NORTHERN CALIFORNIA NoCOUG ORACLE USERS GROUP

ORACLE Platinum Partner

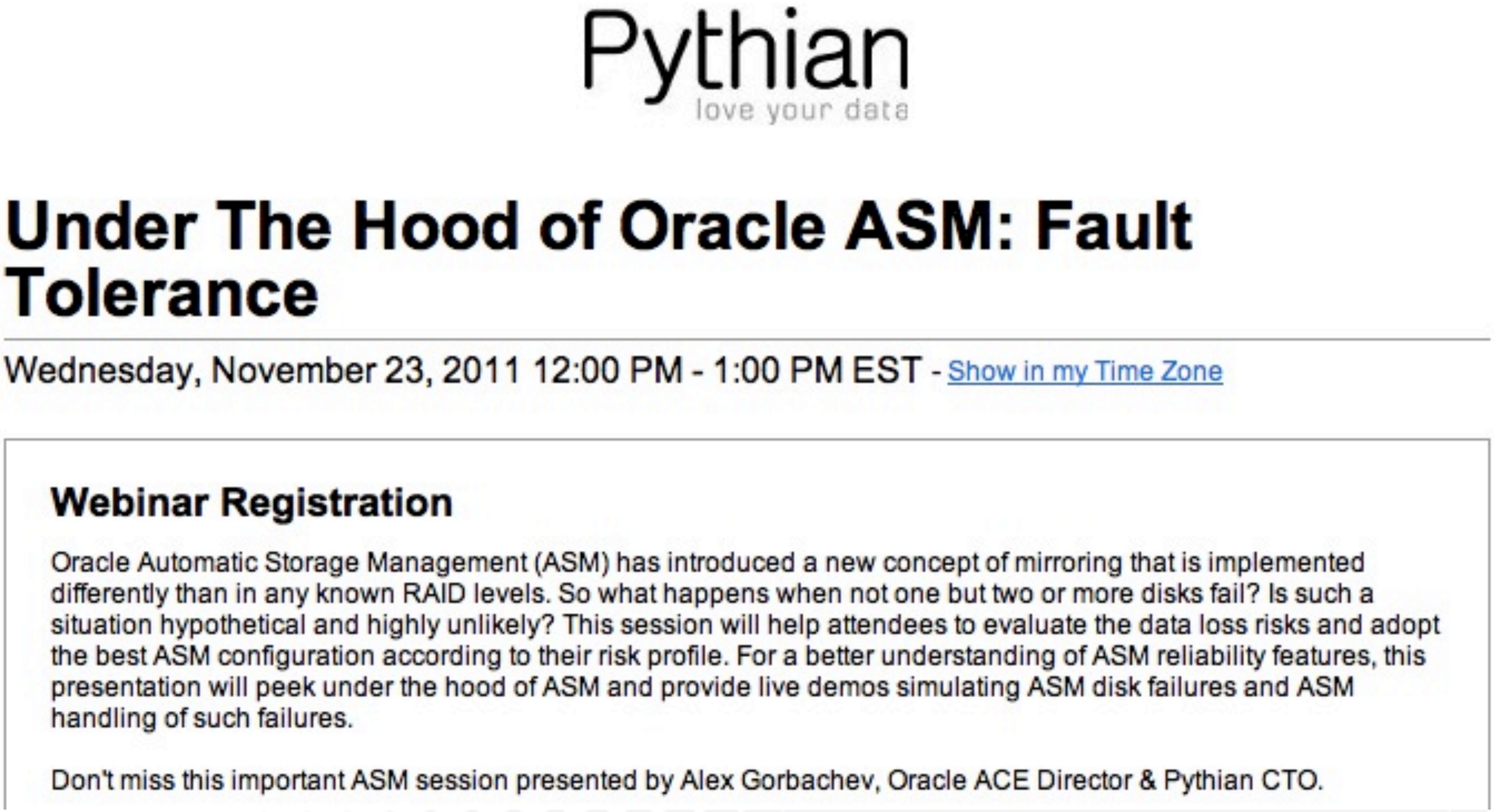
ORACLE PARTNERNETWORK
2011 ORACLE TITAN AWARD NORTH AMERICA

Under the Hood of Oracle Database Appliance

Alex Gorbachev

Mountain View, CA
9-Nov-2011

Pythian
love your data



Under The Hood of Oracle ASM: Fault Tolerance

Wednesday, November 23, 2011 12:00 PM - 1:00 PM EST - Show in my Time Zone

Webinar Registration

Oracle Automatic Storage Management (ASM) has introduced a new concept of mirroring that is implemented differently than in any known RAID levels. So what happens when not one but two or more disks fail? Is such a situation hypothetical and highly unlikely? This session will help attendees to evaluate the data loss risks and adopt the best ASM configuration according to their risk profile. For a better understanding of ASM reliability features, this presentation will peek under the hood of ASM and provide live demos simulating ASM disk failures and ASM handling of such failures.

Don't miss this important ASM session presented by Alex Gorbachev, Oracle ACE Director & Pythian CTO.

http://bit.ly/pythianasmwebinar

Alex Gorbachev

- CTO, The Pythian Group
- Blogger
- OakTable Network member
- Oracle ACE Director
- BattleAgainstAnyGuess.com
- President, Oracle RAC SIG

ORACLE RACSIG

Sydney ORACLE Meetup

Oracle ACE

OakTable.net

Why Companies Trust Pythian

- Recognized Leader:
  - Global industry-leader in remote database administration services and consulting for Oracle, Oracle Applications, MySQL and SQL Server
  - Work with over 150 multinational companies such as Western Union, Fox Interactive Media, and MDS Inc. to help manage their complex IT deployments
- Expertise:
  - One of the world's largest concentrations of dedicated, full-time DBA expertise.
- Global Reach & Scalability:
  - 24/7/365 global remote support for DBA and consulting, systems administration, special projects or emergency response

Harvard University

Rakuten GLOBAL MARKET

TOYOTA

WESTERN UNION

ORACLE Platinum Partner

Penn University of Pennsylvania

FOX SPORTS

NATIONAL GEOGRAPHIC

ORACLE®
PARTNERNETWORK
2011 ORACLE TITAN AWARD
NORTH AMERICA

Oracle Database Appliance

- 2 node RAC cluster-in-a-box with all infrastructure embedded
  - Shared Storage
  - Interconnect
  - Servers

