ORACLE®
EXADATA

A Technical Overview of the Oracle Exadata Database Machine and Exadata Storage Server

ORACLE®

Disclaimer

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Introduction

The Oracle Exadata Database Machine is an easy to deploy solution for hosting the Oracle Database that delivers the highest levels of database performance available. The Exadata Database Machine is a "cloud in a box" composed of database servers, Oracle Exadata Storage Servers, an InfiniBand fabric for storage networking and all the other components required to host an Oracle Database. It delivers outstanding I/O and SQL processing performance for online transaction processing (OLTP), data warehousing (DW) and consolidation of mixed workloads. Extreme performance is delivered for all types of database applications by leveraging a massively parallel grid architecture using Real Application Clusters and Exadata storage. Database Machine and Exadata storage delivers breakthrough performance with linear I/O scalability, is simple to use and manage, and delivers mission-critical availability and reliability.

The Exadata Storage Server is an integral component of the Exadata Database Machine. Extreme performance is delivered by several features of the product. Exadata storage provides database aware storage services, such as the ability to offload database processing from the database server to storage, and provides this while being transparent to SQL processing and database applications. Hence just the data requested by the application is returned rather than all the data in the queried tables. Exadata Smart Flash Cache dramatically accelerates Oracle Database processing by speeding I/O operations. The Flash provides intelligent caching of database objects to avoid physical I/O operations. The Oracle Database on the Database Machine is the first Flash enabled database. Exadata storage provides an advanced compression technology, Exadata Hybrid Columnar Compression, that typically provides 10x, and higher, levels of data compression. Exadata compression boosts the effective data transfer by an order of magnitude. The Oracle Exadata Database Machine is the world's most secure database machine. Building on the superior security capabilities of the Oracle Database, the Exadata storage provides the ability to query fully encrypted databases with near zero overhead at hundreds of gigabytes per second. The combination of these, and many other, features of the product are the basis of the outstanding performance of the Exadata Database Machine.

The Exadata Database Machine has also been designed to work with, or independently of, the Oracle Exalogic Elastic Cloud. The Exalogic Elastic Cloud provides the best

platform to run Oracle's Fusion Middleware and Oracle's Fusion applications. The combination of Exadata and Exalogic is a complete hardware and software engineered solution that delivers high-performance for all enterprise applications including Oracle E-Business Suite, Siebel, and PeopleSoft applications.

Exadata Product Family

The foundation of the Exadata family of products is the Oracle Exadata Database Machine (Database Machine). The Database Machine is a complete and fully integrated database system that includes all the components to quickly and easily deploy any enterprise database delivering the best performance. The Exadata Storage Server (Exadata storage or Exadata cells) is used as the storage for the Oracle Database in the Database Machine and is used to grow existing Database Machine deployments.

Exadata Database Machine

The Database Machine is a pre-configured system ready to be turned on day one, taking significant integration work, cost and time out of the database deployment process. Since it is a well known configuration Oracle Support is very familiar with how to service the system resulting in a superior support experience with the system. The benefit of a common infrastructure to deploy a database for any application, whether OLTP, DW, a mix of the two, or as a platform for consolidation of several databases, creates tremendous opportunities for efficiencies in the datacenter. It is truly a "cloud in box".

Exadata Database Machine X2-8

There are two versions of Exadata Database Machine. The Exadata Database Machine X2-2 expands from 2 twelve-core database servers with 192 GB of memory and 3 Exadata Storage Servers to 8 twelve-core database servers with 768 GB of memory and 14 Exadata Storage Servers, all in a single rack. The Exadata Database Machine X2-8 is comprised of 2 sixty four-core database servers with 2 TB of memory and 14 Exadata Storage Servers, in a single rack. The X2-2 provides a convenient entry point in to the Exadata Database Machine family with the largest degree of expandability in a single rack. The X2-8 is for large deployments with large

memory requirements or a need to consolidate many databases on to a single rack. Both versions run the Oracle Database 11g Release 2 database software.

Exadata Database Machine X2-2

Three versions of the Exadata Database Machine X2-2 are available – the *Full Rack*, *Half Rack,* and *Quarter Rack* – depending on the size, performance and I/O requirements of the database to be deployed. One version can be upgraded online to another ensuring a smooth upgrade path as processing requirements grow. Common to all X2-2 Database Machines are:

- Industry standard Oracle Database 11g database servers preconfigured with: two six-core Intel® Xeon® X5670 processors running at 2.93 GHz, 96 GB memory, four 300 GB 10,000 RPM SAS disks, two 40 Gb/second InfiniBand ports, two 10 Gb/second Ethernet ports, four 1 Gb/second Ethernet ports, and dual-redundant, hot-swappable power supplies. Oracle Linux 5 Update 5 is preinstalled on the database servers as the operating system for the Database Machine. Solaris 11 Express (x86) will be an alternative operating system on the database servers, available in the future.

- Exadata Storage Servers preconfigured with: two socket six-core Intel Xeon L5640 processors running at 2.26 GHz, 24 GB memory, 384 GB of Exadata Smart Flash Cache, twelve SAS disks connected to a storage controller with 512MB battery-backed cache, dual port InfiniBand connectivity, embedded Integrated Lights Out Manager (ILOM) and dual-redundant, hot-swappable power supplies. The Exadata Storage Servers are available with either 600 GB High Performance SAS disks or 2 TB High Capacity SAS disks. All the Exadata Storage Server Software is preinstalled on the Exadata cell.

- Sun Quad Data Rate (QDR) InfiniBand switches and cables to form a 40 Gb/second InfiniBand fabric for database server to Exadata Storage Server communication, and RAC internode communication.

- Ethernet switch for remote administration and monitoring of the Database Machine.

- Keyboard, Video or Visual Display Unit, Mouse (KVM) hardware for local administration of the Database Machine.

- All of these components are packaged in to a custom 42U rack including the Power Distribution Units (PDU) for the system.

The ratio of components to each other has been chosen to maximize performance, deliver a highly available system and provide the best balance of CPU to I/O power for all database applications. The hardware components in each version of the Exadata Database Machine X2-2 are shown in the following table.

	Database Machine X2-2 Full Rack	Database Machine X2-2 Half Rack	Database Machine X2-2 Quarter Rack
Database Servers	8	4	2
Exadata Storage Servers	14	7	3
InfiniBand Switches	3	3	2

Database Machine X2-2 Components

Exadata Database Machine X2-8

The new top-of-the-line Exadata Database Machine X2-8 combines the best of scale-up and scale-out architectures by delivering a grid architecture containing large SMP database servers. Before now, a 64-core SMP required a full rack of equipment by itself, and was difficult to scale out further. The Exadata X2-8 uses two of Sun's new ultra-compact 64-core Intel-based servers to create a high-performance highly-available database grid. Each of the servers includes a terabyte of memory, 40 Gb/second InfiniBand for internal connectivity, and 10 Gb/second Ethernet for connectivity to the data center. The X2-8 has the same storage grid architecture as the X2-2 with 14 Exadata Storage Servers providing intelligent query offload, 10X data compression, 336 TB of raw storage, and up to a 1 million I/Os per second to 5.3 TB of high performance PCI flash. The Exadata X2-8 can be easily expanded to an 8 rack grid with 2368 CPU cores, and 2.6 petabytes of raw storage. The new Exadata X2-8 delivers extreme performance for all business applications, and enables large-scale database consolidation.

