Open storage architecture for private Oracle database clouds

White Paper

rev. 2015-05-18

Abstract

Enterprise IT is transitioning from proprietary mainframe and UNIX systems to private and public clouds based on industry standard x86 servers and GNU/Linux operating system. Oracle Grid Infrastructure and Oracle RAC high-availability architectures make this transition possible even for mission-critical applications and database systems.

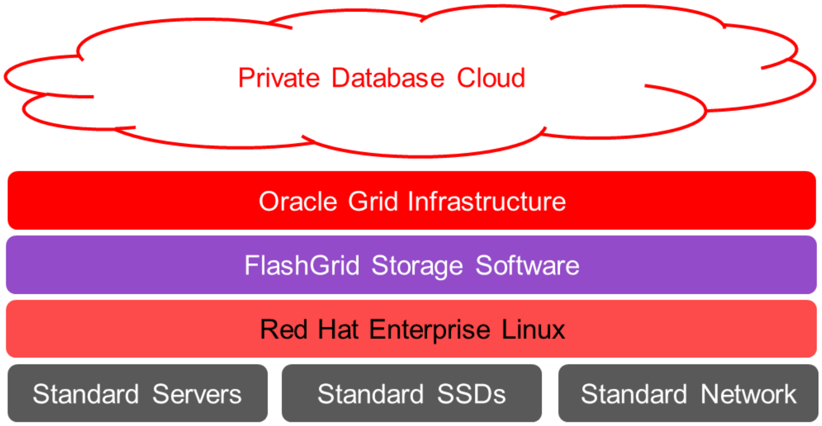
In contrast to the compute systems, transition of the storage systems from proprietary to industry standard hardware lags behind. Even the most technologically savvy enterprise SaaS and hosted application providers are still tied to proprietary storage systems for their database back-end. The dependency on the proprietary storage systems limits the scalability and agility required in the new cloud economy.

Arguably, the biggest challenge with the storage has been reliance on the spinning hard-drives. The low performance and high failure rates of the hard-drives created the need for large and complex storage systems housing hundreds of hard-drives along with complex caching, auto-tiering, QoS, and failure protection mechanisms. However, standardization and commoditization of PCIe NVMe SSDs simply eliminates the need for the complex proprietary storage systems.

In this white paper we introduce an open storage architecture that enables enterprises and cloud service providers to build scalable and flexible storage foundation for Oracle database clouds using commodity x86 servers and PCIe NVMe SSDs.

Architecture highlights

- Primary shared storage based on standard NVMe PCIe SSDs
- Physical storage located inside the database nodes (hyper-converged) or in separate storage nodes
- Standard x86 servers used as database and storage nodes
- Fully distributed architecture with no single point of failure
- FlashGrid software manages SSD devices and connectivity, integrates with Oracle ASM
- Oracle ASM (part of Oracle Grid Infrastructure) manages data, volumes, mirroring, snapshots
- Choice of 10/40/100 GbE or InfiniBand/RDMA for network connectivity
- Red Hat Enterprise Linux 7 provides enterprise-grade stability, security, and support at the OS level



NVMe SSDs

NVMe is an industry standard for PCIe-attached SSDs. The NVMe SSDs deliver outstanding performance of up to 5 GB/s and up to 850,000 random IOPS per SSD. Multiple NVMe SSDs can be installed per server, up to 48 SSDs in some server models. The hot-plug 2.5" disk form-factor makes handling SSDs as easy as handling regular hard-drives.

As of March 2016, NVMe SSDs are available in capacities ranging from 0.4 TB to 6.4 TB in the 2.5" form-factor. Capacities up to 10 TB are expected to become available by the end of 2016. The cost of enterprise-grade NVMe SSDs ranges between $1/GB and $3/GB, similar to enterprise SATA and SAS SSDs. As the SSD capacities increase the cost per GB is expected to decrease proportionally.

NVMe SSDs are available from all major server OEMs, including HPE, Dell, Lenovo, Cisco UCS, Supermicro, and also directly from the SSD vendors, including Intel, Samsung, Toshiba, HGST, and Seagate. The NVMe SSDs have been proven inside Oracle Exadata systems as the highest performance storage for Oracle databases.

Shared access

With the help of FlashGrid software each ASM instance can access each of the SSDs in the cluster. Each SSD is visible in the OS as `/dev/flashgrid/nodename.driveserialnumber` device where *nodename* is the name of the node where the SSD is physically located.

