SGI® InfiniteStorage

The IS5600 Series

Powered By NetApp

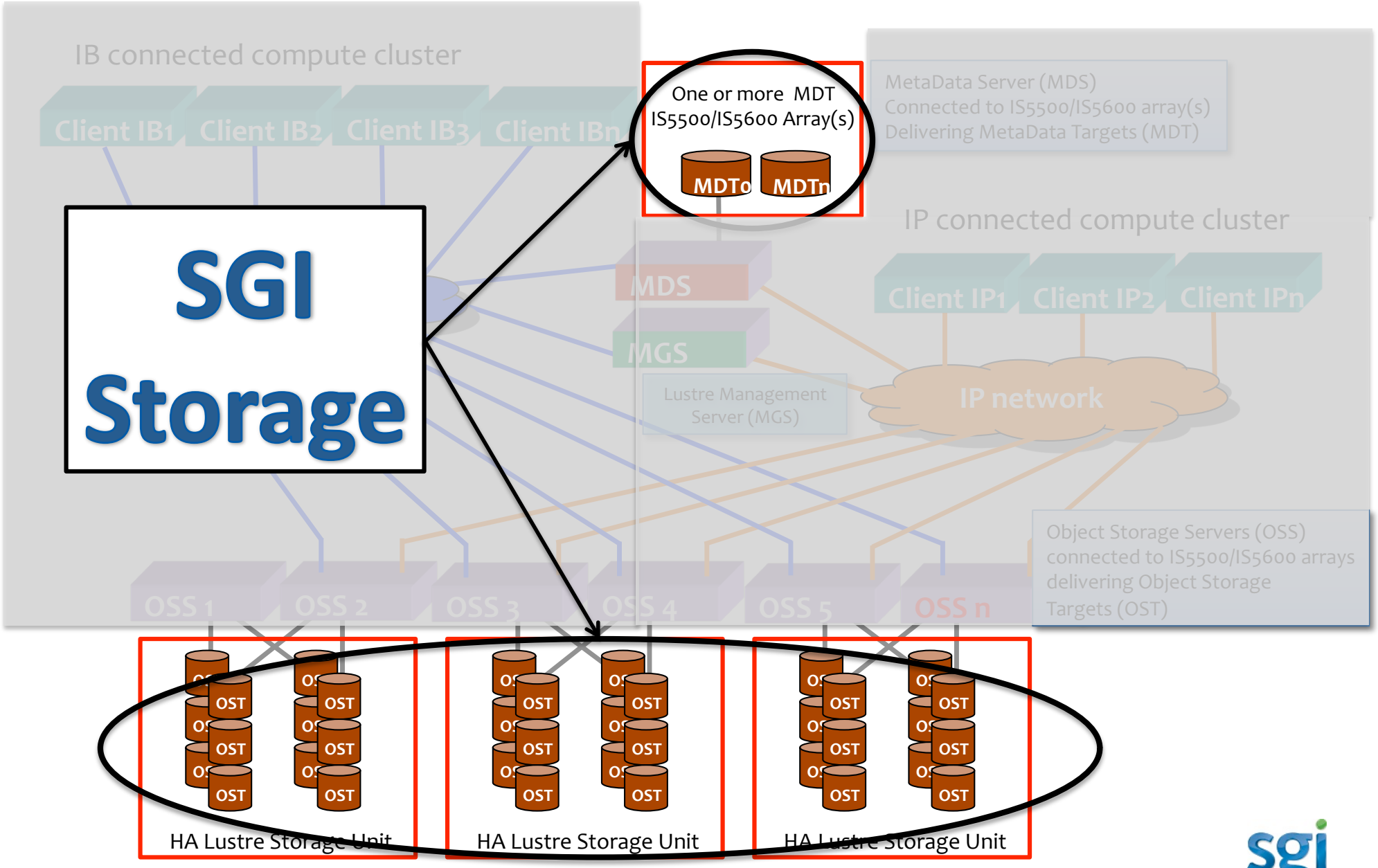
SGI IS5600 Lustre Scalable Units

NetApp – SGI DMF User's Group Update

Agenda

- Overview of the IS5600
- How to architect a Lustre File System to meet customer requirements
  - Performance planning
  - Capacity planning

Basic Lustre Architecture



Lustre Components

- **Clients** – either IP or IB connected compute clusters
- Management Server (**MGS**)
- Metadata Server (**MDS**)
- Metadata Target (**MDT**) provided by IS5500/IS5600 array(s)
- Object Storage Server (**OSS**)
- Object Storage Target (**OST**) provided by IS5500/IS5600 arrays
- Lustre Storage Unit (**LSU**)
  - An SGI term used to describe a scalable unit used to architect a Lustre File System with a specified Bandwidth and Capacity.

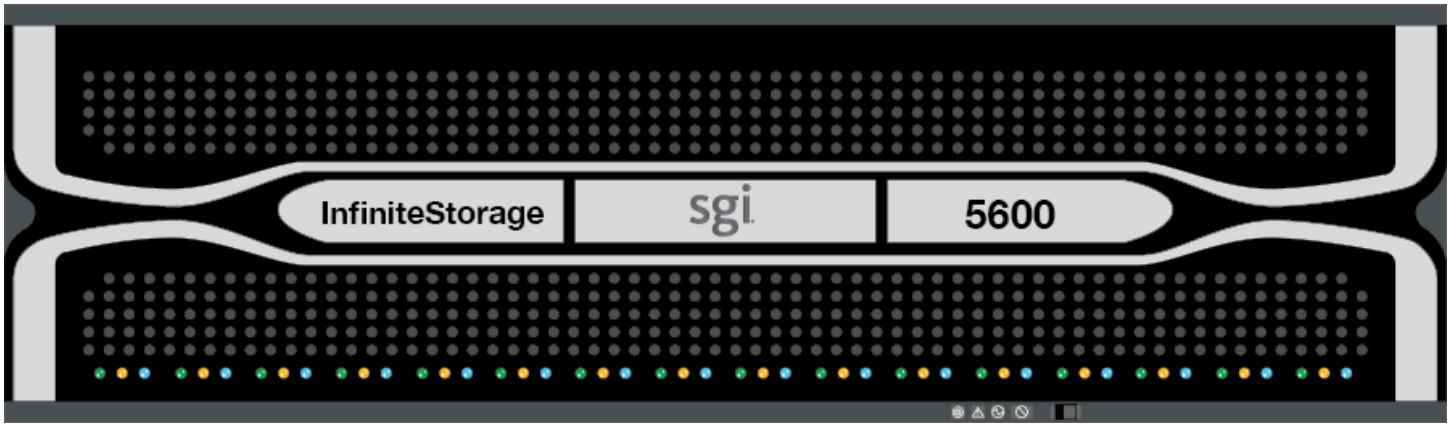
Lustre Metadata Target (MDT)

