ORACLE

Deploying Oracle Maximum Availability Architecture with Exadata Database Machine

July 14, 2020

PURPOSE STATEMENT

This document provides an overview of the high availability and disaster recovery features of the Oracle Database running on the Oracle Exadata Database Machine in the context of Oracle's Maximum Availability Architecture reference tiers. It is intended solely to help assess the business and technical benefits of adapting and configuring applications and databases to best meet Recovery Time Objective (RTO) and Recovery Point Objective (RPO) goals via high availablity and data protection solutions and best practices.

The intended audience is anyone responsible for the maintenance and lifecycle of applications (ranging from critical to development and test systems) that utilize the Oracle Database as part of the architecture. While database administration knowledge is useful in the understanding of some of the deeper concepts, the majority of this document can be read by anyone that has an understanding of basic software and database operations as well as high availability and disaster recovery architecture.

DISCLAIMER

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described in this document remains at the sole discretion of Oracle.

Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.

TABLE OF CONTENTS

OVERVIEW

The integration of Oracle Maximum Availability Architecture (Oracle MAA) operational and configuration best practices with Oracle Exadata Database Machine (Exadata MAA) provides the most comprehensive high availability solution for the Oracle Database on-premise or in the cloud.

Exadata Database Machine, Exadata Cloud at Customer (ExaCC) and Exadata Cloud Service (ExaCS) is a mature, integrated system of software, servers, storage and networking, all pre-configured according to Oracle MAA best practices to provide the highest database and application availability and performance. Mission critical applications in all industries and across both public and private sectors rely upon Exadata MAA. Every Exadata system – integrated hardware and software - has gone through extensive availability testing both internal to Oracle and by mission critical customers worldwide. The lessons learned from the experiences of this global community are channeled back into further enhancements that benefit every Exadata deployment and every Exadata customer.

This paper is intended for a technical audience: database, system and storage administrators and enterprise architects, to provide insight into Exadata MAA best practices for rapid deployment and efficient operation of Exadata Database Machine. The paper is divided into four main areas:

- Exadata MAA Architecture
- Inherent Exadata HA Benefits
- Post Deployment: Exadata MAA Configuration
- Operational Best Practices for Exadata MAA

Exadata MAA best practices documented in this white paper are complemented by the following:

- My Oracle Support Note 757552.1 is frequently updated with input directly from Oracle development to provide customers the latest information gained from continuous MAA validation testing and production deployments.
- Exadata healthcheck (exachk) and its associated Oracle Exadata Assessment Report and MAA score card. This tool is updated quarterly and is leveraged to provide a complete holistic review of your Exadata hardware, software and configuration. Refer to My Oracle Support Note 1070954.1.
- Additional MAA best practice papers that provide a deeper-dive into specific technical aspects of a particular area or topic published at www.oracle.com/goto/maa.

EXADATA MAA REFERENCE ARCHITECTURES

Exadata is the best MAA database platform for all Oracle databases addressing all unplanned outages and planned maintenance activities. Exadata is a pre-optimized, pre-configured, integrated system of software, servers, and storage that comes ready-built to implement Exadata MAA. Refer to Oracle Exadata Database Machine: Maximum Availability Architecture Presentation and Oracle

Cloud: Maximum Availability Architecture Presentation that provide blueprints that align with a range of availability and data protection requirements for Exadata on-premise and Exadata cloud customers.

For real world examples of how Exadata achieves end-to-end application availability and near zero brownout for various hardware and software outages, view the failure testing demonstrated in this Exadata MAA technical video[1] or refer to our many Exadata MAA customer case studies at https://www.oracle.com/database/technologies/ha-casestudies.html.

