WHITE PAPER

FlashGrid® Software
Intel® SSD DC P3700/P3600/P3500
Topic: Hyper-converged Database/Storage

FlashGrid® Software Enables Converged and Hyper-Converged Appliances for Oracle* RAC

Table of Contents

Abstract

FlashGrid® software along with Intel® SSD DC for PCIe* family of drives enables dedicated, high-performance storage for Oracle* Real Application Clusters* (RAC), either inside the database servers or in separate storage servers. FlashGrid-based storage is dedicated to a particular cluster, and it guarantees consistently high performance without interference from other clusters or applications.

Oracle RAC is an industry-leading high-availability technology that allows using Intel® Xeon® processor-based servers for mission-critical databases. Oracle RAC relies on a high-performance, high-availability, shared storage resource that must be concurrently accessible by all database nodes. Historically, most Oracle RAC deployments have used external storage arrays that are shared among multiple clusters or even with other non-database applications. However, complexity, performance bottlenecks, and "noisy neighbors" create serious performance, reliability, and manageability challenges in operation of the mission-critical Oracle databases.

Integrating Intel SSD DC for PCIe family drives with FlashGrid software into a database appliance enables a high-performance, low-latency storage dedicated to one Oracle RAC cluster. This eliminates the need for an external storage array in Oracle RAC installations, silences "noisy neighbors," and gives Oracle database administrators (DBAs) full control of the database storage.

In this white paper we discuss the FlashGrid storage architecture and review storage performance test results for two- and three-node hyper-converged Oracle RAC.

Solving Storage Problems with Database Appliances

Oracle RAC offers a unique capability of running mission-critical databases on industry-standard x86-based server hardware, for example, powered by Intel® Xeon® processor E5 and E7 families.

Traditionally, Oracle RAC installations have been tied to proprietary storage arrays. The storage arrays add complexity and cost, but more importantly, they limit database performance and expose databases to "noisy neighbors." Even the all-flash arrays do not fully solve the storage I/O performance problem because of the bottlenecks in the storage network and in the array controllers. A storage array is typically shared between multiple clusters, or even with other applications, which makes the storage performance bottlenecks even more challenging. One application running an I/O intensive operation can significantly slow down I/O for multiple other databases. But, this is changing today.

IT infrastructure departments, solution providers, and server OEMs can solve the storage challenges for their Oracle RAC customers with high-performance converged or hyper-converged database appliances based on standard servers, high-density NVMe SSDs, and FlashGrid software with seamless Oracle Automatic Storage Management integration.

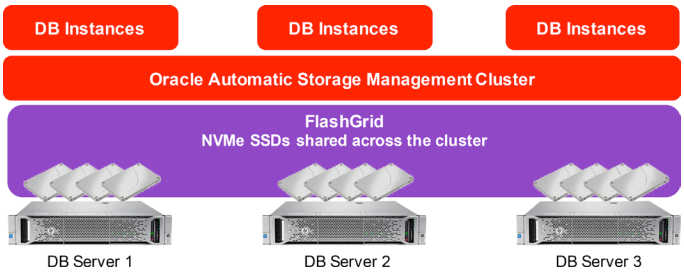


Figure 1. Example of a 3-node Oracle RAC with 4 SSDs in each node.

Introduction to FlashGrid Architecture

FlashGrid software turns standard servers with PCIe-based SSDs, such as the Intel SSD DC P3700, P3600, and P3500 drives, into high-performance, distributed storage for Oracle RAC and private database clouds (Figure 1). The software leverages proven Oracle* Automatic Storage Manager (Oracle ASM) capabilities for volume management, data mirroring, and high availability, while enabling outstanding performance and scalability. With FlashGrid, managing storage in Oracle RAC is an easy task for DBAs and server administrators.

FlashGrid architecture highlights

- Turns standard x86 servers into storage nodes or hyper-converged database/storage nodes.
- Leverages performance of PCIe-based SSDs.
- Manages SSD devices and connectivity, while integrating with Oracle ASM.
- Creates a fully distributed storage with no single point of failure.
- Supports 10/40/100 GbE or InfiniBand*/RDMA for network connectivity.
- 0.4 TB to 104 TB of flash storage per node.

NVMe SSDs – The New Standard of Performance

NVMe is an industry standard for PCIe-attached SSDs. The highly optimized NVMe driver stack minimizes CPU cycles per I/O and delivers high IOPS and low latency. Available 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, with 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. Intel SSD DC for PCIe family of drives offer proven enterprise-grade reliability and are available from major server OEMs.

