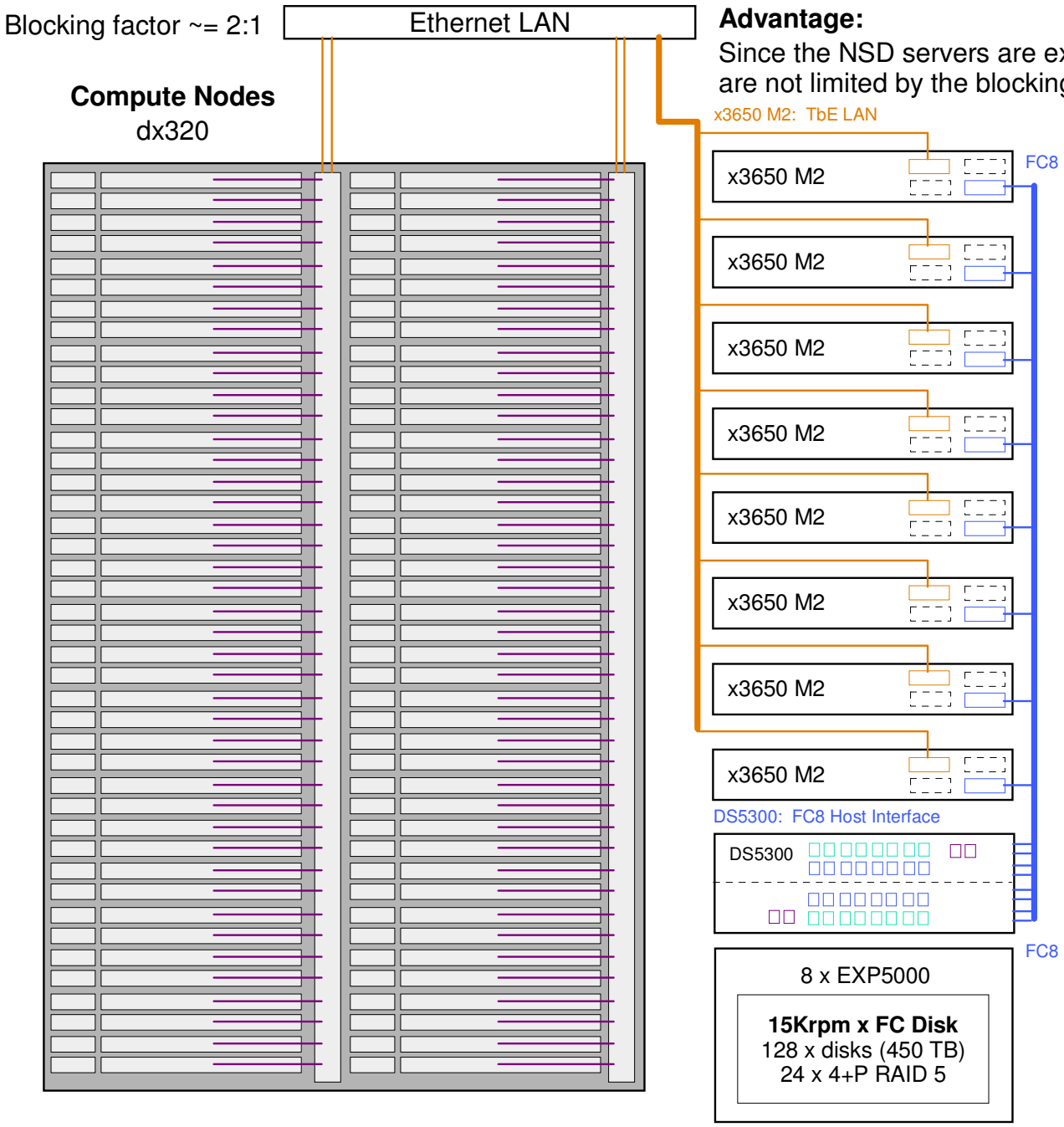
Limitation:
The blocking factor may limit storage server performance.

Caveats and Warnings:
- iDX I/O nodes have a limited number of ports.
- Do not overload the IB switch ASICs by attaching the NSD server IB connections to the same line card.
- If the message passing traffic is heavy, the blocking factor may limit NSD server performance; in the worst case, it may be necessary to reduce or eliminate the blocking factor.

Alternative Solution:
Use the FC8 DCS9900 host interfaces with 2xFC8 adapters in the NSD servers. This will bypass the ASIC and blocking factor issues.

Analysis
- 84 nodes (1 frame)
  - aggregate rates
    write < 5.4 GB/s
    read < 3.6 GB/s
  - peak rate per node
    750 to 1500 MB/s
    depending on blocking
  - average rate per node
    write < 64 MB/s
    read < 40 MB/s
- 336 nodes (4 frames)
  - aggregate rates
    write < 5.4 GB/s
    read < 3.6 GB/s
  - peak rate per node
    750 to 1500 MB/s
    depending on blocking
  - average rate per node
    write < 16 MB/s
    read < 10 MB/s
- Capacity
  - raw < 300 TB
  - usable < 240 TB

Using GPFS in iDataPlex as a Ethernet/LAN File System External NSD Servers



Advantage:
Since the NSD servers are external to the iDX frames, they are not limited by the blocking factor.

