ORACLE

EXADATA

An Oracle White Paper
December 2013

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

ORACLE

Introduction

The Oracle Exadata Database Machine is engineered to be the highest performing and most available platform for running the Oracle Database. Exadata is a modern architecture featuring scale-out industry-standard database servers, scale-out intelligent storage servers, and an extremely high speed InfiniBand internal fabric that connects all servers and storage. Unique software algorithms in Exadata implement database intelligence in storage, PCI based flash, and InfiniBand networking to deliver higher performance and capacity at lower costs than other platforms. Exadata runs all types of database workloads including Online Transaction Processing (OLTP), Data Warehousing (DW) and consolidation of mixed workloads. Simple and fast to implement, the Exadata Database Machine powers and protects your most important databases and is the ideal foundation for a consolidated database cloud.

The Exadata Database Machine is an easy to deploy system that includes all the hardware needed for running the Oracle Database. The database servers, storage servers and network are pre-configured, pre-tuned, and pre-tested by Oracle experts, eliminating weeks or months of effort typically required to deploy a high performance system. Extensive end-to-end testing ensures all components work seamlessly together and there are no performance bottlenecks or single points of failure that can affect the complete system.

Because all Exadata Database Machines are identically configured, customers benefit from the experience of thousands of other users that have deployed the Exadata Database Machine for their mission critical applications. Customer machines are also identical to the machines Oracle Support uses for problem identification and resolution, and the machines Oracle Engineering uses for development and testing of the Oracle Database. Hence, Exadata is the most thoroughly tested and tuned platform for running the Oracle Database and is also the most supportable platform.

The Oracle Exadata Database Machine runs the standard Oracle Database. Therefore, any application that uses the Oracle Database today can be seamlessly migrated to use the Exadata Database Machine with no changes to the application.

The Exadata Database Machine uses a scale-out architecture for both database servers and storage servers. The Exadata configuration carefully balances CPU, I/O and network throughput to avoid bottlenecks. As an Exadata Database Machine grows, database CPUs, storage, and networking are added in a balanced fashion ensuring scalability without bottlenecks. The scale-out architecture accommodates any size workload and allows seamless expansion from small to extremely large configurations while avoiding performance bottlenecks and single points of failure

Exadata also includes a unique technology that offloads data intensive SQL operations into the Oracle Exadata Storage Servers. By pushing SQL processing to the Exadata Storage Servers, data filtering and processing occurs immediately and in parallel across all storage servers as data is read from disk. Exadata storage offload reduces database server CPU consumption and greatly reduces the amount of data moved between storage and database servers. 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 and speeds database logging. Exadata storage provides an advanced compression technology, Hybrid Columnar Compression (HCC), that typically provides 10x, and higher, levels of data compression and 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 Storage Expansion Rack enables the growth of Exadata storage capacity and bandwidth for Exadata Database Machines. It is designed for database deployments that require very large amounts of data beyond what is included in an Exadata Database Machine. Standard Exadata Storage Servers, and supporting infrastructure, are packaged together in the Exadata Storage Expansion Rack to allow an easy to deploy extension of the Exadata storage configuration in an Exadata Database Machine. All the benefits and capabilities of Exadata storage are available and realized when using an Exadata Storage Expansion Rack.

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.

The Oracle SuperCluster incorporates Exadata storage technology for enhancing the performance of the Oracle Database. The SuperCluster can be used to host Oracle's Fusion Middleware, Oracle's Fusion applications, general purpose applications, as well as the Oracle Database and is a high performance integrated platform based on SPARC servers. It is an engineered system designed to host the entire Oracle software solution stack. In addition to the Exadata Storage Servers built in to the SuperCluster, Exadata Storage Expansion Racks can be used to add capacity and bandwidth to the system.

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 and availability. The Exadata Storage Server (Exadata storage or Exadata cells) is used as the storage for the Oracle Database in the Database Machine. It runs the Exadata Storage Server Software that provides the unique and powerful Exadata technology including Smart Scan, Storage Indexes, Smart Flash Cache, Smart Flash Logging, Flash Cache Compression, IO Resource Manager, Network Resource Management, and Hybrid Columnar Compression. The Exadata Storage Expansion Rack is a fast and simple means to grow the Exadata storage capacity and bandwidth of an existing Database Machine or SuperCluster deployment.

The Exadata Engineered System

The Oracle Exadata Engineered Systems is designed and built to be the highest performance and most available platform for running the Oracle Database. Traditional custom systems used to run the Oracle Database do not deliver the performance or availability that an Engineered System can deliver. The components used in custom database systems are often not balanced and misconfigured creating bottlenecks reducing overall system performance. Exadata systems are engineered and optimized end-to-end to deliver optimum performance. The hardware components, database software and libraries, operating system and device drivers, firmware, network configuration, and all the other components in the Engineered System are optimized to work together. Years of tuning have been put in to the Exadata Engineered System to deliver the best performing platform to run the Oracle Database.

Custom database system cannot achieve the availability and up time of an Exadata system. Complex inter-component failures modes are not tested and components are not engineered together as a unit in custom database systems. Exadata systems handle all possible failure modes including node failure, link failure, storage failure and switch failures. This means higher overall system uptime with less deployment and operating risk.

The Exadata 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. Exadata systems are delivered assembled, debugged, and ready-to-run. All Exadata systems are identical with no unique configuration issues so all the Exadata users benefit from the enhanced supportability of a common platform. Given the commonality across Exadata systems, Oracle Support can provide the best possible support for database deployments. Building and operating a custom platform requires top talent. Exadata frees IT personnel to focus on the business needs of their enterprise rather than component integration and testing. With end-to-end single vendor support, and end-to-end unified monitoring of all components, there is less overhead on the IT staff with an Exadata system.

Exadata systems run all existing OLTP and DW applications and the full 30 years of Oracle Database development is available, out of the box. No certification is required for application databases deployed on Exadata systems. Exadata systems can leverage the full Oracle Database ecosystem of IT skills, people, partners and technology. Exadata systems deliver the best functionality, performance and availability transparently even for the most complex applications like the Oracle E-Business Suite, PeopleSoft, Siebel, SAP, as well as custom applications built in-house.

Exadata Database Machine

There are two versions of Exadata Database Machine. The Exadata Database Machine X4-2 expands from 2 x 24 -core database servers with up to 1 TB of memory and 3 Exadata Storage Servers to 8 x 24-core database servers with up to 4 TB of memory and 14 Exadata Storage Servers, all in a single rack. The Exadata Database Machine X3-8 is comprised of 2 eighty-core database servers with 4 TB of memory and 14 Exadata Storage Servers, in a single rack. The X4-2 provides a convenient entry point in to the Exadata Database Machine family with the largest degree of expandability in a single rack. The X3-8 is for large deployments with larger memory requirements or for consolidating multiple databases on to a single system. Both versions run the Oracle Database 11g Release 2 and the Oracle Database 12c database software.

