

An Oracle White Paper
September 2009

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

Sun Oracle Exadata Storage Server and Database Machine	2
Today's Limits On Database I/O	2
Exadata Product Family	4
Sun Oracle Exadata Storage Server	4
Sun Oracle Database Machine	8
Exadata Architecture	12
Database Server Software	13
Enterprise Manager Plug-In For Exadata.....	14
Exadata Software	14
Exadata Smart Scan Processing	15
I/O Resource Management With Exadata.....	19
Accelerated Performance With Exadata	21
Exadata Storage Virtualization.....	22
CONCLUSION	27

Sun Oracle Exadata Storage Server and Database Machine

The Sun Oracle Exadata Storage Server (Exadata) is a storage product optimized for use with Oracle Database applications and is the storage building block of the Sun Oracle Database Machine. Exadata 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 and Exadata Smart Flash Cache to dramatically accelerate Oracle Database processing and speed I/O operations. The Exadata storage products are a combination of software and hardware used to store and access Oracle databases. Exadata 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. The Exadata Storage Servers are also packaged in a complete end-to-end database solution – the Sun Oracle Database Machine. The Sun Oracle Database Machine is an easy to deploy out of the box solution for hosting the Oracle Database for all applications and delivers the highest levels of performance available. 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 to the enterprise.

Exadata is a joint offering from Oracle and Sun Microsystems. Sun is providing the hardware technology used in the Database Machine and Exadata Storage Server. Oracle is providing the software to impart database intelligence to the storage and Database Machine and is tightly integrated with the Oracle Database and all its features. The Sun servers combine the power of the latest generation of Intel® Xeon® processors with Sun's system engineering expertise. These servers offer the needed density and expandability to satisfy the most demanding datacenter applications. The Oracle and Sun partnership makes possible the delivery of the Sun Oracle Database Machine and Exadata Storage Server and the revolutionary capabilities it provides.

Today's Limits On Database I/O

The foundation of the Database Machine and Exadata is smart database software to handle the complex applications deployed to drive the most demanding enterprise business needs. The

Oracle Database provides an incredible amount of functionality to implement the most sophisticated OLTP and DW applications and to consolidate mixed workload environments. But to access terabytes databases with high performance, augmenting the smart database software with powerful hardware provides tremendous opportunities to deliver more database processing, faster, for the enterprise. Having powerful hardware to provide the required I/O rates and bandwidth for today's applications, in addition to smart software, is key to the extreme performance delivered by the Exadata family of products.

Traditional storage devices offer high storage capacity but are relatively slow and can not sustain the I/O rates for the transaction load the enterprise requires for its applications. Instead of hundreds of IOPS (I/Os per second) per disk enterprise applications require their systems deliver at least an order of magnitude higher IOPS to deliver the service enterprise end-users expect. This problem gets magnified when hundreds of disks reside behind a single storage controller. The IOPS that can be executed are severely limited by both the speed of the mechanical disk drive and the number of drives per storage controller.

Traditional storage products provide the Oracle Database a narrow and limited interface to database storage. Many bottlenecks exist in the database I/O path restricting data bandwidth, and hence limiting overall database performance when using traditional storage products. Database servers need many Storage Area Network (SAN) Host Bus Adapters (HBA) to provide the bandwidth necessary to deliver data, from storage to the database, at an adequate rate. Very often the number of HBAs required to deliver adequate performance cannot be supported in the server, or are too costly to provide. And SAN switch cost and complexity increases dramatically to provide the required bandwidth and redundancy. In addition large storage arrays cannot deliver adequate bandwidth to the hundreds of disks they house. This results in the potential performance of disks being artificially capped well below what they are capable of providing. Disk performance is bottlenecked on the Fibre Channel Loops (FCL) to disk and the processing capacity of the storage array.

