ORACLE

ZFS STORAGE APPLIANCE

An Oracle Technical White Paper
December 2013

Configuring a Single Oracle ZFS Storage Appliance into an InfiniBand Fabric with Multiple Oracle Exadata Machines

A configuration best practice guide for implementing a single Oracle ZFS Storage Appliance into an InfiniBand fabric with Oracle Exadata machines, highlighting several architectures.

ORACLE

Introduction

Database, system, and storage administrators are faced with a common dilemma when it comes to backup and recovery of Oracle Exadata databases—how to back up more data, more often, in less time, and within the same budget. The Oracle ZFS Storage Appliance helps administrators meet these challenges by providing a cost-effective and high-bandwidth storage system that combines the high-bandwidth of the InfiniBand network protocol with ZFS-enhanced disk reliability. By integrating the Oracle ZFS Storage Appliance into an Exadata InfiniBand fabric architecture, administrators can reduce the capital and operational costs associated with data protection while maintaining strict service level agreements with end customers.

The Oracle ZFS Storage Appliance is an easy-to-deploy unified storage system uniquely suited for protecting data contained in the Oracle Exadata Database Machine. With native Quad Data Rate (QDR) InfiniBand (IB) connectivity, the Oracle ZFS Storage Appliance is an ideal match for the Exadata internal InfiniBand fabric mesh. The high bandwidth of InfiniBand interconnects reduces backup and recovery time, as well as reduces backup application licensing and support fees, compared to traditional NAS storage systems. When combined with the incremental update backup technology of the Oracle Recovery Manager (Oracle RMAN), Oracle ZFS Storage solutions deliver increases in storage efficiency that can further reduce recovery time and simplify system administration.

When deploying an Oracle ZFS Storage Appliance for backup and recovery of Oracle databases in a multirack Exadata environment, backup window and recovery time objectives (RTO) must be met to ensure timely recovery in the event of a disaster. This paper describes best practices for setting up the Oracle ZFS Storage Appliance for optimal backup and recovery of Oracle databases and includes specific tuning guidelines for integration into an Oracle Database Exadata architecture.

This white paper describes the technical aspects of integrating the Oracle ZFS Storage Appliance into a multirack Exadata architecture. Specifically, it provides an implementation guide to configure the Oracle ZFS Storage Appliance to access an Exadata flexible InfiniBand network, providing a cost-effective RMAN Oracle Database backup and recovery solution.

Highlighted in this paper are:

- Overview of Oracle Exadata with the Oracle ZFS Storage Appliance
- Configuration guide for Exadata data protection with the Oracle ZFS Storage Appliance
- Best practices for configuring multiple Exadata machines to the Oracle ZFS Storage Appliance for database backup and recovery
- Implementation guidelines for using the Oracle ZFS Storage Appliance with multiple Oracle Exadata machines in both a single and multiple subnet InfiniBand fabric

NOTE: References to Sun ZFS Storage Appliance, Sun ZFS Storage 7000, and ZFS Storage Appliance all refer to the same family of Oracle ZFS Storage Appliance products. Some cited documentation or screen code may still carry these legacy naming conventions.

About Oracle ZFS Storage Appliance

The basic architectural features of Oracle ZFS Storage Appliance are designed to provide high performance, flexibility and scalability. The Oracle ZFS Storage Appliance provides multiple connectivity protocols for data access, including: Network File System (NFS), Common Internet File System (CIFS), Internet Small Computer System Interface (iSCSI), InfiniBand (IB), and Fibre Channel (FC). It also supports the Network Data Management Protocol (NDMP) for backing up and restoring data. The Oracle ZFS Storage Appliance architecture also offers the Hybrid Storage Pool (HSP) feature, in which direct random access memory (DRAM), flash and physical disks are seamlessly integrated for efficient data placement (see Figure 1).

A powerful performance monitoring tool called DTrace Analytics provides details about the performance of the various components, including network, storage, file systems, and client access. The tool also offers plenty of drill-down options that allow administrators to monitor specific rates of latency, size of transfer, and utilization of resources. The Oracle ZFS Storage Appliance provides a variety of RAID protections to balance the capacity, protection, and performance requirements of the applications, databases, and virtualized environments.

