Backup and Recovery Performance and Best Practices using the Sun ZFS Storage Appliance with the Oracle Exadata Database Machine
Executive Overview .................................................. 4
Introduction ................................................................ 6
Key Performance Observations and Metrics ......................... 7
Architecture and Description .................................. 9
MAA Test Environment .............................................. 12
Configuration Best Practices ................................... 14
  Sun ZFS Storage Appliance Software Stack .................. 14
  Network Configurations ...................................... 14
  Storage Configuration ........................................ 17
Database Backup and Restore Key Practices ..................... 17
  MAA Principles for Database Backup ............................. 19
  MAA Principles for Database Restore ......................... 23
Database Backup and Recovery Testing ........................ 24
  HA Testing ................................................................ 24
Data Retention .......................................................... 25
Why Is My Backup or Restore Slow? ............................. 26
  Test Exadata Database Machine Read Rates Using Backup Validate 26
  Test Exadata Database Machine Write Rates Using Backup ...... 27
  Evaluating network throughput ................................ 27
  Evaluating bottlenecks on ZFS Appliance ...................... 27
  Other Factors That Can Impact Backup and Restore Rates ..... 31
Conclusion .................................................................. 33
Appendix A: Sun ZFS Storage Appliance Configuration ........ 34
  Network Configuration for InfiniBand Connectivity .......... 34
  Cabling the InfiniBand Connections ............................. 34
  Configuring Network - Devices, DataLinks and Interfaces .... 35
  Network Configuration for 10Gigabit Ethernet Connectivity .... 40
  Cabling the 10Gigabit Ethernet connections .................. 41
  Configuring Network Cluster Resources ...................... 50
  Test Network and Throughput Utilization .................... 60
Appendix B Database Backup Sample Scripts ......................... 61
  Database weekly level 0 backup........................................ 61
  Database daily level 1 backups ...................................... 63
Executive Overview

This paper describes the Maximum Availability Architecture (MAA) best practices on how to configure and use the Sun ZFS Storage Appliance as a backup and recovery target for the Oracle Exadata Database Machine and achieve optimum performance and availability. Following these MAA validated procedures the following results can be achieved:

- Highly available and reliable Exadata Database Machine backup/restore architecture using Sun ZFS Storage Appliance, provides Disk (Exadata Database Machine) to Disk (Sun ZFS Storage Appliance) backup rates up to 9TB/hour to Sun ZFS Storage Appliance

- Restore rates up to 7TB/hour from Sun ZFS Storage Appliance

The MAA best practices for the Exadata Database Machine and Sun ZFS Storage Appliance are most applicable when your applications’ Recovery Point Objective (RPO) and Recovery Time Objective (RTO) requirements can be met with the above restore rates, and your database backups can comply with your required backup windows and with above backup rates. If you require faster backup or restore rates, you need to consider using Exadata Storage Expansion Racks or more internal Exadata storage in the Exadata Database Machine as the target disk storage or a larger Sun ZFS Storage Appliance configuration. Refer to Oracle MAA white paper “Backup and Recovery Performance and Best Practices for Exadata Database Machine - Oracle Database 11.2.0.2” for more information on backing up to Exadata storage. If you require near zero RTO and RPO, use Oracle Data Guard with integrated client failover. Refer to “Oracle Database 11g Release 2 High Availability Best Practices” book and Oracle MAA white papers “Oracle Data Guard: Disaster Recovery Best Practices for Exadata Database Machine” and “Client Failover Best Practices for Data Guard 11g Release 2” for more information.

This Exadata Database Machine and Sun ZFS Storage Appliance backup solution is a balanced high availability and performance configuration that has been validated by the MAA and Sun ZFS Storage Appliance development teams. The recovery time depends upon

- Database size divided by restore rates plus

- Incremental backup size divided by restore / merge rate plus
• Application of redo (archives and online redo).

The key benefits are:

• Highly available and highly performing backup and restore solution

• Cost effective

• Eliminates the configuration and operational guess work to achieve the above backup and restore rates

• Offload backups from Exadata Database Machine so more DATA space is available for database growth or for consolidating additional database on to the Exadata system

• Fast backup and restore times that can meet most RPO and RTO requirements

The Sun ZFS Storage Appliance can share read-only snapshots and read-write clones of backup sets and image copies. Snapshots can be tracked in the RMAN catalog or accessed using Clone DB. Clones can be opened read-write to support alternate processing needs, such as development or reporting systems¹. With Oracle Database 11.2.0.3, Sun ZFS Storage Appliance fully supports Hybrid Columnar Compression (HCC) data and allow for additional capability and flexibility for secondary processing requirements. The Sun ZFS Storage Appliance Remote Replication feature can be used to replicate copies of backup and unstructured data to remote Sun ZFS Storage Appliance systems. These copies can be used for additional processing, such as creating test DB clones, at remote sites or for disaster recovery. The Sun ZFS Storage Appliance can also be part of a disk to disk to tape (D2D2T) solutions and optionally enable compression and encryption. However, this best practice paper focuses on achieving a fast and reliable disk to disk backup and restore solution only. The following prescribed configuration and operational practices maximize the available bandwidth of Sun ZFS Storage Appliance, network bandwidth and a significant

¹ For database clone using Oracle Sun ZFS Storage, refer to the following MAA papers: 1) Database Cloning using Oracle Sun ZFS Storage Appliance and Oracle Data Guard and 2) Database Cloning using Oracle Sun ZFS Storage Appliance and Oracle Recovery Manager
percentage of Exadata IO bandwidth. Customers can also use RMAN “minimize load” option or resource management to reduce impact on critical application performance during backup and restore operations which will be discussed later in the paper.

Introduction

The use of Oracle Maximum Availability Architecture (MAA) best practices on the Exadata Database Machine delivers the most comprehensive and effective HA solution available for the Oracle Database.

A key operational aspect of deploying Exadata Database Machine and Sun ZFS Storage Appliance is to be sure that database backups are performed properly and Oracle Database can be restored quickly if disaster strikes. This white paper is based on Oracle Database release 11.2.0.3 or higher, and it describes the best practices for setting up the optimal backup and recovery strategy to protect mission-critical data.

This paper discusses the following topics:

- Key Backup & Recovery performance observations and metrics are discussed so the DBA better understands what a given configuration is capable of delivering

- Recommended Backup and Restore (B&R) architecture with Exadata Database Machine and Sun ZFS Storage Appliance

- Backup and Restore Strategy and Configuration Practices

- Troubleshooting when performance is below expectations
Key Performance Observations and Metrics

This paper describes the backup and recovery (B&R) performance testing of an Exadata Database Machine X2-2 Full Rack, Half Rack, and Quarter Rack across various configurations. This testing was done to determine what the performance of the various B&R configurations can deliver. From this we provide some "rules of thumb" to enable you to get the most from your ZFS and Exadata system B&R solution.

Disk backup and restore testing was performed with backup set formats using a Sun ZFS Storage 7420c appliance connected to the Exadata Database Machine using InfiniBand and 10GigE infrastructure.

The maximum performance rates for each component of the architecture when using InfiniBand connected Sun ZFS Storage Appliance are shown in Figure 1.

Figure 1: Performance rates per component when using InfiniBand
The maximum performance rates for each component of the architecture using 10GigE connected Sun ZFS Storage Appliance are shown in Figure 2.

Figure 2: Performance rates per component when using 10GigE (LACP)
Table 1 summarizes the performance results for Sun ZFS Storage Appliance based backup and restores performance. For all cases except the Quarter Rack, the ZFS appliance hardware configuration will dictate the maximum B&R performance. For maximum availability and performance two Sun ZFS Storage Appliance heads and a minimum of two trays per head should be deployed, and that is what is shown below.

**TABLE 1 : SUMMARY OF MEASURED ZFS-BASED BACKUP AND RESTORE PERFORMANCE**

<table>
<thead>
<tr>
<th>Instances and Channels</th>
<th>Quarter Rack</th>
<th>Half Rack</th>
<th>Full Rack</th>
</tr>
</thead>
<tbody>
<tr>
<td>X2-2 (11.2.0.3)</td>
<td>4 TB/hour (10GigE)</td>
<td>8 TB/hour (10GigE)</td>
<td>9 TB/hour (10GigE)</td>
</tr>
<tr>
<td>16 RMAN channels spread evenly across instances</td>
<td>4 TB/hour (InfiniBand)</td>
<td>8 TB/hour (InfiniBand)</td>
<td>9 TB/hour (InfiniBand)</td>
</tr>
</tbody>
</table>

**FULL DATABASE RESTORE FROM DISK**

<table>
<thead>
<tr>
<th>Instances and Channels</th>
<th>Quarter Rack</th>
<th>Half Rack</th>
<th>Full Rack</th>
</tr>
</thead>
<tbody>
<tr>
<td>X2-2 (11.2.0.3)</td>
<td>4 TB/hour (10GigE)</td>
<td>7 TB/hour (10GigE)</td>
<td>8 TB/hour (10GigE)</td>
</tr>
<tr>
<td>16 RMAN channels spread evenly across instance</td>
<td>4 TB/hour (InfiniBand)</td>
<td>8 TB/hour (InfiniBand)</td>
<td>9 TB/hour (InfiniBand)</td>
</tr>
</tbody>
</table>

The best Backup and Recovery (B&R) performance is achieved on an Exadata Database Machine X2-2, when backups, restores are run across all the database instances (servers) in the Exadata system with either two or four RMAN channels allocated per instance. Further analysis is being performed regarding the B&R performance on an Exadata Database Machine X2-8. To achieve maximum B&R performance less than 5% CPU was required.