The Exadata Database Machine X2-8 is available in a full rack configuration, runs Oracle Database 11g Release 2, and includes the following technology.

- Two industry standard database servers each preconfigured with: eight socket eight -core Intel® Xeon® X7560 processors running at 2.26 GHz, 1 TB memory, eight 300 GB 10,000 RPM SAS disks, eight 40 Gb/second InfiniBand ports, eight 10 Gb/second Ethernet ports, eight 1 Gb/second Ethernet ports, and dual-redundant, hot-swappable power supplies. Oracle Linux 5 Update 5 with the Unbreakable Enterprise Kernel is preinstalled on the database servers as the operating system for the Database Machine. Solaris 11 Express (x86) will be an alternative operating system on the database servers, available in the future.

- Fourteen Exadata Storage Servers preconfigured with: two socket six-core Intel Xeon L5640 processors running at 2.26 GHz, 24 GB memory, 384 GB of Exadata Smart Flash Cache, twelve SAS disks connected to a storage controller with 512MB battery-backed cache, dual port InfiniBand connectivity, embedded Integrated Lights Out Manager (ILOM) and dual-redundant, hot-swappable power supplies. All of the Exadata Storage Server Software is preinstalled on the Exadata cell.

- Three Sun Quad Data Rate (QDR) InfiniBand switches and cables to form a 40 Gb/second InfiniBand fabric for database server to Exadata Storage Server communication, and RAC internode communication.

- Ethernet switch for remote administration and monitoring of the Database Machine.

- All of these components are packaged in to a custom 42U rack including the Power Distribution Units (PDU) for the system.

Again, the ratio of components to each other has been chosen to maximize performance, deliver a highly available system and provide the best balance of CPU to I/O power for all database applications.

Database Machine Upgradeability

Each model of the Database Machine X2-2 can grow in capacity and power, ensuring a smooth upgrade path, as processing requirements grow. An online field upgrade from the Quarter Rack to the Half Rack and from the Half Rack to Full Rack can be easily performed by Oracle personnel.

Quarter Rack   Half Rack   Full Rack
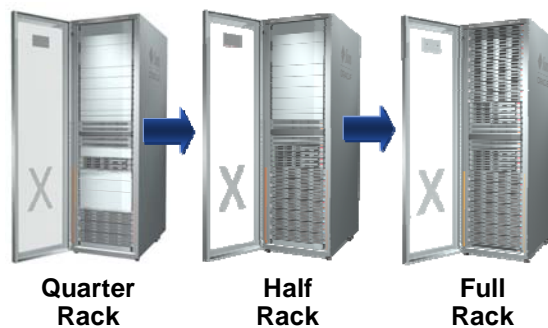
Database Machine X2-2 Upgrades

While an Exadata Database Machine is an extremely powerful system, a building-block approach is used that allows Exadata Database Machines to scale to almost any size. Multiple Database Machine X2-2 Full Rack and Half Rack systems can be connected using the InfiniBand fabric in the system to form a larger single system image configuration. Multiple Exadata Database

Machine X2-8 racks can similarly be connected. This capability is done by connecting InfiniBand cables between the racks as all the InfiniBand infrastructure (switches and port cabling) is designed to provide this growth option. This inherent capability of the Exadata Database Machine to grow enables the support of the largest databases any application would require.

Eight Connected Exadata Database Machine X2-8 Racks Form A Single System

In addition the Exalogic Elastic Cloud connects to an Exadata Database Machine in the same manner using the same InfiniBand fabric. Up to eight full racks of Exalogic and Exadata systems can be connected without the need for any external switches.

Exadata Storage Server

The Exadata Storage Server runs the Exadata Storage Server Software provided by Oracle. The hardware components of the Exadata Storage Server (also referred to as an Exadata *cell*) were carefully chosen to match the needs of high performance database processing. The Exadata software is optimized to take the best possible advantage of the hardware components and Oracle Database. Each Exadata cell delivers outstanding I/O performance and bandwidth to the database.

Building on the high security capabilities in every Oracle Database, the Exadata storage provides the ability to query fully encrypted databases with near zero overhead at hundreds of gigabytes per second. This is done by moving decryption processing from software into the Exadata Storage Server hardware. The Oracle software and the Intel 5600 processors used in the Exadata Storage Server provide Advanced Encryption Standard (AES) support enabling this.

Exadata Storage Server (Exadata Cell)

Exadata Smart Flash Cache

Each Exadata cell comes with 384 GB of Exadata Smart Flash Cache. This means in the Database Machine X2-8 and Full Rack X2-2 there is 5.3 TB of Flash – larger than most databases. This solid state storage delivers dramatic performance advantages with Exadata storage. It provides a ten-fold improvement in response time for reads over regular disk; a hundred-fold improvement in IOPS for reads over regular disk; and is a less expensive higher capacity alternative to memory. Overall it delivers a ten-fold increase performing a blended average of read and write operations.

The Exadata Smart Flash Cache manages active data from regular disks in the Exadata cell – but it is not managed in a simple Least Recently Used (LRU) fashion. The Exadata Storage Server Software in cooperation with the Oracle Database keeps track of data access patterns and knows what and how to cache data and avoid polluting the cache. This functionality is all managed automatically and does not require manual tuning. If there are specific tables or indexes that are known to be key to the performance of a database application they can optionally be identified and pinned in cache.

Exadata Storage Capacity, Performance, Bandwidth and IOPS

The Oracle Exadata Storage Servers comes with either twelve 600 GB 15,000 RPM High Performance SAS disks or twelve 2 TB 7,200 RPM High Capacity SAS disks. The High Performance SAS disk based Exadata Storage Servers provide up to 2 TB of uncompressed user data capacity, and up to 1.8 GB/second of raw data bandwidth. The High Capacity SAS disk based Exadata Storage Servers provide up to 7 TB of uncompressed user data capacity, and up to 1.0 GB/second of raw data bandwidth. When stored in compressed format, the amount of user data and the amount of data bandwidth delivered by each cell significantly increases.

The storage capacity of each model of Database Machine is shown in the following table.

	Database Machine X2-8 and X2-2 Full Rack	Database Machine X2-2 Half Rack	Database Machine X2-2 Quarter Rack
Exadata Smart Flash Cache	5.3 TB	2.6 TB	1.1 TB
Raw Disk Capacity • High Performance SAS • High Capacity SAS	100 TB 336 TB	50 TB 168 TB	21 TB 72 TB
User Data Capacity • High Performance SAS • High Capacity SAS (without data compression)	Up to 28 TB 100 TB	Up to 14 TB 50 TB	Up to 6 TB 21 TB

Database Machine Storage Capacity

User data capacity is an estimate of space available for table rows after mirroring all the disk space, allowing space to recover from disk failures, and setting aside space for database structures like logs, undo, a large TEMP space sufficient for Data Warehousing, and some indexes. This user data capacity represents uncompressed data. When using Exadata Hybrid Columnar Compression the capacity will be many times higher. Actual user data varies by application.

