



---

# **DS5300/GPFS Concepts, Best Practices and Performance**

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

**2 Dec 09, Version 1.1**

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 inquiries, 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.



# Outline

---

## ■ Overview

## ■ DS5300 Architecture

## ■ DS5300/EXP5000

- Architecture and Cabling
- Data Flow Analysis
- Disk to Array Mapping
- Benchmark Results

## ■ DS5300/EXP5060

- Architecture and Cabling
- Disk to Array Mapping
- Benchmark Results

## ■ Example GPFS Building Blocks Using the DS5300

## ■ Architecture of Recommended NSD Servers

## ■ Associated Spreadsheets

- The following spreadsheets can be provided upon request.
- BM\_results\_v1.1.xls
  - detailed performance measurements
- EXP5060 Script Builder
  - Excel spreadsheet tool for the DS5300/EXP5060 that can be used to plan disk to array mappings and generate a script that can be used to build the EXP5060 arrays.
  - Two options
    - EXP5060\_Script\_Builder\_4\_Trays\_24\_LUNs\_Trunking.xls
    - EXP5060\_Script\_Builder\_8\_Trays\_48\_LUNs\_Trunking.xls



# Overview

## ■ Report Goal

- Technical report on the use of the DS5300 with GPFS for an HPC environment
- Provides a conceptual understanding and performance expectations under best practice assumptions

## ■ DS5300 Provides an Effective and Flexible HPC Storage Solution under GPFS

- Balanced<sup>1</sup> Solution Emphasizing Performance
  - DS5300 with 8 x EXP5000 trays (usable capacity < 43 TB over 128 x 15Krpm FC disks)
  - write < 4.4 GB/s
    - capacity:performance ratio = 107 MB/s / TB<sup>2</sup>
  - read < 5.4 GB/s
    - capacity:performance ratio = 132 MB/s / TB<sup>2</sup>
- Balanced<sup>1</sup> Solution Emphasizing Capacity
  - DS5300 with 4 x EXP5060 trays (usable capacity < 192 TB over 240 x 7200 RPM SATA disks)
  - write < 4.0 GB/s
    - capacity:performance = 21 MB/s / TB<sup>3</sup>
  - read < 4.8 GB/s
    - capacity:performance = 26 MB/s / TB<sup>3</sup>
- Disclaimer
  - Performance results are based benchmarks using configurations appropriate for most production environments; specialized tuning techniques (e.g., clustered GPFS allocation maps, direct I/O, short stroking) aimed at peak performance, but of limited generality, are not used. While synthetic benchmark codes are used, the benchmark measurements are obtained using code instrumentation. Thus these measurements generally represent best case results under realistic configurations. So application codes properly written for storage I/O can be expected to achieve these results, though given the varied I/O profiles common in most production environments actual performance may be less.

### Footnotes:

1. These are "balanced" solutions in the sense that the number of disks used are the minimum needed to achieve peak streaming performance; adding more disks will not significantly increase streaming performance.
2. Measurement based on usable capacity with 450 GB/disk.
3. Measurement based on usable capacity with 1 TB/disk

Performance  
will vary  
"according to  
actual driving  
conditions".

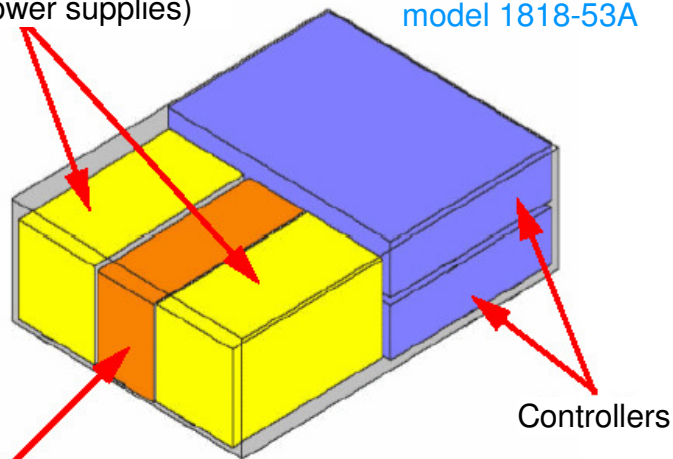




## DS5300

Controller Support Modules  
(Fans, power supplies)

DS5300  
model 1818-53A

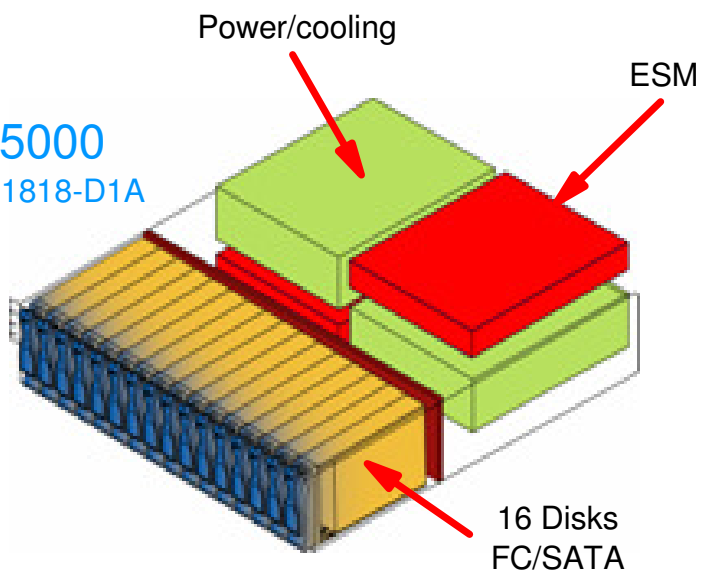


Interconnect Module  
(batteries, midplane)

Controllers

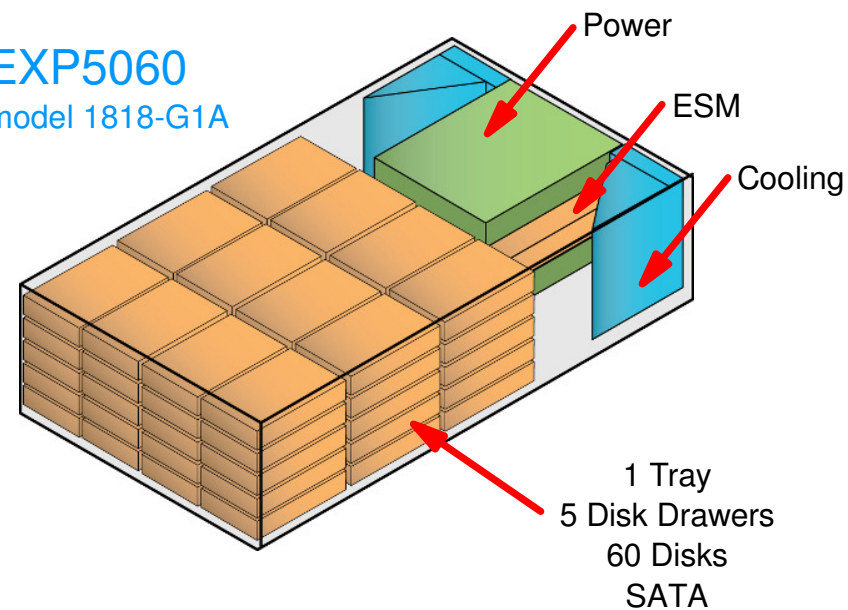


EXP5000  
model 1818-D1A



16 Disks  
FC/SATA

EXP5060  
model 1818-G1A

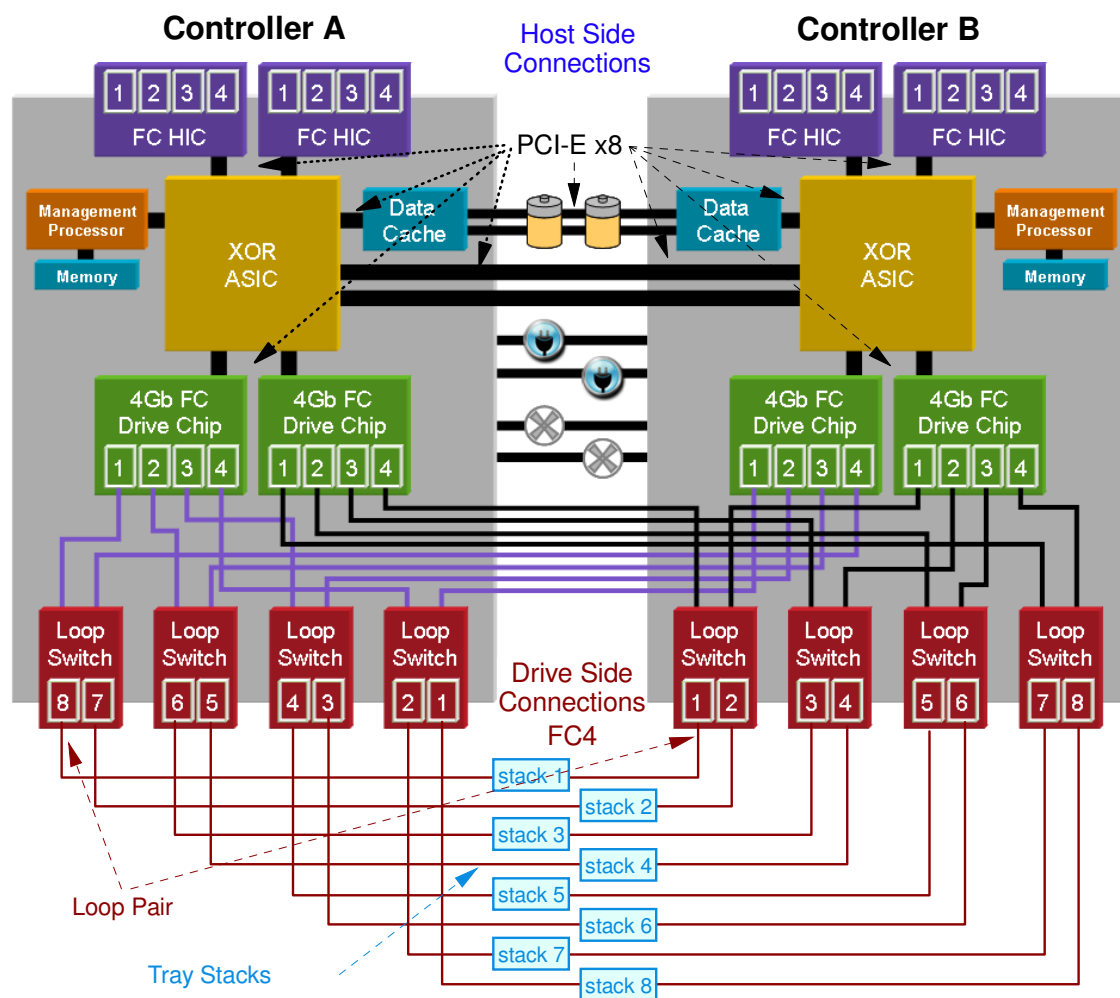


1 Tray  
5 Disk Drawers  
60 Disks  
SATA



# DS5300

## Architectural Overview



A **loop pair** is a set of redundant drive side cables as shown in this diagram. A **stack** is a set of enclosures along a loop pair.

### FOOTNOTES:

- ★ Best practice: Use the EXP5000 for FC disk and the EXP5060 for SATA disk.
- † These are upper bound rates based on lab measurements using specialized tuning parameters and workload assumptions. They demonstrate what the DS5300 can do. Actual performance rates will not exceed these values.

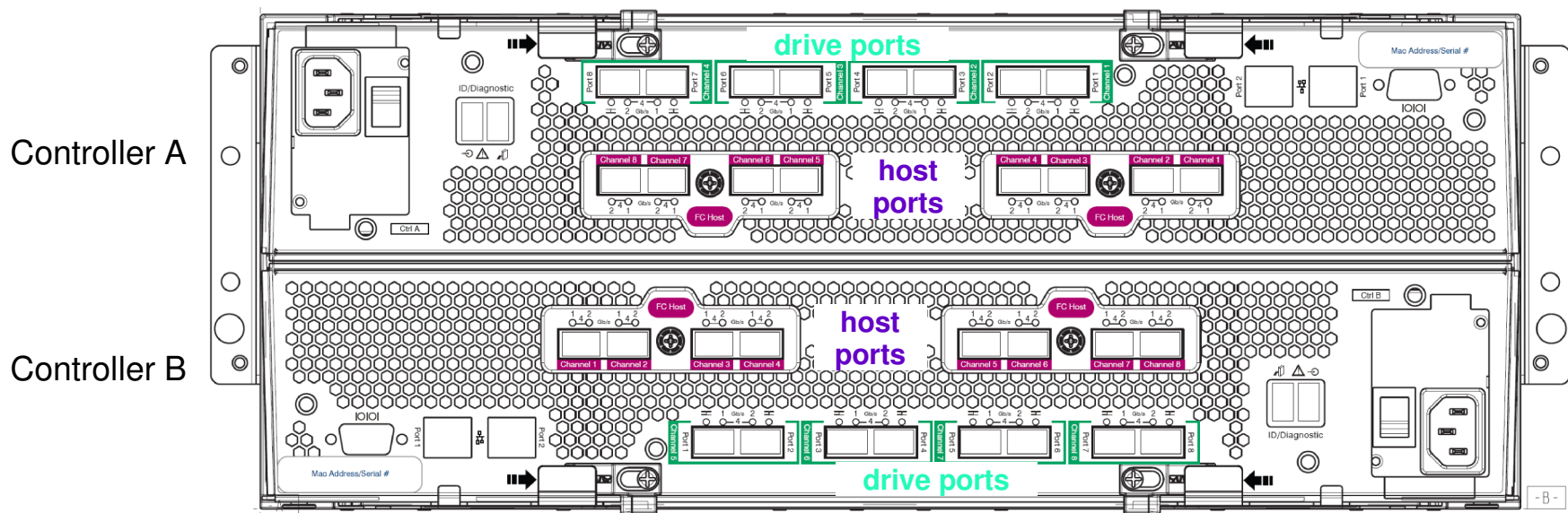
- Dual, redundant RAID controllers
- Dual, redundant power, battery backup and fans
- Internal busses (theoretical)
  - ▶ PCI-E x8 simplex rate = 2 GB/s
- Host side connections (measured)
  - ▶ 16 host-side connections
    - FC4 < 380 MB/s
    - FC8 < 760 MB/s
  - ▶ Active/passive architecture
  - ▶ FCAL
- Drive side connections (measured)
  - ▶ 16 drive-side connections
    - FC4 < 380 MB/s
- Supported trays
  - ▶ EXP5000 (4 Gb/s) - FC/SATA\*
    - up to 28 trays, 448 disks
  - ▶ EXP5060 (4 Gb/s) - SATA only
    - up to 8 trays, 480 disks
- Disk Technology
  - ▶ 15 Krpm FC disk (300, 450 GB)
  - ▶ SATA (750, 1000 GB)
- Peak sustained rates (theoretical<sup>†</sup>)
  - ▶ Streaming rate (144 x 15Krpm FC disk)
    - write < 5.4 GB/s
    - read < 6.4 GB/s
  - ▶ IOP Rate
    - To cache: 700,000 IOP/s (512 B)
    - To media (448 x 15Krpm FC disk)
      - write < 105,000 IOP/s (4096 B)
      - read < 140,000 IOP/s (4096 B)



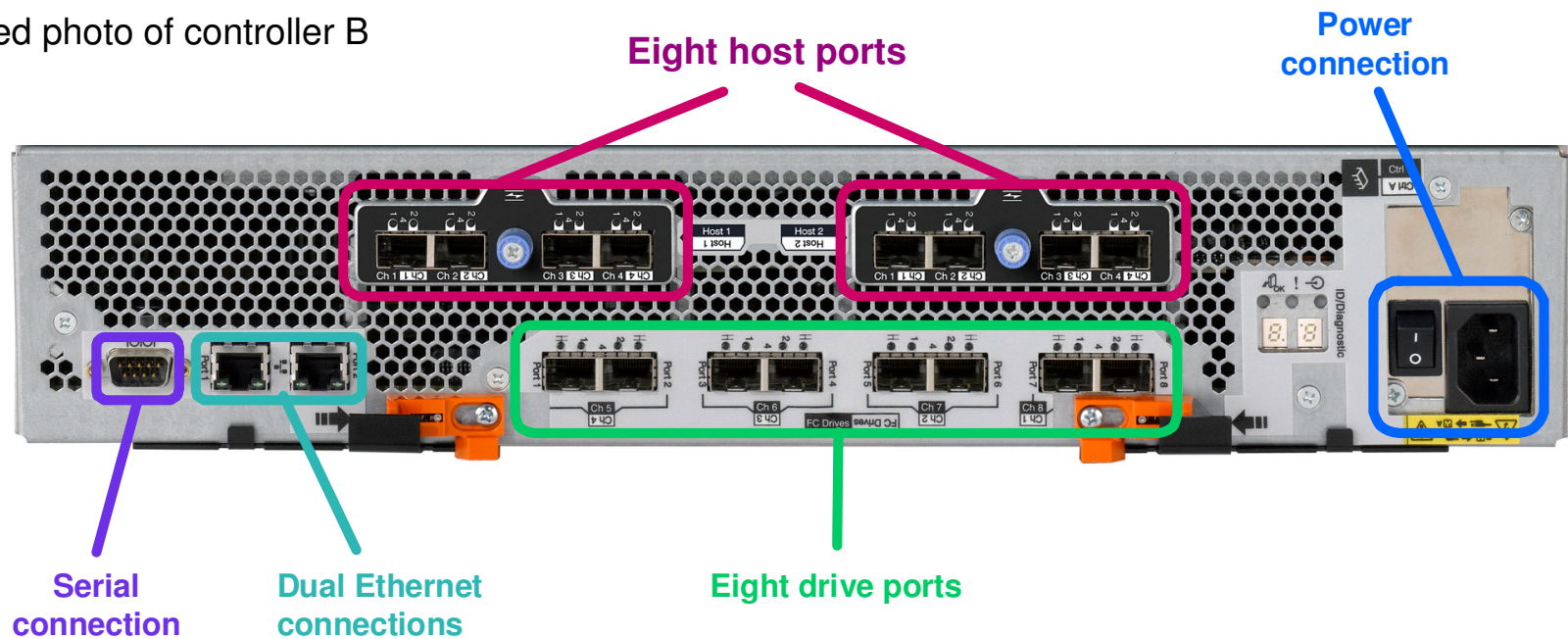


# DS5300

## Rearview



Annotated photo of controller B





## DS5300

### Cabling and Disk to Array Mapping

---

Careful attention must be given to drive side cabling and disk to array mapping on the DS5300 in order to guarantee optimum streaming performance. This issue is less significant for IOP performance.

There are a lot of ways to get it right, but there are also a lot of ways to get it wrong!

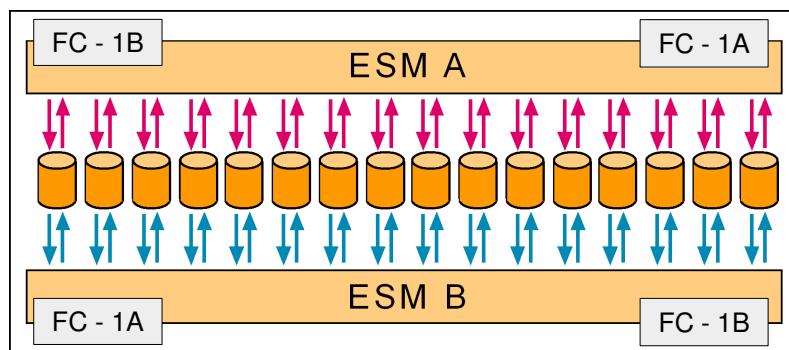
#### **WARNINGS:**

- ▶ Default array mappings (e.g., created by SMClient) are *not* guaranteed to be optimum!
- ▶ Rules and best practices for the DS4800 do *not always* apply to the DS5300.





## DS5300/EXP5000



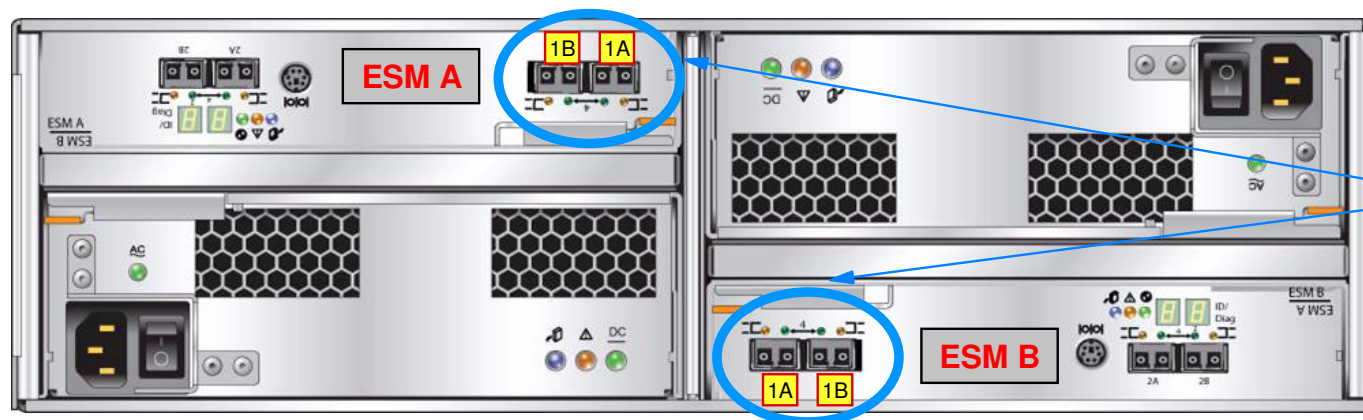
logical layout

- ▶ 16 drives in a 3U tray
- ▶ 2 ESMs per tray
- ▶ 4 x FC4 ports (2 per ESM)
  - High-speed, low-latency interconnect from controllers to drives
- ▶ Supports FC and/or SATA drives
  - Supports both FC and SATA in the same tray.
  - If SATA is preferred, consider using EXP5060 instead.
- ▶ Unique speed-matching technology
  - 3 Gb/s SATA II drives effectively run at 4 Gb/s speeds
- ▶ FC Switched architecture
  - Higher performance, lower latency
  - Drive isolation, better diagnostics
- ▶ RoHS compliant

### ESM - Enclosure Service Module

- ▶ ESM A is the primary path for the odd drives
- ▶ ESM B is the primary path for the even drives

If an ESM fails, the other ESM can access all of the drives.