The above rates are the maximum rates consistently observed. Methods to minimize the impact on your application and essentially reduce the backup rates, such as using the RMAN “duration minimized load” syntax, will be discussed later in the paper.

**Architecture and Description**

The Sun ZFS Storage Appliance can be connected to the Exadata Database Machine using InfiniBand or 10GigE infrastructure.

---

2 For an Exadata Database Machine X2-2 Quarter Rack and Half Rack, the limiting factor is the throughput reading from or writing to the High Redundancy Data Disk Group.
When connecting the Sun ZFS Storage Appliance using the InfiniBand Switches in the Exadata Database Machine, four cables are connected between each head of the Sun ZFS Storage Appliance and the Exadata Database Machine InfiniBand leaf switches.

Figure 3: Sun ZFS Storage Appliance connected to Exadata Database Machine using InfiniBand
Alternatively, the Sun ZFS Storage Appliance can be connected using dedicated InfiniBand Switches that are local to the Sun ZFS Storage Appliance rack. In this implementation, an InfiniBand FatTree topology is created whereby the InfiniBand Leaf Switches are connected to two InfiniBand Spine Switches. Four cables are connected between each head and the InfiniBand Leaf Switches local to the Sun ZFS Storage Appliance, and then a FatTree topology is created between the 4 InfiniBand Leaf Switches and the 2 InfiniBand Spine Switches, 4 cables are connected from each InfiniBand Leaf Switch to each InfiniBand Spine Switch.

Refer to the Exadata Owners Guide for the correct cabling for Multi Rack configurations

Figure 4 : Sun ZFS Storage Appliance connected to Exadata Database Machine via dedicated InfiniBand Switches
When connecting the Sun ZFS Storage Appliance using 10GigE, four cables are connected between each head of the Sun ZFS Storage Appliance and the customer’s existing 10GigE network infrastructure. The Exadata Database Machine is similarly connected into the customer’s existing 10GigE network infrastructure with 2 cables from each DB Node in the Exadata Database Machine. Using the customer’s existing 10GigE network infrastructure allows multiple Exadata Database Machines to be connected to the Sun ZFS Storage Appliance, creating a hub and spoke design.

Figure 5: Sun ZFS Storage Appliance connected via 10Gigabit Ethernet to Exadata Database Machine

MAA Test Environment

The Sun ZFS Storage Appliance that was used in this test was configured as follows:

- Sun ZFS Storage 7420 appliance cluster (two controllers)
- Each head is configured with
  - Processors: 2 x 2Ghz Intel® Xeon® CPU X7550 @ 2.00GHz
  - Memory: 128 GB
  - System Data: 2 x 500 GB
  - Cache Log: 4 x 500 GB
- SAS-2 HBA Controller: 2 x Dual 4x6Gb External SAS-2 HBA
- InfiniBand: 2 x Dual Port QDR IB HCA M2
- 10Gigabit Ethernet: 2 x Dual Ported 10Gb Optical Ethernet
- Cluster Card: 1 x Fishworks CLUSTERON 200
- 4 x Sun Disk Shelf (SAS-2) each tray configured with
  - 2 x STEC/ZeusIOPs 18 GB Log Device
  - 20 x Seagate 1.0 TB 7200 Disk Drive

Figure 6: Sun ZFS Storage 7420 appliance cluster with 4 Trays of Storage

Figure 6 shows a clustered Sun ZFS Storage 7420 with two SAS HBA controllers and the connectivity between each controller port and the four disk trays. For more information refer to the Sun ZFS Storage 7x20 Appliance Installation Guide.
Configuration Best Practices

The best practices for backing up to and restoring from the Sun ZFS Storage Appliance are broken up into the following categories:

- Sun ZFS Storage Appliance Software Stack
- Network configurations
- Storage configuration
- Backup key practices
- Backup Strategy
- Restore key practices
- Restore Strategy
- Data Retention
- Test Practices

Sun ZFS Storage Appliance Software Stack

For the latest details on the recommended software stack for the Sun ZFS Storage Appliance when used in conjunction with the Exadata Database Machine, please refer to My Oracle Support Note 1354980.1. At the time of writing, this recommendation was to use the 2011.1.1.1 release of the Sun ZFS Storage Appliance software stack.

Network Configurations

The network configuration for the Sun ZFS Storage Appliance is different depending upon the environment where the Sun ZFS Storage Appliance will be deployed and the number of Exadata Database Machines that need to be backed up.

- If the Sun ZFS Storage Appliance is to be used to backup one or two Exadata Database Machine systems and the Sun ZFS Storage Appliance can be located within 100 meters of the Exadata Database Machine(s), then it is possible to connect all the systems together using InfiniBand. Each Exadata Database Machine system will reside on its own IB fabric and different IB subnet with our recommended configuration.

- If the Sun ZFS Storage Appliance is to be used to backup more than two Exadata Database Machine systems, or the Sun ZFS Storage Appliance cannot be located within 100 meters of the Exadata Database Machines, then connect the Exadata Database Machines to the Sun ZFS Storage Appliance using 10Gigabit Ethernet.

For configuration simplicity, configure one head first, and then add the second head into the Sun ZFS Storage Appliance cluster.
Network Configuration for InfiniBand Connectivity

To provide the highest availability and performance, it is recommended that each Sun ZFS Storage Appliance controller is configured with two Dual Port QDR InfiniBand HCA cards, providing four ports total per Sun ZFS Storage Appliance controller.

The Sun ZFS Storage Appliance controllers are connected to the InfiniBand Leaf Switches in the Exadata Database Machine rack in a way that provide continuous connectivity in the event of an outage affecting any one of the following:

- InfiniBand HCA card in the Sun ZFS Storage Appliance
- Failure of a Sun ZFS Storage Appliance controller (a.k.a. head)
- Failure of an InfiniBand switch

In order to achieve this, the Sun ZFS Storage Appliance network configuration is configured using an Active / Standby IPMP group. Please refer to Network Configuration for InfiniBand Connectivity in the Appendix.

Figure 7 shows the 2 controllers of a Sun ZFS Storage Appliance connected to two different Exadata Database Machines. This may be the case if there are two different and independent production applications running on different Exadata Database Machines, and the backups are to share a common Sun ZFS Storage Appliance, or if there is a test system that needs to be refreshed from the production system on a regular basis. The requirements for such a configuration are:

- Each Sun ZFS Storage Appliance controller is configured with 4 InfiniBand HCA Cards.
- Two InfiniBand HCA Cards are paired together and are connected into the free ports of the leaf switches in the respective Exadata Database Machine.
- The two Exadata Database Machines MUST be configured with different InfiniBand Subnets.
Network Configuration for 10Gigabit Ethernet Connectivity

To provide the highest availability and performance while not using the InfiniBand network, it is recommended that each Sun ZFS Storage Appliance controller is configured with two Dual Port 10Gb Optical Ethernet, providing four ports total per Sun ZFS Storage Appliance controller.

The Sun ZFS Storage Appliance controllers should also be connected to a pair of redundant 10Gigabit Network Switches in such a way as to provide continuous connectivity in the event of an outage affecting any one of the following:

- 10Gigabit NIC card in the Sun ZFS Storage Appliance
- Failure of a Sun ZFS Storage Appliance controller (aka head)
- Failure of a 10Gigabit Network switch

In order to achieve the highest backup and restore rates, the Sun ZFS Storage Appliance network configuration is configured using an Active / Active IP Multipathing (IPMP) group. Additionally, for performance reasons, the 10Gigabit network should be configured for a large MTU size also known as Jumbo Frames. Please refer to Network Configuration for InfiniBand Connectivity in the Appendix.
Configuring network cluster resources

Once the network resources have been configured, either 10Gigabit Network or InfiniBand, the Interfaces need to be put under the control of the Sun ZFS Storage Appliance cluster.

Clustering of the network resources provides continuous access to the data on the Sun ZFS Storage Appliance in the event of a Sun ZFS Storage Appliance controller head failure.

Refer to the Appendix section, “Configuring Network Cluster Resources,” for more information.