The performance that each cell delivers is extremely high due to the Exadata Smart Flash Cache. The automated caching within Flash enables each Exadata cell to deliver up to 3.6 GB/second bandwidth and 75,000 IOPS when accessing uncompressed data. When data is stored in compressed format, the amount of user data capacity, the amount of data bandwidth and IOPS achievable, often increases up to ten times, or more. This represents a significant improvement over traditional storage devices used with the Oracle Database.

The performance characteristics of each model of Database Machine are depicted in the following table.

	Database Machine X2-8 and X2-2 Full Rack	Database Machine X2-2 Half Rack	Database Machine X2-2 Quarter Rack
Raw Disk Data Bandwidth • **High Performance SAS** • **High Capacity SAS** (without data compression)	Up to 25 GB/sec 14 GB/sec	Up to 12.5 GB/sec 7.0 GB/sec	Up to 5.4 GB/sec 3.0 GB/sec
Effective Data Bandwidth with Flash Cache	Up to 50 GB/sec	Up to 25 GB/sec	Up to 11 GB/sec
Flash Cache IOPS	Up to 1,000,000	Up to 500,000	Up to 225,000
Disk IOPS • **High Performance SAS** • **High Capacity SAS**	Up to 50,000 25,000	Up to 25,000 12,500	Up to 10,800 5,400

Database Machine I/O Performance

Exadata Database Machine Architecture

In the figure below is a simplified schematic of a typical Database Machine Half Rack deployment. Two Oracle Databases, one Real Application Clusters (RAC) database deployed across three database servers and one single-instance database deployed on the remaining database server in the Half Rack, are shown. (Of course all four database servers could be used for a single four node RAC cluster.) The RAC database might be a production database and the single-instance database might be for test and development. Both databases are sharing the seven Exadata cells in the Half Rack but they would have separate Oracle homes to maintain software independence. All the components for this configuration – database servers, Exadata cells, InfiniBand switches and other support hardware are housed in the Database Machine rack.


Database Machine Half Rack Deployment
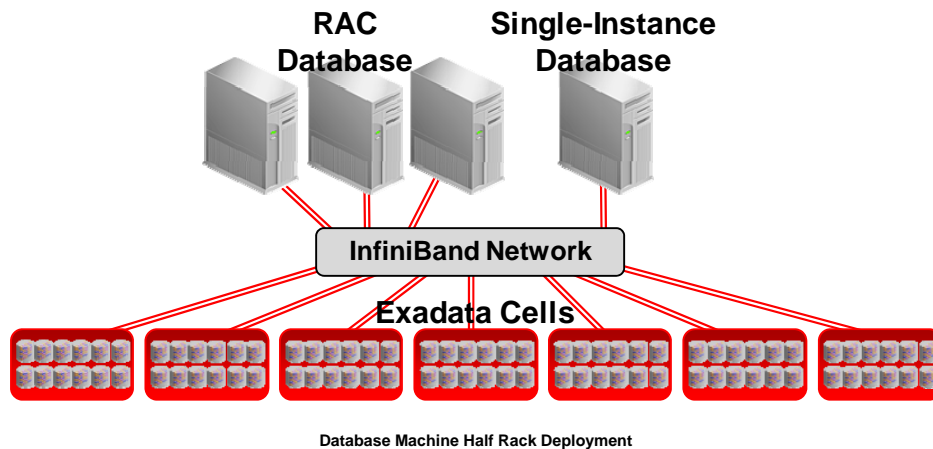
The Database Machine uses a state of the art InfiniBand interconnect between the servers and storage. Each database server and Exadata cell has dual port Quad Data Rate (QDR) InfiniBand connectivity for high availability. Each InfiniBand link provides 40 Gigabits of bandwidth – many times higher than traditional storage or server networks. Further, Oracle's interconnect protocol uses direct data placement (DMA – direct memory access) to ensure very low CPU overhead by directly moving data from the wire to database buffers with no extra data copies being made. The InfiniBand network has the flexibility of a LAN network, with the efficiency of a SAN. By using an InfiniBand network, Oracle ensures that the network will not bottleneck performance. The same InfiniBand network also provides a high performance cluster interconnect for the Oracle Database Real Application Cluster (RAC) nodes.

Oracle Exadata is architected to scale-out to any level of performance. To achieve higher performance and greater storage capacity, additional database servers and Exadata cells are added to the configuration – e.g., Half Rack to Full Rack upgrade. As more Exadata cells are added to

the configuration, storage capacity and I/O performance increases near linearly. No cell-to-cell communication is ever done or required in an Exadata configuration.

The architecture of the Exadata solution includes components on the database server and in the Exadata cell. The software architecture for a Quarter Rack configuration is shown below.
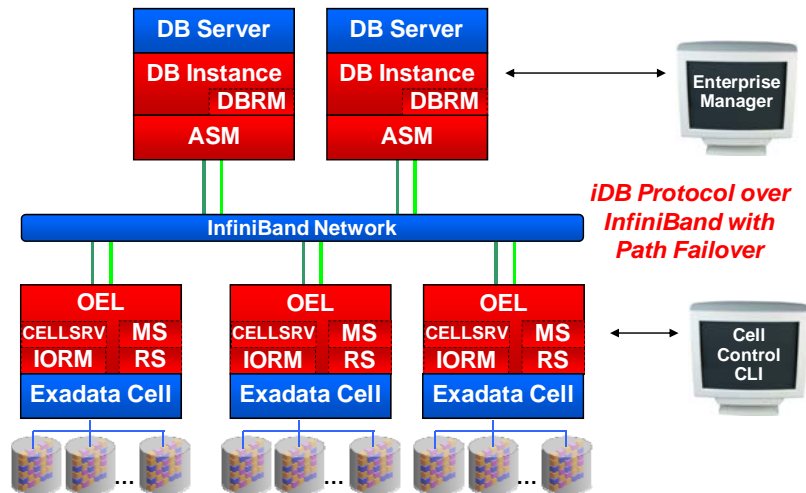
Exadata Software Architecture

When using Exadata, much SQL processing is offloaded from the database server to the Exadata cells. Exadata enables function shipping from the database instance to the underlying storage in addition to providing traditional block serving services to the database. One of the unique things the Exadata storage does compared to traditional storage is return only the rows and columns that satisfy the database query rather than the entire table being queried. Exadata pushes SQL processing as close to the data (or disks) as possible and gets all the disks operating in parallel. This reduces CPU consumption on the database server, consumes much less bandwidth moving data between database servers and storage servers, and returns a query result set rather than entire tables. Eliminating data transfers and database server workload can greatly benefit data warehousing queries that traditionally become bandwidth and CPU constrained. Eliminating data transfers can also have a significant benefit on online transaction processing (OLTP) systems that often include large batch and report processing operations.