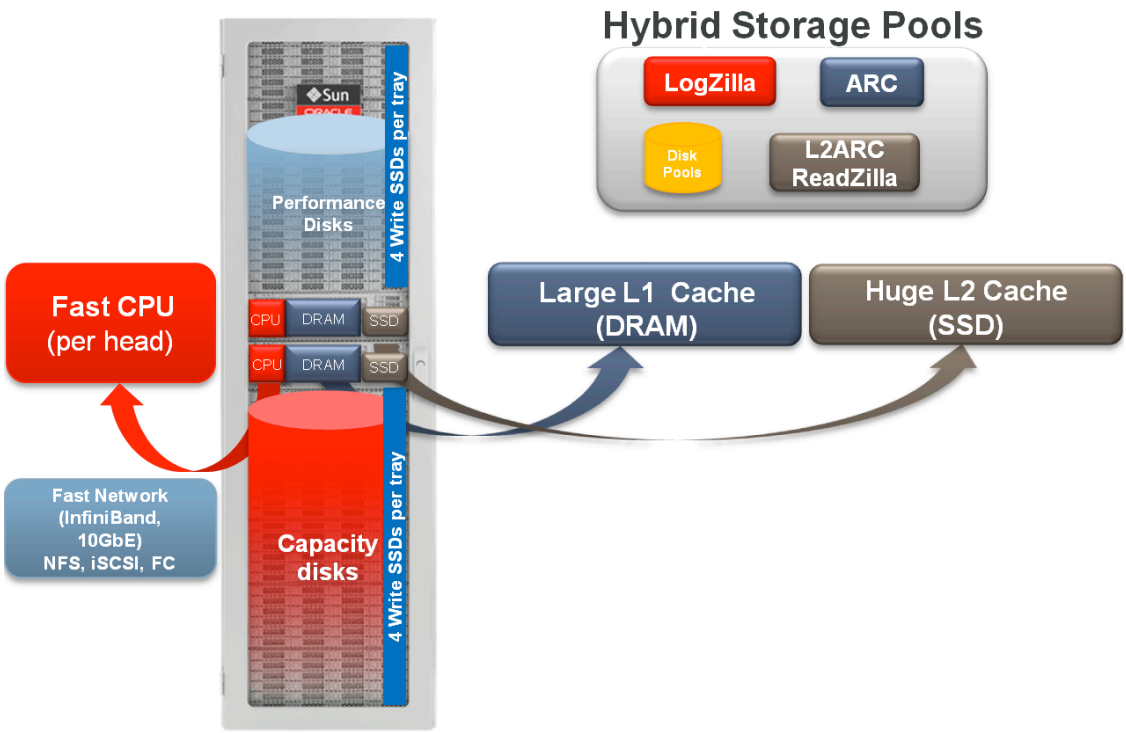

Figure 1. Oracle ZFS Storage Appliance – architecture overview

Overview of Oracle Exadata with the Oracle ZFS Storage Appliance

Oracle Exadata is a complete package of software, servers, storage, and networking, including a high-bandwidth private IB network. Oracle Exadata includes the following technologies to optimize data protection systems based on the Oracle ZFS Storage Appliance:

- QDR IB fabric provides 40 gigabits (Gb) of bandwidth per port between the database servers, storage cells, and the Oracle ZFS Storage Appliance.
- Oracle Recovery Manager (Oracle RMAN) is a native backup and recovery tool for Oracle databases that simplifies backup and recovery administration to the Oracle ZFS Storage Appliance.
- Oracle Direct NFS (dNFS) is an optimized NFS client for Oracle databases that provides a high-bandwidth solution for data transfer with the Oracle ZFS Storage Appliance.
- Backup and restore operations can be automatically parallelized across all database nodes, Exadata storage cells, and Oracle ZFS Storage Appliance interfaces and controllers for maximum scalability and throughput.
- Oracle RMAN incrementally updated backups allow incremental backups to an Oracle ZFS Storage Appliance to be converted to full backups to reduce restore time and efficiently refresh and deploy snapshots and clones of production data.
- An Oracle RMAN clone can produce a complete copy of any database stored in Oracle Exadata on the Oracle ZFS Storage Appliance.

The combined features of Oracle Exadata and the Oracle ZFS Storage Appliance provide IT administrators with advanced tools to meet nearly arbitrary recovery point objectives (RPO) and recovery time objectives (RTO) and ensure business continuity requirements are met.

Overview of System Components for this Architecture

The following tables describe the hardware configuration, operating systems, and software releases utilized by this white paper.

Table 1 shows the minimum recommended hardware components for the architecture.

TABLE 1. HARDWARE USED IN REFERENCE ARCHITECTURE

EQUIPMENT	QUANTITY	CONFIGURATION
Database Server-Storage	Two or more Oracle Exadata racks residing on one or more subnet(s)	Oracle Exadata Database Machine X2-2 or X3-2 half rack (or full rack) Four (8) x Sun Fire X4170 M2 Servers with 2 × six-core Intel Xeon X5675 processors (3.06 GHz); 96 GB memory (expandable to 144 GB with optional memory expansion kit; 4 x 300 GB 10,000 RPM serial attached SCSI (SAS) disks; 2 x QDR (40Gbit/s) ports, and 2 x 10 Gb Ethernet ports based on the Intel 82599 10GbE controller Seven (14) x Oracle Exadata Storage Servers
Network	Six or more	InfiniBand switch with 36 ports
Primary Storage	Two	Oracle ZFS Storage ZS3-4 clustered DRAM – 512 GB Four x 24 - 2 TB SAS-2 disk trays Two x InfiniBand QDR (40 Gb) dual-port host channel adapters (HCAs) per Oracle ZFS Storage ZS3-4 head (Four x InfiniBand QDR (40 Gb) dual-port HCAs per Oracle ZFS Storage ZS3-4 head if multiple subnets are configured)

Other hardware configuration considerations for the Oracle ZFS Storage Appliance include:

Back-end SAS configuration:

- Two SAS-2 host bus adapters (HBAs) per head for a two-tray or four-tray configuration.
- Three SAS-2 HBAs per head for a six-tray configuration.
- Four SAS-2 HBAs per head for an eight-tray configuration.

Write flash configuration: Write-optimized flash may be omitted since this system's functionality is for backup and restore of backup sets only.

Number of trays: Two to eight, depending on throughput and capacity requirements.

Drive size: High capacity option (3 terabytes [TB]).

Read flash configuration: Optionally, two to four read-optimized flash devices per head for systems that support additional processing, such as development or Quality Assurance (QA).

Architecture of Multiple Oracle Exadata Machines to a Single Oracle ZFS Storage Appliance

The architecture outlined in this document is designed to ensure the Oracle ZFS Storage Appliance is properly cabled and configured to access the entire Exadata InfiniBand fabric, providing a fully redundant, high-availability backup and recovery solution.

