A DBA’s Crash Course on Flash-Based Architectures

Roye Avidor
Technical Marketing Engineer, HGST

Agenda

- About HGST
- Our “Street Cred” SSDs and Software
- Technical Details
  - Flash vs. SSD—Why DBAs should care about the difference
  - How Flash changes storage architecture designs
  - How current storage architecture designs compare
  - Two rather special Flash-based offerings from HGST
- Business Benefits
- Q&A

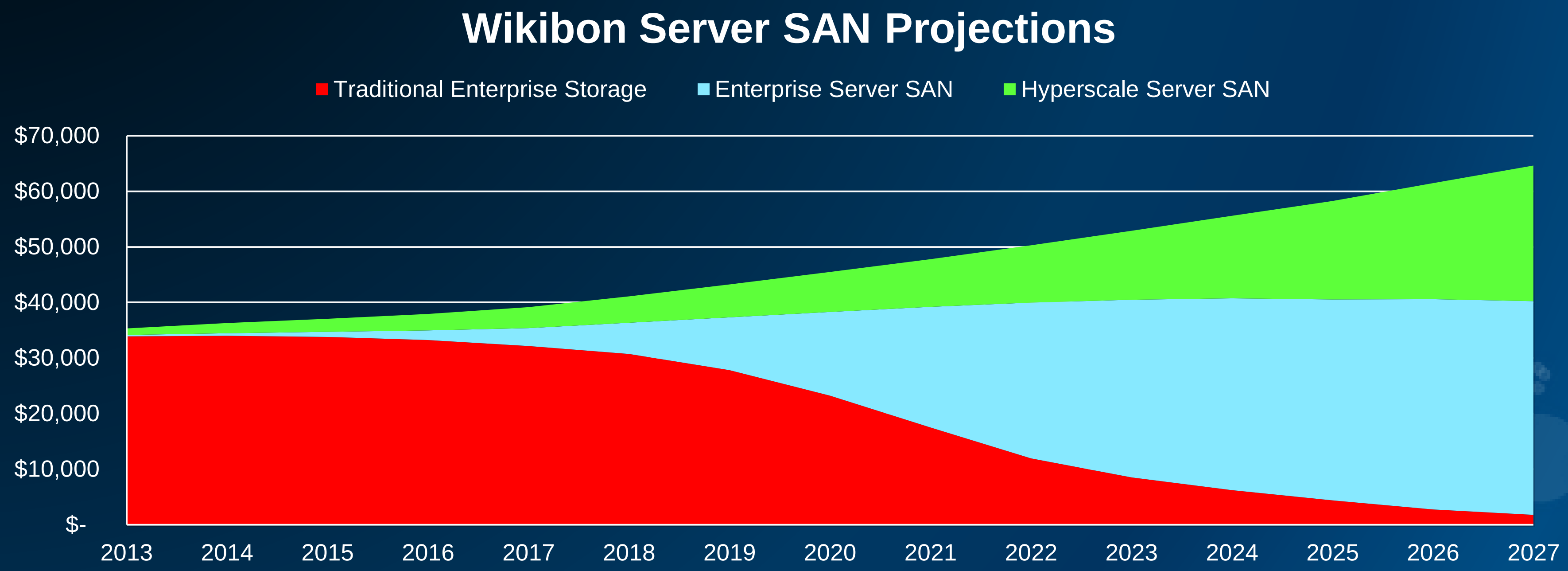
Company Profile

- **Founded in 2003 through the combination of the hard drive businesses of IBM, the inventor of the hard drive, and Hitachi, Ltd. ("Hitachi")**
- **Acquired by Western Digital in 2012**
- Headquartered in San Jose, California
- Approximately 41,000 employees worldwide
- More than 4,700 active worldwide patents (YE2013)

Mission: HGST is optimizing storage efficiency and reliability for today's data-centric economy, delivering technology innovations and enabling new ways to capture and utilize data, and reduce total cost of management.

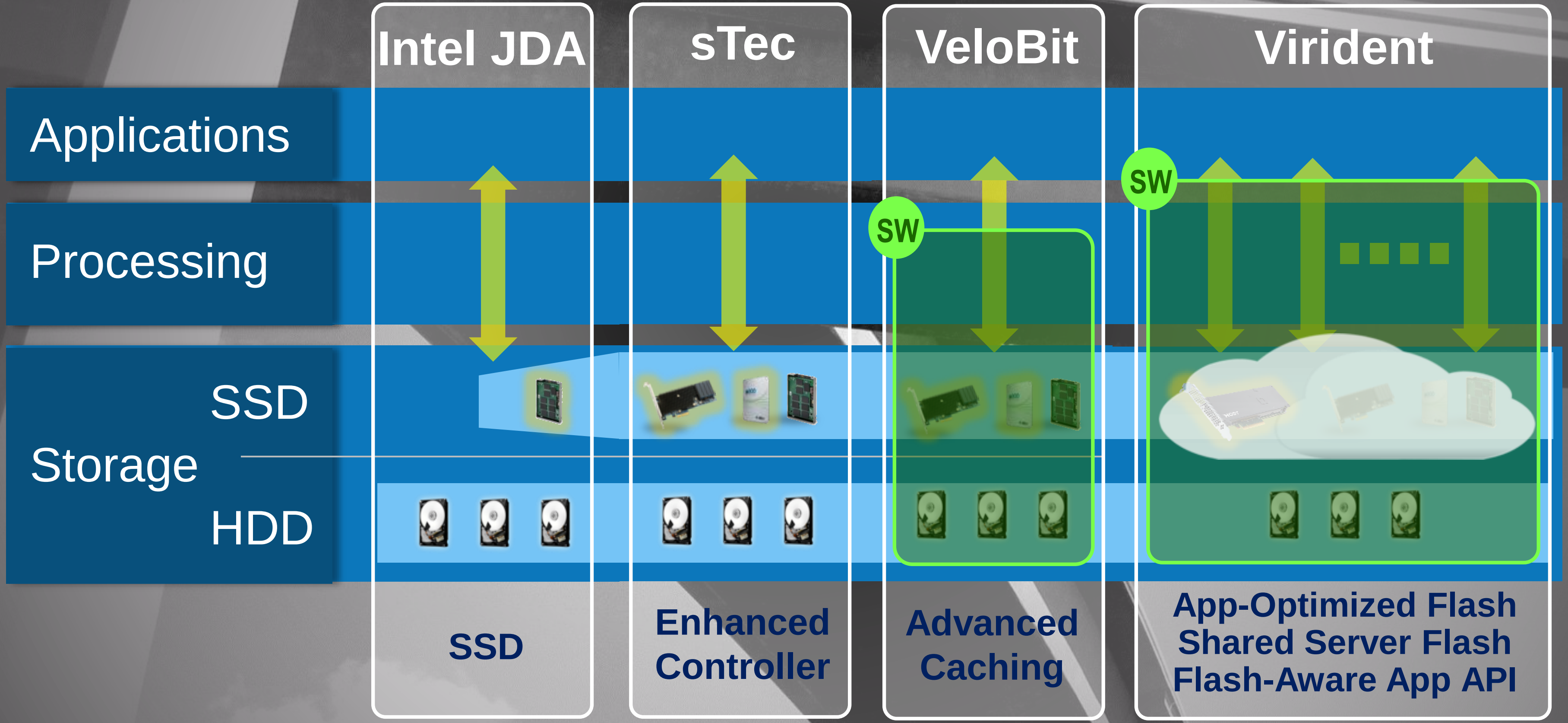
Wikibon Server SAN Projection

Traditional Enterprise Storage, Hyperscale Server SAN & Enterprise Server SAN Revenue Projections 2013-2027



HGST | Long Live Data™
a Western Digital company

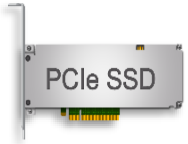
$1B Invested into On-ramping "Hot" Storage



Intel JDA	sTec	VeloBit	Virident
SSD	Enhanced Controller	Advanced Caching	App-Optimized Flash Shared Server Flash Flash-Aware App API

HGST Hardware Technologies

PCIe SSDs



- Big Data Analytics
- Virtualization, VDI, VSAN
- HF Trading
- Databases / OLTP
- Scale-out DBs

SAS SSDs

- Big Data Analytics
- VSAN
- HF Trading
- Databases / OLTP
- Indexing

Performance

- Databases / OLTP
- Content Serving
- Business Intelligence
- Cloud Gaming
- HPC

Capacity

- Cloud Storage
- Virtualized Servers
- VSAN
- Cloud Computing
- Storage Arrays
- Social Networks
- Long-Tail Content
- Big Data Storage

Capacity Scale

- Long-Tail Content
- Video on Demand
- Cloud Storage
- Big Data Storage
- Replicas
- Mail Servers
- Surveillance