Exadata is totally transparent to the application using the database. The exact same Oracle Database 11g Release 2 that runs on traditional systems runs on the Database Machine – but on Database Machine it runs faster. Existing SQL statements, whether ad hoc or in packaged or custom applications, are unaffected and do not require any modification when Exadata storage is

used. The offload processing and bandwidth advantages of the solution are delivered without any modification to the application. All features of the Oracle Database are fully supported with Exadata. Exadata works equally well with single-instance or Real Application Cluster deployments of the Oracle Database. Functionality like Oracle Data Guard, Oracle Recovery Manager (RMAN), Oracle GoldenGate, and other database tools are administered the same, with or without Exadata. Users and database administrators leverage the same tools and knowledge they are familiar with today because they work just as they do with traditional non-Exadata storage.

Since the same Oracle Database and functionality exist on the Database Machine as on traditional systems, the IT staff managing a Database Machine must possess similar knowledge about this same software they will manage on the Database Machine. Oracle Database administration, backup and recovery, RAC and OEL experience are important to possess when managing a Database Machine.

Database Server Software

Oracle Database 11g Release 2 has been significantly enhanced to take advantage of Exadata storage. The Exadata software is optimally divided between the database servers and Exadata cells. The database servers and Exadata Storage Server Software communicate using the iDB – the Intelligent Database protocol. iDB is implemented in the database kernel and transparently maps database operations to Exadata-enhanced operations. iDB implements a function shipping architecture in addition to the traditional data block shipping provided by the database. iDB is used to ship SQL operations down to the Exadata cells for execution and to return query result sets to the database kernel. Instead of returning database blocks, Exadata cells return only the rows and columns that satisfy the SQL query. Like existing I/O protocols, iDB can also directly read and write ranges of bytes to and from disk so when offload processing is not possible Exadata operates like a traditional storage device for the Oracle Database. But when feasible, the intelligence in the database kernel enables, for example, table scans to be passed down to execute on the Exadata Storage Server so only requested data is returned to the database server.

iDB is built on the industry standard Reliable Datagram Sockets (RDSv3) protocol and runs over InfiniBand. ZDP (Zero-loss Zero-copy Datagram Protocol), a zero-copy implementation of RDS, is used to eliminate unnecessary copying of blocks. Multiple network interfaces can be used on the database servers and Exadata cells. This is an extremely fast low-latency protocol that minimizes the number of data copies required to service I/O operations.

Oracle Automatic Storage Management (ASM) is used as the file system and volume manager for Exadata. ASM virtualizes the storage resources and provides the advanced volume management and file system capabilities of Exadata. Striping database files evenly across the available Exadata cells and disks results in uniform I/O load across all the storage hardware. The ability of ASM to perform non-intrusive resource allocation, and reallocation, is a key enabler of the shared grid storage capabilities of Exadata environments. The disk mirroring provided by ASM, combined

with hot swappable Exadata disks, ensure the database can tolerate the failure of individual disk drives. Data is mirrored across cells to ensure that the failure of a cell will not result in loss of data, or inhibit data accessibility. This massively parallel architecture delivers unbounded scalability and high availability.

The Database Resource Manager (DBRM) feature in Oracle Database 11g has been enhanced for use with Exadata. DBRM lets the user define and manage intra and inter-database I/O bandwidth in addition to CPU, undo, degree of parallelism, active sessions, and the other resources it manages. This allows the sharing of storage between databases without fear of one database monopolizing the I/O bandwidth and impacting the performance of the other databases sharing the storage. Consumer groups are allocated a percent of the available I/O bandwidth and the DBRM ensures these targets are delivered. This is implemented by the database tagging I/O with the associated database and consumer group. This provides the database with a complete view of the I/O priorities through the entire I/O stack. The intra-database consumer group I/O allocations are defined and managed at the database server. The inter-database I/O allocations are defined within the software in the Exadata cell and managed by the I/O Resource Manager (IORM). The Exadata cell software ensures that inter-database I/O resources are managed and properly allocated within, and between, databases. Overall, DBRM ensures each database receives its specified amount of I/O resources and user defined SLAs are met.

Two new features of the Oracle Database that are offered exclusively on the Exadata Database Machine are the Oracle Database Quality of Service (QoS) Management and the QoS Management Memory Guard features. QoS Management allows system administrators to directly manage application service levels hosted on Oracle Exadata Database Machines. Using a policy-based architecture, QoS Management correlates accurate run-time performance and resource metrics, analyzes this data with its expert system to identify bottlenecks, and produces recommended resource adjustments to meet and maintain performance objectives under dynamic load conditions. Should sufficient resources not be available QoS will preserve the more business critical objectives at the expense of the less critical ones. In conjunction with Cluster Health Monitor, QoS Management's Memory Guard detects nodes that are at risk of failure due to memory over-commitment. It responds by automatically preventing new connections thus preserving existing workloads and restores connectivity once the sufficient memory is again available.

Enterprise Manager Plug-In For Exadata

Exadata has been integrated with the Oracle Enterprise Manager (EM) Grid Control to easily monitor the Exadata environment. By installing an Exadata plug-in to the existing EM system, statistics and activity on the Exadata server can be monitored, and events and alerts can be sent to the system administrator. The advantages of integrating the EM system with Exadata include:

- Monitoring Oracle Exadata storage

- Gathering storage configuration and performance information

- Raising alerts and warnings based on thresholds

- Providing rich out-of-box metrics and reports based on historical data

All the functions users have come to expect from the Oracle Enterprise Manager work along with Exadata. By using the EM interface, users can easily manage the Exadata environment along with other Oracle Database environments traditionally used with the Enterprise Manager. DBAs can use the familiar EM interface to view reports to determine the health of the Exadata system, and manage the configuration of the Exadata storage.

## Exadata Storage Server Software

Like any storage device the Exadata Storage Server is a computer with CPUs, memory, a bus, disks, NICs, and the other components normally found in a server. It also runs an operating system (OS), which in the case of Exadata is Oracle Enterprise Linux (OEL) 5.3. The Exadata Storage Server Software resident in the Exadata cell runs under OEL. OEL is accessible in a restricted mode to administer and manage the Exadata cell.

CELLSRV (Cell Services) is the primary component of the Exadata software running in the cell and provides the majority of Exadata storage services. CELLSRV is multi-threaded software that communicates with the database instance on the database server, and serves blocks to databases based on the iDB protocol. It provides the advanced SQL offload capabilities, serves Oracle blocks when SQL offload processing is not possible, and implements the DBRM I/O resource management functionality to meter out I/O bandwidth to the various databases and consumer groups issuing I/O.