Exadata Database Machine X4-2

Exadata Database Machine X4-2

Four versions of the Exadata Database Machine X4-2 are available – the *Full Rack*, *Half Rack, Quarter Rack,* and *Eighth Rack* – depending on the size, performance and I/O requirements of the database to be deployed. One version can be expanded online to another ensuring a smooth upgrade path as processing requirements grow. In addition, the Exadata X4-2 can be easily

expanded to an 18 rack grid with 3.456 CPU cores and 12 petabytes of raw storage. Common to all X4-2 Database Machines are:

- Industry standard database servers preconfigured with: two socket twelve-core Intel® Xeon® E5-2697 v2 processors running at 2.7 GHz, up to 512 GB memory, four 600 GB 10,000 RPM disks, two 40 Gb/second InfiniBand ports, two 10 Gb Ethernet ports, four 10/1 Gb Ethernet ports, and dual-redundant, hot-swappable power supplies. Oracle Linux 5 Update 9 (with the Unbreakable Enterprise Kernel 2) and Solaris 11 Update 1are preinstalled on the database servers. At system deployment the desired operating system for the Database Machine is selected.
- Exadata Storage Servers preconfigured with: two socket six-core Intel Xeon E5-2630 v2 processors running at 2.6 GHz, 96 GB memory, 3.2 TB of Exadata Smart Flash Cache, twelve 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 1.2 TB High Performance 10,000 RPM disks or 4 TB High Capacity 7,200 RPM 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.
- 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 X4-2 are shown in the following table.

	Exadata Database Machine X4-2 Full Rack	Exadata Database Machine X4-2 Half Rack	Exadata Database Machine X4-2 Quarter Rack	Exadata Database Machine X4-2 Eighth Rack [1]
Database Servers	8	4	2	2
• **CPU cores for database processing**	192	96	48	24
• **Max Memory (GB)**	4,096	2,048	1,024	1,024
Exadata Storage Servers	14	7	3	3
• **CPU cores for SQL processing**	168	84	36	18
• **Exadata Smart Flash Cache (TB)**	44.8	22.4	9.6	4.8
• **Number of disks for database storage**	168	84	36	18
InfiniBand Switches	2	2	2	2

[1] The hardware in the Eighth Rack is physically identical to the Quarter Rack but with half the components (database processing CPU cores, SQL processing CPU cores, flash and disk) enabled. Upgrades from the Eighth Rack to Quarter Rack are done as a software activation of the hardware.

Exadata Database Machine X4-2 Hardware

Exadata Database Machine X3-8

The Exadata Database Machine X3-8 combines an outstanding scale-up and scale-out architecture by delivering a grid infrastructure containing large SMP database servers and an Exadata storage grid. Before now, a large SMP required a full rack of equipment by itself, and was difficult to scale out further. The Exadata X3-8 uses two of Sun's ultra-compact 80-core Intel-based servers to create a high-performance highly-available database grid. Each of the servers includes 2 terabytes of memory, 40 Gb/second InfiniBand for internal connectivity, and 10 Gb Ethernet for connectivity to the data center. The Exadata X3-8 can be easily expanded to an 18 rack grid with 2,880 CPU cores and 12 petabytes of raw storage. The new Exadata X3-8 delivers extreme performance for all business applications, and enables large-scale database consolidation.

The Exadata Database Machine X3-8 is available in a full rack configuration, runs Oracle Database 11g Release 2 or 12c, and includes the following technology.

- Two industry standard database servers each preconfigured with: eight socket ten-core Intel® Xeon® E7-8870 processors running at 2.40 GHz, 2 TB memory, eight 300 GB 10,000 RPM disks, eight 40 Gb/second InfiniBand ports, eight 10 Gb Ethernet ports, eight 1 Gb Ethernet ports, and dual-redundant, hot-swappable power supplies. Oracle Linux 5 Update 8 with the Unbreakable Enterprise Kernel is preinstalled on the database servers.
- Fourteen Exadata Storage Servers preconfigured with: two socket six-core Intel Xeon E5-2630 v2 processors running at 2.6 GHz, 96 GB memory, 3.2 TB of Exadata Smart Flash Cache, twelve disks (either 1.2 TB High Performance 10,000 RPM disks or 4 TB High Capacity 7,200 RPM 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.

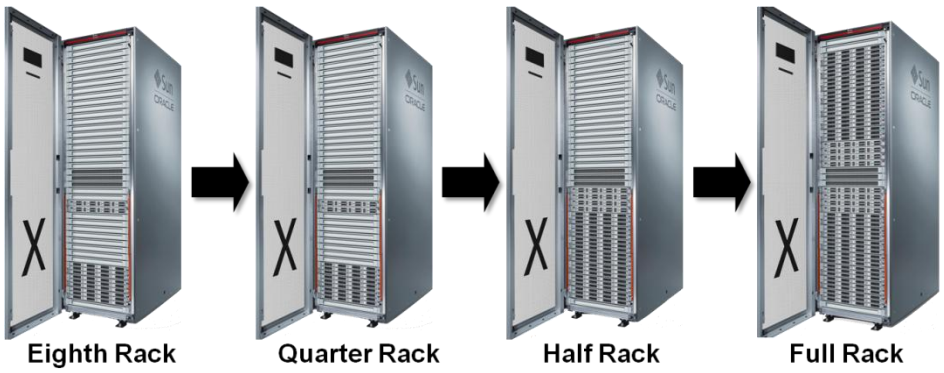
Again, the ratio of components to each other has been chosen to maximize performance, deliver a highly available system and provide the best balance of CPU to I/O power for all database applications.

	Exadata Database Machine X3-8
Database Servers	2
• **CPU cores for database processing**	160
• **Memory (GB)**	4,096
Exadata Storage Servers	14
• **CPU cores for SQL processing**	168
• **Exadata Smart Flash Cache (TB)**	44.8
• **Number of disks for database storage**	168
InfiniBand Switches	2

Exadata Database Machine X3-8 Hardware

Database Machine Upgradeability

Each model of the Exadata Database Machine X4-2 can grow in capacity and powers, ensuring a smooth upgrade path, as processing requirements grow. An online field upgrade from the Eighth Rack, to Quarter Rack to Half Rack to Full Rack can be easily performed by Oracle personnel.