Hyper-Converged Architecture Simplifies Cluster Configuration

FlashGrid software enables hyper-converged nodes in Oracle RAC, integrating the physical storage inside the database nodes. In most environments, the hyper-converged configuration is optimal. However, FlashGrid supports separate storage nodes for configurations that do not have enough room inside the database servers for the required storage capacity.

Enabling Shared Storage Access Across the Cluster

FlashGrid software works seamlessly with Oracle ASM. With the help of FlashGrid software, each Oracle ASM instance can access any of the SSDs in the cluster. Each SSD is visible in the OS as /dev/flashgrid/nodename.drivename, where nodename is the name of the node where the SSD is physically located (Figure 2).

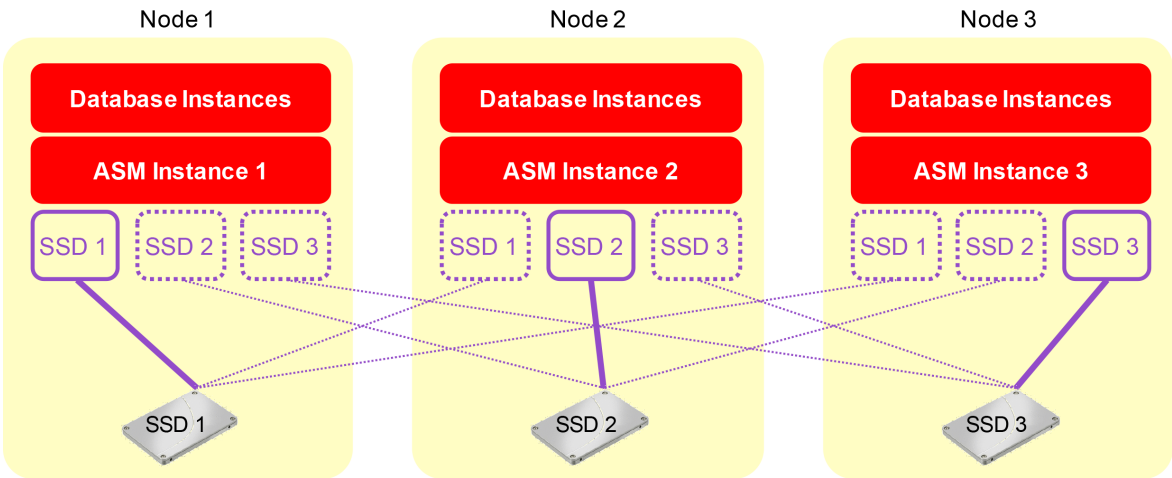


Figure 2. Diagram of shared access to storage in a 3-node cluster with one SSD per node.

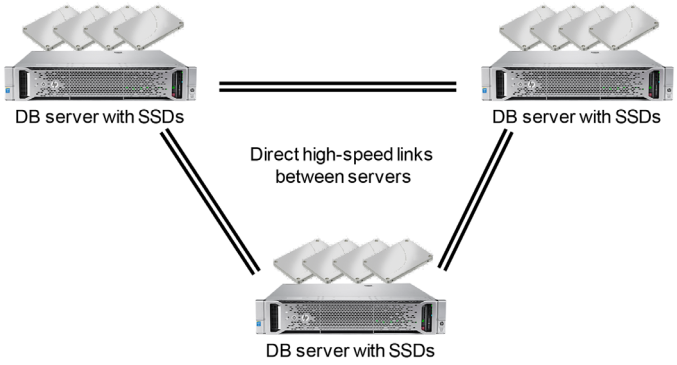


Figure 3. Example of a 3-node hyper-converged cluster topology without network switches.

FlashGrid Direct-Fabric Technology™ Creates Switchless High-Speed Storage Network

In simple two- or three-node clusters, the FlashGrid Direct-Fabric™ Technology allows using direct back-to-back links for storage traffic without a network switch. 100 Gbps storage network speeds can be realized with dual-port network adapters or two network adapters without adding the cost of a switch (Figure 3).