Traditional storage devices are also unaware that a database is residing on the storage and therefore cannot provide any database-aware I/O or SQL processing. When the database requests rows and columns what is returned from the storage are data blocks rather than the result set of a database query. Traditional storage has no database intelligence to discern the particular rows and columns actually requested. So, when processing I/O on behalf of the database, traditional storage consumes bandwidth returning data that is not relevant to the database query that was issued.

Exadata products address the key dimensions of database I/O that can hamper database performance.

- Exadata is based on a massively parallel architecture which provides more pipes to deliver more data faster between the database servers and the storage servers.

- Exadata is built using wider pipes that provide extremely high bandwidth between the database servers and the storage servers.
- Exadata is database aware and can ship just the data required to satisfy SQL requests resulting in less data being sent between the database servers and the storage servers.
- Exadata overcomes the mechanical limits of disk drive technology by automatically caching frequently accessed data delivering unprecedented levels of bandwidth and IOPS.

Exadata Product Family

There are two members of the Exadata product family. The foundation of the Exadata family of products is the Sun Oracle Exadata Storage Server. It is used as the storage for the Oracle Database when building custom database systems. The second member of the Exadata product family is the Sun Oracle 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 application requiring the best performance, and includes Exadata storage.

Sun Oracle Exadata Storage Server

The Sun Oracle Exadata Storage Server is a database storage device running 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.

The Sun Oracle Exadata Storage Server is a fast, reliable, high capacity, industry- standard storage server. Each Exadata cell comes preconfigured with: two Intel Xeon E5540 quad-core processors, 384 GB of Exadata Smart Flash Cache, twelve disks connected to a storage controller with 512MB battery-backed cache, 24 GB memory, dual port InfiniBand connectivity, management interface for remote access, dual-redundant hot-swappable power supplies, all the software preinstalled, and takes up 2U in a typical 19-inch rack.



Figure 1: Exadata Storage Cell

Exadata Smart Flash Cache

Each Exadata cell comes with 384 GB of Exadata Smart Flash Cache. 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 Server Performance, Bandwidth and IOPS

The Sun Oracle Exadata Storage Server comes with either twelve 600 GB Serial Attached SCSI (SAS) disks or twelve 2 TB Serial Advanced Technology Attachment (SATA) disks. SAS based Exadata Storage Servers provide up to 2 TB of uncompressed user data capacity, and up to 1.5 GB/second of raw data bandwidth. SATA based Exadata Storage Servers provide up to 7 TB of uncompressed user data capacity, and up to 0.85 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 increases up to 10 times. User data capacity is computed after mirroring all the disk space, and setting aside space for database structures like logs, undo, and temp space. 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 of the Flash cache 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. This represents a significant improvement over traditional storage devices used with the Oracle Database.

The performance specifications of the Exadata Storage Server are shown below.

	SAS Based Exadata Storage Server	SATA Based Exadata Storage Server
Exadata Smart Flash Cache	384 GB	384 GB
Raw Disk Capacity	7.2 TB	24 TB
User Data Capacity (without data compression)	Up to 2 TB	Up to 7 TB
Raw Disk Data Bandwidth	Up to 1.5 GB/sec	Up to 0.85 GB/sec
Effective Data Bandwidth with Flash	Up to 3.6 GB/sec	Up to 3.6 GB/sec
Effective Data Bandwidth with Flash Cache and Data Compression	Up to 36 GB/sec	Up to 36 GB/sec
Flash Cache IOPS	Up to 75,000	Up to 75,000
Disk IOPS	Up to 3,600	Up to 1,440

InfiniBand and the Exadata Storage Server

Oracle Exadata storage uses a state of the art InfiniBand interconnect between the servers and storage. An 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.

Exadata Storage Server Configuration

In figure 2 below, a small Exadata storage based database environment is shown. Two Oracle Databases, one RAC and one single instance, are sharing three Exadata cells. All the components for this configuration – database servers, Exadata cells, InfiniBand switches, Ethernet switches, and other support hardware – can be housed in, and take up less than half of, a typical 19-inch rack.