Database Machine X4-2 Upgrades

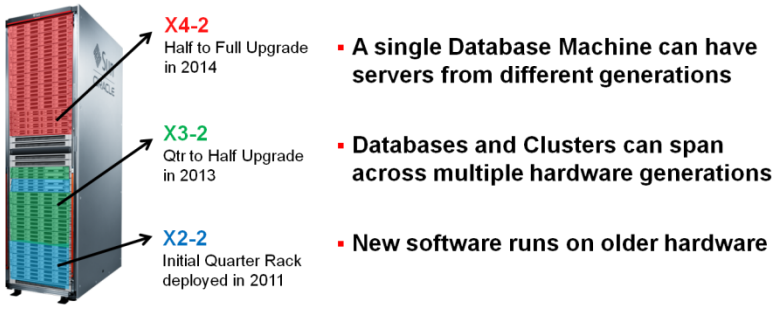
Upgrade kits are available for each of these upgrades. The hardware in the Eighth Rack is physically identical to the Quarter Rack but with half the key components (database processing CPU cores, SQL processing CPU cores, flash and disk) enabled. The upgrade from the Eighth Rack to Quarter Rack is done as a software activation of the hardware that was not active. The Quarter Rack to Half Rack and Half Rack to Full Rack upgrade includes the components necessary to bring the system to the next larger configuration and is installed by Oracle Support.

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 X4-2 systems can be connected using the InfiniBand fabric in the system to form a larger single system image configuration. Multiple Exadata Database Machine X3-8 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. Up to 18 racks can be connected by simply connecting the InfiniBand cables and installing internal InfiniBand switches. Larger configurations can be built with additional external InfiniBand switches. Any combination of X4-2 Full and Half Racks can be connected. Quarter racks can be inter-connected with other racks in two circumstances. Two Quarter Racks can be inter-connected to each other or one Quarter Rack can be connected to a configuration of Full and Half Racks. The inherent capability of the Exadata Database Machine to grow enables the support of the largest databases any application would require.

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

Eight Connected Exadata Database Machine X4-2 Racks Form a Single System

Exadata Database Machines protect your investment by allowing newer generation processors and Exadata storage to be deployed seamlessly into existing Exadata Database Machines. Similarly, new software releases are compatible with older Exadata Database Machines. An example of the upgrades and expansions possible across the generations of Exadata Database Machine follows.



X2-2 Quarter Rack Upgraded To A Half Rack With X3-2 Servers Then To A Full Rack With X4-2 Servers

Exadata Storage Server

The Exadata Storage Server runs the Exadata Storage Server Software and provides the unique and powerful Exadata software technology of the Database Machine including Smart Scan, Smart Flash Cache, Smart Flash Logging, IO Resource Manager, Storage Indexes and Hybrid Columnar Compression.

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 CPU cores in the Exadata Storage Server are dedicated to providing features such as Smart Scan SQL processing that is done in the Exadata storage.

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 processors used in the Exadata Storage Server provide Advanced Encryption Standard (AES) support enabling this.

Exadata Storage Server (Exadata Cell)

Exadata Smart Flash Cache

Exadata systems use the latest PCI flash technology rather than flash disks. PCI flash delivers ultra-high performance by placing flash memory directly on the high speed PCI bus rather than behind slow disk controllers and directors. Each Exadata Storage Server includes 4 PCI flash cards with a total raw capacity of 3.2 TB of flash memory. A full rack Exadata Database Machine X4-2 includes 56 PCI flash cards providing 44.8 TB of raw physical flash memory.

Exadata flash can be used directly as flash disks, but it is almost always configured as a flash cache in front of disk since caching provides flash level performance for much more data than fits directly into flash.

The Exadata Smart Flash Cache automatically caches frequently accessed data in PCI flash while keeping infrequently accessed data on disk drives. This provides the performance of flash with the capacity and low cost of disk. The Exadata Smart Flash Cache understands database workloads and knows when to avoid caching data that the database will rarely access or is too big to fit in the cache. For example, Exadata understands when I/Os are run for backup purposes, for table scans, and for storing temporary results that will be quickly deleted. In addition to automatic caching, administrators can optionally provide SQL directives to ensure that specific

tables, indexes, or partitions are always retained in flash. Tables can be retained in flash without the need to move the table to different tablespaces, files or LUNs as is often required with traditional storage.

Exadata's Smart Flash Cache is designed to deliver flash-level IO rates, throughput, and response times for data that is many times larger than the physical flash capacity in the machine by automatically moving active data that is experiencing heavy IO activity into flash, while leaving cold data that sees infrequent IO activity on disk. It is common for hit rates in the Exadata Smart Flash Cache to be over 90%, or even 98% in real-world database workloads even though flash capacity is more than 10 times smaller than disk capacity. Such high flash cache hit rates mean that Exadata Smart Flash Cache provides an effective flash capacity that is often 10 times larger than the physical flash cache. For example, a full rack Exadata Database Machine X4-2 often has an effective flash capacity of 440 TB.

On top of the capacity benefits provided by smart caching, Exadata Smart Flash Cache Compression dynamically increases the capacity of the flash cache by transparently compressing user data as it is loaded into the flash cache. This allows much more data to be kept in flash memory, and further decreases the need to access data on disk drives. The compression and decompression operations are completely transparent to the application and database. Exadata Smart Flash Cache Compression leverages hardware acceleration to deliver zero performance overhead for compression and decompression, even when running at rates of millions of I/Os per second or 100s of Gigabytes per second.

Flash Cache Compression benefits vary based on the compressibility of the user data. Tables that are uncompressed will see the largest benefits. Indexes will also typically compress very well. Exadata Smart Flash Cache Compression will also provide significant flash cache space expansion on top of the benefits already provided by Advanced Row and Basic table compression. OLTP applications will often see the overall logical size of the flash cache double even if they use Advanced Row Compression. Tables that use Hybrid Columnar Compression or LOB Compression will see minimal additional compression since these are already very highly compressed formats. With flash cache compression turned on, a full rack Exadata Database Machine X4-2 provides up to 90 TB of logical flash cache capacity (before database level compression is factored in).

Flash performance is often limited and bottlenecked by traditional storage architectures. In contrast, Exadata uses a combination of scale-out storage, InfiniBand networking, database offload, and PCI flash to deliver extremely high performance rates from flash. A single full rack Exadata Database Machine X4-2 achieves up to 100 GB per second of data scan bandwidth, and up to 2,660,000 random 8K read I/O operations per second (IOPS) when running database workloads. This performance is orders of magnitude faster than traditional database architectures. It is important to note that these are real-world end-to-end performance figures measured running SQL workloads with realistic IO sizes inside a single rack Exadata system. They are not component level measurements based on low level IO tools.