Storage Configuration

For the purpose of database backup and recovery to the Sun ZFS Storage Appliance, it is recommended that the Sun ZFS Storage Appliance be configured in two equally sized pools. Additionally, for the backup strategy recommended for Exadata Database Machine, the write flash (log) devices are not required for backup and restore performance as these operations do not cache data in the Sun ZFS Storage Appliance head. However, for systems that support other processing, such as unstructured data, general NFS access or secondary database processing write flash devices should be considered and should be evenly distributed between the two pools. Each pool should be created on the local node.

The pool configuration used for this testing consisted of two pools. Half of the drives from each disk tray and half of the number of write flash devices from each disk tray were assigned to the pool, which was then configured with a Data Profile of “Single parity, narrow stripes” and a Log Profile of “Mirrored Log”. If the configuration permits, select the No Single Point of Failure option for “Mirrored Log” profile.

Once the pools have been created, they are put under the control of the Sun ZFS Storage Appliance cluster. By configuring the network and pool resources, a database backup or restore being performed by RMAN will continue to run in the event that a failure occurs within the Sun ZFS Storage Appliance controller or a network component supporting the Sun ZFS Storage Appliance.

Database Backup and Restore Key Practices

These key practices that should be followed in order to obtain maximum availability and performance from the Sun ZFS Storage Appliance:

- Create Shares on the Sun ZFS Storage Appliance
- Enable Direct NFS
- Configure oranfstab
- Adopt MAA principles for Database Backup and Restore
Create Shares on the Sun ZFS Storage Appliance

The Oracle Exadata Backup Configuration Utility v1.0 is a plug-in that can be downloaded from Oracle Technology Network (OTN) at the Sun NAS Storage Downloads page. The utility automates the procedure of creating the Sun ZFS Storage Appliance Projects and Shares and creates the corresponding entries on the Database Nodes in the cluster. Finally, it creates RMAN command files that can be used for backing up the database.

Prior to running the Oracle Exadata Backup Configuration Utility, the two pools of storage should be created as previously discussed.

The utility incorporates all of the best practices of using the Sun ZFS Storage Appliance as a backup destination for an Exadata Database Machine:

- Create two projects, one per pool
- Configure the projects to use a write bias of “throughput”
- Configure the projects NFS Exceptions to allow access to the shares by the Exadata Database Machine.
- Create at least eight shares per project – for optimal backup and restore rates

Refer to the Sun NAS Storage Download page on the Oracle Technology Network for more details.

Enable Direct NFS

The Oracle Database is backed up to the Sun ZFS Storage Appliance using the Direct NFS option available within the Oracle Database version 11g Release 2 running on the Exadata Database Machine. By default Direct NFS option is not enabled, and to enable Direct NFS requires the Oracle kernel to be relinked, requiring a shutdown of the databases running on the Exadata Database Machine.

By enabling Direct NFS, the Linux kernel NFS processes are bypassed and the Direct NFS utility opens up 1 MB network buffers solely for use by RMAN to speed up the backup and restore of the database, bypassing the system configured TCP network buffers. To enable Direct NFS on a specific Oracle Home, first all databases that are using the Oracle Home need to be stopped. Login as the owner of the Oracle software, typically “oracle” and then

---

3 A ZFS Share maps to an NFS Mount Point that is then mounted by the database nodes on the Exadata Database Machine. A ZFS Project provides a logical grouping of ZFS Shares and also allows for the central configuration of properties that affect all the corresponding shares including identifying the hosts or networks that are allowed to write to the ZFS Shares.


5 One customer observed a 25% performance increase by enabling Direct NFS for their database restore operations and the throughput was more consistent with Direct NFS enabled.
$ srvctl stop database -d <dbname>

Then on each database node, after sourcing the ORACLE environment variables (ORACLE_HOME) enter the following commands:

$ cd $(ORACLE_HOME)/rdbms/lib
$ make -f ins_rdbms.mk dnfs_on

The databases can now be restarted, but for databases that will be backed up using Direct NFS, the /etc/oranfstab file must be created.

Refer to the Oracle® Database Installation Guide 11g Release 2 (11.2) for Linux – Section 5.3.9 “Configuring and Using Direct NFS Client” for more details.

Configure /etc/oranfstab

The previous section in the Installation Guide introduces the procedure for creating the /etc/oranfstab file. It is not essential to create an oranfstab file for the Oracle DBMS to utilize Direct NFS. However, under the following configurations, an oranfstab file must be created to fully utilize the network resources between the database nodes in the Exadata Database Machine and the Sun ZFS Storage Appliance.

- Exadata Database Machine X2-8 is to be connected to the Sun ZFS Storage Appliance using InfiniBand OR Exadata Database Machine is running Solaris OR an Active/Active IPMP group is created on the Sun ZFS Storage Appliance

Refer to the Appendix for sample /etc/oranfstab files.

MAA Principles for Database Backup

The recommended backup strategy when backing up the Oracle Database to the Sun ZFS Storage Appliance is to use a combination of Level 0 and Level 1 backupsets.

- The Oracle Database Fast Recovery Area should not be put on the Sun ZFS Storage Appliance, and should remain on the Exadata Storage Servers that are part of the Exadata Database Machine
- Weekly a full level 0 backup will be taken and written to the Sun ZFS Storage Appliance
- Daily incremental level 1 backup will be taken and written to the Sun ZFS Storage Appliance.

The frequency of the level 0 and level 1 backup is dictated by your RTO and RPO requirements when recovering from failures. However, for most the above recommendation should be sufficient.

Oracle MAA development team does not recommend an incrementally updated backup strategy which uses a combination of data file copies and incremental backupsets that are subsequently merged into the data file copies when using the Sun ZFS Storage Appliance. When an incrementally updated backup strategy is selected, the backup solution becomes I/O bound during the merge process, and to
alleviate this bottleneck a significant number of disk spindles must be made available in order to achieve the required IOPS rate. Additional testing is expected to be conducted by the Exadata MAA and Sun ZFS Storage Appliance development teams in order to determine the Sun ZFS Storage Appliance backup and restore rates that are achievable with specific configurations. The configuration is likely to consist of 92 High Performance 600GB drives and a number of write flash devices.

**RMAN Channels**

The recommendation for both the weekly Level 0 and daily Level 1 backups is to allocate a maximum of 16 RMAN Channels which are configured to write to one of the 16 Sun ZFS Storage Appliance Shares created by the Oracle Exadata Backup Configuration Utility in the previous step.

**RMAN Compression**

If you intend to use RMAN compression to preserve space on the ZFS Storage Appliance, but the database consists of a mixture of uncompressed data and compressed data or data that cannot be compressed further (which might be the case with LOB data stored in the database), then two backup jobs should be submitted. One of the backup jobs will create an uncompressed backup, and the second a compressed backup. Attempting to compress data that doesn’t typically yield good compression results lengthens the backup window significantly, consumes unnecessary CPU resources, and results in very little space savings.

For instance, if the database consists of predominantly uncompressed data, but the database also stores LOB data in a small number of tablespaces, then the backup should be broken into two parts: an *include* part listing the tablespaces where the LOB data is stored, and an *exclude* part for the rest of the database.

The extract from the weekly level 0 RMAN backup job might look like the following:

```sql
configure exclude for tablespace 'LOB_DATA1';
configure exclude for tablespace 'LOB_DATA2';
run {
allocate channel ch01 device type disk ....
backup as backupset incremental level 0 section size 32g filesperset 8
tablespace LOB_DATA1, LOB_DATA2 tag 'FULLBACKUPSET_L0';
backup as compressed backupset incremental level 0 section size 32g filesperset 8 database tag 'FULLBACKUPSET_L0' plus archivelog;
...}
```
Note that in many cases, enabling compression lengthens your backup and restore elapsed times so tests are required to ensure that you can still meet your RTO requirements.

**RMAN Duration Minimize Load**

A second consideration is using the RMAN option to amortize the backup over a time period to minimize the performance impact to your critical application. For example, if there is an 8 hour backup window available for the weekly level 0 database backup of 20 TB database, but critical functions are still running during that duration, you can use the RMAN duration minimize load option to spread the work over the 8 hour window.

So instead of completing the backup in less than 3 hours, and possibly impacting your application during that time, use RMAN backup option duration minimize load to minimize the impact throughout the entire backup window.

Example RMAN backup job consists of:

```sql
run {
allocate channel ...
backup as backupset incremental level 0 section size 32g duration 8:00 minimize load database tag 'FULLBACKUPSET_L0' plus archivelog;
...
}
```

**RMAN and BIGFILE Tablespaces**

The use of certain RMAN options such as sectionsize are recommended and included in the backup scripts that are generated by the Oracle Exadata Backup Configuration Utility. The sectionsize parameter is especially important to break up BIGFILE tablespaces into more manageable amounts. It is recommended that the largest tablespaces are broken up into equal size sections for each RMAN channel. For instance, if there are two tablespaces of 10TB and 14TB, and we have 16 RMAN channels, then the section size should be set to approx 320GB. Samples of the generated RMAN backup scripts can be found in the Appendix.