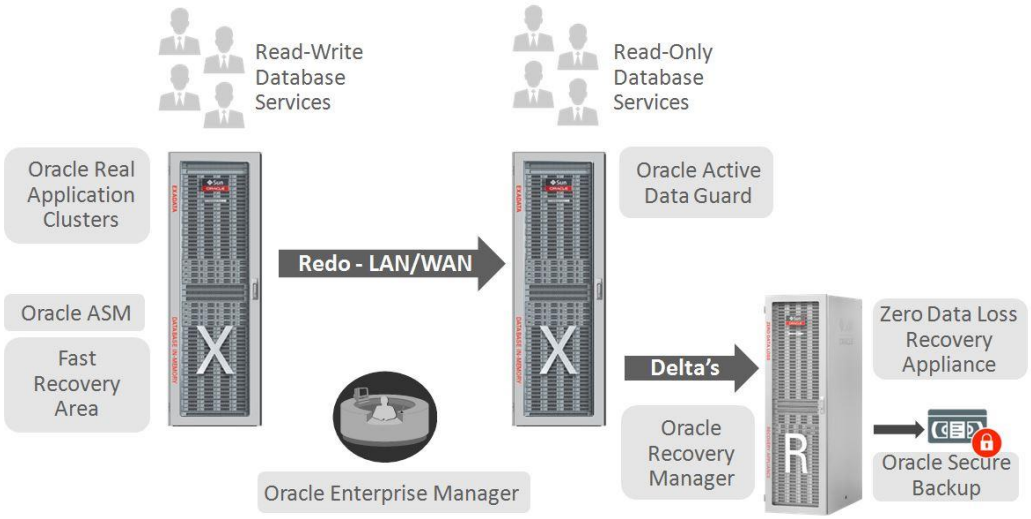

Figure 1. Basic Oracle Exadata Database Machine Configuration

Exadata MAA architecture "Gold Reference Architecture" consists of the following major building blocks:

- A production Exadata system (primary). The production system may consist of one Exadata elastic configuration or one or more interconnected Exadata Database Machines as needed to address performance and scale-out requirements for data warehouse, OLTP, or consolidated database environments.
- A standby Exadata system that is a replica of the primary. Oracle Data Guard is used to maintain synchronized standby databases that are exact, physical replicas of production databases hosted on the primary system. This provides optimal data protection and high availability if an unplanned outage makes the primary system unavailable. A standby Exadata system is most often located in a different data center or geography to provide disaster recovery (DR) by isolating the standby from primary site failures. Configuring the standby system with identical capacity as the primary also guarantees that performance service-level agreements can be met after a switchover or failover operation. For many Active Data Guard benefits, refer to Oracle Data Guard section in High Availability Overview documentation.

Note that Data Guard is able to support up to 30 standby databases in a single configuration. An increasing number of customers use this flexibility to deploy both a local Data Guard standby for HA and a remote Data Guard standby for DR. A local Data Guard standby database complements the internal HA features Exadata by providing an additional layer of HA should unexpected events or human error make the production database unavailable even though the primary site is still operational. Low network latency enables synchronous replication to a local standby resulting in zero data loss if a failover is required and fast redirection of application clients to the new primary database

- A development/test Exadata system that is independent of the primary and standby Exadata systems. This system will host a number of development/test databases used to support production applications. The test system may even have its own standby system to create a test configuration that is a complete mirror of production. Ideally the test system is configured similar to the production system to enable:

1 http://vimeo.com/esgmedia/exadata-maa-tests

- Use of a workload framework (e.g. Real Application Testing) that can mimic the production workload.
- Validation of changes in the test environment, including evaluating the impact of the change and the fallback procedure, before introducing any change to the production environment.
- Validation of operational and recovery best practices.

Exadata also supports space-efficient database snapshots that can be used to create test and development environments.

Some users will try to reduce cost by consolidating these activities on their standby Exadata system. This is a business decision with trade-offs around cost, operational simplicity and flexibility. In the case where the standby Exadata is also used to host other development and test databases, additional measures may be required at failover time to conserve system resources for production needs. For example, non-critical test and development activities may have to be deferred until failed system is repaired and back in production.

HA BENEFITS INHERENT TO EXADATA

Exadata is engineered and preconfigured to enable and achieve end-to-end application and database availability with every hardware fault such FANs, PDUs, batteries, switch, disk, flash, database server, motherboards, and DIMMs. Extensive engineering and integration testing validates every aspect of the system, including hundreds of integrated HA tests performed on a daily basis. The HA characteristics inherent in Exadata are described in the following sections.

Hardware Components

The following hardware and component redundancy is common to all models of Exadata: X8M, X8, X7, X6, X5, X4, X3, X2 and future Exadata generations.

Redundant database servers

Exadata arrives at a customer site with multiple preconfigured industry-standard Oracle Database servers running Oracle RAC and your select Oracle database release such as Oracle Database 19c Release. Oracle engineering and testing teams ensure the firmware, software, and hardware configuration is tuned and pre-configured to provide high availability and scalability. Database servers are clustered, and they communicate with each server using the high bandwidth, low latency Remote Direct Memory Access (RDMA) Network fabric. With this configuration, applications can tolerate a database server or Oracle RAC instance failure with minimal impact.

