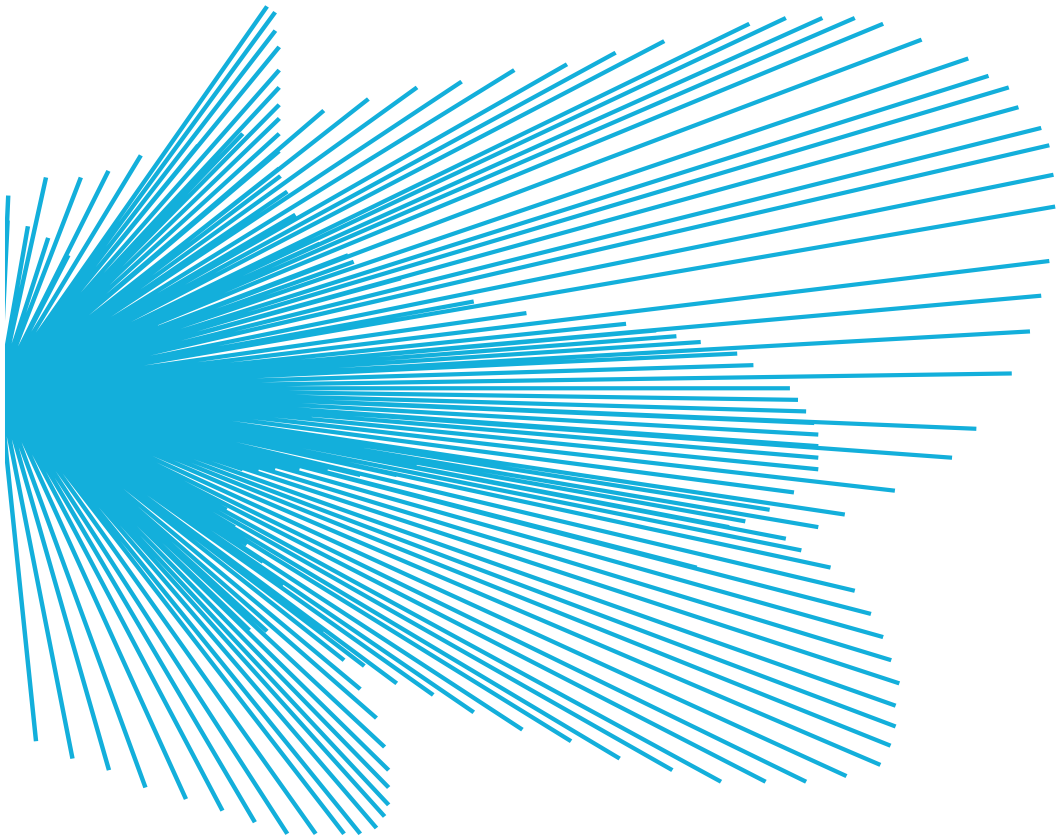
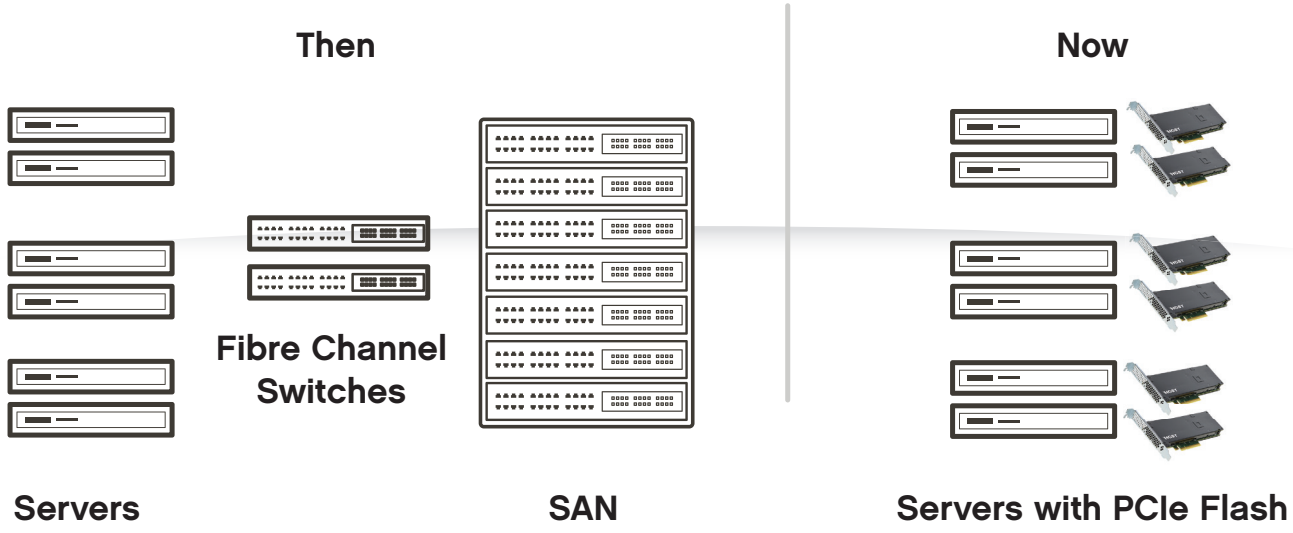