To take full advantage of the high bandwidth InfiniBand fabric, and to provide the most configuration flexibility, the recommended hardware components include two Oracle Exadata racks (full or half) configured on the same InfiniBand subnet and a clustered Oracle ZFS Storage ZS3-4 with two InfiniBand host channel adapters (HCAs) per head.

Depending on the throughput needs and the backup/recovery windows, the Oracle ZFS Storage Appliance can be configured to have access to the entire InfiniBand network and can be utilized as the backup/recovery storage for every Exadata within the InfiniBand subnet. With additional InfiniBand HCAs, the Oracle ZFS Storage Appliance can be configured to access two infiniBand networks, furthering the flexibility of the solution.

Additionally, if maximum redundancy and high availability are not requirements, the Oracle ZFS Storage ZS3-4 cluster can be configured with one dual-port HCA per head.

Multirack Exadata on a Single Subnet InfiniBand Fabric

Figure 2 shows how an Oracle ZFS Storage Appliance can be integrated into a single subnet InfiniBand fabric.

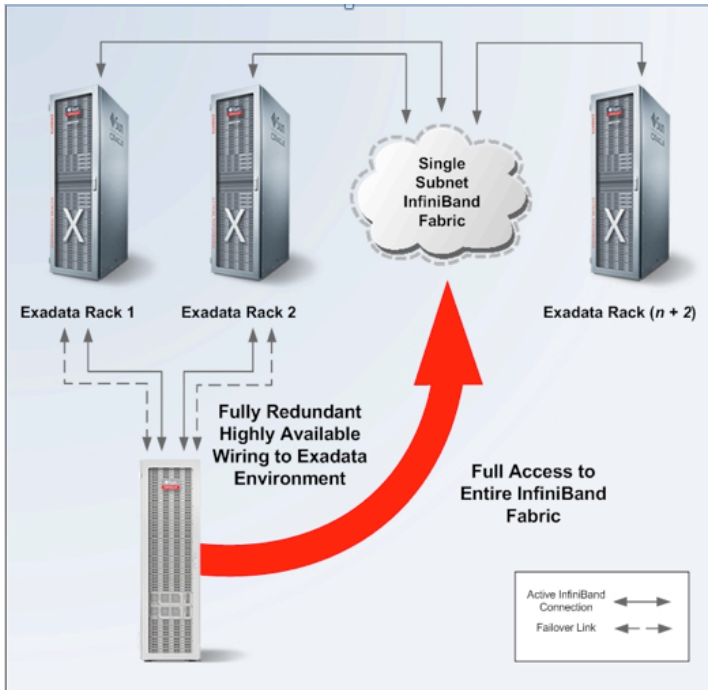

Figure 2. Oracle ZFS Storage ZS3-4 Cluster integration into a multiple rack Exadata architecture

Multirack Exadata with Multiple Subnet InfiniBand Fabrics

Figure 3 shows how an Oracle ZFS Storage Appliance can be integrated into a single subnet InfiniBand fabric.

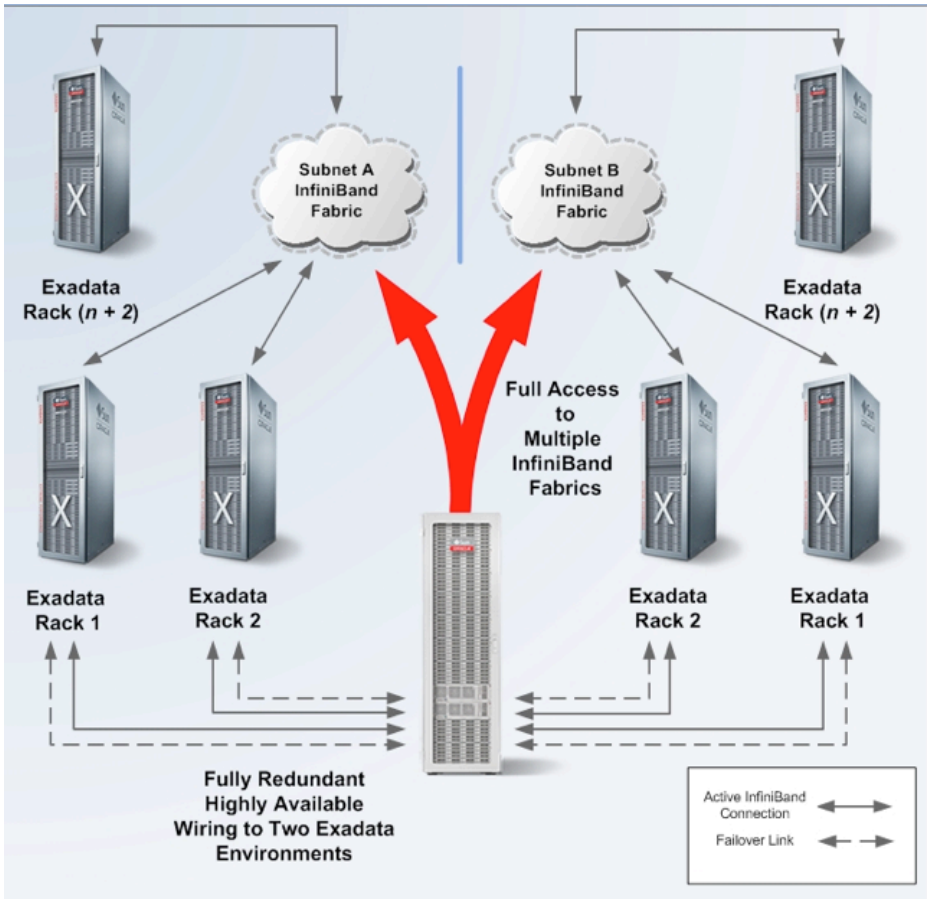

Figure 3. Architecture for an Oracle ZFS Storage ZS3-4 cluster integration into a multirack Exadata and multiple IB subnets

Cabling and Configuring the InfiniBand Network

The following section describes the procedures for cabling an Exadata InfiniBand fabric and cabling and configuring the Oracle ZFS Storage Appliance to access the InfiniBand fabric.