Cold Storage

- Archives
- Regulatory & Compliance
- Surveillance
- Medical Records

Hot → Warm → Cold

HGST | Long Live Data™

HGST Software Solutions

Device Manager

Discover

Monitor

Manage

Report

For Standalone Windows, Linux, & Solaris

SSD

PCIe SSD

Profiler

Capacity Planning

Optimize Flash Usage

Caching Analysis

Any Application

For Standalone Windows & Linux

SSD

PCIe SSD

ServerCache

Application Acceleration

Read Caching & Writeback Caching

For Standalone Windows & Linux

SSD

PCIe SSD

HA

Synchronous Replication

Failover

Low Latency

InfiniBand

Linux

PCIe SSD

Share

Shared Flash

Low Latency

High Performance

Linux

Oracle® RAC

PCIe SSD

ClusterCache

Clustered Server Caching

Endurance

Ultimate Performance

Linux

Oracle® RAC

PCIe SSD

Space

Server SAN Volume Manager

Add Spaces
Replicate
Share
Manage
Linux

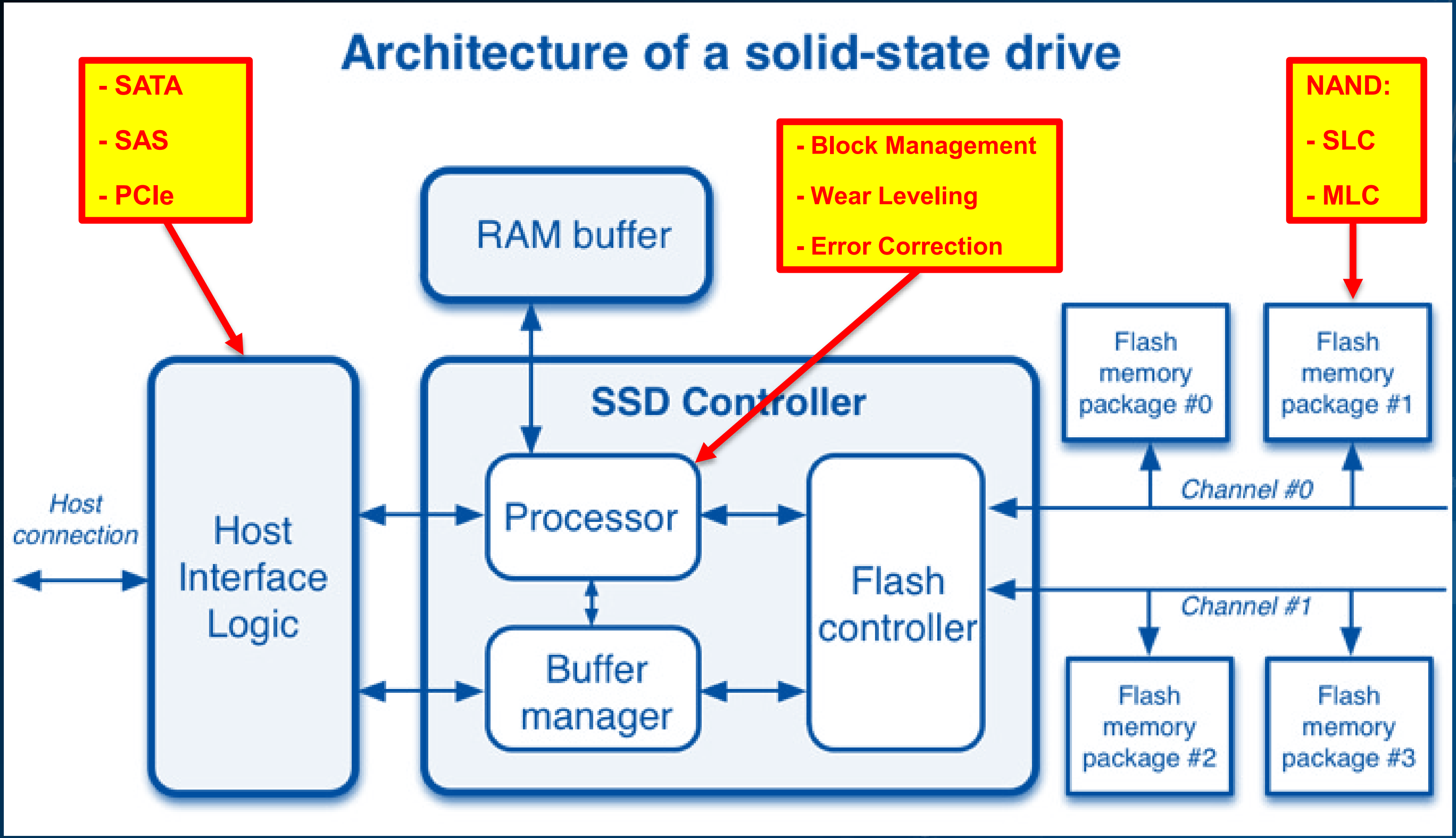
PCIe SSD

Enterprise IT Solutions

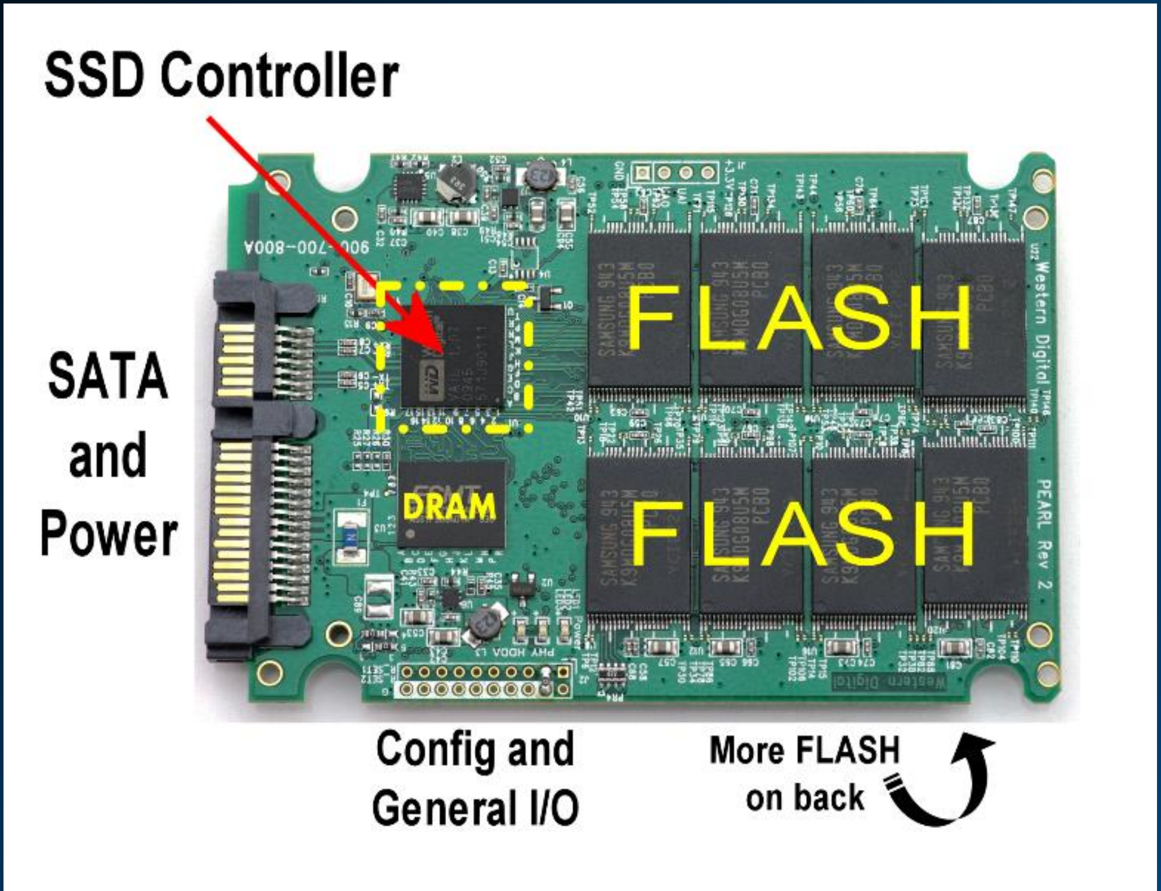
“3x server consolidation on MySQL”	**“5x IOPS improvement on Oracle® RAC”**	**“10X Latency Reduction for Exchange”**	**“46X faster report generation on MS SQL”**	**“7X Increase in VDI Instances”**

Microsoft® Exchange
SharePoint
Microsoft® SQL Server®
ORACLE®
hadoop
MySQL®
vmware®
Windows Server® Hyper-V™
DataCore™ SOFTWARE
GRIDSTORE
Fluidata
established brands
VAIL
HOSTED. ENHANCED. TELEPHONY.
中国平安 PING AN
in
IMPERIAL PFS
YAHOO! JAPAN
KAYAK®

What’s Inside of SSD/Flash



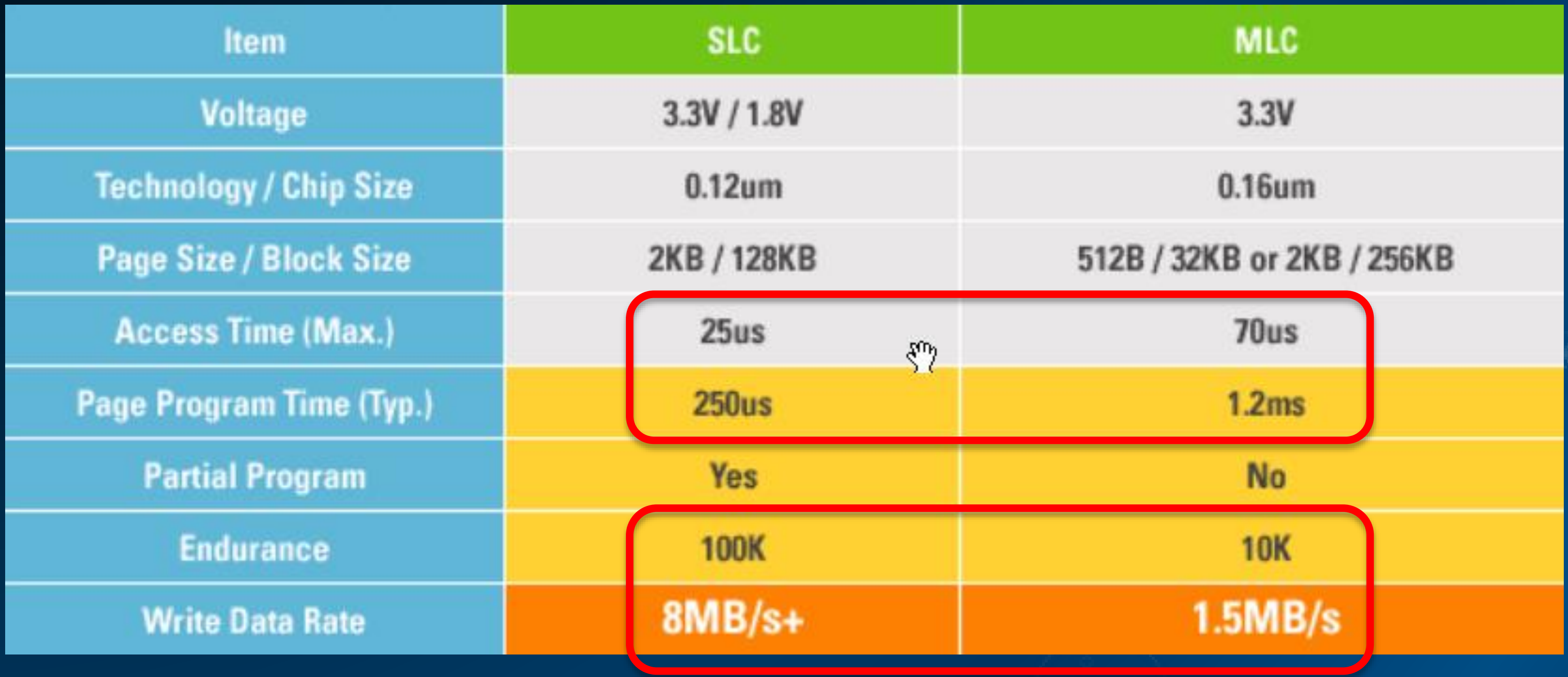
Flash Example



NAND: SLC vs. MLC

Single-level cell (SLC) SSD drives are faster and more reliable.

Multi-level cell (MLC) SSD drives are slower, cheaper, but less reliable.



Item	SLC	MLC
Voltage	3.3V / 1.8V	3.3V
Technology / Chip Size	0.12um	0.16um
Page Size / Block Size	2KB / 128KB	512B / 32KB or 2KB / 256KB
Access Time (Max.)	25us	70us
Page Program Time (Typ.)	250us	1.2ms
Partial Program	Yes	No
Endurance	100K	10K
Write Data Rate	8MB/s+	1.5MB/s

