Hyper-converged storage for Oracle RAC based on NVMe SSDs and standard x86 servers

White Paper

rev. 2016-05-18

Gold Partner

Abstract

Oracle Real Application Clusters (RAC) is an industry leading high-availability technology that allows using commodity server hardware for mission-critical databases. An important part of the Oracle RAC architecture is a shared storage that must be accessible concurrently by all database nodes and must also provide high performance and high availability of data. However, implementing such storage solution using industry standard hardware has been a challenge due to lack of comprehensive software for managing hyper-converged infrastructure. As a result, majority of Oracle RAC deployments has been done on proprietary storage array hardware, which increases complexity and costs of the Oracle RAC deployments.

In this white paper we discuss a new storage architecture for Oracle RAC that is based on standard x86 servers with Intel SSD DC P3700 under Oracle Linux 7. Management of the storage is done by FlashGrid software integrated with Oracle Automatic Storage Management (ASM). We also provide storage performance measurement results for 2- and 3- node Oracle RAC clusters and compare them to performance specifications of a popular flash storage array.

Introduction to FlashGrid Architecture

FlashGrid software turns standard servers with NVMe PCIe SSDs into high-performance distributed storage for Oracle RAC clusters and private database clouds.

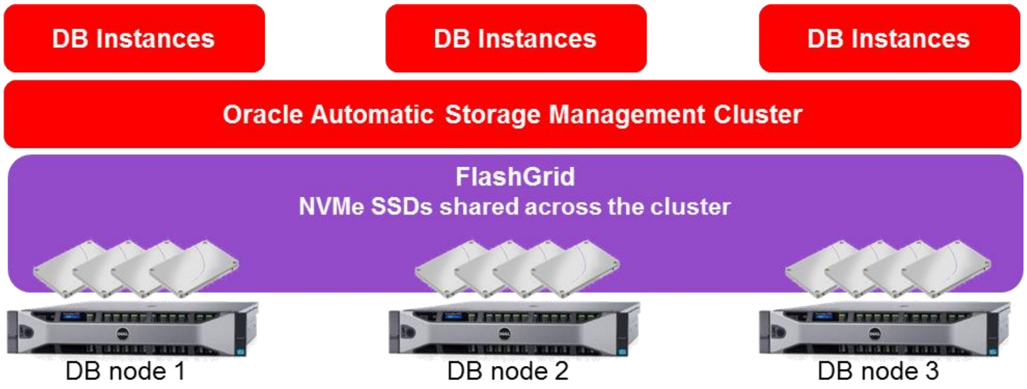

DB Instances
DB Instances
DB Instances
Oracle Automatic Storage Management Cluster
FlashGrid
NVMe SSDs shared across the cluster
DB node 1
DB node 2
DB node 3

Figure 1. Example of a 3-node Oracle RAC cluster with 4 SSDs in each node

FlashGrid Architecture Highlights

- Primary shared storage based on standard NVMe PCIe SSDs
- Optional SAS HDDs for capacity tier
- Physical storage located inside the database nodes (hyper-converged) or in separate storage nodes
- Standard x86 servers used as database and storage nodes
- FlashGrid software manages SSD devices and connectivity, integrates with Oracle ASM
- Oracle ASM manages data, volumes, mirroring, snapshots
- 2-way or 3-way mirroring of data across separate nodes
- Fully distributed architecture with no single point of failure
- Choice of 10/40/100 GbE or InfiniBand/RDMA for network connectivity

- FlashGrid Read-Local Technology minimizes network overhead by serving reads from local SSDs at the speed of PCIe

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 18 SSDs in some server models. The hot-plug 2.5" disk form-factor makes handling SSDs as easy as handling regular hard-drives. NVMe SSDs are available from major server OEMs including Oracle Sun servers and Exadata systems. For this paper we are testing and analyzing NVMe SSDs from Intel.

Hyper-converged node architecture

In this paper we focus on an architecture using hyper-converged nodes with the physical storage located inside the database nodes. In most environments the hyper-converged configuration is optimal. However, separate storage nodes can also be used, for example, when there is not enough room inside the database servers for the required storage capacity.

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.drivename` device where *nodename* is the name of the node where the SSD is physically located.