- MDT = IS5600 array(s) providing one or more LUNs
- Contains mapping of Lustre files to objects

Lustre Object Storage Target (OST)

- An OST is a LUN
- A LUN can be created on a standard RAID Volume Group or on a Dynamic Disk Pool
- A typical configuration would have 2 or more OSS servers connected to an IS5x00 array
- For this presentation we will discuss how to size the storage system that contains the OSTs that will be connected to the compute nodes (OSS)

SGI® InfiniteStorage IS5600 Lustre Storage Units (IS5600 LSU)



SGI® InfiniteStorage™ 5600

-- Next Gen RAID Platform

- Simple: Easy to use, flexible and highly configurable architecture
  - Dynamic Disk Pools (DDP) that allow for easy addition of capacity to existing
  - Supports a combination of RAID-6 Volume Groups and Dynamic Disk Pools simultaneously
  - Interoperates in Existing SGI 12, 24 & 60 Bay Enclosures

- Seamless: A dynamically scalable storage well suited for DMF Auto-Tiering with CXFS
  - Create large pools of drives that can dynamically expand and contract to match the capacity needs of a given tier of storage
  - DDP offering “No Emergency Storage” so performance continues even during failures and no immediate need to swap out failed drives
  - Dramatically accelerated drive rebuild times as compared to RAID-6 for large capacity drives

- Streamlined: Price/Performance optimized
  - Highest data throughput per footprint / spindle / watt / $ with leading mixed workload performance
  - Optimized all SSD arrays for pure IOPS or SSD + HDD Hybrid array options
  - Streamlined Support using AutoSupport “Phone Home” (ASUP)

