FUSION-iO®

Oracle RAC 12c Reference Architecture with ION Accelerator

Table of Contents

Contents

Table of Contents 2

Executive Summary 3

About this Reference Architecture 3

About the ION Accelerator 4

Hardware Components and Specifications 5
Database Nodes *5*
Storage Nodes *5*
Network Components *5*

Dell PowerEdge R620 Hardware Configuration – Database Nodes 6
PCIe Slot Configuration *6*
ConnectX-3 Port Configuration *6*
BIOS Settings *6*

ION Accelerator Hardware Configuration – Storage Nodes 7
PCIe Slot Configuration *7*
ConnectX-3 Port Configuration *7*
BIOS Settings *7*

Network Configuration 8
Network Overview *8*
Network Topology *8*
Configuring the Mellanox SX6036 Switch – Private and Storage Network *8*
Configuring the Mellanox SX6036 Switch – Public Network *9*

ION Accelerator Storage Configuration 10
High Availability Overview *10*
Storage Pool Configuration *10*

Database Node Configuration 12
Multipath I/O, IRQ and NUMA Configuration *12*

Performance Verification 14

Oracle RAC and ASM Configuration 15

Scaling Database and Storage Nodes 16
Maximum Node Configurations *16*

Benchmarking the Oracle RAC Solution 17
Calibrate_IO Results *17*

Testing ION High Availability with Oracle RAC 18

Summary 19

Executive Summary

This document describes a small footprint reference architecture for Oracle Real Application Clusters 12c on Red Hat Enterprise Linux powered by Dell PowerEdge servers and Fusion ION Accelerators, all connected with Mellanox InfiniBand. This compact yet powerful reference architecture includes a 4-node Oracle RAC cluster with over 19 TB of fully redundant flash memory storage that fits in under 18 rack units and delivers in excess of 2.5 million 8K random read IOPS and 41 GB/s on long sequential reads, as measured by the application database. The half rack reference architecture is scalable to full rack and multi-rack configurations. The open concept architecture allows customers to attach 3rd party components as needed. Every aspect of the reference architecture is redundant and designed for zero downtime and zero data loss.

Hard disk based solutions with flash added as an afterthought or merely as a caching layer require many more servers and processors. This has the negative consequence of dramatically raising processor based software license fees and on-going support and maintenance fees. The Fusion-io all-flash reference architecture described here slashes processor based license fees and support fees by up to 66%, and cuts hardware costs by up to 50%, while delivering the same performance with just 64 processor cores that might otherwise require up to 192 processor cores in hard disk based systems.

A key element of this reference architecture is the ION Accelerator all flash storage appliance with the High Availability option. Each ION Accelerator in this reference architecture uses a Dell PowerEdge R720 with three Mellanox ConnectX-3 InfiniBand cards and four industry-leading Fusion ioDrive2 2.4TB flash storage devices for a total of 9.6TB of storage. Each of these ION Accelerator units fits in 2 rack units (2RU) and delivers well over 645,000 8K database IOPS and up to 12 GB/s sustained throughput, as seen by the database. Using the High Availability option ION Accelerators are deployed in clustered pairs providing full redundancy and fault tolerance. I/O traffic is intelligently balanced across the clustered ION Accelerators to provide double the IOPS and bandwidth. This reference architecture uses two pairs of ION Accelerator High Availability clusters to demonstrate performance scales as the solution expands.

The Oracle RAC nodes consist of Dell PowerEdge R620 servers. Each two-socket server consumes a mere 1U of rack space and yet is capable of pulling up to 1.4 million 8K IOPS from the ION Accelerator storage layer as measured by the Flexible IO Tester utility. The Oracle RAC nodes are connected to the ION Accelerators through redundant Mellanox switches. Each Oracle RAC node sees the ION Accelerator storage as multipath storage. The multipath devices are aggregated by Oracle ASM to create a large and powerful diskgroup. Expanding the size and performance of the database is as easy as adding more ION Accelerator devices to the ASM diskgroup.

About this Reference Architecture

This reference architecture utilizes the Fusion ION Accelerator product with High Availability option to provide a zero-downtime ultra-fast Oracle RAC 12c solution with minimal data center footprint. The half-rack solution described in this reference architecture can be scaled up or down to satisfy data processing requirements, to the full extent supported by Oracle RAC 12c.

This document is beneficial to IT Managers, Oracle DBAs and Storage Architects who are responsible for planning, designing and maintaining a high performance Oracle RAC environments for their business stakeholders. While some of the document describes important concepts and provides key metrics, this document also provides numerous and detailed performance tuning techniques that were used to optimize the I/O traffic.

For more information about how Fusion-io can accelerate your data contact Fusion-io on the Web at www.fusionio.com or by calling (800) 578-6007.

About the ION Accelerator

The ION Accelerator appliance combines ION Accelerator software, proven ioMemory flash, and industry-leading Dell R720 servers. ION Accelerator appliances are deployed across industries that need to scale application performance and includes the following features:

- ***High performance*** delivering in excess of 850,000 4K random read IOPS per ION with 56 microsecond access latency and 11GB/s bandwidth for transactional and sequential workload acceleration
- ***Efficient Density*** with incremental scalable storage of up to 32TB ioMemory capacity in 1U – 4U of space
- ***Flexible*** support for multiple applications, deployment models and storage interfaces
- ***Simple*** to deploy, configure and manage
- ***Durability*** with proven Fusion ioMemory data protection technology like Adaptive Flashback

See the ION Accelerator High Availability white paper for more information.

