High performance Oracle database workloads with the Dell Acceleration Appliance for Databases 2.0

A Dell Reference Architecture

Dell Database Solutions Engineering
June 2015

A Dell Reference Architecture

Revisions

Date	Description
June 2015	Initial release of Fibre Channel-based DAAD 2.0 reference architecture

Contents

Executive Summary

The latest second generation of Dell Acceleration Appliance for Databases (DAAD 2.0) storage node comprises of a Dell PowerEdge R730 server, four Fusion Atomic SX300 series ioMemory PCIe flash adapters, and an updated DAAD ION Accelerator software. DAAD 2.0 storage node will be available in two different capacity flavors—12.8 TB using four 3.2 TB ioMemory PCIe flash adapters and 25.6 TB using four 6.4 TB ioMemory PCIe flash adapters. Both the capacities of DAAD 2.0 will support three different fabrics—Fibre Channel, iSCSI, and Infiniband/SRP.

This whitepaper provides an overview and the configuration details of the Fibre Channel-based DAAD 2.0 reference architecture. This whitepaper is an update to the reference architecture that was done as part of the previous study with DAAD 1.0 that can be found here. It also provides the performance results, analysis, and the improved results of both the capacity flavors of DAAD 2.0 over DAAD 1.0.

The solution described here is a high-availability-enabled, two-node, DAAD storage that uses the Fibre Channel storage protocol. DAAD 2.0 was able to deliver the following performance with the described reference architecture in this whitepaper:

- With Iometer synthetic benchmarking, DAAD 2.0 scaled up to nearly two million IOPS for the 4K Random Read workload, which is nearly double the IOPS performance when compared to DAAD 1.0 while keeping the latency under 3 ms.
- With CalibrateIO benchmarking tool, DAAD 2.0 scaled up to 12 GB per second (GBps) read throughput for the 1 MB sequential read workload, which is about 1.5 times the performance of DAAD 1.0.
- With CalibrateIO benchmarking tool, DAAD 2.0 scaled up to nearly 6 GBps write throughput for the 1 MB sequential write workload, which is about 1.4 times the performance of DAAD 1.0.
- Throughput (GBps) scaled linearly with up to four database nodes stressing a pair of highly available DAAD 2.0 storage nodes.
- Using HammerDB benchmarking tool, DAAD 2.0 scaled over a million New Orders per Minute (NOPM) for TPC-C-like workload with an Average Response Time (ART) of about 10 ms.
- About 15–20% improvement in NOPM performance over DAAD 1.0

1 Introduction

It is very common that the primary performance bottleneck of databases is storage. Due to the storage bottleneck, CPU resources are underutilized because they spend a high percentage of time waiting for IO calls to be serviced. This is an important consideration for CPU-based software licenses such as Oracle: underperforming storage prevents customers from realizing the full potential of their software license investments. Therefore, investing in a high performance storage solution will deliver a high Return on Investment (ROI) in the following ways:

- Improved response time with low latency IO operations, improving customer experience and allowing more on-line orders to be processed per unit of time
- Improved server CPU utilization and software licenses for increased cost savings.
- Ability to consolidate database applications with more storage IO throughput without the added cost of adding CPU resources and software licenses.

Dell Acceleration Appliance for Databases (DAAD) provides a pre-integrated and easy-to-deploy, high performance storage solution that enables IT organizations to quickly and cost-effectively boost database performance. This storage combines Dell PowerEdge servers, Dell Storage, and Dell Networking with innovative flash storage technology from SanDisk to significantly improve Online Transactional Processing (OLTP) or Online Analytical Processing (OLAP) database performance, latency, and throughput. This pre-integrated storage appliance is designed to accelerate leading database environments such as Oracle database, MySQL, Sybase, Microsoft SQL, and MongoDB.

This whitepaper explores the reference architecture and the performance studies of running OLTP workloads on Oracle 12c RAC database and the latest second generation of DAAD, referred to as DAAD 2.0. DAAD 2.0 is based on the latest Dell PowerEdge R730 servers, latest Fusion Atomic CX300 series ioMemory PCIe flash adapters and an updated DAAD ION Accelerator software. It will be offered in two capacity flavors—12.8 TB using four 3.2 TB ioMemory PCIe flash adapters and 25.6 TB using four 6.4 TB ioMemory PCIe flash adapters. Both the capacities will support Fibre Channel, iSCSI, and Infiniband/SRP storage fabrics.

The sections that follow provide the overview, configuration details, and performance results of the Fibre Channel-based DAAD 2.0 reference architecture described in this whitepaper. It also compares its results against the first generation of DAAD, referred to as DAAD 1.0 that was done as part of the previous study that can be found here.

2 Objective

This document provides the reference architecture design and configuration best practices of Oracle 12c RAC Database on the Fibre Channel-based Dell Acceleration Appliance for Databases and covers the following topics:

- Overview of DAAD
- Architecture overview of Oracle 12c RAC database with DAAD
- System hardware/software components and specifications
- Configuration best practices of storage and Oracle RAC database
- Oracle ASM and Oracle 12c RAC database configuration
- Highly available database architecture with the DAAD ION Accelerator software's High Availability (HA) option

This whitepaper also showcases the storage IO performance of the DAAD for the following performance metrics:

- IOPS
- MBps
- Latency

The paper also provides the performance results of DAAD as measured in an Oracle 12c RAC database and studies its scalability. It also discusses the TPC-C-like performance achieved on the Oracle 12c RAC database that is measured in terms of New Orders per Minute (NOPM).

3 Audience

The intended audience of this whitepaper includes IT managers, IT architects, Oracle DBAs, and storage architects who are responsible for planning, designing, implementing, and maintaining Oracle RAC database systems. The reader is expected to have some general knowledge of Oracle RAC, Oracle ASM, and Fibre Channel storage.

4 Dell Acceleration Appliance for Databases (DAAD) 2.0 overview

The Dell Acceleration Appliance for Databases 2.0 (DAAD 2.0) is the second generation of the extremely high performance, low latency and all flash-based DAAD storage appliance. It offers newer and faster hardware, and with multiple storage capacities than DAAD 1.0. It is pre-built by integrating

- Dell PowerEdge R730 server running Intel's latest Haswell processors (E5-2600 v3),
- SanDisk's latest Fusion Atomic CX300 series ioMemory PCIe flash adapters,
- Fibre Channel, iSCSI, or Infiniband front-end fabric adapters,
- DAAD ION Accelerator software.

With this storage appliance as a part of the database infrastructure, IT organizations can take advantage of the following features and benefits for their database applications:

- **Extreme performance**: Extremely high IOPS performance, sub-millisecond latency for OLTP transactional type workloads, and multiple gigabytes per second throughput for OLAP sequential reads and writes in Decision Support Systems and Data Warehouses.
- **High Availability**: Data mirroring across two DAAD nodes setup as an HA pair with seamless database integration with Oracle, Oracle RAC, and other database systems.
- **Multi-fabric**: Multiple front-end fabric connectivity options – Fibre Channel, iSCSI, or Infiniband/SRP – to integrate and work in varied existing or new customer and data center environments
- **Multi-capacity**: Multiple capacities – 12.8 TB or 25.6 TB – to accommodate different starting capacity needs
- **Multi-OS**: Ability to support host-servers running Red Hat Enterprise Linux (RHEL), Oracle Linux running Unbreakable Enterprise Kernel (UEK), Microsoft Windows Server 2012 R2, VMware ESXi, Microsoft Hyper-V, and more
- **Scalable**: Ability to linearly scale in performance and capacity by simply adding more DAAD either with standalone units or as HA pairs with growing business needs.

4.1 DAAD 2.0 components

High-level of the components in DAAD 2.0:

- It is built on the industry leading Dell PowerEdge R730.
- It supports high-speed and high bandwidth front-end connectivity in multiple fabric options—Fibre channel, iSCSI or Infiniband—between database servers and the DAAD nodes.
- Unlike DAAD 1.0 that was offered in only one capacity flavor of 12.8 TB per DAAD node, DAAD 2.0 will be offered with two starting capacity flavors
    - **12.8 TB DAAD**: Built with four Fusion 3.2 TB ioMemory PCIe flash adapters, and
    - **25.6 TB DAAD**: Built with four Fusion 6.4 TB ioMemory PCIe flash adapters

- The DAAD ION Accelerator software installed on each DAAD node enables management and configuration, including HA.
- The highly available DAAD storage cluster configuration consists of a pair of DAAD nodes connected with a private 40GbE point-to-point Interconnection network. DAAD provides scalability across multiple standalone DAAD nodes or across multiple DAAD HA pairs.

NOTE: A DAAD HA pair would require two DAAD nodes of the same capacity.

For the Fibre Channel-based DAAD used in this reference architecture whitepaper, each Dell PowerEdge R730 based DAAD node is pre-integrated with the following components:

- Two Intel Xeon E5-2667 v3 3.2 GHz 8c CPUs and 384 GB of RAM.
- Four Fusion 3.2TB or 6.4 TB ioMemory PCIe flash adapters.
- Two QLogic QLE2662 Dual-Port 16 Gb Fibre Channel HBAs for front-end connectivity.
- One Mellanox ConnectX-3 Dual-Port 40GbE Network Adapter for high availability interconnection.
- DAAD ION Accelerator software version 2.5.1 with the HA feature enabled.

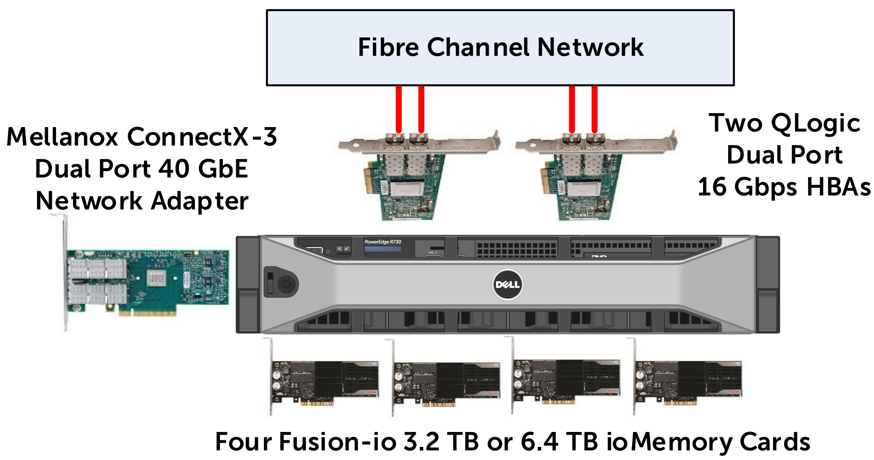

Figure 1 Fibre Channel-based DAAD 2.0 PCIe components

Similar to DAAD 1.0, this appliance uses seven Gen3 PCIe slots in the R730 servers for the ioMemory adapters and for the network cards, as shown in Figure 2:

- PCIe slot 1 is fitted with a Mellanox ConnectX-3 40 GbE card for the high availability connectivity.
- PCIe slots 2-3 are each fitted with QLogic QLE2662 Dual-Port 16 Gb Fibre Channel HBAs, and
- PCIe slots 4-7 are each fitted with either Fusion 3.2 TB or 6.4 TB ioMemory PCIe flash adapters.