With the Write Back Flash Cache feature, the Exadata Smart Flash Cache also caches database block writes. Write caching eliminates disk bottlenecks in large scale OLTP and batch workloads. The flash write capacity of a single full rack Exadata Database Machine X4-2 exceeds 1,960,000 8K write I/Os per second. The Exadata Write cache is transparent, persistent, and fully redundant. The I/O performance of the Exadata Smart Flash Cache is comparable to dozens of enterprise disk arrays with thousands of disk drives.

To further accelerate OLTP workloads, the Exadata Smart Flash Cache also implements a special algorithm to reduce the latency of log write I/Os called Exadata Smart Flash Logging. The time to commit user transactions or perform critical updates is very sensitive to the latency of log writes. Smart Flash Logging takes advantage of the flash memory in Exadata storage combined with the high speed RAM memory in the Exadata disk controllers to greatly reduce the latency of log writes and avoid the latency spikes that frequently occur in other flash solutions. The Exadata Smart Flash Logging algorithms are unique to Exadata.

Exadata uses only enterprise grade flash that is designed by the flash manufacturer to have high endurance. Exadata is designed for mission critical workloads and therefore does not use consumer grade flash that can potentially experience performance degradations or fail unexpectedly after a few years of usage. The enterprise grade flash chips used in Exadata X4 have an expected endurance of 10 years or more for typical database workloads.

The automatic data tiering between RAM, flash and disk implemented in Exadata provides tremendous advantages over other flash-based solutions. When third-party flash cards or flash disks are used directly in database servers, the data placed in flash is only available on that server since local flash cannot be shared between servers. This precludes the use of RAC and limits the database deployment to the size of a single server handicapping performance, scalability, availability, and consolidation of databases. Any component failure, like a flash card, in a single server can lead to a loss of database access. Local flash lacks the intelligent flash caching and Hybrid Columnar Compression provided in Exadata and is much more complex to administer.

Real world experience has shown that server local flash cards and flash disks can become crippled without completely failing leading to database hangs, poor performance, or even corruptions. Flash products have been seen to intermittently hang, exhibit periodic poor performance, or lose data during power cycles, and these failures often do not trigger errors or alerts that would cause the flash product to be taken offline. Worse, these issues can cause hangs inside the Operating System causing full node hangs or crashes. Exadata software automatically detects and bypasses poorly performing or crippled flash. When an unusual condition is detected, Exadata will automatically route I/O operations to alternate storage servers.

Many storage vendors have recognized that the architecture of their traditional storage arrays inherently bottleneck the performance of flash and therefore have developed new flash-only arrays. These flash-only arrays deliver higher performance than traditional arrays but give up the cost advantages of smart tiering of data between disk and flash. Therefore the overall size of data that benefits from flash is limited to the size of expensive flash. Exadata smart flash caching

often provides flash level performance for data that is 10 times larger than physical flash since it automatically keeps active data that is experiencing heavy IO activity in flash while leaving cold data that sees infrequent IO activity on low-cost disk. Database and Flash Cache Compression further extend the capacity of Exadata flash. Third party flash arrays will also not benefit from Exadata Hybrid Columnar Compression.

Exadata not only delivers much more capacity than flash-only arrays, it also delivers better performance. Flash-only storage arrays cannot match the throughput of Exadata's integrated and optimized architecture with full InfiniBand based scale-out, fast PCI flash, offload of data intensive operations to storage, and algorithms that are specifically optimized for database.

Exadata Storage Capacity, Performance, Bandwidth and IOPS

Each Oracle Exadata Storage Servers comes with either twelve 1.2 TB 10,000 RPM High Performance disks or twelve 4 TB 7,200 RPM High Capacity disks. The High Performance disk based Exadata Storage Servers provide up to 6 TB of uncompressed useable capacity, and up to 1.75 GB/second of raw data bandwidth. The High Capacity disk based Exadata Storage Servers provide up to 20 TB of uncompressed useable capacity, and up to 1.5 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.

	Exadata Database Machine X3-8 and X4-2 Full Rack	Exadata Database Machine X4-2 Half Rack	Exadata Database Machine X4-2 Quarter Rack	Exadata Database Machine X4-2 Eighth Rack
Exadata Smart Flash Cache	44.8 TB	22.4 TB	9.6 TB	4.8 TB
Raw Disk Capacity				
• **High Capacity Disk**	672 TB	336 TB	144 TB	72 TB
• **High Performance Disk**	200 TB	100 TB	43.2 TB	21.6 TB
Useable Capacity	Up to	Up to	Up to	Up to
• **High Capacity Disk**	300 TB	150 TB	63 TB	30 TB
• **High Performance Disk**	90 TB	45 TB	19 TB	9 TB
(without data compression)				

Exadata Database Machine Storage Capacity

Note: When calculating raw disk capacity, 1 TB = 1 trillion bytes. Actual formatted capacity is less. Useable capacity available for databases is computed after mirroring (ASM normal redundancy) and leaving one empty disk to automatically handle disk failures.

The performance each cell delivers is extremely high due to the Exadata Smart Flash Cache. The Exadata software can simultaneously scan from Flash and disk to maximize bandwidth. The automated caching within Flash enables each Exadata cell to deliver up to 5.4 GB/second bandwidth and 190,000 database 8K 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.

	Exadata Database Machine X3-8 Full Rack	Exadata Database Machine X4-2 Full Rack	Exadata Database Machine X4-2 Half Rack	Exadata Database Machine X4-2 Quarter Rack	Exadata Database Machine X4-2 Eighth Rack
Raw Flash Data Bandwidth (without data compression)	Up to 100 GB/sec	Up to 100 GB/sec	Up to 50 GB/sec	Up to 21.5 GB/sec	Up to 10.7 GB/sec
Database Flash Read IOPS [1]	Up to 1,500,000	Up to 2,660,000	Up to 1,330,000	Up to 570,000	Up to 285,000
Database Flash Write IOPS [1]	Up to 1,000,000	Up to 1,960,000	Up to 980,000	Up to 420,000	Up to 210,000
Raw Disk Data Bandwidth • **High Capacity Disk** • **High Performance Disk** **(without data compression)**	Up to 20 GB/sec 24 GB/sec	Up to 20 GB/sec 24 GB/sec	Up to 10 GB/sec 12 GB/sec	Up to 4.5 GB/sec 5.2 GB/sec	Up to 2.25 GB/sec 2.6 GB/sec
Database Disk IOPS [1] • **High Capacity Disk** • **High Performance Disk**	Up to 32,000 50,000	Up to 32,000 50,000	Up to 16,000 25,000	Up to 7,000 10,800	Up to 3,500 5,400

Exadata Database Machine X4-2 and X3-8 I/O Performance

Exadata Storage Expansion Rack

The Oracle Exadata Storage Expansion Rack X4-2 is engineered to be the simplest, fastest and most robust way to add additional storage capacity to an Exadata Database Machine or SuperCluster. A natural extension of the Exadata Database Machine, the Exadata Storage Expansion Rack can be used to satisfy the Big Data requirements of the largest mission critical databases.