The Oracle Exadata Backup Configuration Utility does not enable RMAN or Sun ZFS Storage Appliance compression, as this is something that is dependent on the data to be backed up and the database licensing options purchased. Additionally, when RMAN compression is enabled, more CPU processing power on the DB Node is required, with a maximum of each RMAN channel consuming a CPU core on the DB Node where it is running.

For any backup operation or for any changes to your backup practices, ensure that the backups complete within the backup window available and that the database can be restored and recover to meet your RTO requirements in the case of a failure.
MAA Principles for Database Restore

Maximizing System Resources for Restore

When a database needs to be restored due to an unexpected failure, this is often the primary objective for the IT department of a corporation. To facilitate the database restore as fast as possible, the following actions should be taken:

- Suspend all other backup jobs that might be using the same resources. In the case of the Sun ZFS Storage Appliance, all the system and network resources should be dedicated to the restore. Therefore other backup jobs should be suspended and if cloned environments are being used for testing purposes, these environments should also be shutdown.

- Utilize all nodes on the database machine where the database instances normally run. See “Utilizing All Nodes on the Database Machine” for more information.

In a consolidated environment, where multiple databases or applications are using the database machine, it might not be acceptable to use all the system resources for the restore operation. In this case, the number of RMAN channels used for the restore can be reduced so as to use less of the system resources and allow SLAs to be met on the running applications.

Utilizing All Nodes on the Database Machine

Ideally, all the database nodes in the Exadata Database Machine should be used to evenly utilize the processing power of the Exadata Database Machine. This is easily achievable on a database that requires individual datafiles or tablespaces to be restored, but more difficult if the entire database has been lost, and a full database restore needs to occur.

If the database is still running, then the RMAN channels can be allocated; similar to the backup examples previously, or the RMAN session may connect to a database service that is started on all nodes in the Oracle Cluster where the database is running. In the later case, the degree of parallelism can be supplied which will cause the RMAN command to allocate channels across the running database nodes.

Periodic Database Restore Validation

As with any backup and recovery environment the solution is only valid if the database can be restored in the event of an outage. Therefore, the restore procedure should be validated periodically which includes entire database restore and recovery and application restart.

Ideally this would be tested on a test system identical to production. If a test system does not exist, the restore validate command can be used to ensure the database backup can be read and restored but does not validate the database and application recovery.

The restore validate command can also be used to evaluate how fast the data can also be read from the Sun ZFS Storage Appliance into the database nodes of the Exadata Database Machine.
For example, if a database service called orcl_restore is running on all four nodes in an Exadata Database Machine Half Rack, then the following commands would allocate 8 channels between the database nodes and perform a database restore validate.

$ rman target sys/welcome1@dm01-scan/orcl_restore
RMAN> configure device type disk parallelism 8;
RMAN> restore validate database;

A common practice is to do a complete restore and recovery test at a minimum every month and restore validate weekly.

Database Backup and Recovery Testing

The presence of a database backup does not guarantee that it will be usable in the event of a failure. Therefore it is a best practice to have a documented database restore and recovery procedure that is validated on a regular basis, to ensure that the procedure is still valid and that the application’s RTO and RPO can still be met.

This might not be the case if the database to be restored grows due to application consolidation or business growth or, if application changes occur that utilize compressed or encrypted data. If the database backup takes advantage of RMAN compression, then restoring and recovering the newly compressed or encrypted data might adversely affect the ability to meet the applications RTO. Therefore, to ensure that restore / recovery operations work and that they continue to meet RTO/RPO requirements:

- Restore validate tests weekly
- Restore and recovery tests periodically
- Always re-evaluate immediately after adding a new datafile / tablespace or changing backup options such as adding or changing compression options, or encryption.

In fact any change to the database size or configuration or ZFS architecture and configuration, the restore and recovery operations have to be re-evaluated to ensure the data is recoverable and RTO requirements are met.

HA Testing

The Sun ZFS Storage Appliance is configured for high availability. The network connections, either InfiniBand or 10GigE are connected to redundant switches to provide availability in the event of a failure, and the cluster configuration steps performed ensure that the network and disk resources fail over to the surviving ZFS controller in the event of a failure affecting the head.
The following graph plots the aggregated network throughput for the two database nodes of an Exadata Database Machine X2-8, and shows the performance drop that occurs when one of the heads is rebooted. The throughput initially drops but then recovers to approx 50% of the original throughput achieved when the ZFS Storage Array is fully operational.

While the backup rate is reduced, the RMAN backup is not affected and the backup job completes without error.

Data Retention

Database backup retention policy is configured in the same way that it is configured when using a traditional Fast Recovery Area, but the purging of the backups that become obsolete is a manual operation when backups are written to the Sun ZFS Storage Appliance.
The retention policy is either specified as a number of copies, or as a period of time. For instance, if two level 0 backups are to be maintained at all times, the retention policy would be configured with a redundancy level of 2.

RMAN> configure retention policy to redundancy 2;

Deleting old Database Backups

Database backups that are no longer required to support the desired retention policy are considered obsolete in RMAN. As part of the RMAN script, the backups no longer required can be purged with the delete obsolete command.

RMAN> delete obsolete;

Why Is My Backup or Restore Slow?

This section covers key troubleshooting, monitoring, and testing requirements.

Test Exadata Database Machine Read Rates Using Backup Validate

The first question which is usually asked when backups appear to be running slow is “Is the data coming off the cells fast enough?”

This can easily be answered by running the RMAN backup command, such as the weekly level 0 backup, but instead of writing the data to the backup location on the Sun ZFS Storage Appliance, the addition of the validate clause means that all of the RMAN code is executed with the exception of the creating and writing of the data to the Sun ZFS Storage Appliance. Similarly, if RMAN compression or encryption is being used, the effect of these options can be evaluated by the inclusion or removal of the options from the RMAN backup validate command.

On completion of the job, you can analyze the RMAN logfile to determine the duration of the backup job by looking at the Start Backup and Finish Backup times.

Starting backup at 07-Feb-2012 06:30:47

Finished backup at 07-Feb-2012 07:15:19

Note that RMAN writes the time to the log file using the format defined by the environment variable NLS_DATE_FORMAT.

Additionally, you can query the fixed view v$backup_async_io to get the “effective bytes per second” for each RMAN channel, and the OSWatcher logs can be analyzed to see the load that the backup put on the DB Nodes and the Exadata Storage Cells.
Test Exadata Database Machine Write Rates Using Backup

Having verified the backup read rates, the next task is to verify the write rates to the Sun ZFS Storage Appliance, by performing a database backup operation.

Again, the RMAN logfile can be analyzed to determine the duration of the backup job, by looking at the Start and Finish Backup Times. Additionally, the log file might identify a particular phase of the backup that took an excessive amount of time.

Starting backup at 07-Feb-2012 08:43:53
Finished backup at 07-Feb-2012 10:31:19

From a database machine perspective, this can be performed by again analyzing the v$backup_async_io fixed view and the OSWatcher logs.

In addition to analyzing the load that the backup puts on the Database Nodes and the Exadata Database Machine storage cells, the OSWatcher logs can also be used to analyze the load that is put on the network stack writing out of the Database Nodes.

Also, see the section below on Sun ZFS Storage Appliance Real time Monitoring and Alerting on how to use the Analytics product on the Oracle Sun ZFS Storage Appliance.

Evaluating network throughput

In the previous section on Database Restore Key Practices, we introduced the concept of performing periodic database restore validate tests to ensure the integrity of the backup on the Sun ZFS Storage Appliance. The same procedure can be used to evaluate the network throughput.

Evaluating bottlenecks on ZFS Appliance

The Sun ZFS Storage Appliance has a performance monitoring feature called Analytics as well as an alerting system built in.

Real Time Monitoring

When you first login to the Sun ZFS Storage Appliance you will land on the Status > Dashboard screen by default. This screen offers a quick view of the Sun ZFS Storage Appliance, including the storage usage, health of the services as well as the health of the hardware components. Additionally, there are a number of charts showing the 7 day, 24 hour, 60 minute and real time data, for different components including CPU, Network and Disk.

The following graphic shows the Sun ZFS Storage Appliance dashboard for one of the two heads in the cluster.
Highlighted on the CPU chart is the one hour view, and the 42% indicates that the CPUs have an average utilization of 42% over the course of the last hour.

Next to the Disk heading is a lightning bolt. The Sun ZFS Storage Appliance is pre-configured with certain thresholds for each component, and each threshold has a different “storm” indicator that starts sunny, passes through cloudy and rain into different hurricane strengths. By looking at the graphic pictured next to each component the administrator can see how busy the system is currently.