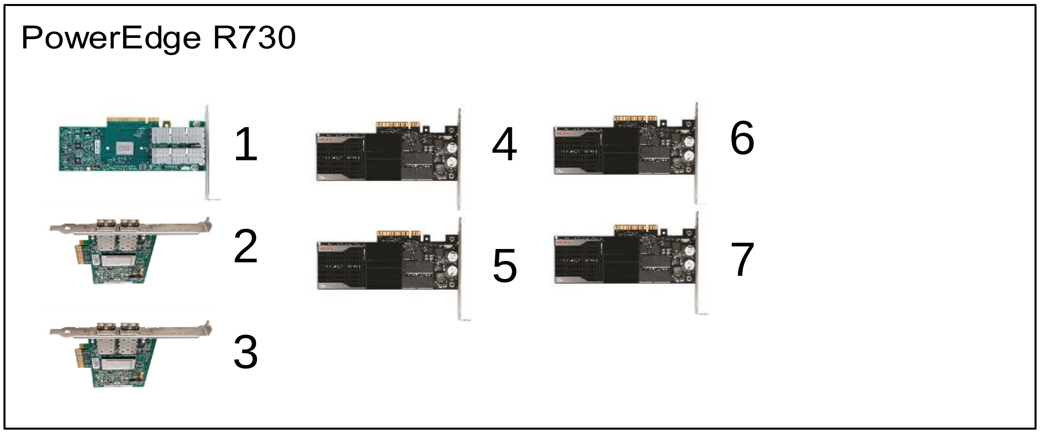

Figure 2 Fibre Channel-based DAAD 2.0 PCIe slot population

NOTE: For full configuration details please see Appendix A.2.

4.2 High availability configuration on DAAD 2.0

Similar to DAAD 1.0, the DAAD 2.0 offers a standalone option that consists of a single DAAD node. With this option, we can achieve high availability at the ioMemory adapter level with a RAID 10 configuration by mirroring two ioMemory adapters with another two ioMemory adapters. This configuration prevents any data loss due to the unlikely failure of an ioMemory adapter. As a result of the mirroring, the usable capacity of each node will be reduced to 50% of the original capacity, namely 12.8 TB to 6.4 TB or 25.6 TB to 12.8 TB per DAAD node depending on the capacity flavor of the DAAD 2.0 in use. However, a single DAAD node presents a single point of failure. So, high availability at the DAAD node level is highly recommended.

High availability at the DAAD nodes can be implemented in two ways:

- Host-based mirroring, or
- DAAD ION Accelerator HA feature

In a host-based mirroring configuration, the mirroring of the storage volumes from two DAAD nodes are implemented by a traditional host-based mirroring method. For Oracle database configurations, Oracle Automatic Storage Manager (ASM) mirroring and failure group configuration is used to implement this data redundancy. The mirrored Logical Unit Numbers (LUNs) from the pair of clustered appliance nodes are put into separate ASM failure groups. Oracle ASM performs synchronous writes to these two LUNs on both DAAD nodes in parallel. To ensure that the Oracle Clusterware functions properly in an event that one of DAAD nodes is offline in an Oracle RAC configuration, the disk group for the Oracle Clusterware voting disk must have either external redundancy set between the DAAD nodes or have an odd number of voting disk copies present on separate DAAD nodes (3 copies for normal redundancy and 5 copies for high redundancy). A two node DAAD configuration with the host-based mirroring option doesn't provide either of these options.

In contrast to the host-based mirroring method, for mission–critical database applications, the DAAD ION Accelerator software provides the HA feature, a powerful and effective storage array-based solution for high availability. This high availability configuration consists of two DAAD nodes that store redundant copies of the data and communicate to each other over a redundant private 40 GbE point-to-point interconnect link. The DAAD ION Accelerator software's HA feature replicates all the data block changes between the two clustered DAAAD nodes. The advantage of this method is that HA is provided at the storage level. Dell recommends the DAAD ION Accelerator HA clustering to implement high availability for an Oracle RAC database configuration as it provides the external redundancy for the Oracle Grid Infrastructure voting disk needed to ensure that the voting disk is accessible to the Oracle RAC Database nodes even when one of the DAAD nodes is offline.

Figure 3 shows the architecture of the Fibre Channel-based DAAD that is configured as an HA pair.

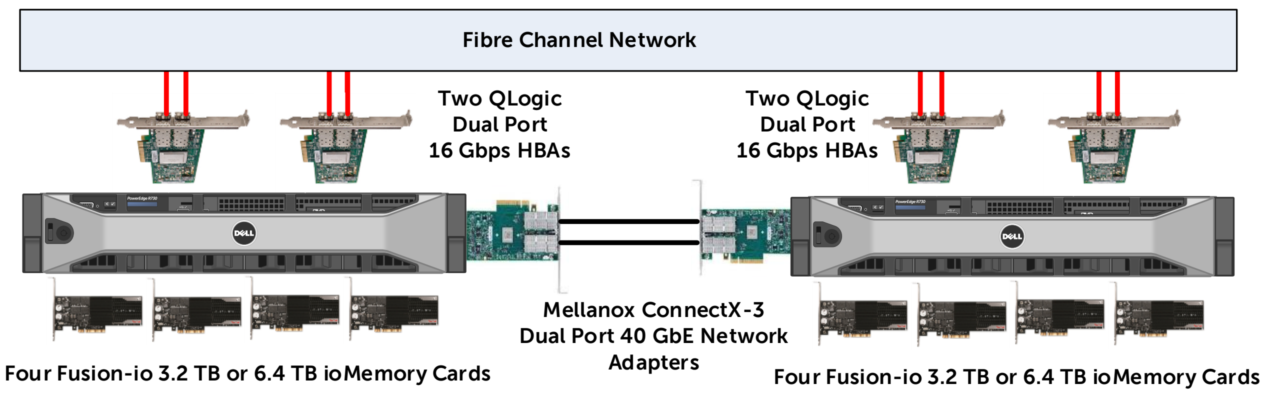

Figure 3 The Fibre Channel-based DAAD 2.0 configured as an HA pair

5 Solution design and configuration best practices

This section discusses the combination of an Oracle 12c database with DAAD and the configuration method that is used to implement this solution. This arrangement is designed to establish a solution for Oracle 12c RAC databases by using the Fibre Channel-based DAAD with the HA option enabled. However, this architecture and configuration method also applies to a single node Oracle standalone database system.

5.1 Solution architecture

In the solution, Oracle 12c RAC database uses a clustered pair of Fibre Channel-based DAAD nodes as the shared storage to store Oracle database files as well as the Oracle Clusterware Registry (OCR) and voting disk files of Oracle 12c Grid Infrastructure. The DAAD ION Accelerator HA clustering enables the HA storage architecture that prevents the Oracle 12c Grid Infrastructure and the Oracle 12c RAC Database from a single point of failure of hardware and software components in the appliance.

Figure 4 shows the architecture of an Oracle 12c RAC database infrastructure with DAAD as the SAN shared storage. The Oracle database nodes connect to the DAAD HA pair through a Fibre Channel network. The database nodes are also connected to private network for the Oracle cluster heartbeat and for the Oracle RAC data synchronization between the Oracle RAC database nodes. All the database nodes are connected to the public network which allows applications and DBAs to connect to the database nodes. The public network is usually connected to the main corporate network.

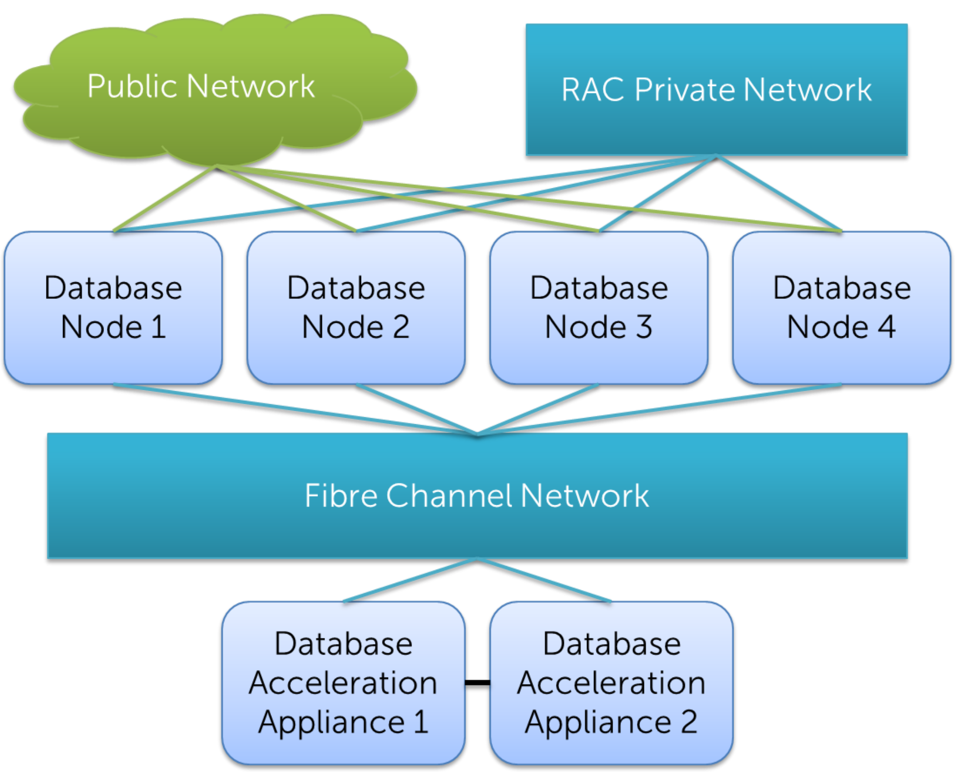

Figure 4 Architecture of Oracle 12c RAC database with DAAD as the SAN shared storage

5.2 Physical architecture and network configuration

To ensure the high availability of the infrastructure, Dell recommends that the network and the storage IO paths are redundant. The following diagram shows the physical implementation of this reference architecture.

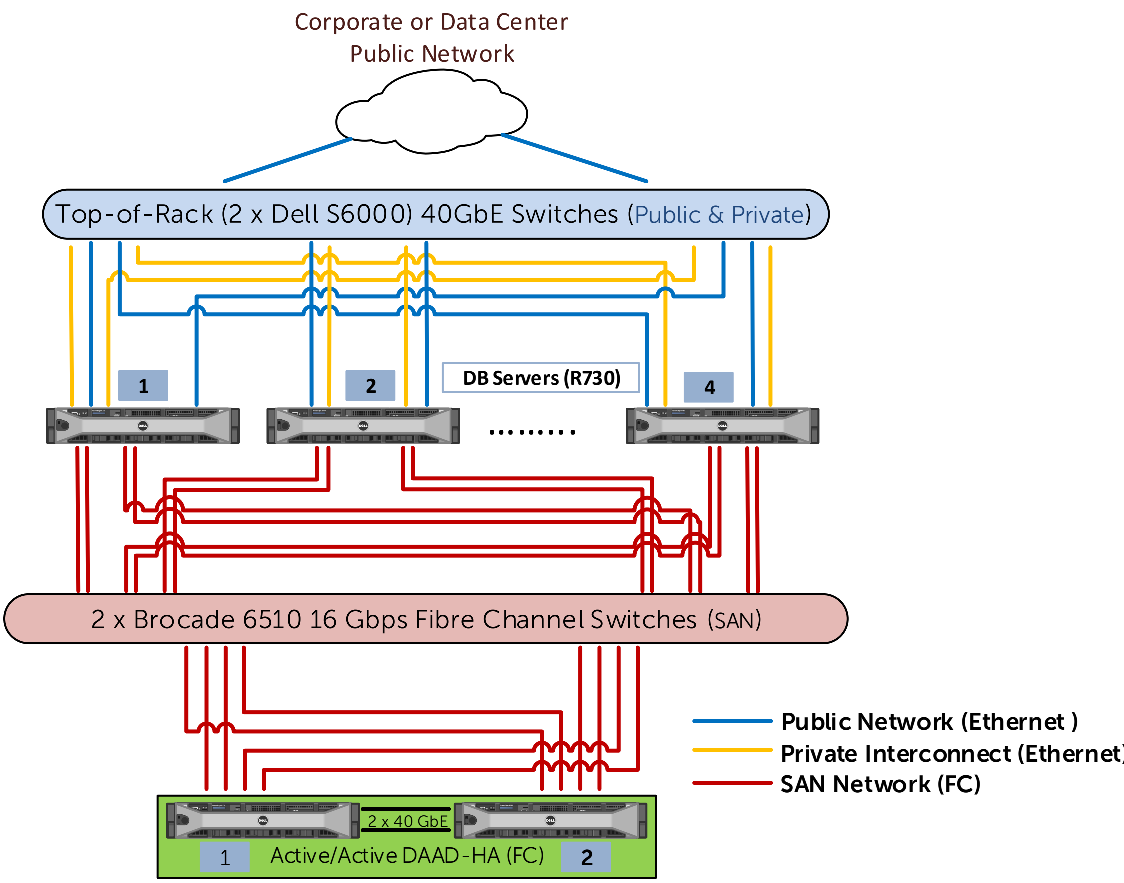

Figure 5 Oracle 12c RAC database reference architecture using DAAD

From the top down in Figure 5, the two top-of-the-rack switches are connected to the RAC database nodes and the network backbone of the data center or the corporate network to which all the applications are also connected. Through this public network, applications send the SQL queries to the database server nodes and the database server nodes send the database query results back to the applications. Dell recommends two dedicated switches for the private network, but if that is not an option then Dell recommends to segregate the private traffic by creating a separate virtual network like VLAN. In this reference architecture, the public and the private network share the same redundant high bandwidth, low latency Dell Networking S6000 switches. However, both the networks are configured on separate VLANs to ensure the segregation of the two network traffic. To test the performance capability of the two DAAD HA nodes, the database nodes were scaled from one all the way up to four nodes by using the Dell PowerEdge R730 servers.