Maintaining a Consistent, Reliable 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 Linux* operating system:

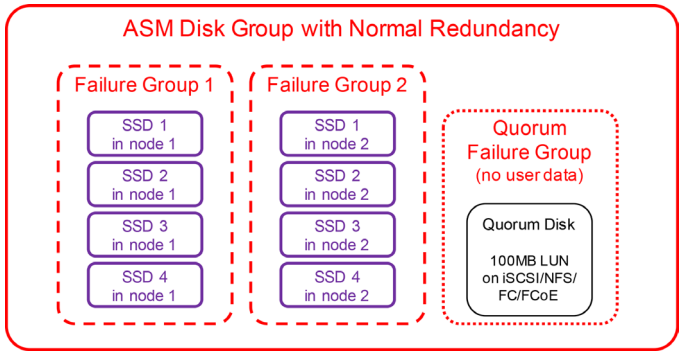
- NVMe device driver
- iSCSI/iSER target and initiator
- DM-Multipath driver

These data path components are developed and tested by an extensive industry ecosystem. FlashGrid software does not introduce any proprietary or new components in the data path to alter the already reliable and consistent behavior in the open source code. Instead, FlashGrid software automates configuration and management of the existing components to achieve maximum reliability and performance in Oracle RAC environments.

FlashGrid Leverages Oracle ASM for High Availability

The FlashGrid architecture leverages capabilities of Oracle ASM for mirroring data. In Oracle ASM's 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. Oracle ASM ensures that mirrored copies of a block are placed in different failure groups (Figure 4).

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.



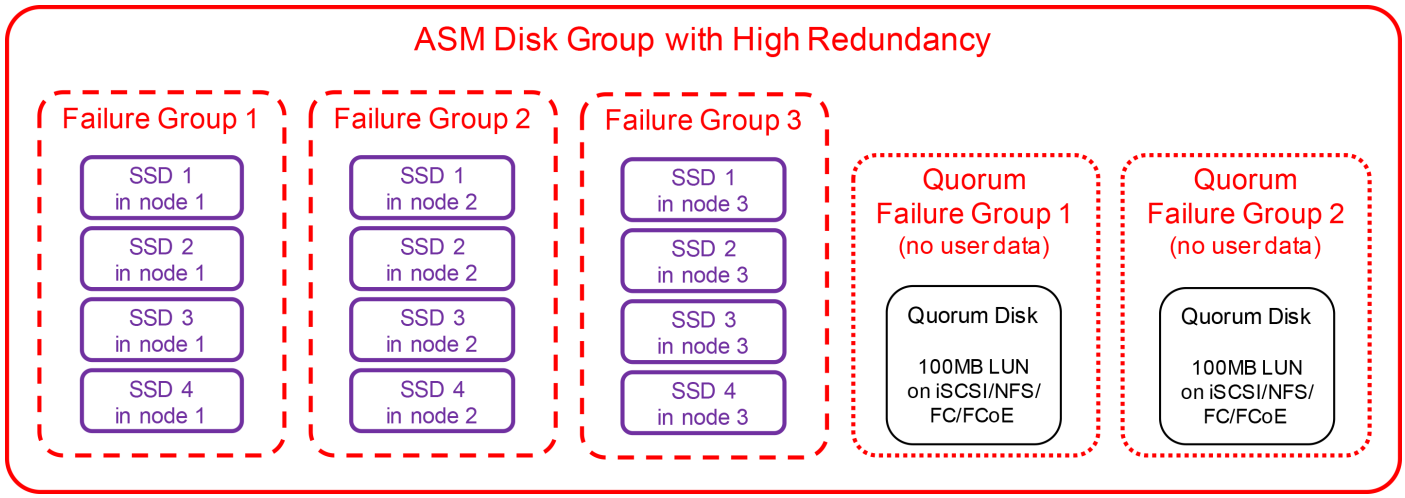


Figure 4. Example of an ASM disk group configuration in a 2-node cluster (above) and 3-node cluster (below) with 4 SSDs per node.