Hardware Components and Specifications

To make this solution a reality, industry-leading hardware from Dell, Mellanox, and Fusion-io was brought together to provide the performance and scalability in a footprint smaller than 18U.

DATABASE NODES

4 each Dell PowerEdge R620 with the following specifications:

- 1 rack unit
- 256GB RAM using 16GB RDIMM, 1333 MHz, Low Volt, Dual Rank, x4 Data Width
- 2 Intel® Xeon® E5-2667 v2 3.30GHz, 25M Cache, 8.0GT/s QPI, Turbo, HT, 8C, 130W
- 3 Mellanox ConnectX®-3 Dual-Port Adapter with VPI

STORAGE NODES

4 each Dell PowerEdge R720 with the following specifications:

- 2 rack units
- 128GB RAM – 16GB RDIMM, 1333MT/s, Low Volt, Dual Rank, x4 Data Width
- 2 Intel® Xeon® E5-2667 v2 3.30GHz, 25M Cache, 8.0GT/s QPI, Turbo, HT, 8C, 130W
- 3 Mellanox ConnectX®-3 Dual-Port Adapter with VPI
- 4 Fusion ioDrive2 Duo 2.4TB
- ION Accelerator software version 2.2 with High Availability Option

NETWORK COMPONENTS

4 each Mellanox SX6036 switches

- 36 ports Non-blocking Managed 56Gb/s InfiniBand/VPI SDN Switch System
- Mellanox Cables

Dell PowerEdge R620 Hardware Configuration – Database Nodes

PCIE SLOT CONFIGURATION

Each Oracle RAC database node is a Dell PowerEdge R620 server. This server has three PCIe 3.0 slots. For this reference architecture all three slots were fitted with Mellanox ConnectX-3 cards as detailed in the table below:

Slot Number	Slot Specification	Peripheral
1	PCIe 3.0 x16 connector with x8 bandwidth; half-length, half-height	Mellanox ConnectX-3 Dual Port Low Profile bracket
2,3	PCIe 3.0 x16 connector with x16 bandwidth; half-length, half-height	Mellanox ConnectX-3 Dual Port Low Profile bracket

Figure 1. Database Node PCIe Slot Utilization

CONNECTX-3 PORT CONFIGURATION

The figure below illustrates the port assignments within each database node. For optimal performance the ports assigned to InfiniBand traffic between the database and storage nodes were distributed across both CPU domains.

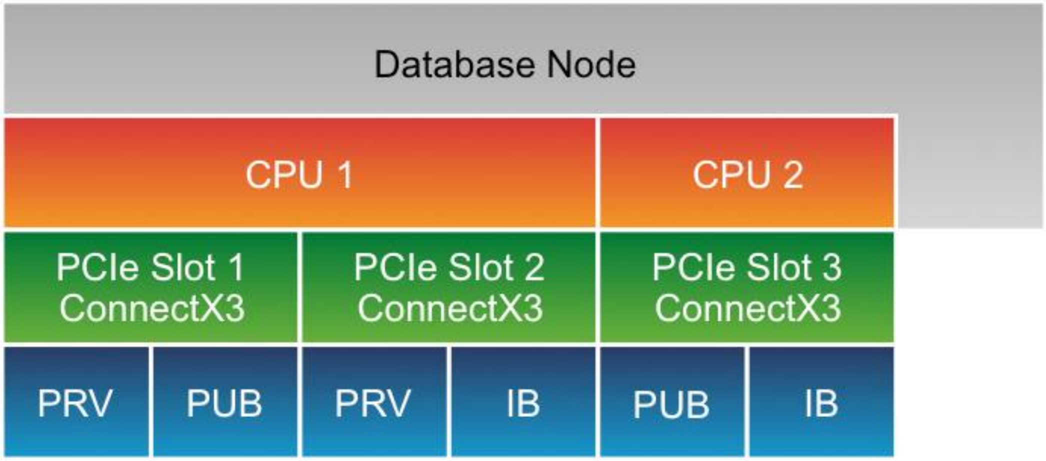

Figure 2. Overview of ConnectX-3 Port Configuration for Database Nodes (PRV – Private, PUB – Public, IB = InfiniBand)

BIOS SETTINGS

The following adjustments were made to the BIOS to increase performance:

- Applied the latest BIOS update to ensure NUMA information was correctly provided from the BIOS to Linux
- Ensured all firmware was up-to-date using the Dell LifeCycle Controller
- Set the BIOS system profile to **Performance**
- Memory profile was set to **Optimize**
- Enabled Hyper-threading
- Disabled C-States

ION Accelerator Hardware Configuration – Storage Nodes

PCIE SLOT CONFIGURATION

Each ION Accelerator storage node is a Dell PowerEdge R720 server. This server has seven PCIe 3.0 slots. For this reference architecture the first three slots were fitted with Mellanox ConnectX-3 cards and the last four slots were fitted with ioMemory as detailed in the table below:

Slot Number	Slot Specification	Peripheral
1,2,3	PCIe 3.0 x8 connector with x8 bandwidth; half-length, half-height	Mellanox ConnectX-3 Dual Port Low Profile bracket
4	PCIe 3.0 x16 connector with x16 bandwidth; half-length, half-height	Fusion-io ioDrive2 Duo 2.4TB
5,6,7	PCIe 3.0 x8 connector with x8 bandwidth; half-length, half-height	Fusion-io ioDrive2 Duo 2.4TB

Figure 3. Storage Node PCIe Slot Configuration

CONNECTX-3 PORT CONFIGURATION

The figure below illustrates the port assignments within each ION Accelerator storage node. The ports labeled “ETH” are directly connected to another ION Accelerator for high availability. The ports labeled “IB” are used to connect ION Accelerator (Storage Nodes) to the Database nodes over SCSI RDMA Protocol (SRP). SRP is the protocol used to present LUNs from ION Accelerator to the database nodes. ION Accelerator with auto enable port 2 on the ConnectX-3 in slots 1 and 2 for Ethernet mode.