Common Technology

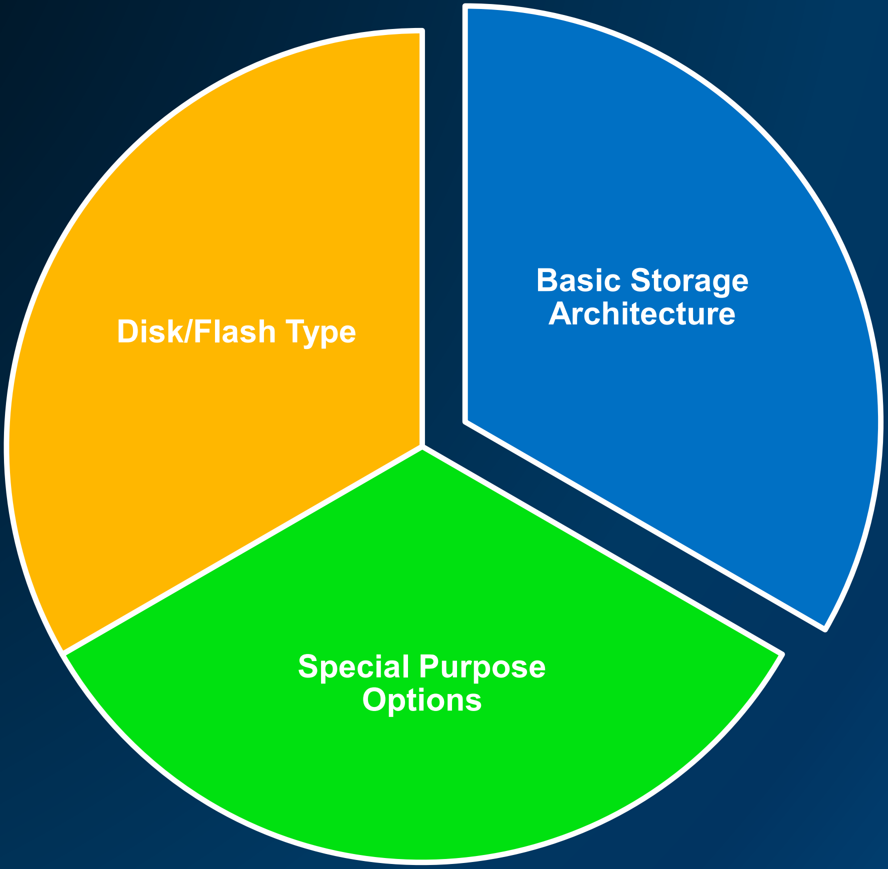
- **Consumer**
- **Enterprise**

PCIe Flash

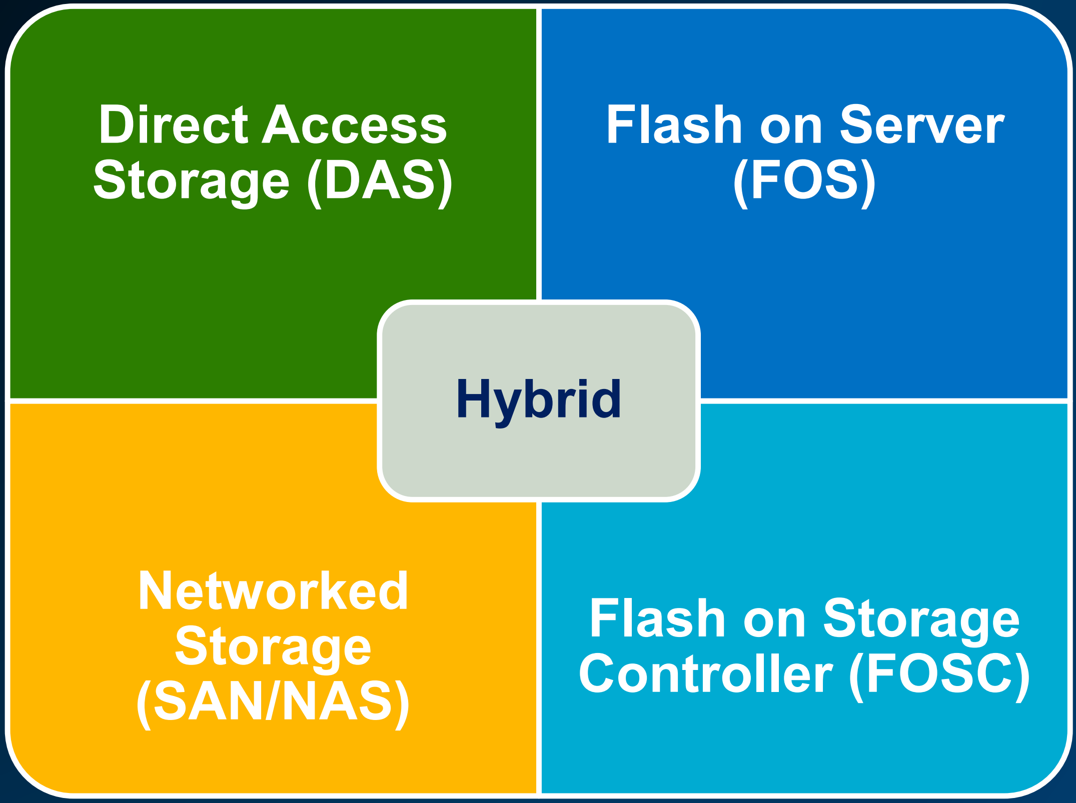
SAS SSDs

SATA SSDs

Storage Base Characteristics

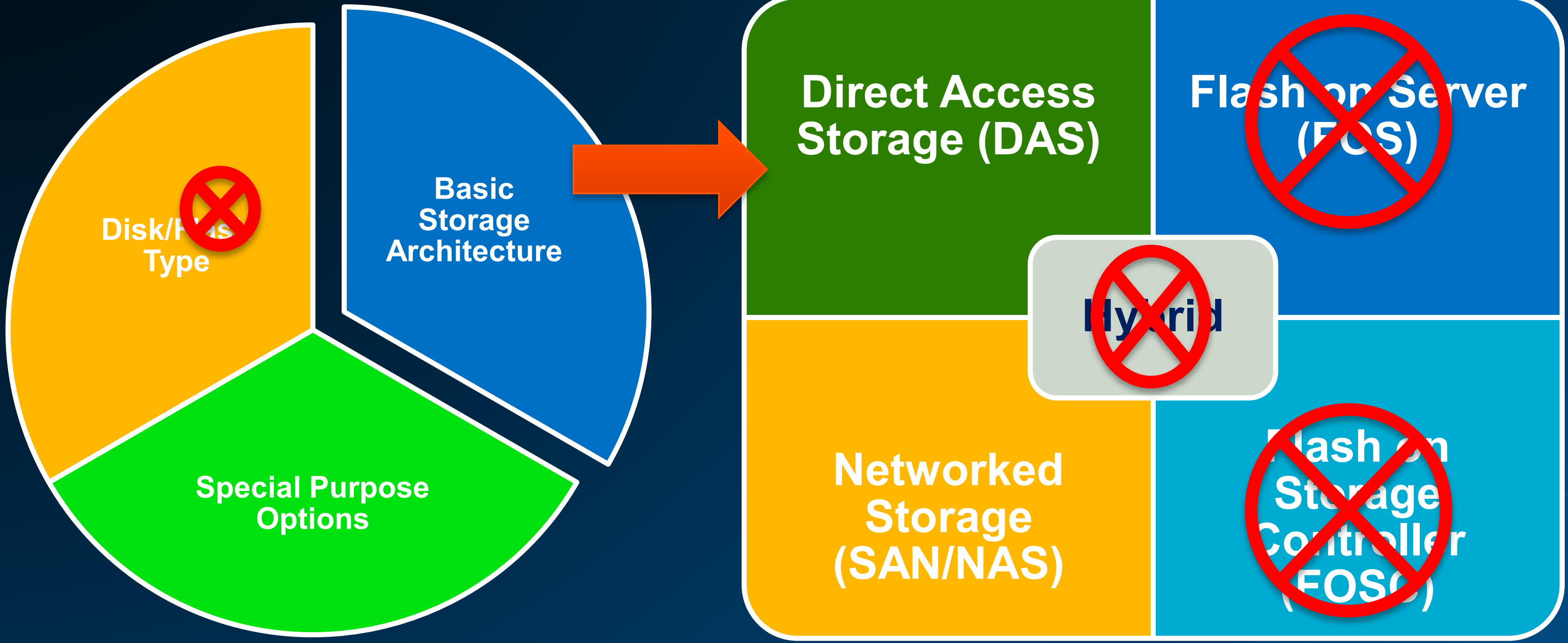


Basic Storage Architecture


Direct Access Storage (DAS)
Flash on Server (FOS)
Hybrid
Networked Storage (SAN/NAS)
Flash on Storage Controller (FOSC)

- http://wikibon.org—The Impact of Flash on Future System and Storage Architectures

Options in the “Good Old Days” (<2004)



Disk/Flash Type

Basic Storage Architecture

Special Purpose Options

Direct Access Storage (DAS)

Flash on Server (FOS)

Hybrid

Networked Storage (SAN/NAS)

Flash on Storage Controller (FOSC)

RAID and/or Caching

Options in the “Good Old Days” (<2004)

HBA

SCSI

Fibre Channel

RAID and/or Caching

Transfer Speeds

Technology	Rate (byte/s)
SCSI (Narrow SCSI) (5 MHz)	5 MB/s
Fast SCSI (8 bits/10 MHz)	10 MB/s
Fast Wide SCSI (16 bits/10 MHz)	20 MB/s
Ultra SCSI (Fast-20 SCSI) (8 bits/20 MHz)	20 MB/s
Ultra Wide SCSI (16 bits/20 MHz)	40 MB/s
Ultra-2 SCSI 40 (Fast-40 SCSI) (8 bits/40 MHz)	40 MB/s
Ultra-2 wide SCSI (16 bits/40 MHz)	80 MB/s
Ultra-3 SCSI (Ultra 160 SCSI; Fast-80 Wide SCSI)	160 MB/s
Ultra-320 SCSI (Ultra4 SCSI)	320 MB/s
Ultra-640 SCSI	640 MB/s

Technology	Rate (byte/s)
Fibre Channel 1GFC (1.0625 GHz)	106.25 MB/s
Fibre Channel 2GFC (2.125 GHz)	212.5 MB/s

Disk

Cache

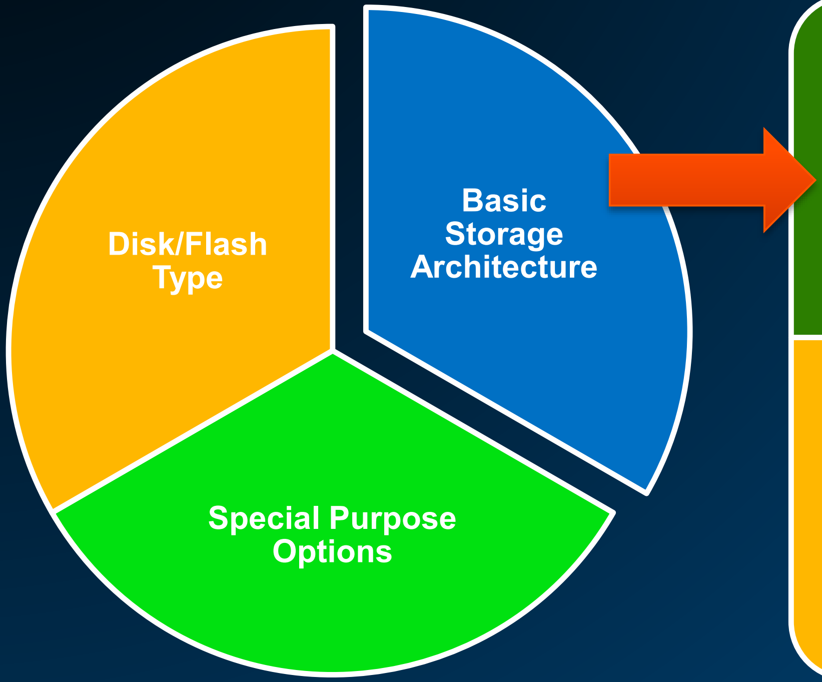
Size

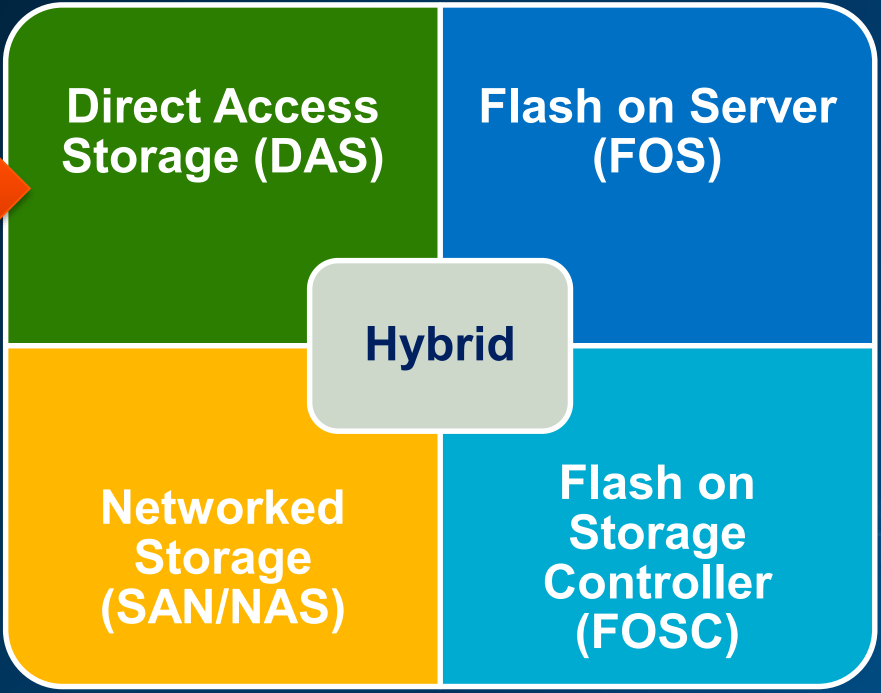
Spinning Magnetic

Media

RPM’s

Plethora of Options Today




Direct Access Storage (DAS)
Flash on Server (FOS)
Hybrid
Networked Storage (SAN/NAS)
Flash on Storage Controller (FOSC)

RAID and/or Caching

Plethora of Options Today

SATA SSDs

SAS SSDs

SAS & SATA

Magnetic Disks

PCIe Flash

Plethora of Options Today

10GbE

InfiniBand

Fibre Channel

RAID and/or Caching

Technology	Rate (byte/s)
SATA revision 1.0	150 MB/s
Serial Attached SCSI (SAS)	300 MB/s
SATA Revision 2.0	300 MB/s
SATA Revision 3.0	600 MB/s
Serial Attached SCSI (SAS) 2	600 MB/s
Serial Attached SCSI (SAS) 3	1,200 MB/s
SATA revision 3.2 - SATA Express	2,000 MB/s
Serial Attached SCSI (SAS) 4 (prelim spec)	2,400 MB/s

Technology	Rate (byte/s)
Fibre Channel 4GFC (4.25 GHz)	425 MB/s
Fibre Channel 8GFC (8.50 GHz)	850 MB/s
Fibre Channel 16GFC (17.0 GHz)	1,500 MB/s

Technology	Rate (byte/s)
iSCSI over Fast Ethernet	12.5 MB/s
iSCSI over gigabit Ethernet	125 MB/s
iSCSI over 10GbE	1,250 MB/s
FCoE over 10GbE	1,250 MB/s
iSCSI over InfiniBand 4×	4,000 MB/s
iSCSI over 100G Ethernet (hypothetical)	12,500 MB/s
FCoE over 100G Ethernet (hypothetical)	12,500 MB/s

Advantages of PCIe Flash

Technology	Rate (byte/s)
PCI Express 1.0 (×1 link)	250 MB/s
PCI Express 1.0 (×2 link)	500 MB/s
PCI Express 2.0 (×1 link)	500 MB/s
PCI Express 3.0 (×1 link)	984.6 MB/s
PCI Express 1.0 (×4 link)	1,000 MB/s
PCI Express 1.0 (×8 link)	2,000 MB/s
PCI Express 2.0 (×4 link)	2,000 MB/s
PCI Express 3.0 (×4 link)	3,934 MB/s
PCI Express 1.0 (×16 link)	4,000 MB/s
PCI Express 2.0 (×8 link)	4,000 MB/s
PCI Express 3.0 (×8 link)	7,880 MB/s
PCI Express 1.0 (×32 link)	8,000 MB/s
PCI Express 2.0 (×16 link)	8,000 MB/s
PCI Express 3.0 (×16 link)	15,7500 MB/s
PCI Express 2.0 (×32 link)	16,000 MB/s
PCI Express 3.0 (×32 link)	31,500 MB/s

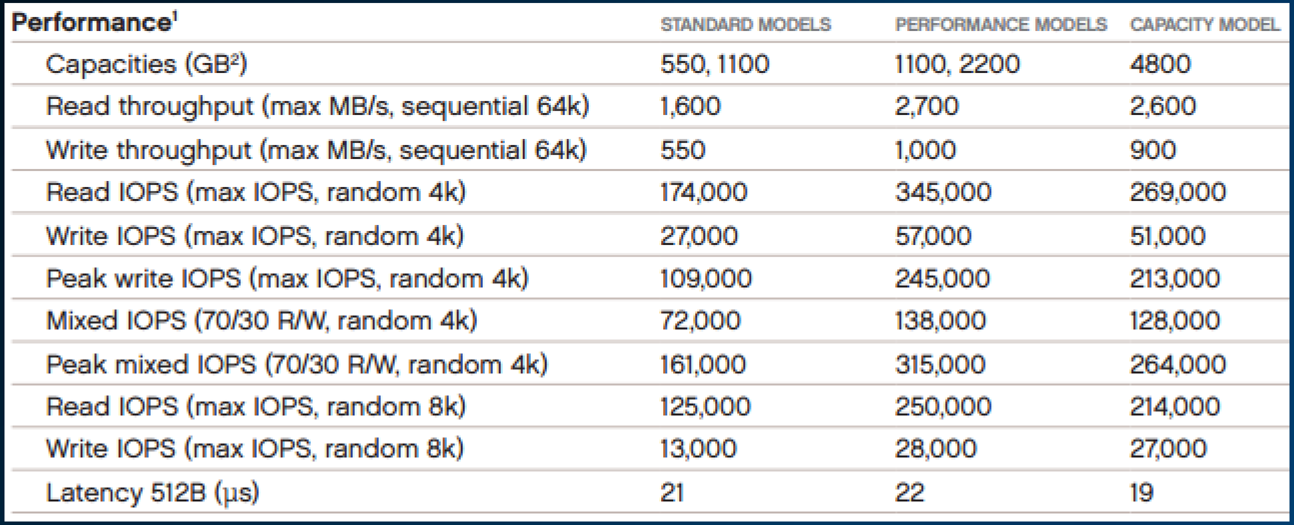
- Performance: The biggest benefit is increased performance. Not only does the PCIe interface have low latency for data transfer, it also bypasses any storage area networking to store or retrieve data. It is, therefore, the fastest way to access data. It delivers microsecond latencies versus millisecond latencies for traditional SAN-based storage.

- Energy Savings: Server-attached PCIe SSDs eliminate the need for additional storage servers, thus saving power on cooling. Traditional storage solutions for high throughput, low latency, and high IOPS need hundreds of hard disk drives, Fibre Channel controllers, and significant amounts of power and cooling.

- Space Savings: PCIe SSDs are compact and fit into the PCIe slot of a server. They eliminate the need for rack space, cooling, and power for storage servers.