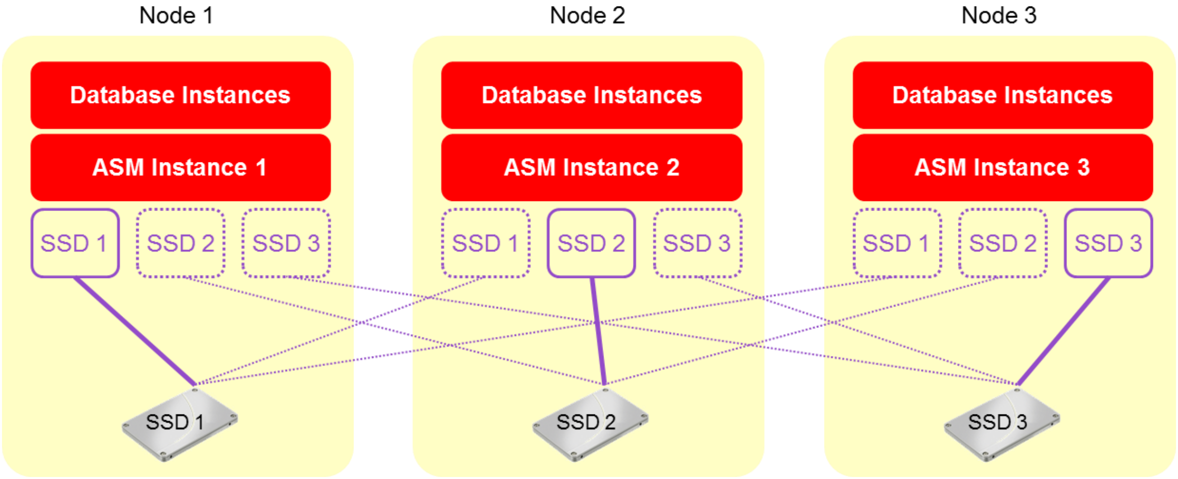


Figure 2. Diagram of shared access to storage in a 3-node cluster with one SSD per node

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 Oracle 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 components to achieve maximum reliability and performance in Oracle RAC environments.

Data mirroring

The FlashGrid architecture leverages capabilities of Oracle ASM for mirroring data. 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.

In Normal Redundancy mode the cluster can withstand loss of one node without interruption of service. In High Redundancy mode the cluster can withstand loss of two nodes without interruption of service.

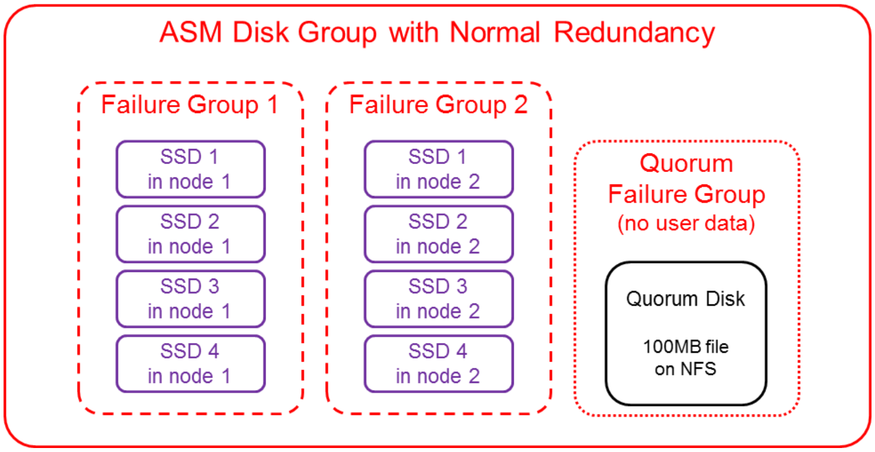


Figure 3. Example of an ASM disk group configuration in a 2-node cluster with 4 SSDs per node