- 2 x dual-socket Oracle Linux servers

– 24 Intel Xeon processor X5675 cores

– 192 GB main memory

- 12 TB raw disk storage
- 292 GB solid state storage
- Built-in redundancy

– Server, storage, network, power and cooling

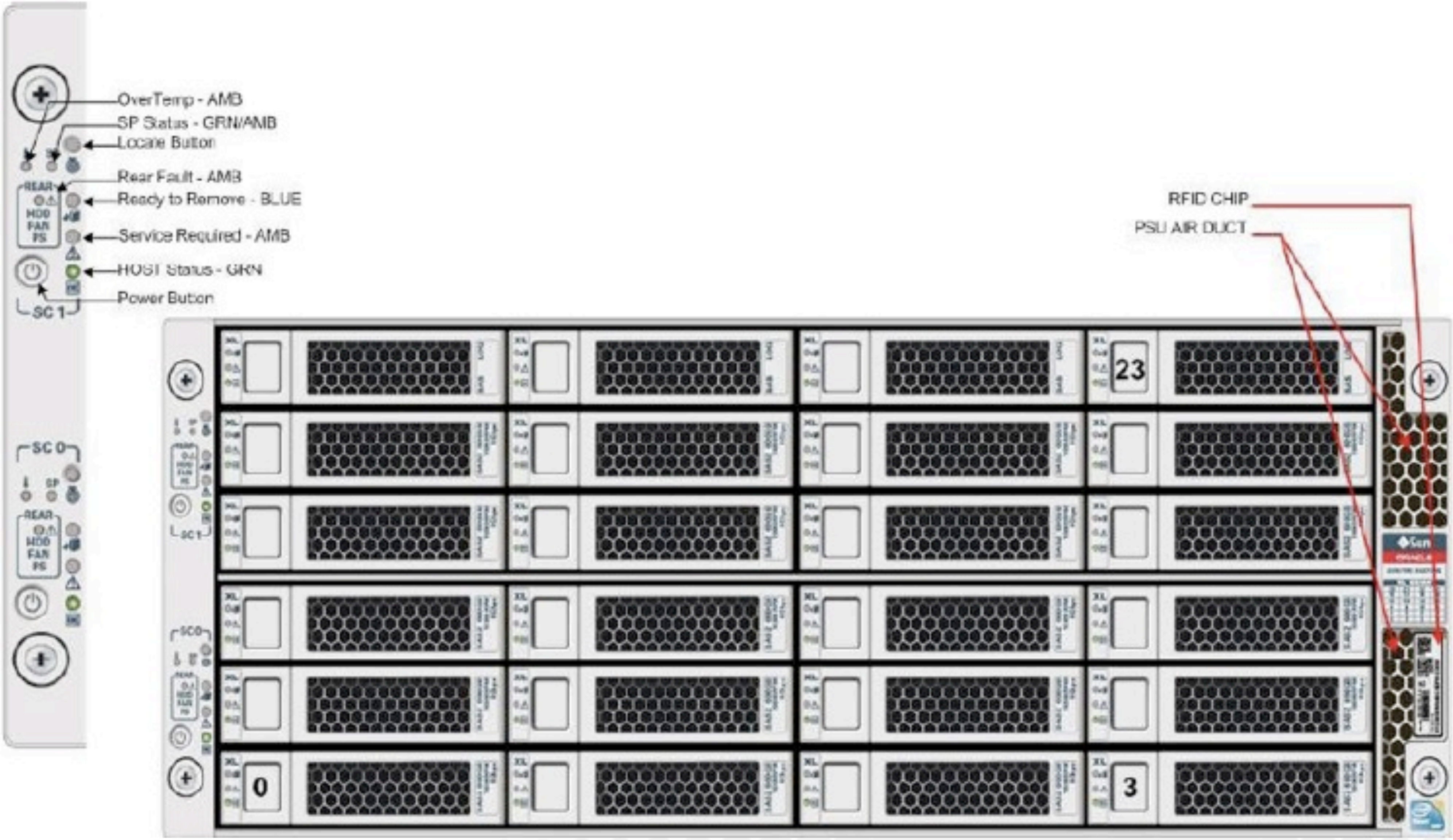
Sun Fire X4370 M2 Overview

FRONT VIEW

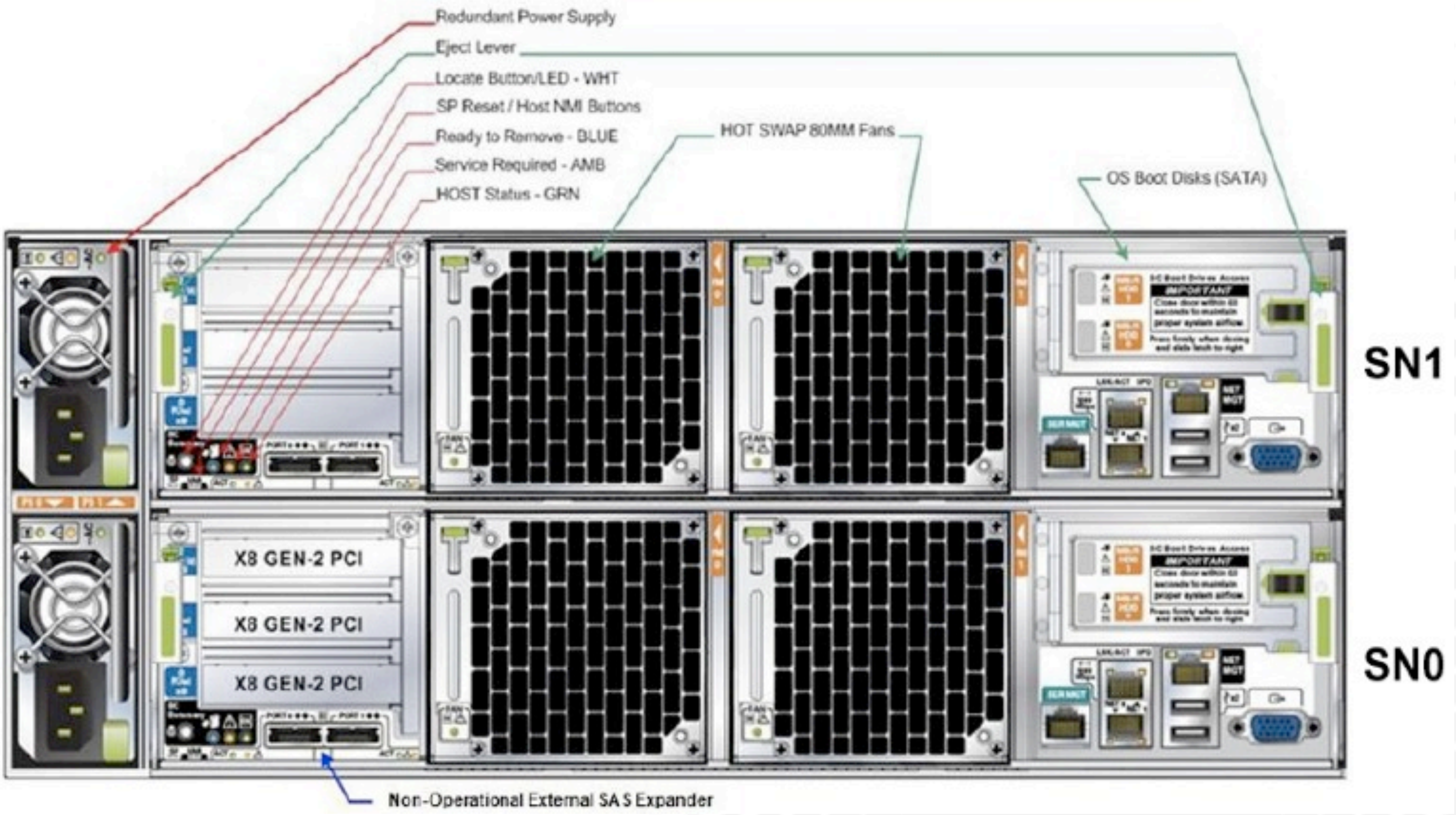
REAR VIEW

- 4U Redundant Storage Server
- 2 Server Nodes (SN)
- 24 3.5" dual ported SAS/SATA/SSD disk slots
  - 20x 600GB 15K RPM SAS (slots 0-19) (Triple-mirrored: 12 TB RAW, 4 TB Usable)
  - 4x 73GB STEC GEN3 SSD (slots 20-23) for redo logs (Triple-mirrored)
- 2 Hot-swap redundant power supplies (A249)
- Redundant 5V and 12V disk backplane power
- Independent power, locate buttons and status per SN
- fixed configuration
- one order number for the hardware + another for the power cord

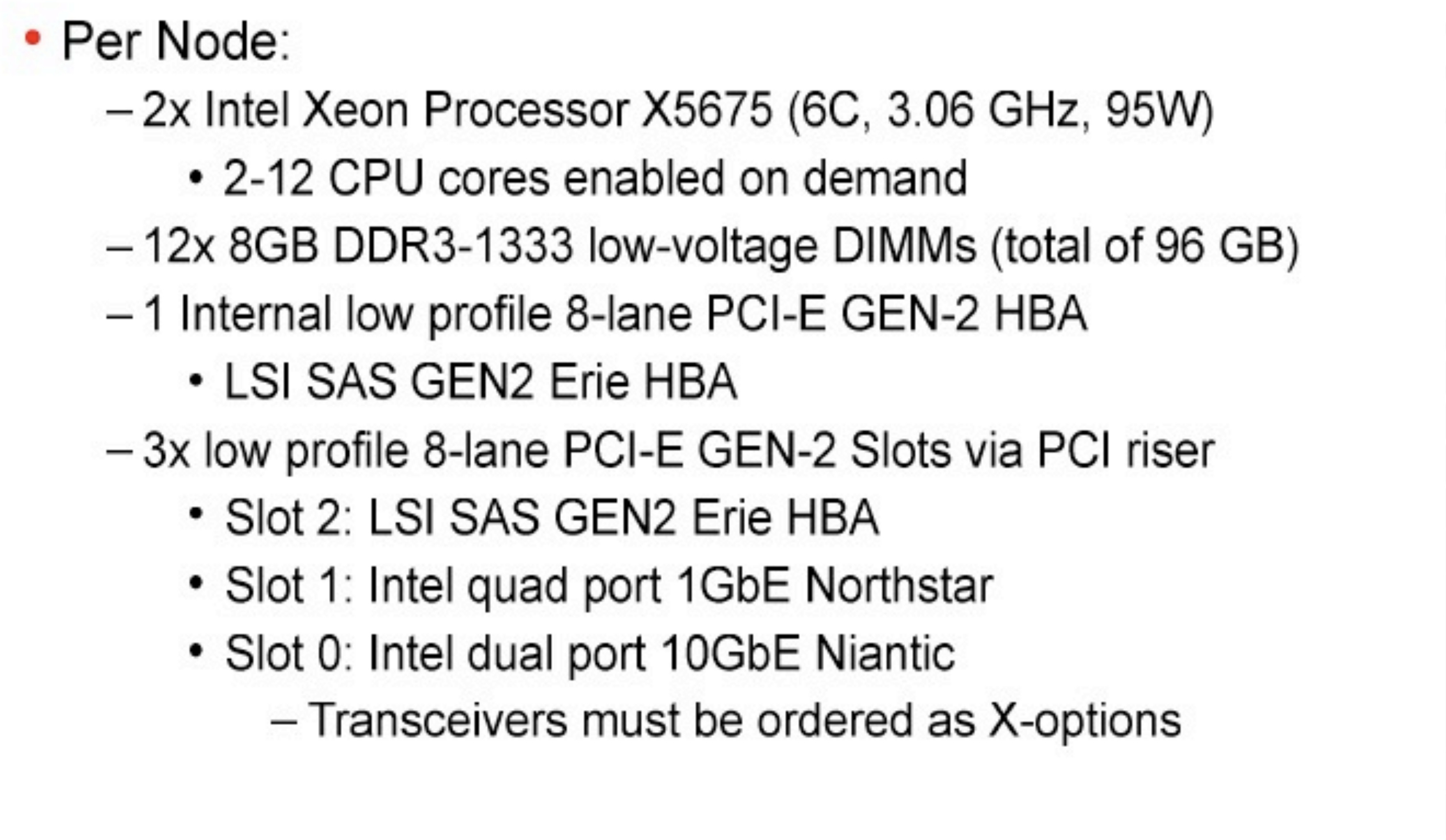
ODA Front View



ODA Rear View

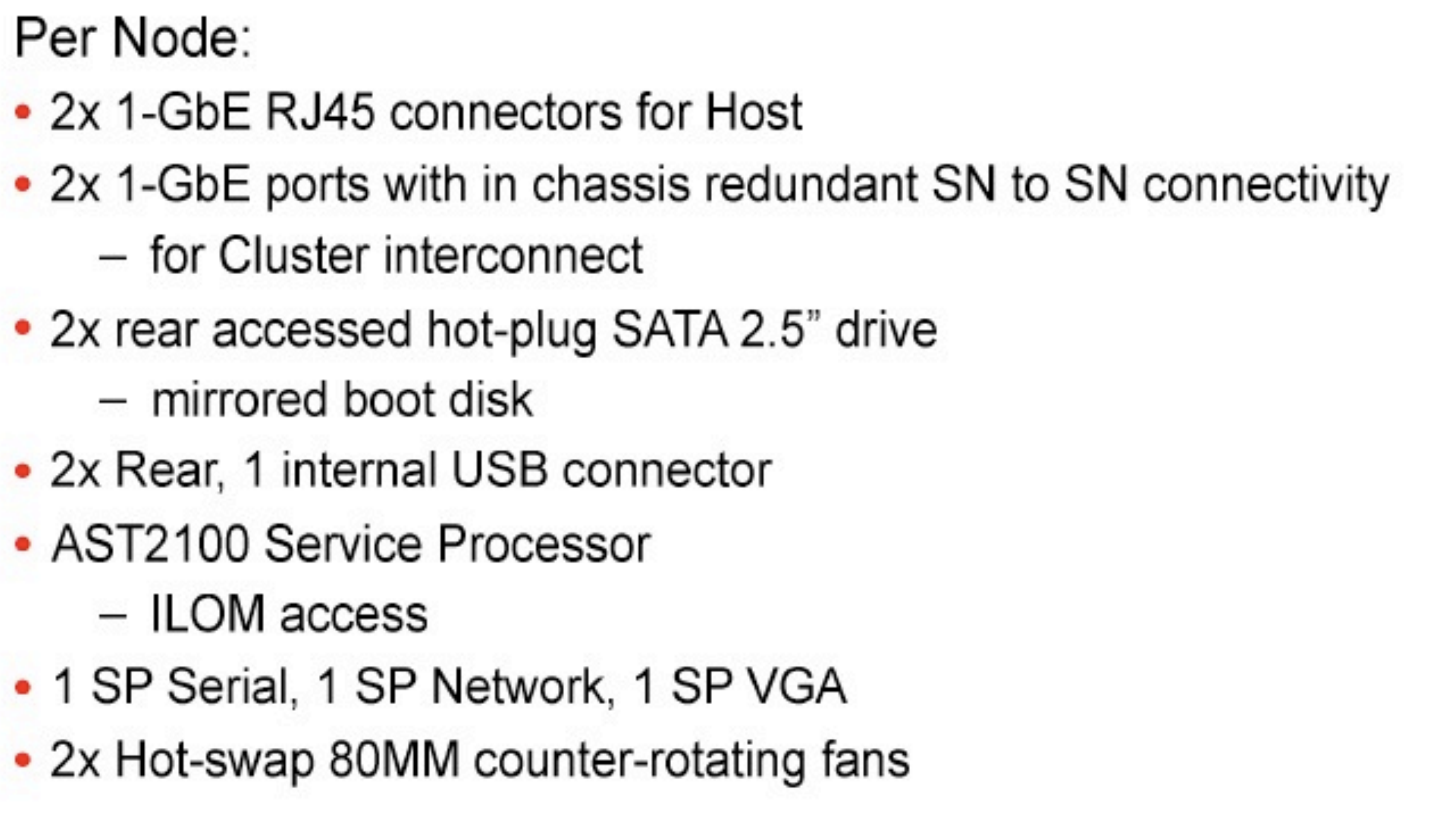


Each Server Node (SN) / System Controller (SC)



- Per Node:
  - 2x Intel Xeon Processor X5675 (6C, 3.06 GHz, 95W)
    - 2-12 CPU cores enabled on demand
  - 12x 8GB DDR3-1333 low-voltage DIMMs (total of 96 GB)
  - 1 Internal low profile 8-lane PCI-E GEN-2 HBA
    - LSI SAS GEN2 Erie HBA
  - 3x low profile 8-lane PCI-E GEN-2 Slots via PCI riser
    - Slot 2: LSI SAS GEN2 Erie HBA
    - Slot 1: Intel quad port 1GbE Northstar
    - Slot 0: Intel dual port 10GbE Niantic
      - Transceivers must be ordered as X-options