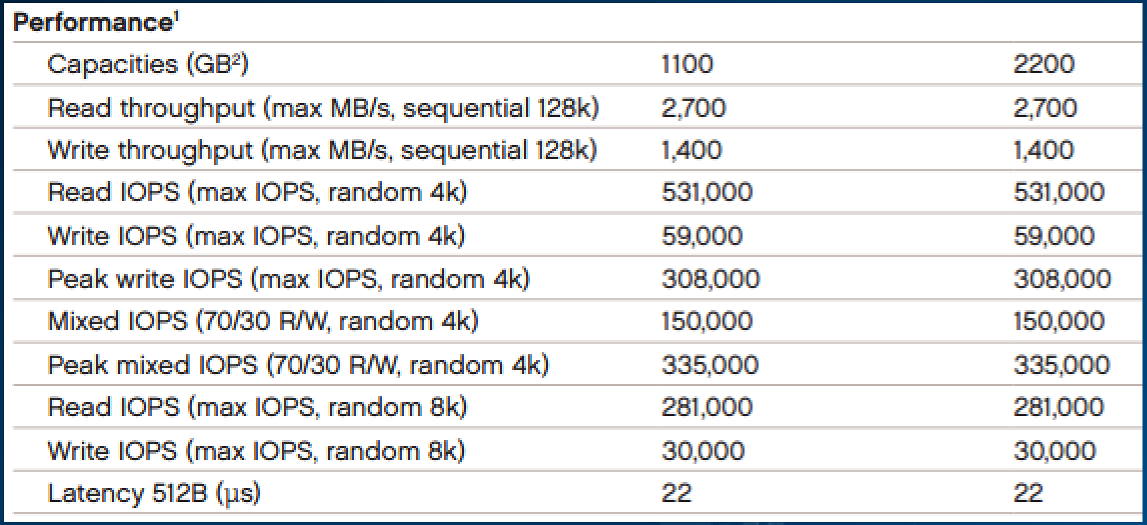
PCIe Flash Speed (IOPS)

HGST FlashMAX II



Performance[1]	STANDARD MODELS	PERFORMANCE MODELS	CAPACITY MODEL
Capacities (GB[2])	550, 1100	1100, 2200	4800
Read throughput (max MB/s, sequential 64k)	1,600	2,700	2,600
Write throughput (max MB/s, sequential 64k)	550	1,000	900
Read IOPS (max IOPS, random 4k)	174,000	345,000	269,000
Write IOPS (max IOPS, random 4k)	27,000	57,000	51,000
Peak write IOPS (max IOPS, random 4k)	109,000	245,000	213,000
Mixed IOPS (70/30 R/W, random 4k)	72,000	138,000	128,000
Peak mixed IOPS (70/30 R/W, random 4k)	161,000	315,000	264,000
Read IOPS (max IOPS, random 8k)	125,000	250,000	214,000
Write IOPS (max IOPS, random 8k)	13,000	28,000	27,000
Latency 512B (µs)	21	22	19

HGST FlashMAX III



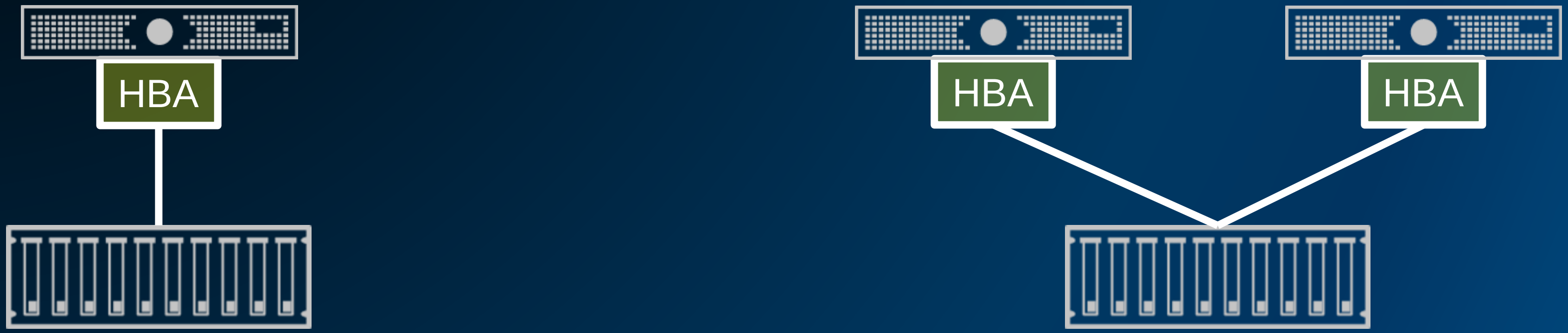
Performance[1]		
Capacities (GB[2])	1100	2200
Read throughput (max MB/s, sequential 128k)	2,700	2,700
Write throughput (max MB/s, sequential 128k)	1,400	1,400
Read IOPS (max IOPS, random 4k)	531,000	531,000
Write IOPS (max IOPS, random 4k)	59,000	59,000
Peak write IOPS (max IOPS, random 4k)	308,000	308,000
Mixed IOPS (70/30 R/W, random 4k)	150,000	150,000
Peak mixed IOPS (70/30 R/W, random 4k)	335,000	335,000
Read IOPS (max IOPS, random 8k)	281,000	281,000
Write IOPS (max IOPS, random 8k)	30,000	30,000
Latency 512B (µs)	22	22

DAS – Still a Viable Option

Single Instance Database

RAC Instance #1

RAC Instance #2



All Disk

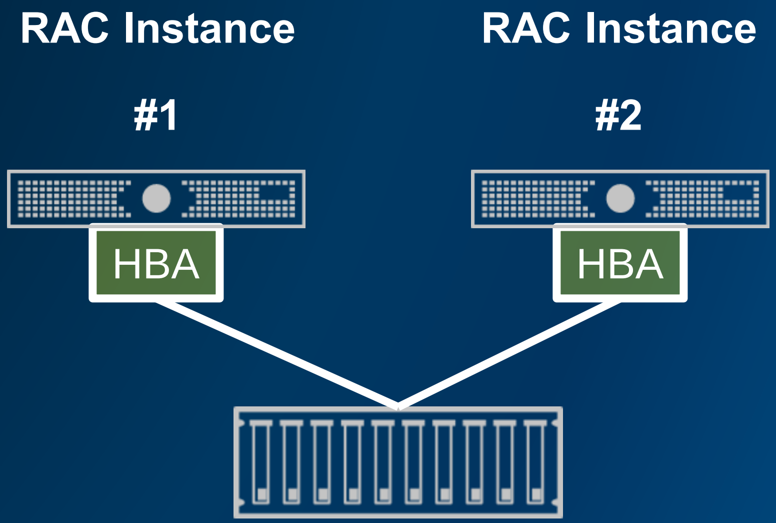
Disk + SSD

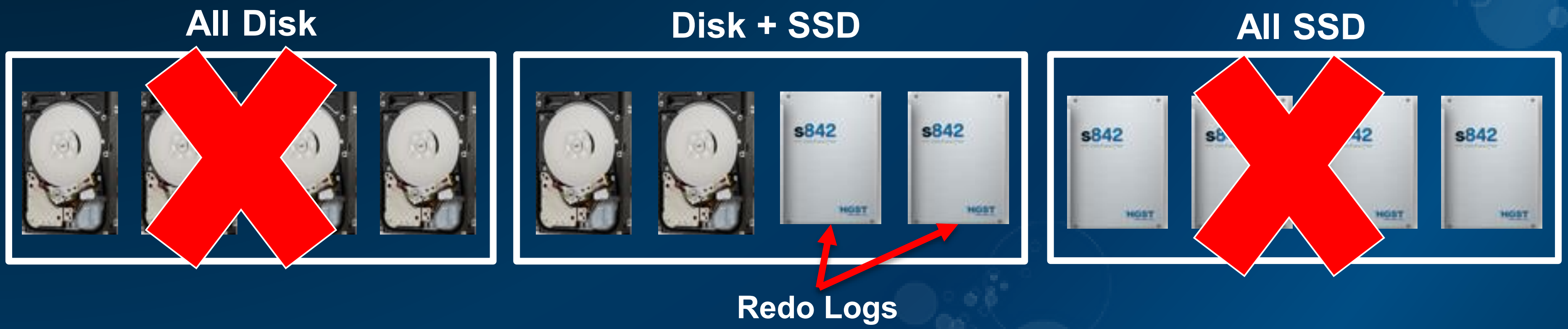
All SSD

Oracle® Data Appliance (ODA X3-X4)

Many Experts' Blogs:

- First thing to go on Flash/SSD should be data
- Redo logs = many sequential writes where spinning disk good enough

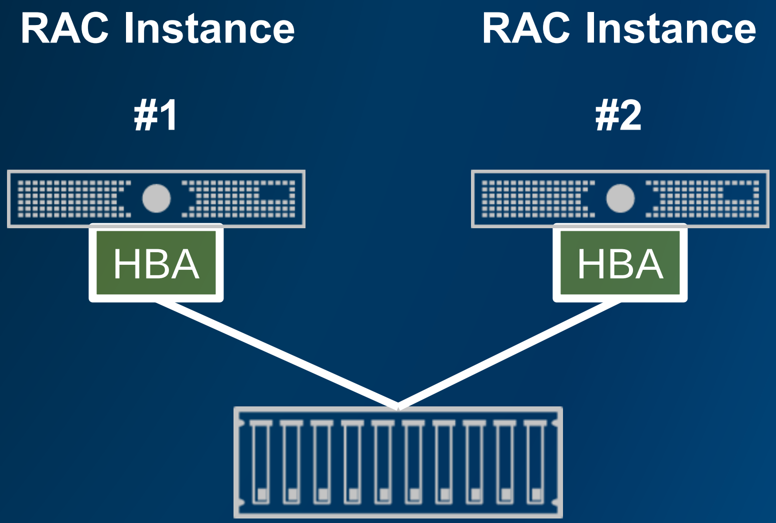


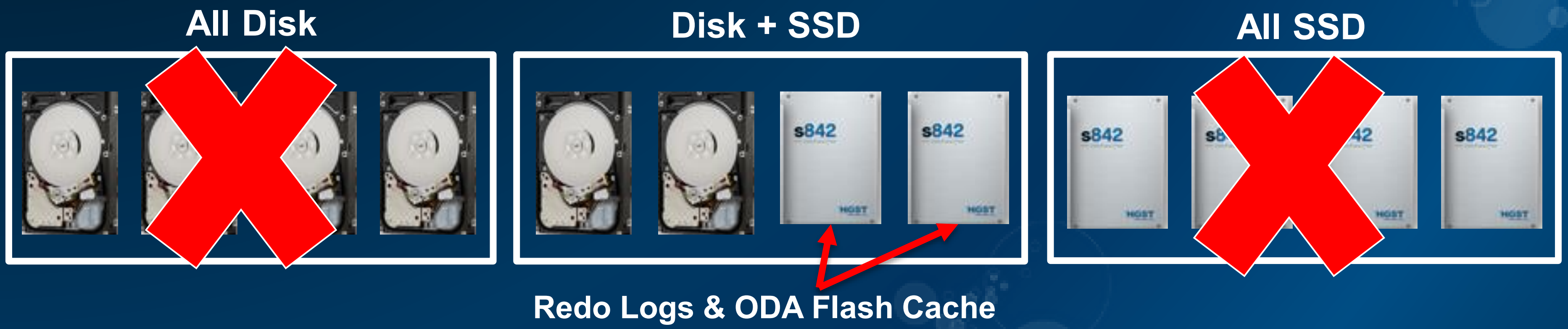


Oracle® Data Appliance (ODA X5)

Many Experts' Blogs:

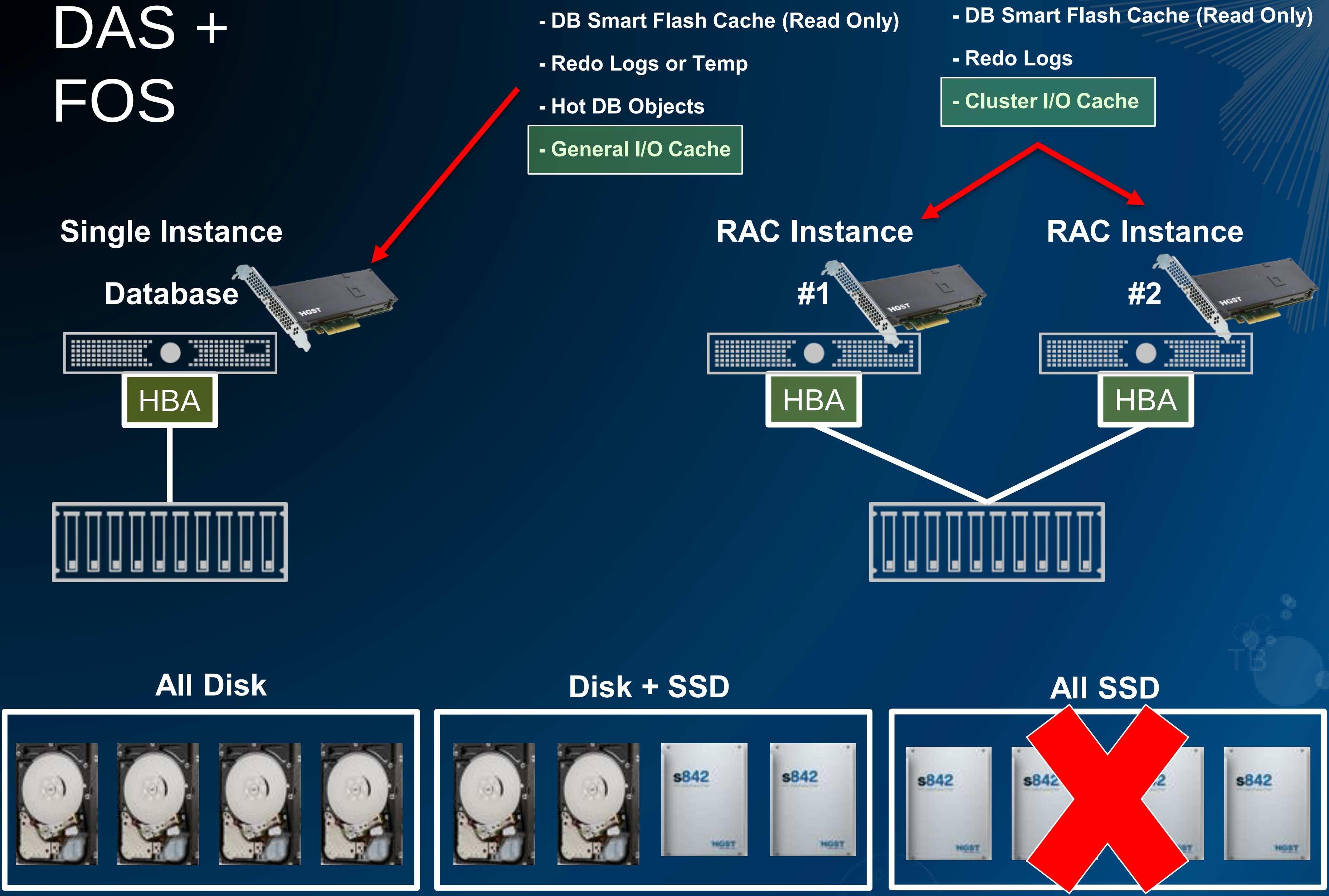
- First thing to go on Flash/SSD should be data
- Redo logs = many sequential writes where spinning disk good enough