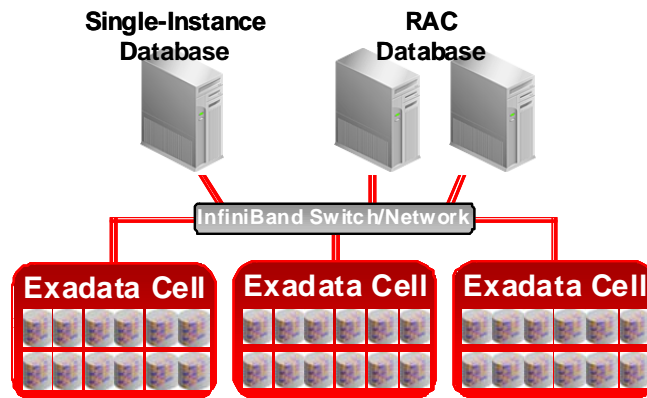


Figure 2: Exadata Storage Cell Based Configuration

Oracle Exadata is architected to scale-out to any level of performance. To achieve higher performance and greater storage capacity, additional Exadata cells are added to the configuration. As more cells are added, capacity and performance increases linearly. No cell-to-cell communication is ever done or required in an Exadata configuration.

Oracle Automatic Storage Management (ASM) is used as the file system and volume manager for Exadata. 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.

When using Exadata, SQL processing is offloaded from the database server to the Exadata server. 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 storage is totally transparent to the application using the database. 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 your application. And 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 Streams, 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. Both Exadata and non-Exadata storage may be concurrently used for database storage to facilitate migration to, or from, Exadata storage.

The nature of traditional storage products encourages inefficient deployments of storage for each database in the IT infrastructure. The Exadata architecture ensures all the bandwidth and I/O resources of the Exadata storage subsystem can be made available whenever, and to whichever, database or class of work needs it. I/O bandwidth is metered out to the various classes of work, or databases, sharing the Exadata server based on user defined policies and service level agreements (SLAs). The Oracle Database Resource Manager (DBRM) has been enhanced for use with Exadata storage to manage user-defined intra and inter-database I/O resource usage to ensure customer defined SLAs are met. The I/O resource management capabilities of Exadata storage enable tailoring the I/O resources to the business priorities of the organization, and to build a shared storage grid for the Oracle databases in the environment.

Sun Oracle Database Machine

Oracle is also offering a fully integrated platform for all your database applications. The Sun Oracle Database Machine is an easy to deploy out of the box solution for hosting the Oracle Database. A fully integrated solution ready to be turned on day one takes a lot of integration work, cost and time, out of the application deployment process. The benefit of a common infrastructure to deploy any application on, whether OLTP, DW or a mix of the two, creates tremendous efficiencies in the datacenter.



Figure 3: Database Machine Full Rack

There are four models of the Database Machine – Database Machine *Full Rack*, Database Machine *Half Rack*, Database Machine *Quarter Rack*, and Database Machine *Eighth Rack* – are offered. Depending on the size and purpose of the database to be deployed, and the processing and I/O bandwidth required there is a system available to meet any need.

Each Database Machine runs the same software, is upgradeable and includes common hardware components. Common to all Database Machines are:

- Exadata Storage Servers, either SAS or SATA.
- Industry standard Oracle Database 11g database servers with: two Intel Xeon dual-socket quad-core E5540 processors running at 2.53 Ghz processors, 72 GB RAM, four 146 GB SAS drives, dual port InfiniBand Host Channel Adapter (HCA), four 1 Gb/second Ethernet ports, and dual-redundant, hot-swappable power supplies.
- 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.

The ratio of components to each other has been chosen to maximize performance and ensure system resiliency. The hardware composition of each model of Database Machine is depicted in the following table.