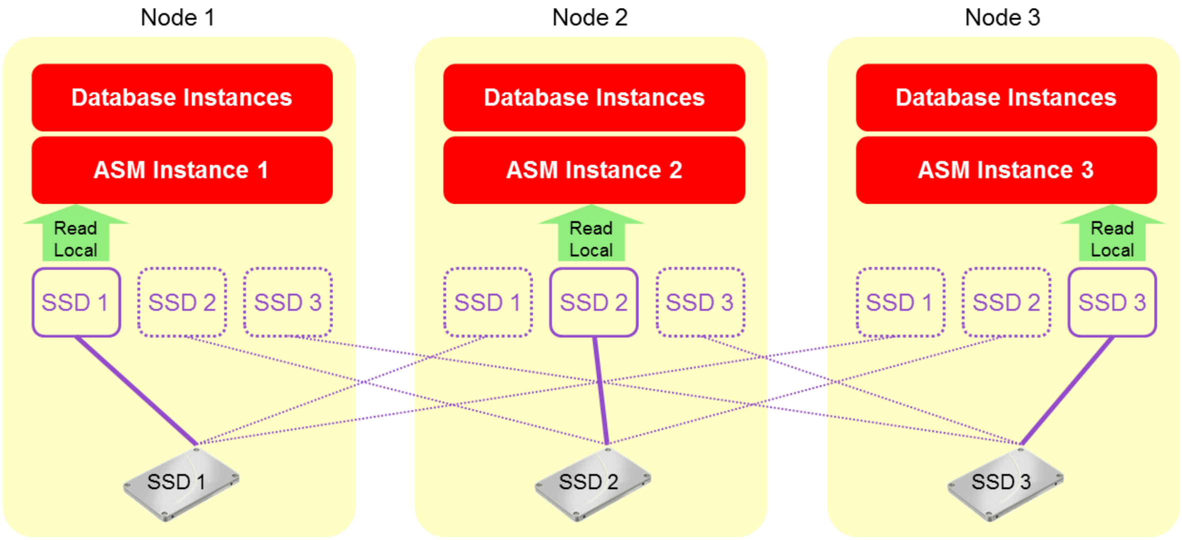


Figure 1. Shared access to storage in a 3-node hyper-converged cluster with one SSD per node

Data path

Data path reliability is critical for error-free operation and high availability. For data access and transfer the FlashGrid architecture leverages existing open-source components included in the Red Hat Enterprise Linux 7 operating system:

- NVMe device driver

- iSCSI/iSER target and initiator
- DM-Multipath driver

These data path components are developed and tested by an extensive industry ecosystem, including major storage vendors. FlashGrid software does not introduce any proprietary or new components in the data path. Instead, FlashGrid software automates configuration and management of the existing Red Hat Enterprise Linux 7 components to achieve maximum reliability and performance in Oracle RAC environments.

Data mirroring

For maximum performance and availability the proposed architecture leverages existing capabilities of Oracle ASM for mirroring data across nodes. No additional data management layers are introduced. In Normal Redundancy mode each block of data has two mirrored copies. In High Redundancy mode each block of data has three mirrored copies. Each ASM disk group is divided into failure groups – one failure group per node. Each disk is configured to be a part of a failure group that corresponds to the node where the disk is physically located. ASM ensures that mirrored copies of a block are placed in different failure groups.

Hyper-converged architecture or separate storage nodes

For each cluster the FlashGrid software provides flexibility to choose between hyper-converged architecture (storage inside the database nodes) or separate storage and database nodes.

For smaller database sizes the hyper-converged architecture typically is optimal. The FlashGrid Read-Local™ Technology minimizes network and CPU overhead by serving all read operations from local SSDs at the speed of the PCIe bus.

For larger database sizes separate database and storage nodes can be used. InfiniBand or 40/100 GbE fabric with RDMA provide up to 20 GB/s of bandwidth per node without consuming CPU on the database nodes. It is easy to add new storage servers when the capacity needs grow.