Only use highlighted ports  
(EXP5000 does not support "trunking")

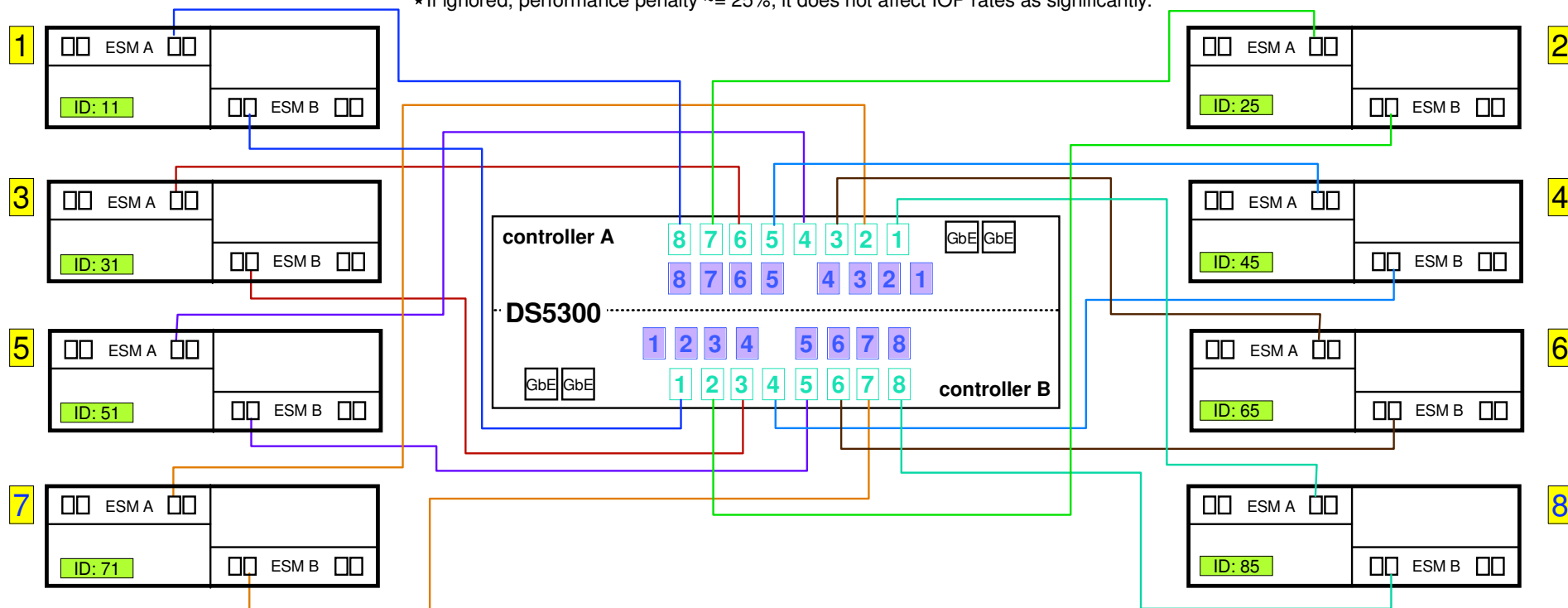


# DS5300/EXP5000

## Drive Side Cabling - 8 Trays

**Balance\*:** Optimum streaming performance is achieved using a multiple of 8 x EXP5000 trays with the same number of trays per stack. Optimum performance is achieved using 8, 16 or 24 trays.

\* If ignored, performance penalty  $\sim 25\%$ ; it does not affect IOP rates as significantly.



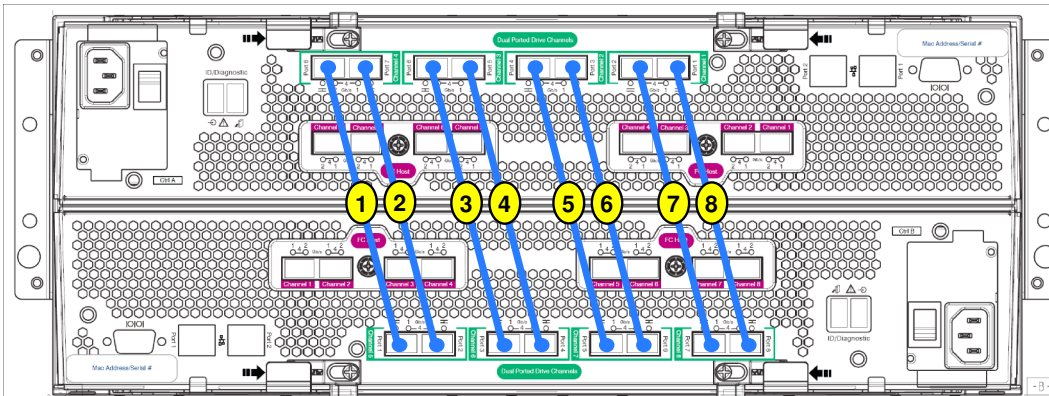
### Stacks

When attaching enclosures, drive loops are configured as redundant pairs (*i.e.*, loop pairs) utilizing one port from each controller; the enclosures along a loop pair are called a stack.

### Tray ID

Tray ID is assigned during system configuration. The values are not arbitrary. **Best practice:**

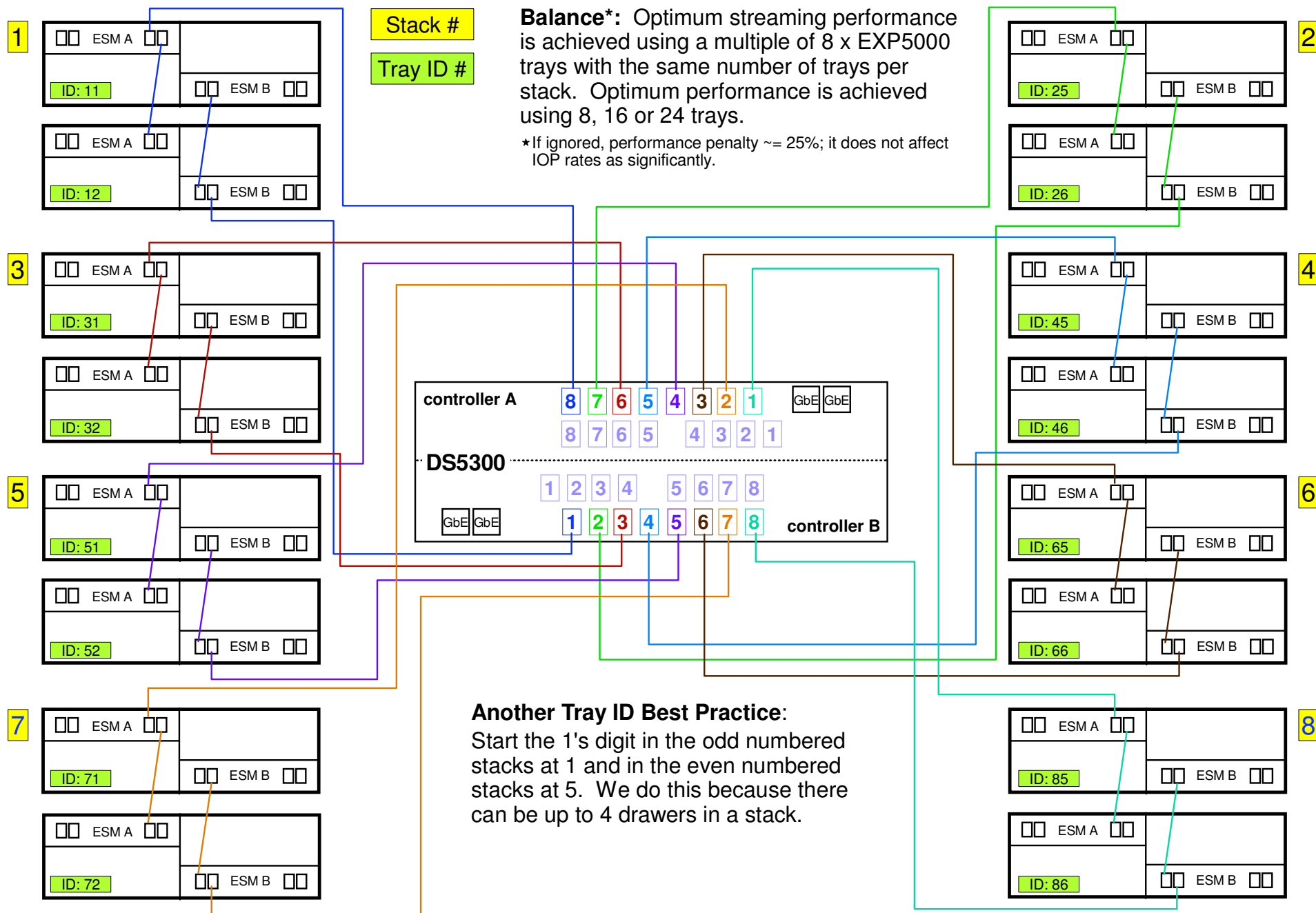
- ▶ 10's digit: stack number
- ▶ 1's digit: ordinal number within a stack





# DS5300/EXP5000

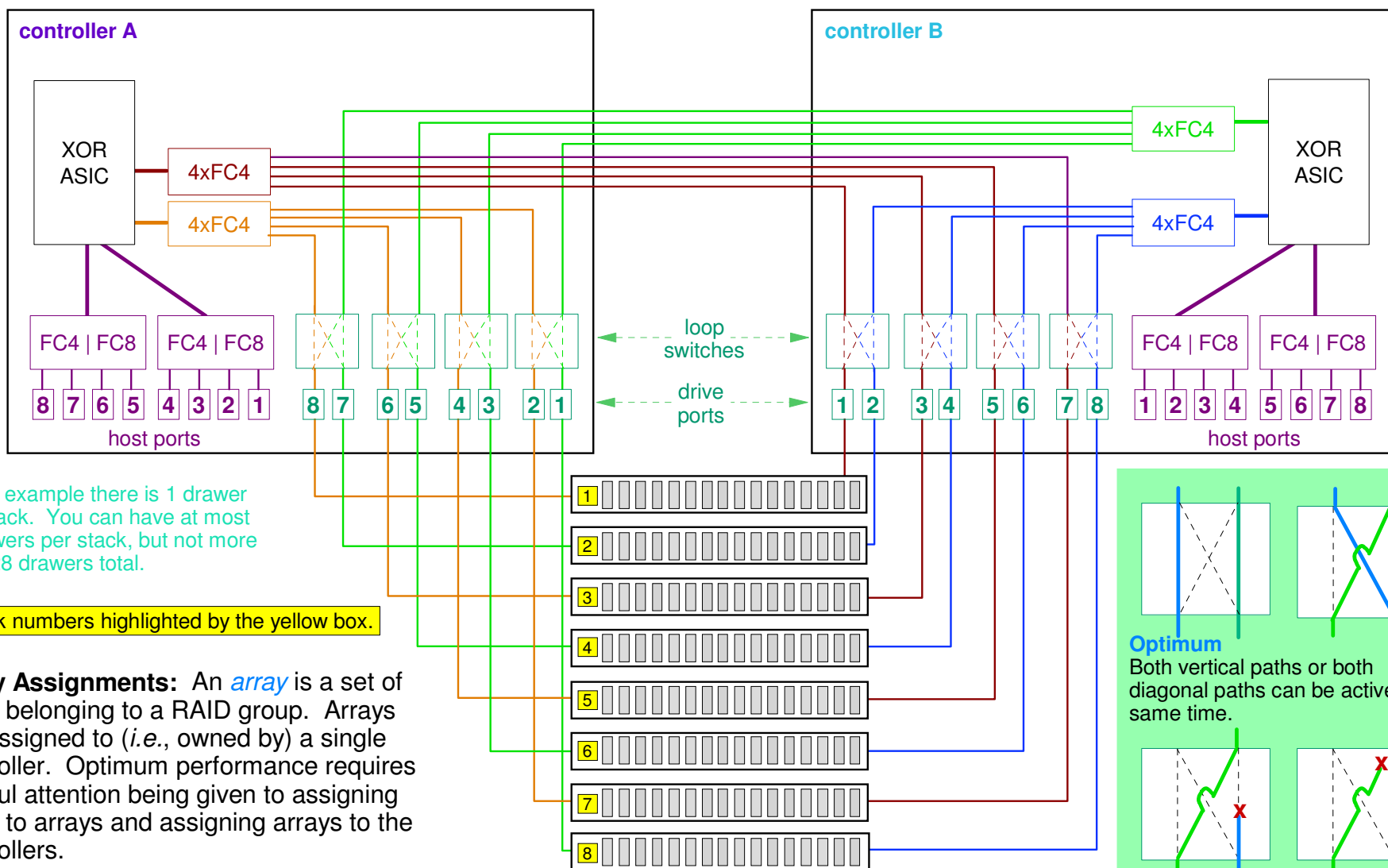
## Drive Side Cabling - 16 Trays





# DS5300/EXP5000

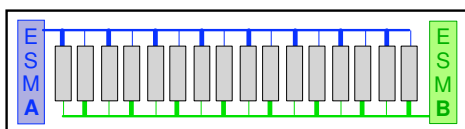
## Drive Side Cabling and Disk to Array Mapping



In this example there is 1 drawer per stack. You can have at most 4 drawers per stack, but not more than 28 drawers total.

Stack numbers highlighted by the yellow box.

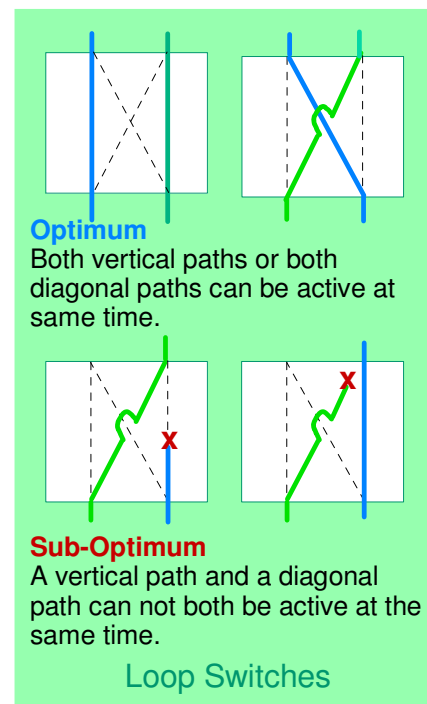
**Array Assignments:** An *array* is a set of disks belonging to a RAID group. Arrays are assigned to (*i.e.*, owned by) a single controller. Optimum performance requires careful attention being given to assigning disks to arrays and assigning arrays to the controllers.



**Remember:** by default

- ESM-A accesses odd disks
- ESM-B accesses even disks

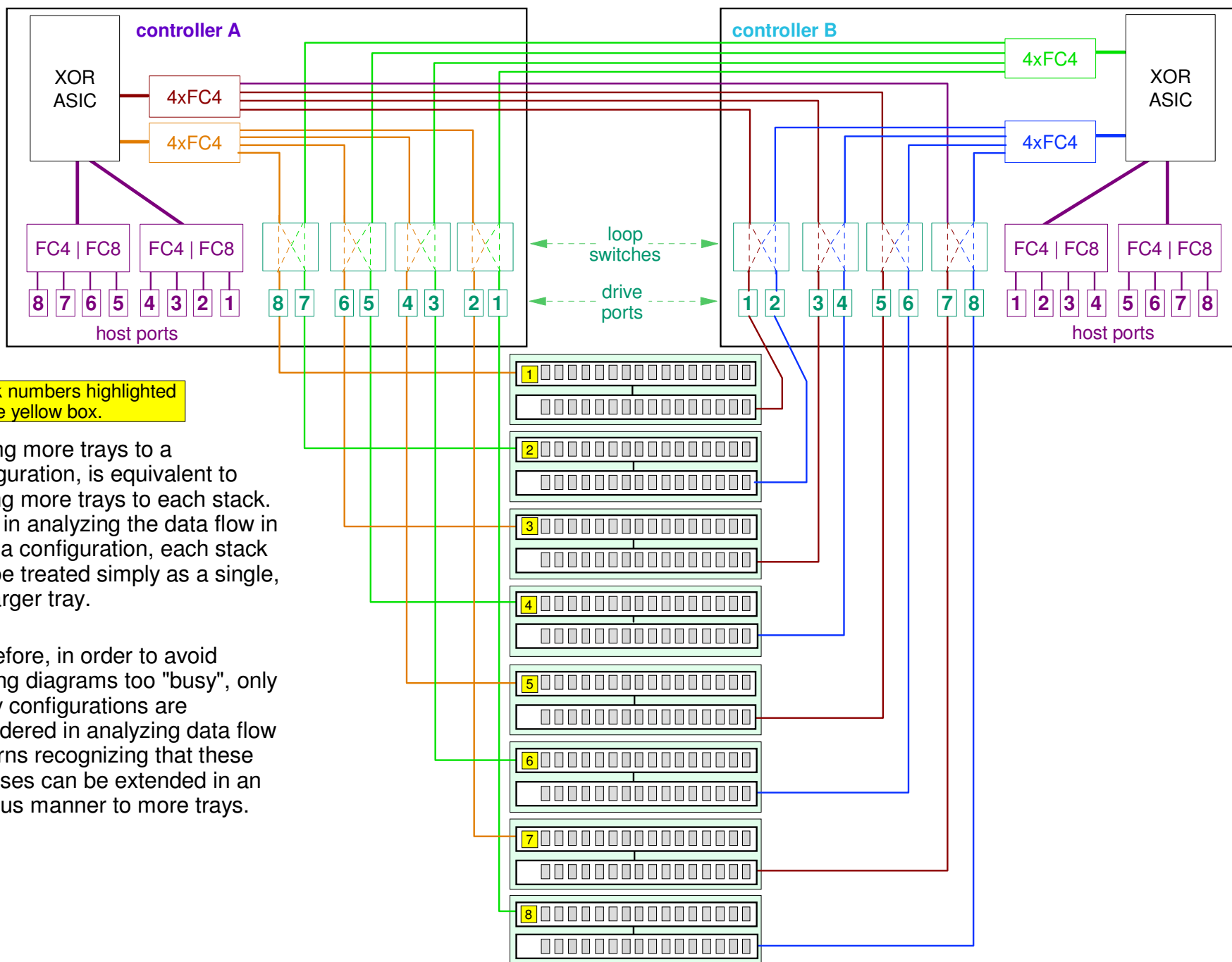
This is independent of controller preference.





# DS5300/EXP5000

## Drive Side Cabling and Disk to Array Mapping - Adding More Trays



Stack numbers highlighted by the yellow box.

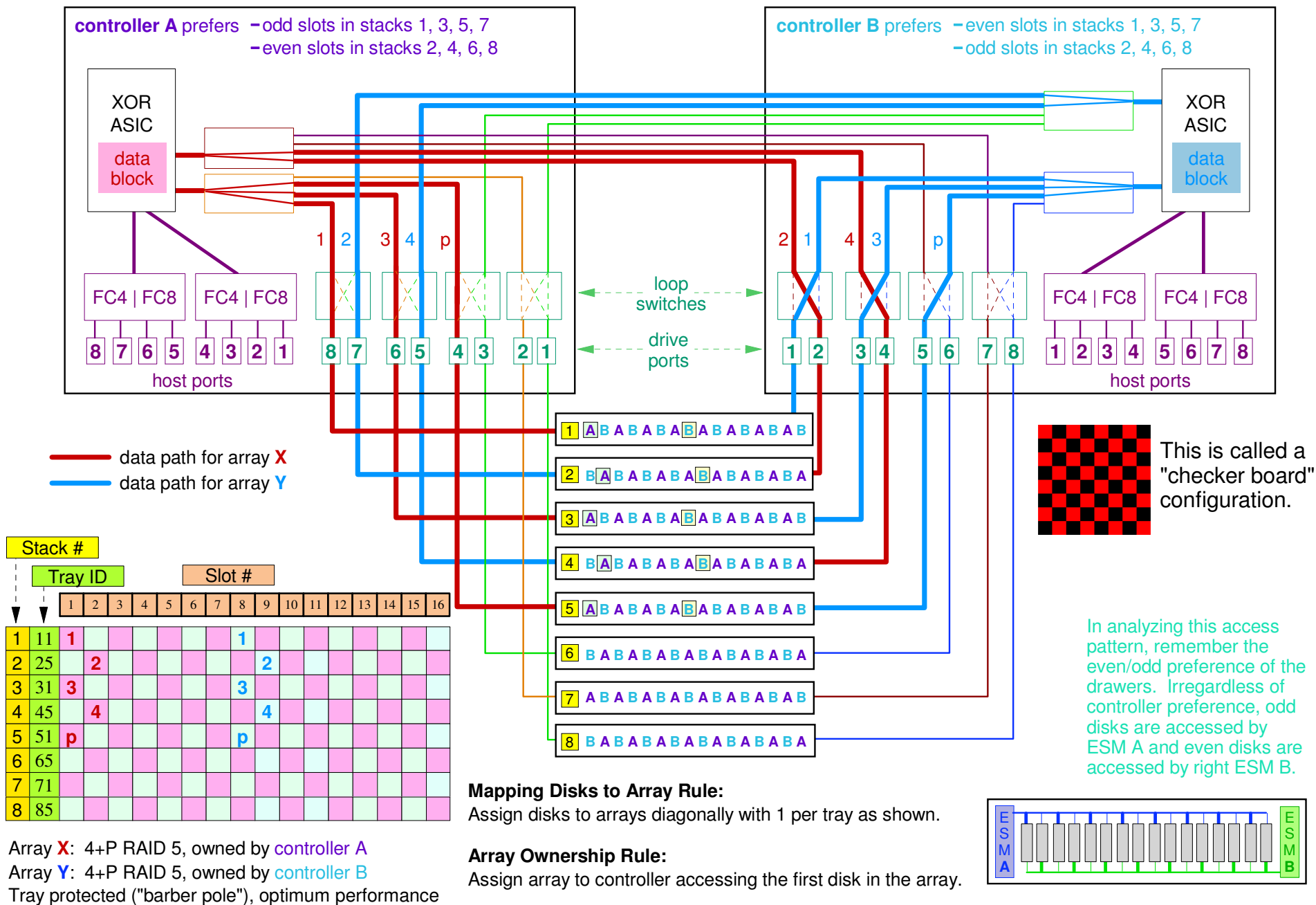
Adding more trays to a configuration, is equivalent to adding more trays to each stack. Thus in analyzing the data flow in such a configuration, each stack can be treated simply as a single, but larger tray.

Therefore, in order to avoid making diagrams too "busy", only 8 tray configurations are considered in analyzing data flow patterns recognizing that these analyses can be extended in an obvious manner to more trays.