IS5x00 Platform Options

Controllers

IS5000
Entry System
Scale to 192 drives, 576TB

IS5500
Midrange Performance
Scale to 384 drives, 1.2PB

IS5600
HPC Performance
Scale to 384 drives, 1.2PB

Enclosures

12bay
Entry System
2U / 3.5” 12 drives
NL-SAS, SSD

24bay
Performance System
2U / 2.5” 24 drives
SAS, SSD

60bay
Dense Performance
4U / 3.5” 60 drives
NL-SAS, SAS, SSD

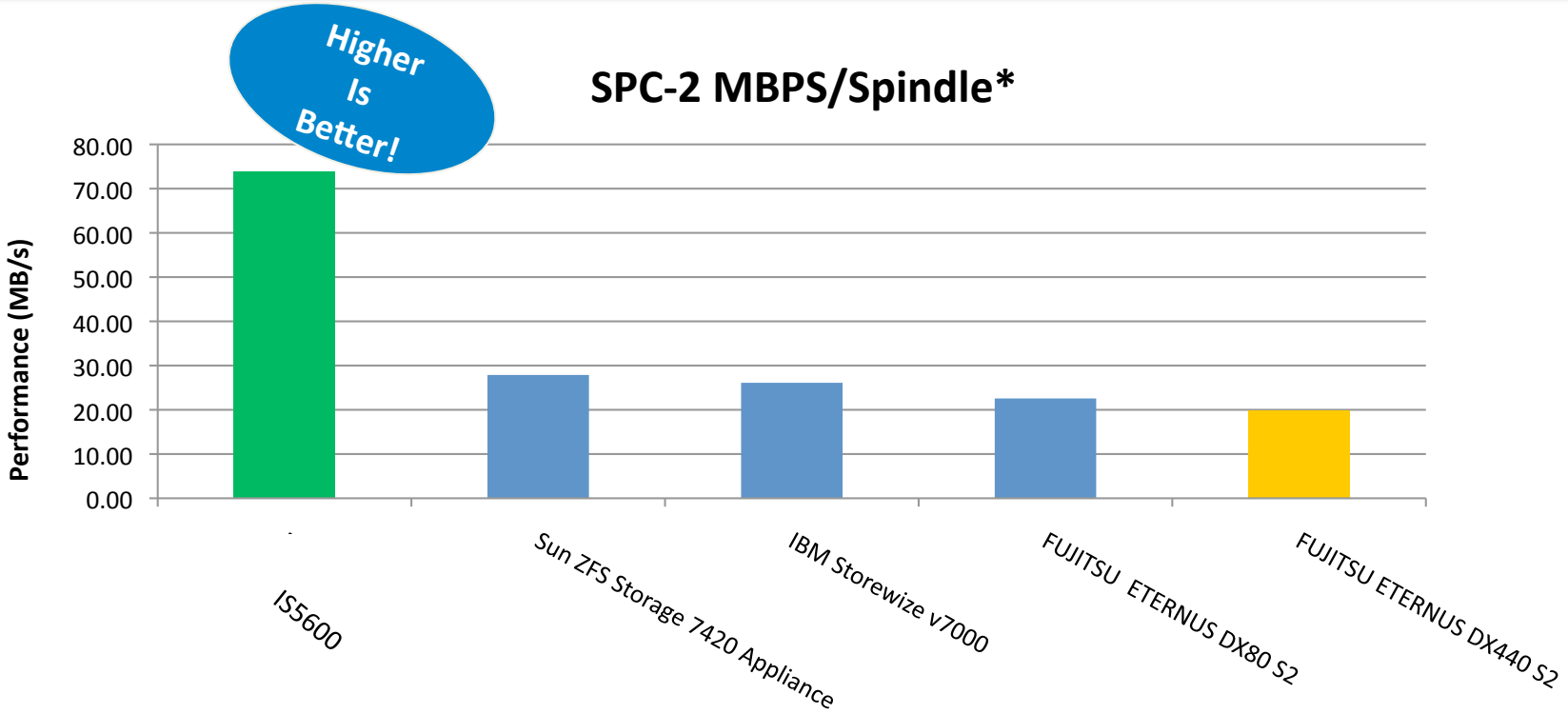
IS5600 Product Specifications

Feature	E5500 Storage Controller
Controller Processor	**Intel Xeon E5-2418L** (“Sandy Bridge”) 2.0 GHz quad core
Controller Type	**Duplex**
Controller Flash	**16GB** for SANtricity OS & cache offload
System Memory	**12** (6/24 GB Cache per controller – (3) (8 GB DIMMS Q3CY2014)
Host Interface (base)	**None**
Host Interface Cards	Quad port **6G SAS** Quad port **16G FC** (GA Q1CY2014) Quad port **12G SAS** (GA Q3CY2014) Dual port **56G IB**(GA Q3CY2014)
Drive Expansion	**Dual 6G SAS**
Drives	**384** (48 2.5” SSD in (2) 24 bay Enclosures)
Power	**125W** (base), **150W** (with HIC)
Enclosure Support	**DE1600 (2u/12)** **DE5600 (2u/24)** **DE6600 (4u/60)**
Performance	**900,000 IOPs*** **6,000 MB/sec****

** RAID6, SAS HIC, 512byte block size max burst Random Read, 384 10k RPM drives*
*** RAID6, SAS HIC, 512K block size Sequential Write, 384 10k RPM drives, Cache Mirroring Enabled*