Below the Storage Usage pie-charts and the Services status, is the Hardware Status. In the event of a failure affecting one of these components, the green indicator will turn amber, indicating that a component has a problem. More information can be found in the Problems tab under Maintenance.

At the bottom of the screen are the most recent alerts to have occurred on the system.

The second graphic shows the Sun ZFS Storage Appliance dashboard for one of the two heads in the cluster, just after the other head had failed.

At the top of the screen a window has revealed the most recent alert, and the bottom of the screen also shows “The appliance has taken over from its cluster peer”.

This screen capture was taken approx 2 minutes after the failover had occurred, and the network “storm” indicator is showing a Cat-4 Hurricane. If you click on the icon, the configure dashboard
Clicking on any of the active graphs opens Analytics, the Sun ZFS Storage Appliance real-time monitoring functionality. Analytics can also be opened by clicking on the menu across the top of the screen.

It is possible to open a previously saved Analytics worksheet, or you can selectively choose different statistics by clicking on the + sign next to Add statistics.
The following graphic shows one metric already present on the screen, along with the different statistics class, metric and sub metric. In the case of this screen shot, Disk IO Bytes broken down by type of operation (read / write).

For more information about Sun ZFS Storage Appliance Analytics, refer to the online help that is accessible from the Sun ZFS Storage Appliance Analytics page.

Real time Alerting

The Sun ZFS Storage Appliance has a built in alerting system which can send monitoring alerts using email, SNMP, or can write to a centralized logging system using SYSLOG. The Alerting system can monitor different categories such as Network, which can show when a link becomes degraded or fails, Hardware faults, Cluster events, such as a cluster failover or takeover.

The system can also report on the progress of NDMP and Remote replication operations if applicable.
Other Factors That Can Impact Backup and Restore Rates

Overlapping Backups

While it might seem obvious, it is worth noting that the backup and restore rates noted in this paper are limited by the Sun ZFS Storage Appliance and not the Exadata Database Machine. If the Sun ZFS Storage Appliance is being deployed to provide a centralized backup solution for a number of Exadata and non-Exadata databases, then backup schedules need to be planned so as to not result in overlapping backups, or in such a way that the competing backups are planned and the SLA for the backups are still met.

For instance, the Sun ZFS Storage Appliance can write at a rate of approx 9TB/hour, but if two backups are run concurrently, then the maximum 9TB/hour rate will be distributed at best between the two backups evenly. Therefore, if there is a database of 9TB and another database of 18TB are scheduled to run concurrently, the SLA for the backup for both databases needs to specify a backup window of 3 hours or more.

Compressed Backups

Backup compression is usually considered when there is a desire to conserve space on a backup target such as the Sun ZFS Storage Appliance. While compression can reduce space utilization, the consequences of enabling compression is a longer running backup in most cases, due to the need to perform the CPU cycles to compress the data. Compression can be enabled at many points in the backup solution:

- At the RMAN level using either BASIC compression that comes with the Oracle 11g Release 2 Enterprise Edition
- At the RMAN level using LOW, MEDIUM or HIGH compression that comes with the Advanced Compression Option (ACO) of Oracle Database 11g Release 2
- At the Sun ZFS Storage Appliance project or share level using LZJB, GZIP, GZIP-2 or GZIP-9 compression algorithms

Regardless of the location where compression is enabled (RMAN or Sun ZFS Storage Appliance) the Database Node or Sun ZFS Storage Appliance controller performing the compression can become CPU-bound. This can be observed by analyzing the OSWatcher logs on the Database Node and looking at the top or mpstat log in particular, or by analyzing the Analytics on the Sun ZFS Storage Appliance and looking at the CPU utilization there.

- If compression is enabled in RMAN on the Database Node, then this results in less data being sent to the Sun ZFS Storage Appliance.
- If compression is enabled on the Sun ZFS Storage Appliance, then there will be no additional CPU load on the Database Nodes of the database machine, but the CPUs on the Sun ZFS Storage Appliance will be busier or the system may become CPU bound.
Decompressed Restores

Assuming that compression was enabled on the backup, decompression will occur at the same location on the restore.

However, during the database restores, the primary goal may be to complete the restore as quickly as possible so as to meet RTO and RPO. This might mean that the degree of parallelism within RMAN might need to be increased.

- If the compression occurs at the RMAN level, then this would mean that each additional RMAN channel would consume another CPU core – assuming the data can be supplied fast enough.

- If the compression occurs at the Sun ZFS Storage Appliance level, then increasing the number of RMAN channels might result in a bottleneck on the Sun ZFS Storage Appliance controller head, meaning the additional RMAN channels might not receive streaming data.

Testing should be conducted at both the database backup and database restore level to ensure that SLAs for the backup and RTO / RPO for the restore can be met if compression is enabled.

Contending I/O Resources

It is not always possible for database backups to be scheduled at quiet times when no other activities are occurring on the system. Data loads might be occurring on the Exadata Database Machine, or testing might be occurring on the Sun ZFS Storage Appliance snapshots or clones if they have been created. The Sun ZFS Storage Appliance may be hosting application data or logs with its own IO requirements and behavior conflicting with backups. These additional operations need to be taken into account when scheduling backups.

On the Exadata Database Machine, it is possible to use Resource Manager to help prioritize the I/O’s from the different operations, either giving prioritization to the backup operations so as to complete the backup in a specific window, or prioritize the data loads to ensure the system is available when the “online” users need to access the system in the morning.

Free Space and Fragmentation on the Sun ZFS Storage Appliance

As with other copy on write solutions, it is not advisable to use up all of the available free space on the disks in the Sun ZFS Storage Appliance pool. The recommendation from Oracle Exadata MAA and Sun ZFS Storage Appliance development team is to keep 20% free space on each disk pool created, allowing the copy and write operations to run optimally. This can be enabled by establishing quotas on the Sun ZFS Storage Appliance projects.

If the free space is allowed to reduce past the 20% recommendation, then write operations to the Sun ZFS Storage Appliance might be impacted because the appliance will be forced to spend more time looking for free space.

Additionally, fragmentation will cause both read and write operations to be impacted. When the Sun ZFS Storage Appliance is initially deployed, the appliance will be able to write data sequentially to the
empty disks, but once the data retention policy has been meet (which might be four weeks of weekly and daily incremental backups and archive logs on the Sun ZFS Storage Appliance), disk space might become fragmented as the old data is purged and new data is written. The effects of fragmentation are exaggerated when snapshots and clones are used for testing purposes, or if an incrementally updated backup solution is implemented against the recommendations of this paper.

Conclusion

The combination of Exadata Database Machine and Sun ZFS Storage Appliance can provide a very fast and cost effective backup and restore solution. With the integrated MAA and ZFS configuration and operational practices, customers should be able to configure this solution and achieved its desired HA benefits while avoiding and detecting the common pitfalls that may lead to slower backup and restore rates.
Appendix A: Sun ZFS Storage Appliance Configuration

Network Configuration for InfiniBand Connectivity

To provide the highest availability and performance, it is recommended that each Sun ZFS Storage Appliance controller is configured with two Dual Port QDR InfiniBand HCA cards, providing four ports total per Sun ZFS Storage Appliance controller.

The two InfiniBand cards will probably be installed into PCIe Slots 4 and 5 in the Sun ZFS Storage Appliance controller, and will be identified as ibp0 & ibp1 in PCIe Slot 4 and ibp2 & ibp3 in PCIe Slot 5. You can think of the two slots as the top and bottom port on the vertically installed PCIe Slot. Within the Sun ZFS Storage Appliance User Interface Under Configuration > Network, we will bind ibp0 & ibp3 as an Active/Standby IPMP group, and ibp2 & ibp1 as a seconds Active/Standby IPMP group.

Cabling the InfiniBand Connections

Run four cables from each Sun ZFS Storage Appliance controller head to the two InfiniBand Leaf Switches in the Exadata Database Machine. On the InfiniBand Leaf Switches, there are 6 ports that are free for customer expansion of the InfiniBand fabric for situations like connects the Sun ZFS Storage Appliance to the Exadata Database Machine for the purposes of backup and recovery. For the purpose of this white paper we will use the ports marked 5A, 6A, 6B and 7A. However, if you already have some devices using one or two of these ports, then you can use any of the 6 ports (5B, 6A, 6B, 7A, 7B and 12A) that are available on the InfiniBand Switch.

If the InfiniBand switch is full due to the presence of other devices such as media servers, data mules or middle tier application systems, then additional InfiniBand switches can be purchased and the two racks can be interconnected using the cabling method discussed in the Database Machine Owners Guide.

Connect the cables as described in the following table.