# DS5300/EXP5000

## Data Flow Example #1







# DS5300/EXP5000

## Sample Disk to Array Mappings Using Example #1



### 4+P RAID 5, Tray Protected with optimum performance

11	A1	B2	A3	B4	A7	B8	A9	B10	A13	B14	A17	B18	A19	B20	HS	B24
25	B2	A1	B4	A3	B8	A7	B12	A11	B14	A13	B18	A17	B20	A19	HS	A23
31	A1	B2	A5	B6	A7	B8	A11	B12	A13	B14	A17	B18	A21	B22	A23	HS
45	B2	A1	B6	A5	B8	A7	B12	A11	B16	A15	B18	A17	B22	A21	B24	HS
51	A1	B2	A5	B6	A9	B10	A11	B12	A15	B16	A17	B18	A21	B22	HS	A23
65	B4	A3	B6	A5	B10	A9	B12	A11	B16	A15	B20	A19	B22	A21	HS	B24
71	A3	B4	A5	B6	A9	B10	A13	B14	A15	B16	A19	B20	A21	B22	A23	HS
85	B4	A3	B8	A7	B10	A9	B14	A13	B16	A15	B20	A19	B24	A23	B24	HS

A<int>:

Arrays owned  
by controller A

B<int>:

Arrays owned  
by controller B

HS: Hot Spare

X: Extra

### 8+2P RAID 6, Tray Protected with optimum performance

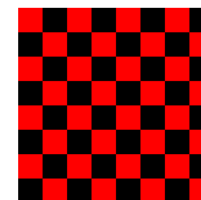
11	A1	B2	A5	B6	A9	B10	A13	B14	A17	B18	A21	B22	A1	B2	X	B18
12	A3	B4	A7	B8	A11	B12	A15	B15	A19	B20	A23	B24	A3	B4	A19	X
25	B2	A1	B6	A5	B10	A9	B14	A13	B18	A17	B22	A21	B2	A1	X	A17
26	B4	A3	B8	A7	B12	A11	B15	A15	B20	A19	B24	A23	B4	A3	B20	X
31	A1	B2	A5	B6	A9	B10	A13	B14	A17	B18	A21	B22	A5	B6	X	B18
32	A3	B4	A7	B8	A11	B12	A15	B15	A19	B20	A23	B24	A7	B8	A19	X
45	B2	A1	B6	A5	B10	A9	B14	A13	B18	A17	B22	A21	B6	A5	X	A17
46	B4	A3	B8	A7	B12	A11	B15	A15	B20	A19	B24	A23	B8	A7	B20	X
51	A1	B2	A5	B6	A9	B10	A13	B14	A17	B18	A21	B22	A9	B10	X	B22
52	A3	B4	A7	B8	A11	B12	A15	B15	A19	B20	A23	B24	A11	B12	A23	X
65	B2	A1	B6	A5	B10	A9	B14	A13	B18	A17	B22	A21	B10	A9	X	A21
66	B4	A3	B8	A7	B12	A11	B15	A15	B20	A19	B24	A23	B12	A11	B24	X
71	A1	B2	A5	B6	A9	B10	A13	B14	A17	B18	A21	B22	A13	B14	X	B22
72	A3	B4	A7	B8	A11	B12	A15	B15	A19	B20	A23	B24	A15	B15	A23	X
85	B2	A1	B6	A5	B10	A9	B14	A13	B18	A17	B22	A21	B14	A13	X	A21
86	B4	A3	B8	A7	B12	A11	B15	A15	B20	A19	B24	A23	B15	A15	B24	X

Disks labeled X are "extra". Traditionally they would have been used as "hot spares", but there is little need for hot spares under RAID 6. These disks could be configured as 2 x 4+4 RAID 10 arrays.

**Best Practice:** Adopt tray protection using the following configurations.

8 trays using 4+P RAID 5 or 4+2P RAID 6

16 trays using 4+P or 8+P RAID 5, or 8+2P RAID 6

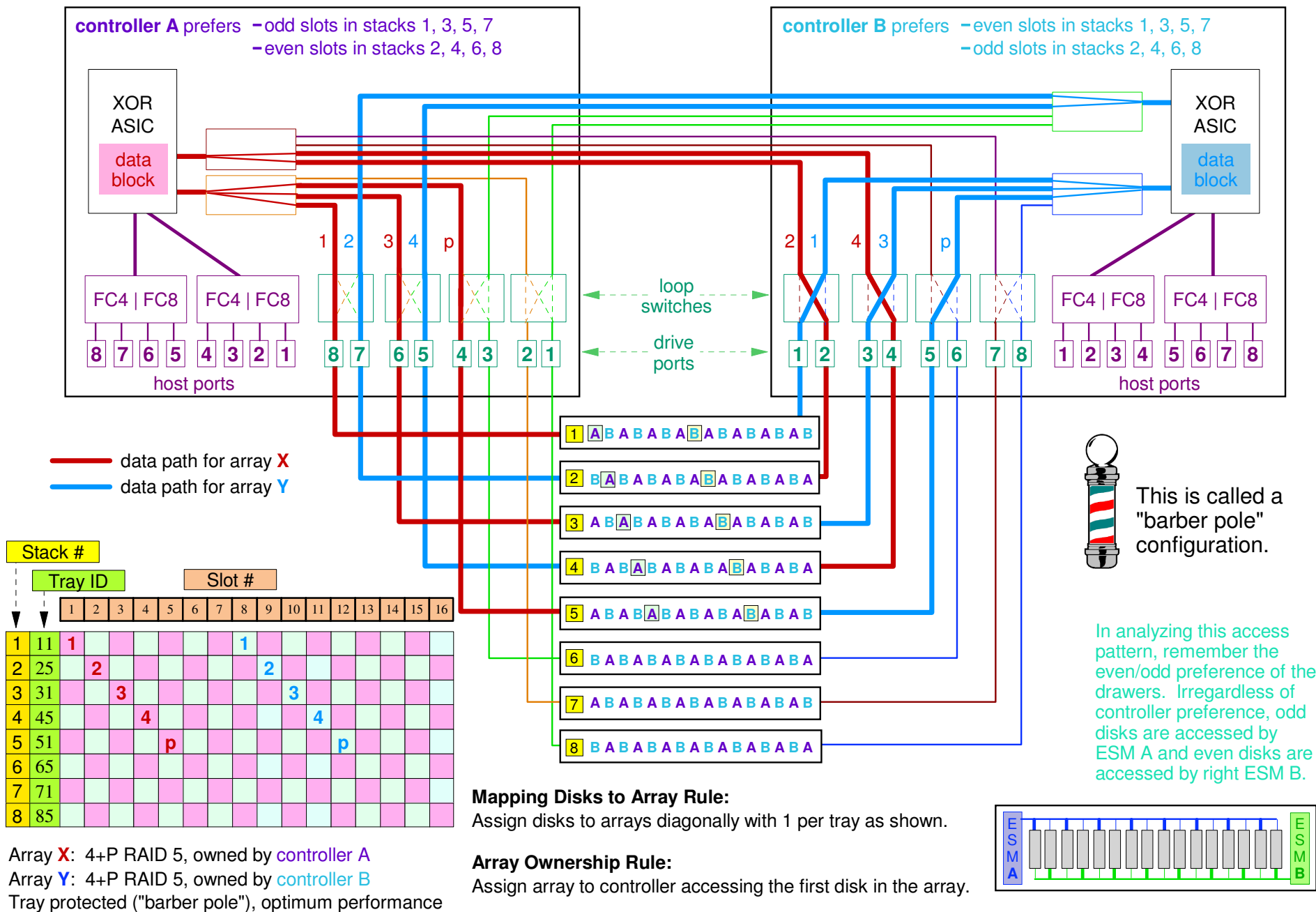


This is called a "checker board" configuration.



# DS5300/EXP5000

## Data Flow Example #2





# DS5300/EXP5000

## Sample Disk to Array Mappings Using Example #2



### 4+P RAID 5, Tray Protected with optimum performance

11	A1	B2	A3	B4	A7	B8	A9	B10	A13	B14	A17	HS	A19	B20	A23	B24
25	B24	A1	B2	A3	B4	A7	B8	A11	B12	A13	B14	A17	HS	A19	B20	A23
31	A23	B24	A1	B2	A5	B6	A7	B8	A11	B12	A13	B14	A17	HS	A21	B22
45	B22	A23	B18	A1	B2	A5	B6	A7	B8	A11	B12	A15	B16	A17	HS	A21
51	A21	B22	HS	B24	A1	B2	A5	B6	A9	B10	A11	B12	A15	B16	A17	B18
65	B20	A21	B22	HS	B18	A3	B4	A5	B6	A9	B10	A11	B12	A15	B16	A19
71	A19	B20	A21	B22	HS	B18	A3	B4	A5	B6	A9	B10	A13	B14	A15	B16
85	B16	A19	B20	A23	B24	HS	B18	A3	B4	A7	B8	A9	B10	A13	B14	A15

A<int>:  
Arrays owned  
by controller A

B<int>:  
Arrays owned  
by controller B

HS: Hot Spare  
X: Extra

### 8+2P RAID 6, Tray Protected with optimum performance

11	A1	B22	A5	B10	A13	B18	A17	B2	A21	X	A1	B6	A9	B14	A17	B22
12	A3	B24	A7	B12	A15	B20	A19	B4	A23	X	A3	B8	A11	B16	A19	B24
25	B22	A1	B10	A5	B22	A13	B2	A21	B14	X	B6	A1	B18	A9	B6	A17
26	B24	A3	B12	A7	B24	A15	B4	A23	B16	X	B8	A3	B20	A11	B8	A19
31	A17	B10	A1	B22	A5	B2	A13	B14	A21	B6	A21	X	A5	B18	A9	B18
32	A19	B12	A3	B24	A7	B4	A15	B16	A23	B8	A23	X	A7	B20	A11	B20
45	B10	A17	B18	A1	B2	A5	B14	A13	B6	A21	B18	X	B6	A5	B22	A9
46	B12	A19	B20	A3	B4	A7	B16	A15	B8	A23	B20	X	B8	A7	B24	A11
51	A9	B14	A17	B2	A1	B14	A9	B6	A13	B22	A21	B18	A21	X	A5	B10
52	A11	B16	A19	B4	A3	B16	A11	B8	A15	B24	A23	B20	A23	X	A7	B12
65	B14	A9	B2	A17	B14	A1	B6	A9	B18	A13	B2	A21	B22	X	B10	A5
66	B16	A11	B4	A19	B16	A3	B8	A11	B20	A15	B4	A23	B24	X	B12	A7
71	A5	B2	A13	B14	A17	B6	A1	B10	A9	B18	A13	B22	A21	B10	A17	X
72	A7	B4	A15	B16	A19	B8	A3	B12	A11	B20	A15	B24	A23	B12	A19	X
85	B2	A5	B14	A13	B6	A17	B10	A1	B2	A9	B22	A13	B10	A21	B18	X
86	B4	A7	B16	A15	B8	A19	B12	A3	B4	A11	B24	A15	B12	A23	B20	X

Disks labeled X are "extra". Traditionally they would have been used as "host spares", but there is little need for hot spares under RAID 6. These disks could be configured as 2 x 4+4 RAID 10 arrays.

**Best Practice:** Adopt tray protection using the following configurations.

8 trays using 4+P RAID 5 or 4+2P RAID 6

16 trays using 4+P or 8+P RAID 5, or 8+2P RAID 6

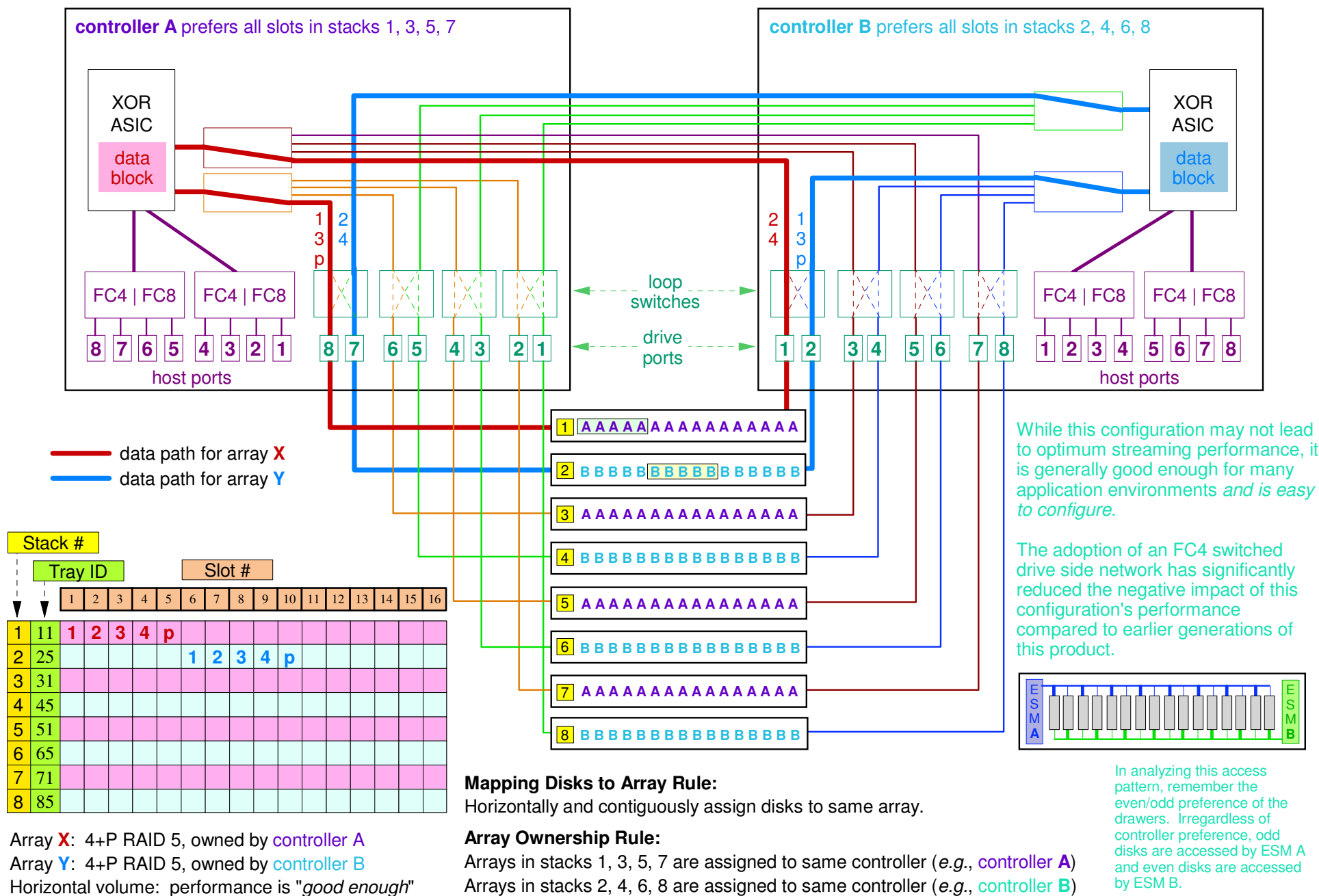


This is called a "barber pole" configuration.



# DS5300/EXP5000

## Data Flow Example #3





# DS5300/EXP5000

## Sample Disk to Array Mappings for Example #3



### 4+P RAID 5, Stack Oriented with "good enough" performance

11	A1	A1	A1	A1	A1	A3	A3	A3	A3	A3	A5	A5	A5	A5	A5	HS
25	B2	B2	B2	B2	B2	B4	B4	B4	B4	B4	B6	B6	B6	B6	B6	HS
31	A7	A7	A7	A7	A7	A9	A9	A9	A9	A9	A11	A11	A11	A11	A11	HS
45	B8	B8	B8	B8	B8	B10	B10	B10	B10	B10	B12	B12	B12	B12	B12	HS
51	A13	A13	A13	A13	A13	A15	A15	A15	A15	A15	A17	A17	A17	A17	A17	HS
65	B14	B14	B14	B14	B14	B16	B16	B16	B16	B16	B18	B18	B18	B18	B18	HS
71	A19	A19	A19	A19	A19	A21	A21	A21	A21	A21	A23	A23	A23	A23	A23	HS
85	B20	B20	B20	B20	B20	B22	B22	B22	B22	B22	B24	B24	B24	B24	B24	HS

A<int>:  
Arrays owned  
by controller A

B<int>:  
Arrays owned  
by controller B

HS: Hot Spare  
X: Extra

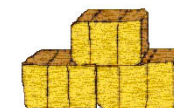
### 8+2P RAID 6, Stack Oriented with "good enough" performance

11	A1	A1	A1	A1	A1	A3	A3	A3	A3	A3	A5	A5	A5	A5	A5	X
12	A1	A1	A1	A1	A1	A3	A3	A3	A3	A3	A5	A5	A5	A5	A5	X
25	B2	B2	B2	B2	B2	B4	B4	B4	B4	B4	B6	B6	B6	B6	B6	X
26	B2	B2	B2	B2	B2	B4	B4	B4	B4	B4	B6	B6	B6	B6	B6	X
31	A7	A7	A7	A7	A7	A9	A9	A9	A9	A9	A11	A11	A11	A11	A11	X
32	A7	A7	A7	A7	A7	A9	A9	A9	A9	A9	A11	A11	A11	A11	A11	X
45	B8	B8	B8	B8	B8	B10	B10	B10	B10	B10	B12	B12	B12	B12	B12	X
46	B8	B8	B8	B8	B8	B10	B10	B10	B10	B10	B12	B12	B12	B12	B12	X
51	A13	A13	A13	A13	A13	A15	A15	A15	A15	A15	A17	A17	A17	A17	A17	X
52	A13	A13	A13	A13	A13	A15	A15	A15	A15	A15	A17	A17	A17	A17	A17	X
65	B14	B14	B14	B14	B14	B16	B16	B16	B16	B16	B18	B18	B18	B18	B18	X
66	B14	B14	B14	B14	B14	B16	B16	B16	B16	B16	B18	B18	B18	B18	B18	X
71	A19	A19	A19	A19	A19	A21	A21	A21	A21	A21	A23	A23	A23	A23	A23	X
72	A19	A19	A19	A19	A19	A21	A21	A21	A21	A21	A23	A23	A23	A23	A23	X
85	B20	B20	B20	B20	B20	B22	B22	B22	B22	B22	B24	B24	B24	B24	B24	X
86	B20	B20	B20	B20	B20	B22	B22	B22	B22	B22	B24	B24	B24	B24	B24	X

Disks labeled X are "extra". Traditionally they would have been used as "host spares", but there is little need for hot spares under RAID 6. These disks could be configured as 2 x 4+4 RAID 10 arrays.

In this case, however, it would be best to arrange the extra disks in a more optimal pattern, using a "barber pole" configuration for example.

**WARNING:** While stack oriented configurations are simple and generally provide good enough performance, they do **not** provide tray protection.



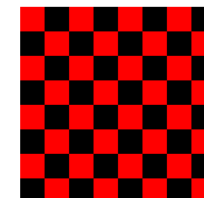
A simple, but good enough solution.



# Comparing Array Mapping Strategies

## ■ Checker Board

- Optimal streaming performance<sup>1</sup>
- Tray protection
- Simpler to configure than the barber poll strategy



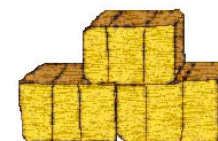
## ■ Barber Poll

- Optimal streaming performance<sup>1</sup>
- Tray protection
- Tedious to configure properly



## ■ Stack Oriented Configuration

- Optimal streaming performance<sup>1</sup> when using all LUNs uniformly
- No tray protection
- Simplest configuration strategy



Footnote:

1. Small record IOP performance is not as sensitive to the disk to array mapping strategies as is streaming performance.





## Notes on Disk to Array Mapping

---

Configuring arrays for optimal performance, especially for strategies preserving tray protection, can be tedious and error prone.

Here is a convenient way to do this.

1. Work out the disk to array mappings using a spread sheet.
2. Create a SMcli script to build the arrays.
  - a. SMcli is a command line tool for managing the DS5300. Scripts can be created or uploaded and executed using the Tools >> Execute Script menu from the Storage Manager GUI.
3. Validate that the disk to array mappings yield optimal performance by running a quick benchmark.
  - a. Suggestion: Setting writeZerosFlag=1 allows the arrays to be used immediately for performance testing by skipping the time consuming low level formatting step. This can only be done using the DS5300 control monitor. **But do NOT do this in production since user data will be corrupted!** Therefore, before putting the system into production, rebuild the arrays with writeZerosFlag=0.



## Notes on Using the "Extra" Disks with 8+2P RAID 6

---

In the 8+2P RAID 6 array mapping examples, 16 disks are designated as extra (X) since assigning them to yet another 8+2P array would create an imbalance.

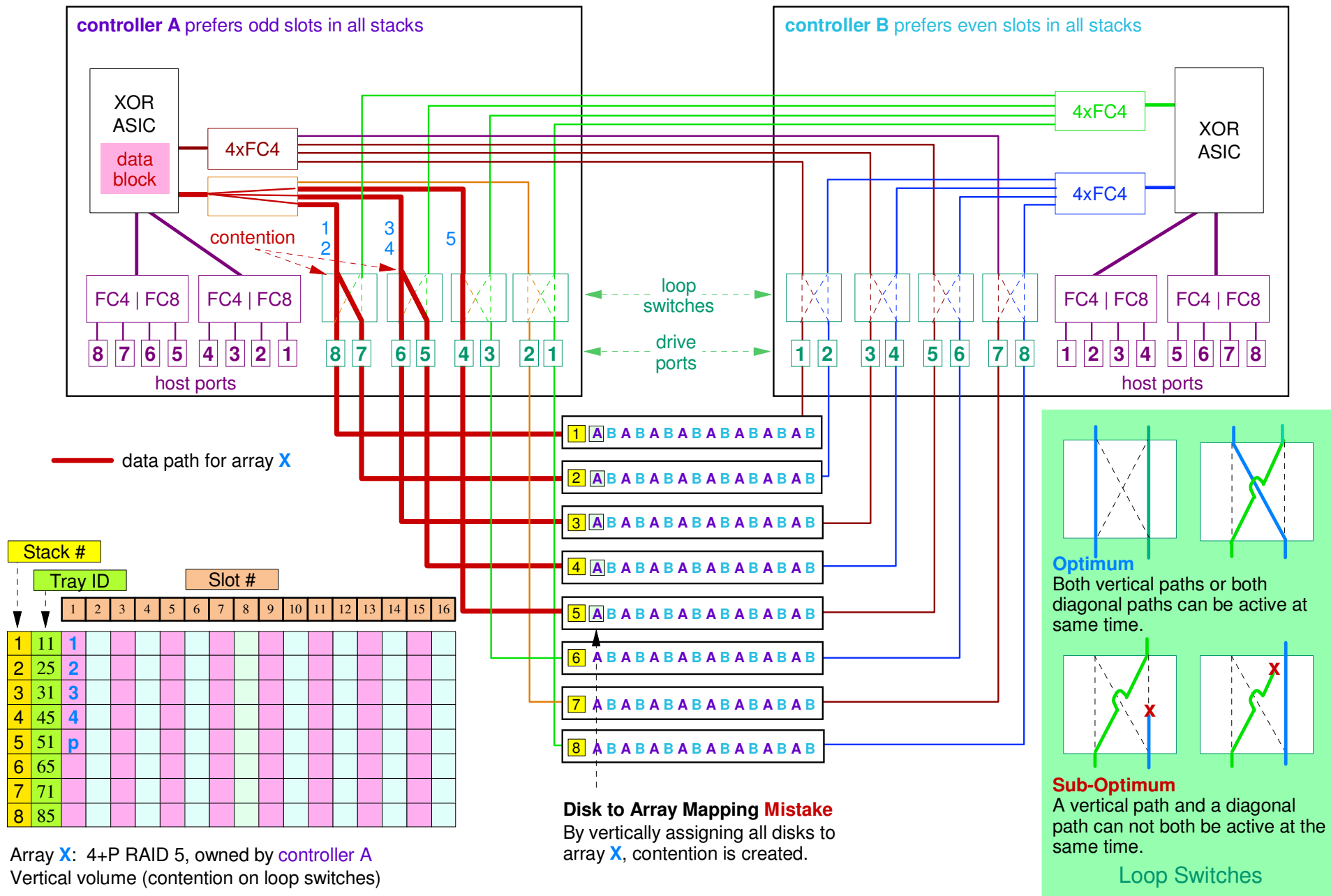
As a best practice, there is no need for hot spares using RAID 6, so an alternative use for these disks is to deploy them as metadataOnly arrays in GPFS.