IS5600 is a leader in Performance Efficiency

IS5600 provides 2.5 times MBps over nearest non-SGI based benchmark

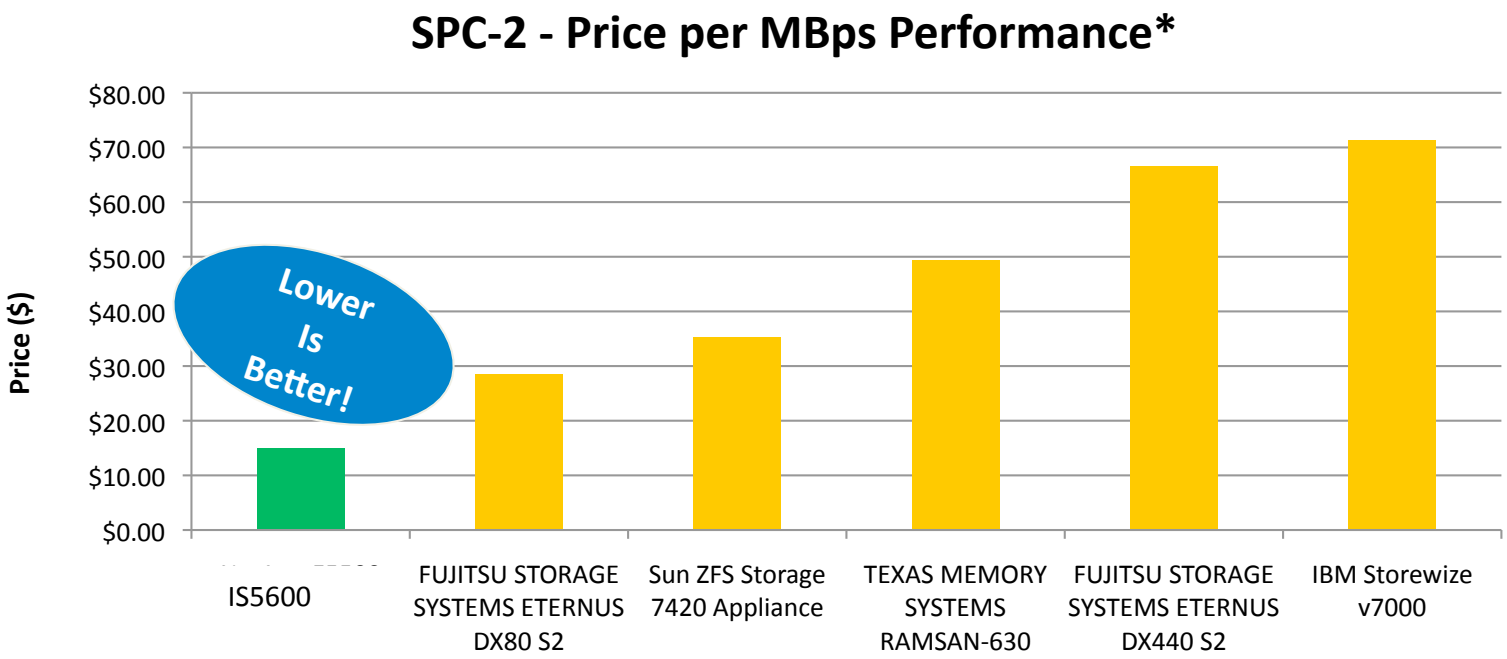


http://www.storageperformance.org/results/benchmark_results_spc2#b00065

** SPC-2 Publications in 2011 or later; Total Price $500K or less*
Comparing IS5600 to non-SGI technology

IS5600 Leads in Performance per $

IS5600 doubles the performance per $

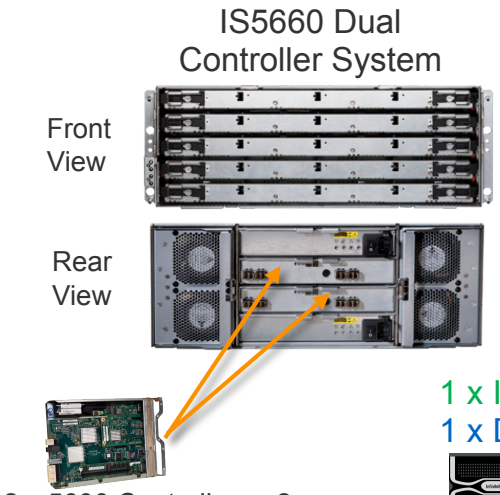


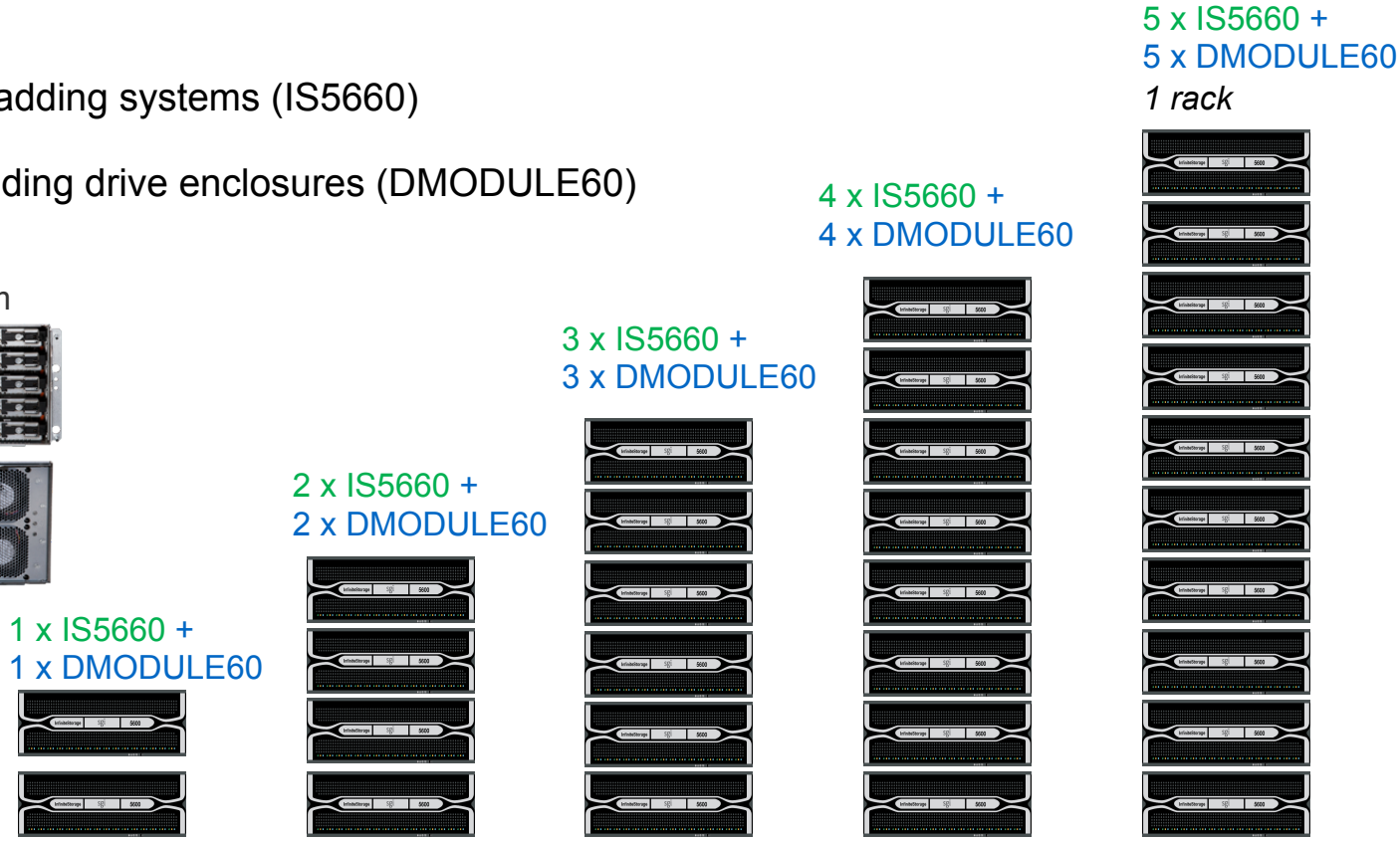
http://www.storageperformance.org/results/benchmark_results_spc2#b00065

** SPC-2 Publications in 2011 or later; Total Price $500K or less*
Comparing IS5600 to non-SGI technology

Lustre Bandwidth Scaling

- Scale bandwidth by adding systems (IS5660)
- Scale capacity by adding drive enclosures (DMODULE60)