Merging Multiple Exadata Racks Together

Two Exadata racks can be cabled together to share the same IB fabric mesh. The merged Exadata racks can then be connected to a single clustered Oracle ZFS Storage Appliance. Successful setup requires adherence to some critical prerequisites and instructions, including physical cabling procedures. Be sure to reference the *Oracle Exadata Database Machine Owner's Guide*, Part IV: Extension

of the Oracle Database Machine and Oracle Exadata Storage Expansion Rack. Refer to the References section at the end of this document for links to documentation.

Prior to merging multiple Exadata machines, carefully consider the following:

- Cabling the two Exadata racks can be done with no downtime but there will be some performance degradation that occurs while the cabling procedure is executed.
- The environment will not be a high-availability environment because one leaf switch will need to be off. All traffic will go through the remaining leaf switch.
- Only the existing rack will be operational, and any new rack that will be added will be powered down.
- The software running on the systems cannot have problems related to InfiniBand restarts.
- The existing spine switch will be set to priority 10 during the cabling procedure. This setting gives the spine switch a higher priority than any other switch in the fabric, and will be first to take the Subnet Manager Master role whenever a new Subnet Manager Master is being set during the cabling procedure.
- The new racks have been configured with the appropriate IP addresses to be migrated into the expanded system prior to any cabling, and there are no duplicate IP addresses. If merging existing Exadata racks with the different InfiniBand subnets or if two or more Exadata racks have the same InfiniBand IP addresses configured, the merger will not complete successfully. Reference the *Oracle Exadata Database Machine Owner's Guide*, Chapter 7, section 21, on how to change InfiniBand IP addresses in an Exadata rack.

When connecting multiple racks together, remove the seven existing inter-switch connections between each leaf switch, as well as the two connections between the leaf switches and the spine switch. From each leaf switch, distribute eight connections over the spine switches in all racks. In multirack environments, the leaf switches inside a rack are no longer directly interconnected, as shown in the following graphic.

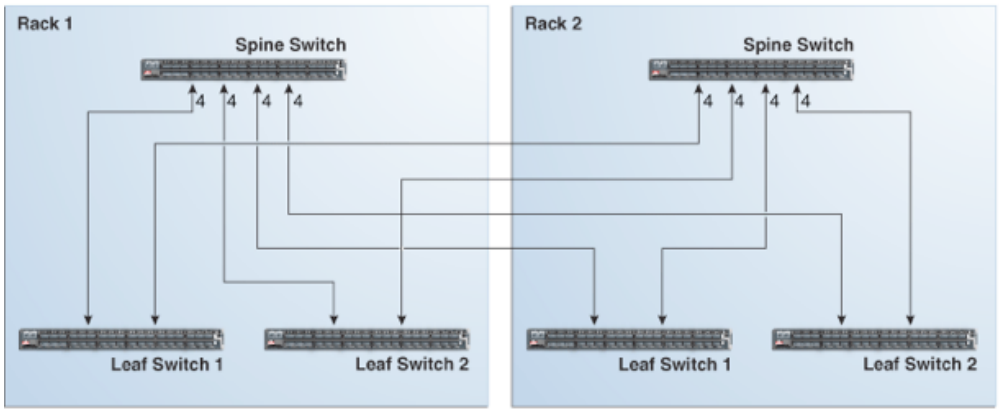

Figure 4. Two-rack Exadata spine and leaf switch cabling configuration

As shown in the preceding graphic, each leaf switch in rack 1 connects to the following switches:

- Four connections to its internal spine switch
- Four connections to the spine switch in rack 2

The spine switch in rack 1 connects to the following switches:

- Eight connections to both internal leaf switches
- Eight connections to both leaf switches in rack 2

As the number of racks increases from two to eight, the pattern continues as shown in the following graphic.

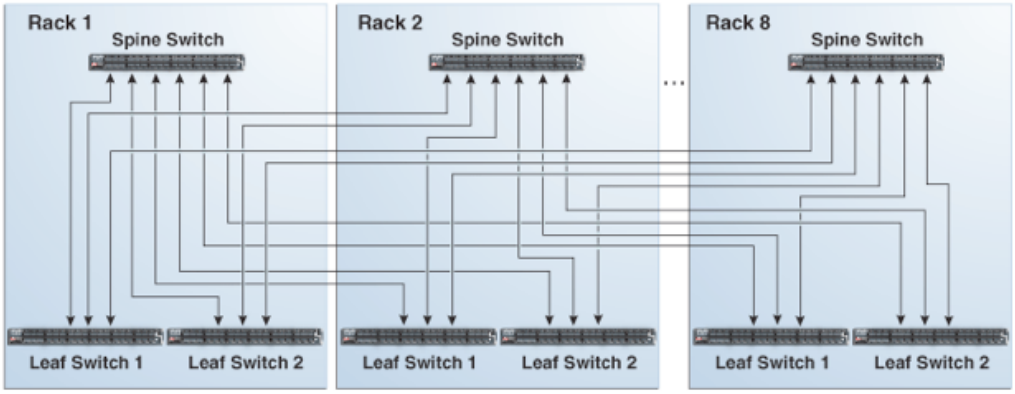

Figure 5. Multirack Exadata spine and leaf switch cabling configuration

As shown in the preceding graphic, each leaf switch has eight inter-switch connections distributed over all spine switches. Each spine switch has 16 inter-switch connections distributed over all leaf switches. The leaf switches are not directly interconnected with other leaf switches, and the spine switches are not directly interconnected with the other spine switches.