As shown in Figure 5, the Fibre Channel network that connects the Oracle RAC database server nodes and the DAAD consists of two 16 Gbps Fibre Channel switches (Brocade 6510). Each database server node is connected to both of these Fibre Channel switches, as is each DAAD node. This configuration ensures that there is no single point of failure in the storage network.

5.3 SAN configuration

As shown in Figure 6, in this configuration, two DAAD nodes are connected to two Fibre Channel switches. To increase the storage network bandwidth, each DAAD node is configured with two dual-port 16 Gb Fibre Channel QLogic QLE2662 HBAs that together provide a total of four Fibre Channel links split between the two Fibre Channel switches.

Figure 6 shows the connectivity between the two dual-port HBAs and the two Fibre Channel switches: each of the ports on a single card connect to a separate Fibre Channel switch. This configuration provides four distinct links between the Fibre Channel network and each DAAD node, which ensures that the entire storage network is free of a single point of failure.

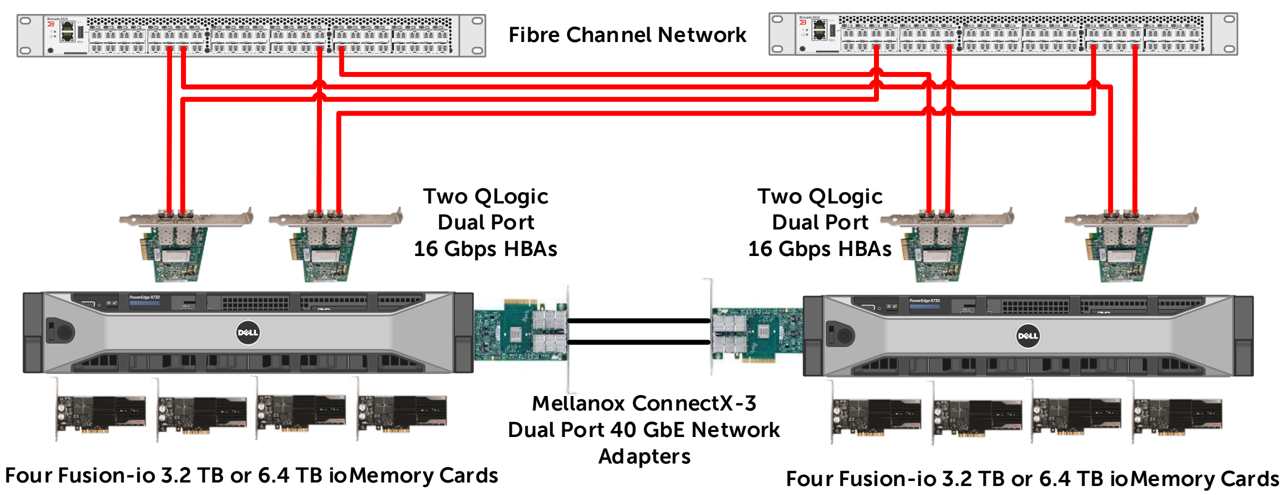

Figure 6 DAAD Fibre Channel SAN connectivity

DAAD is shipped from the Dell factory with the DAAD ION Accelerator software pre-installed. When a DAAD node is booted for the first time, it boots up with the first boot setup configuration process, during which the following tasks can be accomplished:

- Initiate a scan of the existing network cards and controllers
- Enter license agreement
- Configure the network, including management node and HA cluster
- Setup the date and time
- Cluster setup (if the HA feature is licensed)
- Set the password for `admin` user

For further information about these tasks, refer to the *Dell Acceleration Appliance for Databases 2.0 Configuration Guide* that can be found at http://support.dell.com.

After the first boot configuration process completes, connect to the DAAD nodes through the management network to configure the storage flash devices. DAAD ION Accelerator software provides two ways to connect to the DAAD: a command line interface (CLI), or a browser-based graphical user interface (GUI). Either connectivity method can be used to perform the storage configuration tasks including setting up the storage profile and pools, creating volumes, adding initiators, and managing the ioMemory adapters.

- Use an SSH tool to log in to the DAAD using the management IP as an `admin` user. This will allow access to the CLI to run the DAAD ION Accelerator commands to perform the configuration tasks.
- Alternatively, use a web browser to log in to the DAAD GUI using the management IP as an `admin` user to perform the configuration tasks.

Figure 7 shows the GUI of a two-node HA clustered Fibre Channel-based DAAD 2.0.

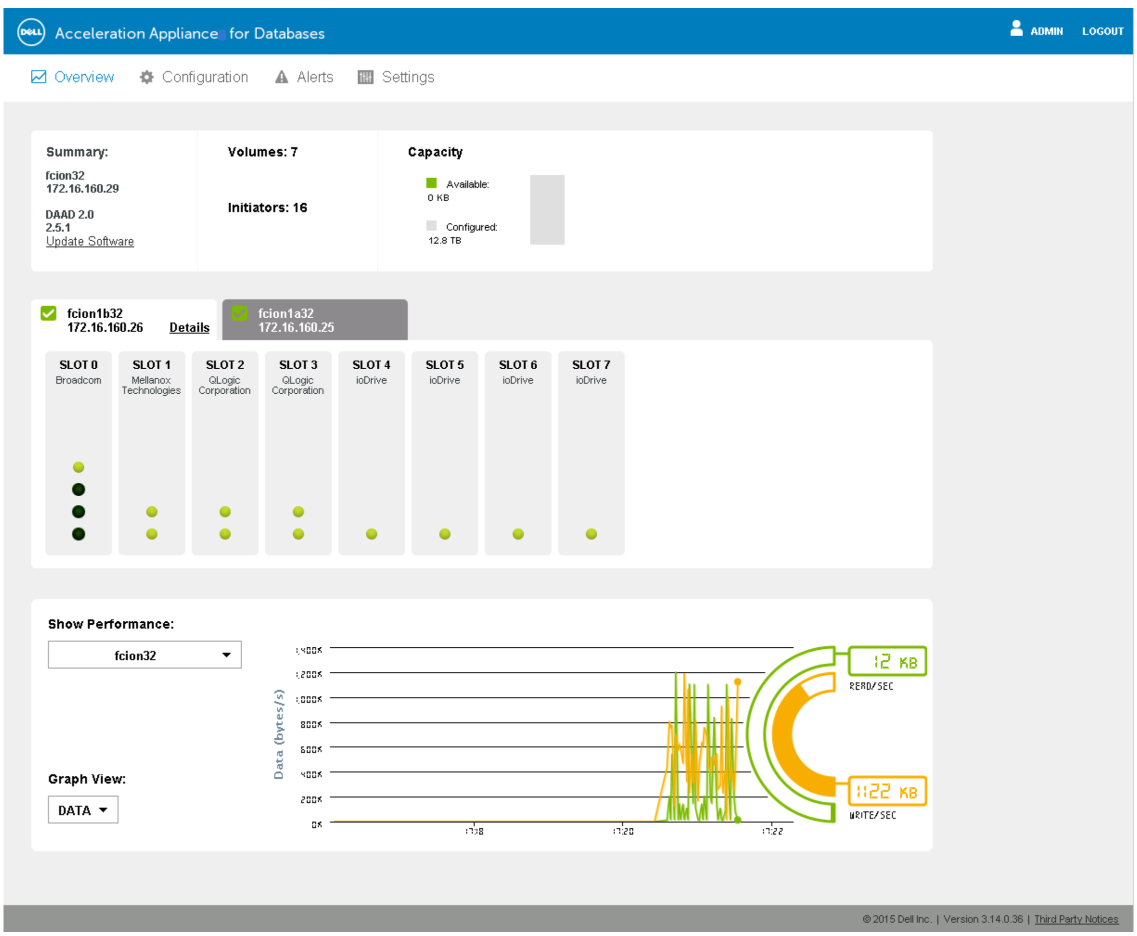

Figure 7 DAAD ION Accelerator software GUI

5.4 Storage configuration

Depending on the capacity flavor in use, each DAAD node has either 3.2 TB or 6.4 TB ioMemory adapters. In this reference configuration with the DAAD ION Accelerator HA clustering enabled, the four ioMemory adapters (ioDrive1-4) of one DAAD node are mirrored with the four ioMemory adapters (ioDrive1-4) in the other node of the HA pair, as shown in Figure 8.