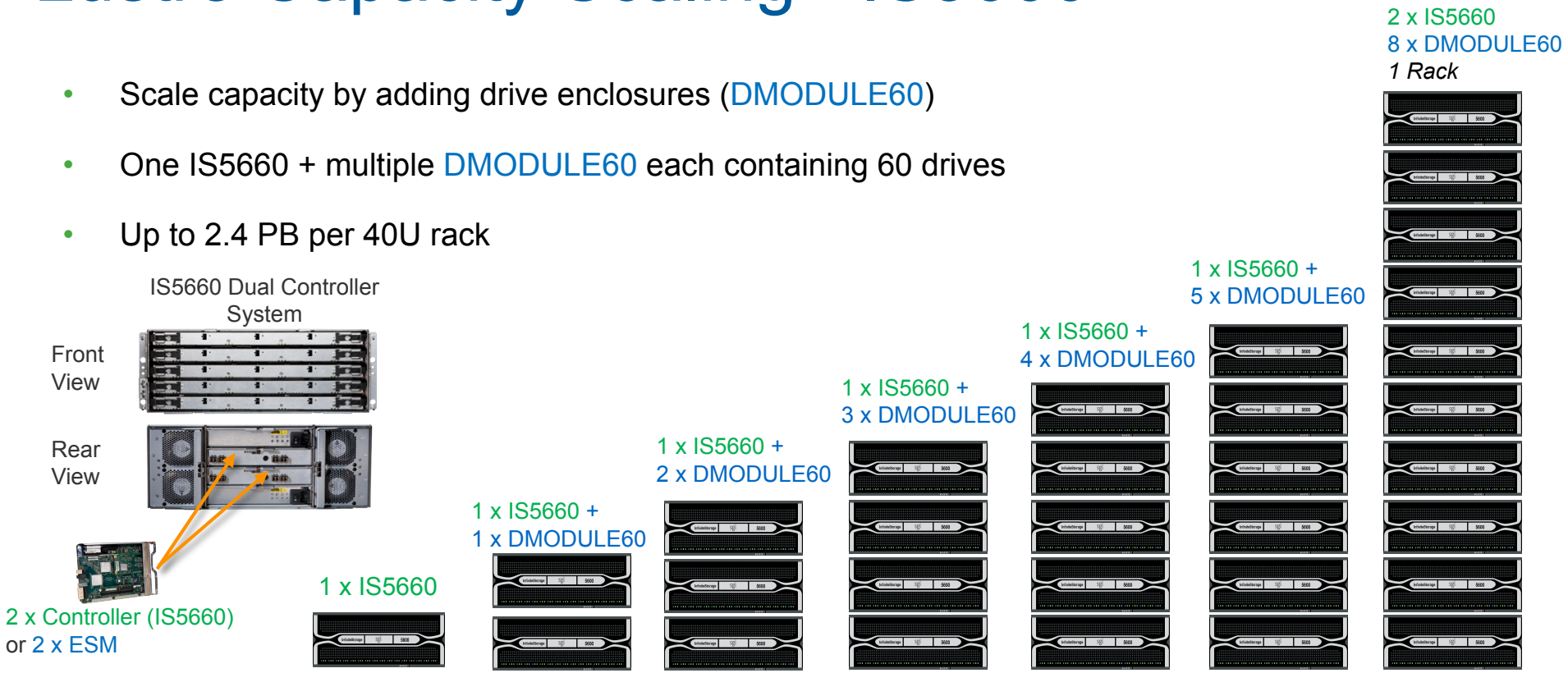


NL-SAS Drives(n)	120	240	360	480	600
Capacity (TB)*	480	960	1440	1920	2400
Bandwidth (GB/s, CME Writes)	5**	10	15	20	25

** 4TB drive RAW capacity*
***Direct IO, IOZone or IOR RAID-6 8+2 100% Sequential Writes CME, NL-SAS 7.2K RPM HDD*

Lustre Capacity Scaling - IS5600

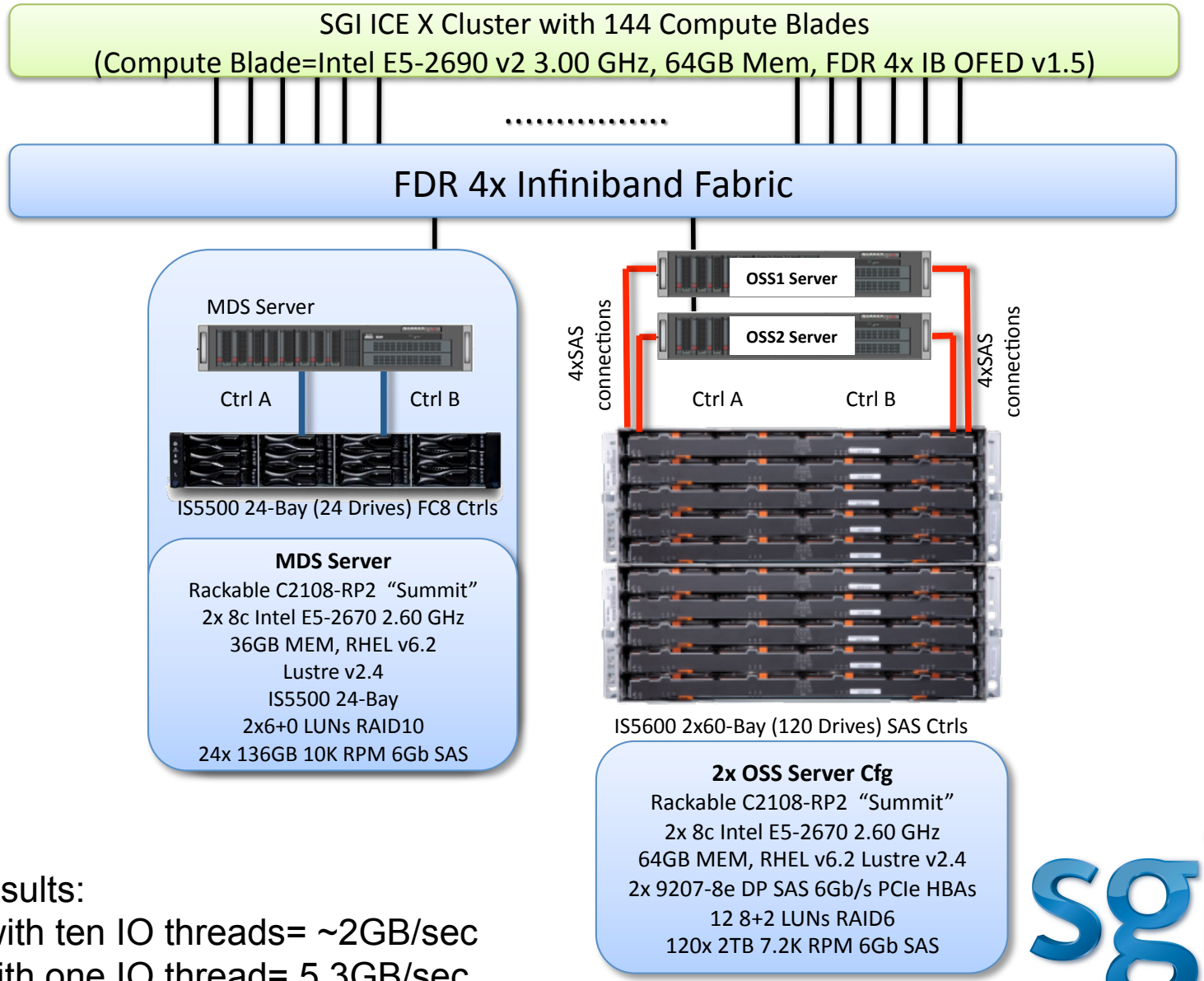
- Scale capacity by adding drive enclosures (DMODULE60)
- One IS5660 + multiple DMODULE60 each containing 60 drives
- Up to 2.4 PB per 40U rack