Traditionally a typical database node eviction caused by a database node failure will result in waiting on CSS misscount (defaulted to 30 or 60 seconds in most systems) before even declaring a database node has failed. During that time the entire cluster freezes and there's an application blackout. Exadata's unique Instant Failure Detection mechanism an ultra fast and safe node eviction to reduce brownout to 2 seconds or less.

In the test results shown in Figure 2 there was just two seconds of application brownout because of the Instant Failure Detection feature On non-Exadata systems, customers will observer 30 or 60 seconds of application brownouts.

Furthermore, with Exadata high bandwidth and low latency pmem cache and write back flash cache, customers can tune database initialization parameter FAST_START_MTTR_TARGET more aggressively reducing application brownout even further for instance and node failures overall. For any database parameter changes, it is still recommended to evaluate the performance impact on comparable test system prior to making production change.

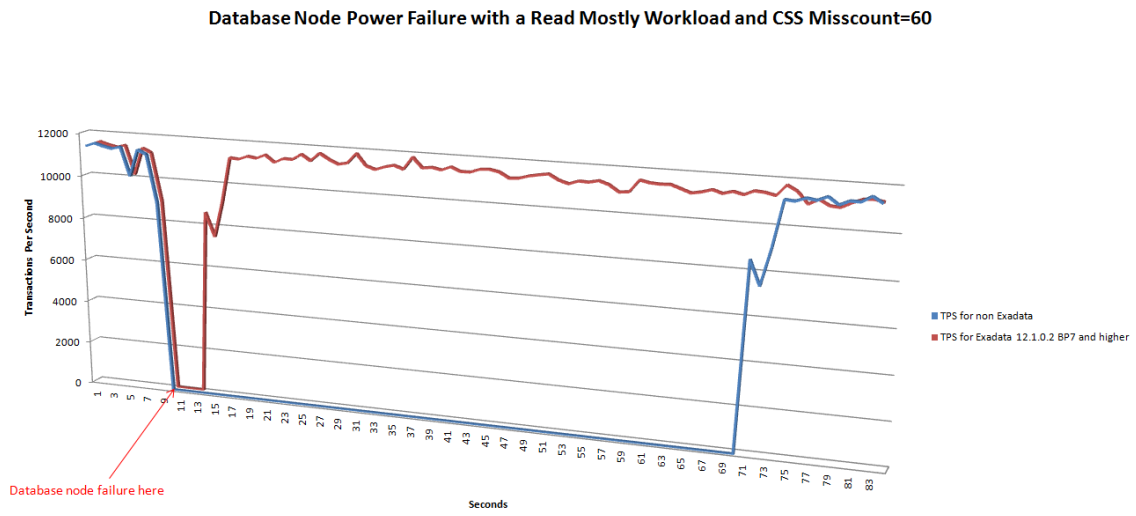

Figure 2: Database Node Power Failure

Redundant storage

Exadata storage components – database server disk drives, Exadata Storage Server disk drives, Exadata Storage Server Flash, M.2 drives, Exadata Persistent Memory Modules and Oracle Exadata Storage Servers (Exadata cell) are all redundant. Exadata Storage Servers are managed with ASM and configured to tolerate hard disk, flash disk, flash card, and complete storage server failures. Exadata Storage Servers are network-accessible storage devices with Oracle Exadata Storage Server Software pre-installed. Database data blocks and metadata are mirrored across cells to ensure that the failure of any component in an Exadata cell, or the whole cell, does not result in loss of data or availability. M.2 drives, flash drives, and hard disk drives are hot pluggable.

Exadata storage hardware and software have been engineered for the lowest application brownout for storage failures and provide extensive data protection with Exadata HARD, Exadata disk scrubbing, and ASM scrubbing. Compared to other traditional storage failures on other platforms, Exadata's application impact for disk, flash or storage server failure is significantly lower. For example, Exadata storage server failure can have less than 1 second application blackout and brownout versus seconds and minutes with other storage running Oracle databases and applications.