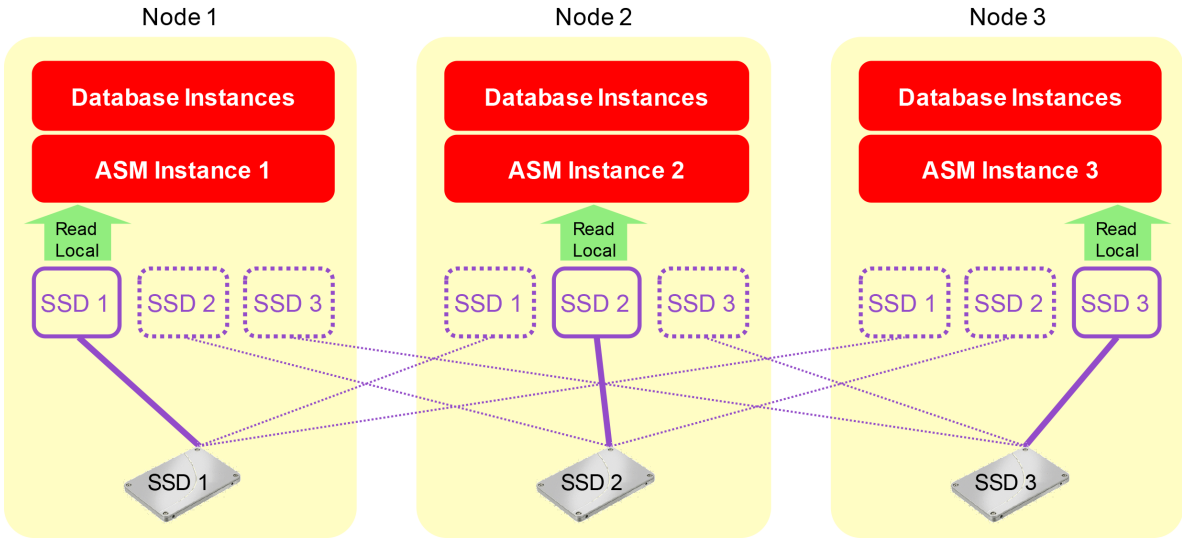


Figure 5. FlashGrid Read-Local Technology in a 3-node cluster with 3-way mirroring.

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, four 1 TB SSDs would be configured instead of three. The extra SSD is not a "dedicated" spare drive; all four SSDs work in parallel. 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 a resync operation can be started that will restore the full redundancy by copying the missing blocks of data (that were on the failed SSD) from other nodes. The resync operation can be started manually, but it can also be configured to start automatically after a period of time.

Using Quorum Disks Increases Reliability

In certain disk group configurations one or two additional quorum disks may be required depending on the number of nodes in the cluster. Oracle 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 two-node cluster and two quorum disks for the three-node cluster. In both cases the quorum disks were located on external NFS storage.

FlashGrid Read-Local™ Technology Accelerates Read and Write Operations

With Oracle ASM, each node has a full copy of all data. In two-node or three-node hyper-converged clusters, 100% of the read traffic is served from local SSDs at the speed of the PCIe bus (Figure 5) instead of travelling over the network at slower wire speeds. This reduces read-side network traffic, and, thus, the write operations are faster. As a result, even a 10 GbE network fabric can be sufficient for achieving outstanding performance in such clusters for both data warehouse and OLTP workloads.

Performance Testing Oracle RAC with FlashGrid Software

To assess the performance and price-performance of FlashGrid software on an Oracle RAC, testing was carried out using the following configuration:

- Number of nodes: two and three hyper-converged nodes (database compute + storage)
- SSDs per server: 1, 2, or 4 of Intel SSD DC P3700 800 GB
- 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® processor E5-2667 v3, 8 cores @ 3.20 GHz
- Network (per node): 2x 10 Gbps Intel® Ethernet controller for storage, 2x 10 Gbps Intel Ethernet controller for RAC interconnect, 2x 1 Gbps Intel Ethernet controller for public network

Figure 6 provides a listing of the disks in the cluster.