NL-SAS Drives(n)	60	120	180	240	300	360	600
Capacity (TB)*	240	480	720	960	1200	1440	2400
Bandwidth (GB/s, CME writes)	4	5	5	5	5	5	10

** 4TB drive RAW capacity*
***Direct IO, IOZone or IOR RAID-6 8+2 100% Sequential Writes CME, NL-SAS 7.2K RPM HDD*

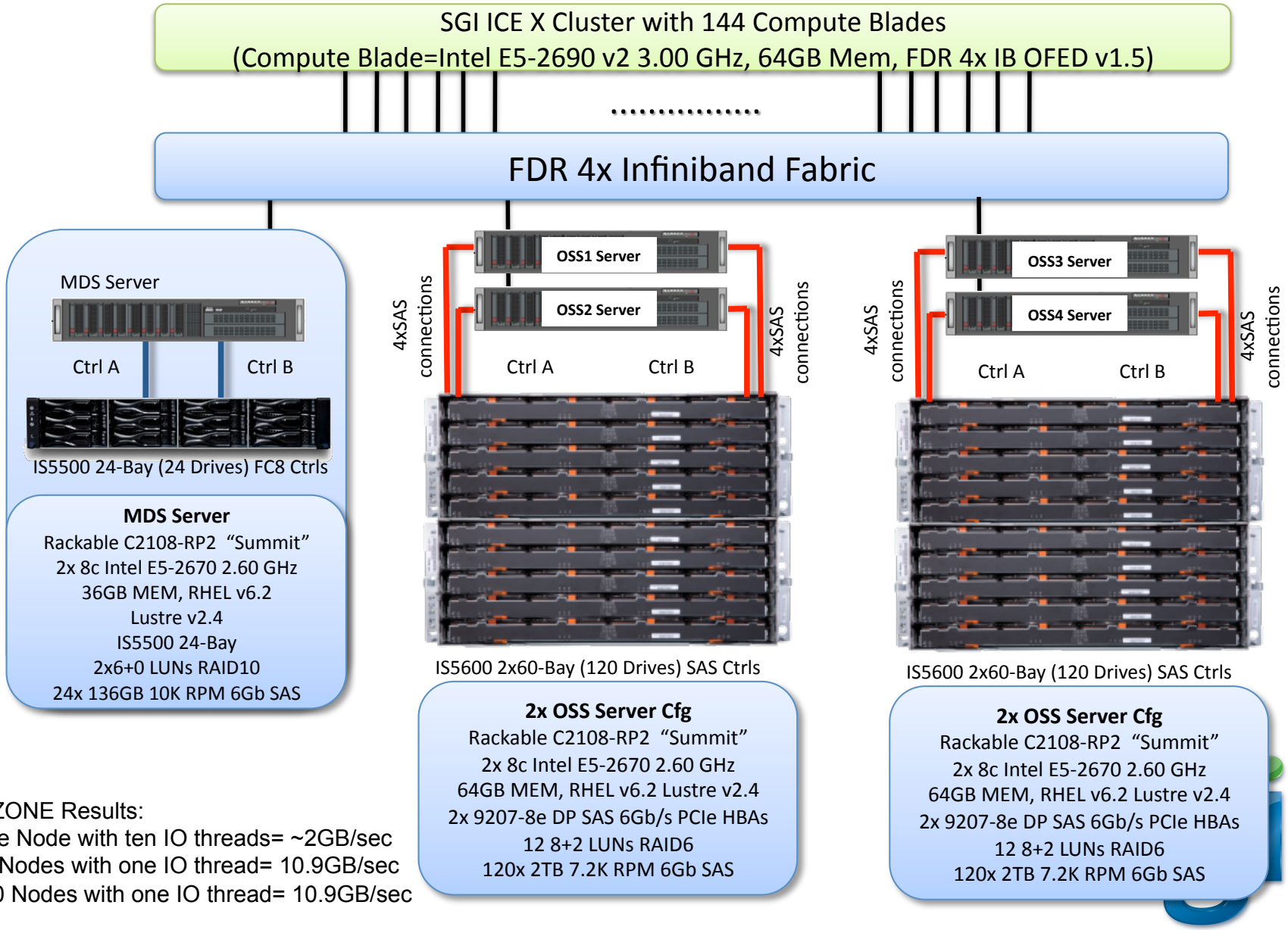
Test 1: One IS5600 with 120 drives with two OSSs, demonstrate max write perf



IOZONE Results:
One Node with ten IO threads= ~2GB/sec
36 Nodes with one IO thread= 5.3GB/sec