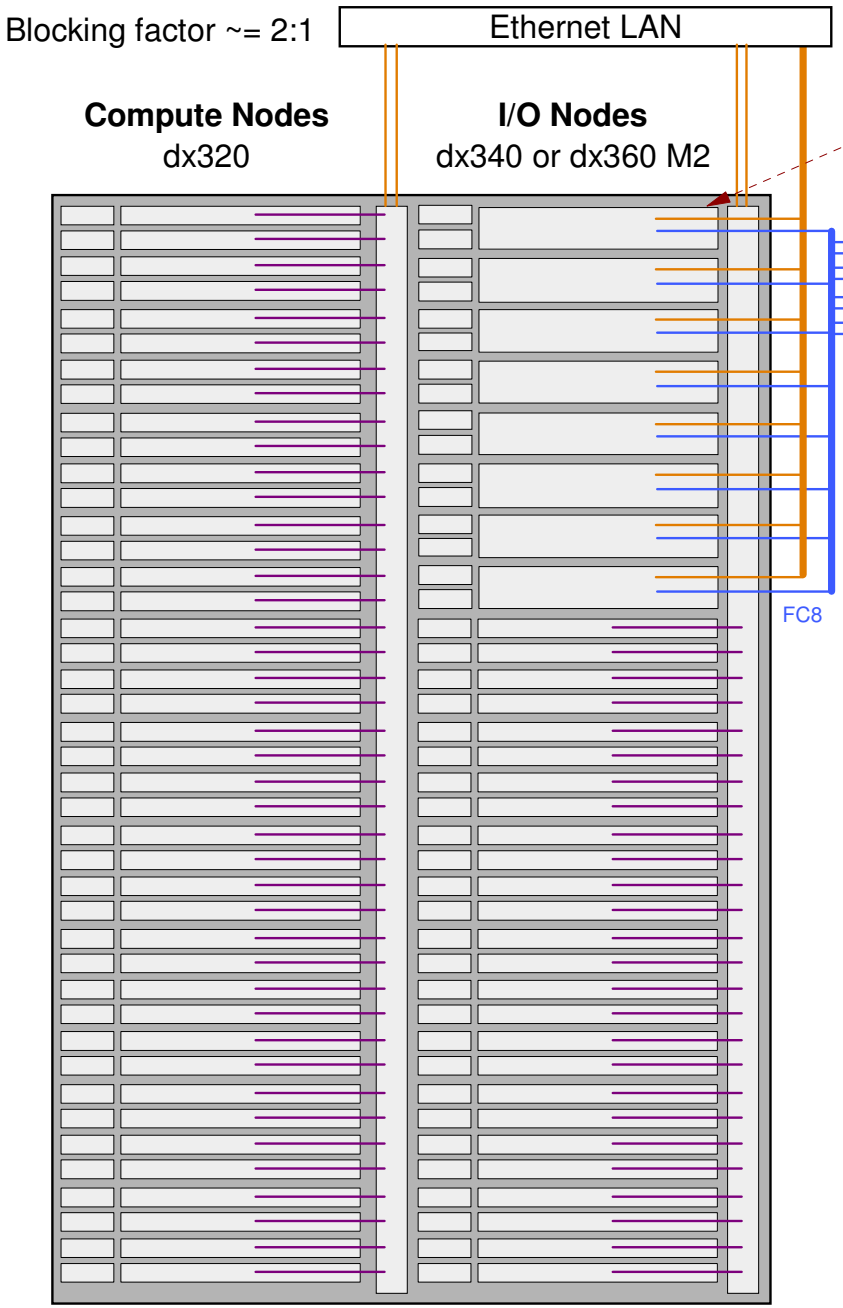
Analysis

- 84 nodes (1 frame)
  - aggregate rates
    write < 4.5 GB/s
    read < 5.5 GB/s
  - peak rate per node < 80 MB/s
  - average rate per node < 33 MB/s
    can not fully utilize controller BW*
- 336 nodes (4 frames)
  - aggregate rates
    write < 4.5 GB/s
    read < 5.5 GB/s
  - peak rate per node < 80 MB/s
  - average rate per node
    write < 13 MB/s
    read < 16 MB/s
- Capacity
  - raw < 56 TB
  - usable < 42 TB

* The DS5300 can deliver up to 5.5 GB/s, but with only 4 TbE uplinks, only 1400 MB/s can be used assuming 50% of the TbE uplinks are devoted to GPFS. At least 4 frames are needed for this solution to be meaningful.

Alternative Solution:
Use 4 NSD servers each with a 2xTbE; in theory this should work, but in practice load balancing over bonded ports is difficult to manage.

Using GPFS in iDataPlex as a Ethernet/LAN File System Internal NSD Servers



Limitation:

Must use I/O nodes in a nonstandard way to avoid blocking factor limitations.

Look! Bypass the BNT switches.

Caveat:

Bypassing the BNT switch avoids the bandwidth restrictions caused by the blocking factor.

Alternative Solution:

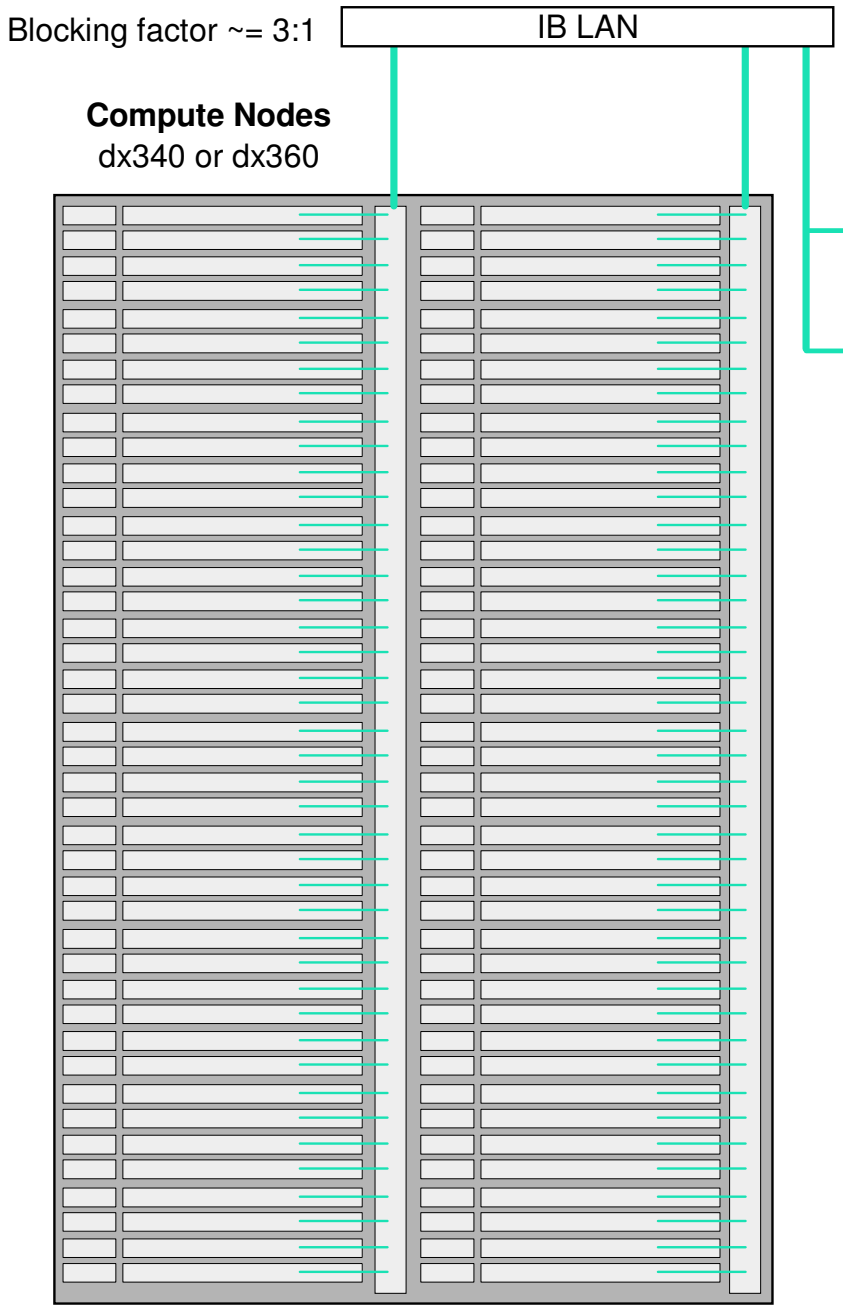
Use 4 NSD servers each with a 2xTbE; in theory this should work, but in practice load balancing over bonded ports is difficult to manage.