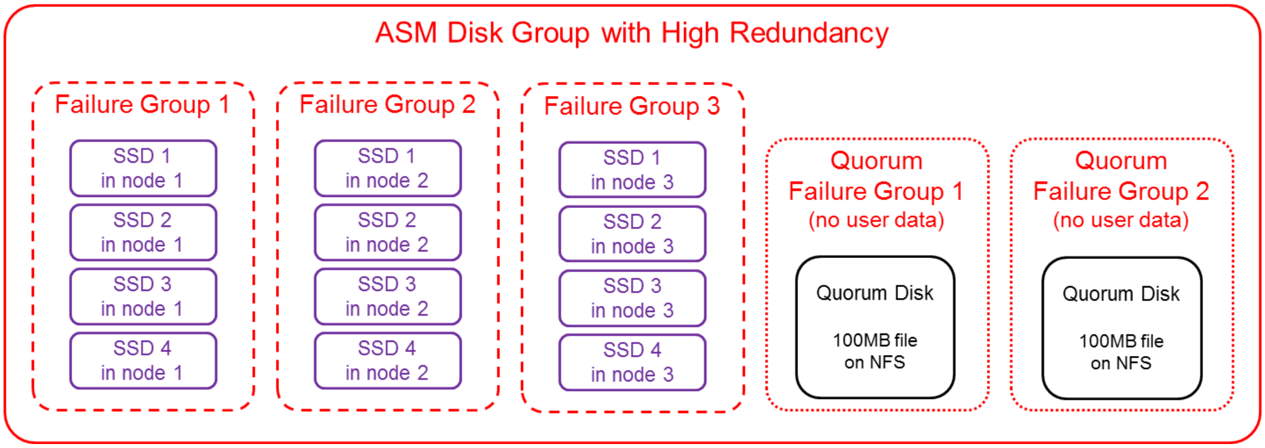


Figure 4. Example of an ASM disk group configuration in a 3-node cluster with 4 SSDs per node

Minimizing exposure to double failures

Configuring one extra SSD per node allows implementing "hot spare" functionality that minimizes exposure to double failures even when it is impossible to quickly replace a failed SSD. For example, for a 3 TB data set we would configure four 1 TB SSDs instead of three. The extra SSD is not a "dedicated" spare drive. All four SSDs work in parallel. But if one of the four SSDs fails then the remaining three SSDs have sufficient capacity to accommodate all data. If it is impossible to replace the failed SSD immediately then we can start a rebalancing operation that will restore the full redundancy by copying the missing blocks of data (that were on the failed SSD) from the other node. The rebalancing operation can be started manually. But it can also start automatically after a period of time configured in the *disk_repair_time* parameter of the disk group.

Quorum disks

In certain disk group configurations one or two additional quorum disks may be required depending on the number of nodes in the cluster. ASM uses quorum disks to store additional copies of metadata that can be used in some failure scenarios. One quorum disk requires less than 100 MB of space. Quorum disks generate little storage traffic and can reside on any type of external shared storage including FC/FCoE/iSCSI SAN or NFS servers.

In the test setups used for this paper we configured one quorum disk for the 2-node cluster and two quorum disks for the 3-node cluster. In both cases the quorum disks were located on external NFS storage.

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.