Two other components of Oracle software running in the cell are the Management Server (MS) and Restart Server (RS). The MS is the primary interface to administer, manage and query the status of the Exadata cell. It works in cooperation with the Exadata cell command line interface (CLI) and EM Exadata plug-in, and provides standalone Exadata cell management and configuration. For example, from the cell, CLI commands are issued to configure storage, query I/O statistics and restart the cell. Also supplied is a distributed CLI so commands can be sent to multiple cells to ease management across cells. Restart Server (RS) ensures the ongoing functioning of the Exadata software and services. It is used to update the Exadata software. It also ensures storage services are started and running, and services are restarted when required.

Exadata Smart Scan Processing

With traditional, non-iDB aware storage, all database intelligence resides in the database software on the server. To illustrate how SQL processing is performed in this architecture an example of a table scan is shown below.
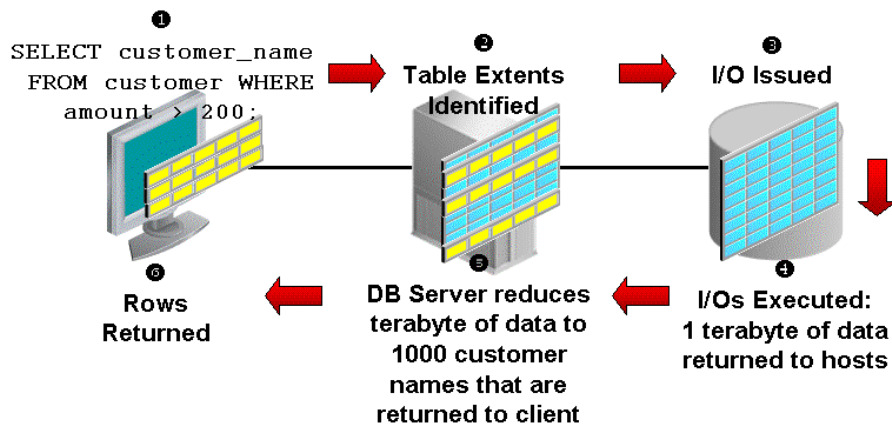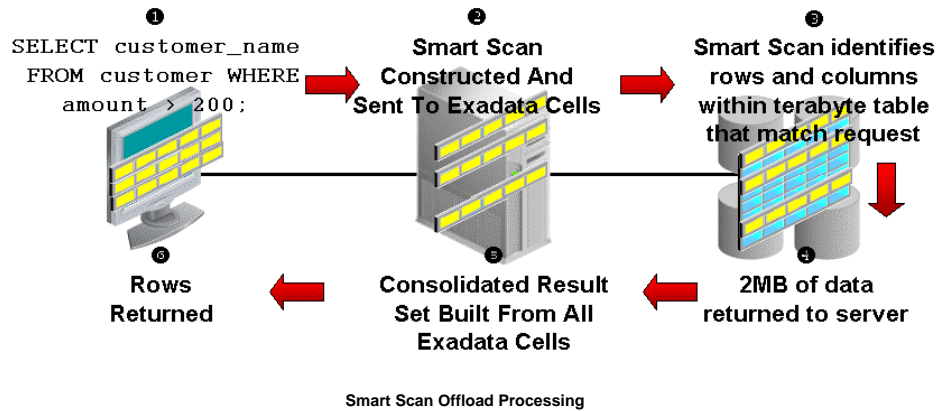
Traditional Database I/O and SQL Processing Model

❶ The client issues a SELECT statement with a predicate to filter and return only rows of interest. ❷ The database kernel maps this request to the file and extents containing the table being scanned. ❸ The database kernel issues the I/O to read the blocks. ❹ All the blocks of the table being queried are read into memory. ❺ Then SQL processing is done against the raw blocks searching for the rows that satisfy the predicate. ❻ Lastly the rows are returned to the client.

As is often the case with the large queries, the predicate filters out most of the rows read. Yet all the blocks from the table need to be read, transferred across the storage network and copied into memory. Many more rows are read into memory than required to complete the requested SQL operation. This generates a large number of data transfers which consume bandwidth and impact application throughput and response time.

Integrating database functionality within the storage layer of the database stack allows queries, and other database operations, to be executed much more efficiently. Implementing database functionality as close to the hardware as possible, in the case of Exadata at the disk level, can dramatically speed database operations and increase system throughput.

With Exadata storage, database operations are handled much more efficiently. Queries that perform table scans can be processed within Exadata storage with only the required subset of data returned to the database server. Row filtering, column filtering and some join processing (among other functions) are performed within the Exadata storage cells. When this takes place only the relevant and required data is returned to the database server.

In the figure below illustrates how a table scan operates with Exadata storage.

```
❶
SELECT customer_name
FROM customer WHERE
   amount > 200;
```

❷
**Smart Scan
Constructed And
Sent To Exadata Cells**

❸
**Smart Scan identifies
rows and columns
within terabyte table
that match request**

❻
**Rows
Returned**

❺
**Consolidated Result
Set Built From All
Exadata Cells**

❹
**2MB of data
returned to server**

Smart Scan Offload Processing

❶ The client issues a SELECT statement with a predicate to filter and return only rows of interest. ❷ The database kernel determines that Exadata storage is available and constructs an iDB command representing the SQL command issued and sends it the Exadata storage. ❸ The CELLSRV component of the Exadata software scans the data blocks to identify those rows and columns that satisfy the SQL issued. ❹ Only the rows satisfying the predicate and the requested columns are read into memory. ❺ The database kernel consolidates the result sets from across the Exadata cells. ❻ Lastly, the rows are returned to the client.

Smart scans are transparent to the application and no application or SQL changes are required. The SQL EXPLAIN PLAN shows when Exadata smart scan is used. Returned data is fully consistent and transactional and rigorously adheres to the Oracle Database consistent read functionality and behavior. If a cell dies during a smart scan, the uncompleted portions of the smart scan are transparently routed to another cell for completion. Smart scans properly handle the complex internal mechanisms of the Oracle Database including: uncommitted data and locked rows, chained rows, compressed tables, national language processing, date arithmetic, regular expression searches, materialized views and partitioned tables.

The Oracle Database and Exadata server cooperatively execute various SQL statements. Moving SQL processing off the database server frees server CPU cycles and eliminates a massive amount of bandwidth consumption which is then available to better service other requests. SQL operations run faster, and more of them can run concurrently because of less contention for the I/O bandwidth. We will now look at the various SQL operations that benefit from the use of Exadata.

Smart Scan Predicate Filtering

Exadata enables predicate filtering for table scans. Only the rows requested are returned to the database server rather than all rows in a table. For example, when the following SQL is issued only rows where the employees' hire date is after the specified date are sent from Exadata to the database instance.

```
SELECT * FROM employee_table WHERE hire_date > '1-Jan-2003';
```

This ability to return only relevant rows to the server will greatly improve database performance. This performance enhancement also applies as queries become more complicated, so the same benefits also apply to complex queries, including those with subqueries.