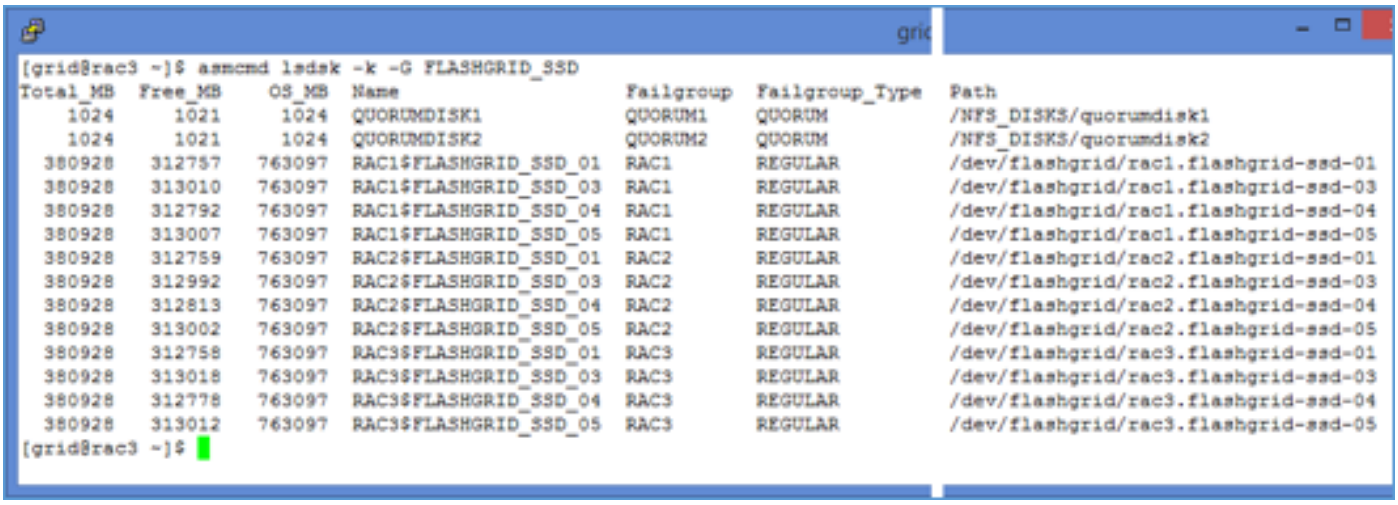


Figure 6. Configuration of the ASM disk group containing database files in a 3-node cluster with 4 SSDs per node.

Test Setup

The DBMS_RESOURCE_MANAGER.CALIBRATE_IO (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; 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 on only three parameters:

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

The following is the SQL code that was used for running CALIBRATE_IO on the test database. The underlined value is the number of SSDs configured in each test. Figure 7 shows the terminal interaction.

```
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;
/
```

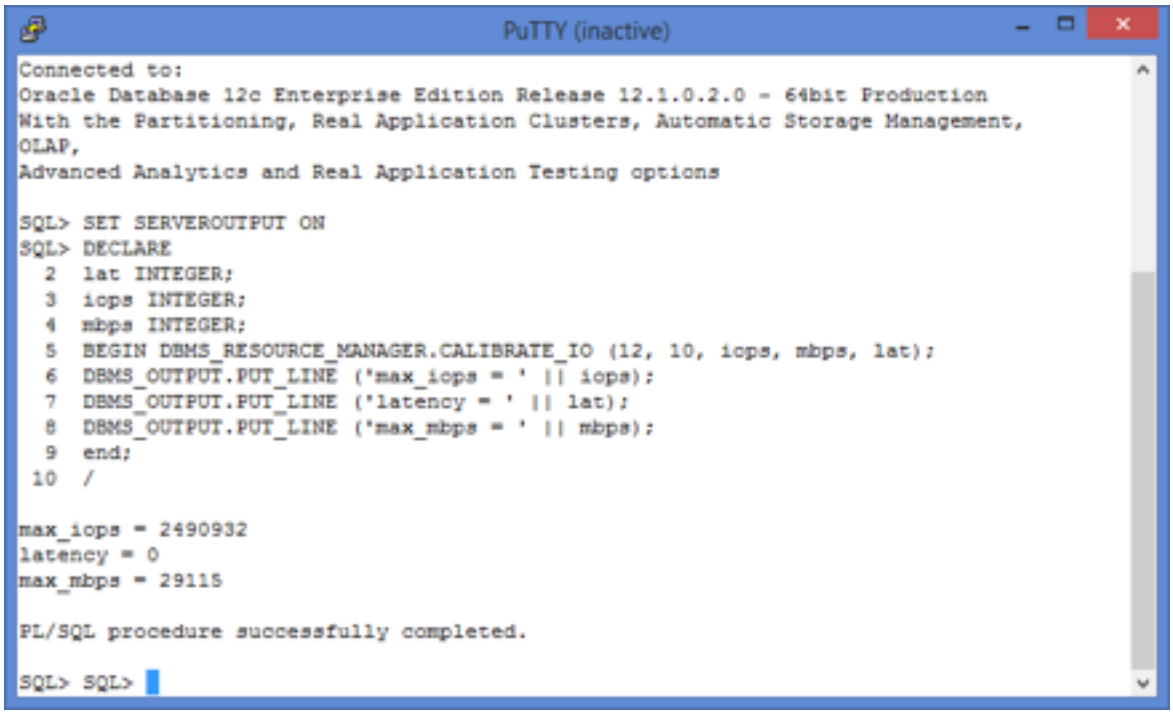


Figure 7. Example of running CALIBRATE_IO on a 3-node cluster with four NVMe SSDs per node.