	Sun Oracle Database Machine Full Rack	Sun Oracle Database Machine Half Rack	Sun Oracle Database Machine Quarter Rack	Sun Oracle Database Machine Basic System
Database Servers	8	4	2	1
Exadata Storage Servers	14	7	3	1
InfiniBand Switches	3	2	2	1
Upgradability	Connect multiple Full Racks via included InfiniBand fabric	Field upgrade from Half Rack to Full Rack	Field upgrade from Quarter Rack to Half Rack	Custom field upgrade

The performance and capacity characteristics of each model of Database Machine is depicted in the following table.

	Sun Oracle Database Machine Full Rack	Sun Oracle Database Machine Half Rack	Sun Oracle Database Machine Quarter Rack	Sun Oracle Database Machine Basic System
Exadata Smart Flash Cache	5.3 TB	2.6 TB	1.1 TB	384 GB
Raw Disk Capacity				
• SAS	100 TB	50 TB	21 TB	7.2 TB
• SATA	336 TB	168 TB	72 TB	24 TB
User Data Capacity				
• SAS	Up to 28 TB	Up to 14 TB	Up to 6 TB	Up to 2 TB
• SATA (without data compression)	100 TB	50 TB	21 TB	7 TB
Raw Disk Data Bandwidth				
• SAS	Up to 21 GB/sec	Up to 10.5 GB/sec	Up to 4.5 GB/sec	Up to 1.5 GB/sec
• SATA	12 GB/sec	6.0 GB/sec	2.5 GB/sec	0.85 GB/sec
Effective Data Bandwidth with Flash Cache	Up to 50 GB/sec	Up to 25 GB/sec	Up to 11 GB/sec	Up to 3.6 GB/sec
Effective Data Bandwidth with Flash Cache and Data Compression	Up to 500 GB/sec	Up to 250 GB/sec	Up to 110 GB/sec	Up to 36 GB/sec
Flash Cache IOPS	Up to 1,000,000	Up to 500,000	Up to 225,000	Up to 75,000
Disk IOPS				
• SAS	Up to 50,000	Up to 25,000	Up to 10,800	Up to 3,600
• SATA	20,000	10,000	4,300	1,440

In summary the Exadata products address the key dimensions of database I/O that can hamper performance.

- More pipes: Exadata is based on a massively parallel architecture which provides more pipes to deliver more data faster between the database servers and storage servers. As Exadata servers are added to the database configuration bandwidth scales linearly.
- Wider pipes: InfiniBand is 8 times faster than Fibre Channel. Exadata is built using wider InfiniBand pipes that provide extremely high bandwidth between the database servers and storage servers.
- More IOPS: With the intelligent and automatic use and management of Exadata Smart Flash Cache to avoid physical I/O effective IOPS scale to handle the largest most demanding applications.
- Smart software: With the Smart Scan processing less data needs to be shipped through the pipes by performing data processing in storage. Exadata is database aware and can ship just the data required to satisfy SQL requests resulting in less data being sent between the database servers and the storage servers.

Exadata Architecture

The hardware environment for a typical Exadata based storage grid was shown in Figure 2. Each Exadata cell is a self-contained server which houses disk storage and runs the Exadata software provided by Oracle. Databases are deployed across Exadata cells, and multiple databases can share Exadata cells. The database and Exadata cells communicate via a high-speed InfiniBand interface.

The collection of Exadata cells shared between a set of databases is referred to as an Exadata Realm. The set of three cells in figure 2 is an example of a realm. Realms ensure the isolation, and hence protection, across a given set of databases. Mechanisms are provided to move disks and whole cells between realms in a controlled and safe manner.

The architecture of the Exadata solution includes components on the database server and in the Exadata cell. The overall architecture is shown below.