- SAS 6Gb/s connection
- FDR Infiniband Network with OFED v1.5.x

Test 2: Two IS5600 with 120 drives with four OSSs, demonstrate max write perf



IOZONE Results:
One Node with ten IO threads= ~2GB/sec
72 Nodes with one IO thread= 10.9GB/sec
100 Nodes with one IO thread= 10.9GB/sec

- SAS 6Gb/s connection
- FDR Infiniband Network with OFED v1.5.x

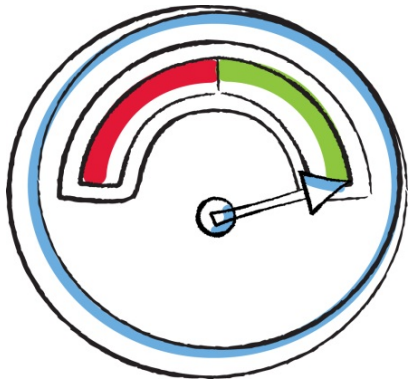
Notable HPC Sites with IS5x00

Where	Customer	Capacity	Systems
Australia	Pawsey Center – Square Kilometer Array	6PB	(**27**) IS5600
Northern California	NASA AMES Advanced Super Computing Division	4PB	(**30**) IS5500
Maryland, USA	NASA Goddard	13PB	(**10**) IS5500
Write Patterson AFB Ohio, USA	TI-11/12 – Wright Patterson AFB	6PB	(**112**) IS5500
Japan	CREIPE	3PB	(**4**) IS5500

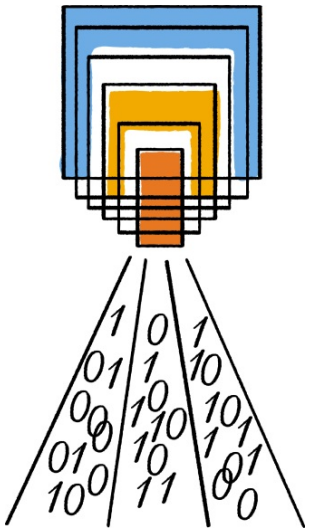
Dynamic Disk Pools (DDP)

Delivering New Levels of Performance and Protection

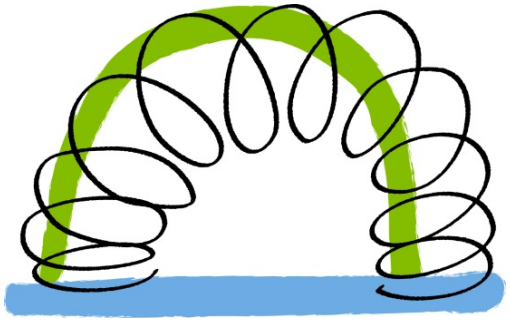
Consistent Performance



Increased Data Protection



Powerful Versatility

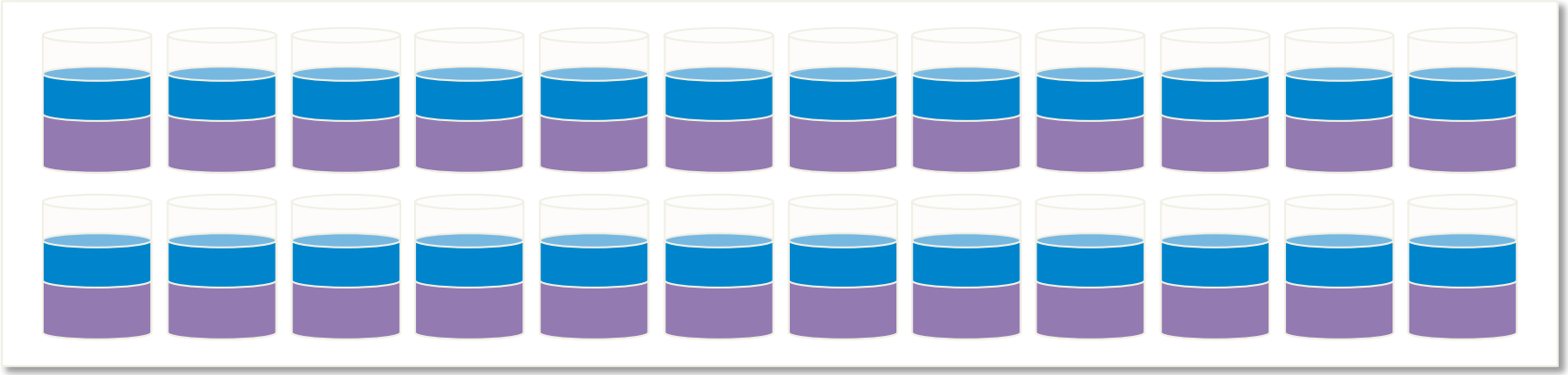


Dynamic Disk Pools Overview

- DDP dynamically distributes data, spare capacity, and parity information across a pool of drives

- Intelligent algorithm defines which drives should be used for data placement (seven patents pending)

- Data is dynamically recreated/redistributed as needed to maintain protection/distribution

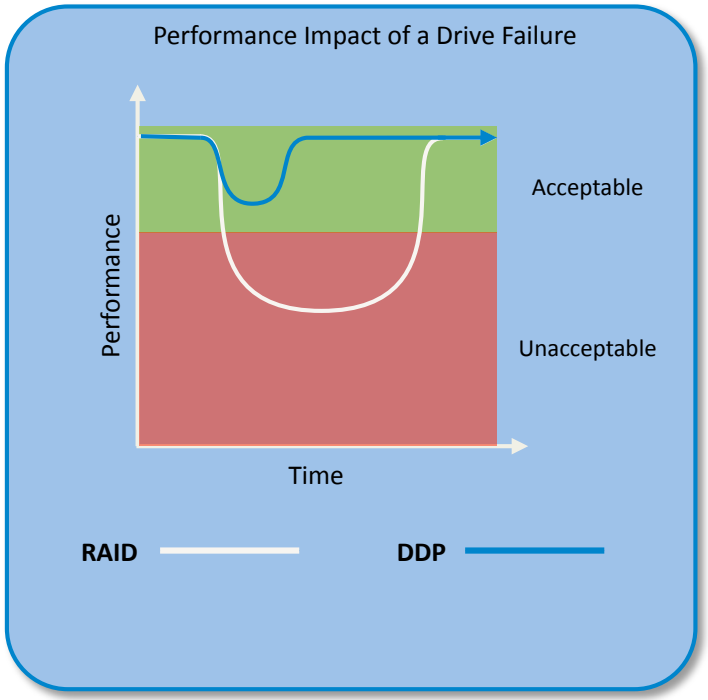


Consistent Performance

Designed to maintain high-speed data delivery even during a drive failure and reconstruction