Performance Results

As shown in these tests, a three-node Oracle RAC with four Intel SSD DC P3700 PCIe-based storage devices per node delivered 29 GB/s of bandwidth and 2.5 million IOPS. Comparing to a traditional FC SAN array architecture, the 29 GB/s of bandwidth would require 18 x 16 Gbps FC links or 36 x 8 Gbps FC links per cluster, which is beyond practical limits of typical FC SAN deployments. (Maximum theoretical bandwidth of a 16 Gbps FC link is 1.6 GB/s and of an 8 Gbps FC link is 0.8 GB/s.) As a result, the FlashGrid architecture provides not only performance isolation between databases and other applications, but also provides significantly higher bandwidth for important database operations such as complex analytical queries, full table scans, or backup.

In all tests, CALIBRATE_IO reported the latency as zero, which means that the actual latency was lower than 1 ms.

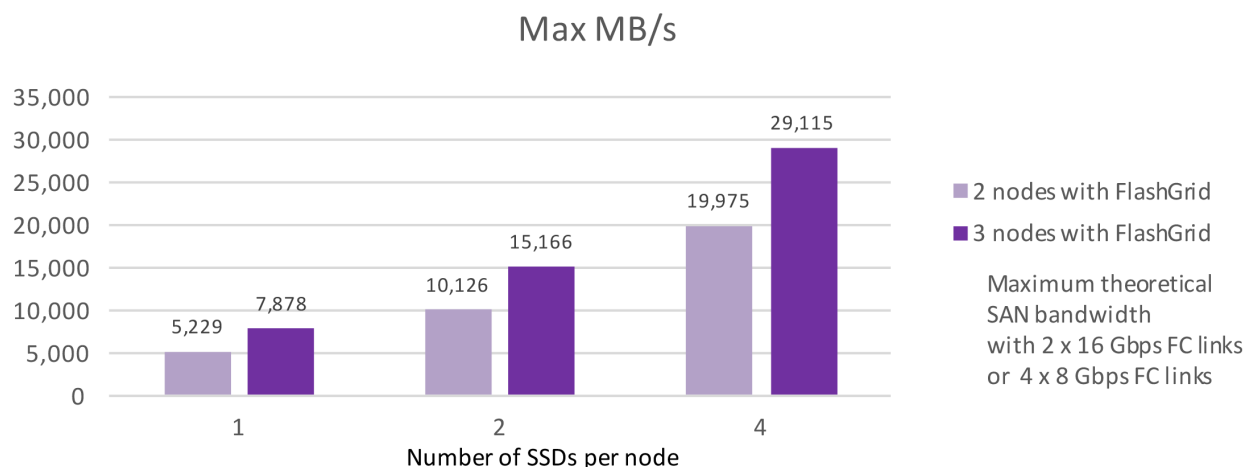


Figure 8. Maximum database storage bandwidth in the test clusters as reported by DBMS_RESOURCE_MANAGER.CALIBRATE_IO.