However, the effectiveness of this strategy for metadata performance is an open question at this time. GPFS will get the greatest performance benefit by being able to access metadata records cached by the DS5300; therefore it will largely depend on whether there is good temporal locality for the metadata. On the other hand, for noncached accesses, if the number of metadata transactions relative to the number of user data accesses is large, then the smaller number of disks required to contain the metadata may not be adequate to sustain performance. (n.b., GPFS uses  $\sim 1.5\%$  capacity for metadata.)



# DS5300/EXP5000

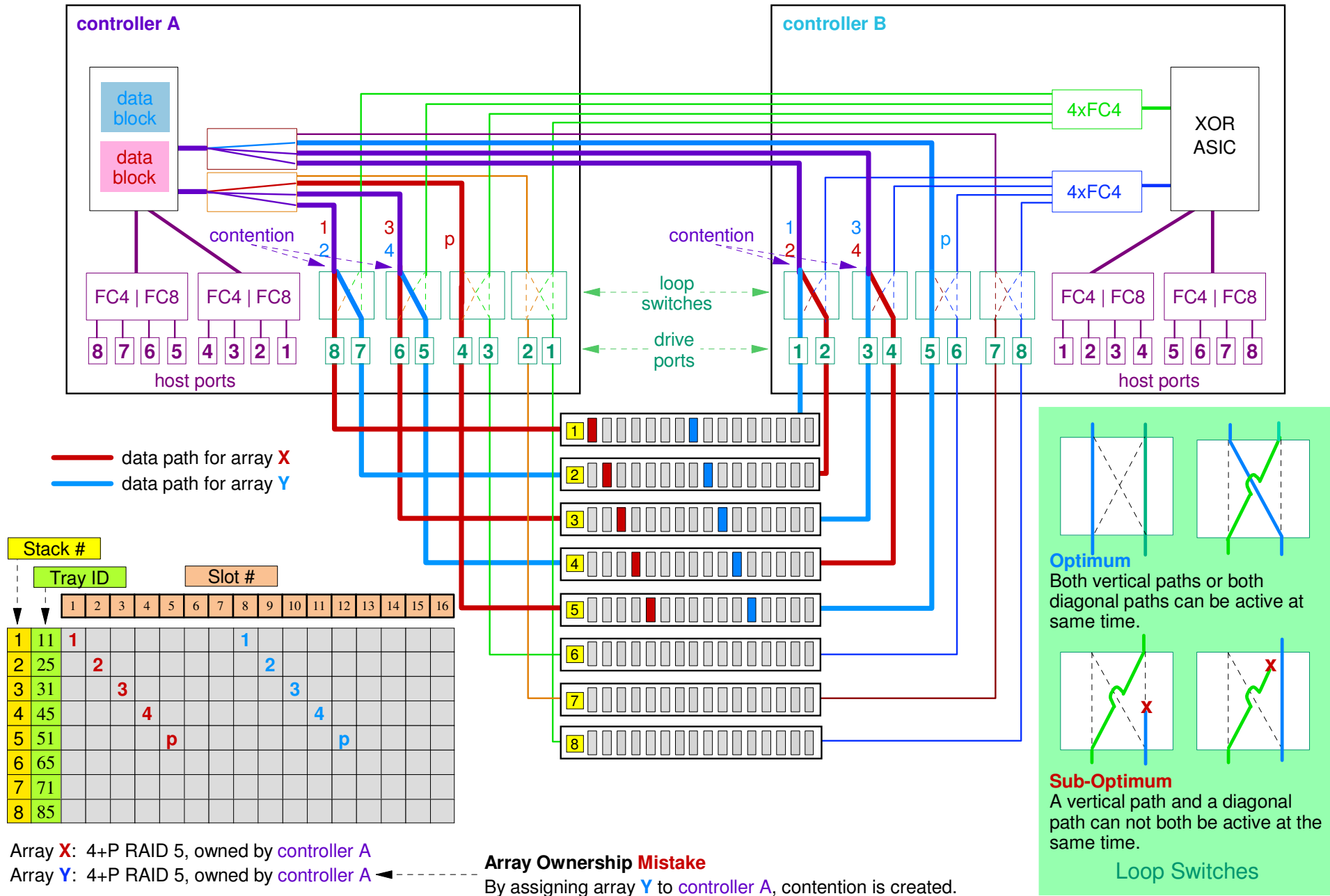
## Data Flow Example #4 (Doing It Wrong)





# DS5300/EXP5000

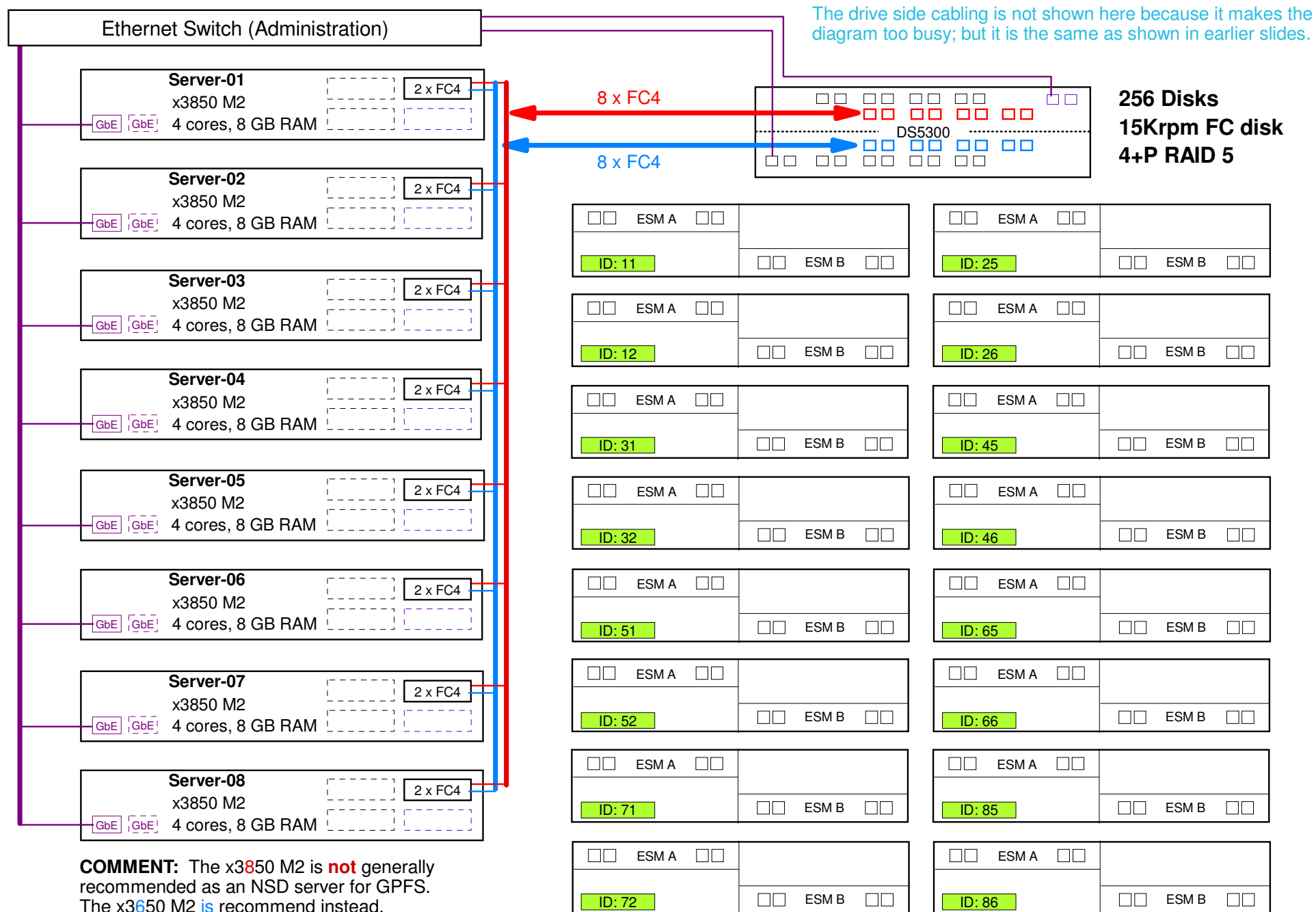
## Data Flow Example #5 (Doing It Wrong)





# DS5300/EXP5000

## Benchmark Configuration Using 16 Trays with No Trunking





# DS5300/EXP5000

## Disk to Array Mappings for Benchmark System

Slot	Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Array Owner	Tray ID																
A	11	0	0	1	2	3	4	4	5	6	7	8	8	9	10	11	HS
B	25	12	12	13	14	15	16	16	17	18	19	20	20	21	22	23	HS
A	31	0	1	1	2	3	4	5	5	6	7	8	9	9	10	11	HS
B	45	12	13	13	14	15	16	17	17	18	19	20	21	21	22	23	HS
A	51	0	1	2	2	3	4	5	6	6	7	8	9	10	10	11	HS
B	65	12	13	14	14	15	16	17	18	18	19	20	21	22	22	23	HS
A	71	0	1	2	3	3	4	5	6	7	7	8	9	10	11	11	HS
B	85	12	13	14	15	15	16	17	18	19	19	20	21	22	23	23	HS
A	12	24	24	25	26	27	28	28	29	30	31	32	32	33	34	35	HS
B	26	36	36	37	38	39	40	40	41	42	43	44	44	45	46	47	HS
A	32	24	25	25	26	27	28	29	29	30	31	32	33	33	34	35	HS
B	46	36	37	37	38	39	40	41	41	42	43	44	45	45	46	47	HS
A	52	24	25	26	26	27	28	29	30	30	31	32	33	34	34	35	HS
B	66	36	37	38	38	39	40	41	42	42	43	44	45	46	46	47	HS
A	72	24	25	26	27	27	28	29	30	31	31	32	33	34	35	35	HS
B	86	36	37	38	39	39	40	41	42	43	43	44	45	46	47	47	HS

Table entries are the LUN numbers.

### COMMENTS:

- Tests using 1, 2, 4, 8, 16, 24 LUNs use only trays 11, 25, 31, 45, 51, 65, 71, 85 effectively making them an 8 tray test.
- Tests using 32 or 48 LUNs use all 16 trays.
- Tests using 2, 4, 8, 16, 32 LUNs distribute the arrays uniformly over the 2 controllers (i.e., the same number of LUNs for each controller)





# DS5300/EXP5000

## Streaming Benchmark Results for 8/16 Trays with **No Trunking**

### DS5300 Parameters

- Firmware version: 07.60.24.00
- Size of cache: 16 GB
- Arrays: 4+P RAID 5, 15Krpm disks, 1 LUN per array using full capacity of disks
- segment size = 256K, cache page size = 32K
- wc\_on: Write Test with Write Cache = enabled, Write Cache Mirroring = enabled
- rc\_on: Read Test with Read Cache = enabled, Prefetch = disabled
- wc\_off: Write Test with Write Cache = disabled
- rc\_off: Read Test with Read Cache = disabled

More performance information can be found in the associated spreadsheet. See the Outline slide for details.

### GPFS 3.2.1.14

- SAN configuration
- blocksize = 4096K
- pagepool = 1G
- Allocation map type = scatter

### RHEL 5.2 (kernel 2.6.18-128.el5 x86\_64)

- Transfer Size: 2M
- Driver: MPP (RDAC)

### Benchmark Code: ibm.v4b

- Number of tasks: 64 (8 per node)
- Record size = 4M
- Access pattern = sequential<sup>1</sup>
- File size: variable but large<sup>2</sup>
- Direct I/O = off

### Footnotes:

1. While the application access is sequential, GPFS randomly distributes the data blocks on disk.
2. File sizes were large enough to negate unnatural caching effects.
3. Performance is measured in GB/s where G = 2<sup>30</sup>.
4. Tests using ≤ 24 LUNs use only 8 trays; tests using ≥ 32 LUNs use 16 trays.

LUNs <sup>4</sup>	1	2	4	8	16	24	32	48
wc_on	0.602	0.999	1.51	2.70	3.72	4.31	4.61	4.45
rc_on	0.337	0.680	1.18	2.26	3.97	5.38	5.85	5.63
wc_off	0.427	0.854	1.48	2.57	3.69	4.31	4.52	4.43
rc_off	0.568	0.643	1.19	2.31	4.03	5.40	5.88	5.64

These performance measurements are based on code instrumentation in ibm.v4b. They were generally 3% to 5% less than measured by the DS5300 performance monitor, especially for a larger number of LUNs. This is attributed to variance in task termination times which negatively skewed performance.

Performance measured in GB/s<sup>3</sup>



# DS5300/EXP5000

## IOP Benchmark Results with **No** Trunking



### DS5300 Parameters

- Firmware version: 07.50.xx.xx
- Size of cache: N/A
- Arrays: 6+P or 7+P RAID 5, 15Krpm disks, 1 LUN per array using full capacity of disks
- segment size = 128K
- cache page size = 32K
- Write Cache = disabled
- Read Cache = disabled

Evaluating "to media" tests.

Disks	8 <sup>1</sup>	256 <sup>1</sup>	384 <sup>1</sup>	448 <sup>2</sup>
transaction size	4K	4K	4K	4K
write	946	18599	18714	18073
duplex (50/50)	820	17297	34505	31157
read	931	29701	47560	56380

← -- linear scaling (50%) up to 256 disks, then it flattens.

← -- linear scaling (~= 100%)

Raw device results provided by LSI<sup>3</sup>

Performance measured in IOP/s.

#### Footnotes:

1. 7+P RAID 5

2. 6+P RAID 5

3. Further testing is need to evaluate IOP performance for GPFS on the DS5300. These results, based on realistic assumptions similar to those used elsewhere in this report, are included here as an indication of what can be expected using GPFS. The single core configuration of the x3850s available for the benchmarks in this study could not sustain enough threads to fully tax the IOP performance of the DS5300.



## DS5300/EXP5060

Front View



Rear View

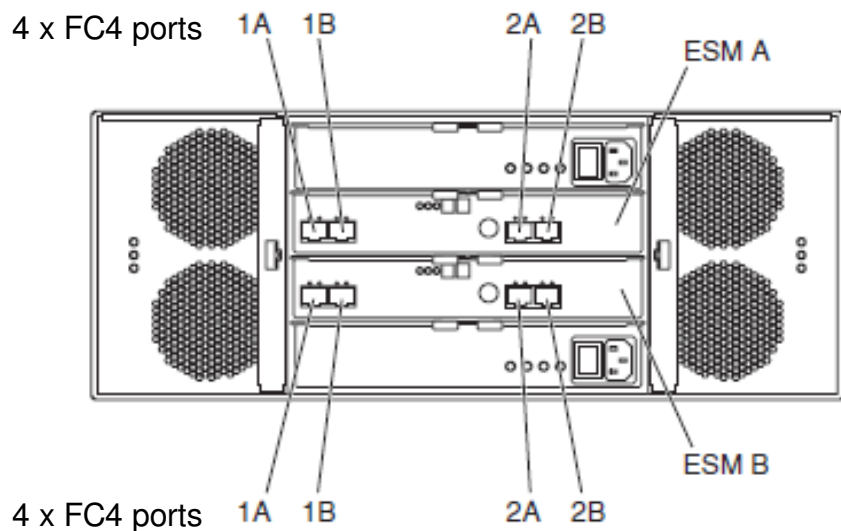


- ▶ High density 4U tray supporting upto 60 x 3.5" SATA disk drives
  - each tray is controlled by 2 ESMs
  - each tray contains 5 drawers
  - each drawer is controlled by 2 DCMs (Drawer Control Module)
  - each drawer contains 12 disk drives
- ▶ 8 x FC4 ports (4 per ESM)
  - supports trunking configuration
- ▶ FC Switched architecture
  - Higher performance, lower latency
  - Drive isolation, better diagnostics
- ▶ RoHS compliant

ESM Scope: when arrays in 4/8 tray configuration span all trays, then

- ▶ ESM A is the primary path for the odd drives
- ▶ ESM B is the primary path for the even drives

--- If an ESM fails, the other ESM can access all of the drives.



Close-up of single ESM

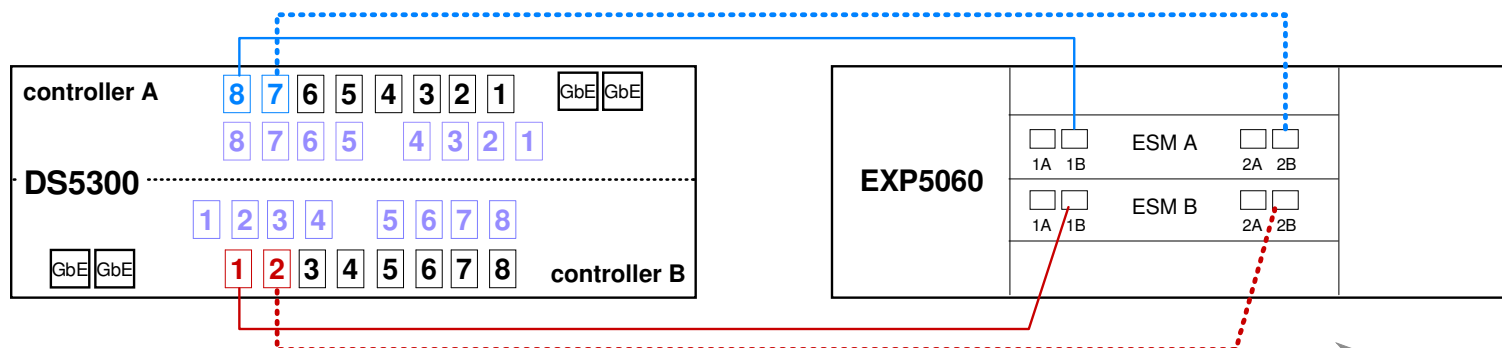




# DS5300/EXP5060

## Trunked vs. Non-trunked Cabling

### Trunked



### Trunked Cabling

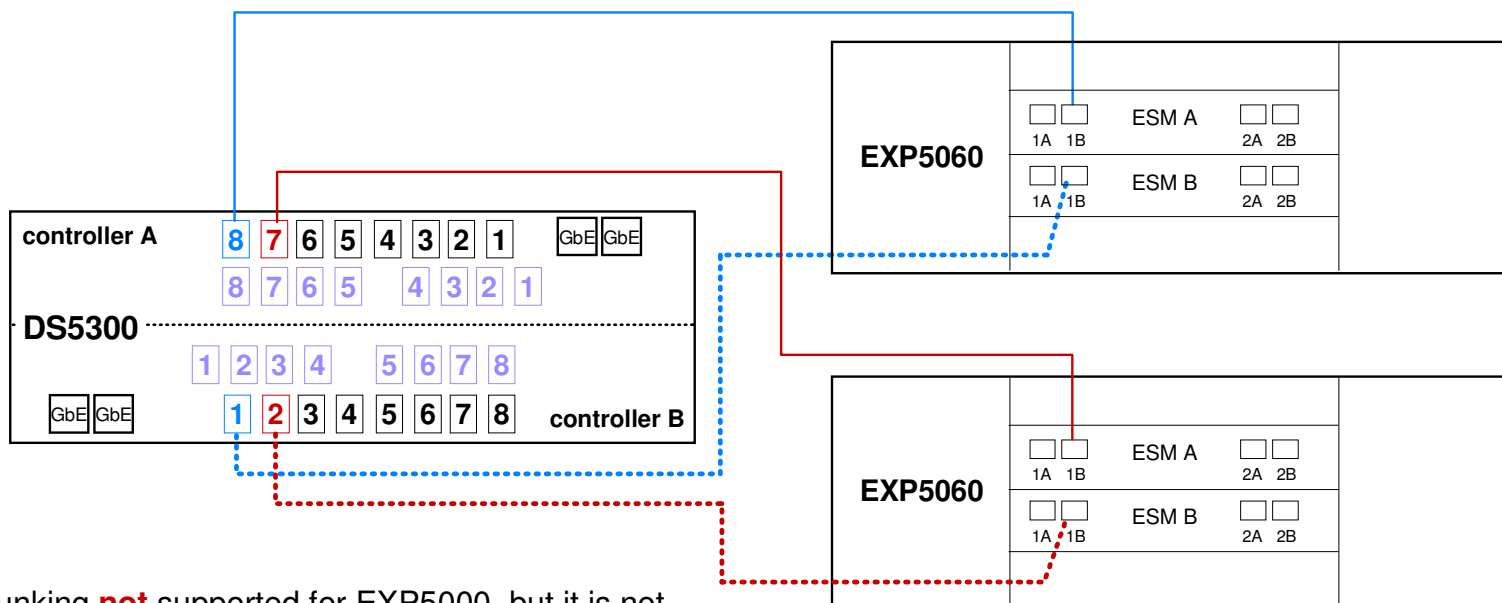
The DS5300/EXP5060 supports a "trunked cabling" configuration which allows 4 simultaneously active paths per tray. Thus 4 separate disk drives can be simultaneously accessed at any time, effectively doubling the performance per tray.

With 60 disk drives per EXP5060 tray, DS5300 performance can be maximized using only 4 trays; without trunking, it would require 8 trays to do the same thing.

#### COMMENT:

Peak streaming performance is the same for each example

### Non Trunked



**COMMENT:** Trunking **not** supported for EXP5000, but it is not as critical in this case since there are only 16 disks per tray.