Refer to the "Multi-Rack Cabling Tables" section of the *Oracle Exadata Database Machine Owner's Guide, 11g Release 2* for further information on cabling more than two Exadata machines.

Cabling and Configuring the Oracle ZFS Storage Network

The following section describes how to wire and configure the Oracle ZFS Storage in three types of Exadata Infiniband architectures, which are common configurations:

- Oracle ZFS Storage, with two InfiniBand HCAs per head, configured to access a single subnet InfiniBand fabric and wired to provide redundancy and high availability.
- Oracle ZFS Storage, with four InfiniBand HCAs per head, configured to access two InfiniBand subnets wired to provide redundancy and high availability.
- Oracle ZFS Storage, with one InfiniBand HCA per head, configured to access a single subnet InfiniBand fabric where high availability is not a requirement.

Cabling Oracle ZFS Storage ports and the InfiniBand fabric

The Oracle ZFS Storage IP ports should be cabled across multiple Exadata leaf switches from two Exadata machines. With IB fabric providing an Inter-Switch Link (ISL) capacity of 280 gigabytes (GB) per second between leaf switches, interconnect link speed of throughput speed should not be a concern. The focus should be on availability. In the event that an Exadata leaf switch fails or an entire Exadata is powered down and brought offline, the Oracle ZFS Storage should be wired so that the remaining Exadata machines in the fabric still have access.

Typically the ibp0 device is assigned to port 1 on the HCA residing on slot 4. The ibp1 device is port 2 and ibp3 is port 1 on the HCA residing in slot 5. To ensure proper cable mapping and specific port assignments, verify which ibp devices are assigned to which ports.

The following example shows how to retrieve the ibp0 and ibp1 ports and Globally Unique Identifiers (GUIDs) using the Oracle ZFS Storage CLI. These steps can be repeated for all the ibp devices on both heads.

```
#ssh root@<7420-head1-name>
Password:
Last login: Thu Jun 27 21:50:22 2013 from <your host>
7420-head1-name:> configuration net devices select ibp0 show
Properties:
                          speed = 32000 Mbit/s
                             up = false
                         active = false
                          media = Infiniband
                    factory_mac = not available
                           port = 1
                           guid = 0x212800013e7c03

7420-head1-name:> configuration net devices select ibp1 show
Properties:
                          speed = 32000 Mbit/s
                             up = false
                         active = false
                          media = Infiniband
                    factory_mac = not available
                           port = 2
                           guid = 0x212800013e7c04
```

Configuring Oracle ZFS Storage Appliance with a single subnet InfiniBand fabric

The Oracle ZFS Storage Appliance cluster configuration should include two dual-port InfiniBand HCAs per head. The multiple HCA within the Oracle ZFS Storage heads should be configured to span multiple Exadata leaf switches from two separate Exadata machines on the same subnet. This will provide for maximum redundancy and high availability.

Install IB HCAs in PCI slots 4 and 5. Additional HCAs can be installed in slots 3 and 6 if the configuration calls for four HCAs. Slots 3-6 are the supported PCIe configuration for the Oracle ZFS Storage ZS3-4, which supports a maximum of four InfiniBand HCAs per head. (The four HCA per head configuration is discussed later in this document.)

Figure 6 depicts the physical PCIe location where the InfiniBAND QDR dual-port HCAs should be installed.

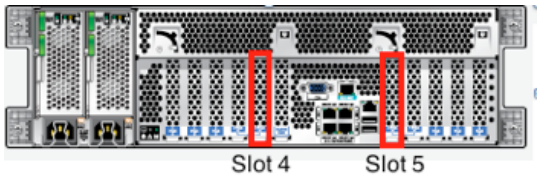

Figure 6. PCIe slot diagram for Oracle ZFS Storage Appliance with two dual-port InfiniBand QDR HCAs

Figure 7 shows a cabling diagram for integrating the Oracle ZFS Storage Appliance into a single subnet InfiniBand fabric. This setup provide redundancy and high availability in the event of an Exadata power outage or Exadata InfiniBand leaf switch failure.