Analysis

- 84 nodes (1 frame)
  - aggregate rates
    write < 4.5 GB/s
    read < 5.5 GB/s
  - peak rate per node < 80 MB/s
  - average rate per node < 33 MB/s
    can not fully utilize controller BW
- 336 nodes (4 frames)
  - aggregate rates
    write < 4.5 GB/s
    read < 5.5 GB/s
  - peak rate per node < 80 MB/s
  - average rate per node
    write < 13 MB/s
    read < 16 MB/s
- Capacity
  - raw < 56 TB
  - usable < 42 TB

★ The DS5300 can deliver up to 5.5 GB/s, but with only 4 TbE uplinks, only 1400 MB/s can be used assuming 50% of the TbE uplinks are devoted to GPFS. At least 4 frames are needed for this solution to be meaningful.

Using GPFS in iDataPlex as a SAN File System



Limitation:

Must reduce the queue depth when scaling out the size of the SAN. May require other special tuning to guarantee stability.

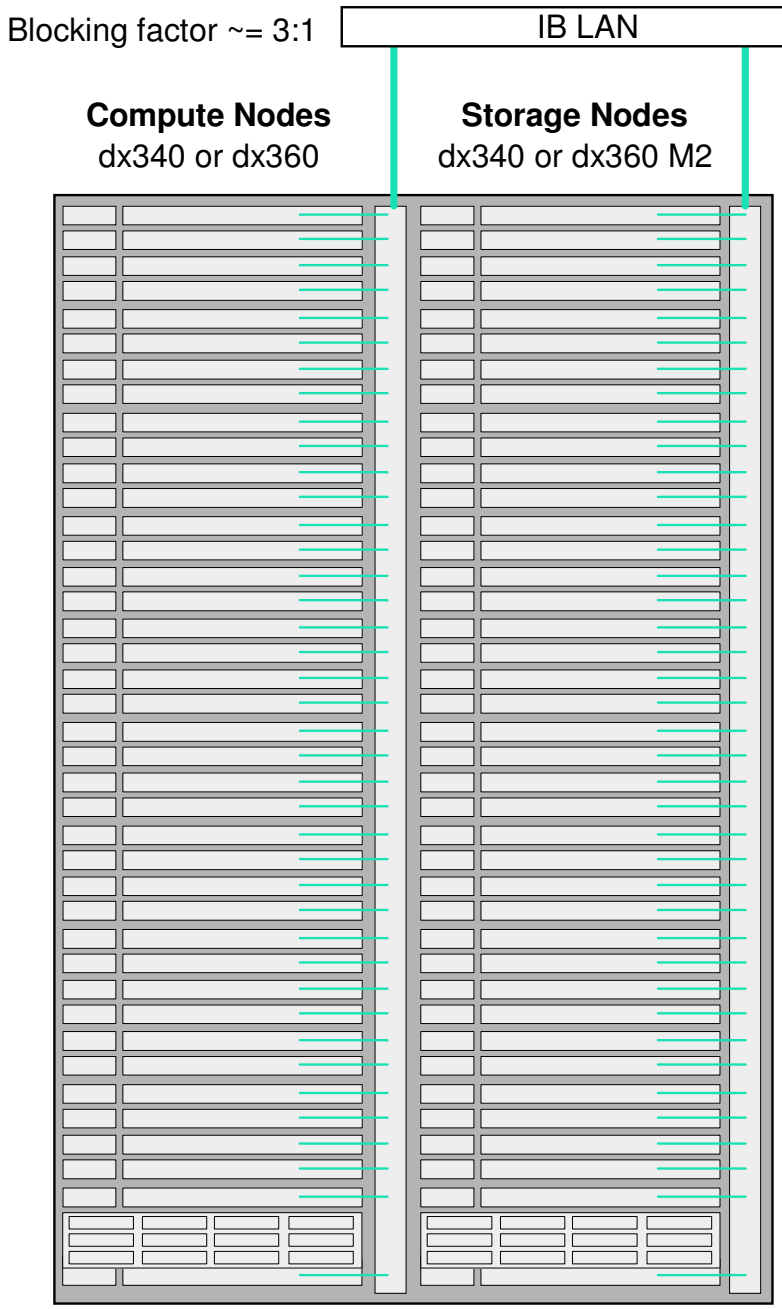
Caveats and Warnings:

- Do **not** scale this solution beyond 3 frames (252 nodes) without assistance from the GPFS development team. (Currently the largest SAN in production is 256 nodes).
- Reduce the queue depth on clusters with more than 32 SAN attachments (e.g., set the queue depth to 1 or 2 on a cluster with 256 nodes).