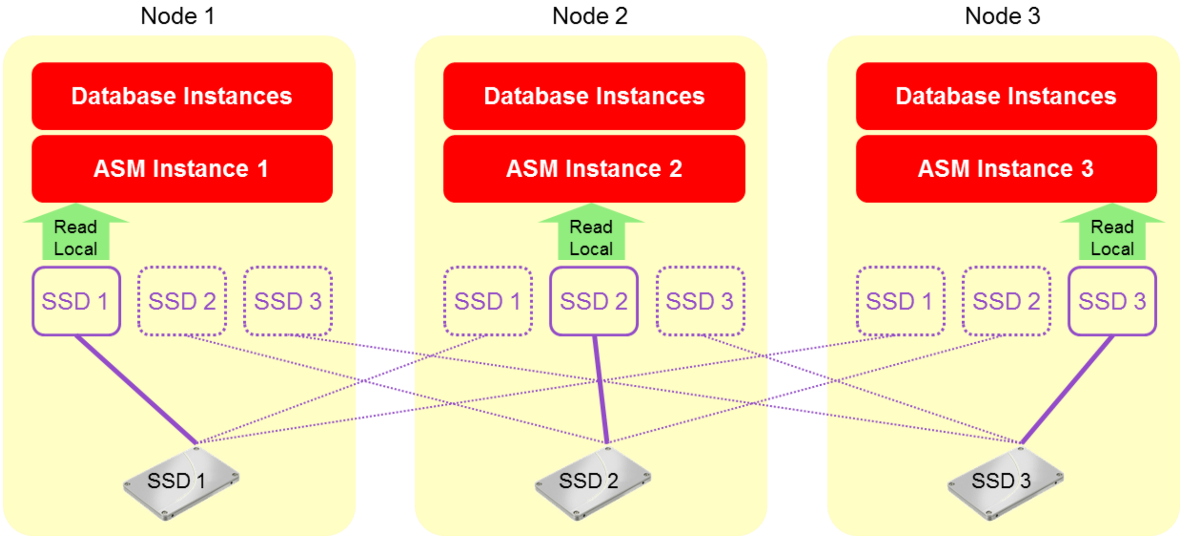
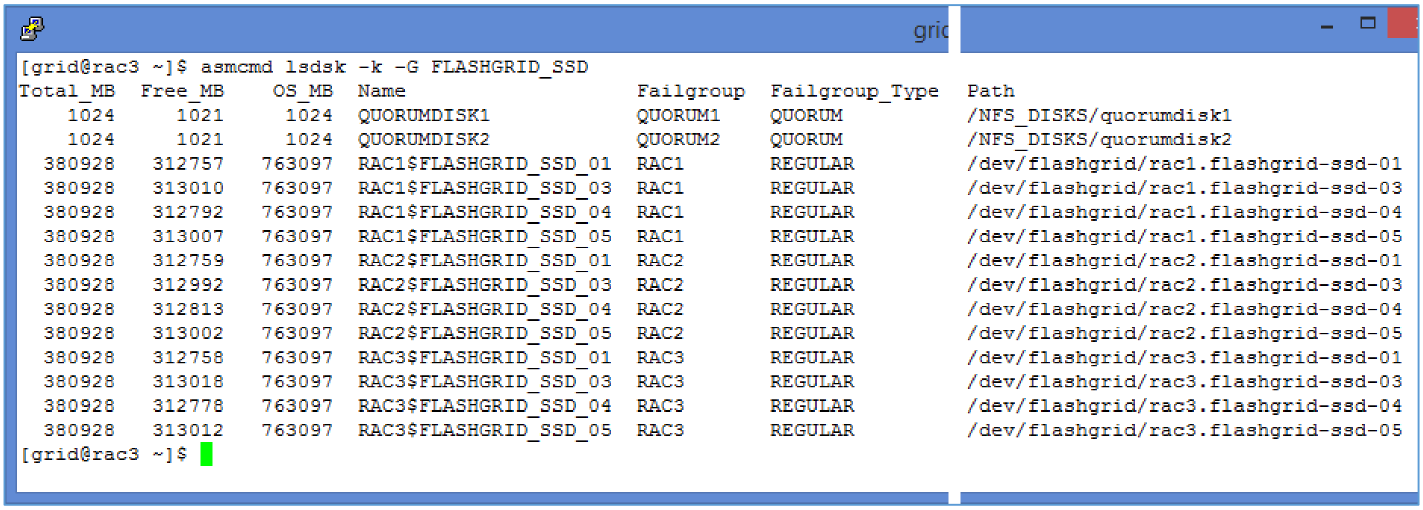


Figure 5. FlashGrid Read-Local Technology in a 3-node cluster with 3-way mirroring