- Performance drop is minimized during rebuild*
- Rebuild completes up to 8x faster than traditional RAID*

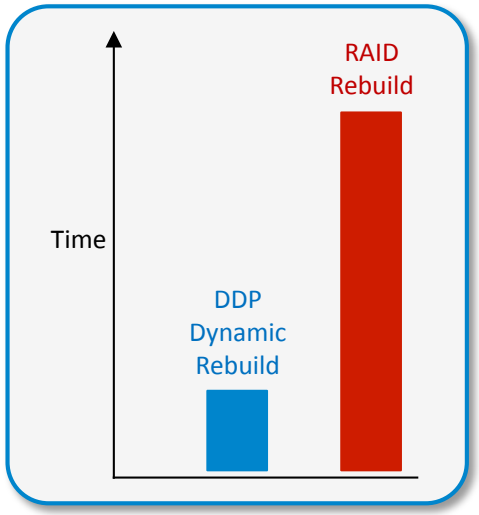
- Result: Significantly more time spent in optimal mode for maximum productivity



Results depend on pool size, drive type, and workload.

Data Protection

- Unmatched protection against drive failures
- Shorter rebuild times reduce exposure to multiple cascading disk failures
- DDP's dynamic rebuild process uses all the drives in a disk pool for the reconstruction of the failed drive
- Critical data within a DDP stripe is given reconstruction priority to protect against data loss should a multiple drive failure occur
- Provides significant improvement in data protection
- Larger pool provides even greater protection



Extreme Versatility

"Right-size" any environment

- Flexible disk pool sizing optimizes shelf utilization
- Multiple ways to implement pools

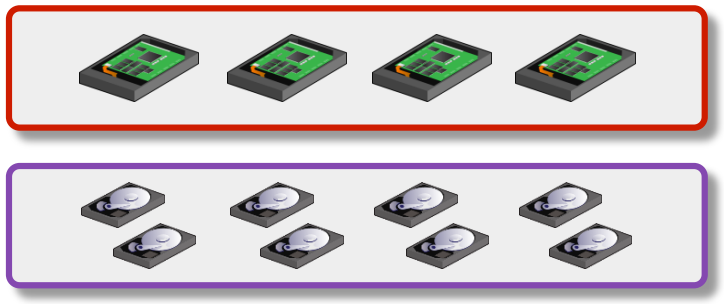
Single pool for all volumes maximizes simplicity, protection, and utilization

Smaller pools with one volume/pool maximize performance for bandwidth applications and clustered file systems

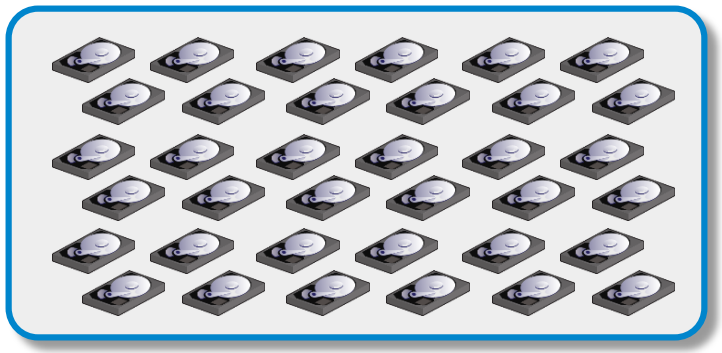
Extreme Versatility

Multiple ways to create storage tiers

- Create multiple pools to meet different requirements
  - Such as: high-performance pool and high-capacity pool
- Intermix traditional RAID and DDP
  - Perfect when smaller RAID 10 or SSD volumes are required for maximum performance



RAID 10 with SSD/15K for Max Performance Volumes



DDP for Capacity Volumes

DDP v RAID-6 Lustre Performance Model

Better Bandwidth Over Time

- Instantaneous bandwidth, RAID-6 is better by approximately 10%
- However factoring in failures that will occur over time, DDP delivers more total performance since you can rebuild a large drive faster in a Disk Pool than in a RAID Group
- Customers are demanding assurances that the storage can recover from failures in a reasonable amount of time

DDP Performance Model Demo

INPUT

Configuration	Optimal Bandwidth	BW reduction due to HDD failure	Average yearly drive failure rate	Time in degraded mode after physical Drive Replacement (Days)
Raid 6 (8+2)	50	30%	2.0%	4
DDP	45	10%	2.0%	0

Do not enter anything below this box.

OUTPUT

Configuration	Total Aggregate Bandwidth Per Year With Expected Drive Failures (PB Per Year)	Total Percentage Increase Per Year in Aggregate Bandwidth Using DDP	Total # Drive Failures Per Year (count)	Time in degraded mode after failure/before replacement. *Note:DDP is rebalancing, RAID 6 is rebuilding (Days)	Total time in degraded mode per drive failure (days)	Total Days Per Year in Degraded Mode (days)
Raid 6 (8+2)	1,355	-0.82%	24	1	5.0	120
DDP	1,344		24	1	1	24

Thank You.

Questions?

NetApp

Thank you