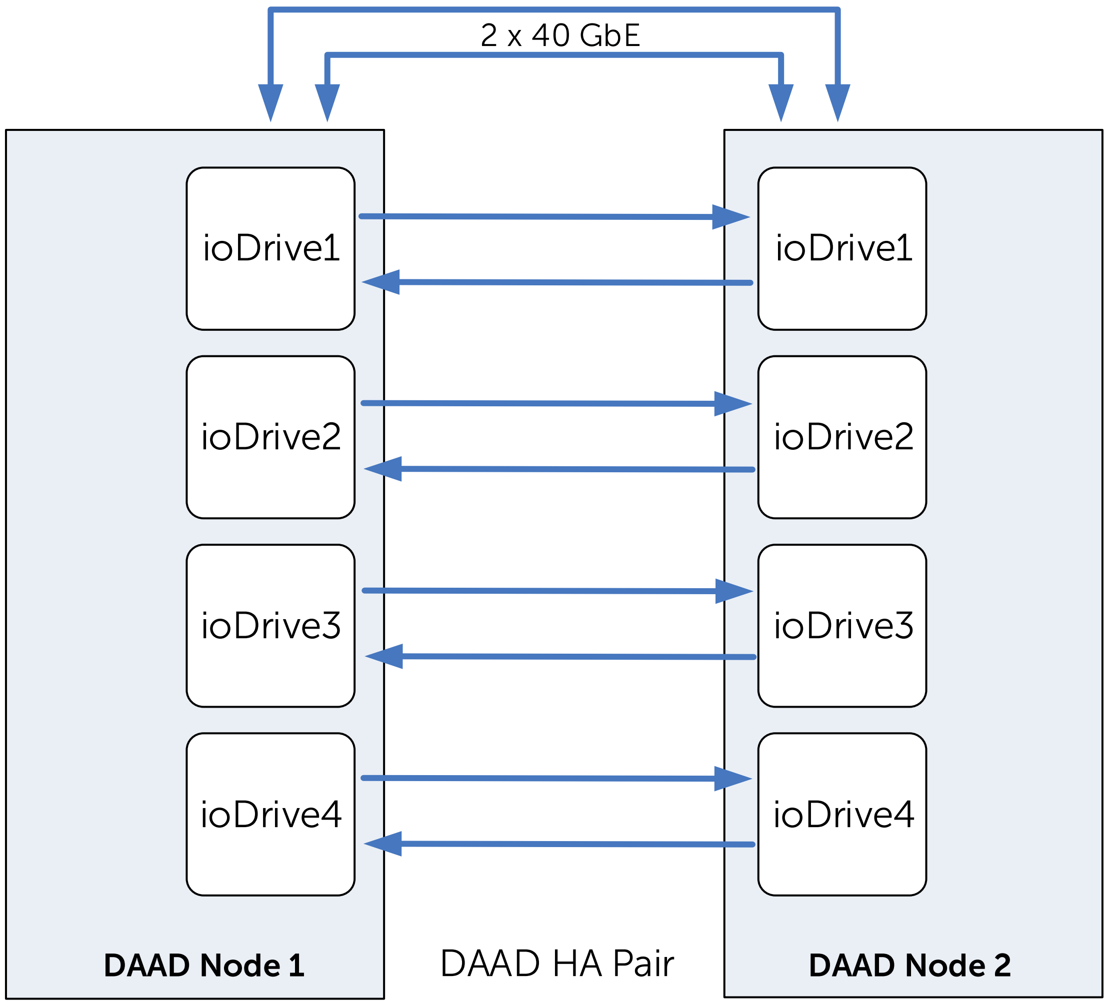

Figure 8 Mirroring between the ioMemory adapters between the DAAD HA nodes

In the storage configuration, a storage pool was created for each ioMemory adapter and then two volumes were created in each storage pool. Each volume was manually assigned a primary DAAD node and a secondary DAAD node. As shown in Figure 9, volumes V1, V3, V5, and V7 use DAAD node 1 as the primary node and volumes V2, V4, V6, and V8 use DAAD node 2 as the primary node to present the volumes to the database nodes. When the database servers update data on the volumes, updates are loaded onto the primary nodes and are then replicated to their mirrored volumes on the secondary node. For example, the update on volume V1 will first come to the V1 of ioDrive1 on DAAD node 1, and then the updates will be replicated through the 40 GbE HA interconnect links to the ioDrive1 on DAAD node 2. This design allows us to balance the workloads evenly across the two DAAD nodes.

Using the 3.2 TB based DAAD HA pair in the second example, the following DAAD ION Accelerator CLI command creates a volume `V1` with size 1600 GB on storage pool `ioDrive1_pool` with `node1` as the primary node and `node2` as the secondary node.

```
admin@node1/> volume:create -n node1 -n node2 V1 1600 ioDrive1_pool
```

Use a similar command to create eight 1600 GB volumes `V1, V3, V5`, and `V7` with `node 1` as the primary node and volumes `V2, V4, V6`, and `V8` with `node2` as the primary node as shown in Figure 9.

Each of these eight volumes is created with the same size as they will be presented to the Oracle database nodes to form Oracle ASM disks of an ASM disk group.

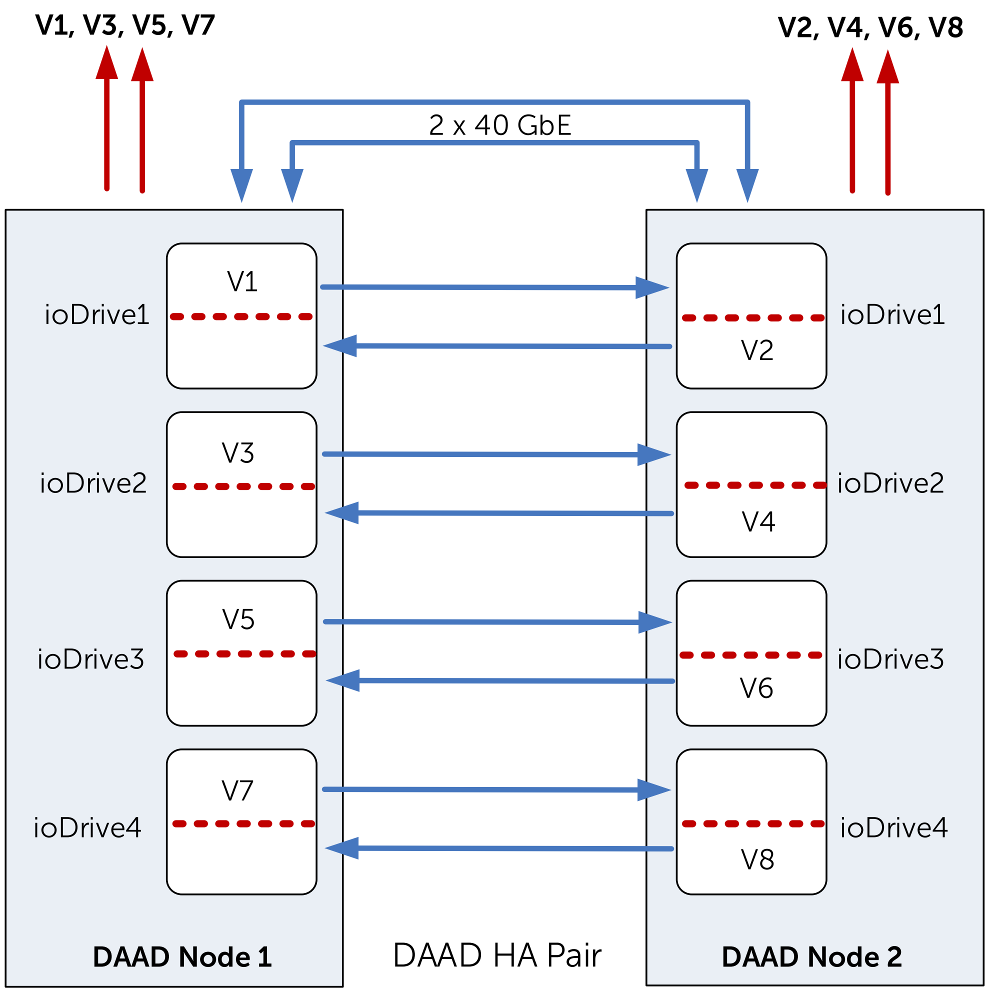

Figure 9 Storage volume setup on the DAAD HA pair

5.5 Database node configuration

The reference configuration includes up to four Dell PowerEdge R730 servers as the Oracle RAC database nodes. Each R730 database node consists of the following hardware components:

- CPU: Two Intel Xeon E5-2687 v3 10c @ 3.1 GHz
- RAM: 128 GB RAM
- Network cards: Two Mellanox ConnectX-3 EN Dual Port 40 GbE Ethernet adapters
- HBAs: Two QLogic QLE2662 16 Gb Fibre Channel adapters

The following software products are installed on each database server node:

- OS: Oracle Linux 6.6 running Unbreakable Enterprise Kernel (UEK) [3.8.13-44.1.1.el6uek.x86_64]
- Oracle Database : Oracle Database 12c Enterprise Edition Release 12.1.0.1.0 - 64bit Production with Oracle Real Application Clusters and Automatic Storage Management options

To provide storage network bandwidth, each database server node is also configured with two dual-port HBA cards that enable four 16 Gb Fibre Channel connections. As in the DAAD nodes, each port of the dual-port HBA cards is connected to a separate Fibre Channel switch as shown in Figure 10.

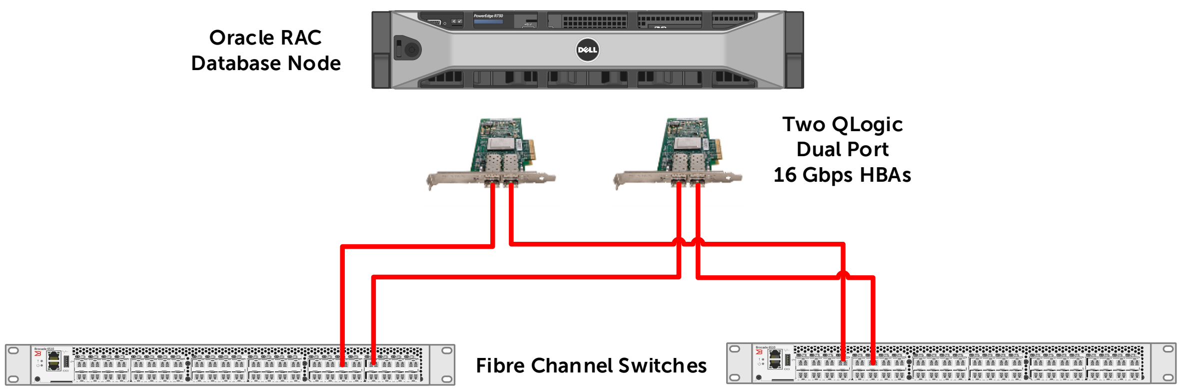

Figure 10 Connecting database server nodes to Fibre Channel switches

5.6 Fibre Channel storage zoning and LUN creation

For the database nodes to access the storage devices, we need to establish the Fibre channel zoning on both Fibre Channel switches. Each zone links the HBA ports of database server nodes with the HBA ports on the DAAD nodes. This zoning method follows standard “Single Initiator” Fibre Channel zoning practices. Figure 11 shows an example of such a zone configuration in one of the Fibre Channel switches. In this example, a zone is established with the database server node HBA port that connects to port 15 of the FC switch with four HBA ports across both the DAAD nodes. This establishes the I/O paths from the database node that connects to port 15 to two DAAD nodes. Because each database node and DAAD node is connected to two ports on each of the two Fibre Channel switches, there are eight distinct I/O paths to the DAAD nodes per switch for a total of 16 distinct I/O paths.

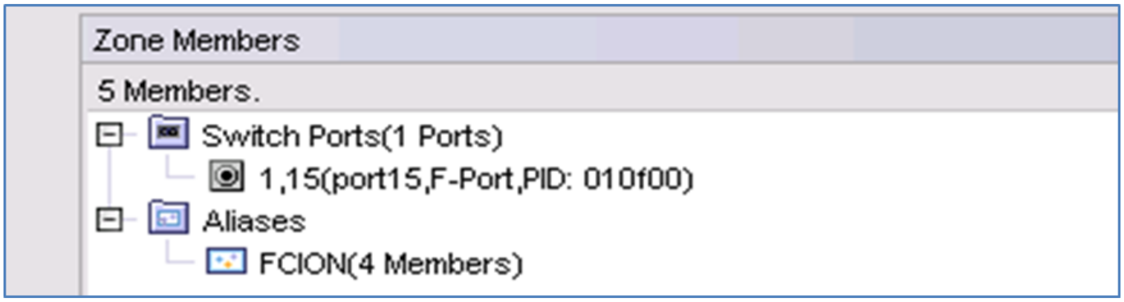

Figure 11 Fibre Channel switch zoning

After the zoning on both the Fibre Channel switches is established, the DAAD ION Accelerator software attempts to auto-discover the database servers' HBA ports that are connected to the DAAD nodes by zoning. These ports' WWPNs will be shown as the initiators on the **INTIATORS** list of the DAAD ION Accelerator software's GUI interface as shown in Figure 12.