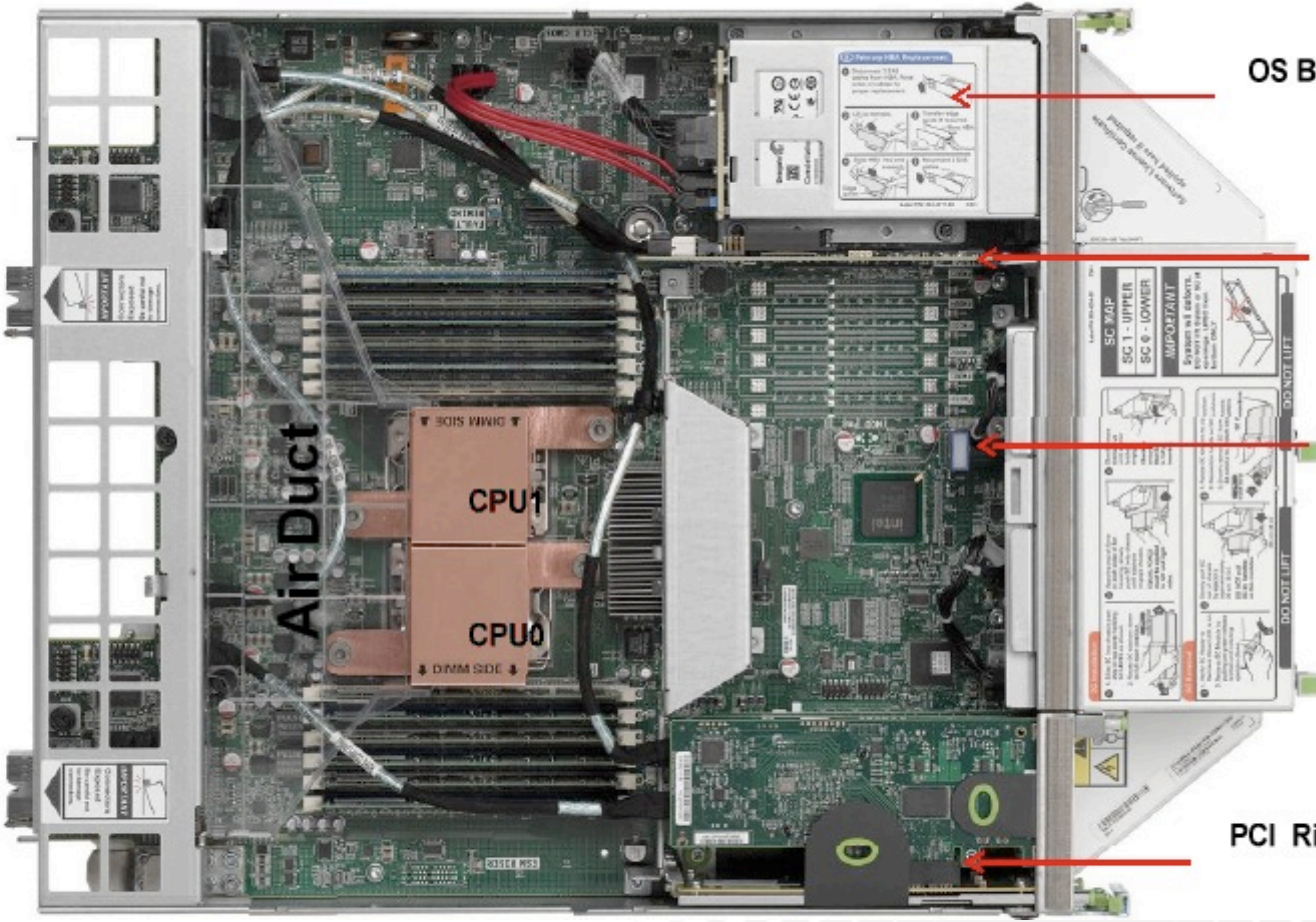
Each Server Node (SN) / System Controller (SC)



Per Node:

- 2x 1-GbE RJ45 connectors for Host
- 2x 1-GbE ports with in chassis redundant SN to SN connectivity
  - for Cluster interconnect
- 2x rear accessed hot-plug SATA 2.5” drive
  - mirrored boot disk
- 2x Rear, 1 internal USB connector
- AST2100 Service Processor
  - ILOM access
- 1 SP Serial, 1 SP Network, 1 SP VGA
- 2x Hot-swap 80MM counter-rotating fans

System Controller View



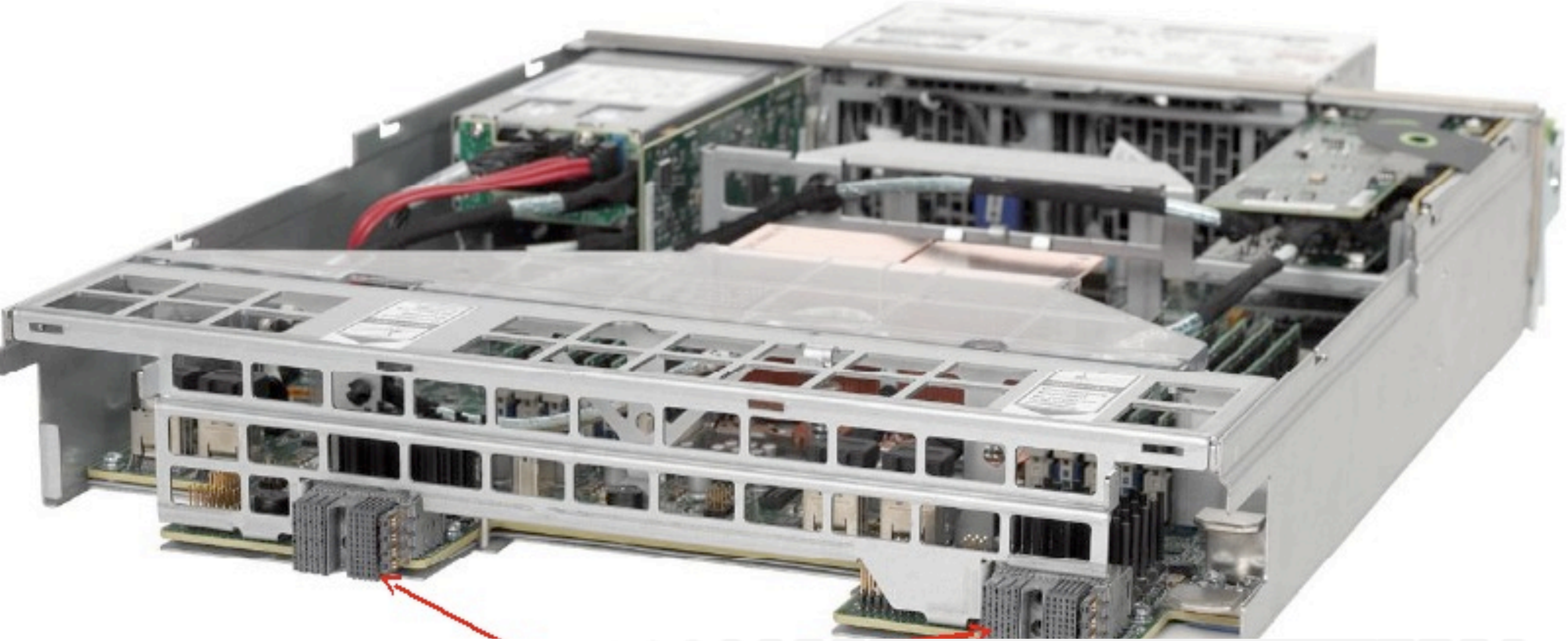
OS Boot Disk

Internal HB

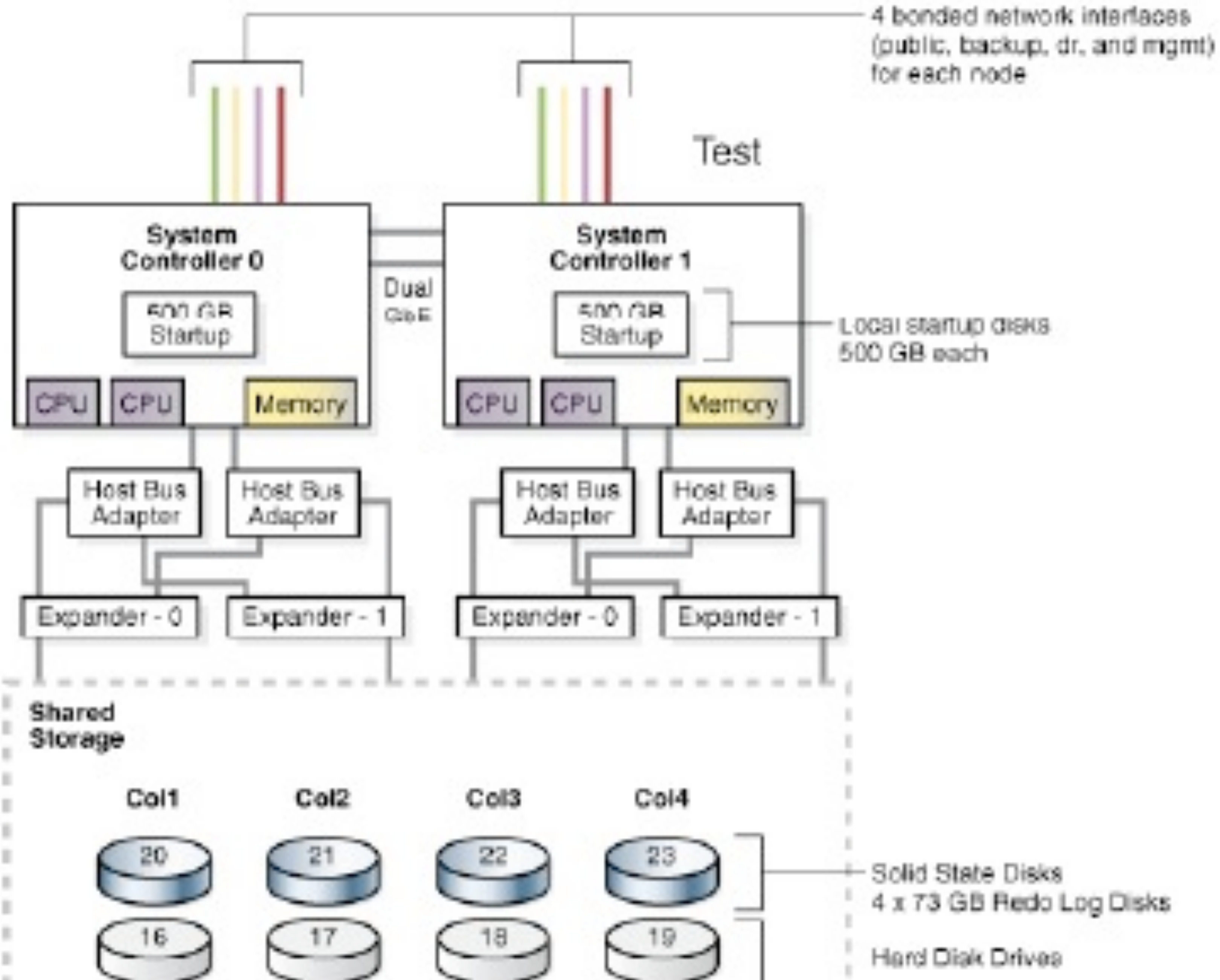
Internal US

PCI Riser

System Controller View



Oracle Database Appliance Architecture



How much?

Only $50K

Why is ODA hardware so inexpensive?