All-Flash Server-Side Storage for Oracle® Real Application Clusters (RAC) on Oracle Linux

HGST
a Western Digital company

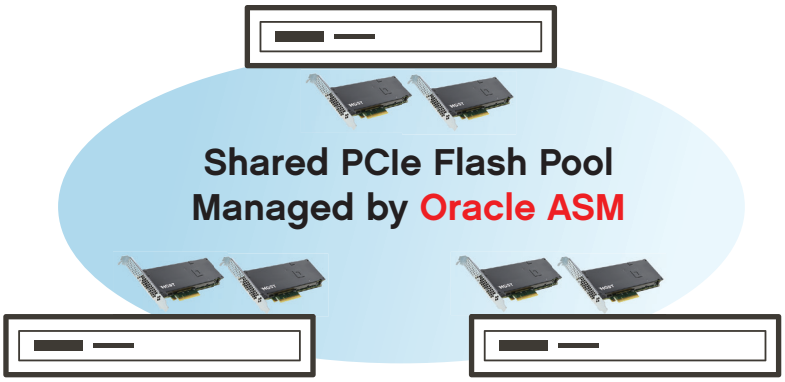
ORACLE Gold Partner

Traditional SAN storage systems cannot keep up with growing application performance needs. The HGST FlashMAX server-side flash platform together with Oracle® Automatic Storage Manager (ASM) offers a new cost-effective approach to implementing high-performance scalable flash storage for Oracle RAC databases.



Solution Architecture Highlights

- HGST FlashMAX® II PCIe cards inside servers used as primary storage with shared access across the cluster
- Oracle ASM performs volume management and data mirroring
- 0.5TB to 72TB of flash in a standard x86 server
- Consistently high performance with linear scalability
- Distributed architecture with no single point of failure
- At a fraction of the SAN cost



Validated by Oracle

Robustness of the novel architecture has been validated with Oracle Linux Test (OLT) suite on Oracle Linux 6.5 with Unbreakable Enterprise Kernel (UEK) Release 3. The validated status ensures that customers get well-coordinated support from both HGST and Oracle.

Validated configuration details: http://linux.oracle.com/pls/apex/f?p=102:2:::NO::P2_VC_ID:649

OLT suite description: https://oss.oracle.com/projects/olt/dist/documentation/OLT_TestCoverage.pdf

Example of clustered database performance reported by Oracle CALIBRATE_IO.

Test Configuration:

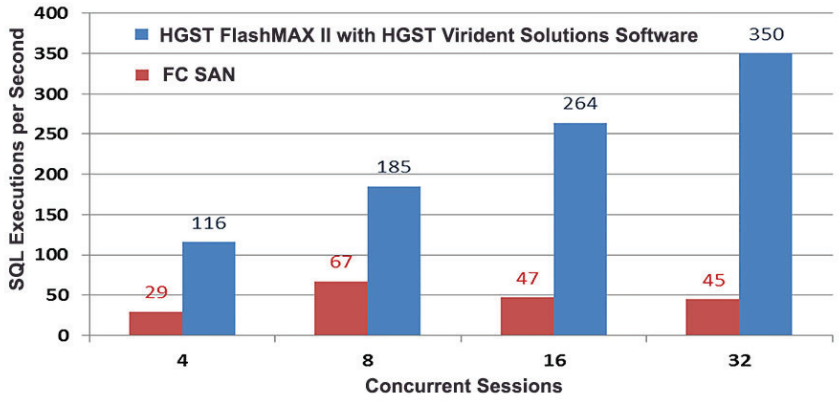
- 3-node RAC cluster
- Two FlashMAX II 2.2 TB cards per node
- CPU: Dual Xeon E5-2690
- 8KB database page size

Calibrate_IO results:

MAX_IOPS = 1076921
LATENCY = 0
MAX_MBPS = 12249

Example of SLOB2 performance improvements in an actual customer environment.