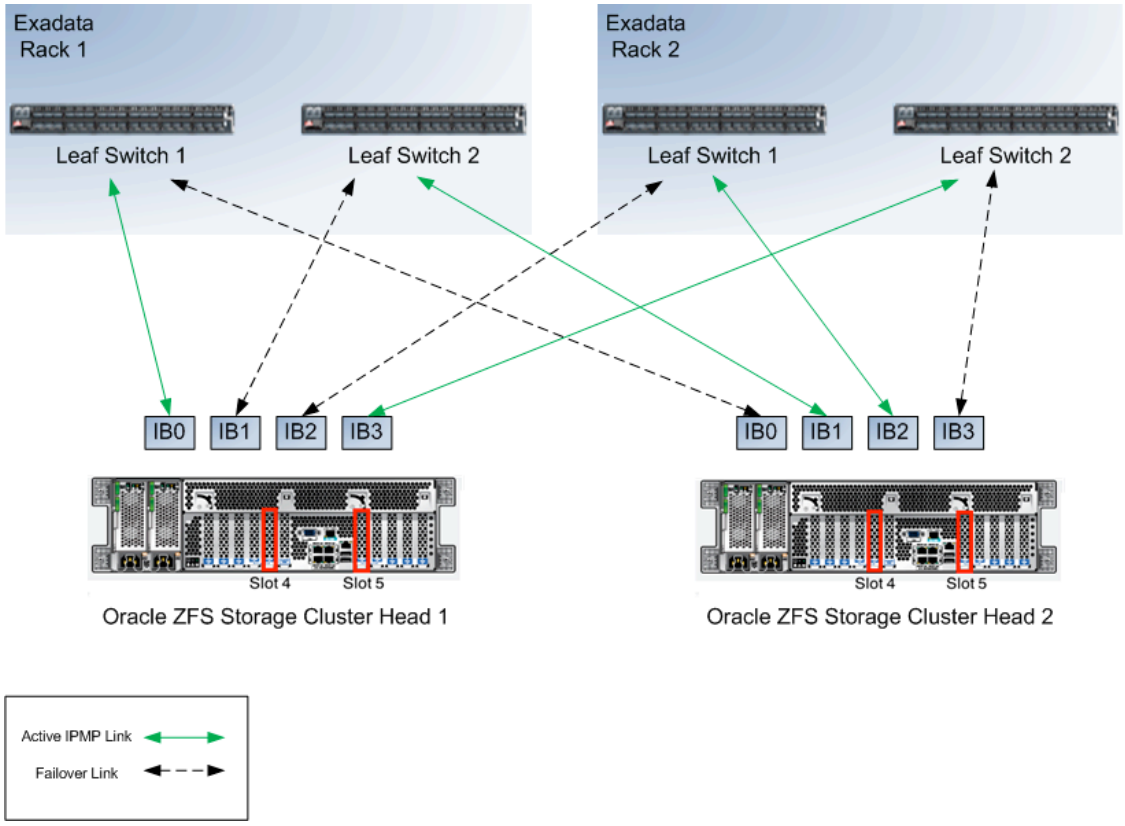

Figure 7. Cabling a highly available clustered Oracle ZFS Storage Appliance to a single subnet InfiniBand fabric

For each Oracle ZFS Storage Appliance head, connect the InfiniBand HCA ports as follows:

- Connect the port assigned ibp0 to one of the following ports on leaf switch 1 from rack 1: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp1 to one of the following ports on leaf switch 1 from rack 1: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp2 to one of the following ports on leaf switch 1 from rack 2: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp3 to one of the following ports on leaf switch 1 from rack 2: 5B, 6A, 6B, 7A, 7B, or 12A.

Connecting Oracle ZFS Storage Appliance with an InfiniBand fabric with two separate subnets

The Oracle ZFS Storage Appliance cluster configuration should include four dual-port InfiniBand HCAs per head. The multiple HCAs within the Oracle ZFS Storage heads should be configured to span multiple Exadata leaf switches from two separate Exadata machines on the same subnet. This will provide for maximum redundancy and high availability.

Install IB HCAs in PCIe slots 4 and 5. Install additional HCAs into slots 3 and 6.

Figure 8 depicts the physical PCIe location where the InfiniBand QDR dual-port HCAs should be installed.

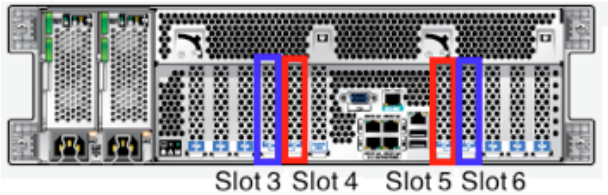

Figure 8. PCIe slots on Oracle ZFS Storage Appliance with four dual-port InfiniBand QDR HCAs

Figure 9 shows a cabling diagram for integrating the Oracle ZFS Storage Appliance into an InfiniBand fabric with multiple subnets.