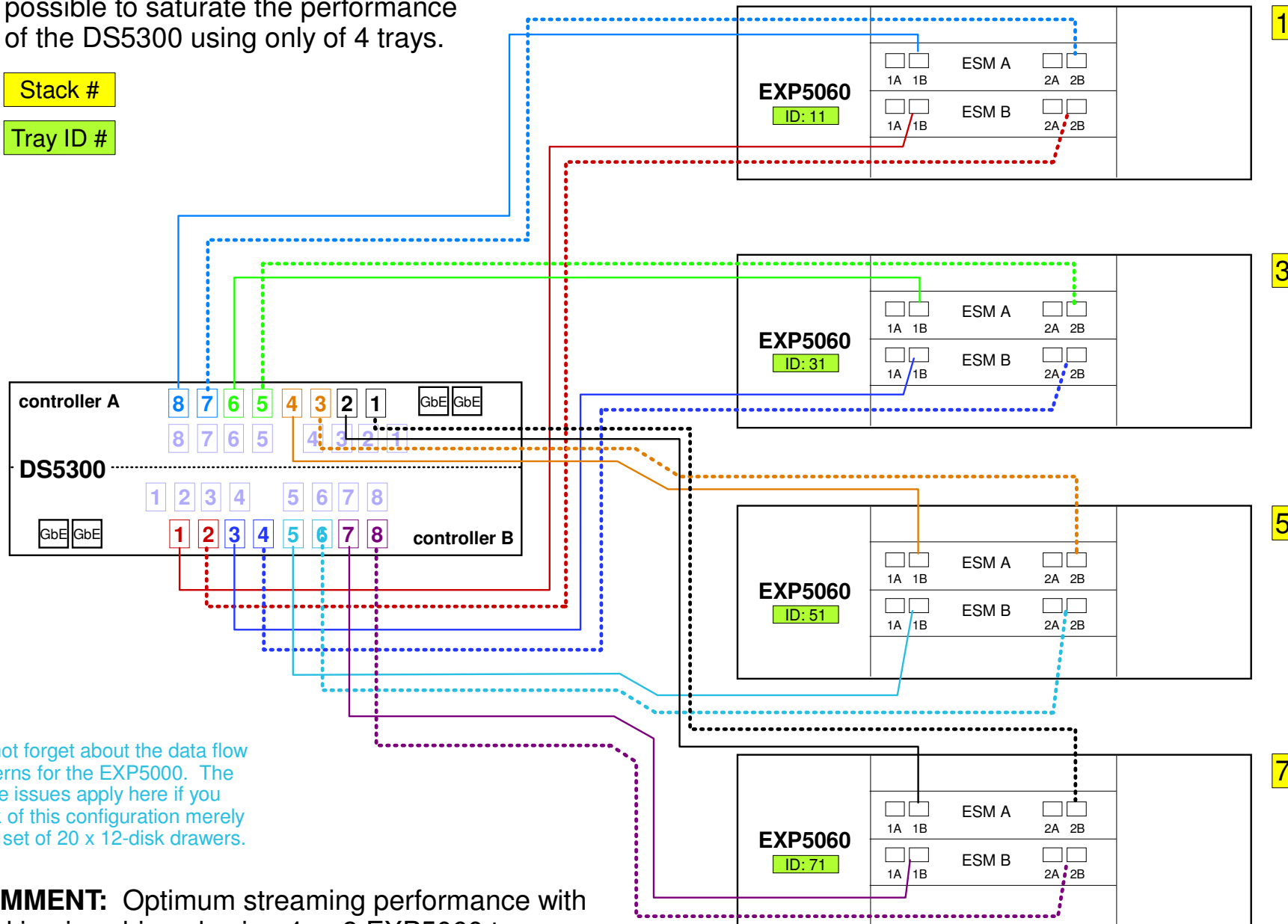
# DS5300/EXP5060

## Drive Side Cabling - 4 Trays with Trunking

**COMMENT:** Trunking makes it possible to saturate the performance of the DS5300 using only of 4 trays.

Stack #

Tray ID #



**COMMENT:** Optimum streaming performance with trunking is achieved using 4 or 8 EXP5060 trays.



# DS5300/EXP5060

## Optimal Disk to Array Mapping - 4 Trays with Trunking

### 24 LUNs

### 8+2P RAID 6

The numbers in the drawers represent the LUN number to which that disk belongs.

Controller A LUNs: 0-11  
Controller B LUNs: 12-23

Slot numbers

3	6	9	12
2	5	8	11
1	4	7	10

When SMClient generates a profile for the DS5300, it references arrays by tray ID, drawer number and slot number.

#### COMMENTS:

- This configuration is drawer protected. Since it is 8+2P RAID 6, it can not be tray protected.
- Think of this as a 20 drawer configuration.



The following LUNs are highlighted to illustrate the pattern:

0 8 15 22

Tray 11	Tray 31	Tray 51	Tray 71																																																																																																																																																																																																																																																
<b>Drawer 1</b> <table> <tr><td>2</td><td>5</td><td>8</td><td>11</td></tr> <tr><td>1</td><td>4</td><td>7</td><td>10</td></tr> <tr><td>0</td><td>3</td><td>6</td><td>9</td></tr> </table> <b>Drawer 2</b> <table> <tr><td>2</td><td>5</td><td>8</td><td>11</td></tr> <tr><td>1</td><td>4</td><td>7</td><td>10</td></tr> <tr><td>0</td><td>3</td><td>6</td><td>9</td></tr> </table> <b>Drawer 3</b> <table> <tr><td>14</td><td>17</td><td>8</td><td>11</td></tr> <tr><td>13</td><td>16</td><td>7</td><td>10</td></tr> <tr><td>12</td><td>15</td><td>6</td><td>9</td></tr> </table> <b>Drawer 4</b> <table> <tr><td>14</td><td>17</td><td>20</td><td>23</td></tr> <tr><td>13</td><td>16</td><td>19</td><td>22</td></tr> <tr><td>12</td><td>15</td><td>18</td><td>21</td></tr> </table> <b>Drawer 5</b> <table> <tr><td>14</td><td>17</td><td>20</td><td>23</td></tr> <tr><td>13</td><td>16</td><td>19</td><td>22</td></tr> <tr><td>12</td><td>15</td><td>18</td><td>21</td></tr> </table>	2	5	8	11	1	4	7	10	0	3	6	9	2	5	8	11	1	4	7	10	0	3	6	9	14	17	8	11	13	16	7	10	12	15	6	9	14	17	20	23	13	16	19	22	12	15	18	21	14	17	20	23	13	16	19	22	12	15	18	21	<b>Drawer 1</b> <table> <tr><td>2</td><td>5</td><td>8</td><td>11</td></tr> <tr><td>1</td><td>4</td><td>7</td><td>10</td></tr> <tr><td>0</td><td>3</td><td>6</td><td>9</td></tr> </table> <b>Drawer 2</b> <table> <tr><td>2</td><td>5</td><td>8</td><td>11</td></tr> <tr><td>1</td><td>4</td><td>7</td><td>10</td></tr> <tr><td>0</td><td>3</td><td>6</td><td>9</td></tr> </table> <b>Drawer 3</b> <table> <tr><td>14</td><td>17</td><td>8</td><td>11</td></tr> <tr><td>13</td><td>16</td><td>7</td><td>10</td></tr> <tr><td>12</td><td>15</td><td>6</td><td>9</td></tr> </table> <b>Drawer 4</b> <table> <tr><td>14</td><td>17</td><td>20</td><td>23</td></tr> <tr><td>13</td><td>16</td><td>19</td><td>22</td></tr> <tr><td>12</td><td>15</td><td>18</td><td>21</td></tr> </table> <b>Drawer 5</b> <table> <tr><td>14</td><td>17</td><td>20</td><td>23</td></tr> <tr><td>13</td><td>16</td><td>19</td><td>22</td></tr> <tr><td>12</td><td>15</td><td>18</td><td>21</td></tr> </table>	2	5	8	11	1	4	7	10	0	3	6	9	2	5	8	11	1	4	7	10	0	3	6	9	14	17	8	11	13	16	7	10	12	15	6	9	14	17	20	23	13	16	19	22	12	15	18	21	14	17	20	23	13	16	19	22	12	15	18	21	<b>Drawer 1</b> <table> <tr><td>2</td><td>5</td><td>8</td><td>11</td></tr> <tr><td>1</td><td>4</td><td>7</td><td>10</td></tr> <tr><td>0</td><td>3</td><td>6</td><td>9</td></tr> </table> <b>Drawer 2</b> <table> <tr><td>2</td><td>5</td><td>8</td><td>11</td></tr> <tr><td>1</td><td>4</td><td>7</td><td>10</td></tr> <tr><td>0</td><td>3</td><td>6</td><td>9</td></tr> </table> <b>Drawer 3</b> <table> <tr><td>2</td><td>5</td><td>20</td><td>23</td></tr> <tr><td>1</td><td>4</td><td>19</td><td>22</td></tr> <tr><td>0</td><td>3</td><td>18</td><td>21</td></tr> </table> <b>Drawer 4</b> <table> <tr><td>14</td><td>17</td><td>20</td><td>23</td></tr> <tr><td>13</td><td>16</td><td>19</td><td>22</td></tr> <tr><td>12</td><td>15</td><td>18</td><td>21</td></tr> </table> <b>Drawer 5</b> <table> <tr><td>14</td><td>17</td><td>20</td><td>23</td></tr> <tr><td>13</td><td>16</td><td>19</td><td>22</td></tr> <tr><td>12</td><td>15</td><td>18</td><td>21</td></tr> </table>	2	5	8	11	1	4	7	10	0	3	6	9	2	5	8	11	1	4	7	10	0	3	6	9	2	5	20	23	1	4	19	22	0	3	18	21	14	17	20	23	13	16	19	22	12	15	18	21	14	17	20	23	13	16	19	22	12	15	18	21	<b>Drawer 1</b> <table> <tr><td>2</td><td>5</td><td>8</td><td>11</td></tr> <tr><td>1</td><td>4</td><td>7</td><td>10</td></tr> <tr><td>0</td><td>3</td><td>6</td><td>9</td></tr> </table> <b>Drawer 2</b> <table> <tr><td>2</td><td>5</td><td>8</td><td>11</td></tr> <tr><td>1</td><td>4</td><td>7</td><td>10</td></tr> <tr><td>0</td><td>3</td><td>6</td><td>9</td></tr> </table> <b>Drawer 3</b> <table> <tr><td>2</td><td>5</td><td>20</td><td>23</td></tr> <tr><td>1</td><td>4</td><td>19</td><td>22</td></tr> <tr><td>0</td><td>3</td><td>18</td><td>21</td></tr> </table> <b>Drawer 4</b> <table> <tr><td>14</td><td>17</td><td>20</td><td>23</td></tr> <tr><td>13</td><td>16</td><td>19</td><td>22</td></tr> <tr><td>12</td><td>15</td><td>18</td><td>21</td></tr> </table> <b>Drawer 5</b> <table> <tr><td>14</td><td>17</td><td>20</td><td>23</td></tr> <tr><td>13</td><td>16</td><td>19</td><td>22</td></tr> <tr><td>12</td><td>15</td><td>18</td><td>21</td></tr> </table>	2	5	8	11	1	4	7	10	0	3	6	9	2	5	8	11	1	4	7	10	0	3	6	9	2	5	20	23	1	4	19	22	0	3	18	21	14	17	20	23	13	16	19	22	12	15	18	21	14	17	20	23	13	16	19	22	12	15	18	21
2	5	8	11																																																																																																																																																																																																																																																
1	4	7	10																																																																																																																																																																																																																																																
0	3	6	9																																																																																																																																																																																																																																																
2	5	8	11																																																																																																																																																																																																																																																
1	4	7	10																																																																																																																																																																																																																																																
0	3	6	9																																																																																																																																																																																																																																																
14	17	8	11																																																																																																																																																																																																																																																
13	16	7	10																																																																																																																																																																																																																																																
12	15	6	9																																																																																																																																																																																																																																																
14	17	20	23																																																																																																																																																																																																																																																
13	16	19	22																																																																																																																																																																																																																																																
12	15	18	21																																																																																																																																																																																																																																																
14	17	20	23																																																																																																																																																																																																																																																
13	16	19	22																																																																																																																																																																																																																																																
12	15	18	21																																																																																																																																																																																																																																																
2	5	8	11																																																																																																																																																																																																																																																
1	4	7	10																																																																																																																																																																																																																																																
0	3	6	9																																																																																																																																																																																																																																																
2	5	8	11																																																																																																																																																																																																																																																
1	4	7	10																																																																																																																																																																																																																																																
0	3	6	9																																																																																																																																																																																																																																																
14	17	8	11																																																																																																																																																																																																																																																
13	16	7	10																																																																																																																																																																																																																																																
12	15	6	9																																																																																																																																																																																																																																																
14	17	20	23																																																																																																																																																																																																																																																
13	16	19	22																																																																																																																																																																																																																																																
12	15	18	21																																																																																																																																																																																																																																																
14	17	20	23																																																																																																																																																																																																																																																
13	16	19	22																																																																																																																																																																																																																																																
12	15	18	21																																																																																																																																																																																																																																																
2	5	8	11																																																																																																																																																																																																																																																
1	4	7	10																																																																																																																																																																																																																																																
0	3	6	9																																																																																																																																																																																																																																																
2	5	8	11																																																																																																																																																																																																																																																
1	4	7	10																																																																																																																																																																																																																																																
0	3	6	9																																																																																																																																																																																																																																																
2	5	20	23																																																																																																																																																																																																																																																
1	4	19	22																																																																																																																																																																																																																																																
0	3	18	21																																																																																																																																																																																																																																																
14	17	20	23																																																																																																																																																																																																																																																
13	16	19	22																																																																																																																																																																																																																																																
12	15	18	21																																																																																																																																																																																																																																																
14	17	20	23																																																																																																																																																																																																																																																
13	16	19	22																																																																																																																																																																																																																																																
12	15	18	21																																																																																																																																																																																																																																																
2	5	8	11																																																																																																																																																																																																																																																
1	4	7	10																																																																																																																																																																																																																																																
0	3	6	9																																																																																																																																																																																																																																																
2	5	8	11																																																																																																																																																																																																																																																
1	4	7	10																																																																																																																																																																																																																																																
0	3	6	9																																																																																																																																																																																																																																																
2	5	20	23																																																																																																																																																																																																																																																
1	4	19	22																																																																																																																																																																																																																																																
0	3	18	21																																																																																																																																																																																																																																																
14	17	20	23																																																																																																																																																																																																																																																
13	16	19	22																																																																																																																																																																																																																																																
12	15	18	21																																																																																																																																																																																																																																																
14	17	20	23																																																																																																																																																																																																																																																
13	16	19	22																																																																																																																																																																																																																																																
12	15	18	21																																																																																																																																																																																																																																																

### Balance is needed for good performance!

- Balance drive access across all available drive channels by making sure that a controller is accessing an equal number of drives from all drive channels.
- Balance drives access across both ESMs in a given tray by insuring that the number of odd slots and even slots used by a controller is equal.
- Avoid sharing a drawer between both controllers by mapping all of the disks in a drawer to LUNs owned by the same controller.





# DS5300/EXP5060

## Alternative View - Optimal Disk to Array Mapping - 4 Trays with Trunking

### 24 LUNs

### 8+2P RAID 6

The numbers in the drawers represent the LUN number to which that disk belongs.

Controller A LUNs: 0-11  
Controller B LUNs: 12-23

The following LUNs are highlighted to illustrate the pattern:

0
8
15
22

#### COMMENTS:

- This configuration is drawer protected. Since it is 8+2P RAID 6, it can not be tray protected.
- Think of this as a 20 drawer configuration.



Tray 11

Drawer 1	0	1	2	3	4	5	6	7	8	9	10	11
Drawer 2	0	1	2	3	4	5	6	7	8	9	10	11
Drawer 3	12	13	14	15	16	17	6	7	8	9	10	11
Drawer 4	12	13	14	15	16	17	18	19	20	21	22	23
Drawer 5	12	13	14	15	16	17	18	19	20	21	22	23

Tray 31

Drawer 1	0	1	2	3	4	5	6	7	8	9	10	11
Drawer 2	0	1	2	3	4	5	6	7	8	9	10	11
Drawer 3	12	13	14	15	16	17	6	7	8	9	10	11
Drawer 4	12	13	14	15	16	17	18	19	20	21	22	23
Drawer 5	12	13	14	15	16	17	18	19	20	21	22	23

Tray 51

Drawer 1	0	1	2	3	4	5	6	7	8	9	10	11
Drawer 2	0	1	2	3	4	5	6	7	8	9	10	11
Drawer 3	0	1	2	3	4	5	18	19	20	21	22	23
Drawer 4	12	13	14	15	16	17	18	19	20	21	22	23
Drawer 5	12	13	14	15	16	17	18	19	20	21	22	23

Tray 71

Drawer 1	0	1	2	3	4	5	6	7	8	9	10	11
Drawer 2	0	1	2	3	4	5	6	7	8	9	10	11
Drawer 3	0	1	2	3	4	5	18	19	20	21	22	23
Drawer 4	12	13	14	15	16	17	18	19	20	21	22	23
Drawer 5	12	13	14	15	16	17	18	19	20	21	22	23

#### ALTERNATIVE VIEW:

The author prefers this view over the one on the previous page. However, most of the tools for working with the EXP5060 are similar to the view from the previous page. So as to stay consistent with these tools, the author uses the view from the previous page else where in this document.

Table entries are the LUN numbers.

1	2	3	...	10	11	12
---	---	---	-----	----	----	----

Each tray can be viewed as an array of 5 rows and 12 columns. Each row is a drawer and each column has a slot number from 1 to 12.

When SMClient generates a profile for the DS5300, it references arrays by tray ID, drawer number and slot number.

### Balance is needed for good performance!

- Balance drive access across all available drive channels by making sure that a controller is accessing an equal number of drives from all drive channels.
- Balance drives access across both ESMs in a given tray by insuring that the number of odd slots and even slots used by a controller is equal.
- Avoid sharing a drawer between both controllers by mapping all of the disks in a drawer to LUNs owned by the same controller.



# DS5300/EXP5060

## Optimal Disk to Array Mapping - 4 Trays with Trunking

### Odd Drive Counts

Tray ID	11	31	51	71
LUN 0	2	2	3	3
LUN 1	0	0	0	0
LUN 2	2	2	3	3
LUN 3	0	0	0	0
LUN 4	2	2	3	3
LUN 5	0	0	0	0
LUN 6	3	3	2	2
LUN 7	0	0	0	0
LUN 8	3	3	2	2
LUN 9	0	0	0	0
LUN 10	3	3	2	2
LUN 11	0	0	0	0
LUN 12	3	3	2	2
LUN 13	0	0	0	0
LUN 14	3	3	2	2
LUN 15	0	0	0	0
LUN 16	3	3	2	2
LUN 17	0	0	0	0
LUN 18	2	2	3	3
LUN 19	0	0	0	0
LUN 20	2	2	3	3
LUN 21	0	0	0	0
LUN 22	2	2	3	3
LUN 23	0	0	0	0
Drawer 1	6	6	6	6
Drawer 2	6	6	6	6
Drawer 3	6	6	6	6
Drawer 4	6	6	6	6
Drawer 5	6	6	6	6
Tray Total	30	30	30	30

### Even Drive Counts

Tray ID	11	31	51	71
LUN 0	0	0	0	0
LUN 1	2	2	3	3
LUN 2	0	0	0	0
LUN 3	2	2	3	3
LUN 4	0	0	0	0
LUN 5	2	2	3	3
LUN 6	0	0	0	0
LUN 7	3	3	2	2
LUN 8	0	0	0	0
LUN 9	3	3	2	2
LUN 10	0	0	0	0
LUN 11	3	3	2	2
LUN 12	0	0	0	0
LUN 13	3	3	2	2
LUN 14	0	0	0	0
LUN 15	3	3	2	2
LUN 16	0	0	0	0
LUN 17	3	3	2	2
LUN 18	0	0	0	0
LUN 19	2	2	3	3
LUN 20	0	0	0	0
LUN 21	2	2	3	3
LUN 22	0	0	0	0
LUN 23	2	2	3	3
Drawer 1	6	6	6	6
Drawer 2	6	6	6	6
Drawer 3	6	6	6	6
Drawer 4	6	6	6	6
Drawer 5	6	6	6	6
Tray Total	30	30	30	30

In normal operation, odd drives are accessed by ESM A and even drives are accessed by ESM B. Therefore, ensure that all disks belonging to a LUN are either all odd or all even.

Number of disks from LUN X in Tray Y

As far as possible ensure that the number of disks per tray for any LUN is balanced.

If the system is balanced, then the **Tray Total** for "odds" and "evens" should be the same for each tray.