The Exadata Storage Expansion Rack is designed for database deployments that require very large amounts of data including: historical or archive data; backups and archives of Exadata Database Machine data; documents, images, file and XML data, LOBs and other large unstructured data. The expansion rack is extremely simple to configure as there are no LUNs or mount points to configure. Storage is configured and added to a database with a few simple commands, completed in minutes.

The unique technology driving the performance advantages of the Exadata Database Machine is the Oracle Exadata Storage Server, and its software. By pushing database processing to the Exadata Storage Servers all the disks can operate in parallel reducing database server CPU consumption while using much less bandwidth to move data between storage and database servers. The Exadata Storage Expansion Rack is composed of standard Exadata Storage Servers and InfiniBand switches to seamlessly integrate with your Exadata Database Machine. The Exadata Storage Expansion Rack is a high-performance, high-capacity, high-bandwidth, scale-out storage solution delivering up to 387 TB of uncompressed, and mirrored, usable capacity with a corresponding improvement in I/O bandwidth for your Exadata Database Machine deployment.

Three versions of the Exadata Storage Expansion Rack are available. From the Full Rack configuration with 18 Exadata Storage Servers; to the Half Rack with 9 Exadata Storage Servers; to the Quarter Rack system with 4 Exadata Storage Servers; there is a configuration that fits any application. One version can be upgraded online to another ensuring a smooth upgrade path as processing requirements grow. All three versions of the expansion rack are delivered with the same Exadata Storage Servers, 1.2 TB High Performance or 4 TB High Capacity disks, and Exadata Smart Flash Cache, used in the Exadata Database Machine.

	Exadata Storage Expansion Rack X4-2 Full Rack	**Exadata Storage Expansion Rack X4-2 Half Rack**	**Exadata Storage Expansion Rack X4-2 Quarter Rack**
Exadata Storage Servers	18	9	4
• **CPU cores for SQL processing**	216	108	48
• **Exadata Smart Flash Cache (TB)**	57.6	28.8	12.8
• **Number of disks for database storage**	216	108	48
InfiniBand Switches	3	3	2

Exadata Storage Expansion Rack Hardware

In addition to upgrading from a small to large Exadata Storage Expansion Rack, Oracle continues to use a building-block approach to connect the Exadata Storage Expansion Rack to the Exadata Database Machine using the integrated InfiniBand fabric to easily scale the system to any size. Exadata Storage Expansion Full, Half and Quarter Racks can be coupled to Exadata Database Machine Full, Half and Quarter Rack systems in almost any combination. Up to 18

Exadata Database Machine racks and Exadata Storage Expansion Racks can be easily connected via InfiniBand cables. An eighteen rack X4-2 configuration with an Exadata Database Machine Full Rack X4-2 and seventeen Exadata Storage Expansion Full Racks has a raw disk capacity of 15,360 TB and 3,840 CPU cores dedicated to SQL processing. Even larger configurations can be built with additional InfiniBand switches.

	Exadata Storage Expansion Rack X4-2 Full Rack	**Exadata Storage Expansion Rack X4-2 Half Rack**	**Exadata Storage Expansion Rack X4-2 Quarter Rack**
Exadata Smart Flash Cache	57.6 TB	28.8 TB	12.8 TB
Raw Disk Capacity			
• **High Capacity Disk**	864 TB	432 TB	192 TB
• **High Performance Disk**	258 TB	129 TB	57 TB
Useable Capacity	Up to	Up to	Up to
• **High Capacity Disk**	387 TB	194 TB	85 TB
• **High Performance Disk**	116 TB	58 TB	25 TB
(without data compression)			

Exadata Storage Expansion Rack Capacity

On example of the Big Data strengths of the Exadata Storage Expansion Rack is when used as a destination for Exadata Database Machine backups. A full database backup can be created at up to 27 TB/hour when backing up uncompressed data that is being written to mirrored disk in an Exadata Storage Expansion Rack. It is capable of backing up hundreds of terabytes per hour when doing incremental database backups and petabytes per hour with incremental backups of Hybrid Columnar Compressed data. A disk backup on an Exadata Storage Expansion Rack is usable directly without loss of performance and without having to do a restore. This is a unique backup capability only available when backing up to an Exadata Storage Expansion Rack. It is by far the fastest and simplest way to backup and recover your Oracle Exadata Database Machine.

As new Exadata Storage Expansion Racks are connected to an Exadata Database Machine the storage capacity and performance of the system grow. The system can be run in single system image mode or logically partitioned for consolidation of multiple databases. Scaling out is easy with Exadata Database Machine and Exadata Storage Expansion Racks. Automatic Storage Management (ASM) dynamically and automatically balances the data across Exadata Storage

Servers, online, evenly spreading the I/O load across the racks, fully utilizing all the hardware and easily integrating the expansion rack into the configuration. The I/O Resource Manager can also be used to apportion I/O bandwidth to different databases and users of the system to deliver on business service level targets.

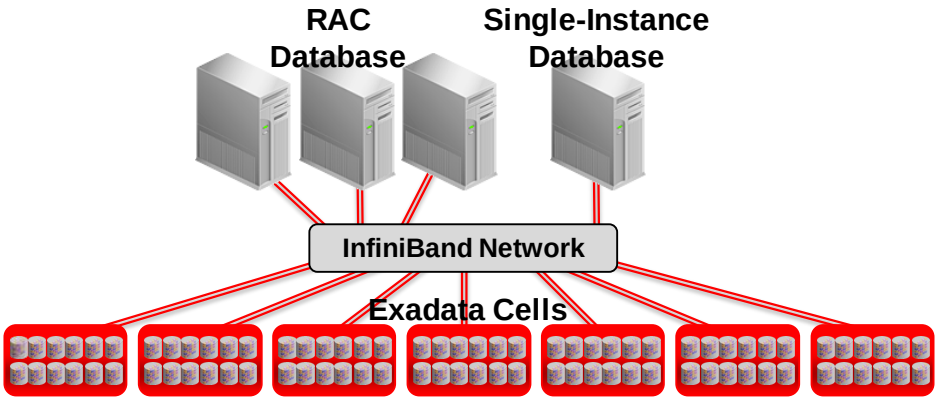
	Exadata Storage Expansion Rack X4-2 Full Rack	Exadata Storage Expansion Rack X4-2 Half Rack	Exadata Storage Expansion X4-2 Rack Quarter Rack
Raw Flash Data Bandwidth (without data compression)	Up to 130 GB/sec	Up to 65 GB/sec	Up to 29 GB/sec
Database Flash Read IOPS [1]	Up to 3,420,000	Up to 1,710,000	Up to 760,000
Database Flash Write IOPS [1]	Up to 2,520,000	Up to 1,260,000	Up to 560,000
Raw Disk Data Bandwidth (without data compression)	Up to	Up to	Up to
• **High Capacity Disk**	26 GB/sec	13 GB/sec	6 GB/sec
• **High Performance Disk**	30 GB/sec	15 GB/sec	7 GB/sec
Database Disk IOPS [1]	Up to	Up to	Up to
• **High Capacity Disk**	42,000	21,000	9,500
• **High Performance Disk**	64,000	32,000	14,400