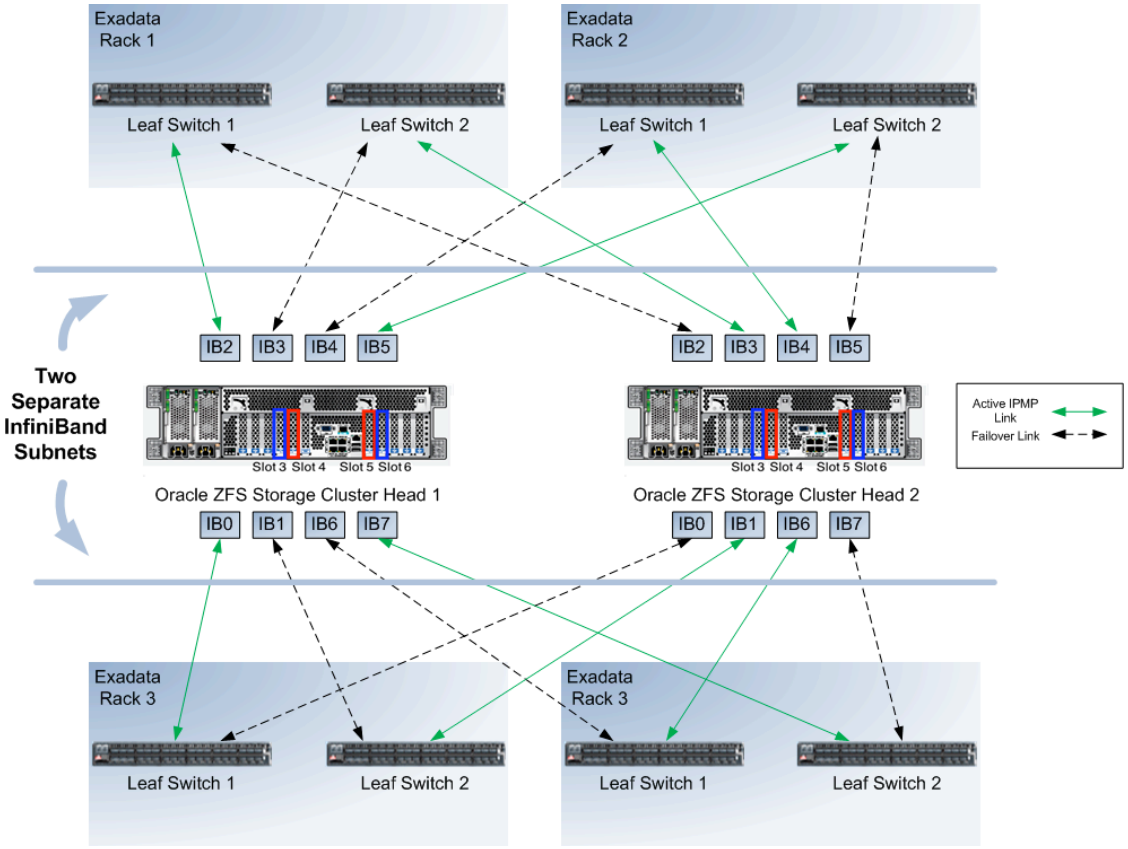

Figure 9. Cabling for a highly available clustered Oracle ZFS Storage ZS3-4 to an InfiniBand fabric with two separate subnets

For example, with the following assignments:

- Devices ibp0/1 are assigned to ports 1 and 2 for the HCA in slot 3.
- Devices ibp2/3 are assigned to ports 1 and 2 for the HCA in slot 4.
- Devices ibp4/5 are assigned to ports 1 and 2 for the HCA in slot 5.
- Devices ibp6/7 are assigned to ports 1 and 2 for the HCA in slot 6.

The configuration would be as follows:

- Connect the port assigned ibp0 to one of the following ports on leaf switch 1 from rack 1: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp1 to one of the following ports on leaf switch 1 from rack 1: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp2 to one of the following ports on leaf switch 1 from rack 2: 5B, 6A, 6B, 7A, 7B, or 12A.

- Connect the port assigned ibp3 to one of the following ports on leaf switch 1 from rack 2: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp4 to one of the following ports on leaf switch 1 from rack 3: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp5 to one of the following ports on leaf switch 1 from rack 3: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp6 to one of the following ports on leaf switch 1 from rack 4: 5B, 6A, 6B, 7A, 7B, or 12A.
- Connect the port assigned ibp7 to one of the following ports on leaf switch 1 from rack 4: 5B, 6A, 6B, 7A, 7B, or 12A.

Oracle ZFS Storage with InfiniBand fabric using a single HCA per head

If high availability in not a requirement, the Oracle ZFS Storage cluster can be configured with just one dual-port InfiniBand HCA per head. Install an IB HCA in each head, preferably in PCI slots 4 or 5.

Figure 10 shows the cabling to integrate the Oracle ZFS Storage Appliance into a single subnet InfiniBand fabric with just one HCA per head.

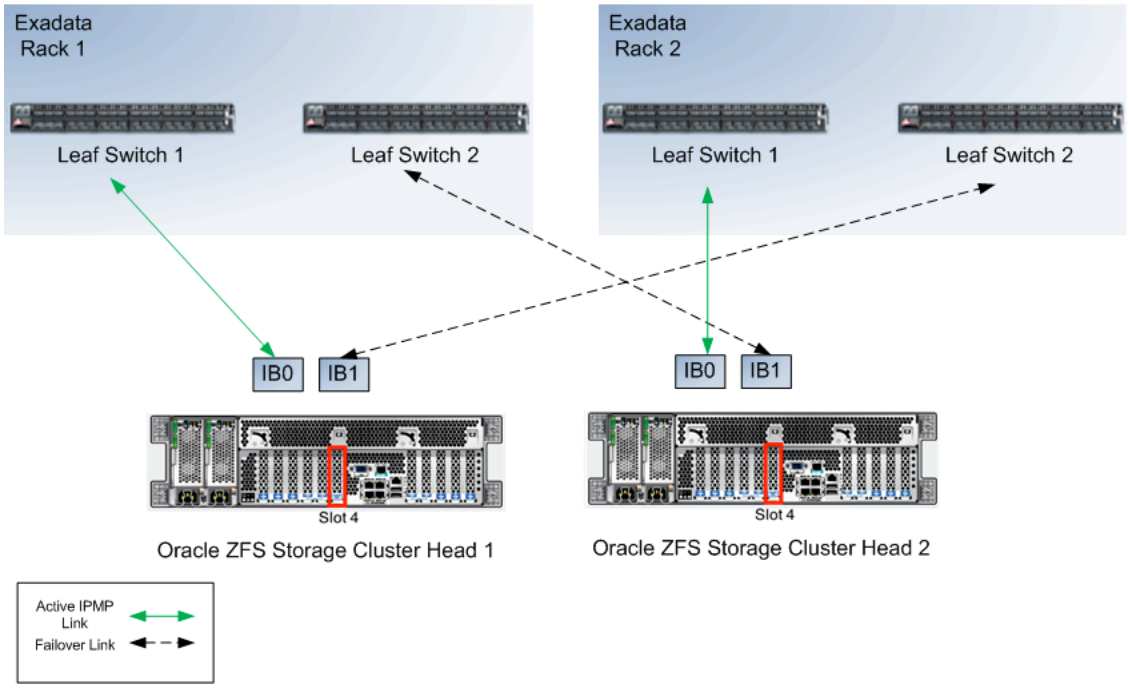

Figure 10. Cabling a clustered Oracle ZFS Storage ZS3-4 to a single subnet InfiniBand fabric

Configuring the IPMP Network with Oracle ZFS Storage

This section describes the process to configure the IP network multipathing (IPMP) groups.