Tested Configuration Details

- 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. 15.9
- Oracle Linux 7.1
- 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



```
[grid@rac3 ~]$ asmcmd lsdsk -k -G FLASHGRID_SSD
Total_MB  Free_MB   OS_MB  Name                   Failgroup  Failgroup_Type  Path
    1024     1021    1024  QUORUMDISK1            QUORUM1    QUORUM          /NFS_DISKS/quorumdisk1
    1024     1021    1024  QUORUMDISK2            QUORUM2    QUORUM          /NFS_DISKS/quorumdisk2
  380928   312757  763097  RAC1$FLASHGRID_SSD_01  RAC1       REGULAR         /dev/flashgrid/rac1.flashgrid-ssd-01
  380928   313010  763097  RAC1$FLASHGRID_SSD_03  RAC1       REGULAR         /dev/flashgrid/rac1.flashgrid-ssd-03
  380928   312792  763097  RAC1$FLASHGRID_SSD_04  RAC1       REGULAR         /dev/flashgrid/rac1.flashgrid-ssd-04
  380928   313007  763097  RAC1$FLASHGRID_SSD_05  RAC1       REGULAR         /dev/flashgrid/rac1.flashgrid-ssd-05
  380928   312759  763097  RAC2$FLASHGRID_SSD_01  RAC2       REGULAR         /dev/flashgrid/rac2.flashgrid-ssd-01
  380928   312992  763097  RAC2$FLASHGRID_SSD_03  RAC2       REGULAR         /dev/flashgrid/rac2.flashgrid-ssd-03
  380928   312813  763097  RAC2$FLASHGRID_SSD_04  RAC2       REGULAR         /dev/flashgrid/rac2.flashgrid-ssd-04
  380928   313002  763097  RAC2$FLASHGRID_SSD_05  RAC2       REGULAR         /dev/flashgrid/rac2.flashgrid-ssd-05
  380928   312758  763097  RAC3$FLASHGRID_SSD_01  RAC3       REGULAR         /dev/flashgrid/rac3.flashgrid-ssd-01
  380928   313018  763097  RAC3$FLASHGRID_SSD_03  RAC3       REGULAR         /dev/flashgrid/rac3.flashgrid-ssd-03
  380928   312778  763097  RAC3$FLASHGRID_SSD_04  RAC3       REGULAR         /dev/flashgrid/rac3.flashgrid-ssd-04
  380928   313012  763097  RAC3$FLASHGRID_SSD_05  RAC3       REGULAR         /dev/flashgrid/rac3.flashgrid-ssd-05
[grid@rac3 ~]$
```

Figure 6. Configuration of the ASM disk group containing database files in a 3-node cluster with 4 SSDs per node

Measuring Performance

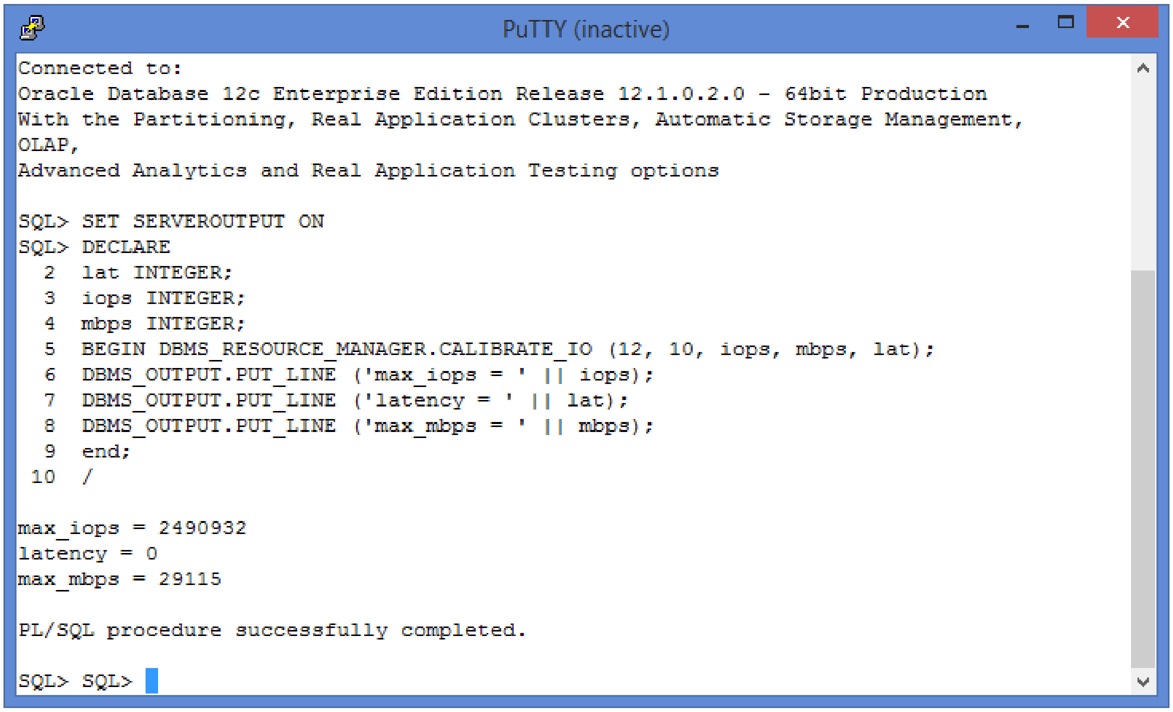
DBMS_RESOURCE_MANAGER.CALIBRATE_IO procedure provides an easy way for measuring storage performance including maximum bandwidth, random IOPS, and latency. The CALIBRATE_IO procedure generates I/O through the database stack on actual database files. The test is read-only and it is safe to run it on any existing database. CALIBRATE_IO results do not depend on any non-storage factors, such as memory size or the number of CPU cores. For a given storage system, the CALIBRATE_IO results depend only on three parameters:

- Database block size parameter (typically 8KB is used)
- Number of disks (should be configured according to the actual number of physical HDDs/SSDs used)
- Maximum allowed latency (the minimum value of 10ms recommended for high-performance systems)

As a result, the CALIBRATE_IO results are easy to replicate and verify. This makes CALIBRATE_IO a good tool for directly comparing performance of different storage systems.

Below is a SQL code we used for running CALIBRATE_IO on the test database. The highlighted parameter was set equal to the number of SSDs configured in each test.

```
SET SERVEROUTPUT ON
DECLARE
lat INTEGER;
iops INTEGER;
mbps INTEGER;
BEGIN DBMS_RESOURCE_MANAGER.CALIBRATE_IO (12, 10, iops, mbps, lat);
DBMS_OUTPUT.PUT_LINE ('max_iops = ' || iops);
DBMS_OUTPUT.PUT_LINE ('latency = ' || lat);
DBMS_OUTPUT.PUT_LINE ('max_mbps = ' || mbps);
end;
/
```



```
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit Production
With the Partitioning, Real Application Clusters, Automatic Storage Management,
OLAP,
Advanced Analytics and Real Application Testing options

SQL> SET SERVEROUTPUT ON
SQL> DECLARE
  2  lat INTEGER;
  3  iops INTEGER;
  4  mbps INTEGER;
  5  BEGIN DBMS_RESOURCE_MANAGER.CALIBRATE_IO (12, 10, iops, mbps, lat);
  6  DBMS_OUTPUT.PUT_LINE ('max_iops = ' || iops);
  7  DBMS_OUTPUT.PUT_LINE ('latency = ' || lat);
  8  DBMS_OUTPUT.PUT_LINE ('max_mbps = ' || mbps);
  9  end;
 10  /

max_iops = 2490932
latency = 0
max_mbps = 29115

PL/SQL procedure successfully completed.

SQL> SQL>
```

Figure 7. Example of running CALIBRATE_IO on a 3-node cluster with four NVMe SSDs per node

Performance Results

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.