[1] Based on 8K IO requests running SQL.

Exadata Storage Expansion Rack 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, in order to offer the best performance and availability on the Half Rack, 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.



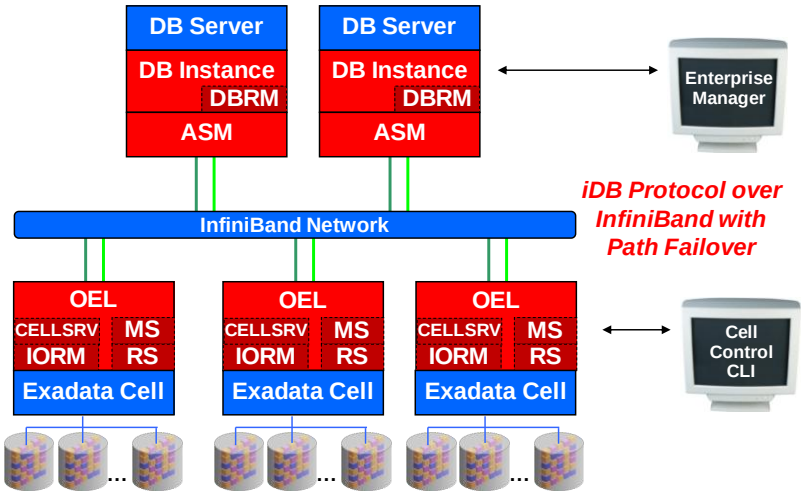
Database Machine Half Rack Deployment

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 or the Oracle Database 12c 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

The Oracle Database software 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). 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 open standard Reliable Datagram Sockets (RDS) 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 of the Oracle Database 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.

The Oracle Database File System (DBFS) was introduced with Oracle Database 11g Release 2 to provide high performance file system access to files stored in an Oracle Database. DBFS is built on SecureFiles, a feature introduced with Oracle Database 11g Release 1 to provide high performance SQL and programmatic access to files stored in an Oracle Database. DBFS and SecureFiles are very effective when used with applications that store files in the database along with metadata about those files, and DBFS in particular enables file-based tools and utilities to allow them to operate on files that are stored in the database.

DBFS has been optimized for use with the Exadata Database Machine: it provides very high throughput for files, comparable to high end storage arrays, and it allows data warehousing ETL (extract, transform and load) staging to be implemented directly on the Exadata system. DBFS also provides space for application files that can be shared across all the databases deployed on the Exadata system. Other vendors' data warehousing products require the customer to purchase a separate extra-cost storage array to deploy and manage to perform these functions. With DBFS, Exadata customers can simplify their data warehousing environments while improving performance.

Two features of the Oracle Database, the Oracle Database Quality of Service (QoS) Management and the QoS Management Memory Guard features, 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 Support for Exadata Database Machine

Oracle Enterprise Manager Cloud Control 12c uses a holistic approach to manage the Exadata Database Machine and provides comprehensive lifecycle management from monitoring to management and ongoing maintenance for the entire engineered system.

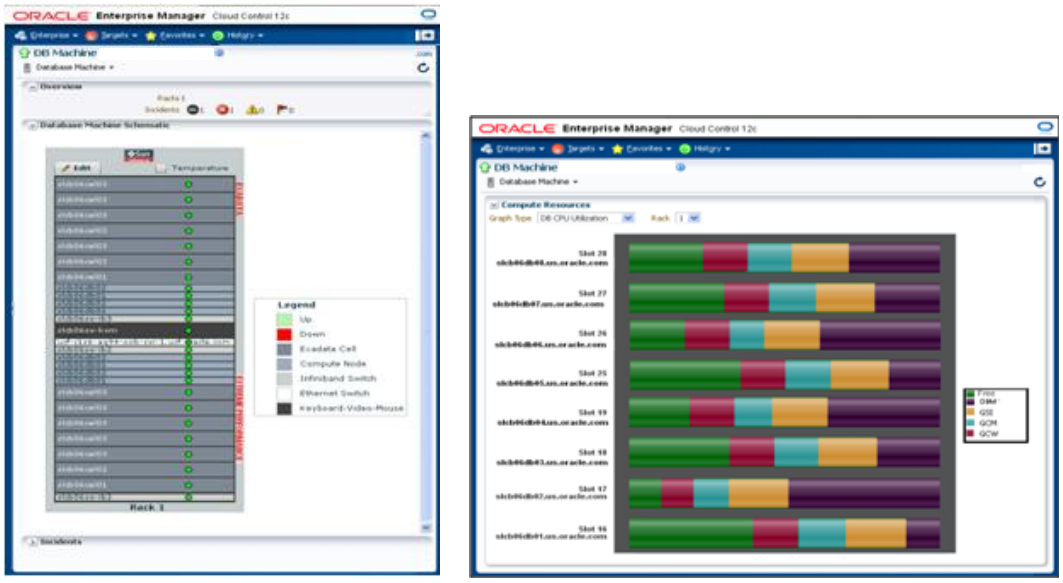
Integrated System Monitoring

Oracle Enterprise Manager provides comprehensive monitoring and notifications to enable administrators to proactively detect and respond to problems with Oracle Exadata Data Machine and its software and hardware components. Administrators can easily adjust these monitoring settings to suit the needs of their datacenter environment. When notified of these alerts, administrators can easily view the history of alerts and associated performance metrics of the problem component, such as the network performance of an InfiniBand port or the disk activity of an Exadata storage cell, to identify the root cause of the problem. With direct connectivity into the hardware components of Exadata, Oracle Enterprise Manager can alert administrators to hardware-related faults and log service requests automatically through integration with Oracle Automatic Service Requests (ASR) for immediate review by Oracle Support.