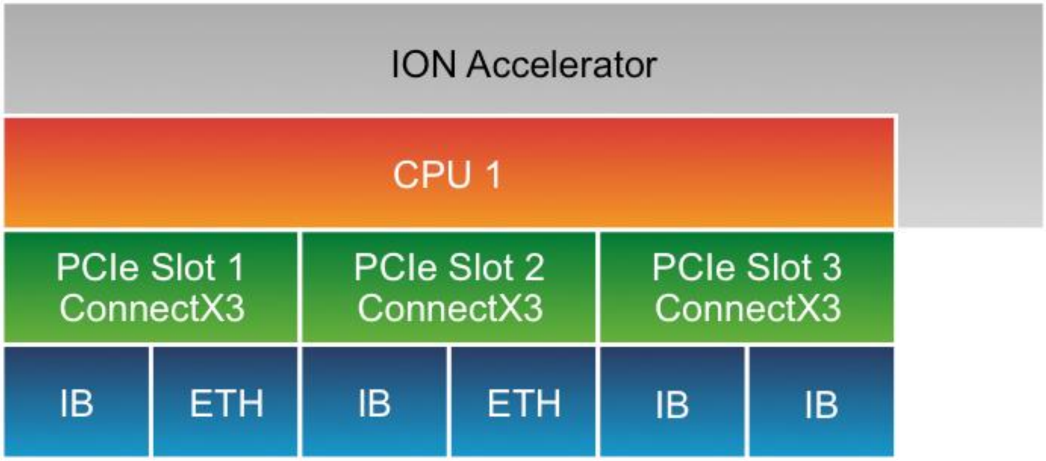

Figure 4. Overview of ConnectX-3 Port Configuration for Storage Nodes

BIOS SETTINGS

The same adjustments were made to the BIOS as previously discussed for the R620 database nodes.

Network Configuration

NETWORK OVERVIEW

Each layer in the solution has redundant components to maximize availability. Two clustered pairs of ION Accelerator appliances (four in total) are mirrored to one another over a 40GbE connection. Each ION Accelerator connects to redundant Mellanox SX6036 switches to extend high availability to the InfiniBand network. Each database node in the Oracle RAC cluster has a total of 6 HCA ports. Two of the ports are designated for InfiniBand/SRP protocol, two ports are designated for Ethernet connectivity and used for end user/downstream applications, and the remaining two ports are used for internode communication. These ports are connected to redundant Mellanox SX6036 switches.