Exadata quarter rack: $330k

2 Compute Servers

3 Storage Servers

2 InfiniBand switches

Sun Rack

Admin Switch

KVM Device

Oracle Database Appliance: $50k

Single 4U appliance

Exadata / ODA Comparison

	Exadata Quarter Rack	Oracle Database Appliance
Hardware list price	$330k	$50k
Storage server software	$360k	$0
Database license list price	$846k	$47.5k - $846k
Usable Disk capacity	7TB	4TB
Hybrid Columnar Compression	Yes	No
Smart Scans	Yes	No
Expandable disk capacity	Storage expansion rack / Half rack upgrade	None *
Expandable compute capacity	Half rack upgrade	None **
Flash memory	Exadata flash cache / ASM diskgroup	REDO ***

2 cores no RAC to 24 cores RAC

* Potential option of iSCSI or NFS but non-standard - it breaks simplicity

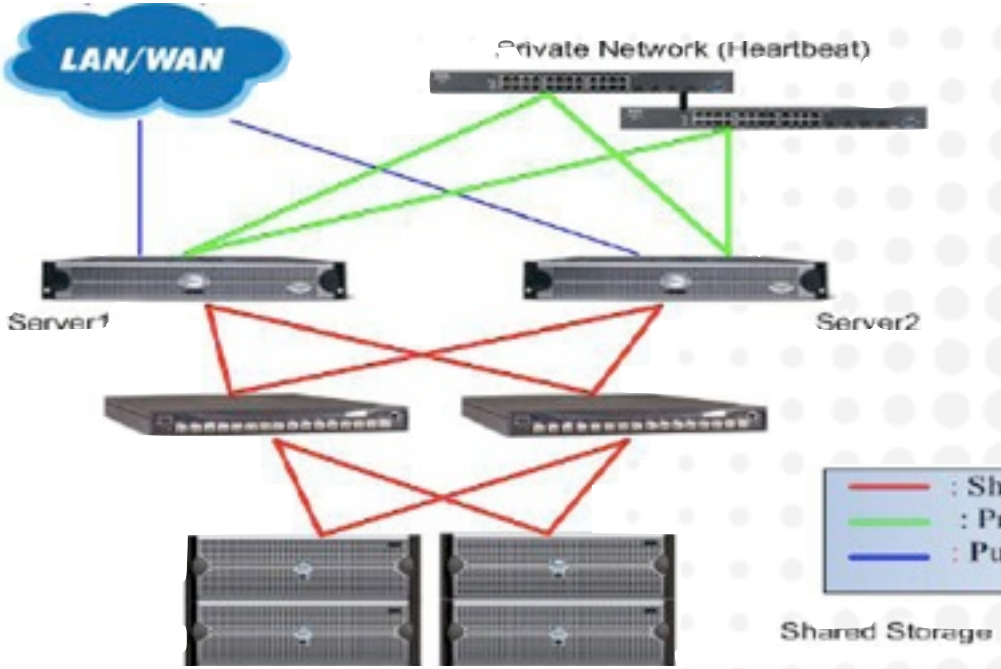
** Scales within single appliance from 2 to 24 cores

*** Technically, can host DB files & even Database Flash Cache but non-standard

Generic x86 RAC platform

vs

Oracle Database Appliance

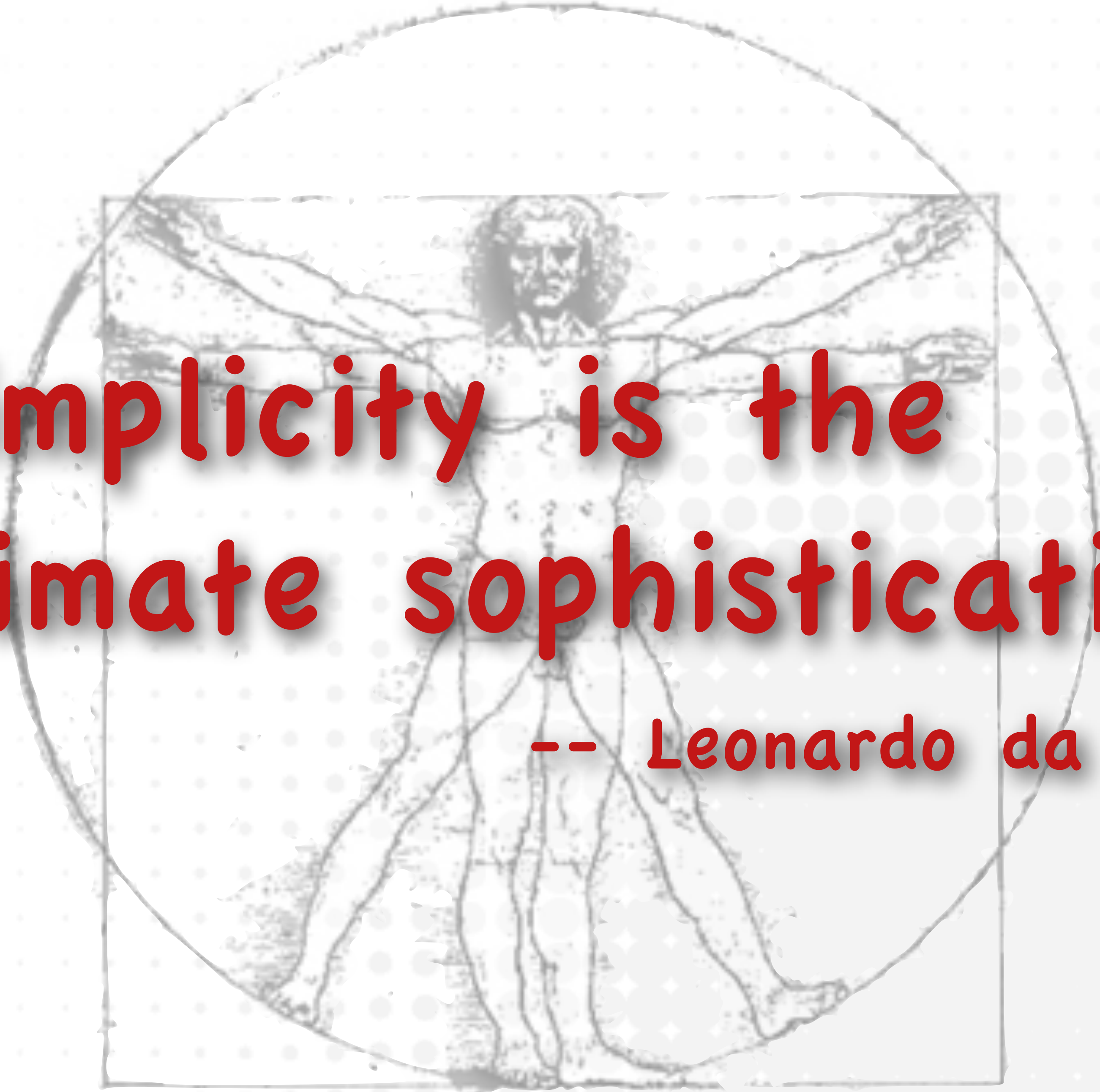


🟥 : Shared storage connection
🟩 : Private interconnect
🟦 : Public network connection

Diagram by Kay Yu

=

27

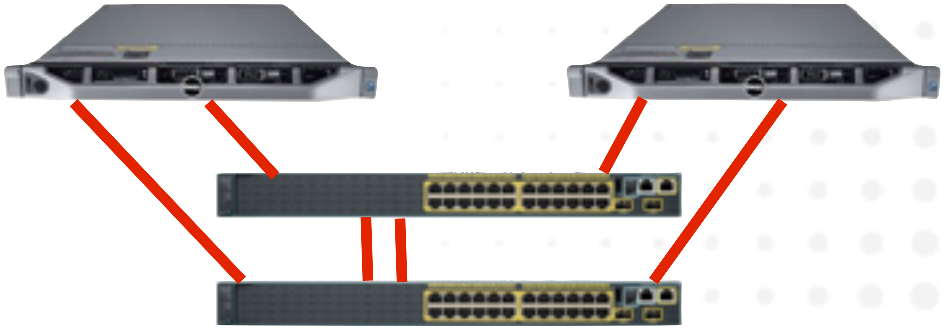


"Simplicity is the ultimate sophistication"

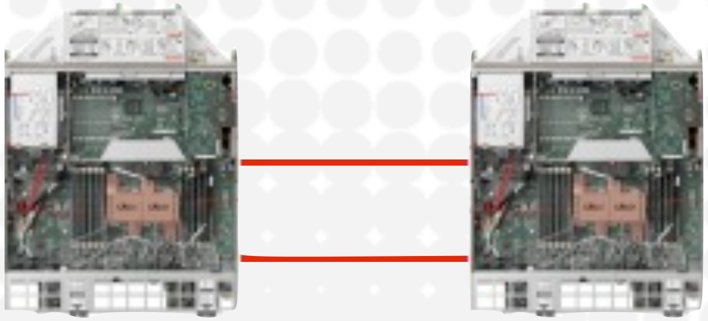
-- Leonardo da Vinci

Cluster Interconnect

Generic x86 vs ODA



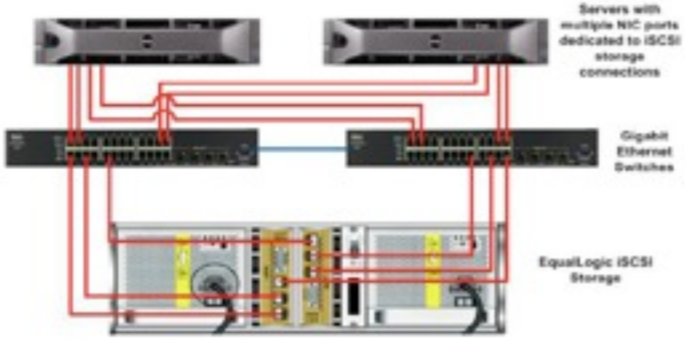
vs



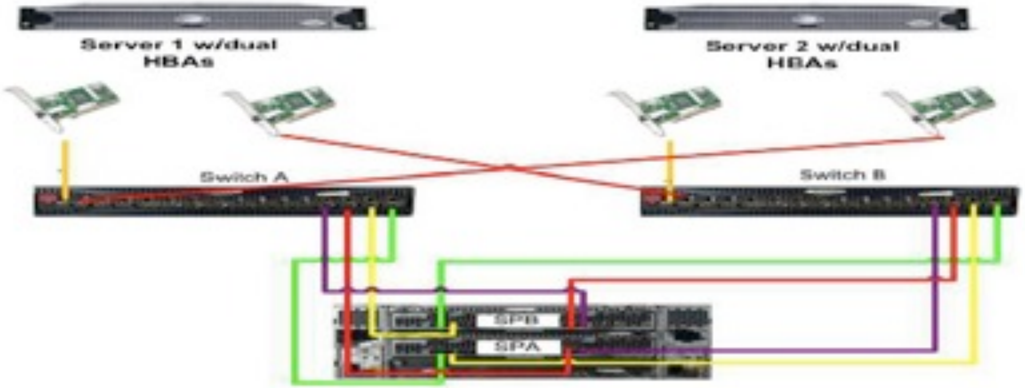
Shared Storage

Generic x86 vs ODA

iSCSI example

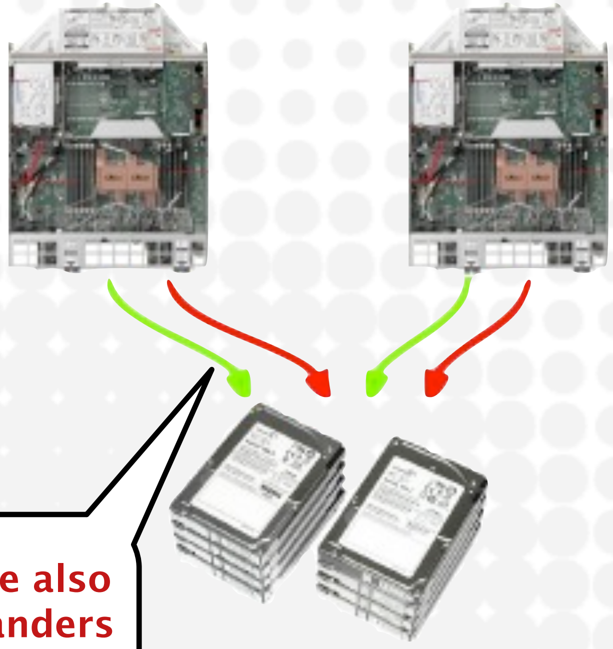

Servers with multiple NIC ports dedicated to iSCSI storage connections
Gigabit Ethernet Switches
EqualLogic iSCSI Storage