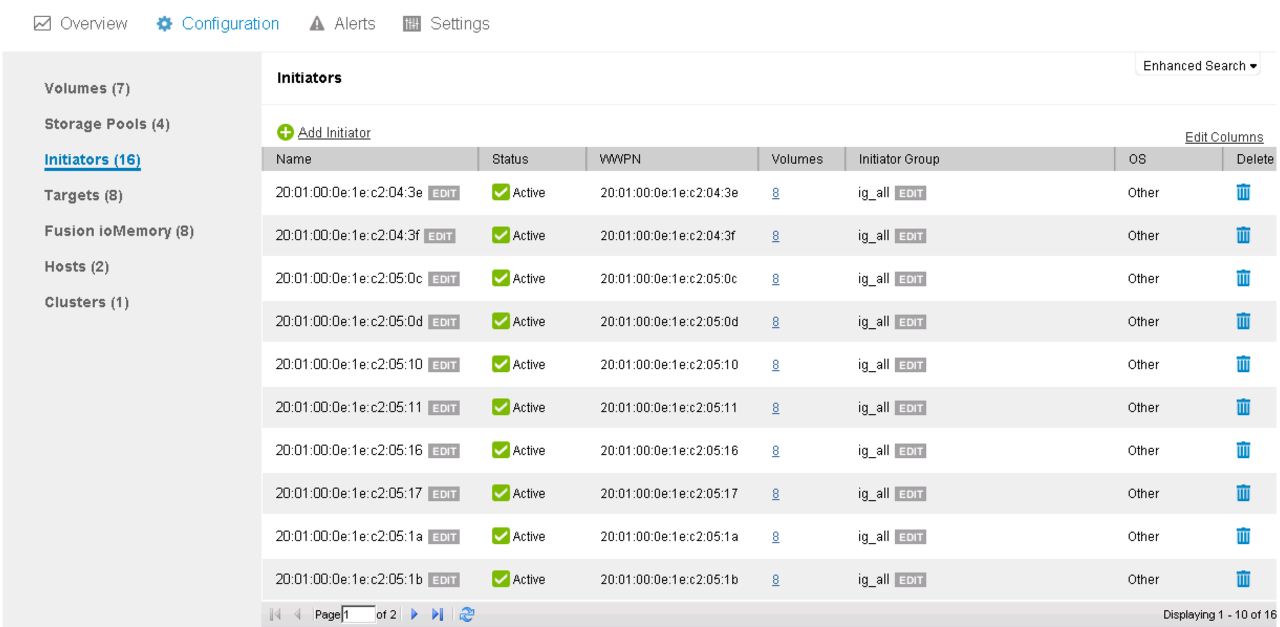

Figure 12 DAAD initiators list

The initiators with similar access privileges are collected into an initiator group. As shown in Figure 12, an initiator group named `ig_all` is created to group all initiators that belong to the database server nodes of the Oracle RAC database. When this initiator group is granted access to the volumes, the corresponding database server nodes are granted access to the volumes.

To give an initiator group access to a volume you can specify the initiator group name during LUN creation. For example, the following CLI command creates a LUN `fcion_v1` and gives it access to the initiator group `ig_all`, which represents all initiators that belong to the four nodes of the Oracle RAC database servers.

```
admin@node1> lun:create fcion_v1 ig_all -b 512 -a
```

Figure 12 shows that initiator group `ig_all` is granted access to the eight volumes, which are all of the volumes of the two DAAD nodes. This indicates that all of the Oracle RAC database server nodes have access to all eight volumes of the DAAD HA pair.

For details on configuring volumes, initiator groups, and LUNs, refer to Dell Acceleration Appliance for Databases CLI Reference and Dell Acceleration Appliance for Databases GUI Guide that can be found at http://support.dell.com.

5.7 Disk or device configuration on database nodes

After the LUNs are created and the database server nodes are able to access the volumes or the disks, the database server nodes must discover the LUNs and map them to aliases so that Oracle databases can access them. This can be achieved as follows:

1. Ensure `sg3_utils` and `device-mapper-multipath` are installed on each node.

   ```
   yum -y install sg3_utils device-mapper-multipath
   ```

2. Run the script `rescan-scsi-bus.sh` (part of the `sg3_utils` package) on each node to discover LUNs.
3. On each database server node, edit `/etc/multipath.conf` and add the following to the `devices` section:

   ```
   device {
       vendor                  "FUSIONIO"
       features                "3 queue_if_no_path pg_init_retries 50"
       hardware_handler        "1 alua"
       no_path_retry           3
       path_grouping_policy    group_by_prio
       path_selector           "queue-length 0"
       failback                immediate
       path_checker            tur
       prio                    alua
       fast_io_fail_tmo        15
       dev_loss_tmo            60
   }
   ```

4. With the LUNs mapped as `/dev/dm-[*]` devices, run `map_wwid_to_iob_alias.py` (refer to Appendix D) to generate the list of multipath aliases and the corresponding LUN WWIDs. Add this list to the `multipath.conf` file on each database node.
5. After the `multipath.conf` file is configured with the appropriate aliases for each device, the `multipathd` service needs to be reloaded for the changes to take effect. This can be achieved by running the following command:

   ```
   #> service multipathd reload
   ```

6. Verify that the aliases work by running the following command and ensure all devices are visible

```
#> ls -l /dev/mapper/ fcion*
```

7. On just one server, delete any old data on the devices by running the Linux `dd` command as follows:

```
#> cd /dev/mapper
#> for x in fcion_v_??; do dd if=/dev/zero of=$x bs=8192 count=128000;
done
```

8. Create partitions on each device. If you are running an Oracle RAC database configuration, you must set two partitions. The first partition must be 15 GB for OCR and voting disks. The second partition must use all remaining disk space for Oracle data files. If you are running an Oracle standalone database, you must create just a single partition on each device for the Oracle data files. To create two partitions for the Oracle RAC database configuration—on just one server—run these commands to create two partitions on each device.

```
#> cd /dev/mapper
#> for x in fcion_v_??; do (parted -s -a optimal $x mklabel gpt mkpart
primary 1 15G mkpart primary 15G 100%); done
```

9. On each server, rescan the partition table

```
#> partprobe
```

10. Permissions need to be set on the devices for them to be used by Oracle database. This is done by creating a new `udev` rules file located in `/etc/udev/rules.d/`. Each device must be owned by the `grid` user and be part of the `asmadmin` group. Here are the steps for creating the `udev` rules to be performed on each database node:

    a. Copy the sample rules file `12-dm-permissions.rules` to the `udev` directory:

    ```
    #> cp /usr/share/doc/*/12-dm-permissions.rules /etc/udev/rules.d/
    ```

    b. Add the following lines to the end of `12-dm-permissions.rules` file:

    ```
    ENV{DM_NAME}=="fcion*1", OWNER:="grid", GROUP:="asmadmin", MODE:="0660"

    ENV{DM_NAME}=="fcion*2", OWNER:="grid", GROUP:="asmadmin", MODE:="0660"
    ```

    c. Reload the rules into memory:

    ```
    #> udevadm trigger
    ```

11. Reboot all nodes at this time to ensure that all changes take effect.
12. After rebooting the nodes, verify on each node if the `udev` rules and permissions got applied correctly by running the following command:

```
#> ls –l /dev/dm-*
```

5.8 Oracle 12c grid infrastructure and RAC database configuration

For Oracle 12c RAC database configuration, the DAAD LUNs or devices that are presented to all the database servers are used for storing the following:

- Oracle OCR and voting disks
- Oracle database files.

As mentioned in the Section *5.7 Disk or device configuration on database nodes*, two partitions are created on each device: first partition (15 GB) and second partition (remaining space on the device). The first partition is for OCR and voting disks. The two node DAAD HA pair in this reference architecture was configured by using best practices recommended by Dell, and as described in Section 5.4 Storage configuration. As a result, there are a total of eight devices resulting in a total of eight 15 GB first partitions and eight second partitions. During the Oracle 12c Grid Infrastructure installation, five of these 15 GB first partitions can be used to create a disk group for the OCR and voting disks. The eight second partitions are used to create the ASM disk group for Oracle data files. As a best practice, ensure that the primary volumes (or partitions) from each DAAD node are used in the ASM disk group in order to distribute the Oracle data files and take advantage of the performance of all the ioMemory adapters across the two DAAD HA pair nodes.

The following shows an example of creating ASM disk group '`+DATA`' by using these eight second partitions:

```
su - grid
sqlplus / as sysasm
create diskgroup DATA external redundancy disk
  '/dev/mapper/fcion_v_1p2',
  '/dev/mapper/fcion_v_2p2',
  '/dev/mapper/fcion_v_3p2',
  '/dev/mapper/fcion_v_4p2',
  '/dev/mapper/fcion_v_5p2',
  '/dev/mapper/fcion_v_6p2',
  '/dev/mapper/fcion_v_7p2',
  '/dev/mapper/fcion_v_8p2';
```

The Oracle database's DBCA utility can be used to create the Oracle 12c RAC database on the "`+DATA`" disk group.

6 Performance Test Methodology

The Oracle RAC database configuration for the performance studies is based on the reference configuration described in this document. Two separate Oracle RAC database clusters were setup:

- One cluster was setup with four Oracle RAC database nodes connected to a pair of DAAD HA nodes with each containing four 3.2 TB ioMemory PCIe flash adapters.
- The other cluster was setup with four Oracle RAC database nodes connected to another pair of DAAD HA nodes with each containing four 6.4 TB ioMemory PCIe flash adapters.

Both the DAAD nodes in the two clusters were enabled with the HA feature. Three different application tools were used to measure the I/O performance of the DAAD 2.0 and the overall performance of the entire Oracle RAC database environment.

- Iometer
- CALIBRATE_IO
- HammerDB

Iometer is a synthetic storage benchmarking application tool and was used to test the I/O performance of the two DAAD 2.0 capacity offerings. Four access patterns were used to ascertain performance.

- 4k Random Read
- 8k Random 70/30 Read/Write (simulates OLTP-like workload)
- 1MB Sequential Read (simulates OLAP-like workload)
- 1MB Sequential Write

CALIBRATE_IO is a procedure in Oracle Database. This procedure was used to measure the performance of an Oracle RAC 12c database and studied the performance scalability by the number of Oracle RAC database nodes for the following metrics.

- Max IOPS
- Max MBps

HammerDB is used to measure the TPC-C-like performance of the Oracle RAC database and the scalability of the performance by the number of Oracle RAC database nodes. The performance metrics that were used to analyze the performance of the system were:

- New Orders per Minute (NOPM)
- Average Response Time (ART)

The scalability testing for both CALIBRATE_IO and HammerDB tests were conducted by running Oracle RAC database across one node, two nodes, and four database nodes. This section also provides the comparative graphs of the performance improvement of DAAD 2.0 over the previous generation DAAD 1.0

Note: For more details about the benchmarking application tools and settings, see Appendix C.

7 Performance Results and Analysis

7.1 Iometer

This section shows the performance results obtained by using Iometer benchmarking tool for the various workloads patterns, especially seen in OLTP and OLAP database environments. The graphs shown in this section show the performance results of both the capacity flavors offered with DAAD 2.0 and also compares its performance with DAAD 1.0.