Redo Logs & ODA Flash Cache

DAS + FOS



- DB Smart Flash Cache (Read Only)
- Redo Logs or Temp
- Hot DB Objects
- General I/O Cache

- DB Smart Flash Cache (Read Only)
- Redo Logs
- Cluster I/O Cache

Single Instance Database

HBA

RAC Instance #1

HBA

RAC Instance #2

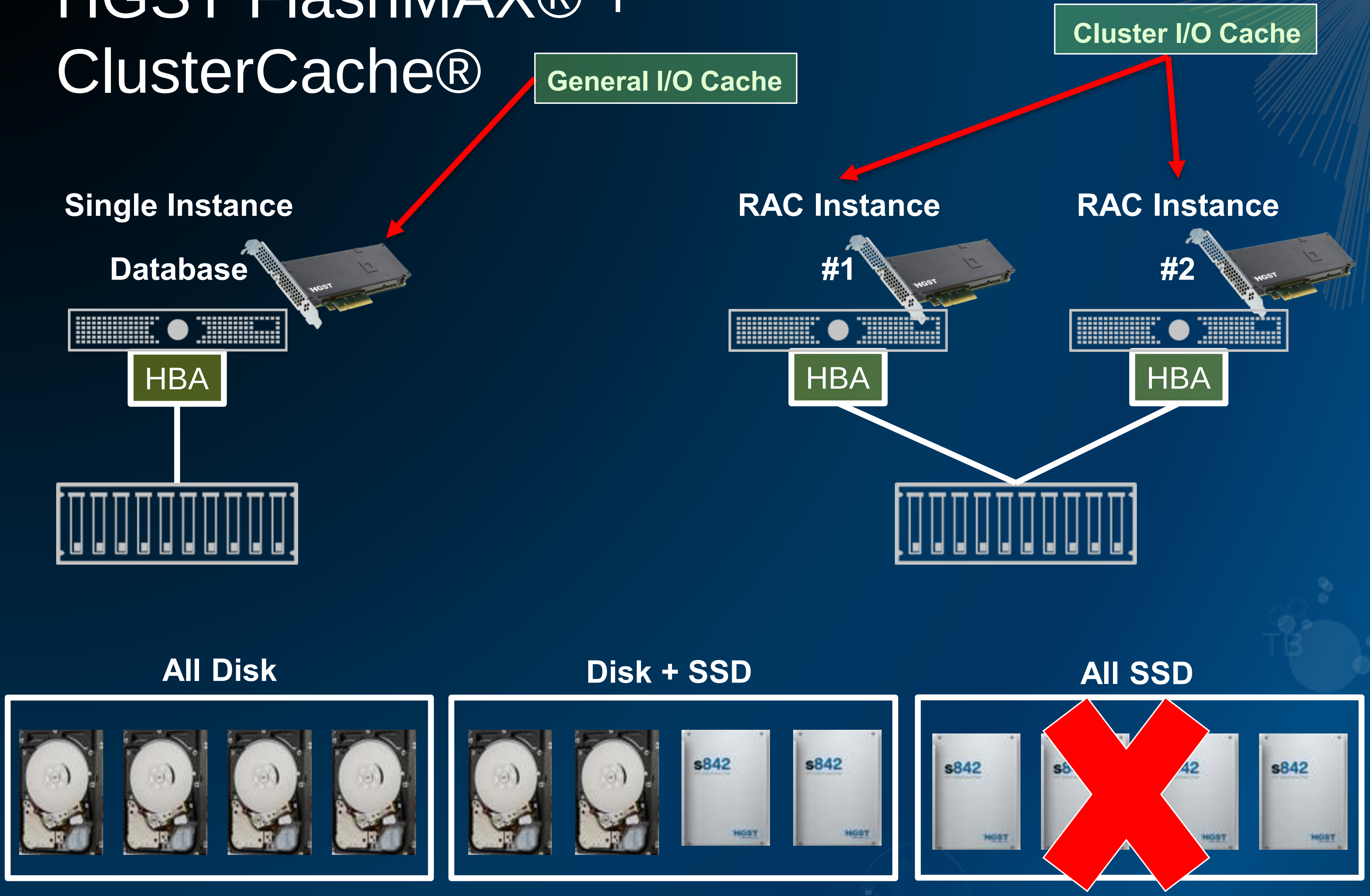
HBA

All Disk

Disk + SSD

All SSD

HGST FlashMAX® + ClusterCache®



General I/O Cache

Cluster I/O Cache

Single Instance Database

RAC Instance #1

RAC Instance #2

HBA

HBA

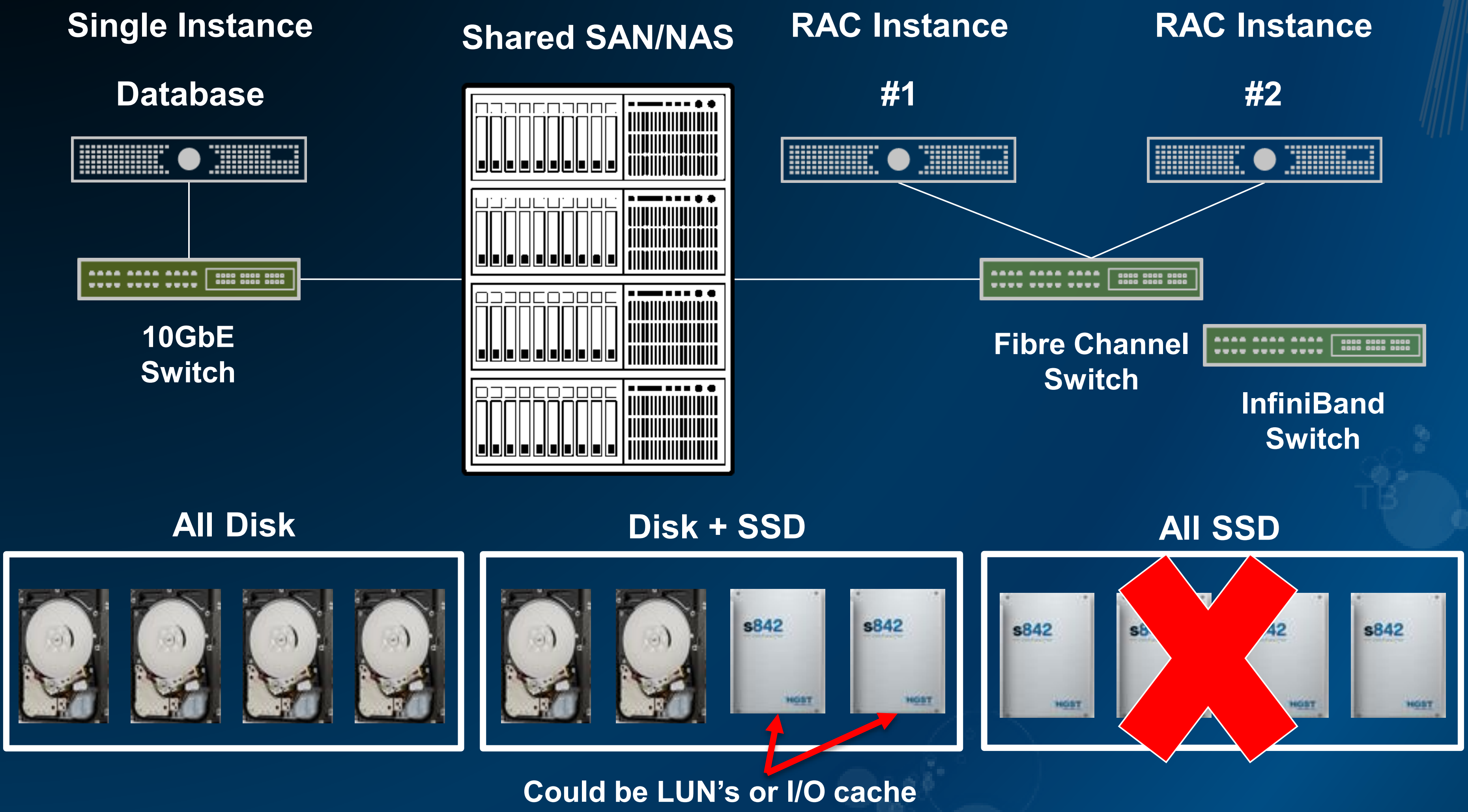
HBA

All Disk

Disk + SSD

All SSD

Networked Storage – Current Mainstay



Single Instance

Database

Shared SAN/NAS

RAC Instance

#1

RAC Instance

#2

10GbE
Switch

Fibre Channel
Switch

InfiniBand
Switch

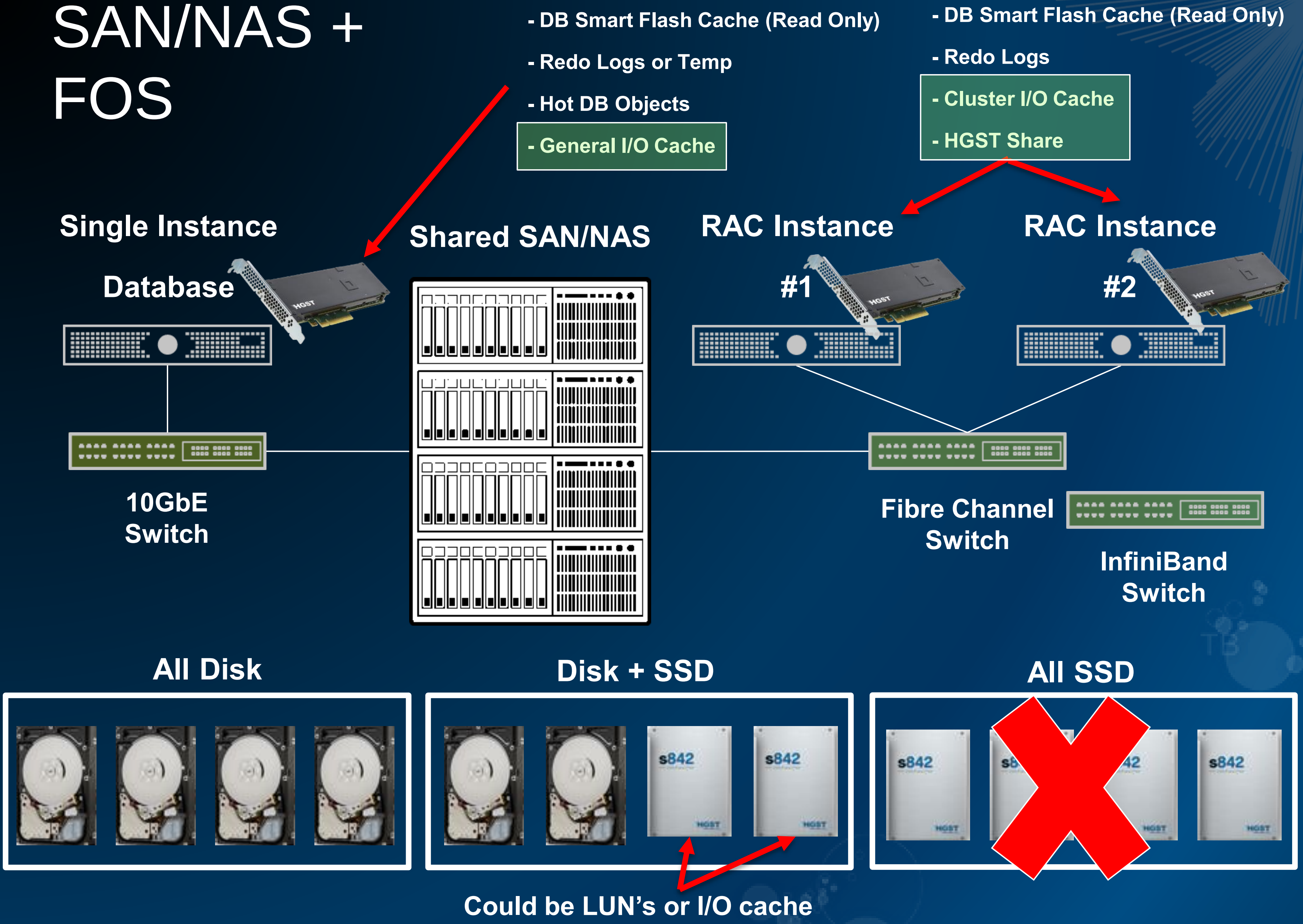
All Disk

Disk + SSD

All SSD

Could be LUN’s or I/O cache

SAN/NAS + FOS



- DB Smart Flash Cache (Read Only)
- Redo Logs or Temp
- Hot DB Objects
- General I/O Cache