Fibre Channel example



Diagrams by Kay Yu

vs



There are also SAS expanders and HBAs

Generic x86 Platform vs Oracle Database Appliance

	Generic RAC Platform	Oracle Database Appliance	Generic non-RAC
Storage	SAN / NAS	“Local” shared disks	Local disks
Interconnect	Network switch	Direct Fiber connect	N/A
Horizontal scalability	High	Medium	None
Storage scalability	Yes	No	No
Config. flexibility	Yes	No	Yes
RAC HA	Yes	Yes	No
DR	Yes	Yes	Yes
Licensing	Node granularity	CPU Core granularity	Full node only

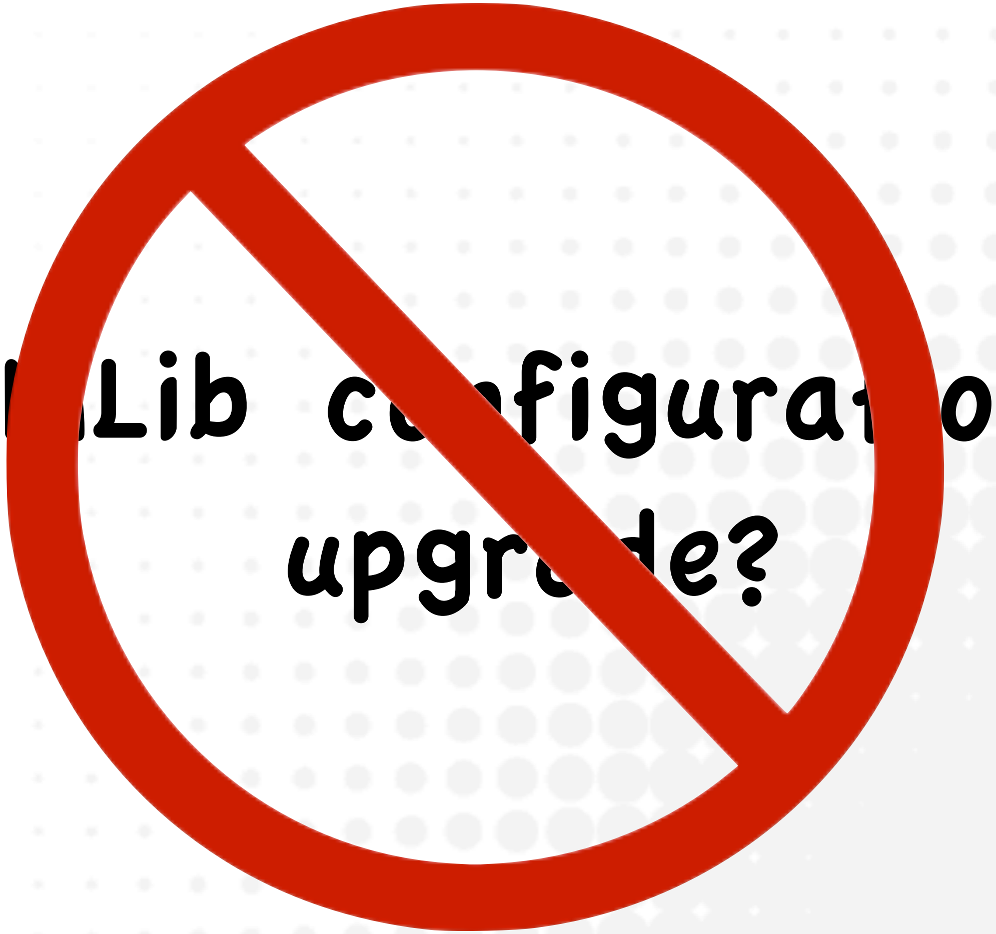
~~Shared storage setup?~~

Interconnect setup?

Multipathing Configuration?

OS pre-requisites?

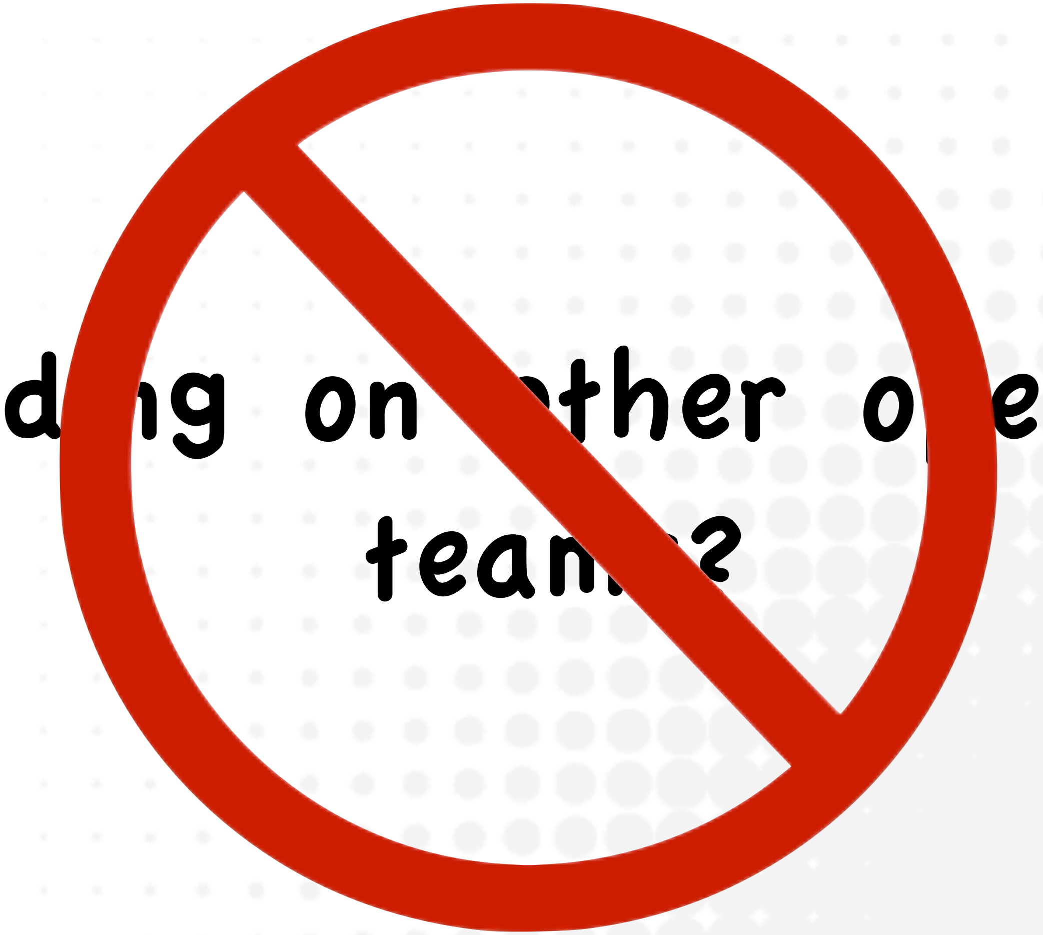
ASMLib configuration & upgrade?



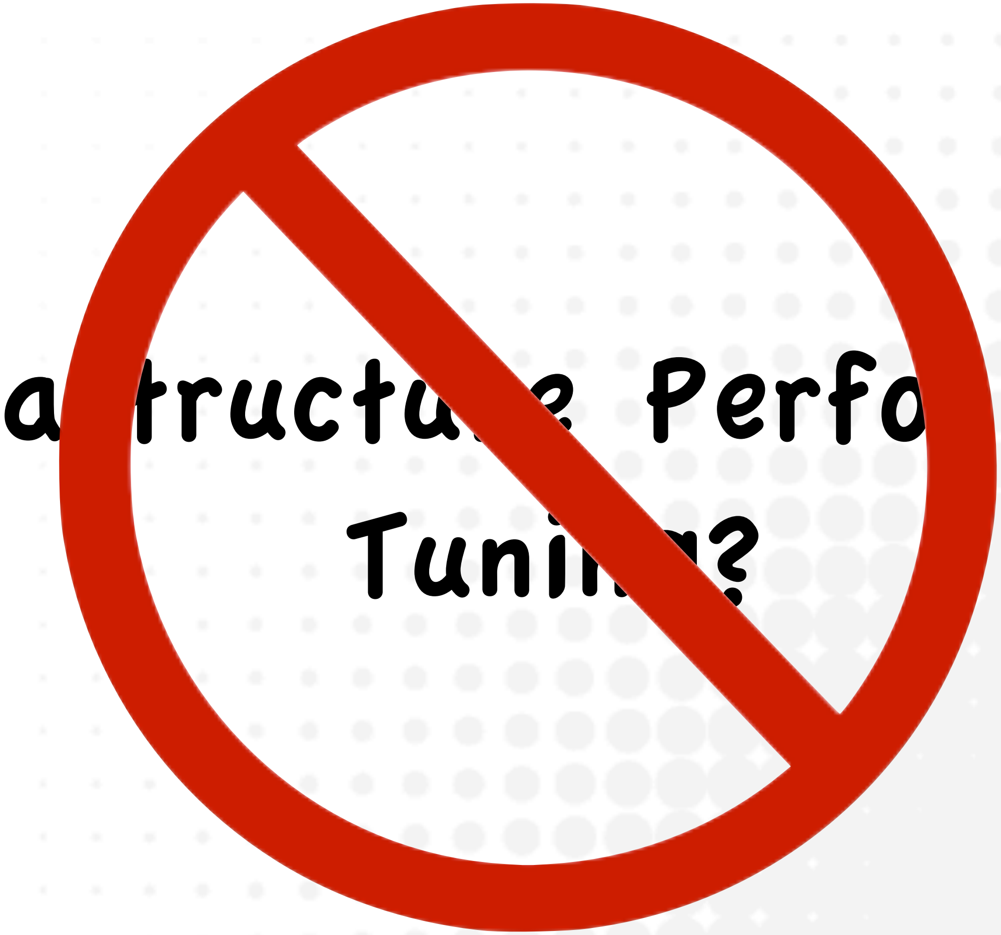
SAN Failures?

Interconnect Failures?