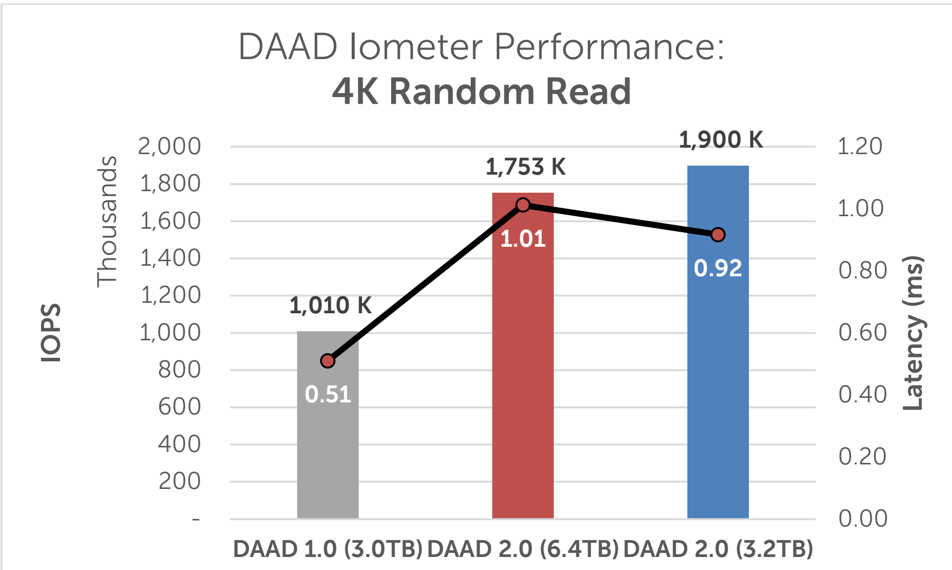

Figure 13 Iometer performance results: small 4K random reads

As shown in Figure 13, the 3.2 TB ioMemory adapter-based DAAD 2.0 scales up to nearly two million IOPS for the 4K Random Read workload which is nearly double the IOPS performance when compared to DAAD 1.0. DAAD 2.0 delivers this performance while keeping the latency under 1 ms.

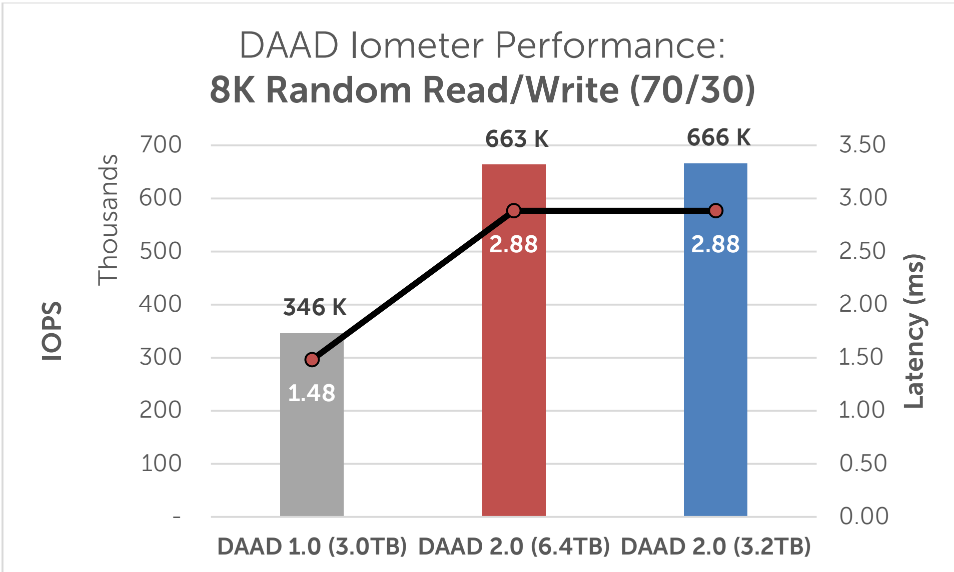

Figure 14 Iometer performance results: small 8K random read/writes (70/30)

As seen in Figure 14, DAAD 2.0 delivers around 665K IOPS when stressed with 8K random workload with 70% reads and 30% writes distribution. This performance number is nearly double when compared to the performance of DAAD 1.0 under the same workload pattern, which is usually seen in an OLTP-like database environment.

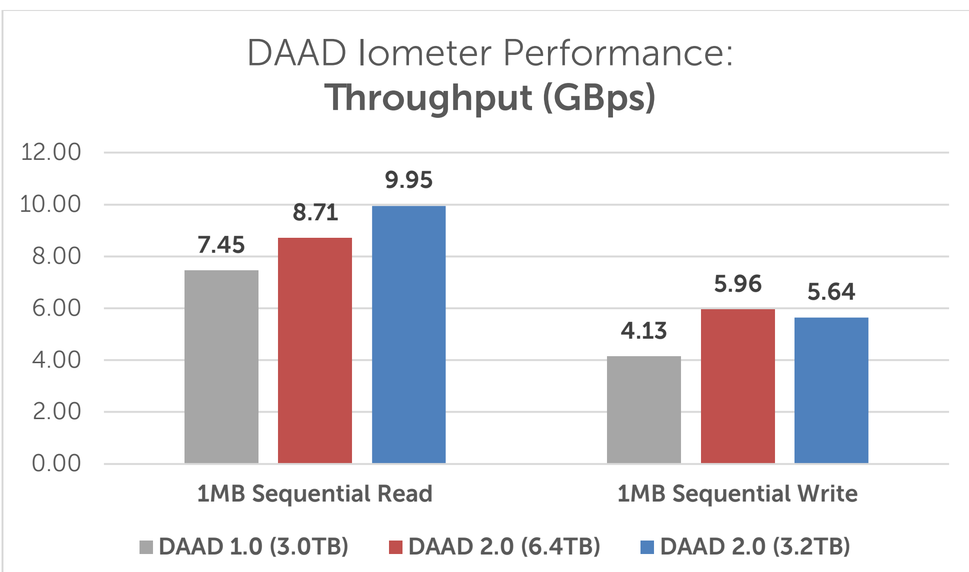

Figure 15 Iometer performance results: large sequential reads and writes

Figure 15 shows the throughput performance of the two generations of the DAAD, measured in GB per second (GBps). As seen in Figure 15, DAAD 2.0 scales up to nearly 10 GBps when stressed with large 1 MB sequential reads and nearly up to 6 GBps when stressed with large 1 MB sequential writes. This is about 1.4 times the performance when compared to DAAD 1.0.

7.2 CALIBRATE_IO:

CALIBRATE_IO provides information similar to that offered by Iometer except that the data access patterns used are provided by Oracle and are intended to be representative of common database workloads. Max IOPS is designed to simulate a transactional database (OLTP), while Max GBps is simulating an analytics database (OLAP).

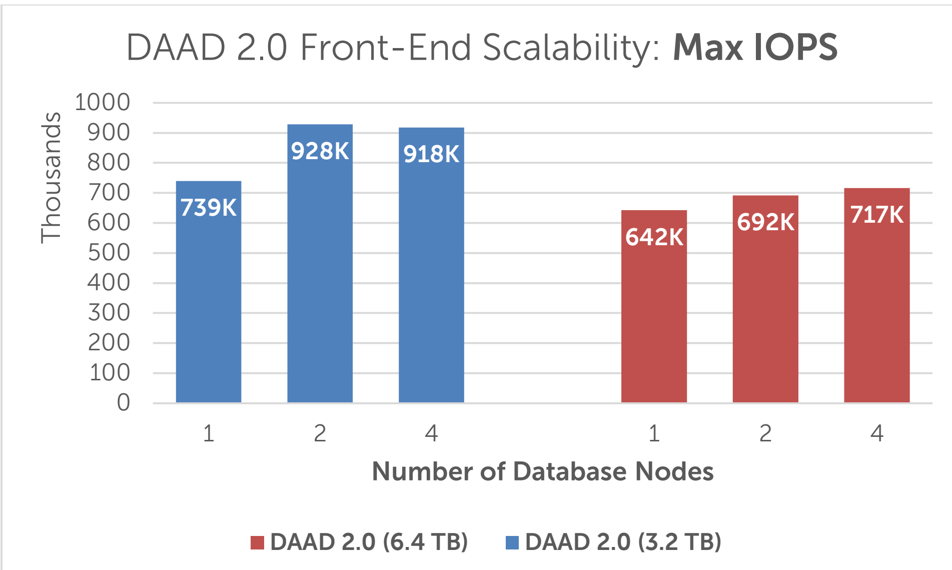

Figure 16 CALIBRATE_IO: IOPS scalability by number of database nodes

Because CALIBRATE_IO measures a combination of disk performance and database performance, it is beneficial to test how the performance scales across multiple nodes. The tests were repeated on three different configurations, with the number of database nodes changing each time. The tests were run with one database node, two database nodes, and four database nodes. With this specific benchmarking tool, we observed that the Max IOPS performance scaled going from one to two database nodes. However, we did not see much improvement going from two to four database nodes, as can be seen in Figure 16.The 3.2 TB ioMemory adapter-based DAAD was able to generate close to a million IOPS.

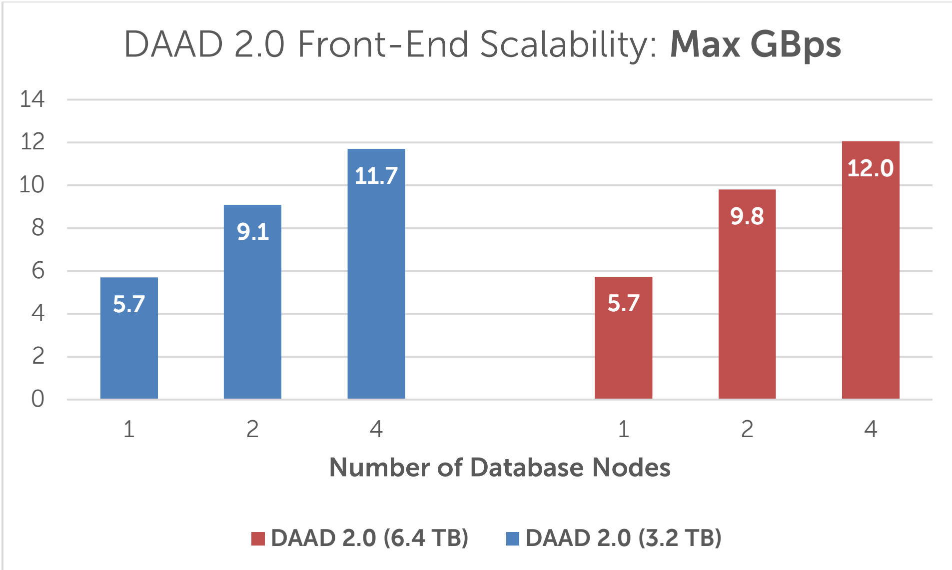

Figure 17 CALIBRATE_IO: Throughput scalability by number of database nodes

As seen in Figure 17, both the storage flavors of DAAD scaled linearly in terms of the throughput or the GBps with increasing number of database nodes. The DAAD 2.0 was able to achieve an impressive peak throughput performance of 12 GBps.

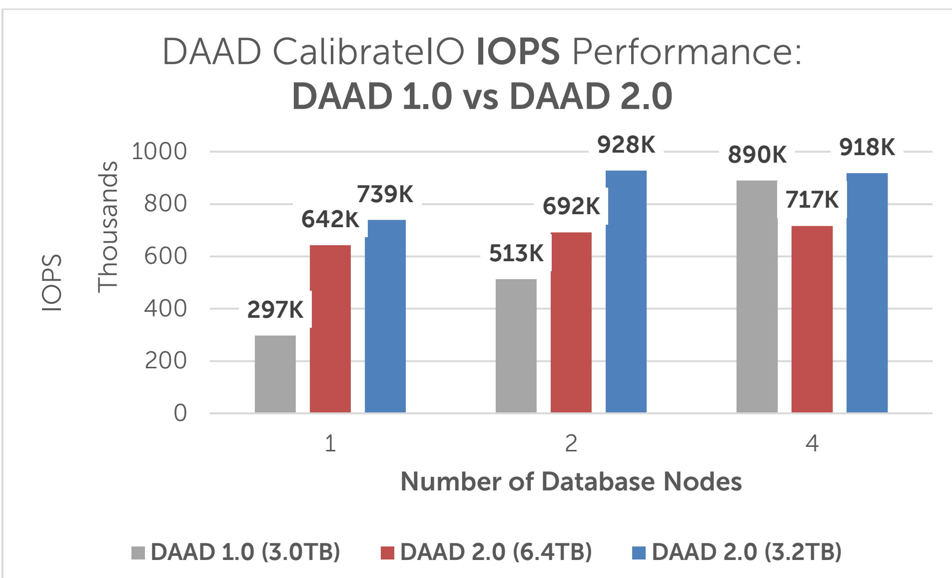

Figure 18 CALIBRATE_IO: DAAD 1.0 vs DAAD 2.0 IOPS scalability

Figure 18 shows the relative IOPS performance of DAAD 1.0 and DAAD 2.0. We observe more than double or near-double performance of DAAD 2.0 compared to DAAD 1.0 for tests run with one and two database nodes. However, the relative IOPS performance for both the generations of DAAD are near about the same with four database nodes stressing them.