NETWORK TOPOLOGY

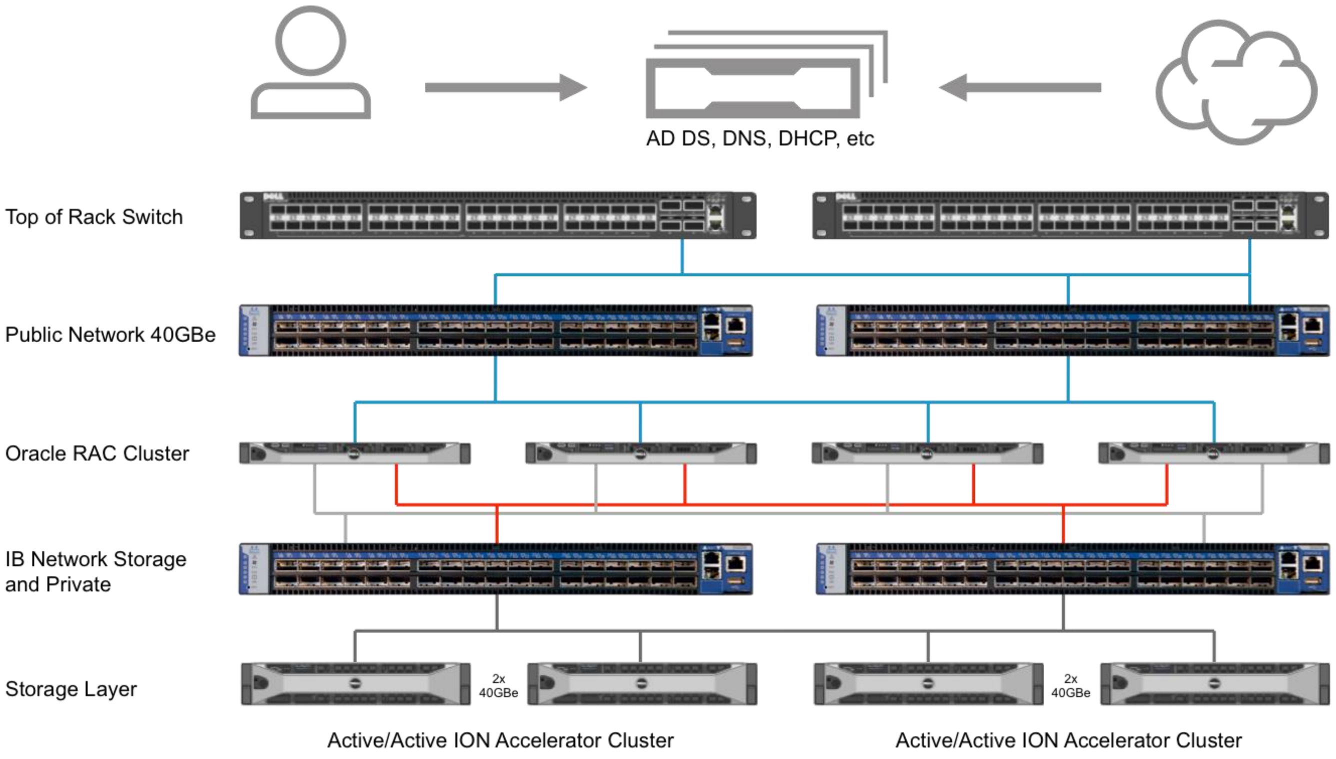

Figure 5*. Overview of Oracle RAC RA with ION Accelerator Solution Network Topology*

CONFIGURING THE MELLANOX SX6036 SWITCH – PRIVATE AND STORAGE NETWORK

The Oracle RAC private network and storage network share redundant switches for high availability as illustrated in Figure ***5***. Figure 6 shows nine 1-meter FDR10 cables connecting the switches and providing sufficient bandwidth for the solution and future growth, and the remaining ports left available for storage and private interconnect traffic.

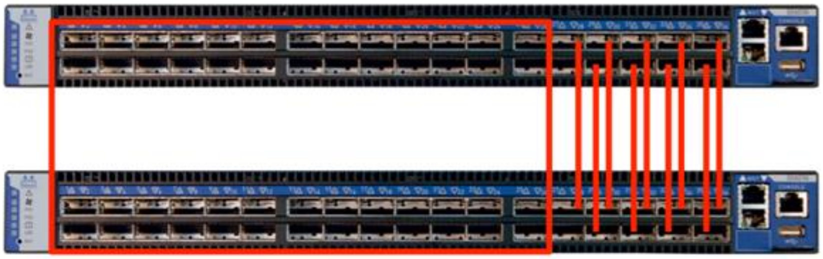

Figure 6. Ports used for storage and cluster internode communication

The subnet manager was configured for high availability and ran on the primary SX6036 switch. IPOIB was used for Oracle RAC internode communication. The MTU packet size was set to 4092.

CONFIGURING THE MELLANOX SX6036 SWITCH – PUBLIC NETWORK

Two Mellanox SX6036 switches were used for the public network. These were configured with Mellanox's gateway software, which can be used to connect the SX6036 to a top-of-rack switch to provide business users, downstream applications and services access to the solution. The public switches are connected to a pair of Dell S4810 switches via 40Gbs Ethernet uplinks.

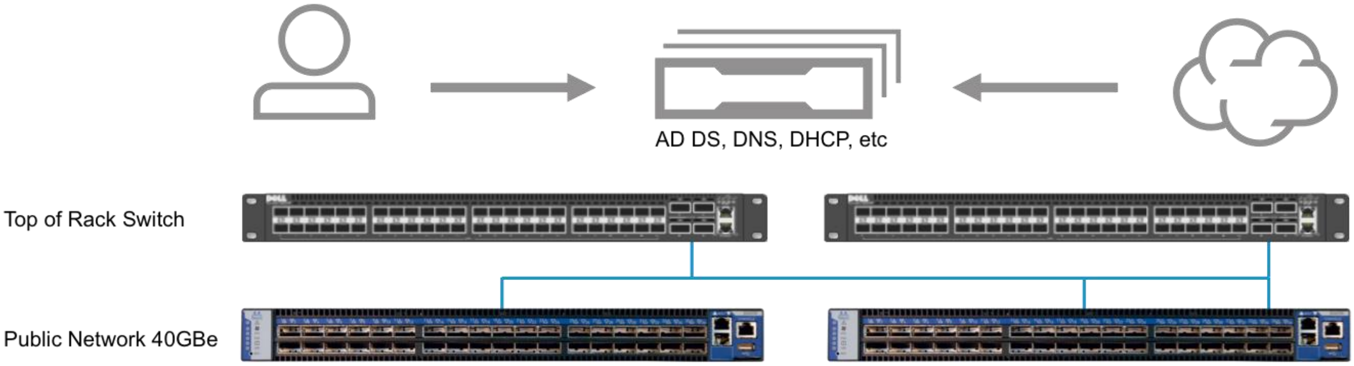

Figure 7.Ports used to connect top-of-rack switch to public network

ION Accelerator Storage Configuration

HIGH AVAILABILITY OVERVIEW

ION Accelerator enables a powerful and effective High Availability (HA) environment for shared storage, when HA licensing is enabled. ION HA clustering provides an important option for customers who prefer array-based HA over host-based mirroring. This can be especially useful if your application does not provide logical volume management, such as with all VMware environments and most implementations of Microsoft Clustering.

The diagram below illustrates the differences between ION HA and traditional host-based mirroring. The same application servers and ION storage nodes are used in both cases. The left side of the diagram shows ION providing HA by replicating all data across a private 40GbE point-to-point connection. The right side of the diagram shows host-based mirroring such as Oracle ASM Redundancy Groups, where data is made highly available by synchronously writing the same data to both ION in parallel. For this Oracle RAC reference architecture we chose ION HA over host-based mirroring. This option provided greater flexibility. For example, you can choose which storage pools to mirror for high availability.