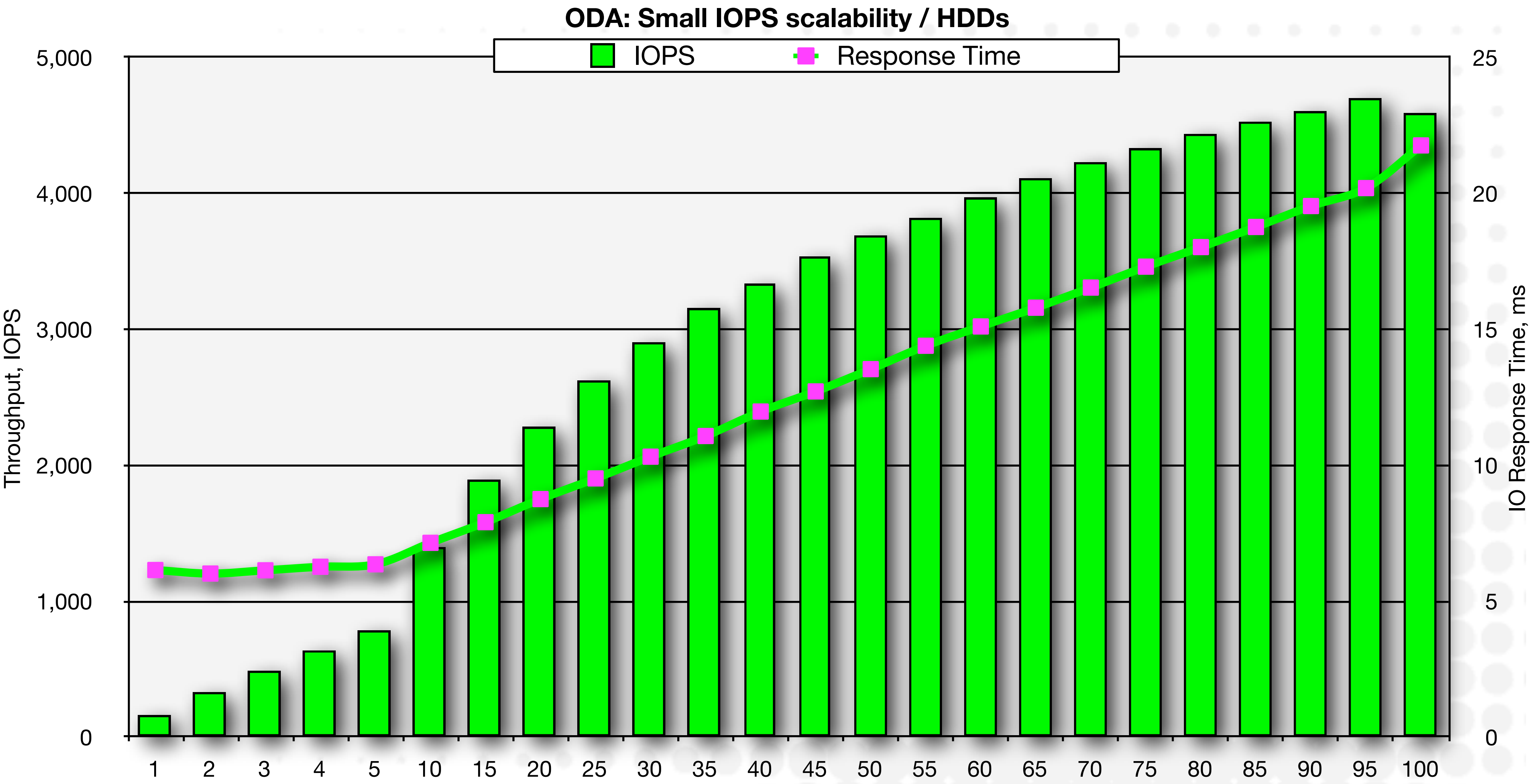
Depending on other operations teams?



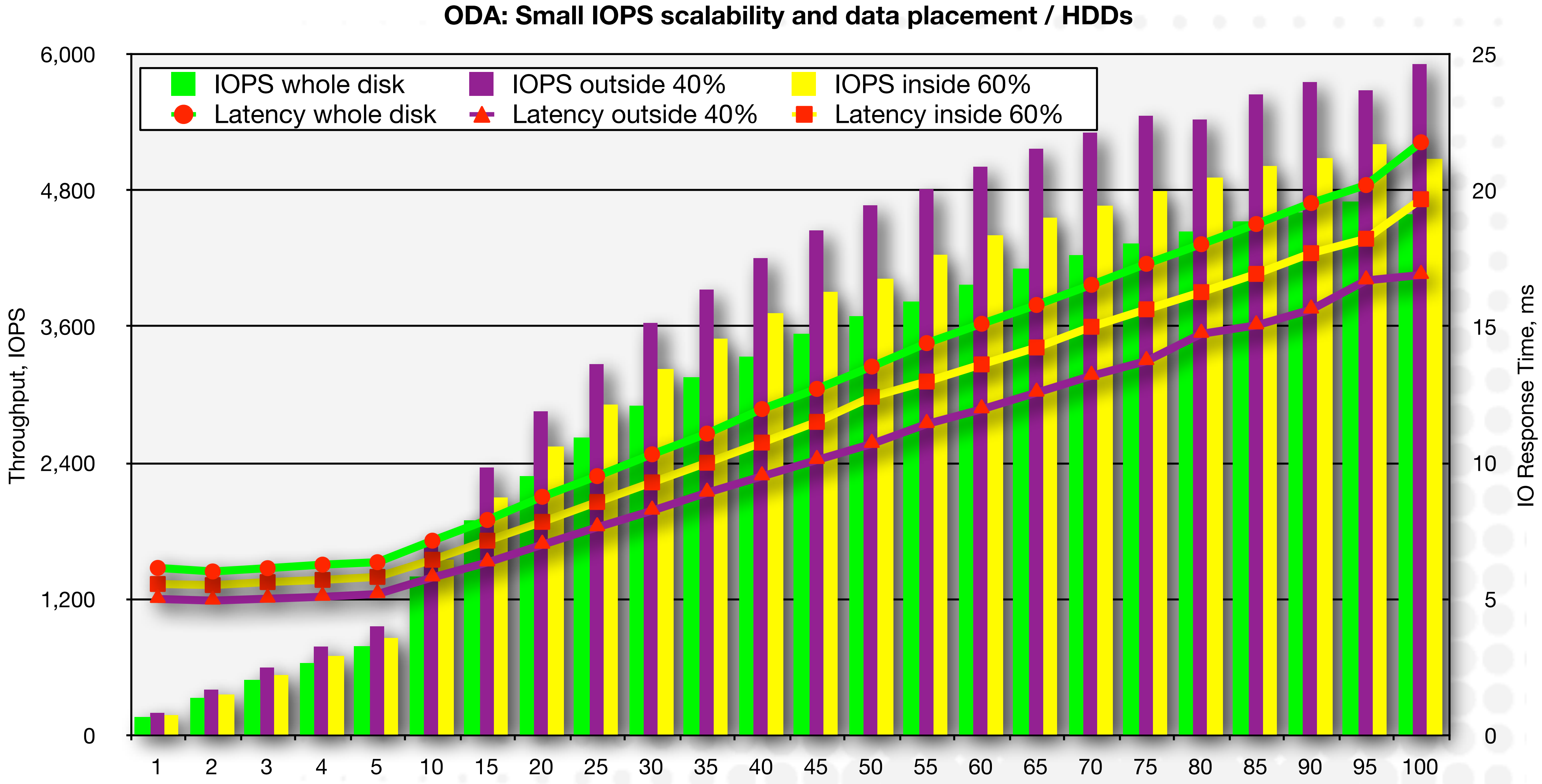
Infrastructure Performance Tuning?



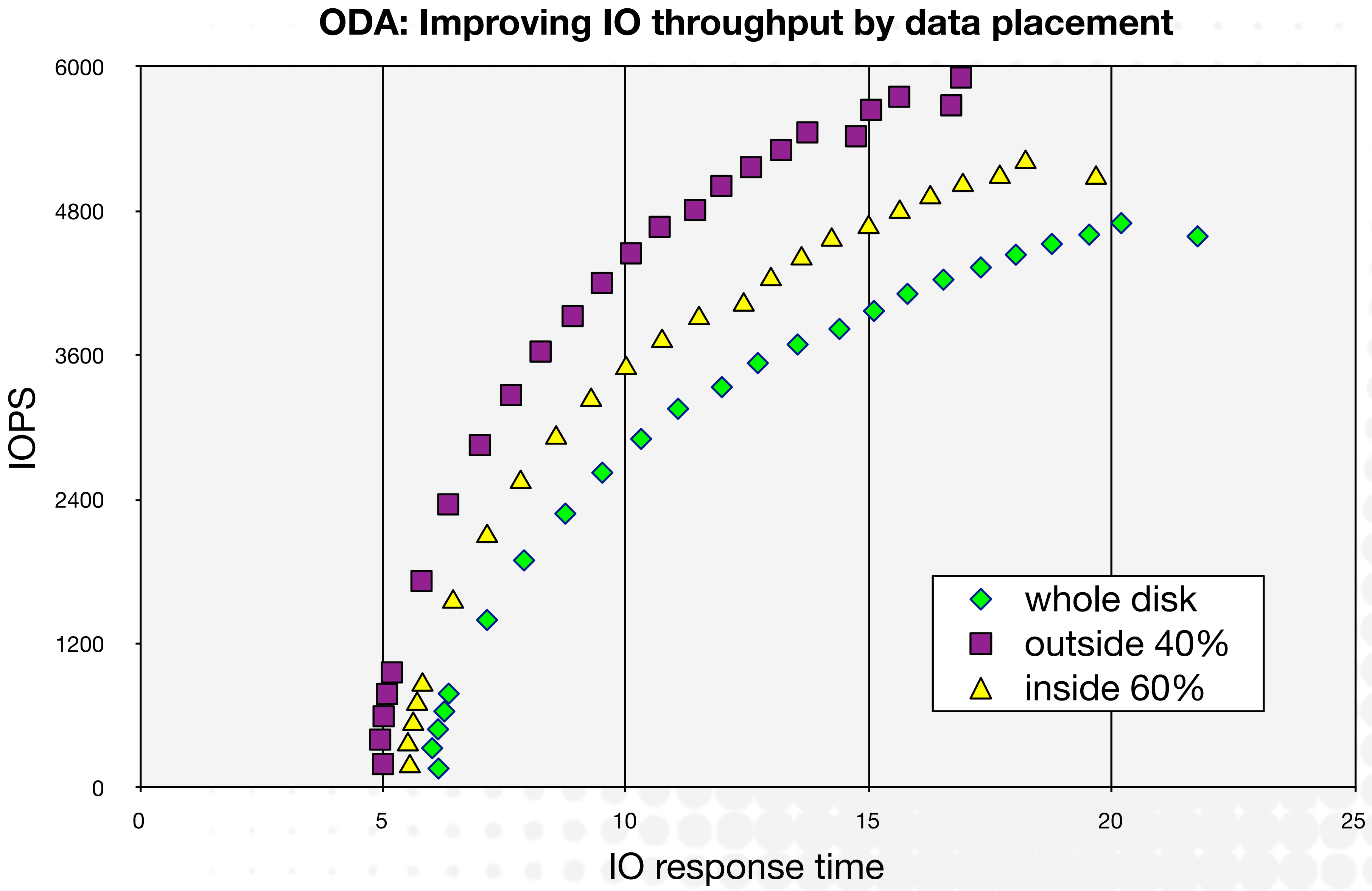
ODA Small Random Reads - HDDs Scalability



ODA Small Random Reads: Data Placement (1)