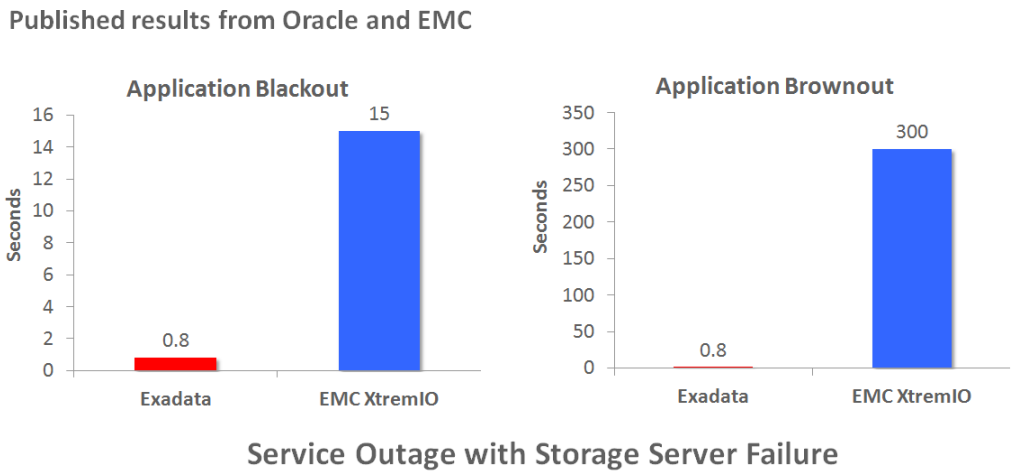

Figure 3. Storage Failure

Redundant connectivity

Redundant RDMA network fabric adapters and redundant RDMA network fabric switches are pre-configured. Configuring network redundancy for client access to database servers using Linux channel bonding is recommended and can be done at deployment time.

For network failures within an Exadata system, the observed application brownout typically ranges from zero to single digit seconds.

Redundant power supply

Exadata has redundant power distribution units (PDUs) and power supply units (PSUs) for high availability. The PDUs accept separate power sources and provide a redundant power supply to PSUs in:

- Oracle Database nodes
- Exadata Storage Cells
- InfiniBand switches
- Cisco network switch

Power supply units for Oracle Database nodes, Exadata Storage Cells, InfiniBand and Cisco switches are all hot swappable.

Software Components:

The following are standard Oracle software components explicitly optimized and validated for Exadata Database Machine.

Firmware and Operating System

All database and Exadata storage servers are packaged with validated firmware and operating system software preinstalled.

Database Server Tier

Grid Infrastructure (Oracle Clusterware and ASM) and Oracle RAC software are installed and patched to recommended software version at deployment, enabling applications to tolerate and react to instance and node failures automatically with zero to near-zero application brownout. As described in Appendix 1, all Grid Infrastructure patches and most database patches can be applied in rolling fashion.

Storage Tier

- Tolerating hard disk, flash disk, flash card and Exadata cell failures
- Applying software changes in a rolling manner
- Exadata storage cells include Oracle Hardware Assisted Resilient Data (HARD) to provide a unique level of validation for Oracle block data structures such as data block address, checksum and magic numbers prior to allowing a write to physical disks. HARD validation with Exadata is automatic (setting DB_BLOCK_CHECKSUM is required to enable checksum validation). The HARD checks transparently handle all cases including ASM disk rebalance operations and disk failures.

High Performance

Oracle Development teams who focus on high performance for OLTP and Data Warehouse applications have optimized the configuration defaults set for Exadata. In some cases, there will be different default settings for different generations of Exadata systems. These settings are the result of extensive performance testing with various workloads, both in Oracle labs and in production deployments.

Additional Exadata HA Features and Benefits

Refer to Table 1 for an overview of Exadata specific HA features and benefits. For a more detailed description of these capabilities and complete list of features, please refer to Exadata documentation such as Oracle Exadata Database Machine System Overview, Exadata Machine Maintenance Guide and Exadata Storage Server Software User's Guide.

TABLE 1: HA FEATURES AND BENEFITS

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
REDUCED HA BROWNOUT	Fast node detection and failover or Instant Failure Detection	Reduced node failure detection from as many as 60 seconds to just 2 seconds or less.	Integrated with X8M with Exadata 19.3 Grid Infrastructure 12.1.0.2 BP7 and higher