The basic network configuration steps are as follows:

Configure ibp0, ibp1, ibp2, and ibp3 with address 0.0.0.0/8 (necessary for IPMP), connected mode, and partition key (the default is ffff). To identify the partition key used by the Oracle Exadata system, run the following command as the root user:

```
# cat /sys/class/net/ib0/pkey
```

Configure the active/active IPMP groups. The active/active IPMP group will be spread across two separate HCAs and across both Oracle ZFS Storage ZS3-4 chassis busses. For example, during the Oracle ZFS Storage cabling portion of configuration, if it is determined that the HCA residing in slot 4 contains ibd2 and ibd3 and that the HCA in slot 5 contains ibd0 active and ibd1, create an active/active IPMP group with ibd0/ibp3 and ibd1/ibp2.

Enable adaptive routing to ensure traffic is load balanced appropriately when multiple IP addresses on the same subnet are owned by the same head. This occurs after a cluster failover.

Figure 11 demonstrates how an IPMP group can be configured through the Oracle ZFS Storage Appliance browser user interface (BUI).

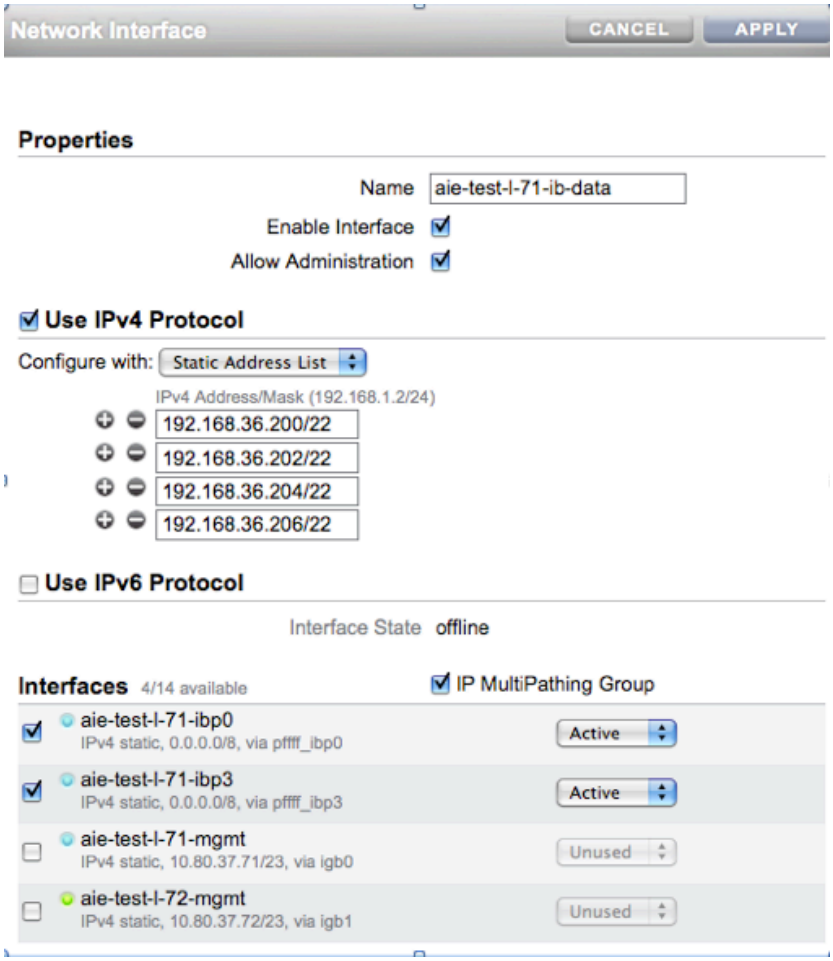

Figure 11. Multirack Exadata spine and leaf switch cabling configuration in the Oracle ZFS Storage BUI

Conclusion

The Oracle ZFS Storage Appliance provides a simple, high-performance, and cost-effective platform to ensure data protection for Oracle Exadata databases. It is easily integrated with existing Exadata database systems with file-based access provided through NFS using the high-bandwidth network protocol InfiniBand (IB). Capital and operational costs are driven down by the elimination of backup application license and support fees associated with disk-based backup, as well as license-free data services, including snapshot, compression, and replication.

This article has demonstrated the technical aspects of integrating the Oracle ZFS Storage Appliance into a multirack Exadata architecture, and has provided an implementation guide to configure the Oracle ZFS Storage Appliance to access the Exadata flexible InfiniBand network. The architecture is designed to ensure Oracle ZFS Storage Appliance access to the entire Exadata InfiniBand fabric, providing a fully redundant, high-availability backup and recovery solution.

The Oracle ZFS Storage Appliance is the preferred general-purpose data protection solution for an Oracle Exadata connected to an InfiniBand fabric, or multiple fabrics, providing a cost-effective RMAN backup and recovery solution for Oracle Database.

References

RESOURCES	LOCATION
Oracle ZFS Storage Appliance Administration Guide and documentation library	http://www.oracle.com/technetwork/documentation/oracle-unified-ss-193371.html
Oracle Exadata Database Machine Documentation	http://wd0338.oracle.com/archive/cd_ns/E13877_01/doc/index.htm

Configuring a Single Oracle ZFS Storage Appliance into an InfiniBand Fabric with Multiple Oracle Exadata Machines
December 2013, Version 1.0
Author: Application Integration Engineering, Joseph Pichette

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together