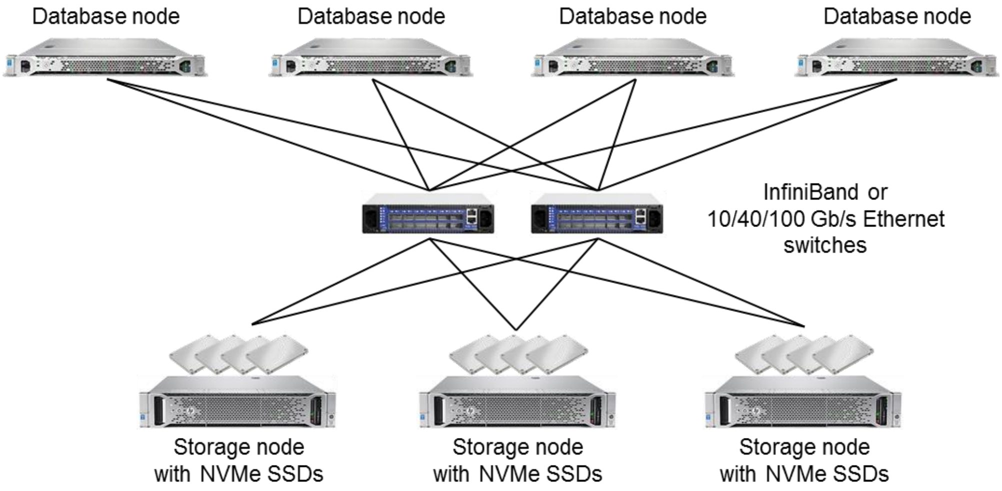


Figure 2. Example of a cluster topology with separate database and storage nodes

FlashGrid Read-Local™ Technology

In hyper-converged clusters the read traffic can be served from local SSDs at the speed of the PCIe bus instead of travelling over the network. In 2-node clusters with 2-way mirroring or 3-node clusters with 3-way mirroring 100% of the read traffic is served locally because each node has a full copy of all data. Because of the reduced network traffic, the write operations are faster too. As a result, even 10 GbE network fabric can be sufficient for achieving outstanding performance in such clusters for both data warehouse and OLTP workloads.

Switchless storage network with FlashGrid DirectFabric™ Technology

The FlashGrid DirectFabric Technology enables high-performance storage network connectivity between the cluster nodes without using the switches. This reduces the total cost of network hardware and also reduces the number of potential points of failure. The switchless configurations are most efficient in 2- or 3-node hyper-converged clusters. However, the switchless configurations can be used in other cluster topologies too.

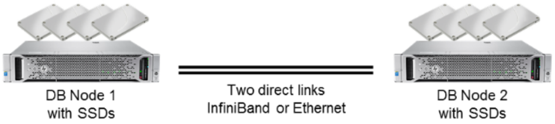


Figure 4. A 2-node cluster with switchless storage network

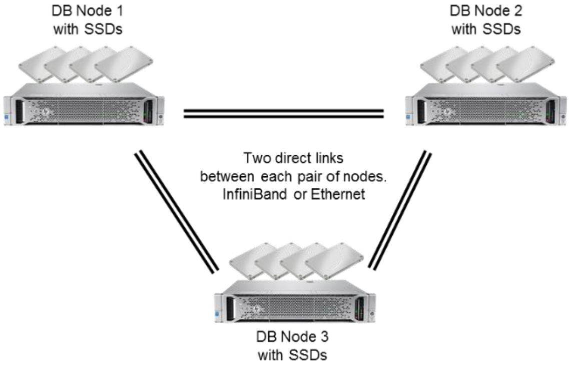


Figure 3. A 3-node cluster with switchless storage network

Data reduction with copy-on-write snapshots

Data de-duplication is a flagship feature of the new generation of proprietary all-flash arrays. However, unlike some non-database environments (e.g. VDI), de-duplication has no benefit with Oracle database files unless there are multiple copies of the same database. Using multiple copies of the same database is typical in Dev/Test environments. For the Dev/Test environments the copy-on-write snapshot capability of the Oracle ASM Cluster File System offers a more efficient and more predictable way of reducing the amount of data than the de-duplication algorithms. The copy-on-write snapshots prevent proliferation of duplicate data blocks in the first place instead of identifying duplicate blocks after they were already created.