Smart Scan Column Filtering

Exadata provides column filtering, also called column projection, for table scans. Only the columns requested are returned to the database server rather than all columns in a table. For example, when the following SQL is issued, only the employee_name and employee_number columns are returned from Exadata to the database kernel.

```
SELECT employee_name, employee_number FROM employee_table;
```

For tables with many columns, or columns containing LOBs (Large Objects), the I/O bandwidth saved can be very large. When used together, predicate and column filtering dramatically improves performance and reduces I/O bandwidth consumption. In addition, column filtering also applies to indexes, allowing for even faster query performance.

Smart Scan Join Processing

Exadata performs joins between large tables and small lookup tables, a very common scenario for data warehouses with star schemas. This is implemented using Bloom Filters, which are a very efficient probabilistic method to determine whether a row is a member of the desired result set.

Smart Scan Processing of Encrypted Tablespaces and Columns

Smart Scan offload processing of Encrypted Tablespaces (TSE) and Encrypted Columns (TDE) is supported in Exadata storage. This enables increased performance when accessing the most confidential data in the enterprise.

Storage Indexing

Storage Indexes are a very powerful capability provided in Exadata storage that helps avoid I/O operations. The Exadata Storage Server Software creates and maintains a Storage Index (i.e., metadata about the database objects) in the Exadata cell. The Storage Index keeps track of minimum and maximum values of columns for tables stored on that cell. When a query specifies a WHERE clause, but before any I/O is done, the Exadata software examines the Storage Index to determine if rows with the specified column value exist in the cell by comparing the column value to the minimum and maximum values maintained in the Storage Index. If the column value is outside the minimum and maximum range, scan I/O for that query is avoided. Many SQL Operations will run dramatically faster because large numbers of I/O operations are automatically replaced by a few lookups. To minimize operational overhead, Storage Indexes are created and maintained transparently and automatically by the Exadata Storage Server Software.

Offload of Data Mining Model Scoring

Data Mining model scoring is offloaded to Exadata. This makes the deployment of data warehouses on Database Machine an even better and more performant data analysis platform. All data mining scoring functions (e.g., prediction_probability) are offloaded to Exadata for processing. This will not only speed warehouse analysis but reduce database server CPU consumption and the I/O load between the database server and Exadata storage.

Other Exadata Smart Scan Processing

Two other database operations that are offloaded to Exadata are incremental database backups and tablespace creation. The speed and efficiency of incremental database backups has been significantly enhanced with Exadata. The granularity of change tracking in the database is much finer when Exadata storage is used. Changes are tracked at the individual Oracle block level with Exadata rather than at the level of a large group of blocks. This results in less I/O bandwidth being consumed for backups and faster running backups.

With Exadata the create file operation is also executed much more efficiently. For example, when issuing a Create Tablespace command, instead of operating synchronously with each block of the new tablespace being formatted in server memory and written to storage, an iDB command is sent to Exadata instructing it to create the tablespace and format the blocks. Host memory usage is reduced and I/O associated with the creation and formatting of the tablespace blocks is offloaded. The I/O bandwidth saved with these operations means more bandwidth is available for other business critical work.

Exadata Hybrid Columnar Compression

Compressing data can provide dramatic reduction in the storage consumed for large databases. Exadata provides a very advanced compression capability called Exadata Hybrid Columnar Compression (EHCC). Exadata Hybrid Columnar Compression enables the highest levels of data compression and provides enterprises with tremendous cost-savings and performance improvements due to reduced I/O. Average storage savings can range from 10x to 15x depending on how EHCC is used. With average savings of 10x IT managers can drastically reduce and often eliminate their need to purchase new storage for several years. For example, a 100 terabyte database achieving 10x storage savings would utilize only 10 terabytes of physical storage. With 90 terabytes of storage now available, IT organizations can delay storage purchases for a significant amount of time.

EHCC is a new method for organizing data within a database block. As the name implies, this technology utilizes a combination of both row and columnar methods for storing data. This hybrid, or best of both worlds, approach achieves the compression benefits of columnar storage, while avoiding the performance shortfalls of a pure columnar format. A logical construct called the compression unit is used to store a set of Exadata Hybrid Columnar-compressed rows. When data is loaded, column values are detached from the set of rows, ordered and grouped together

and then compressed. After the column data for a set of rows has been compressed, it is fit into the compression unit.

Smart Scan processing of EHCC data is provided and column projection and filtering are performed within Exadata. Queries run directly on Exadata Hybrid Columnar Compressed data and do not require the data to be decompressed. Data that is required to satisfy a query predicate does not need to be decompressed, only the columns and rows being returned to the client are decompressed in memory. The decompression process takes place on the Exadata cell in order to maximize performance and offload processing from the database server. Given the typical ten-fold compression of Hybrid Columnar Compressed Tables, this effectively increases the I/O rate ten-fold compared to uncompressed data.

I/O Resource Management With Exadata

With traditional storage, creating a shared storage grid is hampered by the inability to prioritize the work of the various jobs and users consuming I/O bandwidth from the storage subsystem. The same occurs when multiple databases share the storage subsystem. The DBRM and I/O resource management capabilities of Exadata storage can prevent one class of work, or one database, from monopolizing disk resources and bandwidth and ensures user defined SLAs are met when using Exadata storage. The DBRM enables the coordination and prioritization of I/O bandwidth consumed between databases, and between different users and classes of work. By tightly integrating the database with the storage environment, Exadata is aware of what types of work and how much I/O bandwidth is consumed. Users can therefore have the Exadata system identify various types of workloads, assign priority to these workloads, and ensure the most critical workloads get priority.

In data warehousing, or mixed workload environments, you may want to ensure different users and tasks within a database are allocated the correct relative amount of I/O resources. For example you may want to allocate 70% of I/O resources to interactive users on the system and 30% of I/O resources to batch reporting jobs. This is simple to enforce using the DBRM and I/O resource management capabilities of Exadata storage.

An Exadata administrator can create a resource plan that specifies how I/O requests should be prioritized. This is accomplished by putting the different types of work into service groupings called Consumer Groups. Consumer groups can be defined by a number of attributes including the username, client program name, function, or length of time the query has been running. Once these consumer groups are defined, the user can set a hierarchy of which consumer group gets precedence in I/O resources and how much of the I/O resource is given to each consumer group. This hierarchy determining I/O resource prioritization can be applied simultaneously to both intra-database operations (i.e. operations occurring within a database) and inter-database operations (i.e. operations occurring among various databases).

When Exadata storage is shared between multiple databases you can also prioritize the I/O resources allocated to each database, preventing one database from monopolizing disk resources

and bandwidth to ensure user defined SLAs are met. For example you may have two databases sharing Exadata storage. Business objectives dictate that each of these databases has a relative value and importance to the organization. It is decided that database A should receive 33% of the total I/O resources available and that database B should receive 67% of the total I/O of resources. To ensure the different users and tasks within each database are allocated the correct relative amount of I/O resources, various consumer groups are defined.