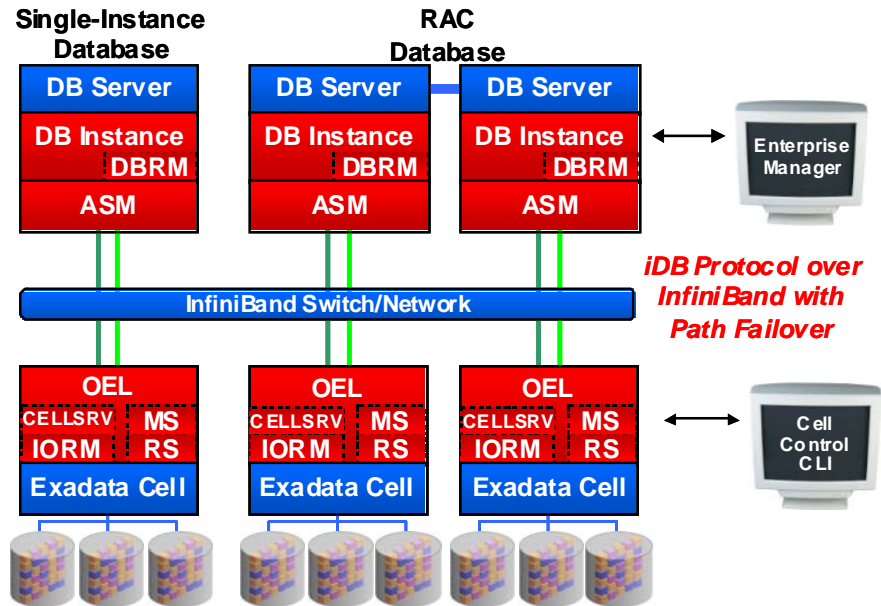


Figure 4: Exadata Software Architecture

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 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.

Automatic Storage Management (ASM) is the storage management foundation of 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. And the ASM mirroring and failure group functionality provides much of the data protection and resiliency across the Exadata environment. With ASM, data is mirrored across cells to ensure high availability in the event of cell failure.

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.

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 Software

Like any storage device the Exadata 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.

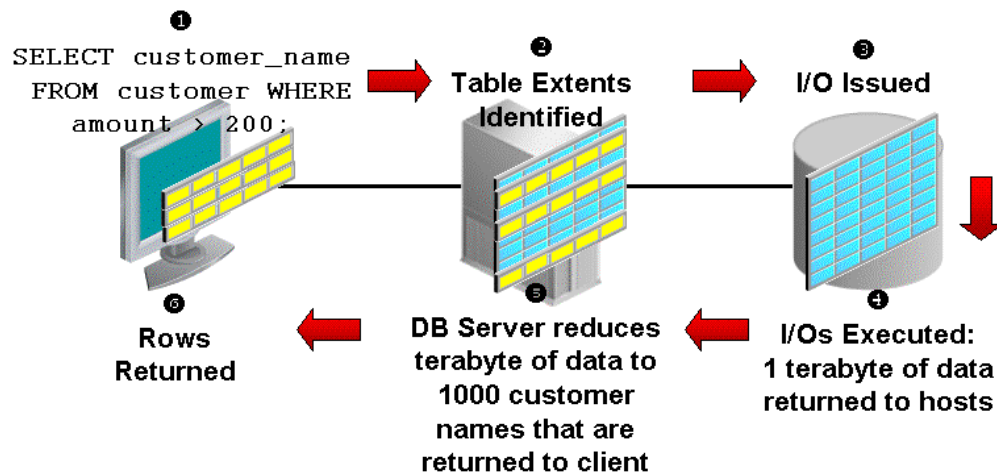


Figure 5: Traditional Database I/O and SQL Processing Model

1 The client issues a `SELECT` statement with a predicate to filter and return only rows of interest. **2** The database kernel maps this request to the file and extents containing the table being scanned. **3** The database kernel issues the I/O to read the blocks. **4** All the blocks of the table being queried are read into memory. **5** Then SQL processing is done against the raw blocks searching for the rows that satisfy the predicate. **6** 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 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.

Figure 6 below illustrates how a table scan operates with Exadata storage.