- DB Smart Flash Cache (Read Only)
- Redo Logs
- Cluster I/O Cache
- HGST Share

Single Instance Database

Shared SAN/NAS

RAC Instance #1

RAC Instance #2

10GbE Switch

Fibre Channel Switch

InfiniBand Switch

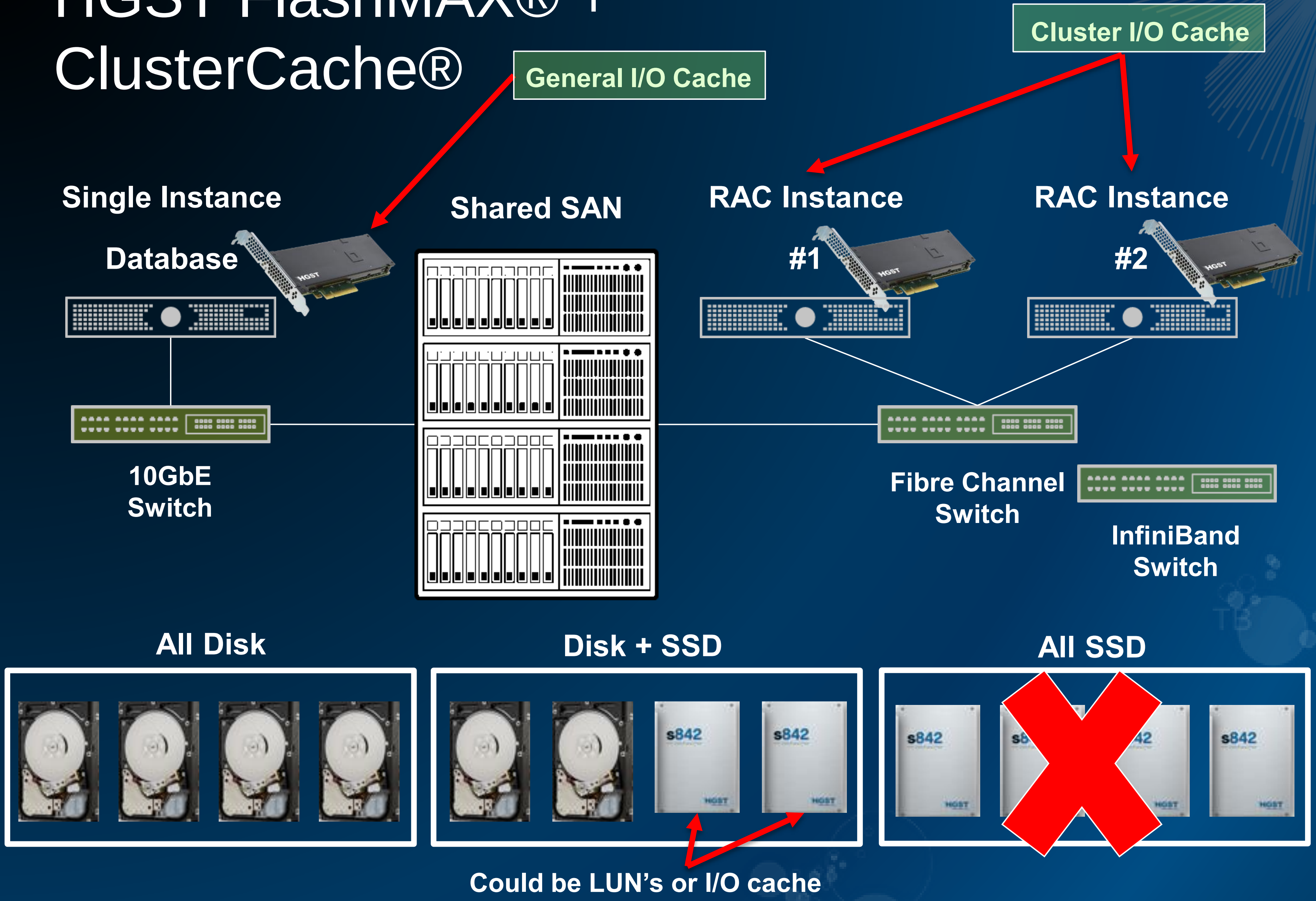
All Disk

Disk + SSD

All SSD

Could be LUN's or I/O cache

HGST FlashMAX® + ClusterCache®



General I/O Cache

Cluster I/O Cache

Single Instance Database

Shared SAN

RAC Instance #1

RAC Instance #2

10GbE Switch

Fibre Channel Switch

InfiniBand Switch

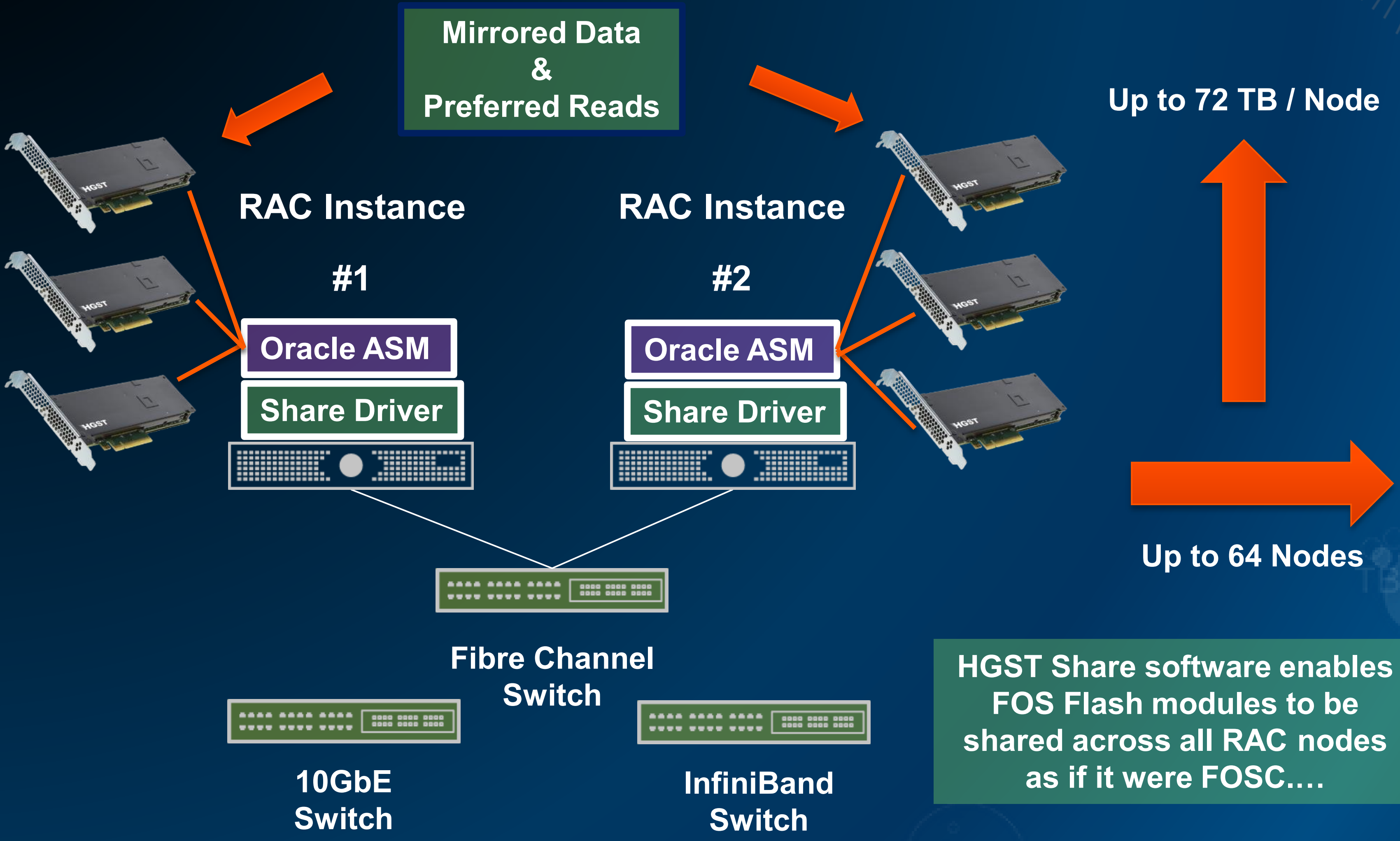
All Disk

Disk + SSD

All SSD

Could be LUN's or I/O cache

HGST FlashMAX® + Share®



Mirrored Data
&
Preferred Reads

RAC Instance

#1

Oracle ASM

Share Driver

RAC Instance

#2

Oracle ASM

Share Driver

Fibre Channel
Switch

10GbE
Switch

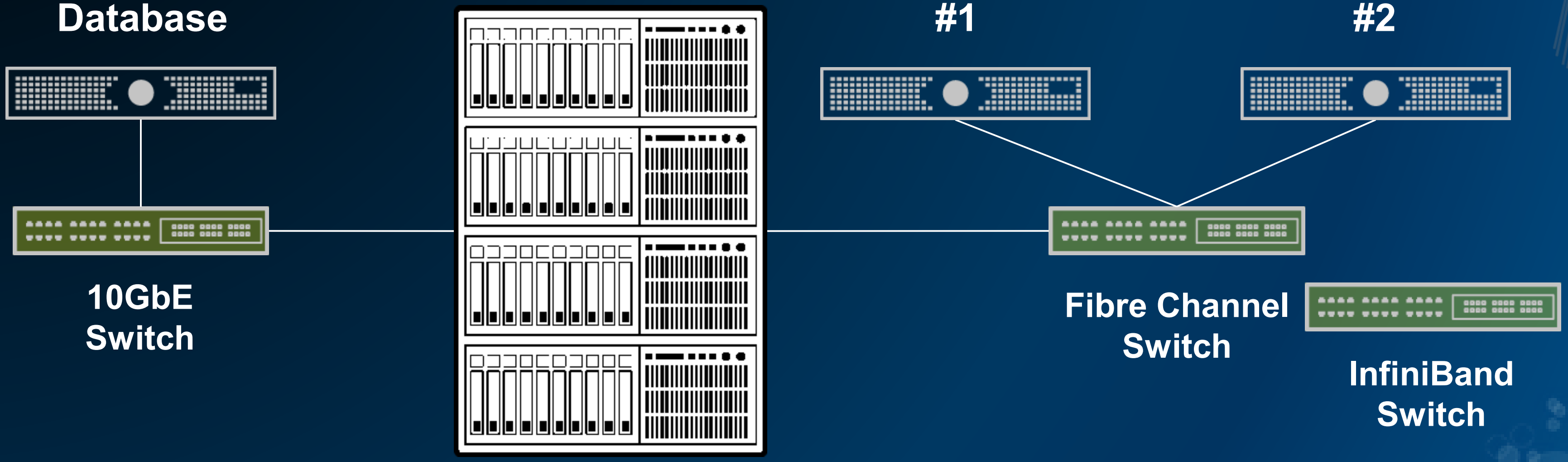
InfiniBand
Switch

Up to 72 TB / Node

Up to 64 Nodes

HGST Share software enables FOS Flash modules to be shared across all RAC nodes as if it were FOSC….

SAN/NAS + FOSC



Single Instance

Database

10GbE Switch

Shared SAN/NAS

RAC Instance

#1

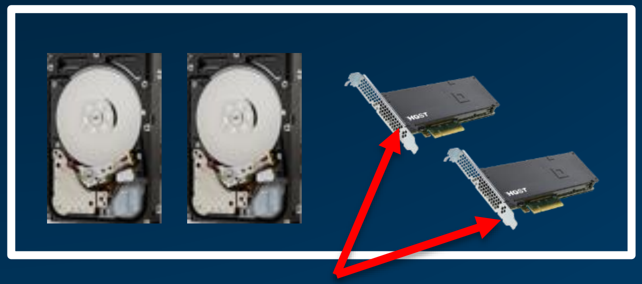
RAC Instance

#2

Fibre Channel Switch

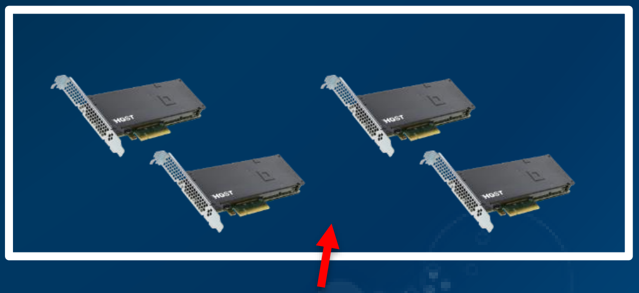
InfiniBand Switch

Disk + Flash



Could be LUN's or I/O cache

All Flash

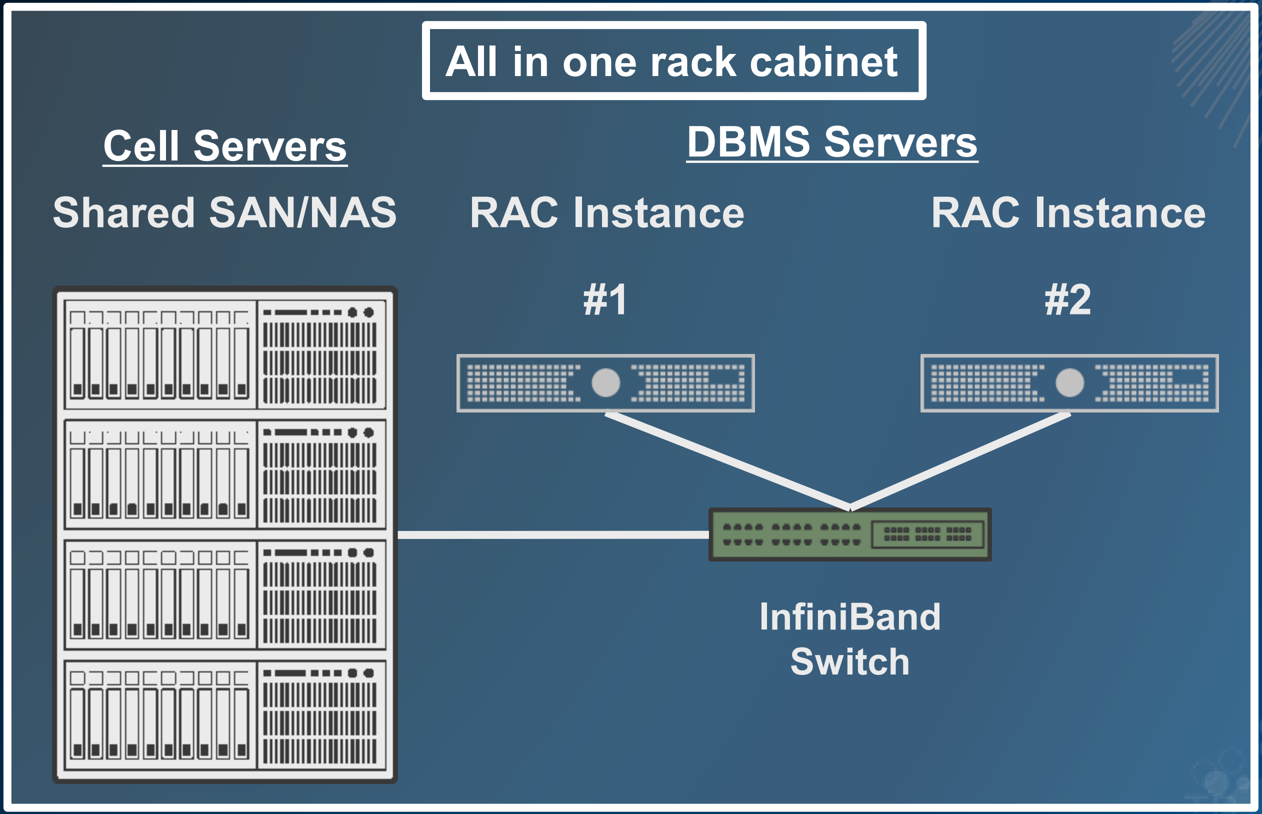


Could be LUN's or I/O cache

Oracle® Exadata X2-X3

Cell-Offloading / Smart-Scan

- Column Filtering
- Row Filtering
- JOIN Filtering
- Storage Indexes
- Function Offload
- Virtual Columns
- HCC Decompress
- Decryption