Number of disks from Tray Y in Drawer Z

Drawers **are** shared by both controllers (*i.e.*, a drawer contains disks from LUNs owned by controller A and LUNs owned by controller B)\*

Drawers are **not** shared by both controllers

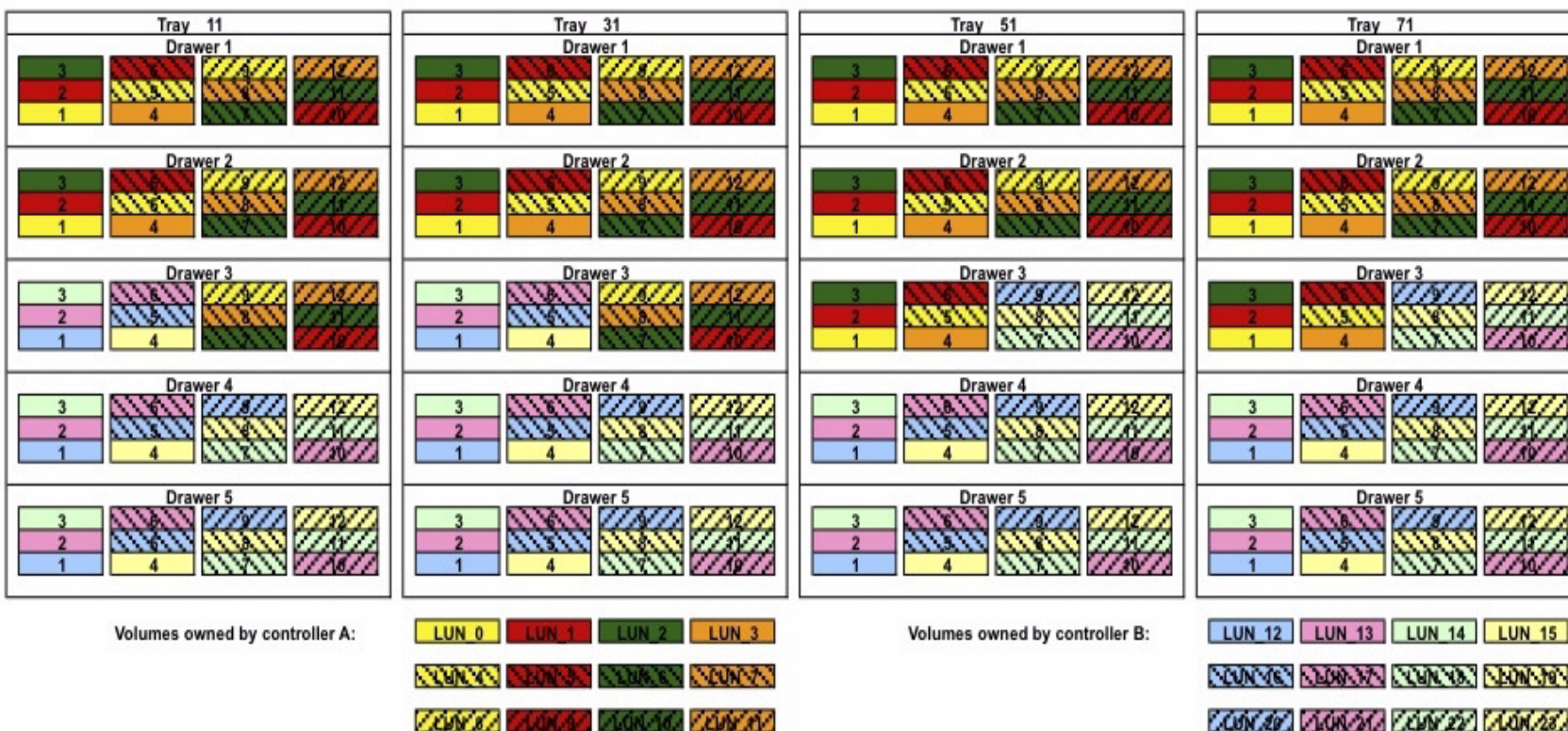
### Footnote:

★ Since there is an odd number of drawers per tray, it is necessary to either share a drawer or have an imbalance across drive channels; the option with the least performance degradation is to share drawers.



# DS5300/EXP5060

## An Alternative View Using 4 Trays with Trunking



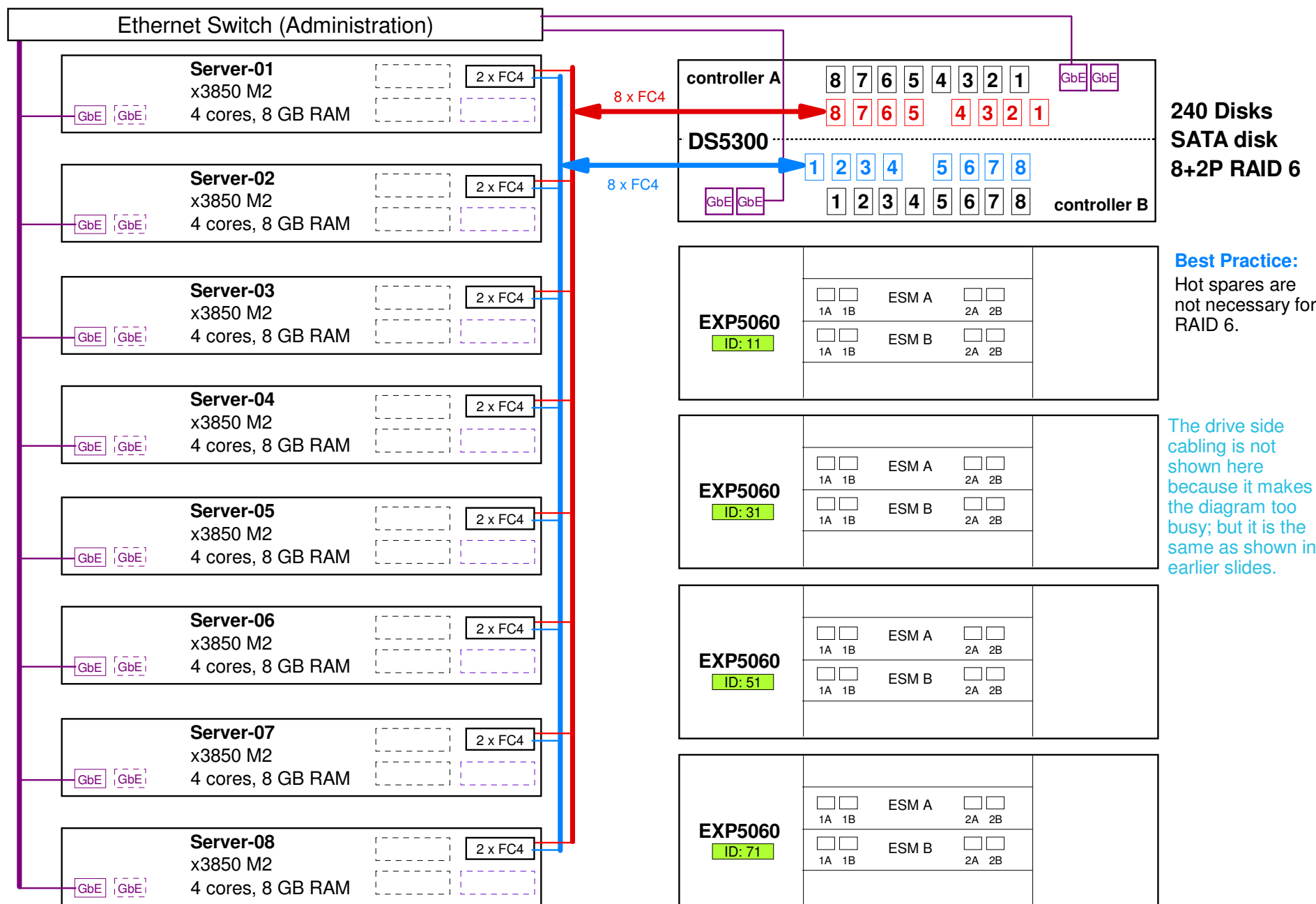
An Excel Spreadsheet tool can be provided upon request that assists with the disk to array mapping task. It provides the following basic features:

- ▶ GUI assisted drive layout (illustrated above)
- ▶ Balance analysis
- ▶ Generate a SMCLI script to build the arrays



# DS5300/EXP5060

## Benchmark Configuration Using 4 Trays with Trunking



**COMMENT:** The x3850 M2 is **not** generally recommended as an NSD server for GPFS. The x3650 M2 **is** recommend instead.



# DS5300/EXP5060

## Streaming Benchmark Results for 4 Trays with Trunking

### DS5300 Parameters

- Firmware version: 07.60.24.00
- Size of cache: 16 GB
- Arrays: 8+2P RAID 6, 1 LUN per array using full capacity of disks
- segment size = 256K, cache page size = 32K
- wc\_on: Write Test with Write Cache = enabled, Write Cache Mirroring = enabled
- rc\_on: Read Test with Read Cache = enabled, Prefetch = disabled
- wc\_off: Write Test with Write Cache = disabled
- rc\_off: Read Test with Read Cache = disabled

More performance information can be found in the associated spreadsheet. See the Outline slide for details.

### GPFS 3.2.1.14

- SAN configuration
- blocksize = 4096K
- pagepool = 1G
- Allocation map type = scatter

### RHEL 5.2 (kernel 2.6.18-128.el5 x86\_64)

- Transfer Size: 2M
- Driver: MPP (RDAC)

### Benchmark Code: ibm.v4b

- Number of tasks: 64 (8 per node)
- Record size = 4M
- Access patter = sequential<sup>1</sup>
- File size: variable but large<sup>2</sup>
- Direct I/O = off

### Footnotes:

1. While the application access is sequential, GPFS randomly distributes the data blocks on disk.
2. File sizes were large enough to guarantee to negate unnatural caching effects.
3. Performance is measured in GB/s where G = 2<sup>30</sup>.

LUNs	1	2	4	8	16	24
wc_on	0.565	0.873	1.39	2.32	2.59	3.89
rc_on	0.429	0.712	1.56	2.65	3.50	4.65
wc_off	0.451	0.754	1.39	2.34	2.58	3.94
rc_off	0.804	0.734	1.51	2.93	3.56	4.72

These performance measurements are based on code instrumentation in ibm.v4b. They were generally 3% to 5% less than measured by the DS5300 performance monitor, especially for a larger number of LUNs. This is attributed to variance in task termination times which negatively skewed performance.

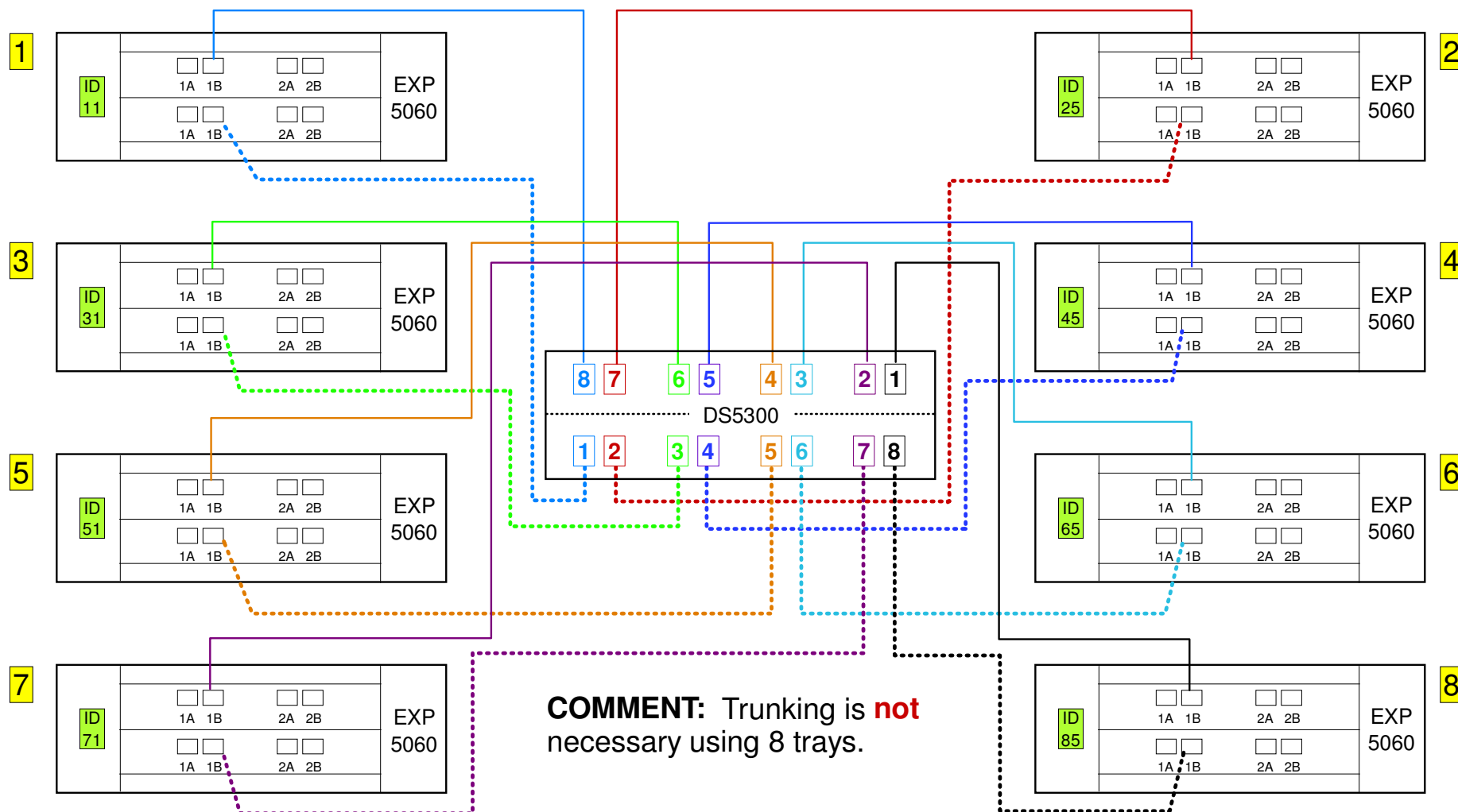
Performance measured in GB/s<sup>3</sup>





# DS5300/EXP5060

## Drive Side Cabling - 8 Trays with **no** Trunking



Stack #    Tray ID #

**COMMENT:** Optimum streaming performance without trunking requires all 8 EXP5060 trays.



# DS5300/EXP5060

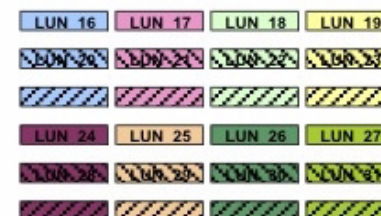
## Optimal Disk to Array Mapping - 8 Trays with no Trunking



Volumes owned by controller A:



Volumes owned by controller B:



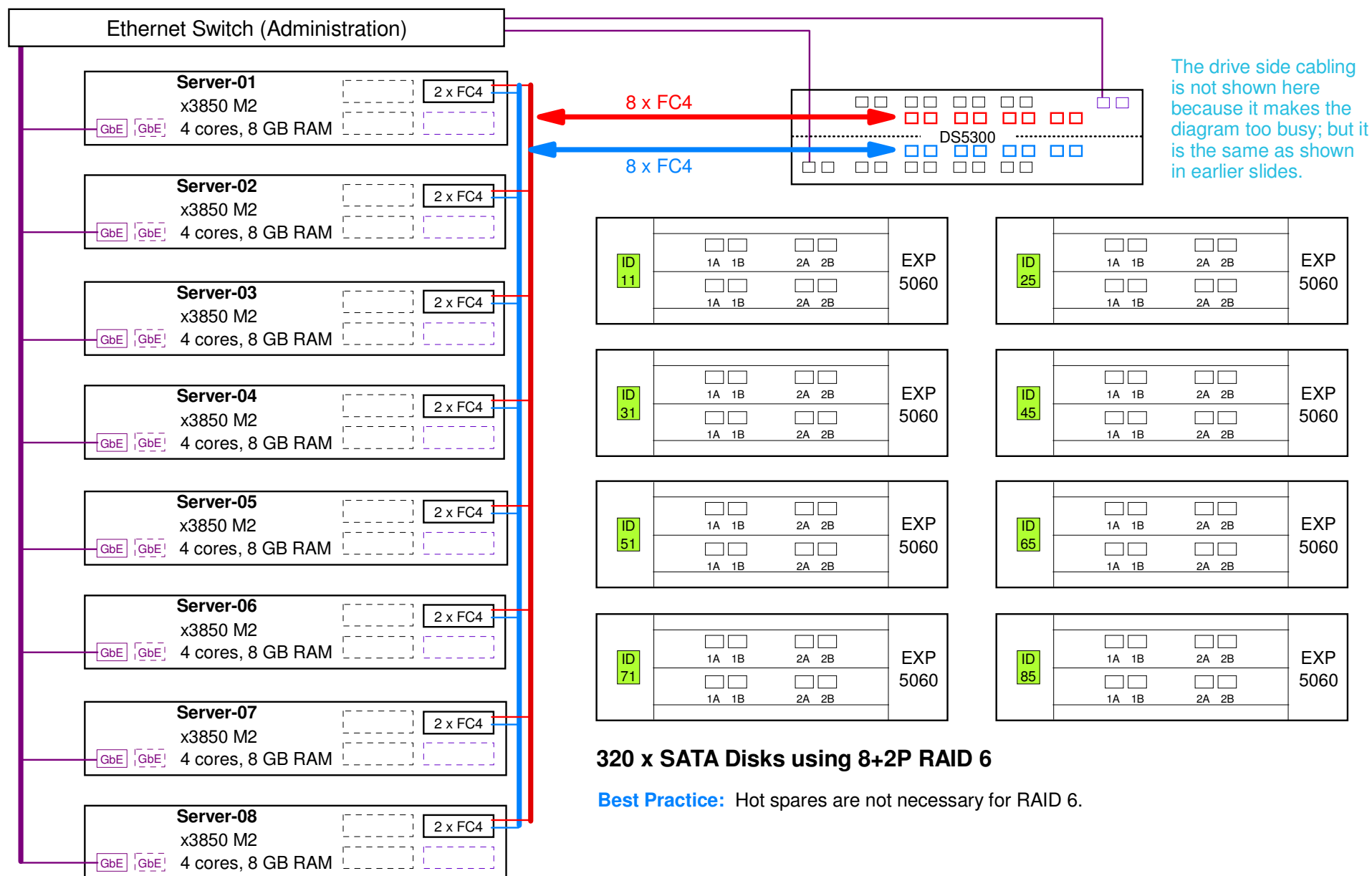
**COMMENT:** This was the configuration that was used for benchmarking since we only had 320 x SATA disks.





# DS5300/EXP5060

## Benchmark Configuration Using 8 Trays with **No** Trunking



**COMMENT:** The x3850 M2 is **not** generally recommended as an NSD server for GPFS. The x3650 M2 **is** recommended instead.



# DS5300/EXP5060

## Benchmark Results for 8 Trays with **No** Trunking

### DS5300 Parameters

- Firmware version: 07.60.24.00
- Size of cache: 16 GB
- Arrays: 8+2P RAID 6, SATA disks, 1 LUN per array using full capacity of disks
- segment size = 256K, cache page size = 32K
- wc\_on: Write Test with Write Cache = enabled, Write Cache Mirroring = enabled
- rc\_on: Read Test with Read Cache = enabled, Prefetch = disabled
- wc\_off: Write Test with Write Cache = disabled
- rc\_off: Read Test with Read Cache = disabled

More performance information can be found in the associated spreadsheet. See the Outline slide for details.

### GPFS 3.2.1.14

- SAN Configuration
- blocksize = 4096K
- pagepool = 1G
- Allocation map type: scatter

### RHEL 5.2 (kernel 2.6.18-128.el5 x86\_64)

- Transfer Size: 2M
- Driver: MPP (RDAC)

### Benchmark Code: ibm.v4b

- Number of tasks: 64 (8 per node)
- Record size = 4M
- Access pattern = sequential<sup>1</sup>
- File size: variable but large<sup>2</sup>
- Direct I/O = off

### Footnotes:

1. While the application access is sequential, GPFS randomly distributes the data blocks on disk.
2. File sizes were large enough to negate unnatural caching effects.
3. Performance is measured in GB/s where  $G = 2^{30}$ .

Performance measured in GB/s<sup>3</sup>

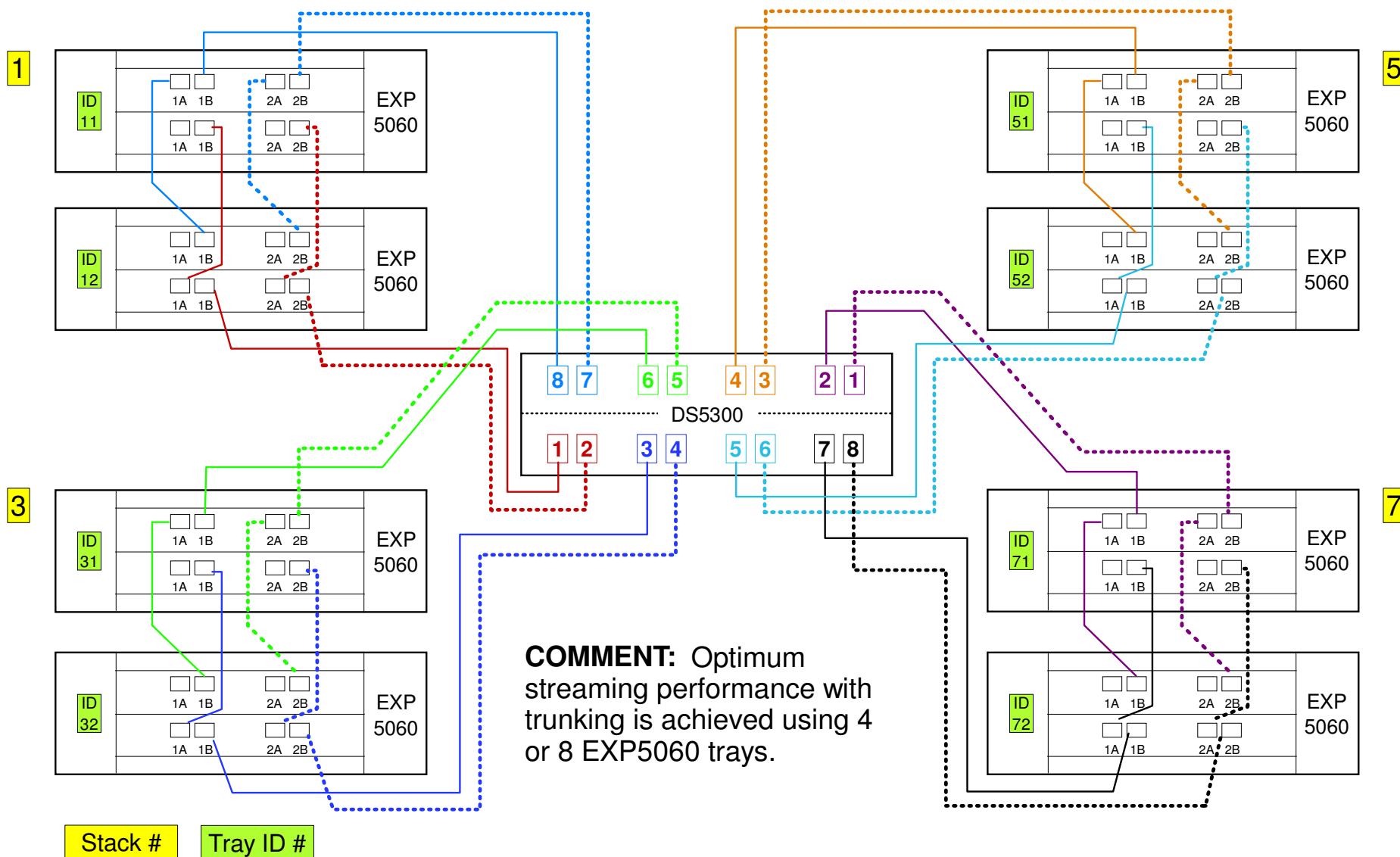
LUNs	1	2	4	8	16	32
wc_on	0.528	0.825	1.52	2.44	2.91	4.23
rc_on	0.396	0.713	1.37	2.63	3.44	5.30
wc_off	0.443	0.694	1.47	2.43	2.89	4.17
rc_off	0.664	0.653	1.44	2.74	3.47	5.31

These performance measurements are based on code instrumentation in ibm.v4b. They were generally 3% to 5% less than measured by the DS5300 performance monitor, especially for a larger number of LUNs. This is attributed to variance in task termination times which negatively skewed performance.