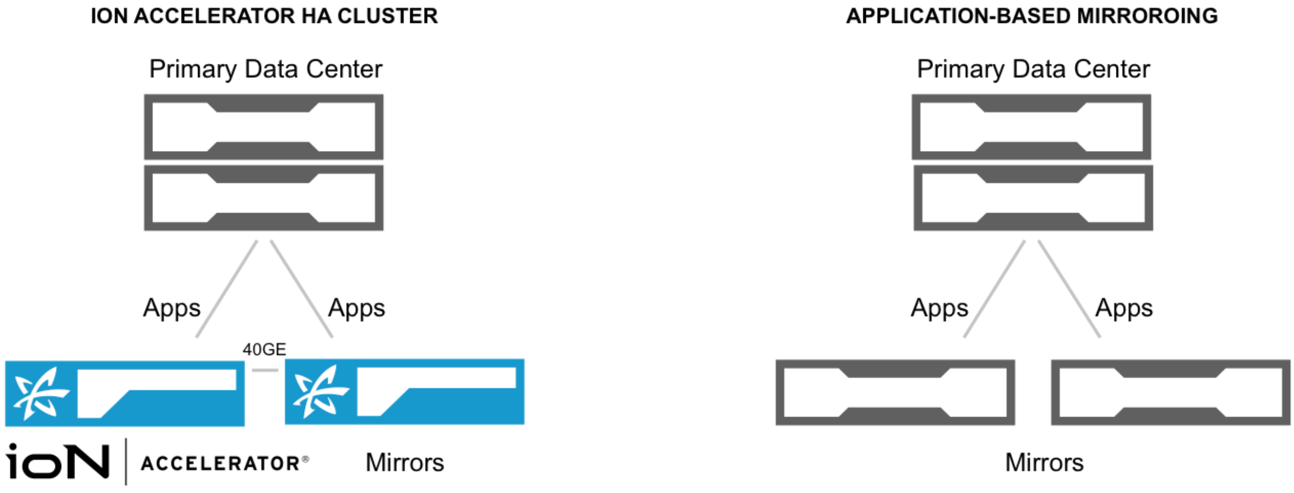

Figure 8. ION HA cluster replication versus host-based mirroring

ION HA replicates all block changes over a private 40GbE interconnect between the two units using a pair of redundant point-to-point connections. The private interconnect utilizes the existing ConnectX-3 adapters in each ION Accelerator as noted in Figure 4. In the event either ION storage node is taken offline, all data will be available from the remaining active unit ensuring users and applications continue their operations.

STORAGE POOL CONFIGURATION

Volume Configuration

Each of the four ION Accelerator storage nodes in this reference architecture has a total of eight ioMemory DIMMs: two DIMMs per ioDrive2 2.4TB Duo (fioa, fiob, fioc, fiod, fioe, fiof, fiog and fioh). Each ioMemory DIMM was given its own storage pool. Two volumes were then created from each pool using the "volume:create" command with the "-n" switch in the ION Accelerator CLI utility FIKON. The "-n" switch provides the ability to specify a primary and secondary storage node for the volume. A round robin technique was used to ensure each ION Accelerator would have primary responsibility for an equal number of volumes, thus ensuring maximum bandwidth and IOPS.

When creating a primary pool, ION HA automatically places the mirror on the secondary ION. ION uses the same ioMemory device name and pool name when creating the mirror on the secondary unit.

Figure 9 shows basic LUN access (exported volumes) in an HA configuration:

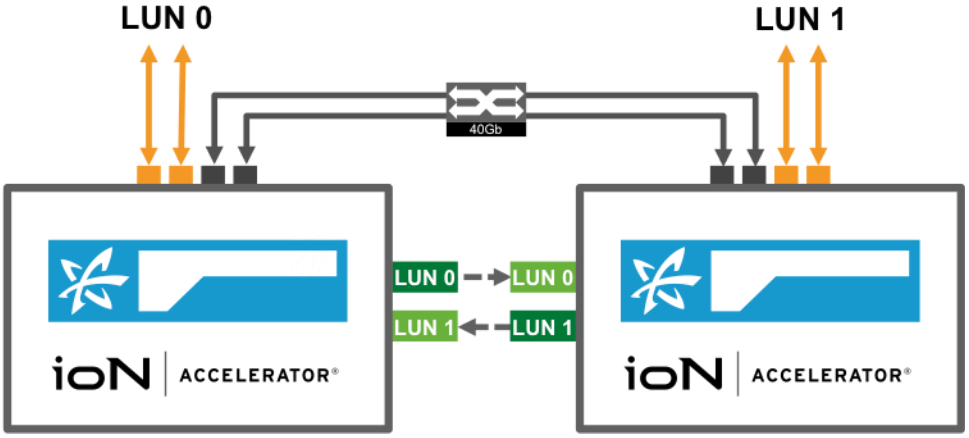

Figure 9. Basic LUN access in an ION HA configuration

In this simplified example, ION Unit 1 presents LUN 0 to the application, while ION Unit 2 presents LUN 1 to the application. Each LUN is visible to all database nodes. Writes to LUN 0 are sent to ION Unit 1, which synchronously replicates the writes to ION Unit 2. In the event of a path failure all operations on LUN 0 are automatically redirected to ION Unit 2. Similarly, all writes to LUN 1 are sent to ION Unit 2, which synchronously replicates the writes to ION Unit 1, but in the event of a path failure all operations to this LUN are automatically redirected to ION Unit 1. When a path redirection occurs all data operations are still permitted. When an ION is taken offline, such as for routine maintenance, all paths to that ION are redirected the remaining active ION which records all block changes. When the ION resumes operations all pending block changes are applied and the paths restored to their original state automatically.