All in one rack cabinet

Cell Servers

Shared SAN/NAS

DBMS Servers

RAC Instance

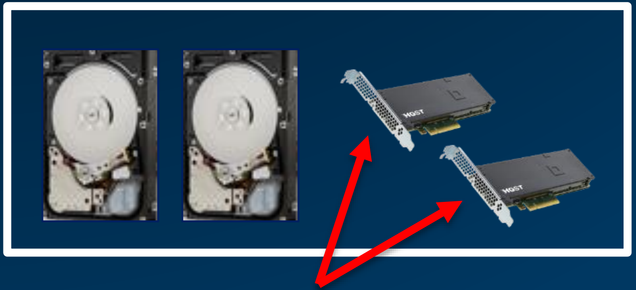
#1

RAC Instance

#2

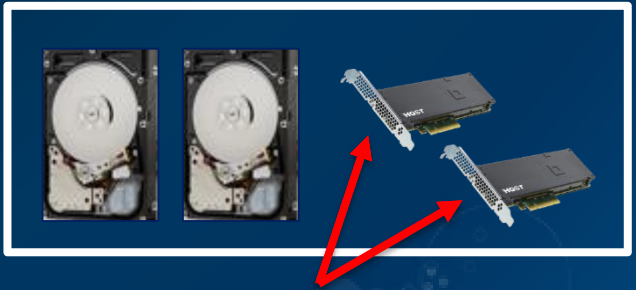
InfiniBand
Switch

HC = 7,200 RPM Disk + Flash



Could be LUN’s or I/O cache

HP = 15,000 RPM Disk + Flash

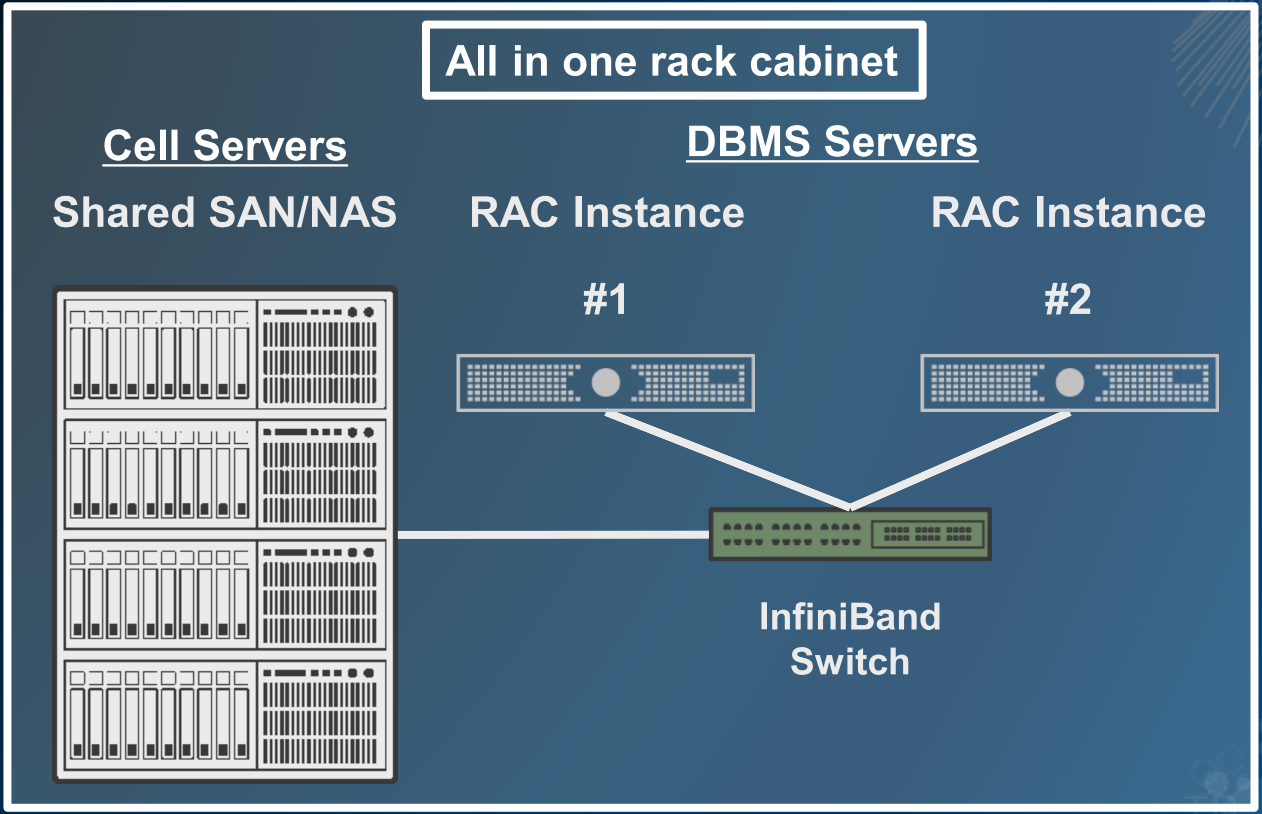


Could be LUN’s or I/O cache

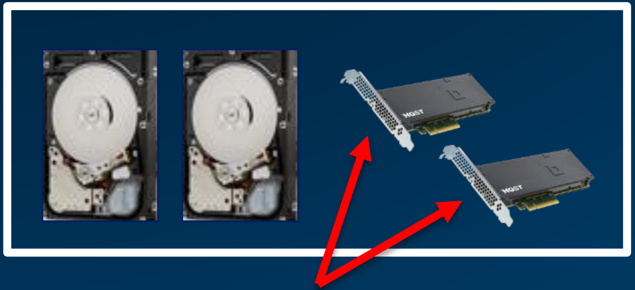
Oracle® Exadata X4-X5

Cell-Offloading / Smart-Scan

- Column Filtering
- Row Filtering
- JOIN Filtering
- Storage Indexes
- Function Offload
- Virtual Columns
- HCC Decompress
- Decryption

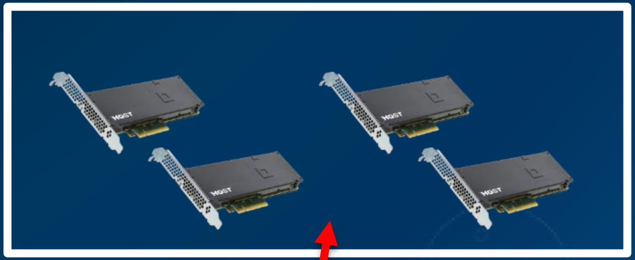


HC = 7,200 RPM Disk + Flash



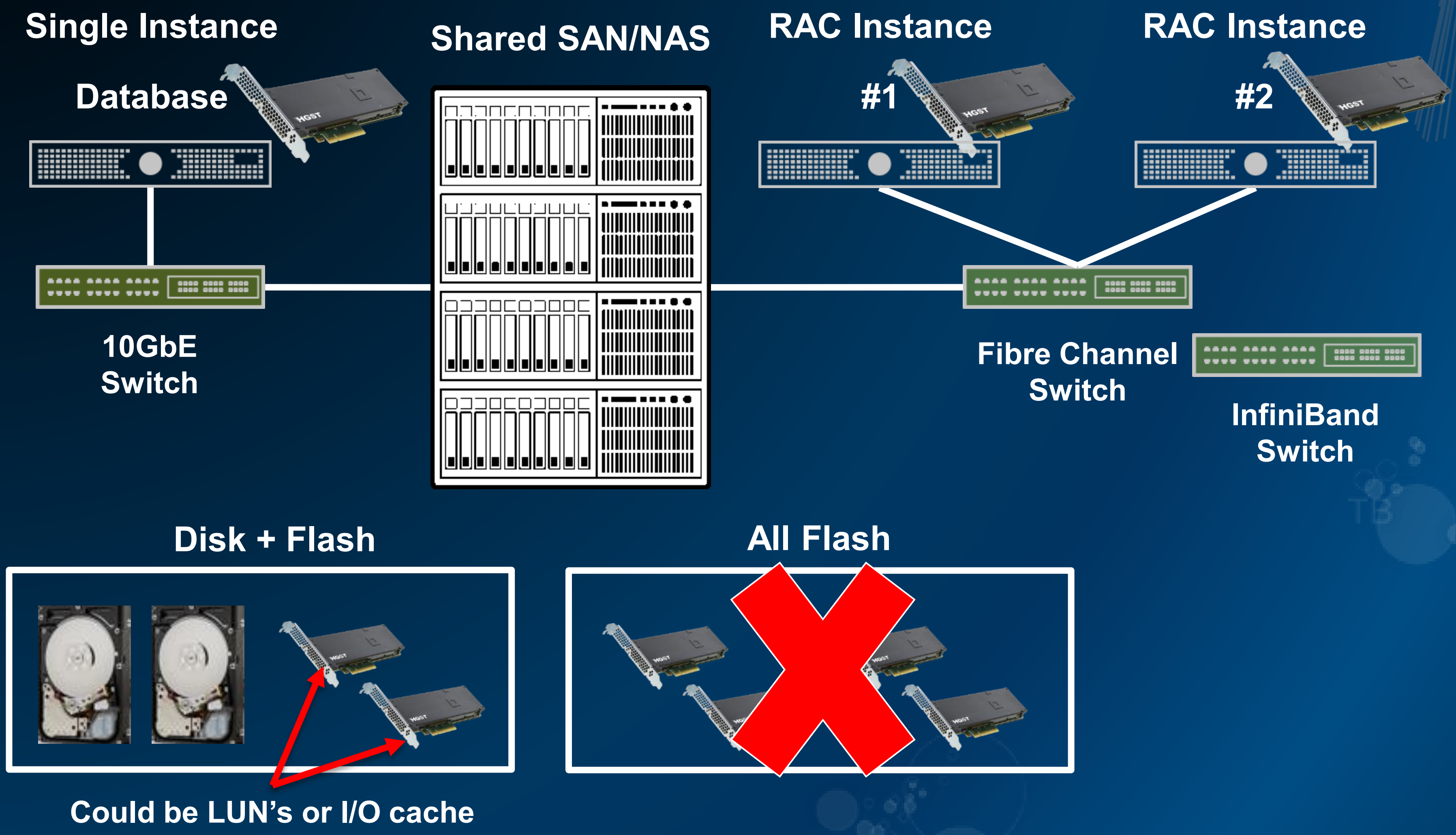
Could be LUN's or I/O cache

HP = All Flash



Could be LUN's or I/O cache

SAN/NAS + FOS + FOSC



Single Instance

Database

10GbE Switch

Shared SAN/NAS

RAC Instance #1

RAC Instance #2

Fibre Channel Switch

InfiniBand Switch

Disk + Flash

All Flash

Could be LUN’s or I/O cache

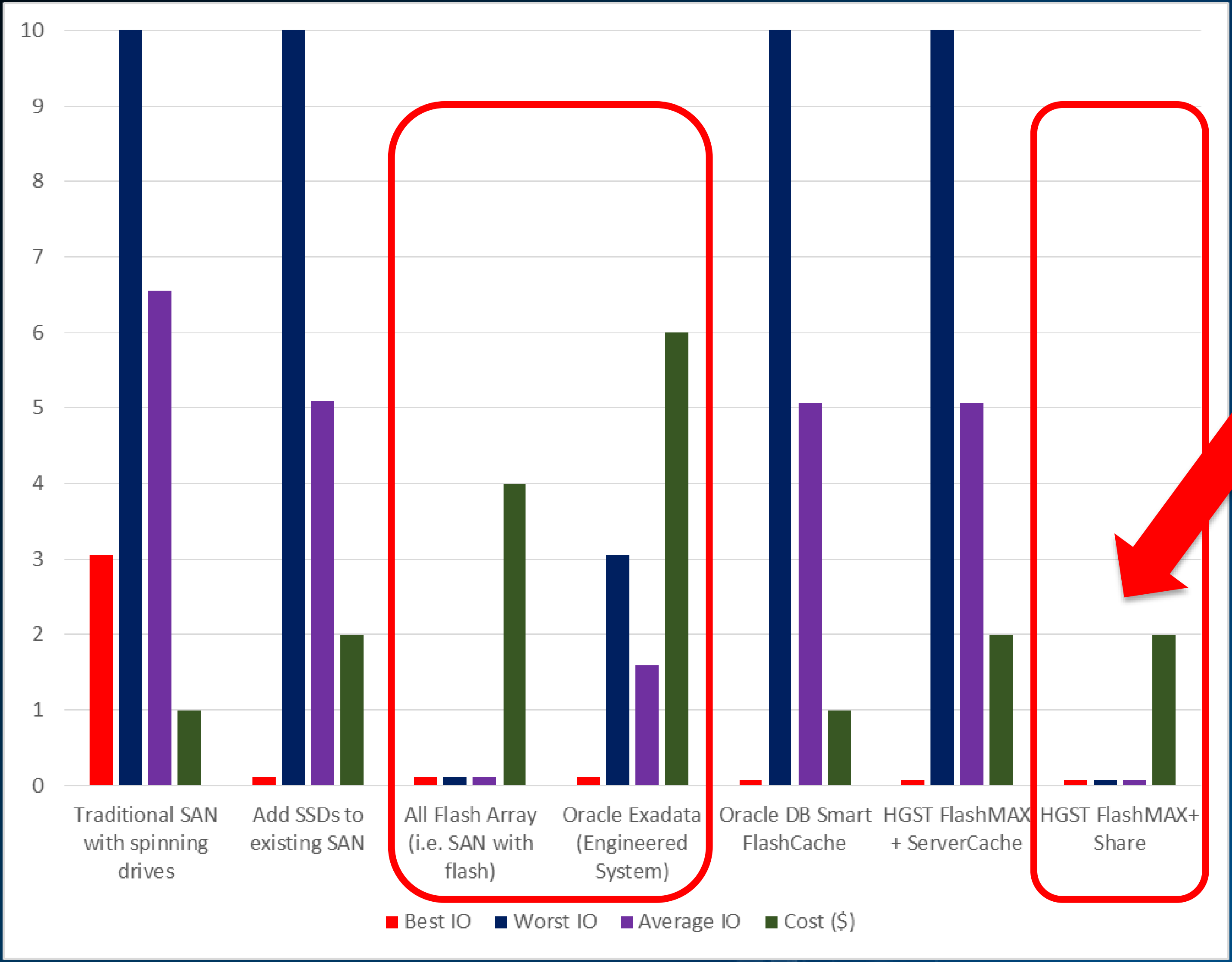
Compare Storage Options for Oracle® RAC

Rough "Theoretical" Comparison

	Best (ms)	Worst (ms)	Cost
Traditional SAN with spinning drives	3 disk + FC	10 disk + FC	$
Add SSDs to existing SAN	.06 flash + FC	10 disk + FC	$$
All-Flash Array (SAN with all flash)	.06 flash + FC	.06 flash + FC	$$$$
Oracle Exadata (Engineered System)	.06 flash + IB	3 disk + IB	$$$$$$
Oracle DB Smart Flash Cache	.06 flash	10 disk + FC	$
HGST FlashMAX® + HGST ClusterCache	.06 flash	10 disk + FC	$$
HGST FlashMAX® + HGST Share	.06 flash	.06 flash	$$