- Two consumer groups are defined for database A
  - 60% of the I/O resources are reserved for interactive marketing activities
  - 40% of the I/O resources are reserved for batch marketing activities
- Three consumer groups are defined for database B
  - 60% of the I/O resources are reserved for interactive sales activities
  - 30% of the I/O resources are reserved for batch sales activities
  - 10% of the I/O resources are reserved for major account sales activities

These consumer group allocations are relative to the total I/O resources allocated to each database.

In essence, Exadata I/O Resource Manager has solved one of the challenges traditional storage technology does not address: creating a shared grid storage environment with the ability to balance and prioritize the work of multiple databases and users sharing the storage subsystem. Exadata I/O resource management ensures user defined SLAs are met for multiple databases sharing Exadata storage. This ensures that each database or user gets the correct share of disk bandwidth to meet business objectives.

Quality of Service (QoS) Management with Exadata

Oracle Exadata QoS Management is an automated, policy-based product that monitors the workload requests for an entire system. It manages the resources that are shared across applications and adjusts the system configuration to keep the applications running at the performance levels needed by your business. It responds gracefully to changes in system configuration and demand, thus avoiding additional oscillations in the performance levels of your applications.

Oracle Exadata QoS Management monitors the performance of each work request on a target system. It starts to track a work request from the time a work request requests a connection to the database using a database service. The amount of time required to complete a work request, or the response time (also known as the end-to-end response time, or round-trip time), is the time from when the request for data was initiated and when the data request is completed. By accurately measuring the two components of response time (the time spent using resources and the time spent waiting to use resources), QoS Management can quickly detect bottlenecks in the

system. It then makes recommendations to reallocate resources to relieve a bottleneck, thus preserving or restoring service levels. System administrators are alerted to the need for this reallocation and it is implemented with a simple button click on the QoS Management dashboard. Full details as to the entire cluster's projected performance impact to this action are also provided. Finally an audit log of all actions and policy changes is maintained along with historical system performance graphs.

Oracle Exadata QoS Management manages the resources on your system so that:

- When sufficient resources are available to meet the demand, business-level performance requirements for your applications are met, even if the workload changes;

- When sufficient resources are not available to meet the demand, Oracle Exadata QoS Management attempts to satisfy the more critical business performance requirements at the expense of less critical performance requirements;

- When load conditions severely exceed capacity, resources remain available.

Benefits of Using Oracle Exadata QoS Management

In a typical company, when the response times of your applications are not within acceptable levels, problem resolution can be very slow. Often, the first questions that administrators ask are: "Did we configure the system correctly? Is there a parameter change that fixes the problem? Do we need more hardware?" Unfortunately, these questions are very difficult to answer precisely; the result is often hours of unproductive and frustrating experimentation.

Oracle Exadata QoS Management provides the following benefits:

- Reduces the time and expertise requirements for system administrators who manage Oracle Real Application Clusters (Oracle RAC) resources

- Helps reduce the number of performance outages

- Reduces the time needed to resolve problems that limit or decrease the performance of your applications

- Provides stability to the system as the workloads change

- Makes the addition or removal of servers transparent to applications

- Reduces the impact on the system caused by server failures

- Helps ensure that service-level agreements (SLAs) are met

- Enables more effective sharing of hardware resources

- Protects existing workloads from over committed memory-induced server failures

- Exadata Storage Virtualization

- Exadata provides a rich set of sophisticated and powerful storage management virtualization capabilities that leverage the strengths of the Oracle Database, the Exadata software, and Exadata hardware.

Exadata Storage Software

As discussed earlier, the Exadata cell is a server that runs the Oracle Enterprise Linux as well as the Oracle provided Exadata software. When first started, the cell boots up like any other computer into Exadata storage serving mode. The first two disk drives have a small Logical Unit Number (LUN) slice called the System Area, approximately 13 GB of size, reserved for the OEL operating system, Exadata software, and configuration metadata. The System Area contains Oracle Database 11g Automatic Diagnostic Repository (ADR) data, and other metadata about the Exadata cell. The administrator does not have to manage the System Area LUN, as it is automatically created. Its contents are automatically mirrored across the physical disks to protect against drive failures, and to allow hot disk swapping. The remaining portion of these two disk drives is available for user data.

Exadata User Storage Virtualization

Automatic Storage Management (ASM) is used to manage the storage in the Exadata cell. ASM volume management, striping, and data protection services make it the optimum choice for volume management. ASM provides data protection against drive and cell failures, the best possible performance, and extremely flexible configuration and reconfiguration options.

A Cell Disk is the virtual representation of the physical disk, minus the System Area LUN (if present), and is one of the key disk objects the administrator manages within an Exadata cell. A Cell Disk is represented by a single LUN, which is created and managed automatically by the Exadata software when the physical disk is discovered.

Cell Disks can be further virtualized into one or more Grid Disks. Grid Disks are the disk entity assigned to ASM, as ASM disks, to manage on behalf of the database for user data. The simplest case is when a single Grid Disk takes up the entire Cell Disk. But it is also possible to partition a Cell Disk into multiple Grid Disk slices. Placing multiple Grid Disks on a Cell Disk allows the administrator to segregate the storage into pools with different performance or availability requirements. Grid Disk slices can be used to allocate "hot", "warm" and "cold" regions of a Cell Disk, or to separate databases sharing Exadata disks. For example a Cell Disk could be partitioned such that one Grid Disk resides on the higher performing portion of the physical disk and is configured to be triple mirrored, while a second Grid Disk resides on the lower performing portion of the disk and is used for archive or backup data, without any mirroring. An Information Lifecycle Management (ILM) strategy could be implemented using Grid Disk functionality.
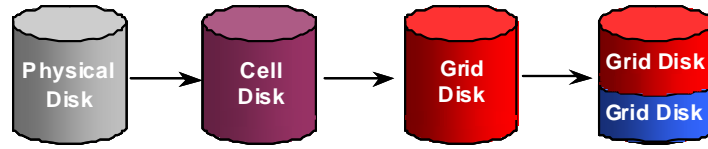
Grid Disk Virtualization

The following example illustrates the relationship of Cell Disks to Grid Disks in a more comprehensive Exadata storage grid.