# DS5300/EXP5060

## Drive Side Cabling - 8 Trays with Trunking

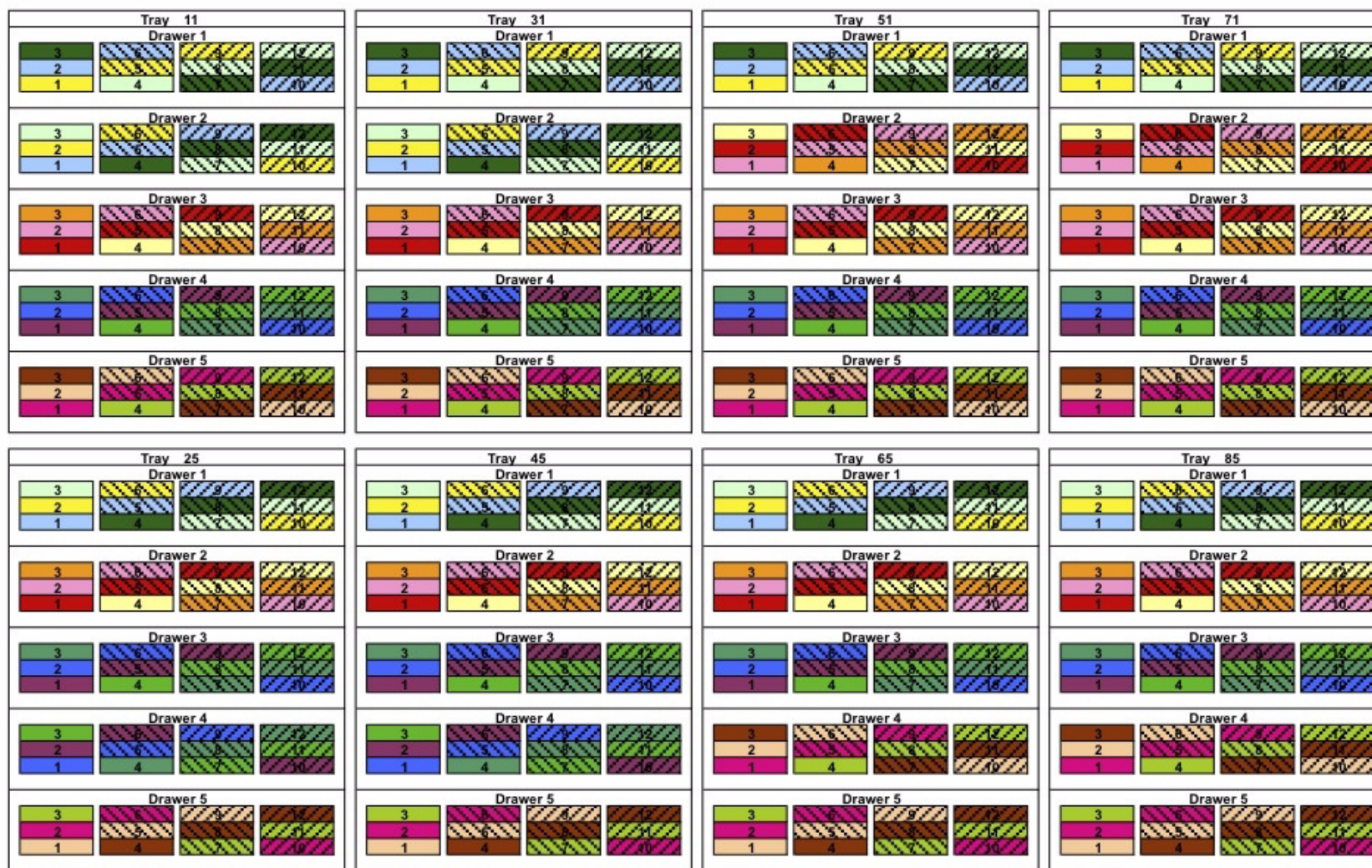


**COMMENT:** This example illustrates how trunking can be used in an 8 tray configuration. However, DS5300 performance can be maximized with **out** trunking when using 8 trays.



# DS5300/EXP5060

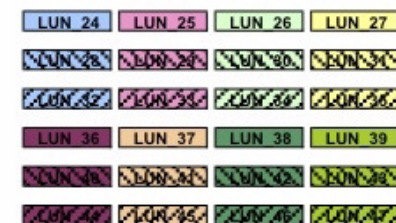
## Optimal Disk to Array Mapping - 8 Trays with Trunking



Volumes owned by controller A:



Volumes owned by controller B:



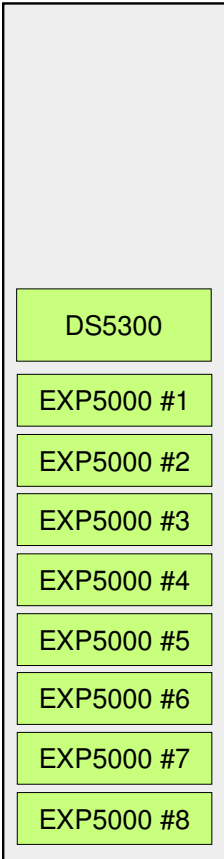
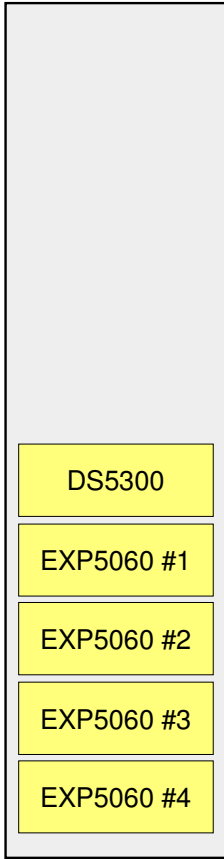
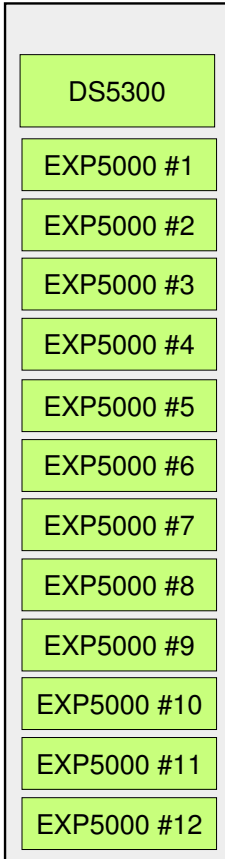
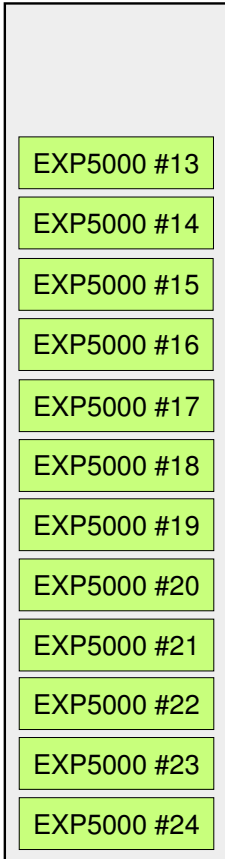
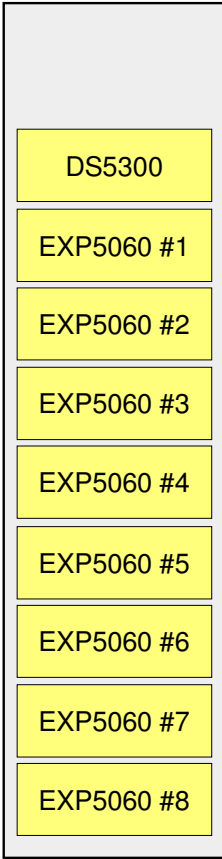
**COMMENT:** Clearly, this configuration should easily yield peak streaming performance similar to the previous configurations.





# Comparing EXP5000 and EXP5060

## Aggregate Solutions with Optimal Streaming Configurations

Performance Optimized	Capacity Optimized	Over Engineered for Streaming		Capacity Optimized
				
8 x EXP5000 128 disks FC disk, 15Krpm, 450 GB/disk Dimension: 28u Raw Capacity: 58 TB Streaming rate ► write < 4.3 MB/s ► read < 5.4 MB/s IOP rate* ► write < 9,000 IOP/s ► read < 15,000 IOP/s	4 x EXP5060 240 disks SATA disk, 7200 RPM, 1 TB/disk Dimension: 20u Raw Capacity: 240 TB Streaming rate ► write < 3.9 MB/s ► read < 4.8 MB/s IOP rate* ► write < TBD+ ► read < TBD+	24 x EXP5000* 384 disks FC disk, 15Krpm, 450 GB/disk Dimension: 76u Raw Capacity: 174 TB Streaming rate ► write < 4.4 MB/s ► read < 5.6 MB/s IOP rate* ► write < 18,000 IOP/s ► read < 30,000 IOP/s	8 x EXP5060 480 disks SATA disk, 7200 RPM, 1 TB/disk Dimension: 36u Raw Capacity: 480 TB Streaming rate ► write < 4.0 MB/s ► read < 5.0 MB/s IOP rate* ► write < TBD+ ► read < TBD+	

### FOOTNOTES:

★ Based on 4K, noncached transactions.

✦ As a first order approximation 7200 RPM SATA disk IOP rates ≈ 33% of 15Krpm FC disk IOP rates.



## Comparing EXP5000 and EXP5060 Aggregate Solutions with Optimal Streaming Configurations

---

16 and 24 tray EXP5000 configurations are **not** recommended as a best practice (***unless** IOP performance must be optimized*). The peak streaming data rate is similar in each case, yet similar streaming performance can be obtained using 240 x SATA disks over 4 x EXP5060 trays, but with significantly greater capacity (e.g., 240 TB with 1 TB/disk SATA drivers).

Also given RAID 6 and today's higher quality SATA disks, the risk of data loss is negligible. However RAID rebuilds in a SATA disk configuration may have a greater impact upon performance than in a FC disk solution since they may occur more frequently and will take more time. For many customers, the cost to capacity ratio of a SATA solution will outweigh this performance issue.



The following pages illustrate a **best practice** strategy (commonly called "building blocks") for designing a GPFS/DS5300 storage solution.





# Building Block Strategy

---



## ■ Concept

- Define a smallest common storage unit consisting of servers, controllers and disks.
- Replicate it multiple times until capacity and performance requirements are satisfied.
- Facilitates a “build out as you grow” strategy.

## ■ Issue

- Building blocks work best with LAN configured file systems; they do not work as well with SAN configured file systems

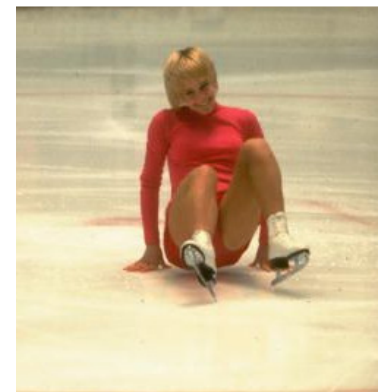
# Building Block Strategy

## Balance

- ▶ Ideally, an I/O subsystem should be balanced
  - Do not make one part of storage system fast and another slow
    - Overtaxing some components of the I/O subsystem may disproportionately degrade performance

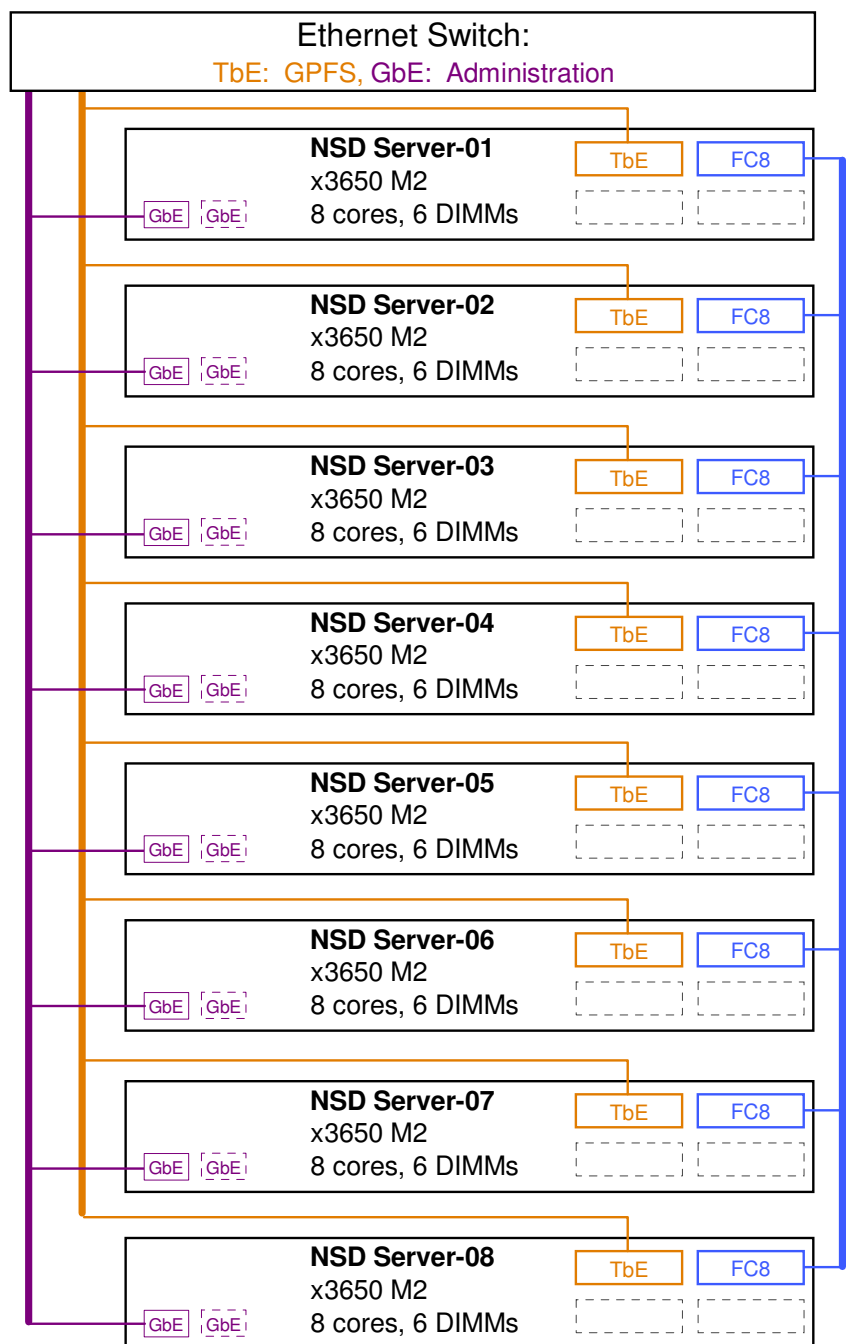
**Warning:** customer requirements may make this goal unachievable

- ▶ “Performance is often inversely proportional to capacity.”  
Todd Virnoche, Business Partner Enablement, IBM
- ▶ Number of disks needed to meet capacity exceeds performance
  - Common example: data warehouses
- ▶ Number of disks needed to meet performance exceeds capacity
  - Common example: national labs, university computing centers



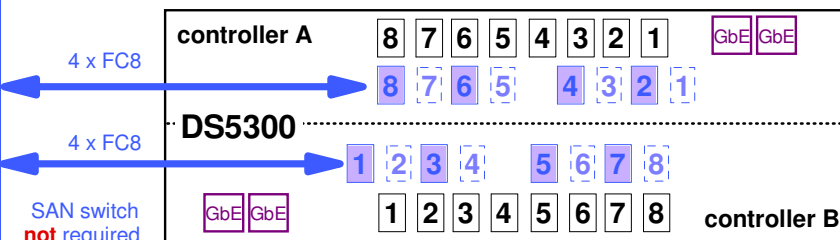


# Building Block #1 (Ethernet)



## Performance Analysis

- ▶ DS5300 streaming data rate
  - 128 x 15Krpm FC disks: write < 4.4 GB/s, read < 5.4 GB/s
    - based on 4M transactions using 4+P RAID 5
  - 240 x 7200 RPM SATA disks: write < 4.0 GB/s, read < 4.8 GB/s
    - based on 4M transactions using 8+2P RAID 6
- ▶ DS5300 IOP rate
  - 448 x 15Krpm disks: write < 18,000 IOP/s, read < 56,000 IOP/s
    - based on to media 4K transactions
  - 240 x SATA disks: **TBD**
- ▶ 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



The GbE administrative network is not illustrated in this diagram.

## Tray Options

- option #1: 8 x EXP5000, 128 x 15Krpm FC disk
- option #2: 4 x EXP5060, 240 x 7200 SATA disks

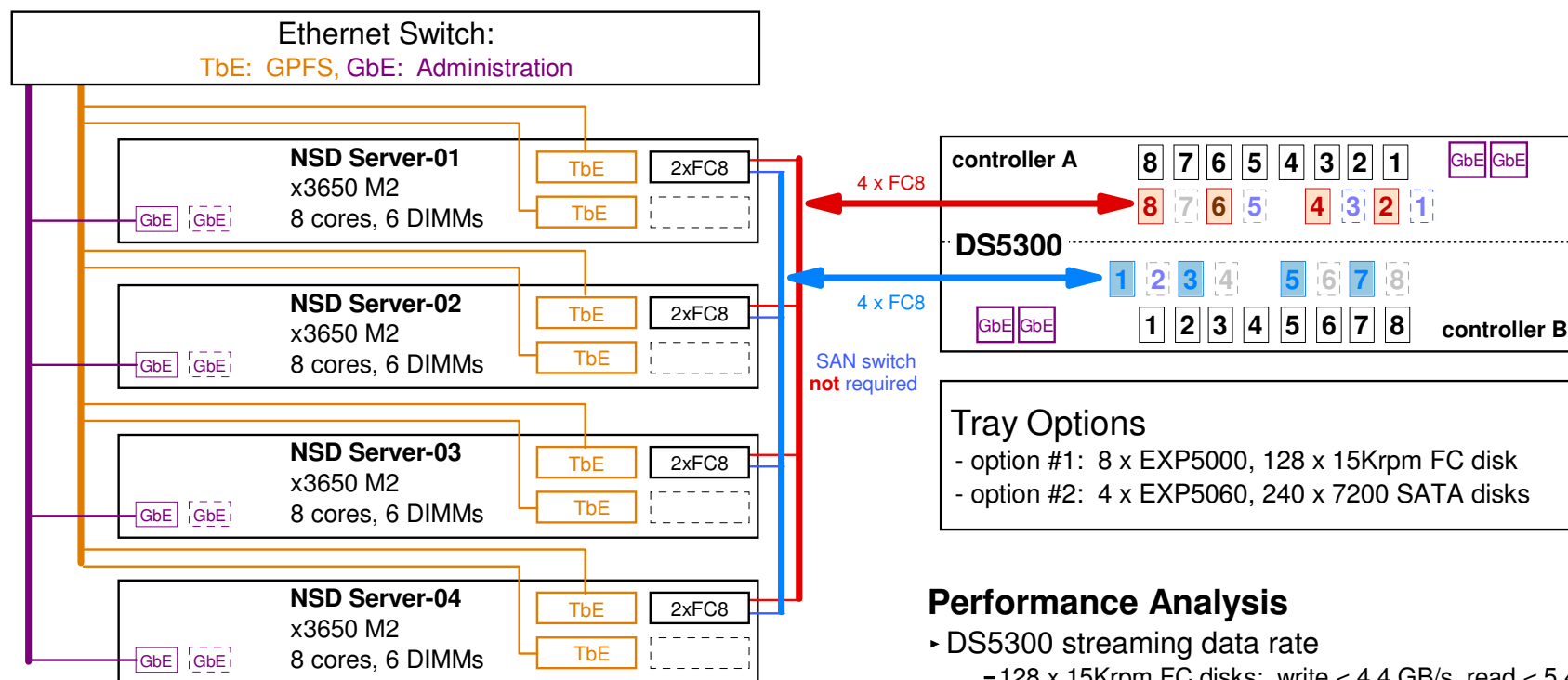
Use appropriate drive side cabling as shown in earlier slides.

## COMMENT:

- ▶ If HBA fail over 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.



## Building Block #2 (Ethernet)



The GbE administrative network is not illustrated in this diagram.

Use appropriate drive side cabling as shown in earlier slides.