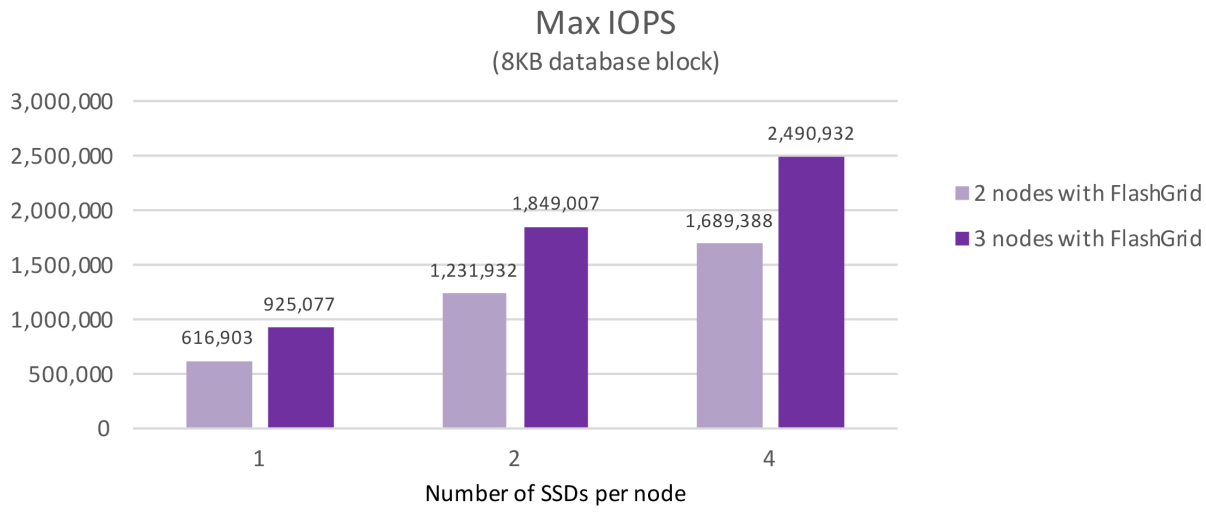


Figure 9. Maximum database storage performance in the test clusters as reported by DBMS_RESOURCE_MANAGER.CALIBRATE_IO.

Capacity and Cost

Besides a significant improvement in performance, the FlashGrid architecture provides flexibility for building very low-cost, entry-level storage to high-capacity, high-performance storage for mission-critical databases or large data warehouses. The smallest storage configuration with two RAC nodes is 400 GB of usable capacity, requiring only two NVMe SSDs. Based on Intel pricing at the time of publication, total storage cost for two Intel SSD DC P3500 is approximately USD 548. For large data set sizes, up to 104 TB of flash can be configured per node.

Intel SSD DC for PCIe storage devices are available in a variety of capacities and with three levels of flash endurance that allow designing highly optimized solutions for various use-cases—all with enterprise-grade reliability. Table 1 provides an overview of the Intel SSD DC for PCIe family. Table 2 provides examples of possible cluster configurations for various use-cases.

Intel SSD model	Intel SSD DC P3500	Intel SSD DC P3600	Intel SSD DC P3608[a]	Intel SSD DC 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
Approx. $/GB[b]	$0.85/GB	$1.0/GB	$2.1/GB	$1.6/GB

Table 1. Overview of Intel SSD DC PCIe-based Family of Storage Devices.

	RAC Standard Edition	Mixed-Use Database	Mission-Critical OLTP Database	Data Warehouse	Large Data Warehouse
Number of nodes	2 nodes	2 nodes	3 nodes	2 nodes	6 nodes
Usable storage capacity	1.6 TB	24 TB	8 TB	96 TB	288 TB
Number of SSDs per node	2	8	4	48	48
SSD type	Intel SSD DC P3600 800 GB	Intel SSD DC P3600 2 TB	Intel SSD DC P3700 2 TB	ntel SSD DC P3500 2 TB	Intel SSD DC P3500 2 TB
Storage bandwidth (est.)[c]	10 GB/s	40 GB/s	29 GB/s	55 GB/s	120 GB/s[d]
Storage performance (est.)[c]	750,000 IOPS	3,000,000 IOPS	2,500,000 IOPS	6,000,000 IOPS	10,000,000+ IOPS[d]
Oracle ASM mirroring	2-way	2-way	3-way	2-way	2-way
Storage HW cost (MSRP)	$3,100[b]	$31,440[b]	$39,180[b]	$108,480[b]	$325,440[b]

Table 2. Examples of storage configurations for various use-cases.

Conclusion

A traditional Oracle RAC implementation, with an external storage array, is complex and creates the problem of "noisy neighbors" affecting database performance. Using FlashGrid software and Intel SSD DC PCIe-based family of storage devices, companies can build hyper-converged Oracle RAC clusters using standard servers to maintain high database performance. FlashGrid seamlessly works with Oracle ASM to provide the needed high availability, reliability, and manageability required for mission-critical workloads. FlashGrid and Intel together create a cost-efficient, high-performance solution for Oracle RAC databases and private database clouds.

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

[b] Based on retail prices reported as of 9/03/2016 at http://www3.intel.com/buy/us/en/catalog/components/ssd/sort,onmarket-d.

[c] The actual performance may vary depending on the server platform and CPUs.

[d] With two 100 Gb/s RDMA storage fabric links per node.

Printed in USA 0916/YMB/HBD/PDF ♻ Please Recycle 334914-001US