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
		Instant failure detection is a unique technology that works transparently and enables incredible availability for OLTP applications.	
	Automatic detection of Exadata storage failures with low application impact Also renamed as Instant Failure Detection	Automatic detection and rebalance. Application impact from 1 to 2 seconds delay	Continual improvements in each Exadata software release.
	Automatic detection of Exadata network failures with low application impact	Automatic detection and failover Application impact from 0 to 5 seconds delay	Continual improvements in each Exadata software release
	Zero Blackout for Exadata Storage Server Restarts and Exadata Storage Software Updates	Optimized DB notification when a storage restart has to occur to ensure zero application blacklout.	Grid Infrastructure 12c and higher. Exadata 12.1 and higher
	Reduce brownout for instance failures	With Exadata high bandwidth and low latency pmem cache and write back flash cache, customers can tune database initialization parameter, FAST_START_MTTR_TARGET, more aggressively without possible impact to the application reducing application brownout even further for instance and node failures.	Continual improvements in each Exadata database software release
	Full high redundancy advantages for Oracle files and Oracle Clusterware voting files with 3 or 4 storage cells	Oracle voting files can be placed in a high redundancy disk group with less than 5 storage server enabling all the data protection and redundancy benefits for both Oracle database and Oracle cluster. This will be done automatically through Oracle Exadata Deployment if you chose to create a high redundancy disk group.	Exadata 12.1.2.3.0 and higher
AD/ZONE FAILURE	Stretched Cluster	With Oracle 12.2 Extended Clusters on Exadata, you can expand and compliment HA benefits by providing availability for a localized site failure. This is particularly beneficial when there are isolated sites or availability domains (sometimes referred to as “fire cells” with independent power, cooling and resources) within a data center or between two metro data centers. With a properly configured Extended Cluster on Exadata, applications and databases can tolerate a complete site failure plus an additional Exadata storage cell or Exadata database server failure	Exadata 12.2.1.1.0 and higher
DATA PROTECTION	Automatic Hard Disk Scrub and Repair	Automatically inspects and repairs hard disks periodically when hard disks are idle. If bad sectors are detected on a hard disk, then	Database and GI 11.2. and 12c Exadata 11.2.3.3 and higher

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
		Exadata automatically sends a request to ASM to repair the bad sectors by reading the data from another mirror copy. By default, the hard disk scrub runs every two weeks. With Adaptive Scrubbing the frequency of scrubbing a disk may change automatically if bad sectors are discovered. If a bad sector is found on a hard disk in a current scrubbing job, Oracle Exadata Storage Server Software will schedule a follow-up scrubbing . When no bad sectors are found in a scrubbing job for that disk, the schedule will fall back to the scrubbing schedule specified by the `hardDiskScrubInterval` attribute.	Exadata 12.1.2.3.0 or higher
	Exadata H.A.R.D.	Exadata Hardware Assisted Resilient Data (HARD) provides a unique level of validation for Oracle block data structures such as data block address, checksum and magic numbers prior to allowing a write to physical disks. HARD validation with Exadata is automatic. The HARD checks transparently handle all cases including ASM disk rebalance operations and disk failures	DB_BLOCK_CHEKSUM = TYPICAL or TRUE to enable all the Exadata HARD checks.
	ASM Scrubbing	ASM provides ability to check the data integrity across all mirror extent sets. Lost write detection is possible with assistance from Oracle Support.	Grid Infrastructure 19c and higher
	Secure Erase	Erases all data on both database servers and storage servers, and resets InfiniBand switches, Ethernet switches, and power distribution units back to factory default. You use this feature when you decommission or repurpose an Oracle Exadata machine. The Secure Eraser completely erases all traces of data and metadata on every component of the machine	Exadata 12.2.1.1.0 and higher
QUALITY OF SERVICE	Cell-to-Cell Rebalance Preserves flash and PMEM Cache Population	Data rebalancing may occur for a variety of reasons. For example, a rebalance operation might happen to maintain data redundancy when a hard disk suffers a real or predictive failure. When a rebalance operation moves data to a different storage server, some of the data might be cached in the write back flash cache and persistent memory (PMEM) cache, also known as Persistent Memory Data Accelerator. Relevant PMEM cache entries are automatically replicated to the target storage server when a rebalance operation moves data to a different storage server. This new feature maintains more consistent application performance after a rebalance operation	Oracle Exadata System Software release 20.1.0. Oracle Exadata Database Machine X8M.
	Enhanced OLTP High Availability During Cell Outages, and Failures	Oracle Exadata System Software automatically populates secondary mirrors into the flash cache when data is evicted from the buffer cache Oracle Exadata System Software automatically manages the secondary mirrors in the flash cache in an optimal way so that newer or more active secondary mirrors replace the cold data in the cache. Thus, this feature provides higher availability and improved application performance by greatly reducing the secondary mirror flash cache	Oracle Exadata System Software release 19.1.0 Oracle Database 19c