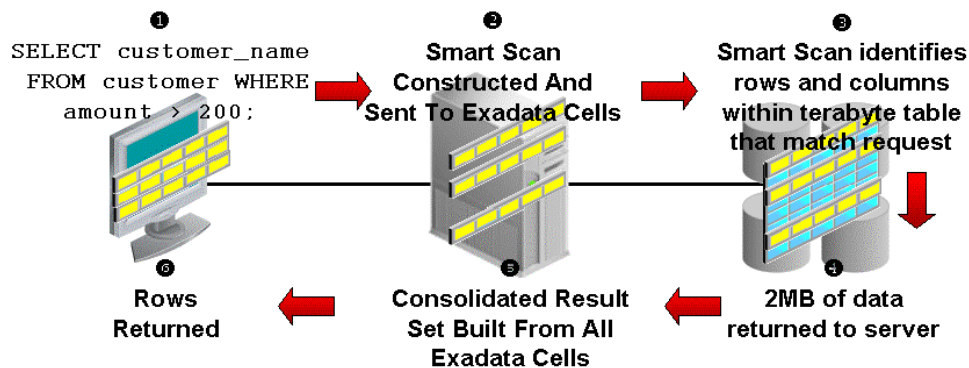


Figure 6: Smart Scan Offload Processing

1 The client issues a `SELECT` statement with a predicate to filter and return only rows of interest. **2** 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. **3** The CELLSRV component of the Exadata software scans the data blocks to identify those rows and columns that satisfy the SQL issued. **4** Only the rows satisfying the predicate and the requested columns are read into memory. **5** The database kernel consolidates the result sets from across the Exadata cells. **6** 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.

New Smart Scan Offload I/O Optimizations and Functionality

With Oracle Database 11g Release 2, several new powerful Smart Scan and offload capabilities are provided with Exadata storage. These include: Storage Indexing technology, Smart Scan offload of new Hybrid Columnar Compressed Tables, Smart Scan offload of encrypted tablespaces and columns, and offload of data mining model scoring.

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 in Exadata memory. 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 exists 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 in-memory lookups. To minimize operational overhead, Storage Indexes are created and maintained transparently and automatically by the Exadata Storage Server Software.

Smart Scan of Hybrid Columnar Compressed Tables

Another new feature of Oracle Database 11g Release 2 is Hybrid Columnar Compressed Tables. These new tables offer a high degree of compression for data that is bulk loaded and queried. Smart Scan processing of Hybrid Columnar Compressed Tables is provided and column projection and filtering are performed within Exadata. In addition, the decompression of the data is offloaded to Exadata eliminating CPU overhead on the database servers. Given the typical ten-fold compression of Hybrid Columnar Compressed Tables, this effectively increases the I/O rate ten-fold compared to uncompressed data.

Smart Scan of Encrypted Tablespaces and Columns

New in Exadata is the Smart Scan offload processing of Encrypted Tablespaces (TSE) and Encrypted Columns (TDE). While the prior release of Exadata fully supported the use of TSE and TDE on Exadata it did not benefit from Exadata offload processing. This enhancement increases performance when accessing confidential data.

Offload of Data Mining Model Scoring

Another new function offloaded to Exadata is Data Mining model scoring. This makes the deployment of data warehouses on Exadata or 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.

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 as depicted below.

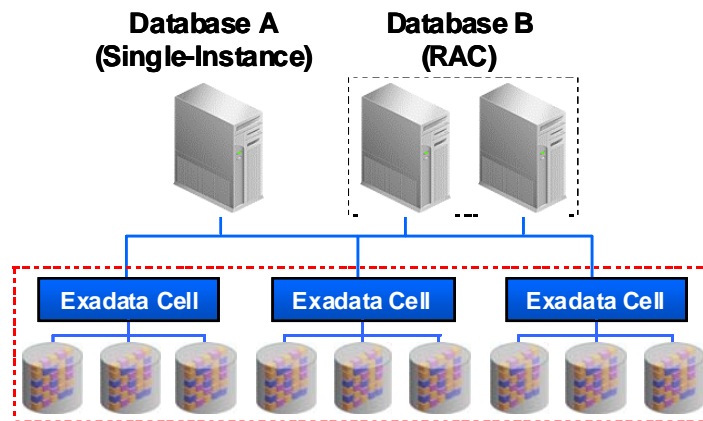


Figure 7: Inter-Database I/O Resource Management with Exadata

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.