Storage performance

Tested configuration

- Number of nodes: 2 and 3 hyper-converged nodes (database compute + storage)
- SSDs per server: 1, 2, or 4 of Intel SSD DC P3700 800GB
- Oracle Grid Infrastructure 12.1.0.2
- Oracle Database 12.1.0.2 RAC
- Database files on ASM
- FlashGrid software ver. 15.12
- Red Hat Enterprise Linux 7.2
- Servers: Dell PowerEdge R730xd
- CPU: Dual Intel Xeon E5-2667 v3, 8 cores @ 3.20GHz
- Network (per node): 2 x 10 GbE for storage, 2 x 10 GbE for RAC interconnect, 2 x 1 GbE public network

The database storage bandwidth and IOPS as reported by CALIBRATE_IO are shown on the two charts below. In all tests the latency was reported as zero, which means that the actual latency was lower than 1ms. Performance of EMC XtremIO Single X-Brick array (based on the vendor's specifications) is provided for comparison.

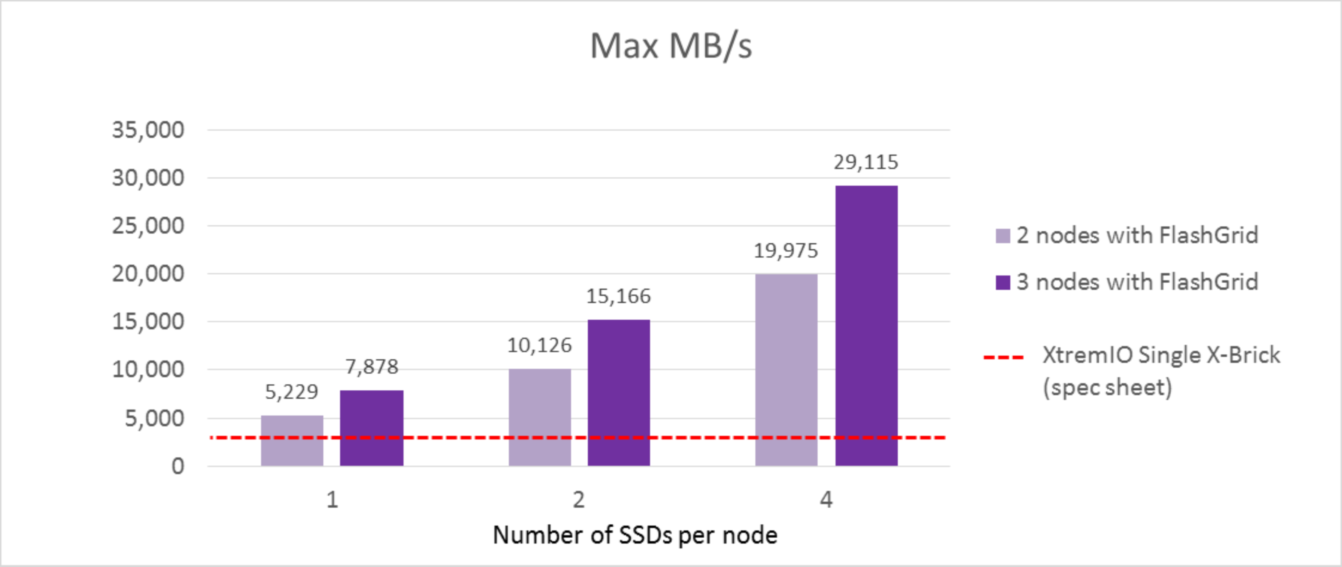


Figure 5. Maximum database storage bandwidth in the test clusters as reported by DBMS_RESOURCE_MANAGER.CALIBRATE_IO