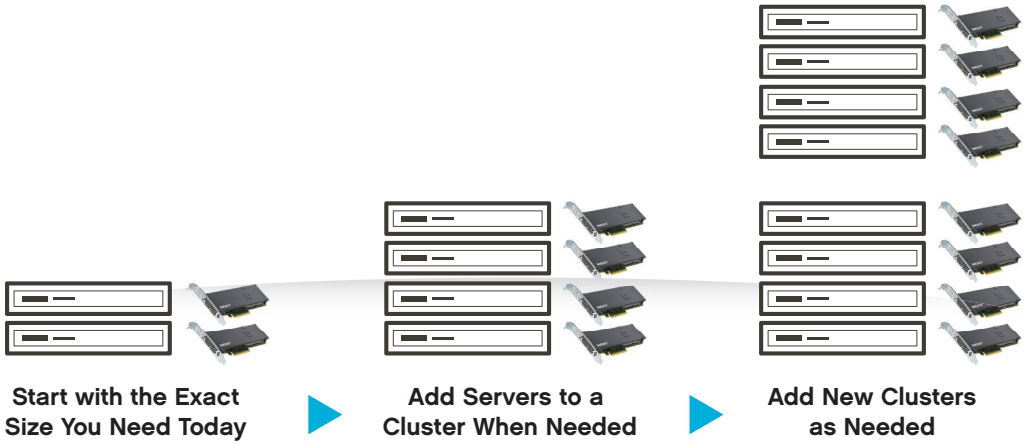
- 2-node Oracle RAC cluster
- One FlashMAX II 1.1 TB card per node
- Compared to existing FC SAN



Consistently High Performance

Having all data on flash ensures that the performance stays consistently high regardless of the hot data size. Placing flash inside the servers with direct PCIe connectivity ensures the lowest latencies and the highest bandwidth. The non-blocking InfiniBand interconnect with Remote Direct Memory Access (RDMA) technology keeps the microsecond latencies while achieving up to 1 million IOPS and up to 10GB/s of bandwidth per node.

Linear Scaling of Storage Performance and Capacity



Linear Scalability

When adding more nodes to a cluster, each new node adds both storage capacity and performance. Separate clusters use their own separate storage inside each cluster without interference between the clusters. The architecture is fully distributed. There is no centralized controller limiting the performance. With the non-blocking InfiniBand interconnect the storage performance can scale linearly up to 64 nodes per cluster with unlimited number of clusters.

Shared Access to PCIe Flash with HGST Virident Share

Shared access to storage from all nodes of a cluster is a mandatory requirement with Oracle RAC. HGST Virident Share software provides shared access to all FlashMAX PCIe flash drives across the cluster. With Virident Share each FlashMAX II card in a cluster is accessible from all nodes in the cluster as if it was a local block storage device. Share comes with additional tools that provide integration with Oracle ASM and simplify configuration and maintenance tasks.

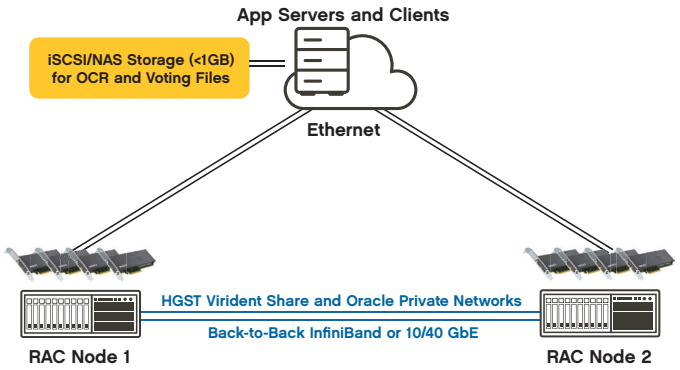
Network Configuration Options

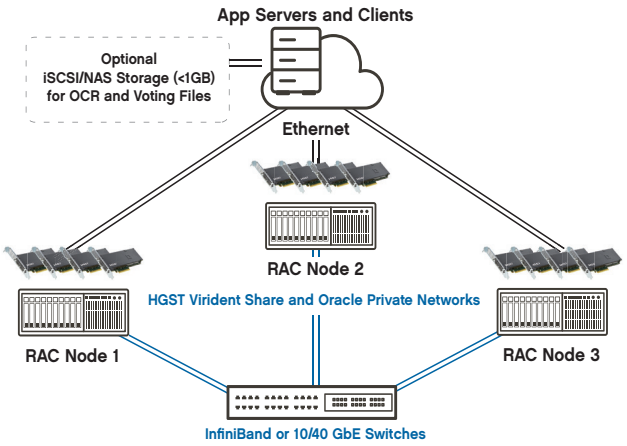
HGST Virident Share provides the following interconnect options:

- InfiniBand 56 Gb/s or 40 Gb/s with RDMA for the best performance
- 10 GbE or 40 GbE for broad interoperability

2-node clusters can use direct back-to-back links without using additional switches. If there are 3 or more nodes in a cluster then using switches is required.