### Tray Options

- option #1: 8 x EXP5000, 128 x 15Krpm FC disk
- option #2: 4 x EXP5060, 240 x 7200 SATA disks

### Performance Analysis

- DS5300 streaming data rate
  - 128 x 15Krpm FC disks: write < 4.4 GB/s, read < 5.4 GB/s
    - based on 4M transactions using 4+P RAID 5
  - 240 x 7200 RPM SATA disks: write < 4.0 GB/s, read < 4.8 GB/s
    - based on 4M transactions using 8+2P RAID 6
- DS5300 IOP rate
  - 448 x 15Krpm disks: write < 18,000 IOP/s, read < 56,000 IOP/s
    - based on to media 4K transactions
  - 240 x SATA disks: **TBD**
- 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
  - 1560 MB/s per 2xFC8 is possible, but 1400 MB/s is required

### Comments on Performance

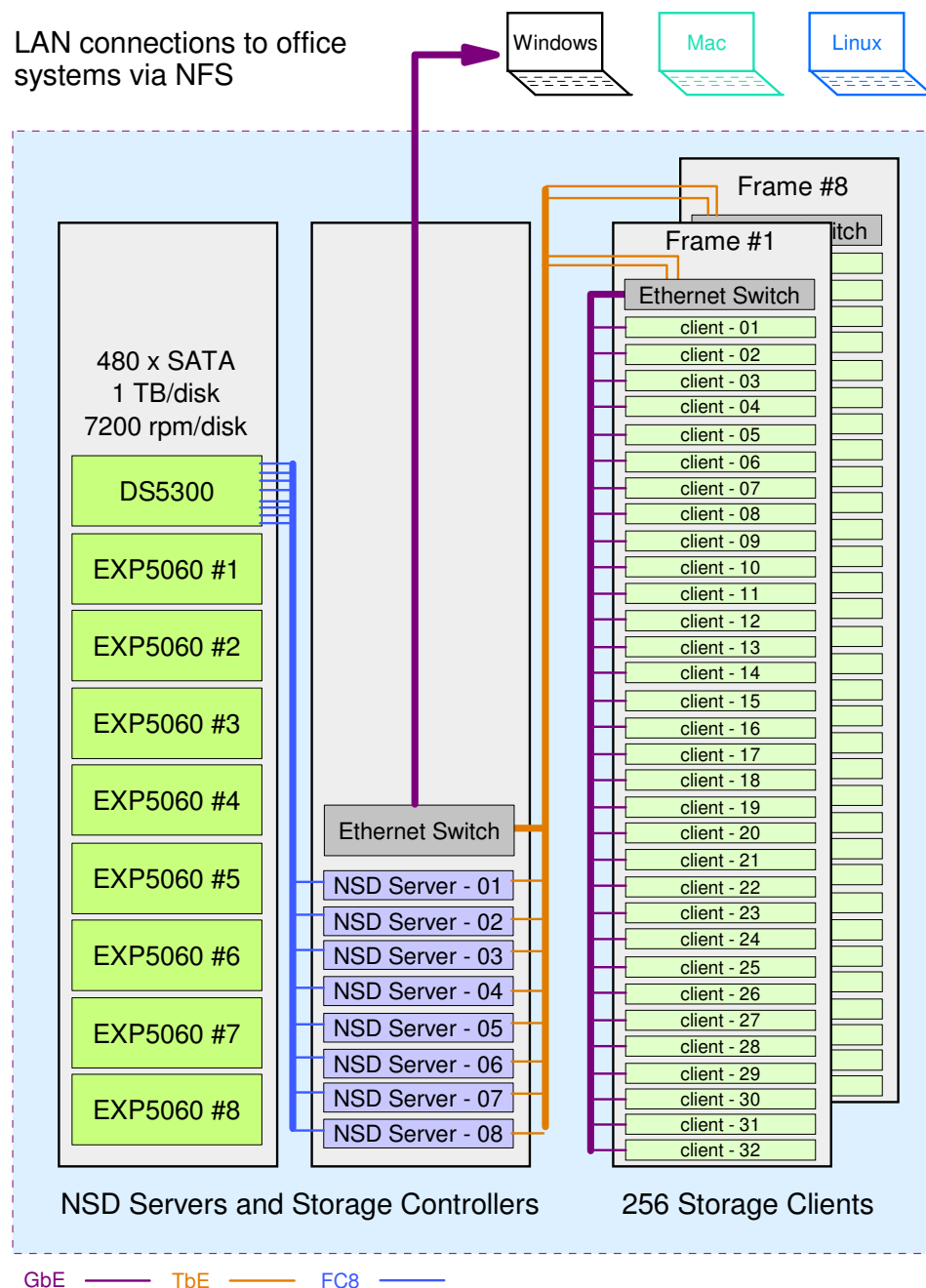
- Each server must sustain 2800 MB/s total I/O BW (n.b., 1400 MB/s over the TbE ports and 1400 MB/s over the 2xFC8 HBA. Lab tests using synthetic benchmarks show that the current x3650 M2 (with PCI-E Gen 1) can deliver 7.3 GB/s duplex BW, so this is **not** likely to be a problem!
- **CAUTION:** In practice it is often difficult to uniformly balance the load over multiple Ethernet links using link aggregation protocols. Since 1400 MB/s is required over dual TbE ports, there is little room for error. **User's are encouraged to validate this solution before putting it into production.** The solution on the previous page is safer since it relies on a single TbE port per NSD server.





# EXAMPLE

## Capacity Optimized Solution Using Building Block #1



### Statistics

- ▶ 1 Building Blocks
- ▶ 256 GPFS Clients
- ▶ Capacity
  - raw  $\approx$  480 TB
  - usable  $\approx$  384 TB
- ▶ Data Rates<sup>1</sup>
  - Streaming performance
    - write
      - rate < 4.0 GB/s
      - average rate per node  $\approx$  16 MB/s
      - ratio  $\approx$  10.5 MB/s per TB
    - read
      - rate < 4.8 GB/s
      - average rate per node  $\approx$  19 MB/s
      - ratio  $\approx$  12.8 MB/s per TB

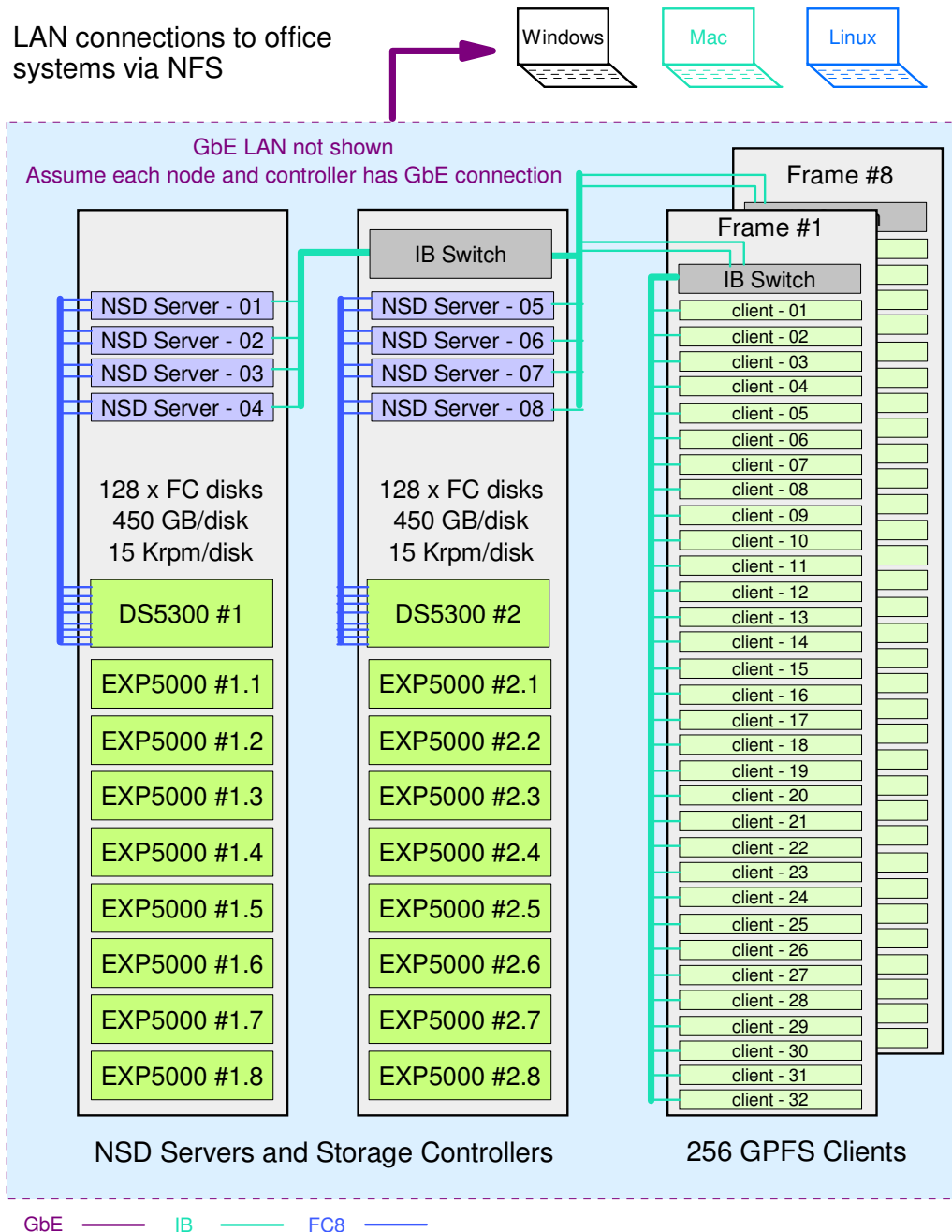
### Footnotes:

1. Data rates are interpolated from earlier benchmark results. Validation testing is recommended.



# EXAMPLE

## Performance Optimized Solution Using Building Block #3



### Statistics

- ▶ 2 Building Blocks
- ▶ 256 GPFS Clients
- ▶ Capacity
  - raw  $\approx$  112 TB
  - usable  $\approx$  84 TB
- ▶ Data Rates<sup>1</sup>
  - Streaming performance
    - write
      - aggregate rate < 8.5 GB/s
      - average rate per node  $\approx$  34 MB/s
      - ratio  $\approx$  104 MB/s per TB
    - read
      - rate < 10.5 GB/s
      - average rate per node  $\approx$  42 MB/s
      - ratio  $\approx$  128 MB/s per TB
  - IOP performance
    - 4K transactions with unfavorable caching
    - write
      - rate < 10,000 IOP/s
      - average rate per node  $\approx$  40 IOP/s
      - ratio  $\approx$  120 IOP/s per TB
    - read
      - rate < 32,000 IOP/s
      - average rate per node  $\approx$  125 IOP/s
      - ratio  $\approx$  380 IOP/s per TB

### Footnotes:

1. Data rates are interpolated from earlier benchmark results.  
Validation testing is recommended.





---

The following pages contain information on common NSD servers used with the DS5300. These servers can generally be interchanged on a 1 to 1 basis, but there are nuanced differences.



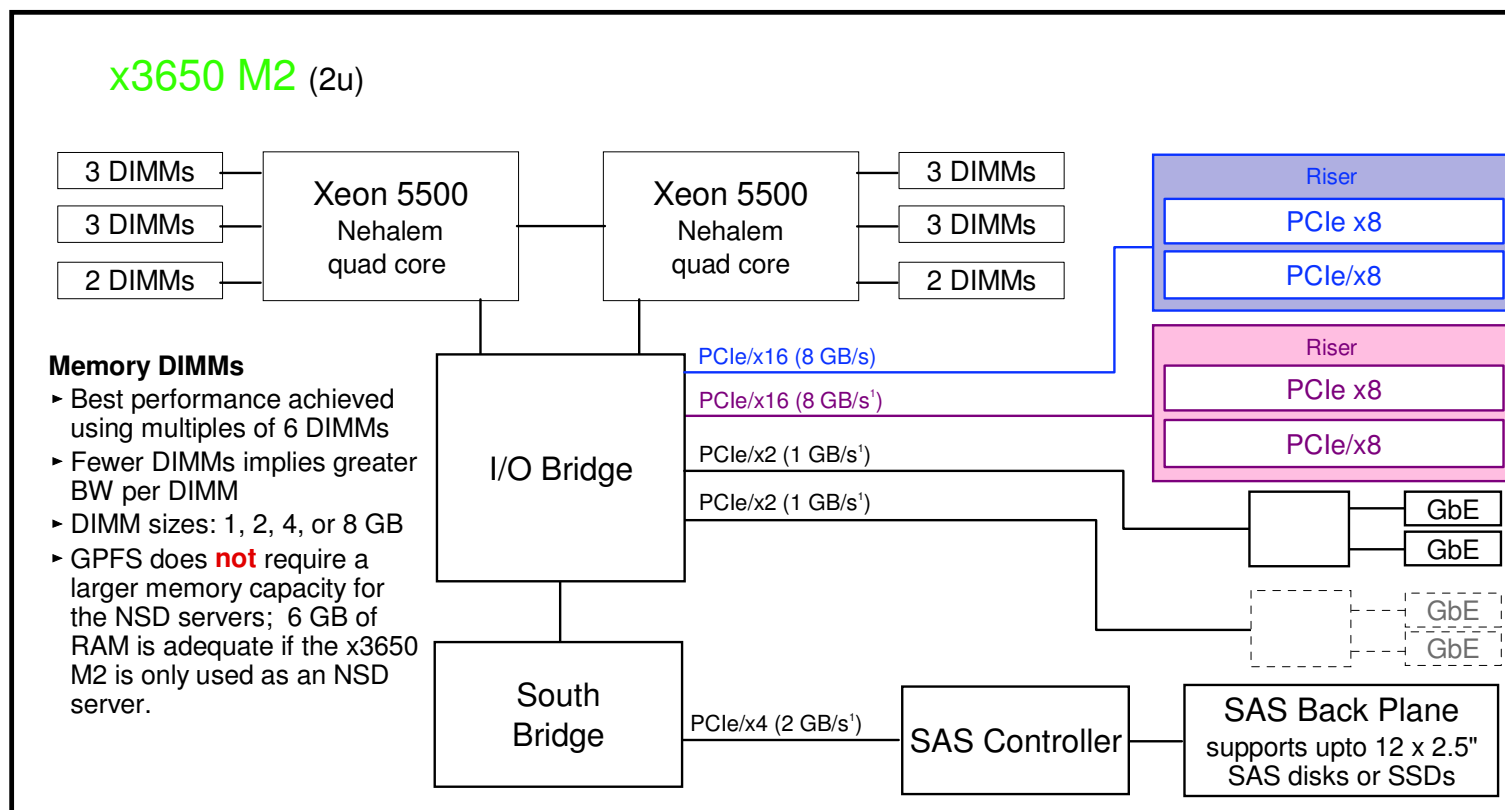
## x3650 M2 System Architecture

The work horse...



The x3650 M2 is a common and cost effective storage server for GPFS in most System X environments (it can even be used with System P).

This diagram illustrates those features most useful to its function as a storage server.



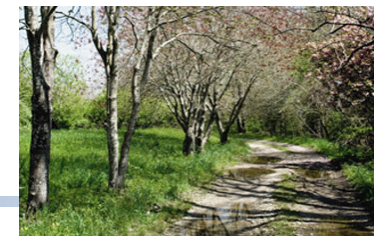
**Data rates are based on the Gen 1 PCIe standard\*.**

1. Listed bus rates are theoretical duplex rates assuming on 512 MB/s per link. Production data will be less.
2. Actual peak duplex rates for PCIe x8 adapter < 3.2 GB/s
3. Measured aggregate over 4 x PCIe x8 adapters < 7.3 GB/s

\* See <http://en.wikipedia.org/wiki/PCIe> for details on the PCI Express standard



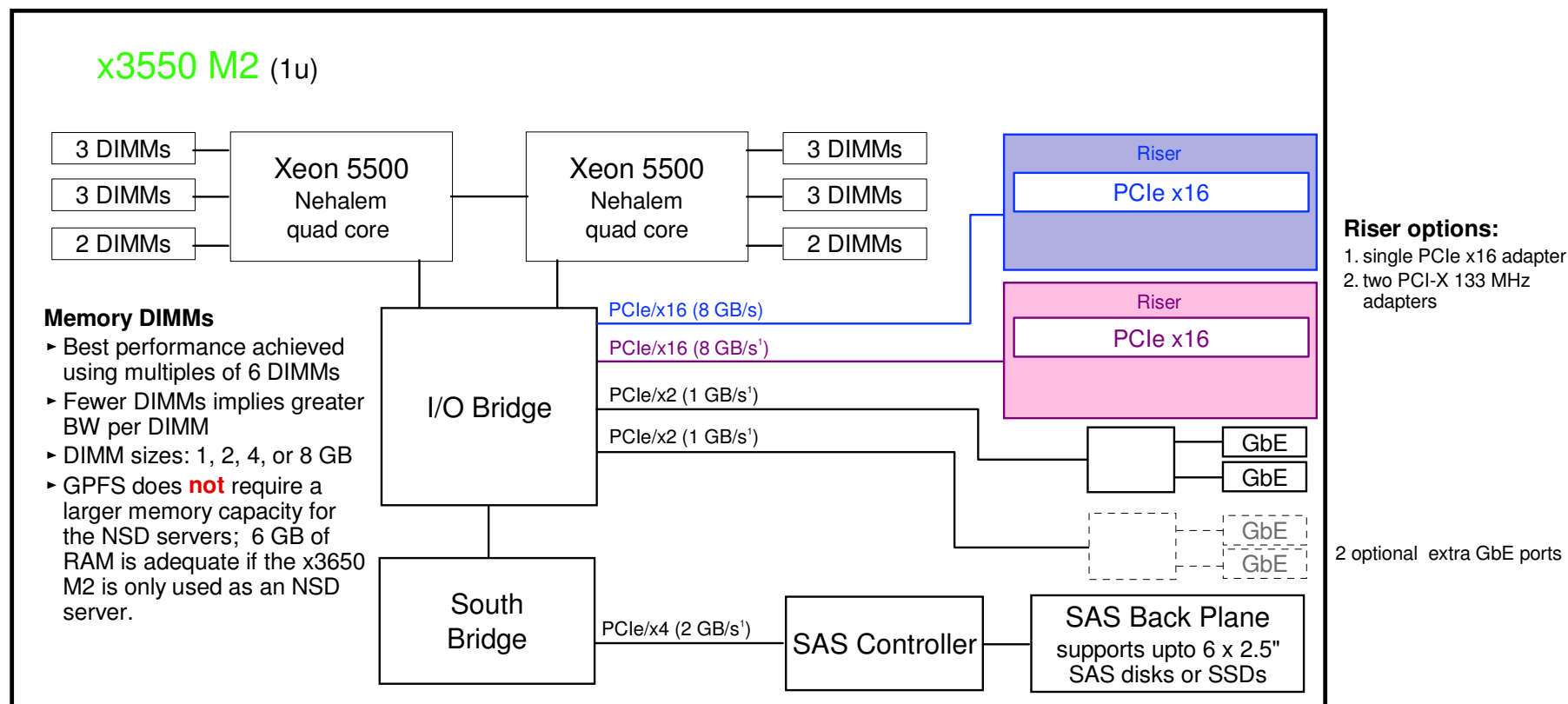
A road less traveled...



## x3550 M2 System Architecture

The x3550 may be a cost effective storage server under some cases for GPFS where it is needed only for disk I/O service; it's main limitation is a lack of PCI-E slots.

This diagram illustrates those features most useful to its function as a storage server.



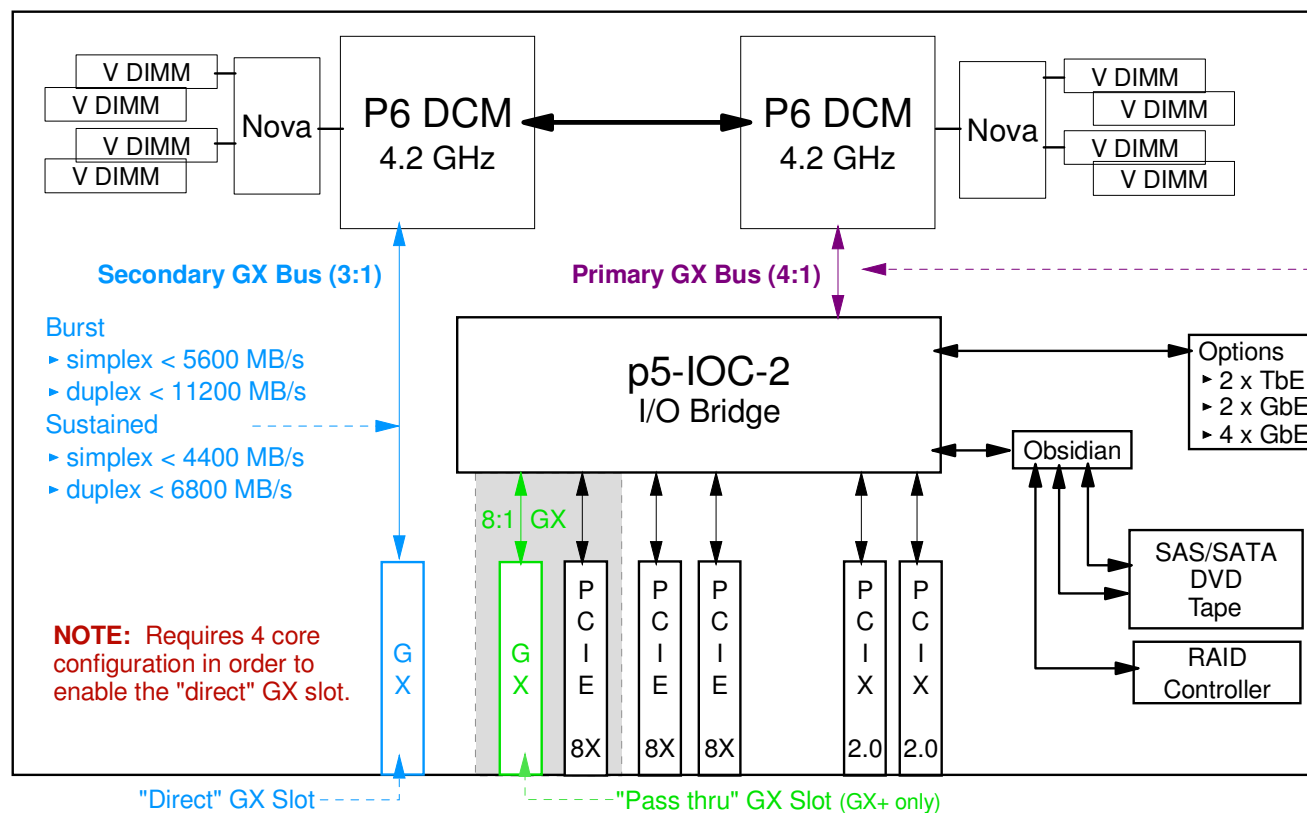
**Data rates are based on the Gen 1 PCIe standard\*.**

1. Listed bus rates are theoretical duplex rates assuming on 512 MB/s per link. Production data will be less.
2. Actual peak duplex rates for PCIe x8 adapter < 3.2 GB/s
3. Measured aggregate over 4 x PCIe x8 adapters < 7.3 GB/s

\* See <http://en.wikipedia.org/wiki/PCIe> for details on the PCI Express standard



# POWER 520 Express System Architecture



## "DIRECT" GX Slot

- IB cards are **only** supported in this slot.
- Card options:
  - dual port, IB 12xSDR @ 6:1 ratio (GX+)
  - dual port, IB 12xDDR @ 3:1 ratio (GX++)
  - RIO2 card @ 8:1 (GX+)
- 12x IB ports 1X and 4X cables
  - requires special "width changer" cable
- GX Bus width: 32 bits
- Rules of thumb:
  - Sustained simplex rates < 80% of simplex burst rate
  - Sustained duplex rates < 60% of duplex burst rate
  - single SDR "lane"
    - burst < 250 MB/s, sustained < 185 MB/s
  - single DDR "lane"
    - burst < 500 MB/s, sustained < 375 MB/s

## "Pass Thru" GX Slot

- The pass thru GX slot occupies the same physical space as the 1st PCI-E slot. Therefore you can not use both of these slots.
- Supports the RIO2 card @ 8:1 (GX+). It does **not** support IB card.

## Single PCI Adapter Data Rates

- PCI-E 8x:
  - Simplex: Burst < 2000 MB/s, Sustained < 1400 MB/s
  - Duplex: Burst < 4000 MB/s, Sustained < 2100 MB/s
- PCI-X 2.0
  - Burst < 2000 MB/s, Sustained < 1400 MB/s (this is not a duplex protocol)

## Overview

This server is a cost effective storage server for GPFS in most System p clusters using Ethernet. This diagram illustrates those features most useful to its function as a storage server.



## Acknowledgments

---

The author would like to thank the following individuals from LSI for their technical assistance and providing the resources used for the benchmarking.

- Tim Chau, Performance Test Engineer
- Robert Dilmore, Principal HPC Architect
- Dean Fairchild, Business Development Executive
- Mark Regester, Storage Performance Analyst
- Con Rice, Modular Storage Product Specialist