Problems that would have required a combination of database, system and storage administrators to detect in traditional systems can now be diagnosed in minutes because of integrated systems monitoring for the entire Exadata Database Machine.

Manage Many as One

Oracle Enterprise Manager provides a unified view of Oracle Exadata hardware and software where you can view the health and performance of all components such as database servers, InfiniBand switches, Exadata storage cells, Oracle databases, ASM, etc.



Monitoring Exadata using Enterprise Manager Cloud Control 12c

Oracle databases run transparently on Oracle Exadata Database Machine without any changes. However, there are times when a DBA needs to drill down from the database to the storage system to identify and diagnose performance bottlenecks or hardware faults. Enterprise Manager's integrated view of the hardware and software of Exadata allows the DBA to navigate seamlessly from the database performance pages to the associated Exadata storage server to isolate the problem, whether they may be caused by a hardware component or other databases running on the same storage subsystem. The SQL Monitoring capability that analyzes the performance of SQL executions in real time is Exadata aware and can pinpoint the plan operations of the execution plan that are being offloaded onto the Exadata storage servers, giving DBAs visibility into the efficiency of the SQL statement.

The Exadata management capabilities in Enterprise Manager are provided in-line with the health and performance features of the specific component being managed. For example, in addition to monitoring the performance of the InfiniBand network, administrators can also alter the port settings if Enterprise Manager detects port degradation. On the Exadata storage cell, administrators can configure and activate I/O resource manager plans within Enterprise Manager if they see excessive I/O resource consumption by one particular database affecting the performance of other databases on the same set of storage cells.

Consolidation Planning

As enterprises increasingly look to consolidate their disparate databases onto the Oracle Exadata infrastructure, administrators can use Consolidation Planner in Oracle Enterprise Manager to determine optimal consolidation strategies for different Exadata configurations. Using the actual hardware configurations and the server workload history stored in Enterprise Manager, Consolidation Planner analyzes the workloads of the source systems and computes the expected utilization for the consolidation plan on the target Exadata systems. Equipped with a rich library of hardware configurations, Consolidation Planner can guide administrators to define consolidation scenarios for even phantom Exadata servers, ranging from the different versions of X4-2 to X3-8. Now, businesses can make smarter and optimal decisions about the exact configurations of Exadata that is right for their database consolidation needs.

For Oracle Exadata Database Machine, management is engineered together with hardware and software to provide not just high performance and availability but also ease of management and consolidation.

Exadata Storage Server Software

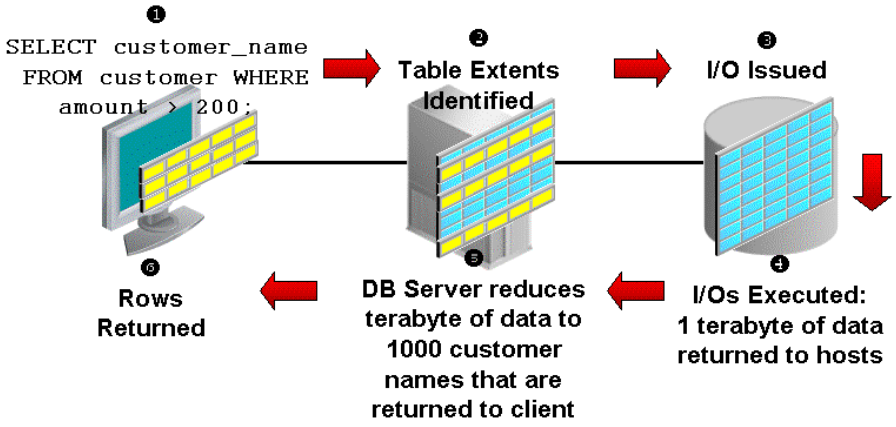
Like any storage device the Exadata Storage Server is a computer with CPUs, memory, a bus, disks, NICs, and the other components normally found in a server. It also runs an operating system (OS), which in the case of Exadata is Oracle Linux. The Exadata Storage Server Software resident in the Exadata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

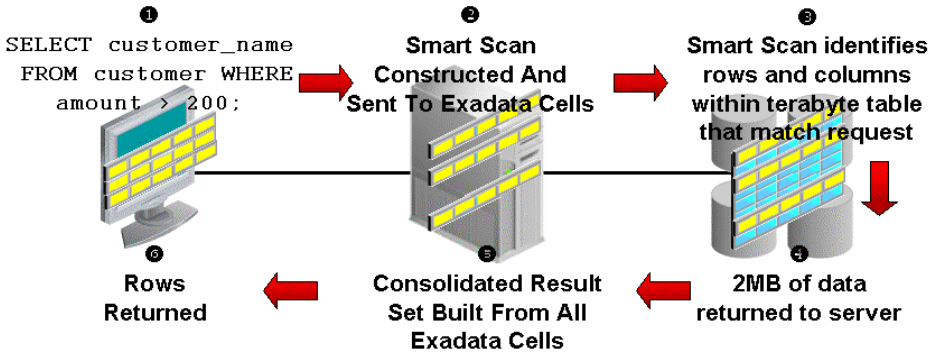
As is often the case with the large queries, the predicate filters out most of the rows read. Yet all the blocks from the table need to be read, transferred across the storage network and copied into memory. Many more rows are read into memory than required to complete the requested SQL operation. This 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.



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.

Hybrid Columnar Compression

Compressing data can provide dramatic reduction in the storage consumed for large databases. Exadata provides a very advanced compression capability called Hybrid Columnar Compression (HCC). 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 HCC 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.

HCC 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 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 HCC data is provided and column projection and filtering are performed within Exadata. Queries run directly on 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.

Exadata Smart Flash Cache Features

Oracle has implemented a smart flash cache directly in the Oracle Exadata Storage Server. The Exadata Smart Flash Cache holds frequently accessed data in very fast flash storage while most of the data is kept in very cost effective disk storage. This happens automatically without the user having to take any action. The Oracle Flash Cache is smart because it knows when to avoid trying to cache data that will never be reused or will not fit in the cache. The Oracle Database and Exadata storage optionally allow the user to provide directives at the database table, index and segment level to ensure that specific data is retained in flash. Tables can be retained in flash without the need to move the table to different tablespaces, files or LUNs like you would have to do with traditional storage and flash disks.