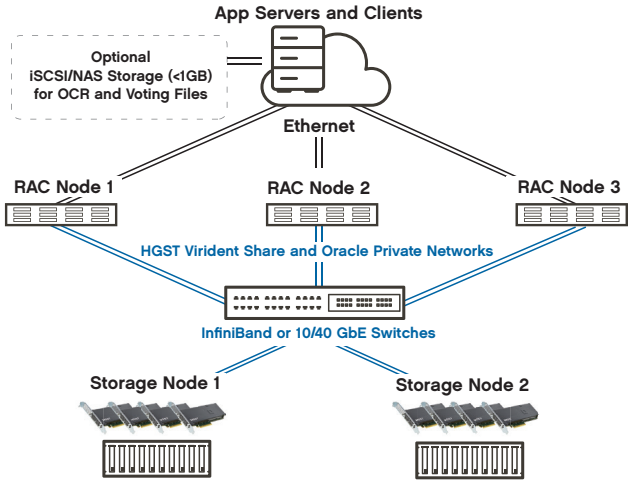
High availability of Virident Share interconnect is achieved by having two links to each node. In configurations with switches these two links should be connected to two separate switches. When using InfiniBand, Virident Share provides embedded dual-path capability. When using 10/40 GbE, standard network bonding mechanisms are available. In either case, there is no need for a separate multipath driver.





Separate Database or Storage Nodes

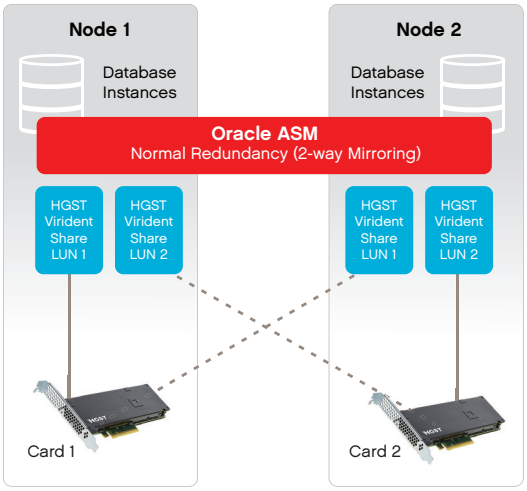
Typically, installing PCIe flash cards inside each database server is the most effective configuration. However, in some scenarios it makes sense to have additional storage nodes, to have additional database nodes, or to have separate database or storage nodes. All of these options are supported.



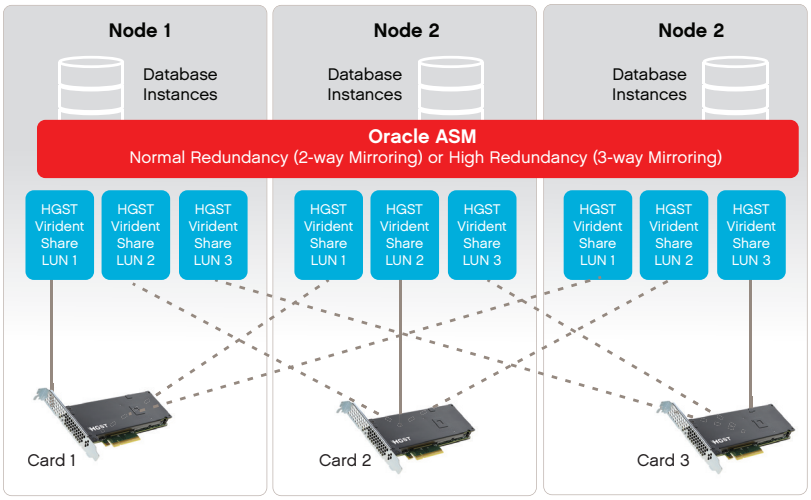
High Availability

The distributed architecture has no single point of failure. Oracle ASM performs volume management including mirroring and striping data across cards and nodes. To ensure mirroring of the data, "Normal Redundancy" (2-way mirroring) or "High Redundancy" (3-way mirroring) must be selected when creating an ASM disk group. Each block of data has mirrored copies on two or three nodes depending on redundancy level selected for the ASM disk group.

Device access mapping with 2 nodes and 1 card per node.



Device access mapping with 3 nodes and 1 card per node.