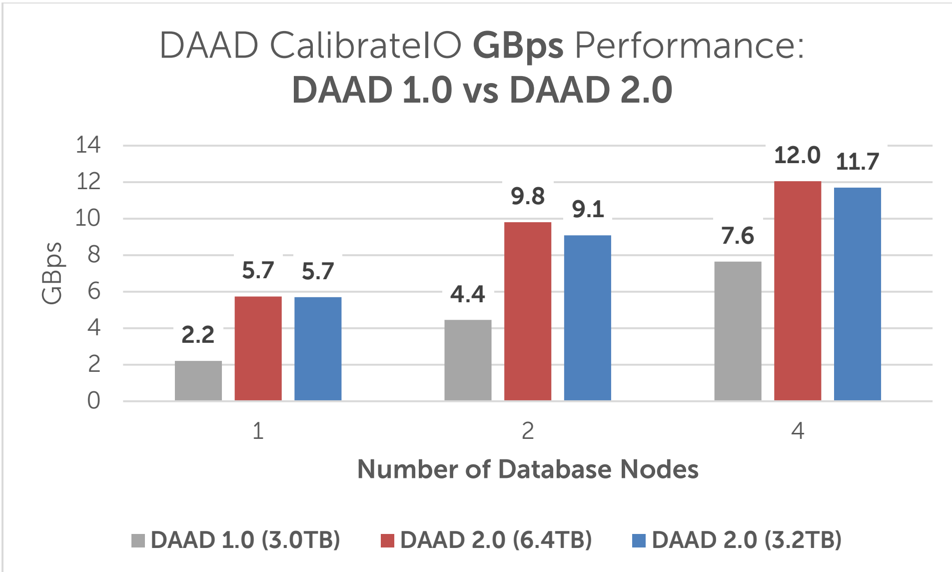

Figure 19 CALIBRATE_IO: DAAD 1.0 vs DAAD 2.0 GBps scalability

As seen in Figure 19, both the capacity flavors of DAAD 2.0 storage arrays completely outperform the first generation of DAAD in terms of throughput or GBps. With one database node stressing the storage arrays, DAAD 2.0 delivers about 2.6 times the throughput performance delivered by DAAD 1.0. Similarly, with four database nodes stressing the storage arrays, DAAD 2.0 delivers about 1.5 times the throughput performance delivered by DAAD 1.0.

7.3 HammerDB:

HammerDB allows us to simulate (though not exactly duplicate) a traditional TPC-C-like database workload and gain an understanding of how the DAAD storage arrays perform in a real world environment. Though the DAAD 2.0 arrays are capable of generating much higher performance numbers, the captured results denoted in this reference architecture were capped at an Average Response Time (ART) of in and around 10 ms.

Similar to the DAAD 1.0, DAAD 2.0 running with Oracle Database shows strong scaling in our primary application benchmark while keeping Average Response Time (ART) very low. The DAAD 2.0 equipped with the newer generation Atomic CX300 series of the ioMemory flash adapters integrated with the latest Dell PowerEdge R730 servers was able to deliver even greater database performance than the first generation of the DAAD.

Figure 20 shows the peak New Orders per Minute (NOPM) performance results delivered by both the capacity flavors of the DAAD 2.0. The graph also compares the DAAD 2.0 results with the DAAD 1.0 NOPM results. As seen in Figure 20, both the flavors of DAAD 2.0 were able to deliver more than a million NOPM with four database nodes driving the TPC-C-like workload. DAAD 2.0 was able to deliver 15% to 23% more NOPM than DAAD 1.0 while keeping the ART less than 10 ms for most of the scalability tests. The ART performance can be seen in Figure 21.

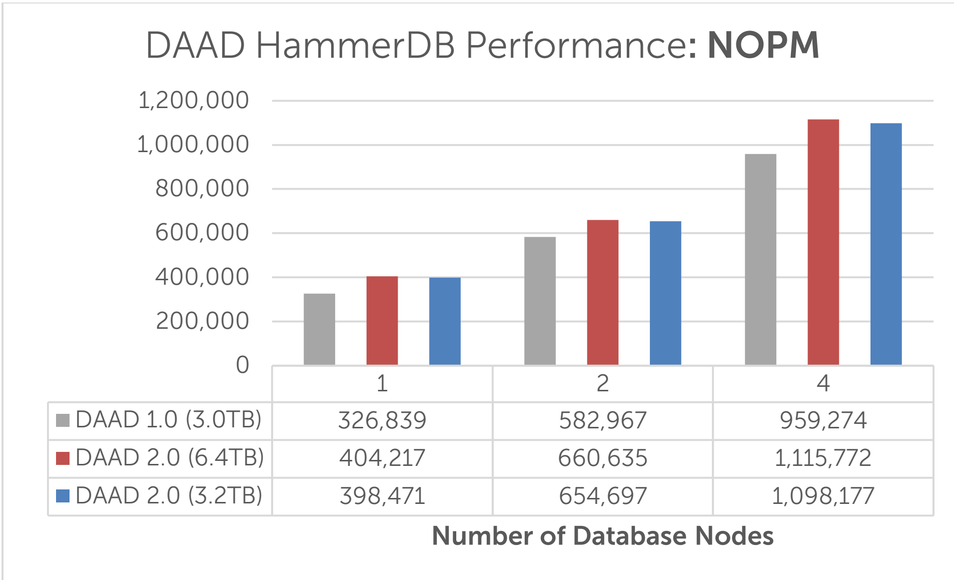

Figure 20 HammerDB: DAAD 1.0 vs DAAD 2.0 NOPM scalability

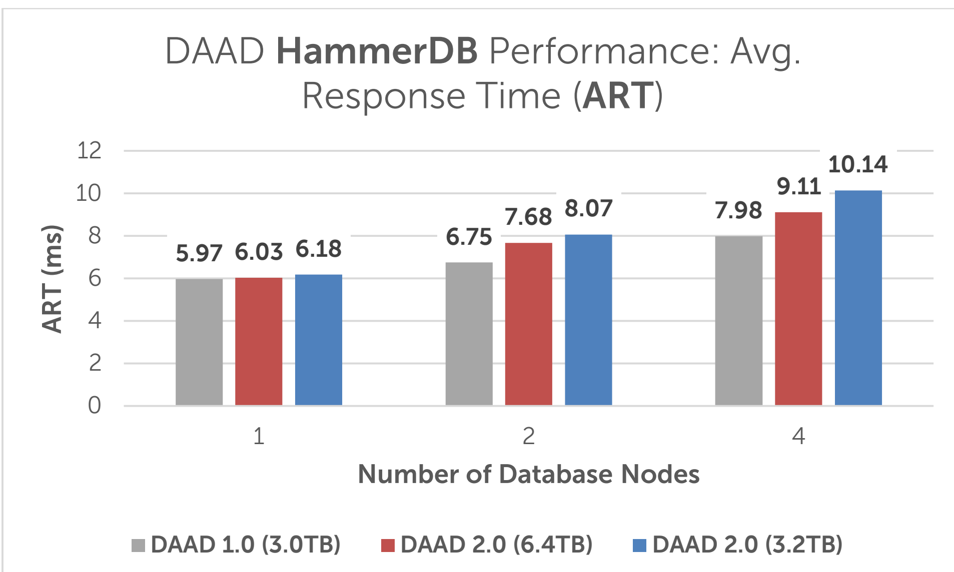

Figure 21 HammerDB: DAAD 1.0 vs DAAD 2.0 Average Response Time (ART) performance

8 Conclusion

This reference architecture is an update to the previous reference architecture with Dell Acceleration Appliance for Databases (DAAD) 1.0. It discussed two major points regarding the acceleration of Oracle Database workloads with this second generation of DAAD; that is, DAAD 2.0:

- An updated reference architecture and configuration best practices, and
- The performance results and analysis

Through the reference architecture discussion, this whitepaper demonstrated how to design and implement an Oracle 12c RAC Database with the Fibre Channel-based DAAD 2.0.

The Fibre Channel-based DAAD 2.0 demonstrated the following performance results and improvements over DAAD 1.0:

- With Iometer synthetic benchmarking, DAAD 2.0 scaled up to nearly two million IOPS for the 4K Random Read workload, which is nearly double the IOPS performance when compared to DAAD 1.0 while keeping the latency under 3 ms.
- With CalibrateIO benchmarking tool, DAAD 2.0 scaled up to 12 GBps read throughput for the 1 MB sequential read workload, which is about 1.5 times the performance of DAAD 1.0.
- With CalibrateIO benchmarking tool, DAAD 2.0 scaled up to nearly 6 GBp/s write throughput for the 1 MB sequential write workload, which is about 1.4 times the performance of DAAD 1.0.
- Throughput (GBp/s) scaled linearly with up to four database nodes stressing a pair of highly available DAAD 2.0 arrays.
- With HammerDB benchmarking tool, DAAD 2.0 scaled over a million New Orders per Minute (NOPM) for TPC-C-like workload with an Average Response Time (ART) of about 10 ms.
- About 15-20% improvement in NOPM performance over DAAD 1.0

A Configuration details

A.1 Database Server Node

Table 1 Database server components

Component	Description
Server	Dell PowerEdge R730
CPU	2 x Intel Xeon E5-2687W v3 (10 cores @ 3.1 GHz)
RAM	8 x 16 GB DDR4 RDIMM 2133MHz
HBA	2 x QLogic QLE2662 DP 16 Gb Fibre Channel
Network Cards	– Integrated Broadcom 5720 QP 1GbE rNDC – 2 x Mellanox CX3 EN DP 40 GbE Network Adapter
Operating system	Oracle Linux 6.6 running Unbreakable Enterprise Kernel (3.8.13-44.1.1.el6uek.x86_64)
Firmware version	BIOS 1.1.4
Oracle Database	Oracle Database 12c Enterprise Edition Release 12.1.0.1.0–64bit Production with Oracle Real Application Clusters and Automatic Storage Management options

A.2 DAAD Node

Table 2 Dell Acceleration Appliance for Databases components

Component	Description
CPU	2 x Intel Xeon E5-2667 v3 (8 cores @ 3.2GHz)
RAM	16 x 16GB DDR-4 RDIMM 2133 MHz
Storage (12.8 TB node)	4 x Fusion 3.2 TB ioMemory PCIe flash drives
Storage (25.6 TB node)	4 x Fusion 6.4 TB ioMemory PCIe flash drives
HBA	2 x dual-port QLogic QLE2662 16 Gb Fibre Channel HBAs
Network Cards	– Mellanox ConnectX-3 40 GbE (for HA link) – Broadcom 5720 QP 1 GbE Network Daughter Card
Operating System	DAAD ION Accelerator v2.4.1

A.3 Switches

Table 3 Switch configuration

Component	Description
Public and Private	2 x Dell Networking S6000 40 GbE Ethernet switches
Storage	2 x Brocade 6510 16 Gb Fibre Channel switches

B Reference resources

Support.dell.com is focused on meeting your requirements with proven services and support.

DellTechCenter.com is an IT Community through which you can connect with Dell Customers and Dell employees for the purpose of sharing knowledge, best practices, and information about Dell products and installations.