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
		misses during cell or flash device failures and flash device replacements.	Exadata Write Back Flash Cache on High Capacity storage servers
		This feature is useful for OLTP workloads only. Oracle Exadata System Software does not cache the secondary mirrors for scan data. Also, this feature is only enabled for write-back Flash Cache. No secondary mirror caching is done for write-through Flash Cache.	Exadata Database Machine X6 and later (due to flash cache size requirements)
	Improved High Availability After Flash Failures	Overall system performance after flash failures has been improved. Previously, after a flash failure, Oracle ASM would start reading from the disks on the affected Exadata Storage Server as soon as flash resilvering completes. However, the Storage Server would still have a fewer than normal number of flash devices, so performance on that Storage Server was affected. Starting with Oracle Exadata System Software 18c (18.1.0), Oracle ASM starts reading from the disks only after all failed flash devices are replaced on that Storage Server and the flash cache is adequately warmed.	Exadata Storage Software 18c (18.1.0) and higher
	Database Side I/O Cancellation	Database Server Read and Write I/Os are bounded to avoid extended blackouts. For Read I/Os, the read I/O is retried on the secondary extent. For Write I/Os that writes to all extents, the target disk is taken offline unless there no redundancy.	Grid Infrastructure 18c and higher for READ I/O Cancellation Grid Infrastructure 19c and higher for Write I/O Cancellation
	I/O Latency Capping for Read Operations	Redirects read I/O operations to another cell when the latency of the read I/O is much longer than expected. This addresses the hung or very slow read IO cases due to device driver, controller, or firmware issues or failing or dying disks, flash or bad storage sectors.	Exadata 11.2.3.3.1 and higher Database and GI 11.2.0.4 BP8 and higher
	I/O Latency Capping for Write Operations	Redirects high latency write I/O operations to another healthy flash device. This addresses the hung or very slow write IO cases.	Exadata 12.1.2.1.0 and higher Database and GI 11.2.0.4 BP8 and higher Write-back flash cache enabled
	Exadata Cell I/O Timeout Threshold	Ability to set I/O timeout threshold that allows for long running I/O to be canceled and redirected to a valid mirror copy.	Exadata 11.2.3.3.1 and higher Database and GI 11.2.0.4 BP8 and higher
	Health Factor for Predictive Failed Disk Drop	When a hard disk enters predictive failure on Exadata Cell, Exadata automatically triggers an ASM rebalance to relocate data from the disk. The ASM rebalances first reads from healthy mirrors to restore redundancy. If all other mirrors are not available, then ASM rebalance reads the data from the predictively-failed disk. This diverts rebalance reads away from the predictively-failed disk when possible to ensure optimal rebalance progress while maintaining maximum data redundancy during the rebalance process. Ability to set I/O timeout threshold that allows for long running I/O to be canceled and redirected to a valid mirror copy.	Exadata storage 11.2.3.3 and higher
	Identification of Underperforming disks and Automatic Removal (aka Disk Confinement)	Underperforming disks affect the performance of all disks because work is distributed equally to all disks. When an underperforming disk is detected, it is removed from the active configuration. Exadata performs internal performance tests. If the problem with the disk is temporary and it passes the tests, then it is brought back into the	Exadata storage 11.2.3.2 and higher