Accelerated Performance With Exadata

Exadata storage provides unmatched performance improvements for typical data warehousing workloads. Full table scans will receive an arbitrarily large improvement due to smart scan filtering and the balanced hardware used for Exadata-based data warehouses. Exadata storage servers deliver a scale-out architecture such that as cells are added to the configuration, bandwidth increases. This, coupled with faster InfiniBand interconnect and the reduction of data transferred due to the offload processing, yields very large performance improvements. Often a ten-fold speed up in these operations is seen when using Exadata storage compared to storage products traditionally used with the Oracle Database — but in many cases a 50-fold, or greater, speedup is achieved.

Two examples of real world performance improvements follow.

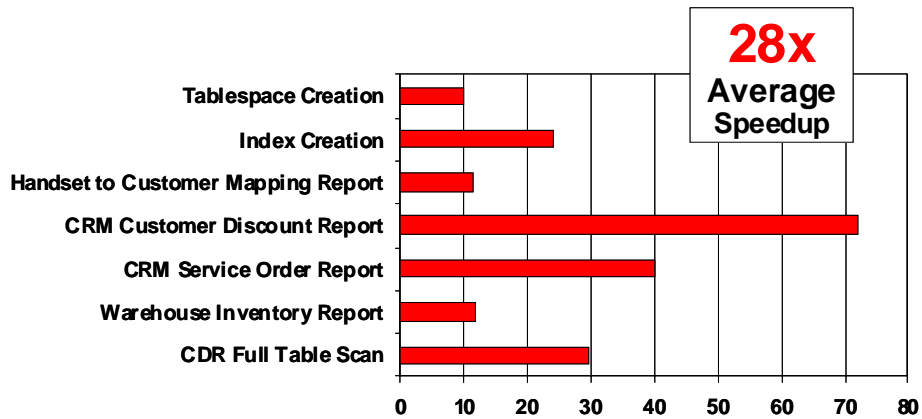


Figure 8: Telecommunications Application Performance Improvement with Exadata of 10X to 72X

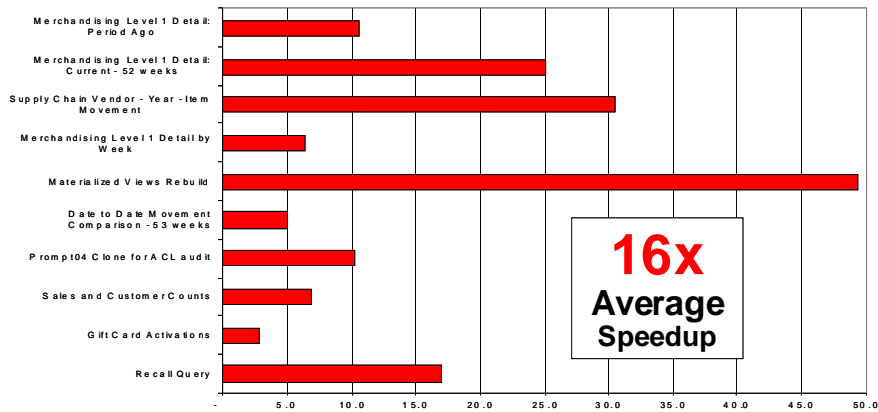


Figure 9: Retail Application Performance Improvement of 3X to 48X

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.