Analysis

- 84 nodes (1 frame)
  - aggregate rates
    write < 5.4 GB/s
    read < 3.6 GB/s
  - peak rate per node
    500 to 1500 MB/s
    depends on impact of blocking
  - average rate per node
    write < 64 MB/s
    read < 40 MB/s
- 252 nodes (3 frames)
  - aggregate rates
    write < 5.4 GB/s
    read < 3.6 GB/s
  - peak rate per node
    500 to 1500 MB/s
    depends on impact of blocking
  - average rate per node
    write < 21 MB/s
    read < 14 MB/s
- Capacity
  - raw < 300 TB
  - usable < 240 TB

Using iDataPlex Storage Nodes as NSD Servers



Limitations:

- Single point of failure design
- Blocking factor may limit storage server performance.

Analysis

- 84 nodes (1 frame)
  - aggregate rates*
    write < 800 MB/s
    read < 1200 MB/s
  - peak rate per node
    750 to 1500 MB/s
    depends on impact of blocking
  - average rate per node
    write < 10 MB/s
    read < 15 MB/s
- Capacity
  - SAS: 450 GB/disk
    raw < 11 TB
    usable (no mirroring) < 7 TB
    usable (mirroring) < 3.5 TB

Analysis

- 336 nodes (4 frames)
- Assume 2 storage nodes per frame
  - aggregate rates*
    write < 3.0 GB/s
    read < 4.5 GB/s
  - peak rate per node
    750 to 1500 MB/s
    depends on impact of blocking
  - average rate per node
    write < 10 MB/s
    read < 15 MB/s
- Capacity
  - SAS: 450 GB/disk
    raw < 44 TB
    usable (no mirroring) < 28 TB
    usable (mirroring) < 14 TB

*Based on the following data rates for a 4+P RAID 5 SAS LUN
write < 200 MB/s
read < 300 MB/s

Warnings:

- iDX 3U storage nodes are a single point of failure risk exposure
  - spanning a GPFS file system across multiple storage nodes increases the risk
  - this is an issue for all file systems and is **not** considered a best practice
  - If this must be done, use GPFS mirroring and storage pools to manage the risk.
- Do not use 5+P RAID 5 or 10+2P RAID 6 configurations
- GPFS has not been tested with ServeRAID controllers

DS3000 Series

DS3200

- 3-Gbps SAS connect to host
- Direct-attach
- For System x

- 2U, 12 disks
- Dual Power Supplies
- Support for SAS or SATA disks
- Expansion via EXP3000

DS3400

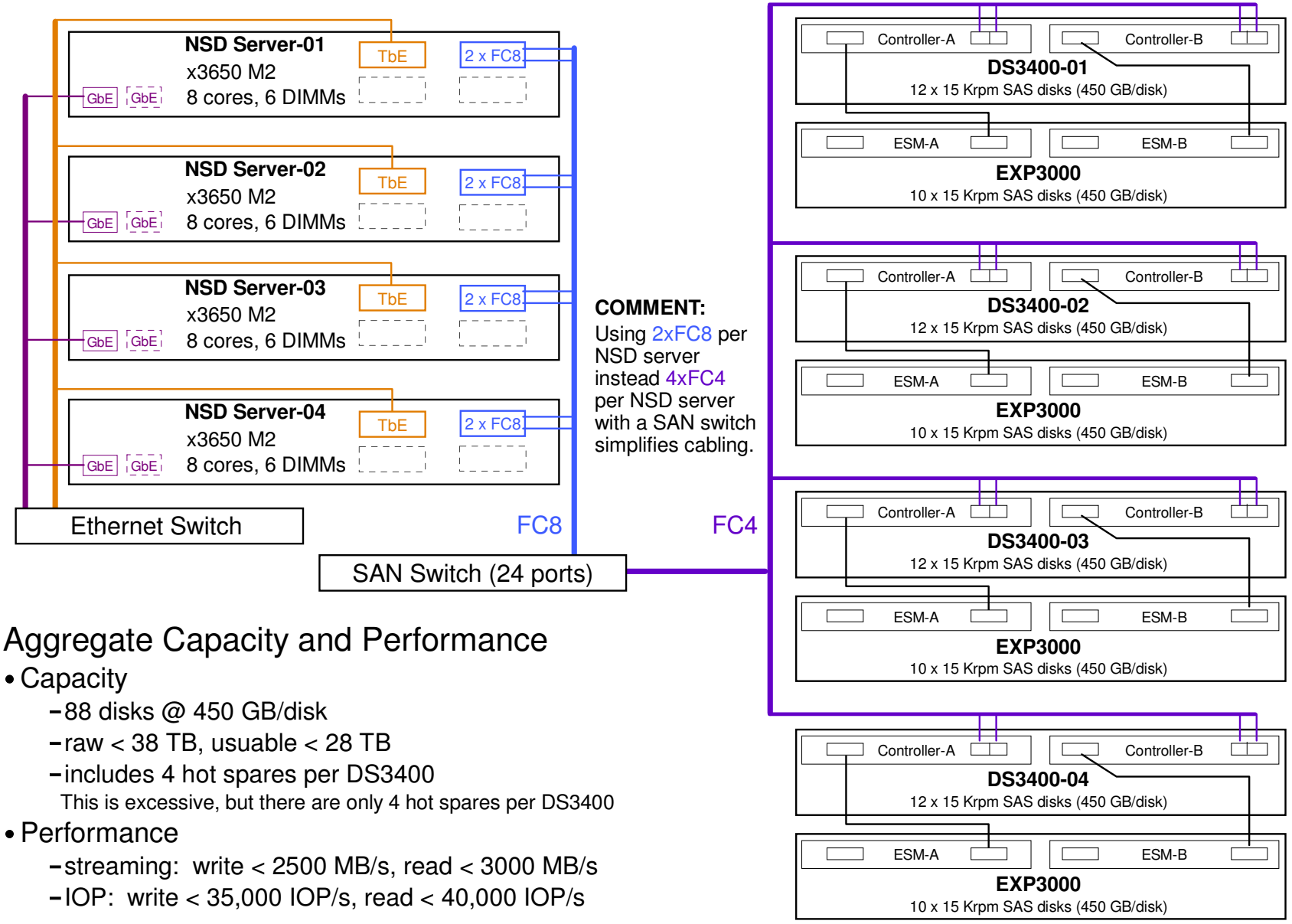
- 4-Gbps Fibre connect to host
- Direct-attach or SAN
- For System x & BladeCenters

- 2U, 12 disks
- Dual Power Supplies
- Support for SAS or SATA disks
- Expansion via EXP3000

WARNINGS: DS3000 controllers are ideal for smaller storage configurations (*n.b.*, do **not** exceed 4 x DS3400s within a single cluster). Also, they do **not** make good metadata stores for GPFS.