Once the Cell Disks and Grid Disks are configured, ASM disk groups are defined across the Exadata configuration. Two ASM disk groups are defined; one across the "hot" grid disks, and a second across the "cold" grid disks. All of the "hot" grid disks are placed into one ASM disk group and all of the "cold" grid disks are placed in a separate disk group. When the data is loaded into the database, ASM will evenly distribute the data and I/O within disk groups. ASM mirroring can be activated for these disk groups to protect against disk failures for both, either, or neither of the disk groups. Mirroring can be turned on or off independently for each of the disk groups.
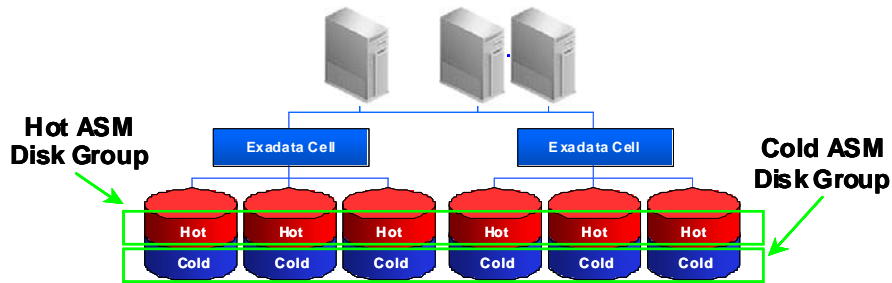
Example ASM Disk Groups and Mirroring

Lastly, to protect against the failure of an entire Exadata cell, ASM failure groups are defined. Failure groups ensure that mirrored ASM extents are placed on different Exadata cells.
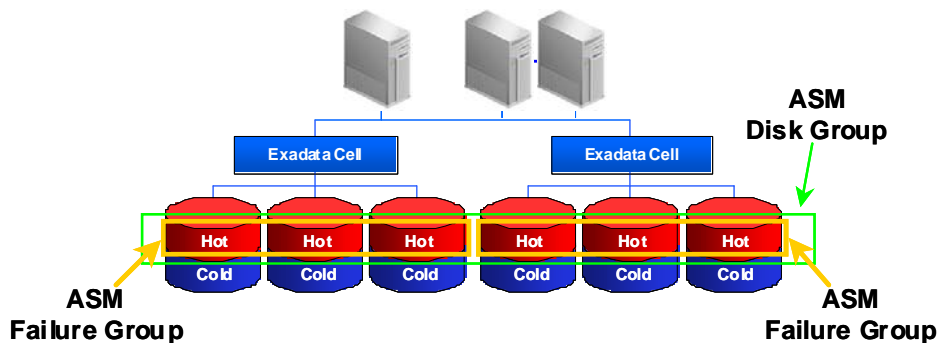
Example ASM Mirroring and Failure Groups

With Exadata and ASM:

- Configuration of Cell Disks (LUN creation) is automated by Exadata software.

- Optionally, multiple Grid Disks can co-exist on the physical disks to tailor performance to the needs of the database application or construct an ILM strategy with Exadata.

- ASM automatically stripes the database data across Exadata disks and cells to ensure a balanced I/O load and optimum performance.

- ASM dynamic add and drop capability enables non-intrusive cell and disk allocation, deallocation, and reallocation.

- ASM mirroring, and the hot swap capability of the Exadata cell, provides transparent data protection and access across disk failures.

- ASM provides for double or triple mirroring to tailor the protection to the criticality of the data.

- ASM failure groups are automatically created with Exadata to provide transparent data protection and access across cell failures.

Migrating to Exadata Storage

There are several techniques for migrating data to a Database Machine. Migration can be done using Oracle Recovery Manager (RMAN) to backup from traditional storage and restore the data onto Exadata. Oracle Data Guard can also be used to facilitate a migration. This is done by first creating a standby database based on Exadata storage. The standby can be using Exadata storage and the production database can be on traditional storage. By executing a fast switchover, taking just seconds, you can transform the standby database into the production database. This provides a built-in safety net as you can undo the migration very gracefully if unforeseen issues arise. Transportable Tablespaces and Data Pump may also be used to migrate to Exadata. Any technique used to move data between Oracle Databases can be used with Exadata.

Additional Data Protection With Exadata

Exadata has been designed to incorporate the same standard of high availability (HA) customers have come to expect from Oracle products. With Exadata, all database features and tools work just as they do with traditional non-Exadata storage. Users and database administrators will use familiar tools and be able to leverage their existing Oracle Database knowledge and procedures. With the Exadata architecture, all single points of failure are eliminated. Familiar features such as mirroring, fault isolation, and protection against drive and cell failure have been incorporated into Exadata to ensure continual availability and protection of data. Other features to ensure high availability within the Exadata server are described below.

Hardware Assisted Resilient Data (HARD) built into Exadata

Oracle's Hardware Assisted Resilient Data (HARD) Initiative is a comprehensive program designed to prevent data corruptions before they happen. Data corruptions are very rare, but when they happen, they can have a catastrophic effect on a database, and therefore a business. Exadata has enhanced HARD functionality embedded in it to provide even higher levels of protection and end-to-end data validation for your data. Exadata performs extensive validation of the data stored in it including checksums, block locations, magic numbers, head and tail checks, alignment errors, etc. Implementing these data validation algorithms within Exadata will prevent corrupted data from being written to permanent storage. Furthermore, these checks and protections are provided without the manual steps required when using HARD with conventional storage.

Data Guard

Oracle Data Guard is the software feature of Oracle Database that creates, maintains, and monitors one or more standby databases to protect your database from failures, disasters, errors, and corruptions. Data Guard works unmodified with Exadata and can be used for both production and standby databases. By using Active Data Guard with Exadata storage, queries and reports can be offloaded from the production database to an extremely fast standby database and ensure that critical work on the production database is not impacted while still providing disaster protection.

Flashback

Exadata leverages Oracle Flashback Technology to provide a set of features to view and restore data back in time. The Flashback feature works in Exadata the same as it would in a non-Exadata environment. The Flashback features offer the capability to query historical data, perform change analysis, and perform self-service repair to recover from logical corruptions while the database is online. In essence, with the built-in Oracle Flashback features, Exadata allows the user to have snapshot-like capabilities and restore a database to a time before an error occurred.

Recovery Manager (RMAN) and Oracle Secure Backup (OSB)

Exadata works with Oracle Recovery Manager (RMAN), a command-line and Enterprise Manager-based tool, to allow efficient Oracle database backup and recovery. All existing RMAN scripts work unchanged in the Exadata environment. RMAN is designed to work intimately with the server, providing block-level corruption detection during backup and restore. RMAN optimizes performance and space consumption during backup with file multiplexing and backup set compression, and integrates with Oracle Secure Backup (OSB) and third party media management products for tape backup.

Conclusion

Businesses today increasingly need to leverage a unified database platform to enable the deployment and consolidation of all applications onto one common infrastructure. Whether OLTP, DW or mixed workload a common infrastructure delivers the efficiencies and reusability the datacenter needs – and provides the reality of grid computing in-house. Building or using custom special purpose systems for different applications is wasteful and expensive. The need to process more data increases every day while corporations are also finding their IT budgets being squeezed. Examining the total cost of ownership (TCO) for IT software and hardware leads to choosing a common high performance infrastructure for deployments of all applications.

By incorporating the Exadata based Database Machine into the IT infrastructure, companies will:

- Accelerate database performance and be able to do much more in the same amount of time.

- Handle change and growth in scalable and incremental steps by consolidating deployments on to a common infrastructure.

- Deliver mission-critical data availability and protection.

ORACLE®

A Technical Overview of the Oracle Exadata
Database Machine and Exadata Storage Server
October 5, 2010
Author: Ronald Weiss

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

SOFTWARE. HARDWARE. COMPLETE.