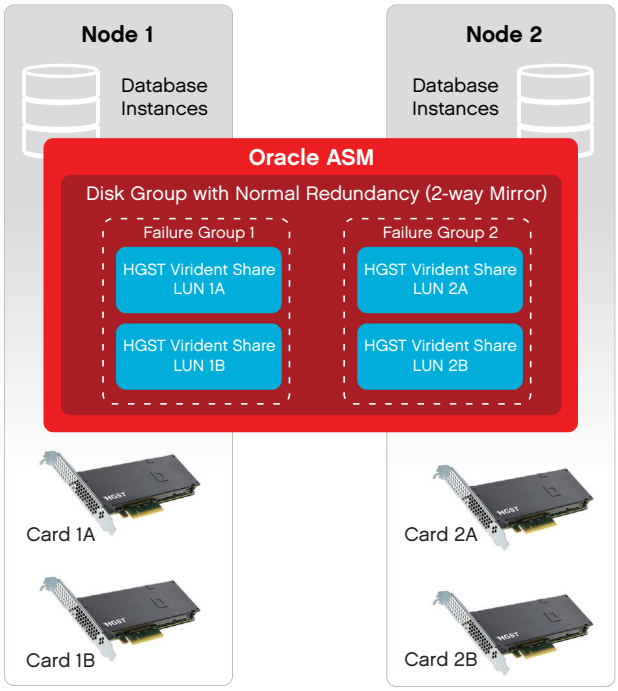


To ensure data safety in case of a whole node failure, all cards residing in the same node must be configured as belonging to the same "Failure Group" when creating an ASM disk group. Each node must have a "Failure Group" corresponding to it. ASM stores mirrored copies of the data in separate "Failure Groups", which means on separate nodes.

Additionally, FlashMAX II provides the following reliability and data protection capabilities on an individual PCIe flash card level:

- Embedded flash-aware RAID
- Global wear leveling
- Extended ECC
- End-to-end data protection with CRC
- Capacitors protecting write data

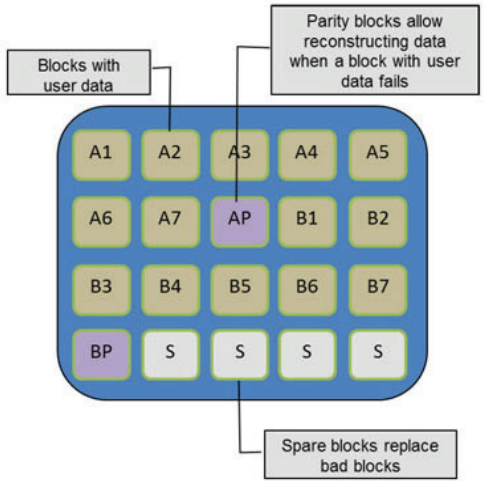
Failure group configuration with 2 nodes and 2 cards per node.



Embedded Flash-Aware RAID

Flash memory is vulnerable to errors that conventional ECC cannot correct, such as bad blocks or die failures. FlashMAX II protects its data residing on flash by using embedded flash-aware RAID that provides 7+1P data redundancy with striping across separate flash channels.

If reading any block of data fails, RAID reconstructs the block from data and parity stored in other channels. Bad blocks are retired and replaced with good blocks from reserved capacity, automatically restoring full redundancy. The flash-aware RAID protects user data against flash block failures, entire flash chip failures, and even multiple flash chip failures.



The flash-aware RAID is always enabled and does not require any configuration. The protection by the flash-aware RAID works completely transparently for the operating system without disrupting applications and without any human intervention. If the amount of flash errors exceeds safe thresholds the system administrator gets a warning and is able to arrange a proactive replacement of the device.

HGST Virident Share Software	HGST FlashMAX II PCIe SSDs	Share Interconnect	Operating Systems	InfiniBand Adapters	Oracle Versions
Version 1.2.X	550 GB Standard 1100 GB Standard 1100 GB Performance 2200 GB Standard	InfiniBand	Oracle Linux 6 with UEK R2 Oracle Linux 6 with UEK R3 Oracle Linux 6 with RH kernel Redhat Enterprise Linux 6	Mellanox ConnectX-2 Mellanox ConnectX-3	11.2 12.1
Version 2.0.X	550 GB Standard 1100 GB Standard 1100 GB Performance 2200 GB Standard 4800 GB Capacity	InfiniBand 10 GbE / 40 GbE	Oracle Linux 6 with UEK R3 Oracle Linux 6 with RH kernel Redhat Enterprise Linux 5/6	Mellanox ConnectX-2 Mellanox ConnectX-3 Mellanox Connect-IB	11.2 12.1

HGST
a Western Digital company