Because ION HA stores two copies of all blocks usable capacity is 50% of the raw capacity of each cluster pair. For this reference architecture, each ION Accelerator HA cluster has a raw capacity of 19.2TB, leaving 9.6TB usable. The solution leverages two ION Accelerator HA clusters for a total of 19.2TB usable capacity.

Database Node Configuration

MULTIPATH I/O, IRQ AND NUMA CONFIGURATION

Fusion-io provides an open source script to tune interconnects on the database nodes. This RPM file can be downloaded from http://www.fusionio.com/files/ion-optimization-scripts free of charge.

The ION tuner (iontuner) file makes the following changes:

Tuning Block Devices

- Sets the I/O scheduler to noop, which reduces latency and CPU usage when queuing I/O to the backend storage compared to the default scheduler, CFQ, which is tuned for traditional disk I/O.
- Sets the I/O request queue affinity to level 2, which forces I/O to complete on the same CPU where the I/O request was initiated.
- Disables disk entropy contribution.

IRQ Pinning

To minimize data transfer and synchronization throughout the system, I/O interrupts were handled on the socket local to their PCIe device. To provide the best results, the irqbalance daemon was disabled.

IRQs were pinned for each driver that handles interrupts for ION Accelerator device I/O as well as for the InfiniBand drivers that support the Oracle RAC private interconnect. Driver IRQs were identified in /proc/interrupts by the matching IRQ numbers to the driver prefix listed in the same row. The IRQs were distributed between the CPU cores within the device's local socket.

Multipath

To configure multipath for ION Accelerator HA, the following parameters and values were added to file /etc/multipath.conf:

- Defaults
  - `user_friendly_names    no`
  - `queue_without_daemon   no`
- Devices
  - Device
    - `vendor              "FUSIONIO"`
    - `features            "3 queue_if_no_path pg_init_retries 50"`
    - `no_path_retry       3`
    - `hardware_handler    "1 alua"`
    - `path_grouping_policy  group_by_prio`
    - `path_selector       "queue-length 0"`
    - `failback            immediate`
    - `path_checker        tur`
    - `prio                alua`

A sample multipath file is shown below. The information from this same can be added to file /etc/multipath.conf and enabled by restarting the multipathd service. The list of multipaths in this example is empty and will be generated by the iontuner package.

```
defaults {
      user_friendly_names        no
      queue_without_daemon       no
}
devices {
      device {
            vendor               "FUSIONIO"
            features             "3 queue_if_no_path pg_init_retries 50"
            no_path_retry        3
            hardware_handler     "1 alua"
            path_grouping_policy group_by_prio
            path_selector        "queue-length 0"
            failback             immediate
            path_checker         tur
            prio                 alua

            # Uncomment if using FC. Do not use for SRP and iSCSI
            #fast_io_fail_tmo    15
            #dev_loss_tmo        60
      }
}
multipaths{

}
```

Additional ION Tuner Options

In addition to IRQ configuration and tuning, ION Tuner provides options to help configure multipathing, Flexible I/O (FIO) jobs, Oracle ORION jobs, and more.

InfiniBand SRP Backport RPM

Fusion-io has developed a RPM that provides higher performance and shorter failover time than what is provided by the SRP initiator included in their enterprise Linux distribution or than the SRP initiator available in one of the various OFED versions. The ib_srp-backport project includes the following additional features:

- Faster failover.
- Configurable dev_loss_tmo, fast_io_fail_tmo and reconnect_delay parameters.
- Support for initiator-side mirroring (by setting dev_loss_tmo to "off").
- Support for Mellanox ConnectX-3 InfiniBand HCA's (via fast registration work requests, a.k.a. FRWR).
- Configurable queue size allowing higher performance against hard disk arrays.
- Improved small block (IOPS) performance.
- Support for lockless SCSI command dispatching on RHEL 6.2 and later.
- Builds against any kernel in the range 2.6.32.3.11.
- RHEL 5.9, RHEL 6.x, SLES 11, Ubuntu 10.04 LTS, Ubuntu 12.04 LTS, OL 6.x, OL UEK 6.x CentOS 6.x and SL 6.x systems are supported.

Your Fusion-io Sales Team can provide access to the RPM packages.

Performance Verification

Before proceeding with the Oracle RAC installation it is important to capture synthetic performance of the solution to verify configuration consistency across all databases nodes, ION Accelerator storage nodes, and network components. For these tests, we used the open source Flexible I/O Tester "FIO" to run various benchmarks against the database nodes. More information on FIO can be found here: http://freecode.com/projects/fio.

You can use the ION tuner command which is installed by the ION optimization script referenced in the Data Node Configuration section to generation FIO job files. Execute "iontuner -help" to see an example of how to generate a FIO job file.