<table>
<thead>
<tr>
<th>SUN ZFS STORAGE APPLIANCE HEAD</th>
<th>SUN ZFS STORAGE APPLIANCE PORT</th>
<th>MODE</th>
<th>INFINIBAND SWITCH</th>
<th>INFINIBAND PORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head A</td>
<td>ibp0</td>
<td>Active</td>
<td>Leaf Switch #2</td>
<td>5B</td>
</tr>
<tr>
<td>Head A</td>
<td>ibp1</td>
<td>Standby</td>
<td>Leaf Switch #3</td>
<td>6A</td>
</tr>
<tr>
<td>Head A</td>
<td>ibp2</td>
<td>Active</td>
<td>Leaf Switch #2</td>
<td>6A</td>
</tr>
<tr>
<td>Head A</td>
<td>ibp3</td>
<td>Standby</td>
<td>Leaf Switch #3</td>
<td>5B</td>
</tr>
<tr>
<td>Head B</td>
<td>ibp0</td>
<td>Active</td>
<td>Leaf Switch #2</td>
<td>6B</td>
</tr>
<tr>
<td>Head B</td>
<td>ibp1</td>
<td>Standby</td>
<td>Leaf Switch #3</td>
<td>7A</td>
</tr>
</tbody>
</table>
Configuring Network - Devices, DataLinks and Interfaces

The Sun ZFS Storage Appliance Network Configuration consists of 3 levels, Devices, DataLinks and Interfaces.

- Devices refer to the physical hardware
- DataLinks manage the physical devices
- Interfaces configure IP Addresses on top of DataLinks

Configuring Network DataLinks

Network DataLinks that will be used by the InfiniBand ports are configured as follows

Navigate to the Configuration > Network screen

Either drag the name of the device across to the Datalinks column, or click the + sign next to Datalinks to configure a new DataLink
Figure 9: Adding a new network datalink

- Check the IB Partition box
- For simplicity, the Name of the DataLink is created the same as the name of the device
- Enter “ffff” for the Partition Key to be used
- For optimal performance, select “Connect Mode” from the drop down box of available Link Modes
- Click the radial button associated with the device to be added

The screen should look like this graphic
Backup and Recovery Performance and Best Practices using Oracle Sun ZFS Storage Appliance and Oracle Exadata Database Machine

Repeat this procedure for the other three new DataLinks that will manage the devices.

Configuring Network Interfaces

There are two phases to configuring the Network Interfaces because you will be using IPMP groups for availability

Phase 1 Configuring Network Interfaces – Static Address

Either drag the name of the datalink across to the Interface column or click the + sign to configure a new Interface, similar to that of configuring the datalink

- For simplicity, use Interface name = Datalink namedatalink
- Check the Use IPv4 Protocol box
- Select “Static Address List” from the drop down box of available “Configure with”
- Enter “0.0.0.0/8” into the address box below the Configure with box
• Click the radial button associated with the datalink to be added

The screen should look like

![Network Interface Screen](image)

- Use IPv4 Protocol
  - Enable Interface
  - Allow Administration
  - Configure with: Static Address List
    - 0.0.0.0/8

- Use IPv6 Protocol

<table>
<thead>
<tr>
<th>Datalink</th>
<th>Mode</th>
<th>IP Address</th>
<th>Interface Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ibp0</td>
<td>Link Mode(cm)</td>
<td>via ibp0</td>
<td>ibp0</td>
</tr>
<tr>
<td>ibp1</td>
<td>Link Mode(cm)</td>
<td>via ibp1</td>
<td>ibp1</td>
</tr>
<tr>
<td>ibp2</td>
<td>Link Mode(cm)</td>
<td>via ibp2</td>
<td>ibp2</td>
</tr>
<tr>
<td>ibp3</td>
<td>Link Mode(cm)</td>
<td>via ibp3</td>
<td>ibp3</td>
</tr>
</tbody>
</table>

![Figure 11: Configuring an InfiniBand network interface](image)

Repeat this procedure for the other three new Interfaces that will manage the datalinks.

Phase 2 Configuring Network Interfaces – IP network multipathing (IPMP) Group

Click the + sign to configure a new IPMP interface

• For simplicity, the Name of the Interface is the name that the IP Address resolves to
• Check the Use IPv4 Protocol box
• Select “Static Address List” from the drop down box of available “Configure with”
- Enter the IP Address and Subnet Mask Bits that you want the interface to operate on into the address box below the Configure with box. For example “192.168.41.189/21”

- Check the IP MultiPathing Group box

- Check the two Interfaces (that were created in the previous step), that will be used by the IPMP group. Refer to the section above where the pairs are documented.

- For the two interfaces that have been checked, ensure the Active / Standby drop box is set correctly

![Network Interface](image)

**Properties**

- Name: `scanas401-priv`
- Enable Interface: checked
- Allow Administration: checked

**Use IPv4 Protocol**

Configure with: `Static Address List`

- `192.168.41.189/21`

**Use IPv6 Protocol**

<table>
<thead>
<tr>
<th>Interfaces</th>
<th>IP MultiPathing Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>ibp0</code></td>
<td>Active</td>
</tr>
<tr>
<td><code>ibp1</code></td>
<td>Unused</td>
</tr>
<tr>
<td><code>ibp2</code></td>
<td>Unused</td>
</tr>
<tr>
<td><code>ibp3</code></td>
<td>Standby</td>
</tr>
<tr>
<td><code>scanas401</code></td>
<td>Unused</td>
</tr>
<tr>
<td><code>scanas402</code></td>
<td>Unused</td>
</tr>
</tbody>
</table>

Figure 12: Configuring an InfiniBand IPMP network group
Repeat this procedure for the other InfiniBand IPMP group.

Once all the changes have been made, ensure that you apply the changes on the main Configuration > Network page. Otherwise these changes will be lost if you close the browser session or navigate away from the screen.

Figure 13: Saving the InfiniBand network configuration

After the changes have been applied, you should be able to ping the network IP Addresses that are created.

Network Configuration for 10Gigabit Ethernet Connectivity

To provide the highest availability and performance, it is recommended that each Sun ZFS Storage Appliance controller is configured with two Dual Port 10Gb Optical Ethernet, providing four ports total per Sun ZFS Storage Appliance controller.
In order to achieve this, the Sun ZFS Storage Appliance network configuration will be configured using an Active / Active IPMP group. Additionally, for performance reasons, the 10Gigabit network should be configured for a large MTU size also known as Jumbo Frames.

The two Dual Port 10Gb Optical Ethernet cards will probably be installed into PCIe Slots 3 and 6 in the Sun ZFS Storage Appliance controller, and will be identified as ixgbe0 & ixgbe1 in PCIe Slot 3 and ixgbe2 & ixgbe3 in PCIe Slot 6. You can think of the two slots as the top and bottom port on the vertically installed PCIe Slot. Within the Sun ZFS Storage Appliance User Interface under Configuration > Network, bind ixgbe0 & ixgbe3 as an Active/Active IPMP group, and ixgbe2 & ixgbe1 as a second Active/Active IPMP group.

Cabling the 10Gigabit Ethernet connections

The 10Gigabit Ethernet connections will be connected into the customers supplied 10Gigabit network switches. Therefore, the recommendation will be more general than for InfiniBand cabling. Ideally, you want to connect the two cables of each IPMP group to two different 10Gigabit Network Switches so as to provide availability in the event of a failure of one of the switches. By using IPMP groups, there is no need to configure LACP also known as LAG or 802.3ad on the network switches. Link Aggregation is achieved using Direct NFS available from the Oracle DBMS. It is recommended though to configure jumbo frames on the network switch to allow larger frame packets to be transmitted which in turn reduces the number of acknowledgements that need to be sent between the database machine and the Sun ZFS Storage Appliance. Jumbo Frames needs to be configured by the network administrator responsible for the 10Gigabit network switches.

Connect the cables as described in the following table.

<table>
<thead>
<tr>
<th>SUN ZFS STORAGE APPLIANCE HEAD</th>
<th>SUN ZFS STORAGE APPLIANCE PORT</th>
<th>10GIGABIT NETWORK SWITCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head A</td>
<td>ixgbe0</td>
<td>Switch #1</td>
</tr>
<tr>
<td>Head A</td>
<td>ixgbe1</td>
<td>Switch #2</td>
</tr>
<tr>
<td>Head A</td>
<td>ixgbe2</td>
<td>Switch #1</td>
</tr>
<tr>
<td>Head A</td>
<td>ixgbe3</td>
<td>Switch #2</td>
</tr>
<tr>
<td>Head B</td>
<td>ixgbe0</td>
<td>Switch #1</td>
</tr>
<tr>
<td>Head B</td>
<td>ixgbe1</td>
<td>Switch #2</td>
</tr>
<tr>
<td>Head B</td>
<td>ixgbe2</td>
<td>Switch #1</td>
</tr>
<tr>
<td>Head B</td>
<td>ixgbe3</td>
<td>Switch #2</td>
</tr>
</tbody>
</table>

Configuring Network – Database Nodes in the Exadata Database Machine
In order to achieve the rates documented in this paper, two network configuration changes were made to each of the DB Nodes in the Exadata Database Machine.