DS3400

Small Performance Optimized Solution with Maximum Scaling



Aggregate Capacity and Performance

- Capacity
  - 88 disks @ 450 GB/disk
  - raw < 38 TB, usuable < 28 TB
  - includes 4 hot spares per DS3400

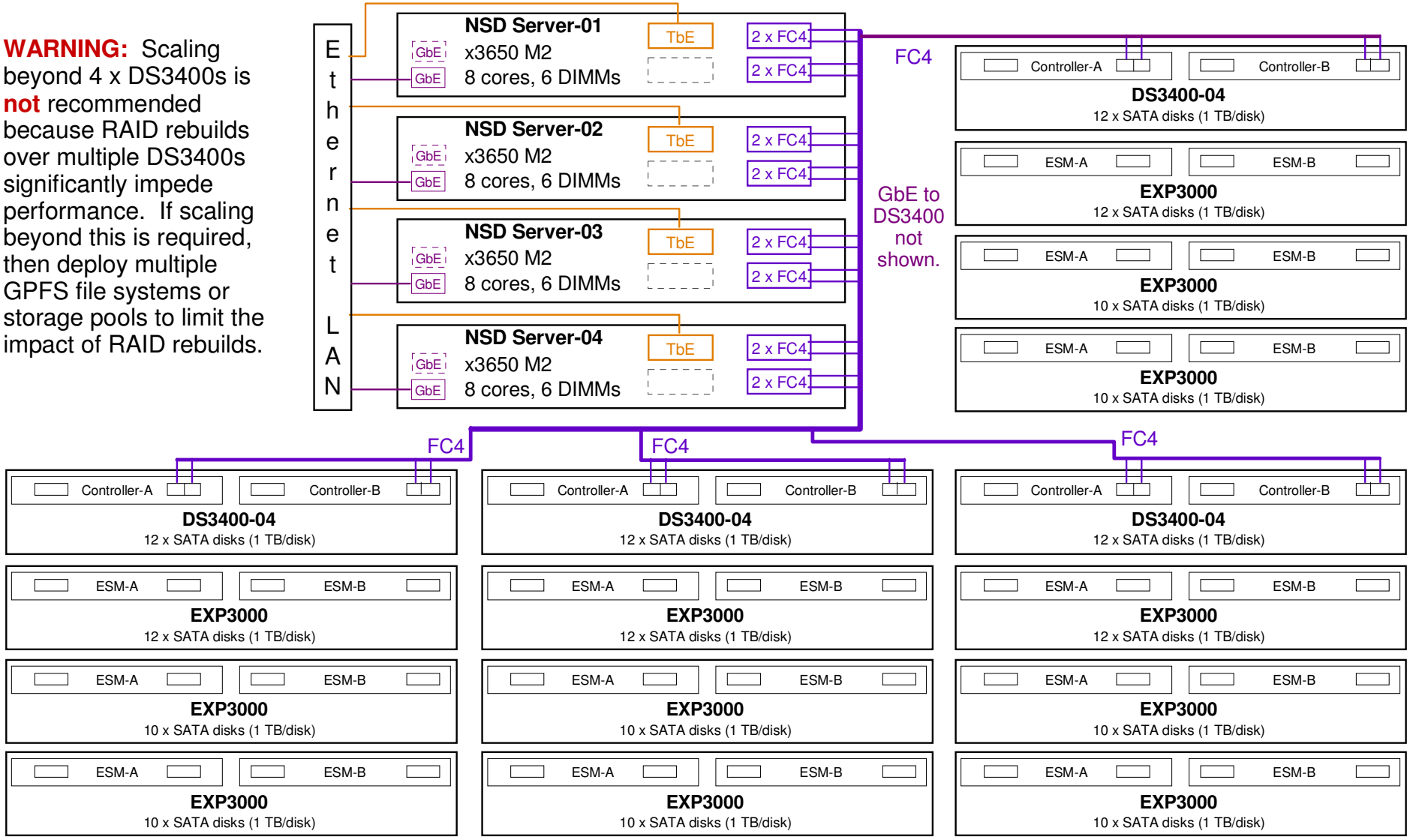
    This is excessive, but there are only 4 hot spares per DS3400
- Performance
  - streaming: write < 2500 MB/s, read < 3000 MB/s
  - IOP: write < 35,000 IOP/s, read < 40,000 IOP/s

WARNING: Scaling beyond 4 x DS3400s is **not** recommended because RAID rebuilds over multiple DS3400s significantly impede performance. If scaling beyond this is required, then deploy multiple GPFS file systems or storage pools to limit the impact of RAID rebuilds.

DS3400

Small Capacity Optimized Solution with Maximum Scaling

WARNING: Scaling beyond 4 x DS3400s is **not** recommended because RAID rebuilds over multiple DS3400s significantly impede performance. If scaling beyond this is required, then deploy multiple GPFS file systems or storage pools to limit the impact of RAID rebuilds.



NSD Server: x3650 M2

- 8 Cores, 6 GB RAM
- 2 dual-port 8 Gb/s FC HBAs (2xFC8)
  - at most 1500 MB/s per adapter
- IB HCA (4xDDR2)
  - at most 1500 MB/s per HCA