Each database node and combination of database nodes was benchmarked to verify raw performance when reading from all four ION Accelerator storage nodes. Each database node has a 3 dual port Mellanox ConnectX-3 HCA adapters (6 ports total) of which 2 ports are designated for storage traffic. The three block sizes most commonly used with Oracle were tested, and the results recorded for throughput and IOPS. The table below shows the test results. IOPS are not relevant and were not reported for 1M block reads.

# of Nodes	# of ION Nodes	Block Size	Read GB/s	Read IOPS
1	4	8K	11.0	1,400,000
1	4	32K	11.4	365,000
1	4	1M	11.5	N/A
2	4	8K	23.7	2,970,000
2	4	32K	24.4	763,000
2	4	1M	24.5	N/A
3	4	8K	32.1	4,000,000
3	4	32K	36.0	1,100,000
3	4	1M	36.0	N/A
4	4	8K	31.7	4,069,000
4	4	32K	44.0	1,380,000
4	4	1M	44.0	N/A

Figure 10. Flexible IO Test Results When Varying # of Database Nodes and Block Sizes

Note: Synthetic performance numbers do not translate into real world application performance. The purpose of this exercise is to verify the storage configuration.

Oracle RAC and ASM Configuration

This reference architecture uses Oracle Automatic Storage Management (ASM) to store all database files and the Oracle Cluster Registry (OCR) files. The ASM configuration consists of one diskgroup named "DATA" configured with External Redundancy. All ION volumes are members of this diskgroup. All ASM attributes use default values. Redundancy is managed by the ION Accelerator HA option. The Fast Recovery Area (FRA) diskgroup can optionally be added to the configuration as needed using ION or non-ION storage.

The RAC configuration uses default values for all settings except for the following:

- The ASM disk discovery path was changed to /dev/mapper/*
- The ASM diskgroup redundancy was changed to External

The DBCA was run in Advanced Mode to access the Custom Database template, and the following non-default selections were made when creating the database:

- The extra cost OLAP and Spatial options were deselected
- Multiplexing was enabled for redo logs, with both copies on +DATA
- The SGA was sized to 16 GB
- The PGA was sized to 16 GB
- The character set was changed to AL32UTF8
- The number of processes was changed to 500

After the database was created the following changes were made:

- The on-line redo logs were replaced with larger files. The final configuration has five groups of redo logs with two members per group, and each member is 5 GB.
- The TEMP tablespace was extended by adding four additional files.
- LGWR was added to the list of high priority processes

Note: We intentionally set the SGA size to 16GB to force work on the I/O subsystem. Production users can increase the SGA to as much as 50% of the physically installed RAM. Each of the Dell PowerEdge R620 database nodes in this reference architecture has 256 GB RAM.

Scaling Database and Storage Nodes

Performance of the Oracle RAC database nodes can be scaled through additional Dell PowerEdge R620 servers. Each database node requires three Mellanox ConnectX-3 HCAs. Adding a database node requires two additional ports on each Mellanox SX6036 switch dedicate to the private and storage network for redundancy. Each database node requires one additional port on each public network switch. The Mellanox SX6036 switches used for Oracle RAC internode connectivity and ION Accelerator have a total of 27 open ports which can be used for scaling. The Mellanox SX6036 switches used for connectivity to infrastructure services and downstream applications have 32 open ports per switch that can be used to scale.

Scaling storage requires adding pairs of ION Accelerator HA storage nodes. Each clustered pair of ION Accelerator HA storage nodes adds up to 24GB/s of bandwidth and 9.6TB of fully redundant and usable storage. Each additional ION HA cluster requires six additional InfiniBand ports connected to the Mellanox SX6036 storage switches and 4 additional ports for cluster replication.

MAXIMUM NODE CONFIGURATIONS

The reference architecture requires pairs of ION Accelerator HA nodes, and a minimum of two Oracle RAC nodes. The maximum number of ION and RAC nodes per full rack is limited by the number of switch ports. There are 72 ports in a default configuration, although this could easily be increased in increments of 36 ports. Based on the number of ports used by each ION and RAC node, and the number of ports used to connect each switch, we can use the following formula to calculate the maximum number of ION and RAC nodes that can be deployed without additional switches:

$$6r + 5i \leq 68$$

Plotting this formula shows the reference architecture can accommodate many permutations to suit each customer’s requirements. As shown in the chart below, the reference architecture supports seven combinations of RAC and ION nodes that can be accommodated within a single rack ranging from a 3-node RAC with 10 ION nodes to maximize storage, all the way to a 9-node RAC with just 2 ION nodes to maximize user scaling. The most balanced full rack configuration uses six RAC nodes and six ION nodes, which would support roughly 3.87 million 8K IOPS and 66 GB/s.