- Increase the TX Queue Length on the 10GigE interfaces
- Increase the RX Queue Length in the Linux kernel

**Increase the TX Queue Length on the 10GigE interfaces**

The TX Queue Length is a configuration setting associated with each of the 10GigE interfaces, and is set via the Linux `ifconfig` command. The parameter needs to be set on both the bonded 10GigE interface as well as the underlying 10GigE slave interfaces. To increase the txqueuelen on a running system run the following commands as the root user:

```
# /sbin/ifconfig eth4 txqueuelen 10000
# /sbin/ifconfig eth5 txqueuelen 10000
# /sbin/ifconfig bondeth1 txqueuelen 10000
```

In order to permanently set these configurations as part of a node or system reboot, enter the same commands in the `/etc/rc.local` file on the DB Node.

**Increase the RX Queue Length in the Linux kernel**

The RX Queue Length is a `sysctl` parameter, and is set via the `sysctl -w` command or via the `/etc/sysctl.conf` configuration file. To set the parameter on a running system run the following command as the root user.

```
# /sbin/sysctl -w net.core.netdev_max_backlog=2000
```

In order to permanently set this parameter as part of a node or system reboot, enter the parameter and value in the `/etc/sysctl.conf` file on the DB Node.

**Configuring Network - Devices, DataLinks and Interfaces**

The Sun ZFS Storage Appliance Network Configuration consists of 3 levels, Devices, DataLinks and Interfaces.

- Devices refer to the physical hardware
- DataLinks manage the physical devices
- Interfaces configure IP Addresses on top of DataLinks

**Configuring Network DataLinks**

Network DataLinks that will be used by the 10Gigabit ports are configured as follows.

Either drag the name of the datalink across to the Interface column or click the + sign to configure a new DataLink.
• Leave the VLAN and IB Partition boxes unchecked
• For simplicity, the Name of the DataLink is created the same as the name of the device
• Select the desired values for Link Speed and Link Duplex from the respective drop down boxes
• Click the radial button associated with the device to be added
Repeat this procedure for the other three new DataLinks that will manage the devices.

**Configuring Network Interfaces**

There are two phases to configuring the Network Interfaces because we will be using IPMP groups for availability.

**Phase 1  Configuring Network Interfaces – Static Address**

Click the + sign to configure a new Interface

- For simplicity, the Name of the Interface is create the same as the name of the datalink
- Check the Use IPv4 Protocol box
- Select “Static Address List” from the drop down box of available “Configure with”
- Enter “0.0.0.0/8” into the address box below the Configure with box
- Click the radial button associated with the datalink to be added
Repeat this procedure for the other three new Interfaces that will manage the datalinks.

Phase 2 Configuring Network Interfaces – IPMP Group

Click the + sign to configure a new IPMP interface
- For simplicity, the Name of the Interface is the first name that the IP Address resolves to
- Check the Use IPv4 Protocol box
- Select “Static Address List” from the drop down box of available “Configure with”
- Enter the IP Address and Subnet Mask Bits that you want the interface to operate on into the address box below the Configure with box. For example “192.168.41.129/21”
• Click the + sign to add a second IP Address and enter the IP Address and Subnet Mask Bits that you want the second interface to operate on. For example “192.168.41.130/21”

• Check the IP MultiPathing Group box

• Check the two Interfaces (that were created in the previous step), that will be used by the IPMP group. Refer to the section above where the pairs are documented.

• For the two interfaces that have been checked, ensure that both interfaces are marked as Active from the drop box
Once all the changes have been made, ensure that you apply the changes on the main Configuration > Network page. Otherwise these changes will be lost if you close the browser session or navigate away from the screen.
After the changes have been applied, you should be able to ping both network IP Addresses that are created for each interface.

**Configuring network routing**

Select the routing tab and enable adaptive routing to ensure that read requests from the Sun ZFS Storage Appliance from IP addresses local to the subnet are serviced from the physical interface that received the request. In cases where the Exadata Database Machine and Sun ZFS Storage Appliance are on different subnets add a network route to the subnet containing the Exadata 10Gb Ethernet interfaces from the Sun ZFS Storage Appliance 10Gb Ethernet interfaces. This should be done on both Sun ZFS Storage Appliance heads in a cluster configuration.

Sophisticated routing, traffic segregation, and quality of service (QoS) filters can affect system throughput. Such tuning is beyond the scope of this paper. Please consult with your network administrator regarding these details.
Configuring Jumbo Frames

The Sun ZFS Storage Appliance User Interface does not allow jumbo frames to be configured, but this can be configured using the Sun ZFS Storage Appliance CLI. Login to the Sun ZFS Storage Appliance using ssh protocol, and configure the MTU size as follows:

- Use the command “configuration net datalinks select ixgbe0 show” to show the current definition for the ixgbe0 interface. There will be four 10Gigabit datalinks if you “configuration net datalinks show”

- Use the command “configuration net datalinks select ixgbe0 set mtu=9000” to set the Maximum Transmission Unit to 9000 bytes. The network engineer will be able to advise what size MTU should be used on your particular network. Repeat this for the three other datalinks that were previously configured.

```
scnas401:~> configuration net datalinks select ixgbe0 show
Properties:
    class = device
    label = ixgbe0
    mac = 0:1b:31:96:e2:60
    links = ixgbe0
    mtu = 1500
    speed = auto
    duplex = auto

scnas401:~> configuration net datalinks select ixgbe1 set mtu=5000
mtu = 5000

scnas401:~> configuration net datalinks select ixgbe2 set mtu=5000
mtu = 9000

scnas401:~> configuration net datalinks select ixgbe3 set mtu=5000
mtu = 9000

scnas401:~> configuration net datalinks select ixgbe0 show
Properties:
    class = device
    label = ixgbe0
    mac = 0:1b:31:96:e2:60
    links = ixgbe0
    mtu = 9000
    speed = auto
    duplex = auto
```

Figure 14: Configuring MTU size on 10 Gigabit datalinks
Configuring Network Cluster Resources

For the purpose of this document, we will describe the procedure for 10 Gigabit Network interfaces, but the procedure is the same for InfiniBand interfaces.

Navigate to the Configuration > Cluster page, and you will see the 6 network interfaces that were previous configured already assigned to the node scansas401. Additionally the management network interfaces are shown running on their respective systems.

Figure 15: Cluster Configuration initial screen

The first step is to issue a cluster takeover, which will reboot the “other” system in the cluster, forcing the resources owned by the other system to be mastered by the local system. With the other system restarting, you can now apply the three resources that will ultimately run on the “other” system. On clicking Apply, a confirmation dialogue will be displayed, with an option to “Apply” or “Failback” the resources. If the “other” system is back, you can elect to “Failback”, but click “Apply” if the system is not back.
Backup and Recovery Performance and Best Practices using Oracle Sun ZFS Storage Appliance and Oracle Exadata Database Machine

Figure 16: Assigning 10 Gigabit resources to "other" system
Once the “other” system has been restarted, the Sun ZFS Storage Appliance UI will show the “other” system as “Ready (waiting for failback)”. All the system resources are running on the local system, and it by clicking failback, the resources will be distributed back to the “other” system.

Figure 17: Failing cluster resources back to owning node

On the successful failback of the resources to the owning node, the Sun ZFS Storage Appliance UI will show the resources correctly distributed.
Figure 18: After cluster network resource configuration

You should be able to logon to the Exadata Database Machine, and “ping” the IP Address of the resources. For 10gigabit networks, you should be able to ping all four interfaces, and for the InfiniBand networks, you should be able to ping both interfaces.

Storage configuration

Login to the first system, and navigate to the Configuration > Storage page, and click the + sign next to Available Pools.
The configuration scripts expect to find the pools to be called pool1 and pool2. On clicking the + sign, enter pool1 in the pop window for the name of the pool.

A two-page dialogue then is displayed. On the first page, select the number of data devices, log devices and cache devices that are to be mapped to the first pool. As two equal pools will be created, allocate the number of resources evenly, and then click commit.
On the second page, there are three profile tabs that need to be completed, Data, Log, and Cache. This paper is focused on simple Oracle Database Backup and Recovery, and not on using the Sun ZFS Storage Appliance for the purpose of testing via Snapshoting and Cloning of data. Therefore, the recommendation is to select “Single Parity, narrow strips” which offers good performance while still maintaining a reasonable amount of space.
If Log devices are present, then if possible select “Mirrored Logs” which offer NSPF, otherwise select “Mirrored Logs”. There is nothing that needs to be selected on the Cache tab. After selecting the appropriate profiles, press the commit button. The Storage Pool will then be created and the space available will be shown on the “Available Pools” screen that is displayed.