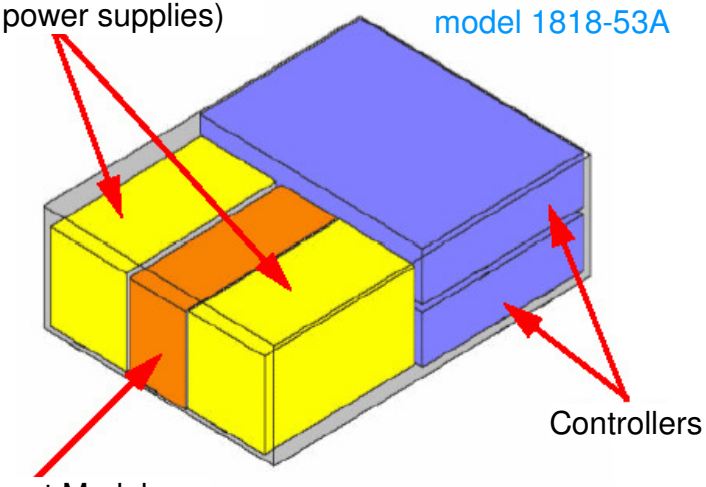
Capacity Optimized

- Use 4 drawers of 1 TB SATA disk per DS3400
- Capacity per DS3400
  - 42 disks @ 1 TB/disk in 8+2P RAID 6 configuration
  - raw = 42 TB, usable = 32 TB
  - includes 2 hot spares
- Aggregate Capacity
  - raw = 168 TB, usable = 128 TB
  - includes 8 hot spares
- Aggregate Performance
  - streaming rate < 3 GB/s

DS5000 Series

Controller Support Modules
(Fans, power supplies)

DS5300
model 1818-53A



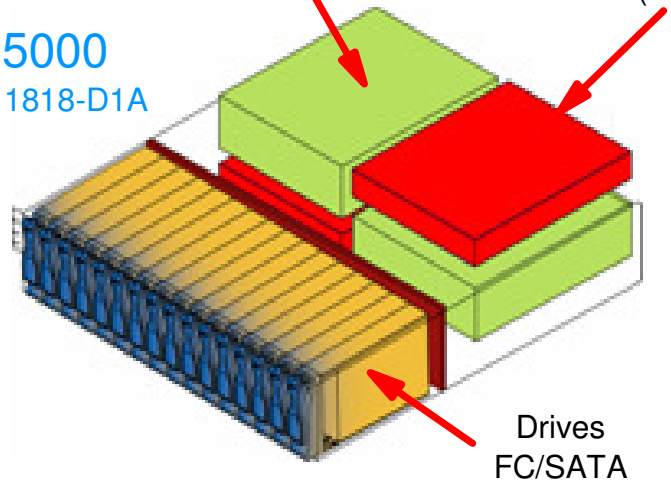
Controllers

Interconnect Module
(batteries, midplane)

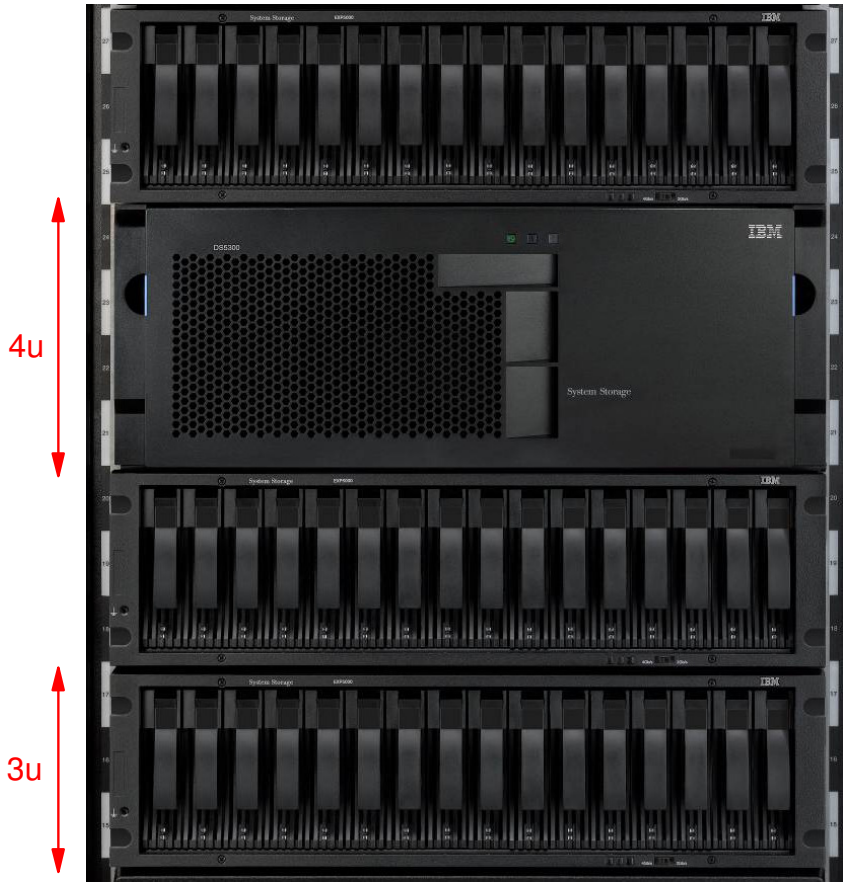
Power/cooling

Controllers
(ESMs)