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
		configuration. If the disk does not pass the tests, then it is marked as poor performance, and an Auto Service Request (ASR) service request is opened to replace the disk. This feature applies to both hard disks and flash disks.	
	I/O Resource Management	I/O Resource Management (IORM) manages disk, flash IOPS and flash cache minimum and maximum flash cache size per pluggable database or physical databases. It also now manages persistent memory. Look for new resource management features with every release.	For I/O Resource Management for Flash and Flash Cache space resource management: Exadata Storage 12.1.2.1.0 and higher and Exadata X2 generation and higher hardware
	Network Resource Management	Network Resource Management automatically and transparently prioritizes critical database network messages through the Exadata network fabric ensuring fast response times for latency critical operations. Prioritization is implemented in the database, RDMA network fabric adapters, Exadata Software, Exadata network adapters, and RDMA network fabric switches to ensure prioritization happens through the entire Exadata internal network fabric. Latency sensitive messages such as Oracle RAC Cache Fusion messages are prioritized over batch, reporting, and backup messages. Log file write operations are given the highest priority to ensure low latency for transaction processing.	Exadata Storage 11.2.3.3 Oracle Database 11.2.0.4 and higher IB switch firmware release 2.1.3-4 and higher Incorporated in new X8M ROCE fabric.
	Cell-to-Cell Rebalance Preserves Flash Cache Population	When a hard disk hits a predictive failure or true failure, and data needs to be rebalanced out of it, some of the data that resides on this hard disk might have been cached on the flash disk, providing better latency and bandwidth accesses for this data. To maintain an application's current performance SLA, it is critical to rebalance the data while honoring the caching status of the different regions on the hard disk during the cell-to-cell offloaded rebalance. The cell-to-cell rebalance feature provides significant performance improvement compared to earlier releases for application performance during a rebalance due to disk failure or disk replacement.	Exadata Storage 12.1.2.2.0 and higher Database and GI 12.1.0.2 BP11 and higher
	Exadata Smart Flash Logging	Exadata smart flash logging ensures low latency redo writes which is crucial to database performance especially OLTP workloads. This is achieved by writing redo to both hard disk and flash where the flash is used as a temporary store (cache) for redo log data to maintain consistently low latency writes and avoid expensive write outliers. Exadata smart flash logging is also needed for Extreme Flash (EF) configuration since flash devices can occasionally be slow. To avoid outliers for EF, redo writes are very selective in choosing and writing to multiple flash drives.	Exadata storage 11.2.2.4 and higher EF is only available for Exadata X5 generations and higher.
PERFORMANCE	Persistent Memory Data Accelerator	Oracle Exadata Storage Server can now use a Persistent Memory (PMEM) Cache in front of Flash Cache. Known as Persistent Memory Data Accelerator, the PMEM cache uses Intel Optane™ DC Persistent Memory Modules (DCPMM). The Database Server uses	Oracle Exadata System Software release 19.3.0

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
		remote direct memory access (RDMA) to enable 10x faster access latency to remote persistent memory. Since the persistent memory is used as a shared cache, caching capacity effectively increases by 10x compared to directly using the persistent memory modules as expensive storage. This arrangement makes it cost effective to apply the benefits of persistent memory to multi-terabyte databases.	Oracle Database 19c
	Persistent Memory Commit Accelerator	Consistent low latency for redo log writes is critical for OLTP database performance, since transactions are committed only when redo logs are persisted. Furthermore, slow redo log persistence affects critical database algorithms. With the persistent memory commit accelerator, Oracle Database 19c uses Remote Direct Memory Access (RDMA) to write redo records in persistent memory on multiple storage servers. By using RDMA, the redo log writes are up to 8x faster, and excellent resilience is provided because the redo log is persisted on multiple storage servers. On each storage server, the persistent memory area contains only the recently written log records and persistent memory space is not required for the entire redo log. Therefore, hundreds of databases can share the persistent memory area, enabling consolidation with consistent performance.	Oracle Exadata System Software release 19.3.0 Oracle Database 19c Oracle Exadata Storage Server X8M-2
	Smart Flash Log Write-Back	The Smart Flash Log Write-Back feature automatically and transparently stores the entire contents of redo log files using Exadata Smart Flash Cache in Write-Back mode, thereby eliminating the HDDs as a potential performance bottleneck. Depending on the system workload, overall log write throughput can improve up to 250%. Smart Flash Log Write-Back works transparently in conjunction with Exadata Smart Flash Log. Smart Flash Log Write-Back boosts overall log write throughput, while Exadata Smart Flash Log continues to prevent log write latency outliers. Applicable for primary and standby databases.	Oracle Exadata System Software release 20.1.0. Oracle Exadata Database Machine X7. Exadata Smart Flash Cache in Write-Back mode.
	Fast In-Memory Columnar Cache Creation	This feature provides a significant performance improvement for columnar cache creation, especially when there are concurrent workloads utilizing hard disk IO bandwidth. For example, a backup that utilizes the hard disk bandwidth no longer needs to share that bandwidth with the in-memory columnar cache creation. As a result, both the backup and the in-memory columnar cache creation run faster.	Oracle Exadata System Software release 20.1.0
	Active Bonding Network	Exadata servers can be configured with active bonding for both ports of InfiniBand card. Active bonding provides much higher network bandwidth when compared to active passive bonding in earlier releases because both InfiniBand ports are simultaneously used for sending network traffic.	Exadata X4 generation and higher hardware Exadata storage 11.2.3.3 and higher
	Exadata Smart Write Back Flash Cache Persistent After Cell Restarts	Exadata Smart Flash Cache transparently and intelligently caches frequently-accessed data to fast solid-state storage, improving database query and write response times and throughput. If there is a problem with the flash cache, then the operations transparently fail	Exadata storage 11.2.3.2 and higher