Referenced or recommended DAAD documents that can found on http://support.dell.com:

- Dell Acceleration Appliance for Databases 2.0 Configuration Guide
- Dell Acceleration Appliance for Databases 2.0 GUI Guide
- Dell Acceleration Appliance for Databases 2.0 Monitoring Guide
- Dell Acceleration Appliance for Databases 2.0 CLI Reference Guide
- Dell Acceleration Appliance for Databases 2.0 Compatibility Guide

C Benchmarking applications

C.1 Iometer

Iometer is an I/O testing application tool originally developed by Intel and announced in early 1998. Intel has since discontinued development on Iometer and has given the project to the Open Source Development Lab. Iometer continues to be developed by developers across the world and is consistently used by many third party storage reviewers.

Iometer provides a simple, yet powerful interface for defining access patterns. We used four access patterns to ascertain performance:

- 4k Random Read
- 8k Random 70/30 Read/Write (simulates OLTP-like workload)
- 1 MB Sequential Read (simulates OLAP-like workload)
- 1 MB Sequential Write

Each test used the same setup:

- I/Os are aligned at 4096 bytes
- Each node ran eight workers with each worker accessing a different LUN
- Each worker ran an access pattern over 5 GB of a volume
- Each node accessed data that was offset from all other nodes to ensure no I/O contention
- Each test had ten seconds for Ramp Up and five minutes for Run Time
- Workers progressed Queue Depth sequentially with values 2, 4, 8, & 16

For each test, the following metrics were measured:

- Total IOPS
- Throughput in MBps
- Latency in microseconds
- Number of Errors

The DAAD HA pair was stressed with four access patterns as shown in Table 4.

Table 4 Access specifications

Test Name	Transfer Size	Read %	Write %	Random %	Sequential %	I/O Align
4K Random Read	4 KB	100%	0%	100%	0%	4 KB
8K Random 70/30 R/W	8 KB	70%	30%	100%	0%	4 KB
1MB Sequential Read	1 MB	100%	0%	0%	100%	4 KB
1MB Sequential Write	1 MB	0%	100%	0%	100%	4 KB

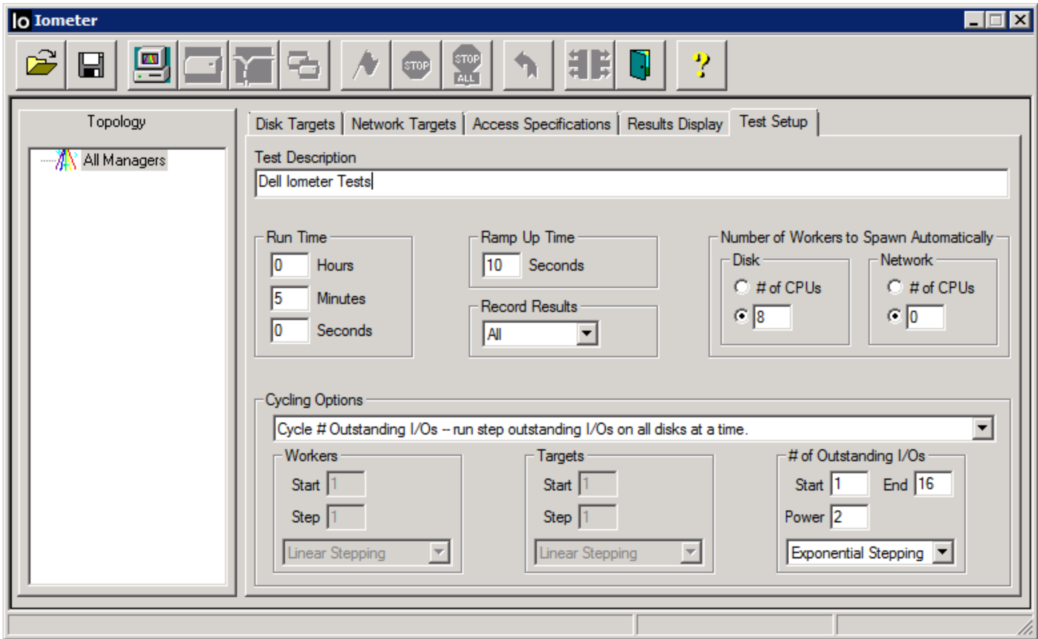

Figure 22 Test setup for Iometer

C.2 CALIBRATE_IO

CALIBRATE_IO is a procedure in Oracle Database that "calibrates the I/O capabilities of storage". This procedure tests a range of access patterns to provide these three metrics that describe the performance of the storage system:

- Max IOPS
- Max MBps
- Latency

CALIBRATE_IO uses two input parameters — <DISKS> and <MAX_LATENCY>. For the reference architecture described in this paper, <DISKS> was set to 16 and <MAX_LATENCY> was set to 10 milliseconds. <MAX_LATENCY> provides a cutoff value that any Max IOPS number must be lesser than.

The script used to run CALIBRATE_IO is:

```
set timing on
DECLARE
   lat  INTEGER;
   iops INTEGER;
   mbps INTEGER;
BEGIN
   DBMS_RESOURCE_MANAGER.CALIBRATE_IO (16, 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;
/
```

The procedure was run five times with the average for each metric reported.

For more information about running CALIBRATE_IO view Oracle's documentation at: docs.oracle.com/cd/E16655_01/server.121/e15857/pfgrf_iodesign.htm#TGDBA94384

C.3 HammerDB

HammerDB is an open source database load testing and benchmarking application tool that can run a TPC-C or TPC-H style test. The reference architecture described in this paper was tested by using only the TPC-C style test.

NOTE: HammerDB implements a workload based on the TPC-C specification. However, it does not implement a full specification TPC-C benchmark and the transaction results from HammerDB cannot be compared with the official published TPC-C benchmarks in any manner. The HammerDB implementation is designed to capture the essence of TPC-C in a form that can be run at low cost on any system bringing professional, reliable, and predictable load testing to all database environments. Therefore, HammerDB results cannot and must not be compared or used with the term tpmC. For more information about TPC-C specification, visit www.tpc.org

C.3.1 Testing Metrics

HammerDB provides three metrics to describe the performance of a system:

- New Orders per Minute (NOPM)
- Transactions per Minute (TPM)
- Average Response Time (ART)

C.3.1.1 New Orders per Minute

The TPC-C benchmark performs five different transactions with each making up a set portion of the total.

- New Order (45%)
- Order Status (43%)
- Delivery (4%)
- Stock Level (4%)
- Payment (4%)

The benchmark uses the New Order transaction as a measure of business throughput. New Orders per Minute (NOPM) is our primary metric of concern. It measures performance at the application level and is designed to be constant regardless of database size (warehouse count) and database type (Oracle, MSSQL, MySQL, and so on). Our goal is to maximize this value while keeping the Average Response Time within specific limits.

C.3.1.2 Transactions per Minute

Transactions per Minute (TPM) is a measure of User Commits and User Rollbacks as measured in the Oracle AWR report. This metric is heavily dependent on the size of the database (warehouse count) and number of nodes in the database cluster. TPM is also specific to TPC-C running on an Oracle database and is not comparable to TPC-C running on MSSQL, MySQL, and others.

C.3.1.3 Average Response Time

Average Response Time (ART) is a measure of the average time for a transaction to complete:

$$\frac{Total\ Completed\ Transactions}{Total\ Transaction\ Time}$$

ART has a low dependency on database size (warehouse count), nodes in the database cluster, or database type (Oracle vs. Non-Oracle).

C.3.2 Testing Parameters

HammerDB offers high customizability and the main parameters are outlined here.

- Warehouse Count
- Table Partitioning
- Keying and Thinking Time
- Virtual User Count
- Run Time

C.3.2.1 Warehouse Count

Warehouse count determines the size of the TPC-C database and the maximum Virtual Users supported. When the “Keying and Thinking Time” parameter is enabled, a minimum of one warehouse per ten users is required. However, as discussed below, we disabled this parameter. For our testing we sized the warehouse count by testing various counts and measuring contention wait events in the database.

We found that more than 1,000 warehouses did not decrease contention wait events while fewer than 1,000 had a negative impact on performance.

C.3.2.2 Table Partitioning

HammerDB offers the option of partitioning the Order Lines table automatically when more than 200 warehouses are used. Table partitioning was disabled for our tests.

C.3.2.3 Keying and Thinking Time

In HammerDB, Keying and Thinking Time is a parameter that can be enabled or disabled.

With Keying and Thinking Time enabled, HammerDB will emulate users entering data into the system. It does this by adding time between different transactions to represent a person taking time to key data on a keyboard and thinking between transactions. With this option disabled, HammerDB sends transactions as quickly as possible with no delay between subsequent transactions. This mode of operation is representative of application servers connecting to the database instead of individual users.

For our testing process, we disabled Keying and Thinking Time.

C.3.2.4 Virtual User Count

The number of Virtual Users used to generate load has a large impact on performance. The goal is to have sufficient users to fully stress the system but not so many that the Average Response Time (ART) is impacted. When running tests with Keying and Thinking time disabled, the number of users must correlate with the number of threads in the database servers.

Each of the R730 servers used as the database nodes in this reference architecture has 20 physical cores and 40 logical cores due to hyper-threading. There were four servers in the Oracle RAC for a total of 80 physical cores and 160 logical cores. The previous study done with the DAAD 1.0 reference architecture showed the optimal Virtual User count per core was four. So, for this reference architecture we ran a series of tests starting with four users per core to six users per core to test the performance scalability of DAAD 2.0.

We found that for this reference architecture 100 users per database node (five users per physical core) gave at least 95% of the peak performance while keeping the response time no greater than 10 ms.

C.3.2.5 Run Time

Run Time has two components: Ramp Up, and Timing Test Period. Ramp Up should allow all Virtual Users to connect to the database and start running transactions before the Timing Test Period begins. We found that after two minutes all users were started and soon after reached a steady-state operation. After rounding up, the Ramp Up time was set to three minutes.

The Timing Test Period needs to be long enough for the transaction mix to become consistent with a predetermined mix. We found that the transaction mix was within one percent of the plan after roughly three minutes of testing. This led us to use a conservative Timing Test Period of five minutes.

HammerDB was used in our testing to see how performance improved with database node scaling, and also to generate a comparison between the two generations of DAAD in a TPC-C-like transaction mix. For details about HammerDB and the metrics used to measure performance, refer to C.3 HammerDB.

D Script to map device WWIDs to aliases

map_wwid_to_ion_alias.py

```
#!/usr/bin/python
import subprocess
from os import listdir
from os.path import isfile

BLOCK = "/sys/block"
VPD = ["/usr/bin/sg_vpd", "-p0x83"]

if (isfile("/usr/bin/sg_vpd") == False):
    print "Please install sg3_utils.rpm is required to run this tool\n"
    sys.exit(1)

luns = {}
for dev in listdir(BLOCK):
    if (dev.startswith("sd")):
        wwid = None
        alias = None
        fusionio = False
        vpd = VPD + ["/dev/" + dev]
        p = subprocess.Popen(vpd, stdout=subprocess.PIPE, close_fds=True)
        stdout, stderr = p.communicate()
        if (p.returncode != 0):
            print "sg_vpd failed for device: /dev/" + dev
        else:
            for l in stdout.splitlines():
                line = l.strip()
                if ("FUSIONIO" in line):
                    fusionio = True
                if (line.startswith("0x") and fusionio):
                    wwid = "2" + line[len("0x"):]
                if ("vendor specific" in line and fusionio):
                    alias = (line.split('-'))[1]
            if ((wwid != None) and (alias != None)):
                luns[wwid] = alias

for wwid in luns.keys():
    print """
    multipath {
        wwid                    %s
        alias                   %s
    }
    """ % (wwid, luns[wwid])
```