EXP5000
model 1818-D1A



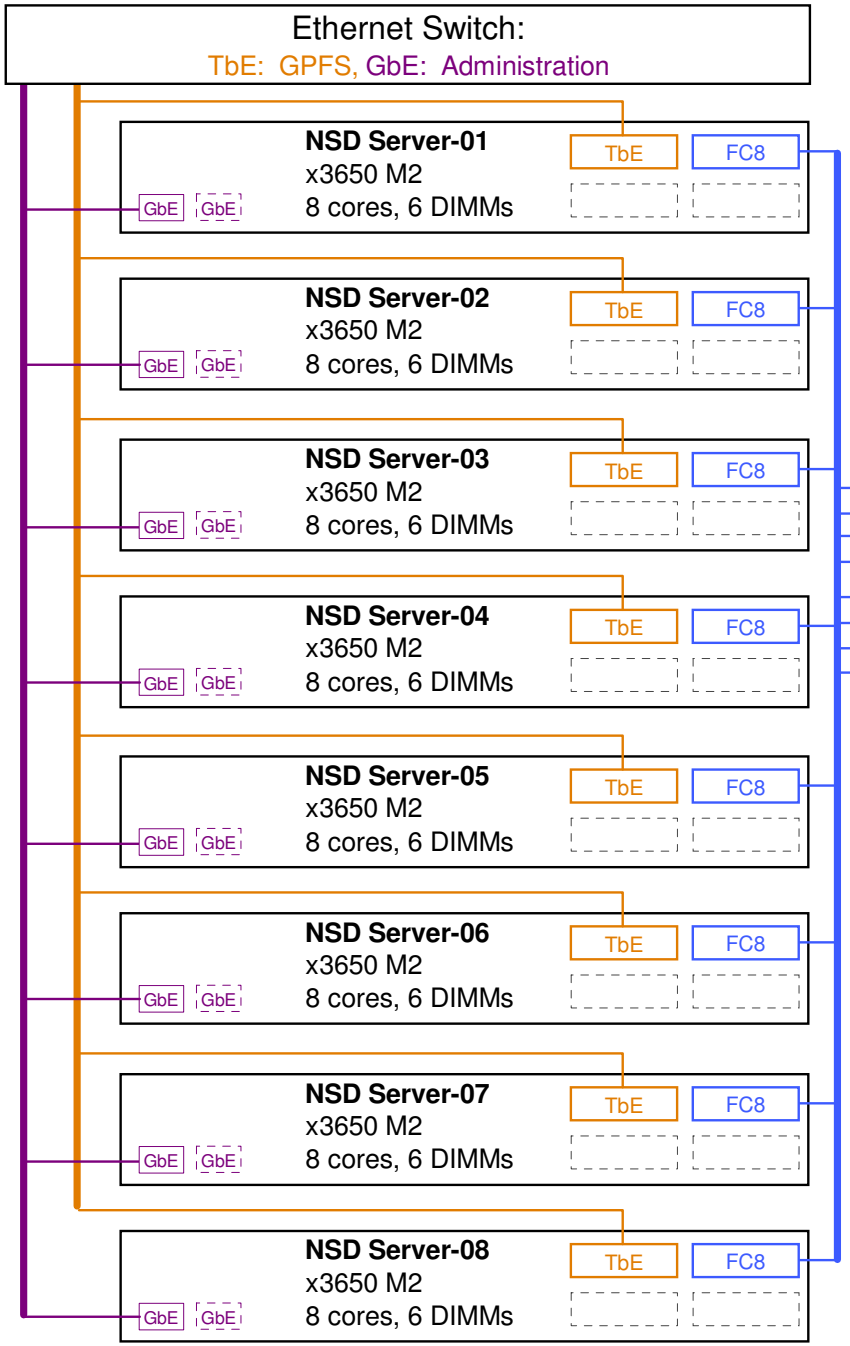
Drives
FC/SATA

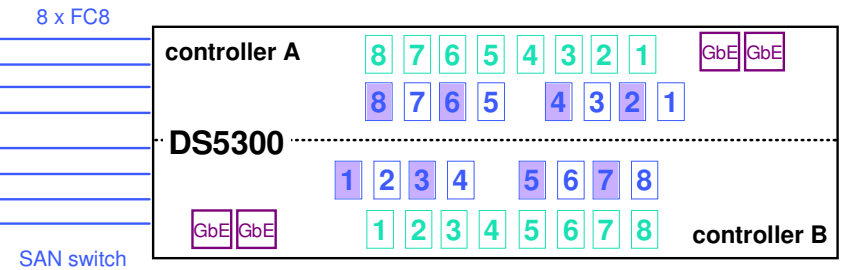


16 Disks per Disk Enclosure

DS5300

Example DS5300 Building Block





The GbE administrative network is not illustrated in this diagram.

SAN switch **not** required

Performance Analysis

- DS5300 streaming data rate
  - 256 x SATA or 128 x 15Krpm disks: write < 4.5 GB/s, read < 5.5 GB/s
- DS5300 IOP rate
  - 256 x SATA disks: write < 3600 IOP/s, read < 12,000 IOP/s
  - 128 x 15Krpm disks: write < 9,000 IOP/s, read < 36,000 IOP/s
- potential aggregate TbE rate: 8 x TbE < 5.6 GB/s
  - 725 MB/s per TbE is possible, but 700 MB/s is required
- potential aggregate FC8 rate: 8 x FC8 < 6.0 GB/s
  - 780 MB/s per FC8 is possible, but 700 MB/s is required

Disk Drawers

- option #1: 128 x 15Krpm FC disk
- option #2: 256 x SATA disks

COMMENT:

- This is a "safe" configuration in the sense that meeting projected performance rates can reasonably be expected (n.b., there are more than enough servers, FC8 and TbE ports to do the job).
- If HBA failover is required, then 8 dual port HBAs may be adopted (thereby requiring a SAN switch). If 2xFC8 adapters are adopted, then peak performance can be maintained during failure conditions.

To Be Completed

Detailed GPFS benchmarks will be available soon and shared with you when they are completed.

Preliminary analysis based on other benchmarks suggests the following can be expected.

- 128 x 15Krpm FC disks @ 4+P RAID 5 (8 x EXP5000)
  - streaming: write < 4.5 GB/s, read < 5.5 GB/s
  - IOP rates: write < 9,000 IOP/s, read < 36,000 IOP/s
- 240 x SATA disks @ 8+2P RAID 6 (4 x EXP5060)
  - streaming: write < 4.5 GB/s, read < 5.5 GB/s
  - IOP rates: write < 3,600 IOP/s, read < 12,000 IOP/s

The EXP5060 will GA by EOY 2009. It supports up to 60 disks per 4U enclosure.

Limitations of the DS5100

- The DS5100 is intended as replacement for the DS4800 with similar performance (*i.e.*, at most ~= 1600 MB/s) to provide an intermediate solution between the DS3400/DS4700 and the DS5300.
- The DS5100 has an "Enhanced Performance" feature that can double its performance (*i.e.*, at most ~= 3200 MB/s)
- A DS5100 with the "Enhanced Performance" feature costs more than an equivalently configured DS5300.