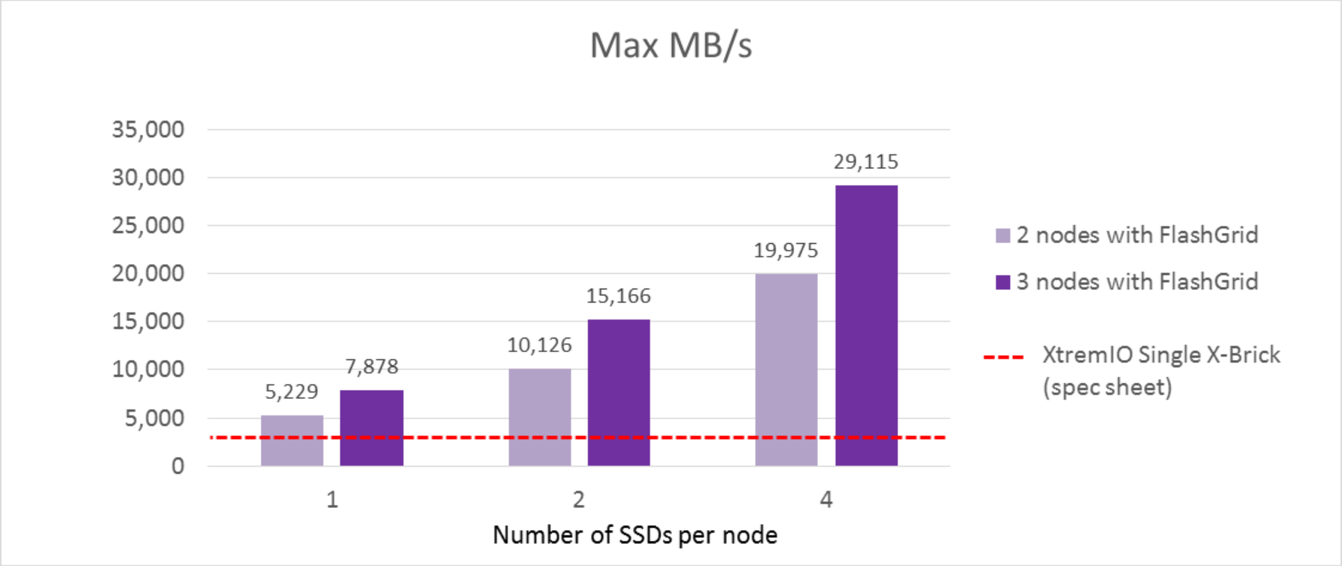


Figure 8. Maximum database storage bandwidth in the test clusters as reported by DBMS_RESOURCE_MANAGER.CALIBRATE_IO

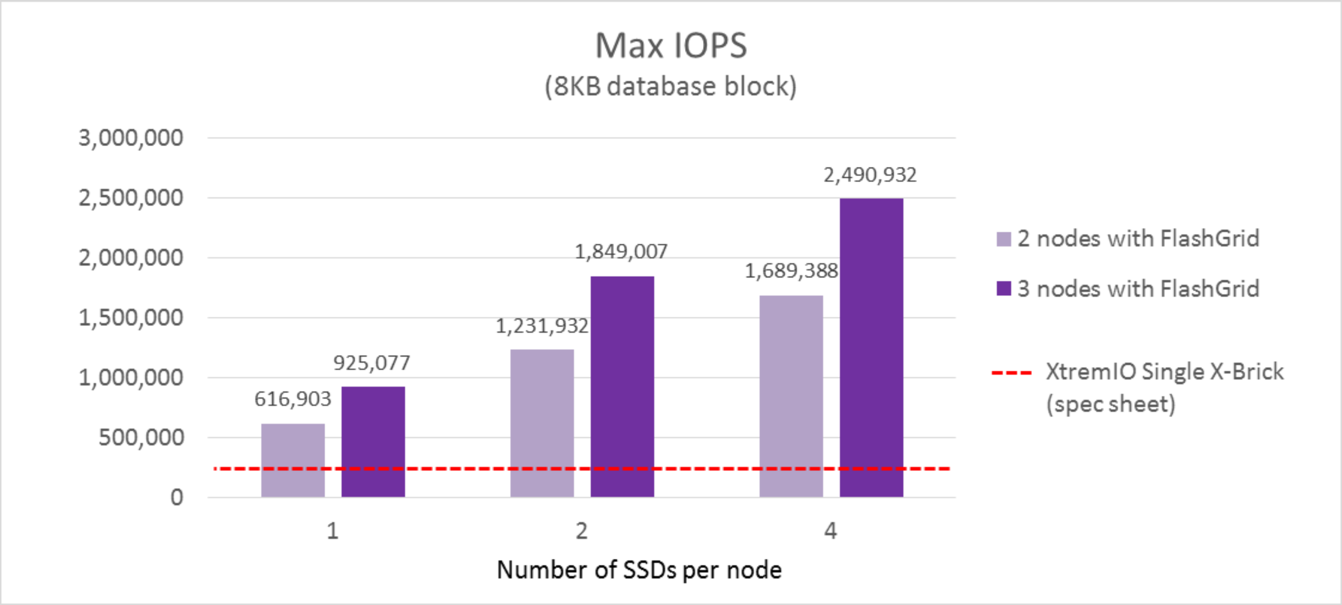


Figure 9. Maximum database storage performance in the test clusters as reported by DBMS_RESOURCE_MANAGER.CALIBRATE_IO

Capacity and Cost

The FlashGrid architecture provides flexibility for building very low-cost entry-level storage or high-capacity high-performance mission-critical storage. The smallest storage configuration with two RAC nodes is 400 GB of usable capacity. It requires only two NVMe SSDs with the total cost of approximately $1050, which makes it affordable even for a home lab. For large data set sizes up to 36 TB of flash can be configured per node.

The following table provides an overview of the enterprise NVMe SSDs available from Intel.