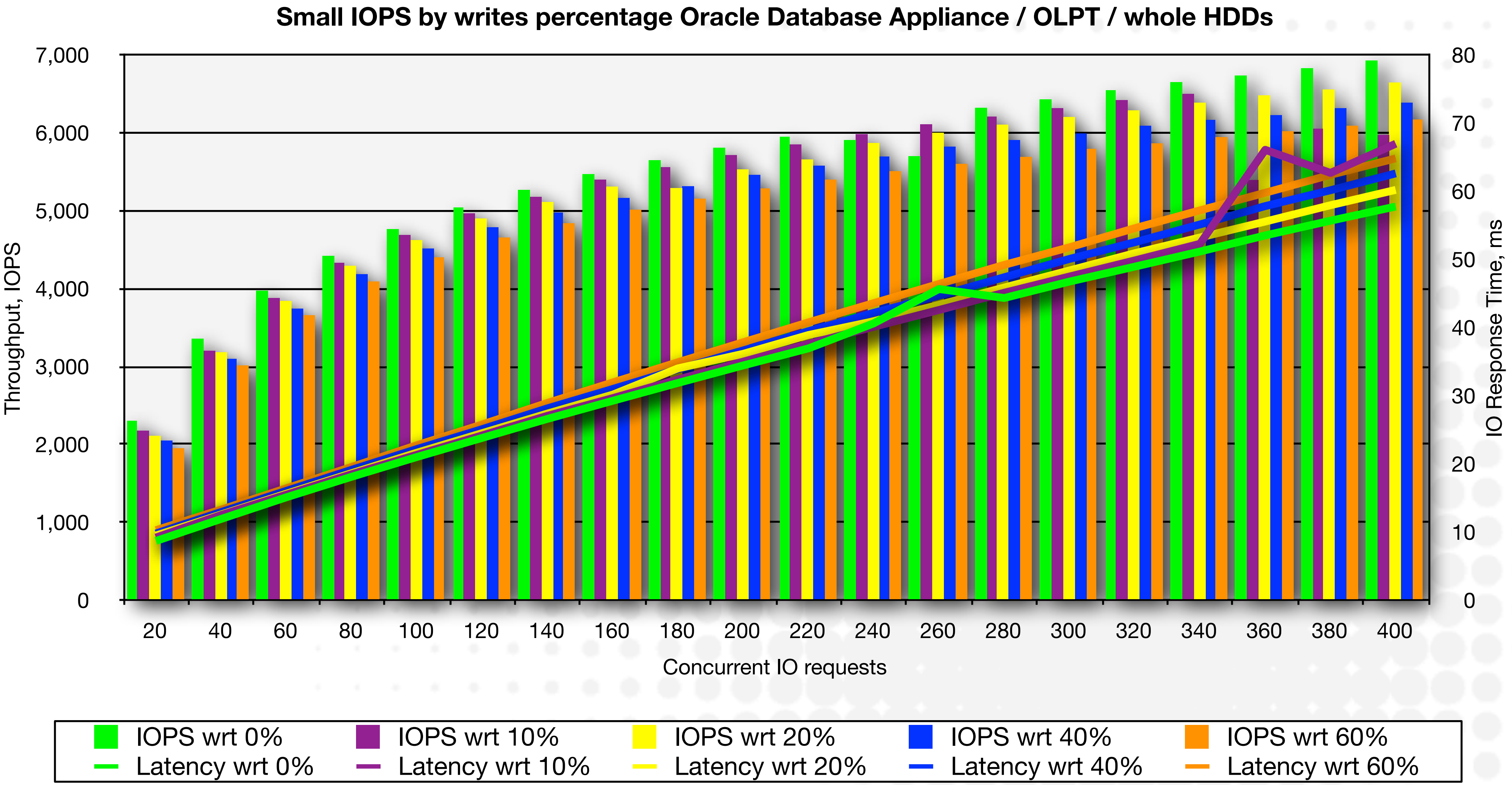
ODA Small Random Reads: Data Placement (2)