Exadata Smart Flash Cache software implements automatic caching of database reads and writes. This software delivers up to 2,660,000 SQL flash 8K read IOPS in a Full Rack X4 Database Machine. With the Write Back Flash Cache feature, the Exadata software ensures ultra high performance for even the most demanding OLTP databases. The write caching technology can deliver up to 1,960,000 SQL flash 8K write IOPS in a Full Rack X4 Database Machine. In addition, Exadata Smart Flash Cache is persistent across Exadata Storage Server restarts and will not require any warm up period.

The Exadata Smart Flash Cache is also used to reduce the latency of log write I/O eliminating performance bottlenecks that might occur due to database logging. The time to commit user transactions is very sensitive to the latency of log writes. Also, many performance critical database algorithms such as space management and index splits are also very sensitive to log write latency. Today Exadata storage speeds up log writes using the battery backed DRAM cache

in the disk controller. Writes to the disk controller cache are normally very fast, but they can become slower during periods of high disk IO. Smart Flash Logging takes advantage of the flash memory in Exadata storage to speed up log writes.

Flash memory has very good average write latency, but it has occasional slow outliers that can be one or two orders of magnitude slower than the average. The idea of the Exadata Smart Logging is to perform redo writes simultaneously to both flash memory and the disk controller cache, and complete the write when the first of the two completes. This literally gives Exadata the best of both worlds. The Smart Flash Logging both improves user transaction response time, and increases overall database throughput for IO intensive workloads by accelerating performance critical database algorithms.

Database writing and logging to the Exadata Smart Flash Cache is handled in a transparent manner. The database and Exadata storage software handles all crash and recovery scenarios without requiring any additional or special administrator intervention beyond what would normally be needed for recovery of the database. From a DBA perspective, the system behaves in a completely transparent manner and the DBA need not be concern themselves with the fact that flash is being used as a temporary store for datafile or redo. The only behavioral difference will be consistently low latencies for database writes.

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.

Consolidating multiple databases on to a single Exadata Database Machine is a cost saving solution for customers. With Exadata Storage Server Software 11.2.2.3 and above, the Exadata I/O Resource Manager (IORM) can be used to enable or disable use of flash for the different databases running on the Database Machine. This empowers customers to reserve flash for the most performance critical databases. With Exadata Storage Server Software 11.2.3.1 and above, there is the option to allocate resources to the different databases sharing the Exadata storage through a share-based policy in addition to the percent-based policy shown above. This mechanism can make it easier to manage resources when a larger number of databases are consolidated on to an Exadata system.

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.

Database Network Resource Management in Exadata

Exadata also implements unique database network resource management to ensure that network intensive workloads such as reporting, batch, and backups don't stall response time sensitive interactive workloads. Latency sensitive network operations such as RAC Cache Fusion communication and Log File Writes are automatically moved to the head of the message queue in server and storage network cards as well as InfiniBand network switches, bypassing any non-latency sensitive messages. Latency critical messages even jump ahead of non-latency critical messages that have already been partially sent across the network, ensuring low response times even in the presence of large network DMA operations.

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 Management and Data Protection

As discussed earlier, the Exadata cell is a server that runs the Oracle Linux (OL) as well as the Oracle Exadata Storage Server 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 30 GB of size, reserved for the OL operating system, Exadata software, and configuration metadata. The System Area contains Oracle Database 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

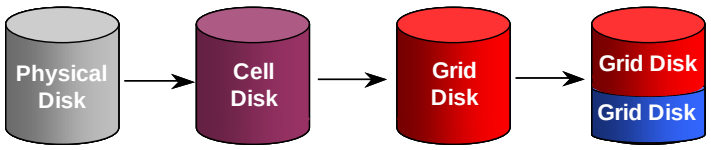
against drive failures, and to allow hot disk swapping. The remaining portion of these two disk drives is available for user data.

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.

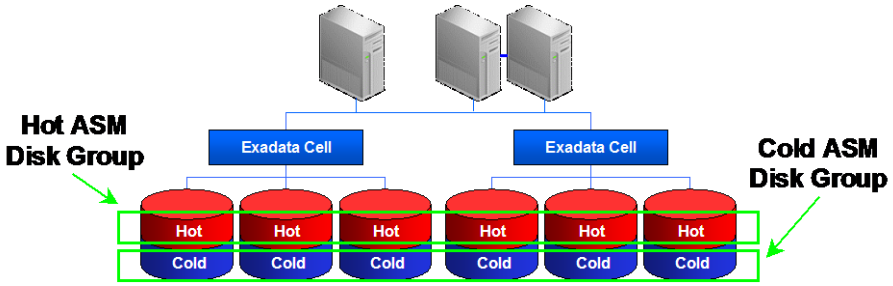


Grid Disk Virtualization

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

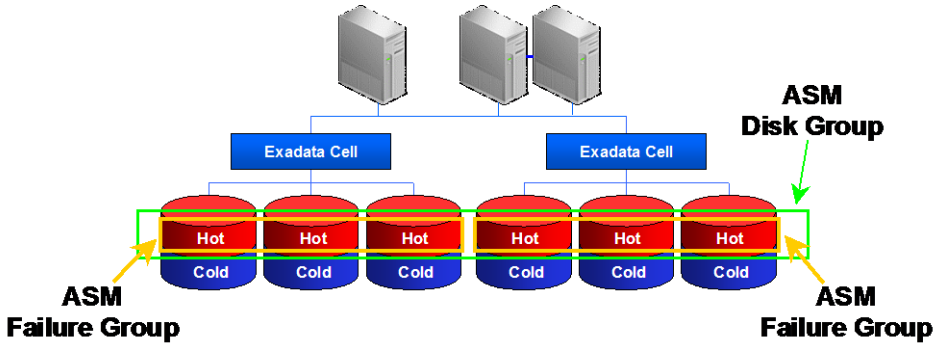
Once the Cell Disks and Grid Disks are configured, ASM disk groups are defined across the Exadata configuration. Two ASM disk groups are defined; one across the "hot" grid disks, and a second across the "cold" grid disks. All of the "hot" grid disks are placed into one ASM disk group and all of the "cold" 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.
- Grid Disks automatically made available in ASM and can be configured extremely easily. There is no need to create mount points or LUNs as with conventional storage. This all happens automatically.

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.

Data Protection in 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. For more information on these technologies and best practices for using them on Exadata Database Machine see the Exadata Maximum Availability Architecture best practice papers at http://www.oracle.com/technetwork/database/features/availability/exadata-maa-best-practices-155385.html.

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

A Technical Overview of the Oracle Exadata Database Machine and Exadata Storage Server
December 2013
Author: Mahesh Subramaniam
Contributing Authors: Juan Loaiza, Tim Shetler

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