DCS9900

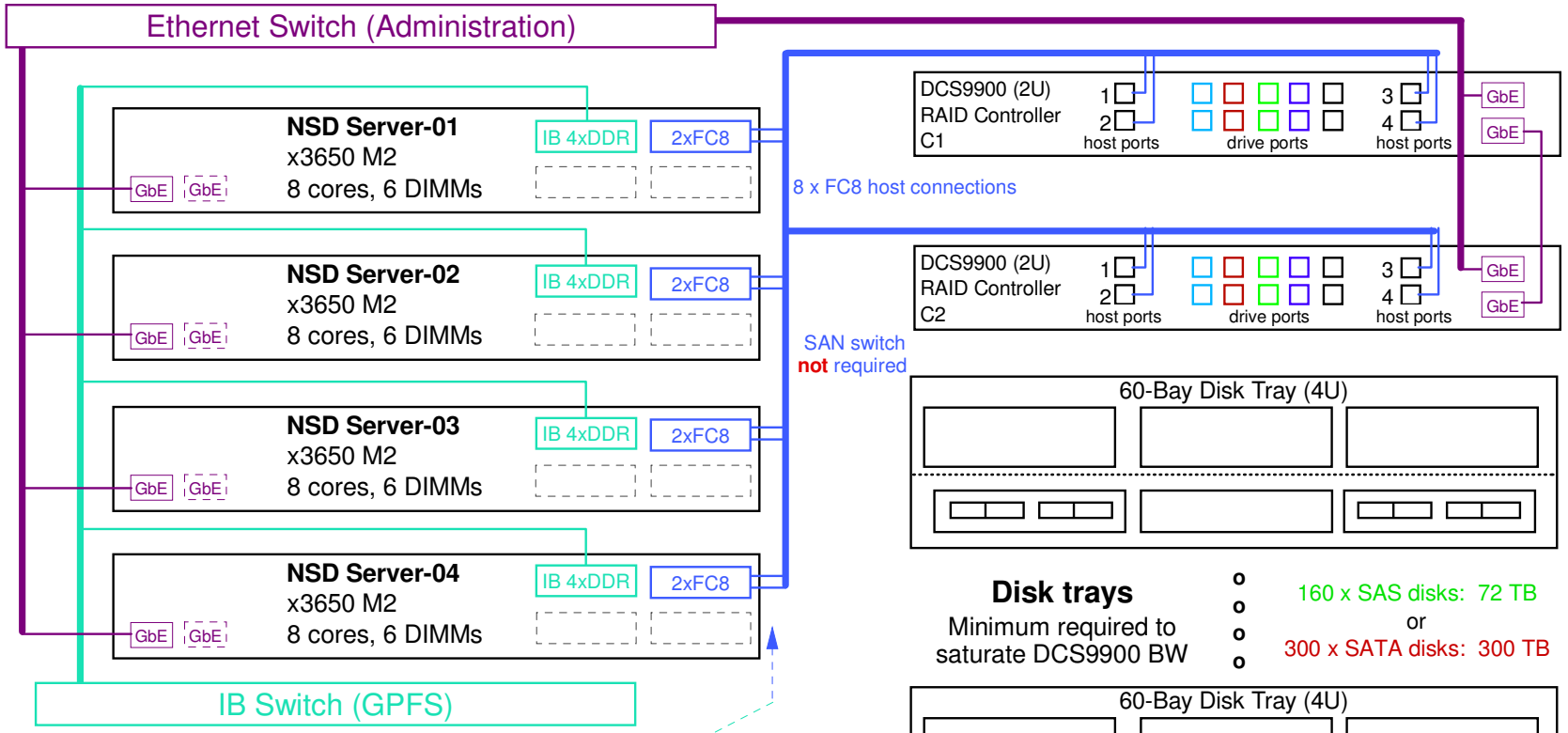
Couplet Front View

Rear View of "Half of a Couplet"

10 Disk Enclosures per Frame
60 Disks per Enclosure
4u per Enclosure

DCS9900

Example DCS9900 Building Block



Disk trays
Minimum required to saturate DCS9900 BW

o
o
o
o

160 x SAS disks: 72 TB
or
300 x SATA disks: 300 TB

The 2xFC8 HBAs can be replaced by dual port 4xDDR IB HCAs using SRP. The IB host ports can either be directly attached to the servers or connected to a dedicated IB SAN switch. It is also possible to use an IB switch for a combined LAN and SAN, but this has been discouraged in the past. As a best practice, it is not recommend to use an IB SAN for more than 32 ports.

COMMENT:
More disks (for a total of 1200) can be added to this solution but it will **not** increase performance.

Peak sustained DCS9900 performance
- streaming data rate < **5.6 GB/s**
- noncached IOP rate < **40,000 IOP/s**

4xDDR IB HCA (Host Channel Adapter)
- Potential peak data rate per HCA < 1500 MB/s
- Required peak data rate per HCA < 1400 MB/s

2xFC8 (dual port 8 Gbit/s Fibre Channel)
- Potential peak data rate per 2xFC8 < 1560 MB/s
- Required peak data rate per 2xFC8 < 1400 MB/s

DCS9900
Benchmark Results

GPFS Parameters
- blocksize(streaming) = 4096K
- blocksize(IOP) = 256K
- pagepool = 1G
- maxMBpS = 4000

COMMENT
The disparity between read and write performance observed below is much less pronounced when using 15Krpm SAS drives. For example, using 160 SAS tiers...

write ~= 5700 MB/s, read ~= 4400 MB/s

This disparity can be removed using cluster block allocation for SATA disk, but this not recommended.

DCS9900 Parameters
- 8+2P RAID 6
- SATA
- cache size = 1024K
- cache prefetch = 0
- cache writeback = ON

4 NSD Servers, no GPFS clients
- P6-p520, 4 cores, 4.2 GHz, 8 GB RAM
- 2xFC8

COMMENT: Benchmarks on a system configured using 160 x 15Krpm SAS drives, delivered the following streaming data rates:
- write < 5.6 GB/s
- read < 4.4 GB/s

Streaming Job
- record size = 4M
- file size = 32G
- number of tasks = 1 to 16
- access pattern = seq

IOP Job
- record size = 4K
- total data accessed = 10G
- number of tasks = 32
- access pattern = small file (4K to 16K)

Access Patt	Tier	1	4	8	16	32	64
Streaming	write (MB/s)	270	790	1400	2700	4800	5400
Streaming	read (MB/s)	220	710	1200	1600	2900	3600
IOP	write (IOP/s)	7,500	13,500	30,000	30,400	41,000	
IOP	read (IOP/s)	3,800	5,900	27,300	27,300	33,500	

Conclusion

- GPFS is a best of class product with good features and broad market acceptance.
- Properly used with careful design, it can provide a best of class storage solution.
- Please contact me with questions if you need technical assistance with GPFS.