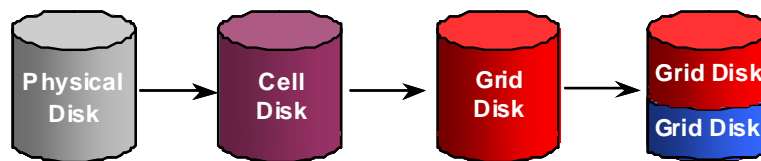


Figure 10: Grid Disk Virtualization

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.

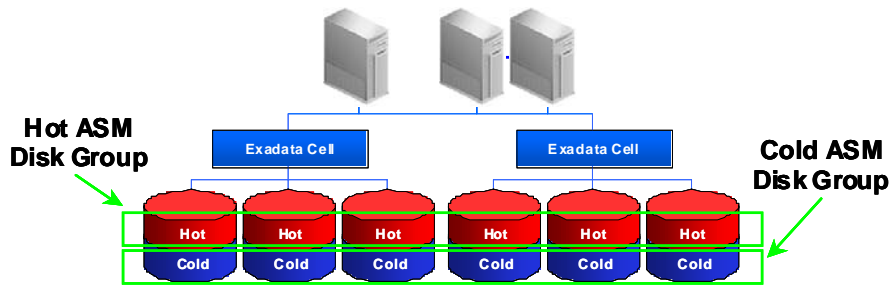


Figure 11: Example ASM Disk Groups and Mirroring

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.

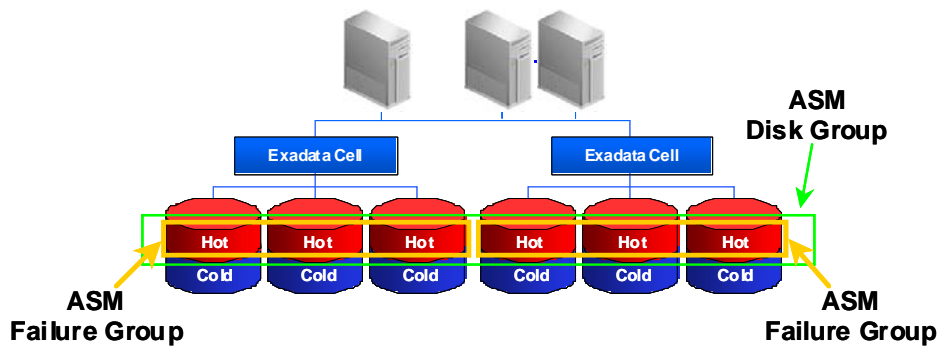


Figure 12: Example ASM Mirroring and Failure Groups

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

Exadata storage can be used in addition to the storage arrays and products traditionally used to store the Oracle database. A single database can be partially stored on Exadata storage and partially on traditional storage devices. Tablespaces can reside on Exadata storage, non-Exadata storage, or a combination of the two, and is transparent to database operations and applications. But to benefit from the Smart Scan capability of Exadata storage, the entire tablespace must reside on Exadata storage. This co-residence and co-existence is a key feature to enable online migration to Exadata storage.

An online non-disruptive migration to Exadata storage can be done for an existing database if the existing database is deployed on ASM and is using ASM redundancy. The steps to accomplish this are:

1. Add an Exadata grid disk to the existing ASM disk group.
2. ASM then automatically rebalances the data within the disk group moving a proportional amount of data to the newly added Exadata grid disk.
3. Then a non-Exadata disk is dropped from the ASM disk group. ASM would then rebalance or migrate the data from the non-Exadata disk to other disks in the disk group.
4. The above is repeated until the entire database has been migrated onto Exadata storage.

In addition, 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. All these approaches provide a built-in safety net as you can undo the migration very gracefully if unforeseen issues arise.

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 needs to leverage a unified database platform to enable the deployment and consolidation of all their 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 cloud 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 lead one to choose high performance common infrastructure for deployments of all applications.

By incorporating Exadata and the 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.

Exadata and the Database Machine provide this solution.



White Paper Title
September 2009
Author: Ronald Weiss
Contributing Authors:

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

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com



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

Copyright © 2009, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle is a registered trademark of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.