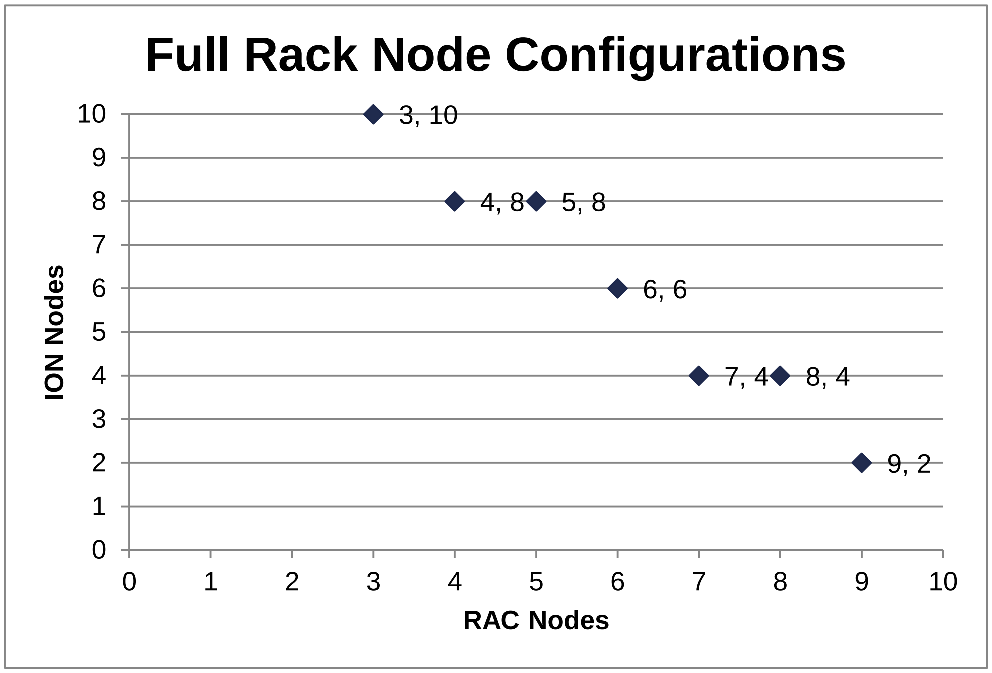

Figure 11. Possible ION and RAC node configurations per-rack

Note: The above discussion addresses this reference architecture only. It assumes a single rack with two switches between the RAC and ION nodes. It is possible to scale the number of switches and full racks. Also note external storage and tape library systems can be added as needed.

Benchmarking the Oracle RAC Solution

Benchmarking was performed on the Oracle RAC solution using the Oracle-supplied stored procedure CALIBRATE_IO, and the open source application HammerDB. This benchmarking was necessary to validate the Oracle RAC installation and configuration and was not intended to demonstrate the maximum performance of an Oracle database. Two databases were created with different block sizes (8K and 32K respectively). CALIBRATE_IO was executed against both databases multiple times and the results recorded. Next, HammerDB was used to execute an OLTP workload against both databases, and a Decision Support System (DSS) workload against the database with the 32K block size. HammerDB results are not provided in accordance with Oracle licensing requirements.

The table below contains the averaged results of three CALIBRATE_IO executions for each database block size.

Calibrate_IO Parameters	Database Block Size	Latency	Max IOPS	Max MBps
Number of Disks128 **Max Latency 10**	8192	0	2,578,000	41,002
Number of Disks =128 **Max Latency = 10**	32768	0	1,354,000	41,636

Figure 12. Calibrate IO Averaged Results per Block Size

Testing ION High Availability with Oracle RAC

To demonstrate the ION Accelerator High Availability feature, a series of hardware failures were simulated while a database workload was running. The HammerDB application was used to place the Oracle RAC database under a heavy multi-user workload throughout all tests. The application, database, and I/O throughput were monitored continuously during each test.

The first test verified all data remains available during a planned outage of an ION storage node. While a heavy database workload was in progress one of the ION storage nodes was shutdown. All I/O was suspended for approximately 30 seconds while the paths were failed over, and then normal I/O resumed. The Oracle RAC database and HammerDB applications showed no errors, and all transactions continued normally once the paths had been failed over.

The second test simulated an unplanned outage where one of the ION Accelerator storage nodes abruptly lost power. The results were identical to those of the planned outage test described earlier.

The third test simulated the unplanned loss of individual paths between the Oracle RAC nodes and the ION Accelerator storage nodes by unplugging individual cables while the database was under load. It was noticed that throughput decreased for approximately 30 seconds until the path had been failed over, but overall system processing continued. Database operations not utilizing the failed path continued, and operations using the failed path paused and then resumed on the new path. No errors were reported by either Oracle RAC or the HammerDB application.

The results demonstrate the ION Accelerator product's ability to provide continuous availability during hardware and/or infrastructure unplanned failures and planned maintenance activities.

Summary

Smaller, faster, lower cost! The all-flash half rack reference architecture based on industry leading technology from Fusion-io, Dell, and Mellanox outperforms larger systems while reducing the overall rack footprint by 50% and reducing core-based license costs by up to 66% as summarized in Figure 13:

	Oracle RAC with ION Accelerator® Solution	Comparable System #1	Comparable System #2
Rack Size	<18U	Full	Full
Cores	64	128	192
Database Servers	4	8	8
Storage Servers	4	14	14
Mirrored Data File Flash Storage	19.2 TB	0	0
Mirrored Data File Disk Storage	0	20 TB	40TB
Read IOPS	2,578,000	1,500,000	2,660,000
Oracle Core Based License Cost	1x	2x	3x
IOPS per Oracle License Cost	2,578,000	750,000	887,000

Figure 13. Up to 66% Cost Reduction 50% Smaller Footprint

The storage capacity of this reference architecture is based on 2.4TB ioDrive2 Duo cards, with four cards per ION and two sets of ION HA clusters with full data redundancy. Fusion-io offers per-card capacities ranging from 365GB to 10.2TB. The storage capacity of the comparable systems is based on published specifications where 60% of the usable capacity is allocated for recovery files and the remaining 40% mirrored down to the sizes listed in the table.

For more information about how Fusion-io can accelerate your data contact Fusion-io on the Web at www.fusionio.com or by calling (800) 578-6007.