Intel SSD model	**P3500**	**P3600**	**P3608****	**P3700**
Typical use-case	Read-intensive databases or dev/test	Mixed use or business-critical databases	Large mixed use or business-critical databases	Write-intensive OLTP or mission-critical databases
Form-factors	add-in HHHL card, 2.5” hot-plug	add-in HHHL card, 2.5” hot-plug	add-in HHHL card	add-in HHHL card, 2.5” hot-plug
Capacities	400 GB, 800 GB, 1.6 TB, 2.0 TB	400 GB, 800 GB, 1.2 TB, 1.6 TB, 2.0 TB	1.6 TB, 3.2 TB, 4.0 TB	400 GB, 800 GB, 1.6 TB, 2.0 TB
Appr. $ / GB *	$1.2 / GB	$1.4 /GB	$2.1 /GB	$2.1 /GB

* Based on retail prices reported as of 11/14/2015 at http://www3.intel.com/buy/us/en/catalog/components/ssd/sort,onmarket-d

** With PCIe 3.0 x8 interface the P3608 provides higher read/write bandwidth of up to 5000/3000 MB/s.

The number of NVMe SSDs that can be configured per node depends on the server model. The table below provides the number of NVMe SSDs and the corresponding capacities that can be configured in some popular server models.

	2.5” hot-plug NVMe SSDs		**Add-in PCIe card NVMe SSDs**		**Max total flash capacity per server**
Server model	**# slots**	**Max flash capacity per server with 2TB 2.5” SSDs**	**# PCIe slots available for NVMe SSDs**	**Max flash capacity per server with 4TB add-in card SSDs**	
Oracle Server X6-2L	9	18 TB	5	20 TB	38 TB
Oracle Server X6-2	4	8 TB	3	12 TB	20 TB
Dell PowerEdge R730xd	4	8 TB	5	20 TB	28 TB
Dell PowerEdge R930	8	16 TB	9	36 TB	52 TB
Dell PowerEdge R630	4	8 TB	2	8TB	16 TB
HPE ProLiant DL380 Gen9	6	12 TB	5	20 TB	32 TB
HPE ProLiant DL560 Gen9	6	12 TB	6	24 TB	36 TB
HPE ProLiant DL580 Gen9	5	10 TB	8	32 TB	42 TB
Supermicro 1028U-TN10RT+	10	20 TB	2	8 TB	28 TB
Supermicro 2028U-TN24R4T+	24	48 TB	2	8 TB	56 TB
Supermicro 2028R-NR48N	48	96 TB	2	8 TB	104 TB

FlashGrid Architecture Compared To a Flash Storage Array

	Intel SSD DC P3700 2TB 4 SSDs per node 2 nodes	Intel SSD DC P3700 2TB 4 SSDs per node 3 nodes	EMC XtremIO 4.0 Single 10TB X-Brick flash storage array **
Storage performance, IOPS	1,700,000	2,500,000	250,000
Storage bandwidth, GB/s	20	29	3
Usable Storage Capacity, TB	8	8	8.33
Rack space	none, inside DB servers	none, inside DB servers	6U
Main data protection mechanism	2-way mirroring across nodes by Oracle ASM	3-way mirroring across nodes by Oracle ASM	Proprietary
Storage HW cost (MSRP)	$34,360 *	$51,540 *	$350,000 ***

* Based on $4,295 retail price of one Intel SSD DC P3700 2TB: http://www3.intel.com/buy/us/en/catalog/search/?q=P3700

** http://xtremio.com/performance , https://www.emc.com/collateral/data-sheet/h12451-xtremio-4-system-specifications-ss.pdf

*** http://www.emc.com/sales/stateoffl/florida-price-list-2014-05.pdf

Conclusion

Oracle RAC can be implemented using standard hardware for both compute and storage. We have demonstrated a hyper-converged compute/storage solution built on standard x86 servers, Intel SSD DC P3700, Oracle Linux 7, and FlashGrid software that leverages data management and high-availability capabilities of Oracle ASM. The solution does not require purchasing a proprietary storage array. The solution allows achieving 10x (or higher) performance compared to a popular flash storage array at approximately 1/6 of the cost.

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.

FlashGrid is a trademark of FlashGrid Inc. Oracle is a registered trademarks of Oracle Corporation. Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. Other names may be trademarks of their respective owners.