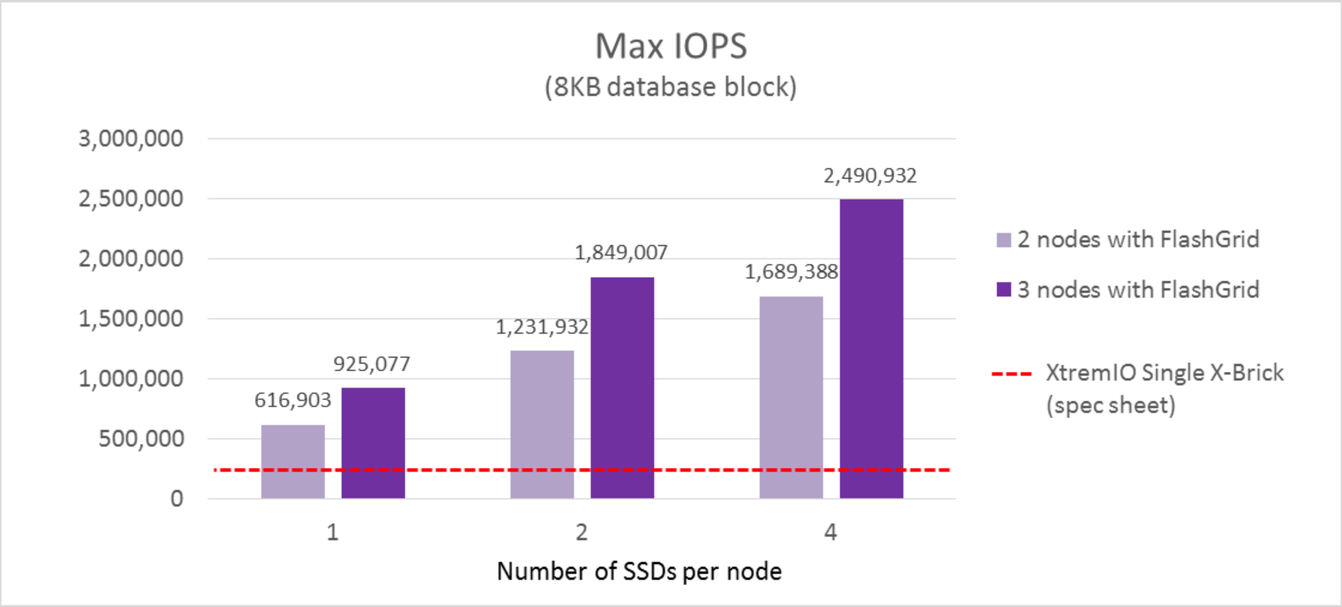


Figure 6. Maximum database storage performance in the test clusters as reported by DBMS_RESOURCE_MANAGER.CALIBRATE_IO

Server models with NVMe SSD support

NVMe SSDs in the add-in card (AIC) form-factor can be installed in most servers that have standard PCIe slots available. However, use of NVMe SSDs of 2.5" SFF form-factor (also known as U.2) is generally preferred because of better manageability. The newest models of rack-mounted servers from most server OEMs include 2.5" SFF NVMe backplane option. The table below provides the number of NVMe SSDs and the corresponding maximum NVMe flash capacities that can be configured in some popular server models.

	2.5" hot-plug NVMe SSDs			**Add-in PCIe card NVMe SSDs**		**Max total NVMe flash capacity per server**
Server model	**# slots**	**Max capacity per SSD**	**Max capacity per server with 2.5" NVMe SSDs**	**# PCIe slots available for NVMe SSDs**	**Max flash capacity per server with 6.4TB add-in card SSDs**	
Oracle Server X6-2L	9	3.2 TB	28.8 TB	5	32 TB	60.8 TB
Oracle Server X6-2	4	3.2 TB	12.8 TB	3	19.2 TB	32 TB
Dell PowerEdge R730xd	4	3.2 TB	12.8 TB	5	32 TB	44.8 TB
Dell PowerEdge R930	8	3.2 TB	25.6 TB	9	57.6 TB	83.2 TB
Dell PowerEdge R630	4	3.2 TB	12.8 TB	2	12.8 TB	25.6 TB
HPE ProLiant DL380 Gen9	6	1.6 TB	9.6 TB	5	32 TB	41.6 TB
HPE ProLiant DL560 Gen9	6	1.6 TB	9.6 TB	6	38.4 TB	48 TB
HPE ProLiant DL580 Gen9	5	1.6 TB	8 TB	8	51.2 TB	59.2 TB
Supermicro 1028U-TN10RT+	10	2 TB	20 TB	2	12.8 TB	32.8 TB
Supermicro 2028U-TN24R4T+	24	2 TB	48 TB	2	12.8 TB	60.8 TB
Supermicro 2028R-NR48N	48	2 TB	96 TB	2	12.8 TB	108.8 TB

Conclusion

With the broadening availability and outstanding performance of NVMe SSDs, enterprises and cloud service providers can implement storage for their database back-end using standard server and SSD hardware, eliminating the need for proprietary storage systems. Red Hat Enterprise Linux 7, FlashGrid software ver. 15.12, and Oracle Grid Infrastructure 12c together provide the software stack for building the open and scalable storage foundation for the most demanding Oracle database environments.

This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document.