Co-locating data onto outer 40% of a disk adds 50% more IOPS

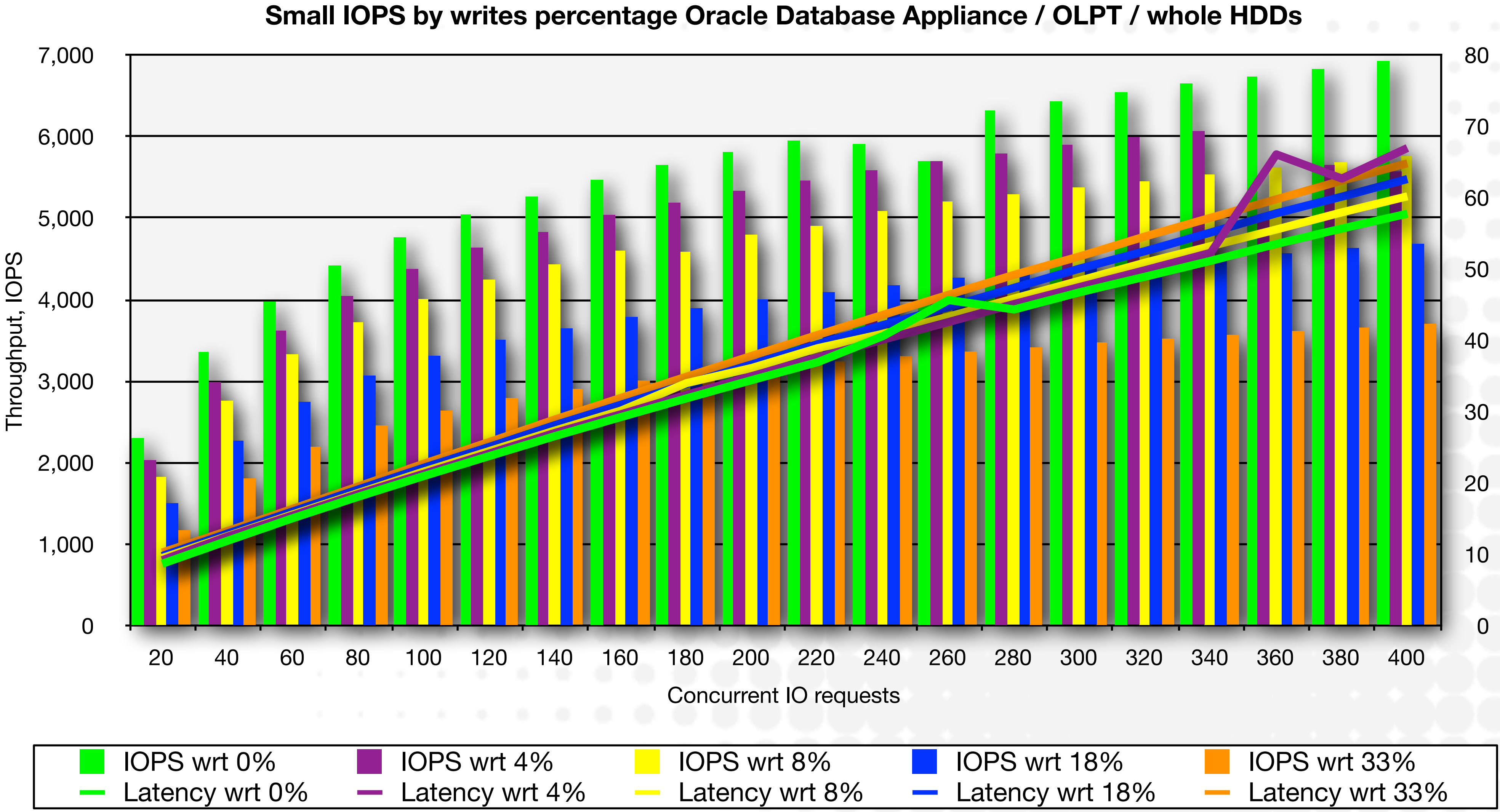
ODA Write IO impact - Minimal

not accounting triple write needs



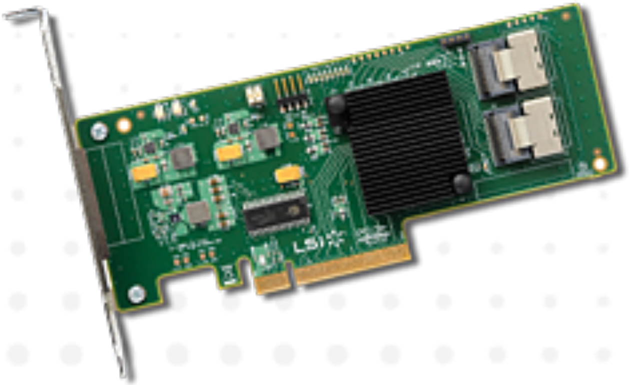
ODA Write IO impact - Minimal

accounting triple writes



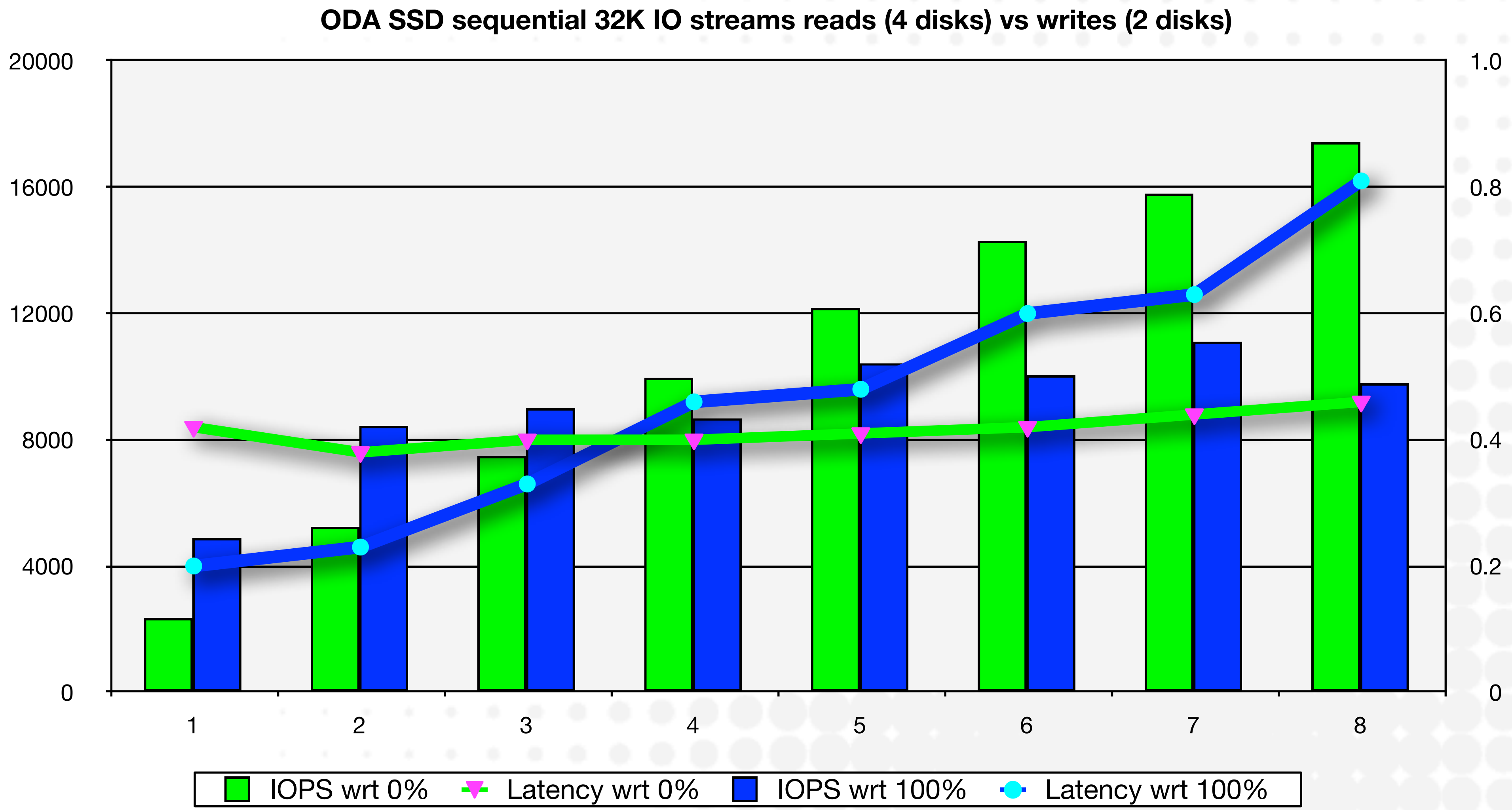
Two LSI SAS9211-8i SAS HBAs

No Cache

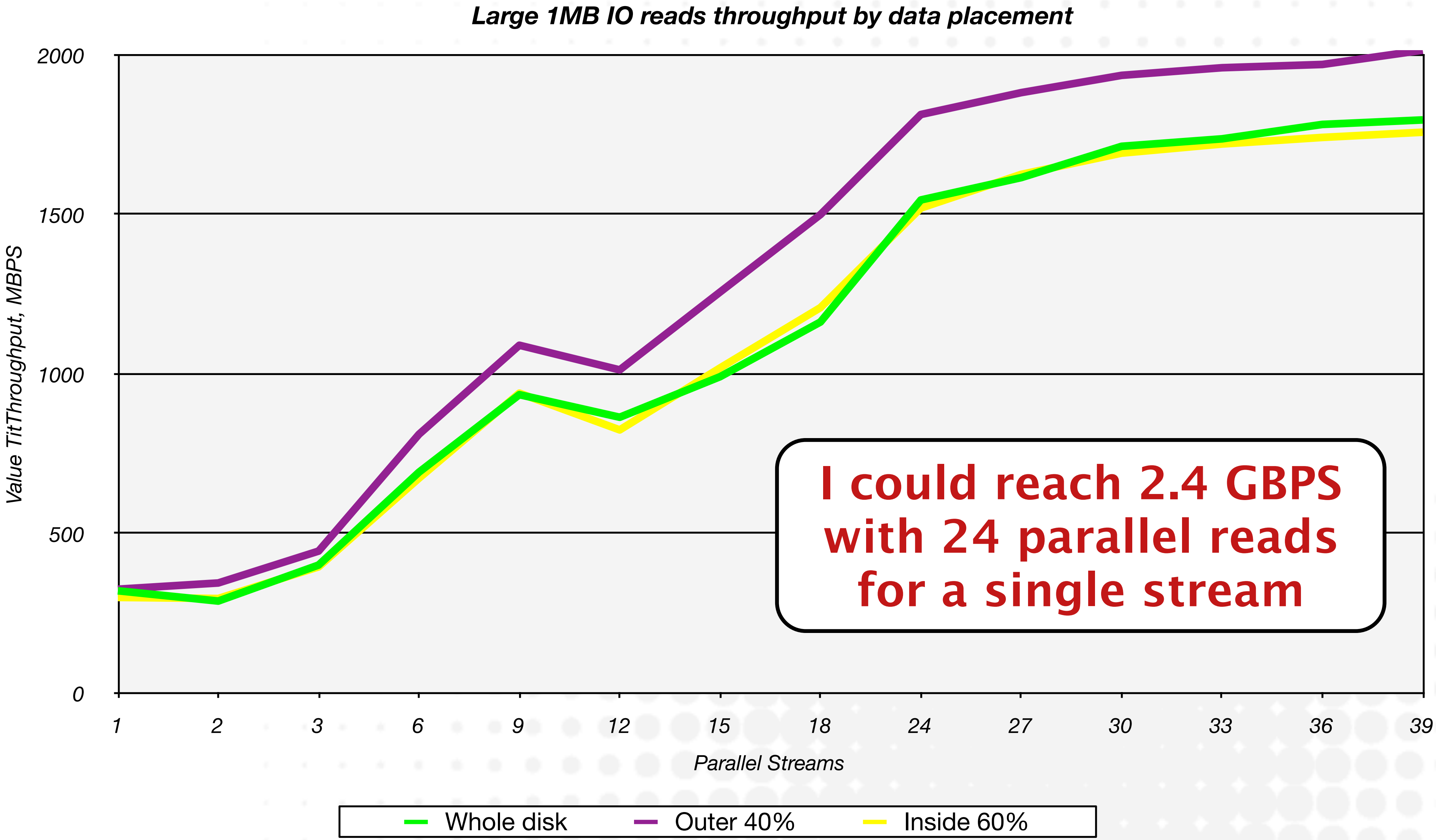


- Cannot use any cache because of shared storage
  - I.e. must go to disk every read or write because of another node
- Be careful not to saturate you IO subsystem with excessive writes
  - Tune aggressiveness of DBWR processes (MTTR target)
  - Direct path loads are OK - sequential writes are not the same
- This is why online redo logs are on SSD!
  - redo write time directly contribute to transactions response time
- 600 MBPS per lane (x8) so theoretical bandwidth 4.8 GBPS

ODA: SSD Performance for LGWR



ODA Sequential Reads Scalability (one node only)



I could reach 2.4 GBPS with 24 parallel reads for a single stream

RMAN Backup Performance

- Backup to FRA in ODA
  - Optimal number of channels - 8
  - 42 GB of data in 1 min 45 seconds (400 MBPS)
    - Should be able to achieve higher rates because RMAN spends too much time managing metadata and etc
  - 1.6 TB full backup in about 1 hour
- Backup to external location
  - BACKUP VALIDATE with 8 channels
  - 42 GB of data in 45 seconds (1 GBPS)
    - Theoretical maximum wire speed for one link 10 GbE
  - 4 TB database can be backed up in 1 hour 15 minutes

Interconnect performance?

- Cache Fusion operations - hundreds of microseconds
  - Like Exadata over Infiniband
  - Don’t need InfniBand => doesn’t need to scale beyond 2 nodes
- Dedicated 2 x 1 GbE Fibre links
  - No bonding - HAIP is used (new in 11.2)

Why High ASM Redundancy for Data on HDDs?

- Triple mirroring is not for paranoids
- Theory of disk failures is based on assumptions that failures happen according to Poisson process
  - Exponentially distributed / non-correlated
- Disk failures in real life are often correlated

Using Device Diversity to Protect Data against Batch-Correlated Disk Failures

Jehan-François Pâris[*]
Department of Computer Science
University of Houston
Houston, TX 77204-3010
+1 713-743-3341
paris@cs.uh.edu

Darrell D. E. Long[*]
Department of Computer Science
University of California
Santa Cruz, CA 95064
+1 831-459-2616
darrell@cs.ucsc.edu

Consider a group of *n* disks all coming from the same production batch. We will consider **two distinct failure processes**:

1. Each disk will be subject to **independent failures** that will be exponentially distributed with rate λ; these independent failures are the ones that are normally considered in reliability studies.
2. The whole batch will be subject to the **unpredictable manifestation of a common defect**. This event will be exponentially distributed with rate $\lambda' \ll \lambda$. It will not result in the immediate failure of any disk but will accelerate disk failures and make them happen at a rate $\lambda'' \gg \lambda$.

After a Failure Caused by a Global Defect

$$P_{surv} = exp(-n\lambda''T_R)$$

λ'' - accelerated rate of failure

λ'' is one failure per week:

$$P_{surv} = 78.813\%$$

λ'' is one failure per month:

$$P_{surv} = 94.596\%$$

(normal redundancy)

After a Failure Caused by a Global Defect

$$P_{surv} = (1+n\lambda''T_R)exp(-n\lambda''T_R)$$

λ'' - accelerated rate of failure

n - 5 hours

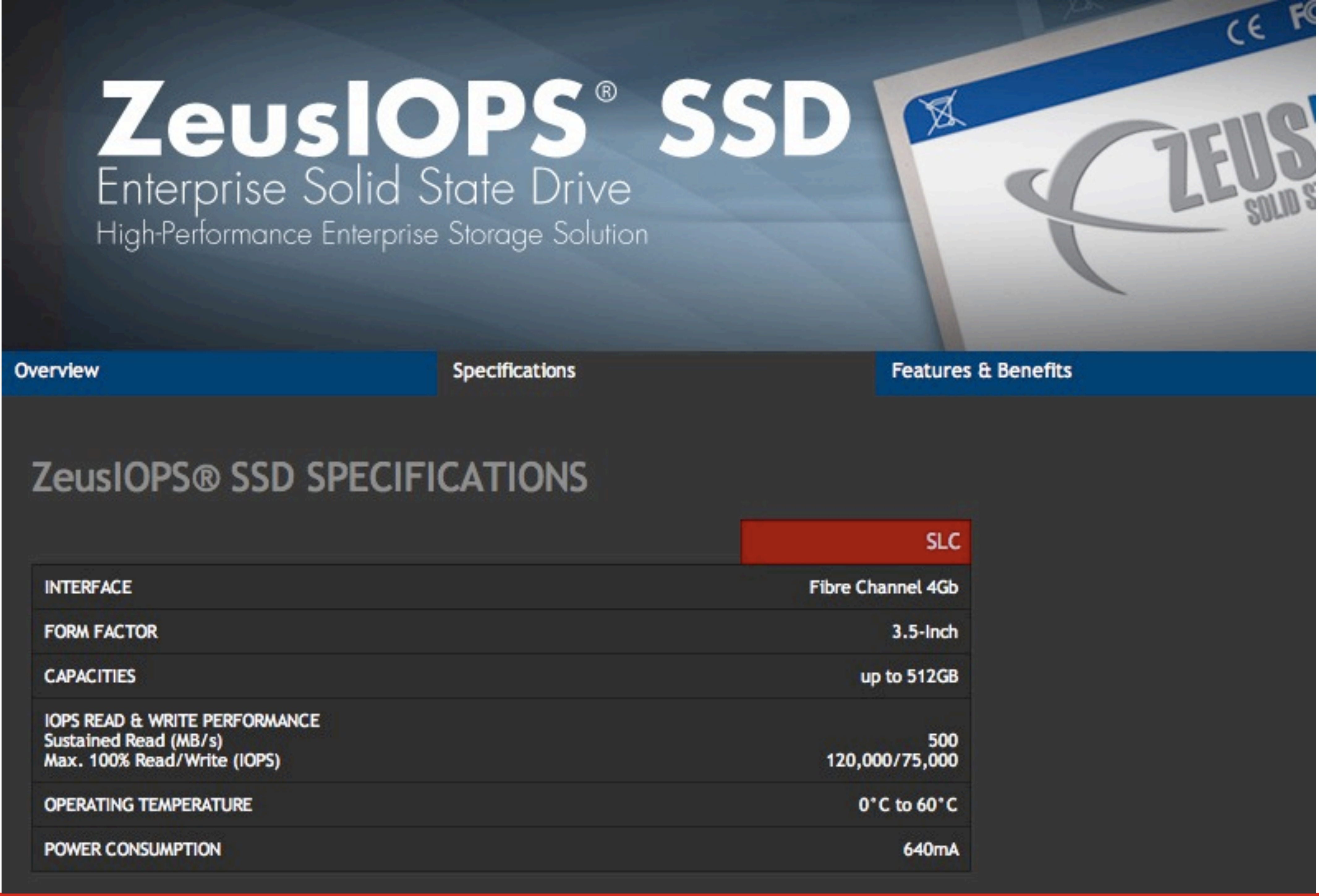
λ'' is one failure per <u>week</u>:

$$P_{surv} = 97.58\%$$

λ'' is one failure per <u>month</u>:

$$P_{surv} = 99.85\%$$

(high redundancy)



ZeusIOPS® SSD

Enterprise Solid State Drive

High-Performance Enterprise Storage Solution

ZeusIOPS® SSD SPECIFICATIONS

	SLC
INTERFACE	Fibre Channel 4Gb
FORM FACTOR	3.5-Inch
CAPACITIES	up to 512GB
IOPS READ & WRITE PERFORMANCE Sustained Read (MB/s) Max. 100% Read/Write (IOPS)	 500 120,000/75,000
OPERATING TEMPERATURE	0°C to 60°C
POWER CONSUMPTION	640mA

Why Normal ASM Redundancy for Redo on SSDs?

- SSD fail less frequently - no moving parts
- Fewer partner disks (n in the formula in previous slides)
- Rebalancing is MUCH faster after a disk failure
  - Window of vulnerability can be much lower

Configuration Worth to Note

- OEL 5.5 without OEK
- Interconnect HAIP (no bonding)
- db_block_checking and db_block_checksum is FULL
- _ENABLE_NUMA_SUPPORT=FALSE
- ACFS is configured (CLoudFS)
- HIGH redundancy ASM for data
- ASMLib is not used

Things Potentially Missing

- FRA is sized 2 GB regardless of database size
- Backups are not configured by default
- Huge pages not used (AMM is in use)
- OS oracle/grid/root environment variables are not set
- BIGFILE tablespaces are not used
- Only two online redo groups per thread
- swapiness cranked up to 100%
- parallel_servers_target=128 (too much?)

Oracle Database Appliance Requires 11.2.0.2

We will upgrade and migrate your DB to ODA **for free**

Pythian

love your data

Upgrade & migrate to Oracle Database Appliance for FREE

Learn More

Q & A

Email me - gorbachev@pythian.com

Read my blog - http://www.pythian.com

Follow me on Twitter - @AlexGorbachev

Join Pythian fan club on Facebook & LinkedIn