Assumption: network access + transfer time = .06 ms

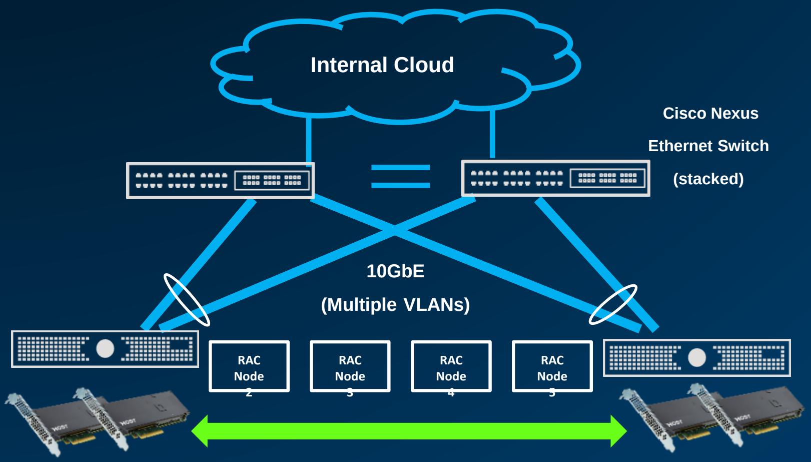
Compare Storage Options for Oracle® RAC



Oracle® RAC Solution—Major Telecom Win

HGST Share software running on HGST FlashMAX “blew away” the incumbent technology.

Reference architecture established for RAC deployments in all of the customer’s business units.



Internal Cloud

Cisco Nexus Ethernet Switch (stacked)

10GbE (Multiple VLANs)

RAC Node 2

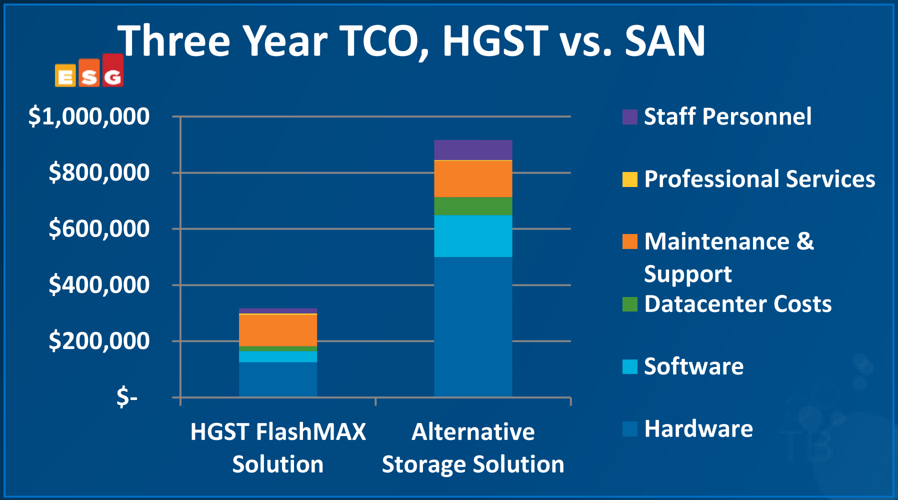
RAC Node 3

RAC Node 4

RAC Node 5

Why we won:

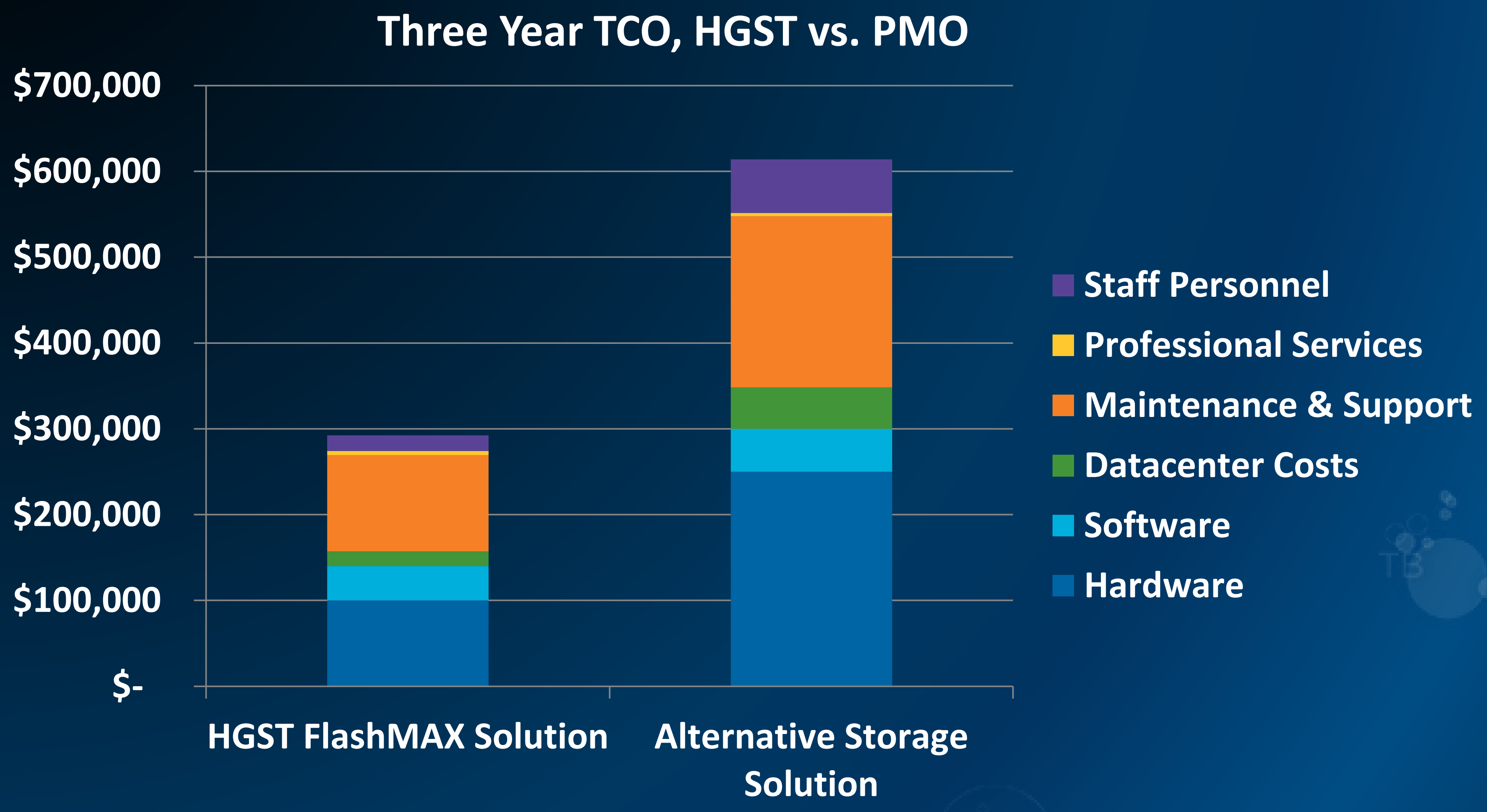
- **Performance**: 6x improvement over SAN
- **Cost**: 1/3rd the cost of SAN



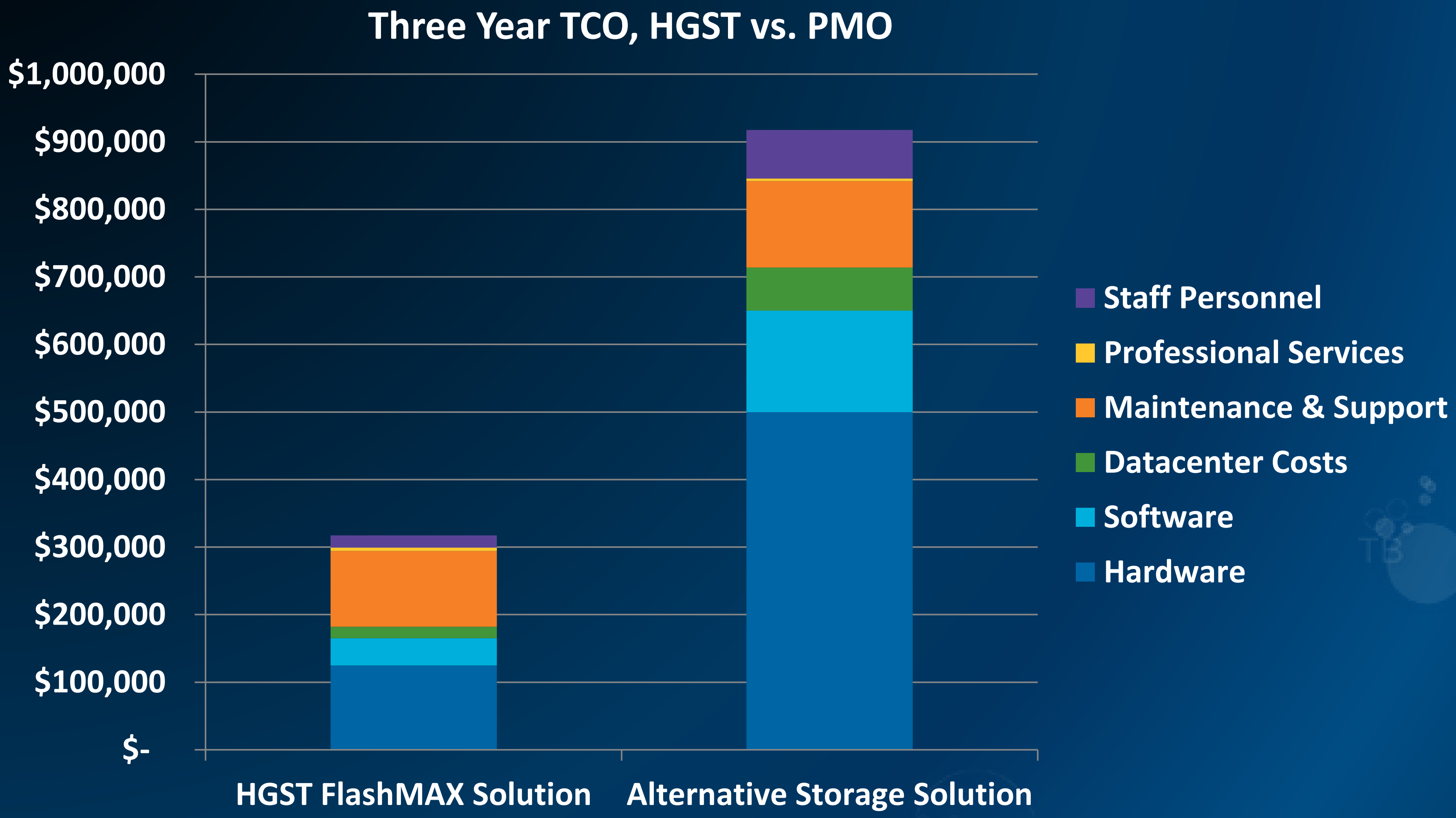
Three Year TCO, HGST vs. SAN

- **Ease of Use**: We look like “any other LUN”
- **Energy Efficient**: Reduced power/cooling
- **Validation**: Solution on Oracle web site

3-Year TCO vs. All-Flash Array—700K IOPs



3-Year TCO vs. Enterprise SAN—700K IOPs



HGST Free Performance Assessment

- Process-driven analysis tied to actual workloads
- Performed by our in house Oracle ACE, Mr. Scalzo
- Completely secure
  - ORAchk, Diagnostics & Tuning Packs
  - Only accesses data dictionary & metadata
  - HGST reviews/parses text output
- 3 steps to actionable insights
  - Collection
  - Analysis
  - Read-out
- Recommendations on tuning and potential benefits of Flash

Q & A

HGST FlashMAX® is a registered trademark of HGST, Inc. and its affiliates in the U.S. and/or other countries. LongLiveData™, HGST, Inc. and its affiliates in the U.S. and/or other countries. Oracle® and Oracle-based names are either registered trademarks or trademarks of Oracle ® Corporation and/or its affiliates. All other marks are the property of their respective owners.

Thank You

Roye Avidor
Technical Marketing Engineer, HGST
roye.Avidor@hgst.com