<table>
<thead>
<tr>
<th>Data Profile</th>
<th>Log Profile</th>
<th>Cache Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TYPE</strong></td>
<td>NSPF</td>
<td>AVAILABILITY</td>
</tr>
<tr>
<td>Double parity</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Double parity</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Mirrored</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Mirrored</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Single parity, narrow stripes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Striped</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Triple mirrored</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Triple mirrored</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Triple parity, wide stripes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Triple parity, wide stripes</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

**Data profile: Single parity, narrow stripes**

Each narrow stripe assigns one parity disk for each set of three data disks, offering better random read performance than wider double parity stripes and moderately increased capacity over mirrored configurations. Workloads characterized by few random accesses can be suitable for this profile. However for large sequential datasets, double parity provides faster throughput and more availability.
Login to the other system and repeat the steps to create the second pool called pool2. Once completed, navigate to the Configuration > Cluster page. You should see something similar to the output below, showing that each system is the owner of a Sun ZFS Storage Appliance Pool as well as the respective network interfaces that were previously configured.
Sample /etc/oranfstab files

Included below are the two oranfstab files, one for 10Gigabit Ethernet and the other for InfiniBand, for a database with a name of “orcl”.

The following sample files are for one particular node in a cluster. The local IP Address referenced in the samples refers to the IP Address associated with the interface on the DB Node that the data is to be sent over. Ensure the file on each node in the cluster correctly identifies the local IP Address.

10Gigabit /etc/oranfstab file

server: scanas401
local: 10.133.62.27 path: 10.133.62.15
dontroute
export: /export/orcl/backup1 mount: /zfssa/orcl/backup1
export: /export/orcl/backup2 mount: /zfssa/orcl/backup2
export: /export/orcl/backup3 mount: /zfssa/orcl/backup3
export: /export/orcl/backup4 mount: /zfssa/orcl/backup4
export: /export/orcl/backup5 mount: /zfssa/orcl/backup5
export: /export/orcl/backup6 mount: /zfssa/orcl/backup6
export: /export/orcl/backup7 mount: /zfssa/orcl/backup7
export: /export/orcl/backup8 mount: /zfssa/orcl/backup8

server: scanas402
local: 10.133.62.27 path: 10.133.62.16
dontroute
export: /export/orcl/backup9 mount: /zfssa/orcl/backup9
export: /export/orcl/backup10 mount: /zfssa/orcl/backup10
export: /export/orcl/backup11 mount: /zfssa/orcl/backup11
export: /export/orcl/backup12 mount: /zfssa/orcl/backup12
export: /export/orcl/backup13 mount: /zfssa/orcl/backup13
export: /export/orcl/backup14 mount: /zfssa/orcl/backup14
export: /export/orcl/backup15 mount: /zfssa/orcl/backup15
export: /export/orcl/backup16 mount: /zfssa/orcl/backup16

Note that there are two local / path parameter combinations for the 10Gigabit network configuration, because the Sun ZFS Storage Appliance is utilizing Active/Active IPMP groups.

InfiniBand /etc/oranfstab file (for an Exadata Database Machine X2-8)

server: scanas401
local: 192.168.20.208 path: 192.168.20.189
dontroute:
export: /export/orcl/backup1 mount: /zfssa/orcl/backup1
export: /export/orcl/backup2 mount: /zfssa/orcl/backup2
export: /export/orcl/backup3 mount: /zfssa/orcl/backup3
export: /export/orcl/backup4 mount: /zfssa/orcl/backup4
export: /export/orcl/backup5 mount: /zfssa/orcl/backup5
export: /export/orcl/backup6 mount: /zfssa/orcl/backup6
export: /export/orcl/backup7 mount: /zfssa/orcl/backup7
export: /export/orcl/backup8 mount: /zfssa/orcl/backup8
server: scanas402
local: 192.168.20.208 path: 192.168.20.191
local: 192.168.20.209 path: 192.168.20.191
local: 192.168.20.211 path: 192.168.20.191
dontroute:
export: /export/orcl/backup9 mount: /zfssa/orcl/backup9
export: /export/orcl/backup10 mount: /zfssa/orcl/backup10
export: /export/orcl/backup11 mount: /zfssa/orcl/backup11
export: /export/orcl/backup12 mount: /zfssa/orcl/backup12
export: /export/orcl/backup13 mount: /zfssa/orcl/backup13
export: /export/orcl/backup14 mount: /zfssa/orcl/backup14
export: /export/orcl/backup15 mount: /zfssa/orcl/backup15
export: /export/orcl/backup16 mount: /zfssa/orcl/backup16

Note that there are four sets of local / path parameters on an Exadata Database Machine X2-8, because there are four InfiniBand ports on each DB Node.

Test Network and Throughput Utilization

Having completed the setup of the system, configuring the Sun ZFS Storage Appliance, running the Exadata Backup Configuration Utility and enabling Direct NFS on the database, it is possible to run a test backup to verify network configuration and throughput utilization.

Using the sample level 0 backup script (see below), modify the backup clause to restrict the backup to a relatively small subset of the database consisting of between 500GB and 1TB of database objects.

Before running the RMAN backup, initiate the following monitoring:

- On each of the DB Nodes run the command
  \[ \text{sar} \ -n \ \text{DEV} \ 1 \ 1000 \]
- On each of the Sun ZFS Storage Appliance controller heads, open an Analytics screen that is monitoring
• Network: interface bytes per second broken down by interface

When the RMAN backup script is run, you should check that all the network interfaces that are expected to be used, as defined by the local / path parameter pairing in the oranfstab file are being utilized.

Additionally, execute an RMAN “restore validate” command to ensure that when reading from the Sun ZFS Storage Appliance, that all the network interfaces are still be used.

Appendix B Database Backup Sample Scripts

Database weekly level 0 backup

The script generated by the Oracle Exadata Backup Configuration Utility for the level 0 backupset will look like the following example.

```sql
run
{
  sql 'alter system set "_backup_disk_bufcnt"=64 scope=memory';
  sql 'alter system set "_backup_disk_bufsz"=1048576 scope=memory';

  allocate channel ch01 device type disk connect
  'system/welcome1@dm01-scan/orcl_bkup1' format
  '/zfssa/orcl/backup1/%U';
  allocate channel ch02 device type disk connect
  'system/welcome1@dm01-scan/orcl_bkup2' format
  '/zfssa/orcl/backup9/%U';
  allocate channel ch03 device type disk connect
  'system/welcome1@dm01-scan/orcl_bkup3' format
  '/zfssa/orcl/backup2/%U';
  allocate channel ch04 device type disk connect
  'system/welcome1@dm01-scan/orcl_bkup4' format
  '/zfssa/orcl/backup10/%U';
  allocate channel ch05 device type disk connect
  'system/welcome1@dm01-scan/orcl_bkup5' format
  '/zfssa/orcl/backup3/%U';
  allocate channel ch06 device type disk connect
  'system/welcome1@dm01-scan/orcl_bkup6' format
  '/zfssa/orcl/backup11/%U';
```

allocate channel ch07 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup7' format
'/zfssa/orcl/backup4/%U';
allocate channel ch08 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup8' format
'/zfssa/orcl/backup12/%U';
allocate channel ch09 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup1' format
'/zfssa/orcl/backup5/%U';
allocate channel ch10 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup2' format
'/zfssa/orcl/backup13/%U';
allocate channel ch11 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup3' format
'/zfssa/orcl/backup6/%U';
allocate channel ch12 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup4' format
'/zfssa/orcl/backup14/%U';
allocate channel ch13 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup5' format
'/zfssa/orcl/backup7/%U';
allocate channel ch14 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup6' format
'/zfssa/orcl/backup15/%U';
allocate channel ch15 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup7' format
'/zfssa/orcl/backup16/%U';
allocate channel ch16 device type disk connect
'system/welcome1@dm01-scan/orcl_bkup8' format
'/zfssa/orcl/backup16/%U';
configure snapshot controlfile name to
'/zfssa/orcl/backup1/snapcf_orcl.f';
backup as backupset incremental level 0 section size 32g database
tag 'FULLBACKUPSET_L0' plus archivelog tag 'FULLBACKUPSET_L0';

The script starts by setting two init.ora parameters which optimize the writing of data to the Sun ZFS Storage Appliance. It then proceeds to allocate 16 RMAN Channels, each one being told to write the
backupsets to a specific share on the Sun ZFS Storage Appliance. The automatic controlfile snapshot that is created by RMAN is also pointed to the first share on the Sun ZFS Storage Appliance.

Then the main backup command is run. The important parts are the explicit creation of an incremental level 0 backupset, and that any datafile that is greater than 32GB is broken up into chunks of 32GB, so as to better utilize all the resources. The script also backs up the archivelogs that are present on the system.

Database daily level 1 backups

The script generated by the Oracle Exadata Backup Configuration Utility for the level 1 backupset is very similar to that for the level 0 backup. The only difference is in the main backup command that is run.

RMAN> backup as backupset incremental level 1 database tag 'FULLBACKUPSET_L1' plus archivelog tag 'FULLBACKUPSET_L1';