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
		over to the mirrored copies on flash. No user intervention is required. Exadata Smart Flash Cache is persistent through power outages, shutdown operations, cell restarts, and so on. Data in flash cache is not repopulated by reading from the disk after a cell restarts. Write operations from the server go directly to flash cache. This reduces the number of database I/O operations on the disks..	
	Data Guard Redo Apply Performance increase of 10+ X	Data Guard redo apply performance takes advantage of Exadata smart flash cache and overall I/O and network bandwidth enabling observed redo apply rates of up to 500 MB/sec for OLTP workloads and up to 1000 MB/sec for batch and load workloads. Traditional storage tends to bottlenecked with network or storage IO bandwidth restricting redo apply performance typically below 50 MB/sec.	Rates observed in-house MAA testing and with real world customers. Rates may vary depending on the amount of database consolidation and available system bandwidth and Exadata generation.
	In-Memory OTLP and Consolidation Acceleration	Exadata Storage Servers add a new memory cache in front of flash memory. This is similar to how the current flash cache is in front of hard disks. This feature provides 100 microsecond (µs) online transaction processing (OLTP) read IO latency, which is 2.5 times lower than the 250 µs flash OLTP read IO latency. You can use existing memory upgrade kits to add more memory to storage servers to take advantage of this feature.	Exadata Storage 18c (18.1.0) and higher Exadata X6 or X7 and higher generations Patch for bug 26923396 applied to the Oracle Database home
	In-Memory Columnar Caching on Storage Servers	Oracle Exadata System Software release 12.2.1.1.0 introduced the support for In-Memory Columnar Caching on Storage Servers for Hybrid Columnar Compressed (HCC) tables. Oracle Exadata System Software 18c (18.1.0) extends the support for In-Memory Columnar Caching on Storage Server for additional table types, specifically uncompressed tables and OLTP compressed tables. By extending the Database In-Memory format for uncompressed tables and OLTP compressed tables, smart scan queries on more table types can benefit from fast vector-processing in-memory algorithms on data stored in the storage flash cache. With this format, most in-memory performance enhancements are supported in Smart Scan including joins and aggregation. Database In-Memory format is space efficient and usually takes up less space than uncompressed or OLTP compressed formats. Storing data in Database In-memory format results in better Storage flash cache space utilization.	Exadata Storage 18c (18.1.0) and higher Oracle Database release 12c release 1 (12.1.0.2) version 12.1.0.2.161018DBBP or Oracle Database 12c release 2 (12.2.0.1) and higher Patch for bug 24521608 if using Oracle Database 12c release 1 (12.1.0.2) Recommended Patch for bug 26261327 (Enables better reverse offload functionality for complex queries)
	Patching of Exadata Storage Cells, Exadata database nodes, and Exadata Switches	Patchmgr utility (and dbnodeupdate.sh) provides patching orchestration and automation for patching Exadata Storage Cells, Exadata database nodes and Exadata switches for both online and offline options.	Patchmgr supports Exadata Storage Cells Patchmgr extended to support InfiniBand Switches with Oracle Exadata Storage 11.2.3.3.0 and higher

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
			Patchmgr to support orchestration of updates for the entire rack Exadata Storage 18c (18.1.0) and higher
	Automated Cloud Scale Performance Monitoring	Exadata Database Machine provides automated, cloud-scale performance monitoring covering a wide range of sub-systems, including CPU, memory, file system, IO, and network. This feature combines artificial intelligence, years of real-world performance triaging experience, and best practices. Oracle Exadata System Software can automatically detect performance issues and figure out the root cause without human intervention. Examples of how this feature operates include: If a spinning process is taking up all the resources on the system and impacting database performance, Oracle Exadata System Software automatically detects the CPU spin, pinpoints the exact process that is causing the spin, and generates an alert. If the Oracle database is not properly configured with huge pages according to the best practice recommendation, Oracle Exadata System Software automatically detects the misconfiguration and generates an alert for the affected database instances. There is no configuration required for this feature. To receive alerts, you must configure the notification mechanism. See Monitoring Requests and Alerts for Oracle Exadata Storage Server and ALTER DBSERVER.	Oracle Exadata System Software 19.1.0
MANAGEMENT	Online Flash Disk Replacement Exadata X7 Storage Servers	Starting with Exadata Database Machine X7-2L and X7-8, flash disks in High Capacity Storage Server can also be replaced online without server downtime.	Exadata Extreme Flash Storage Server OR Oracle Exadata System Software 18c or higher and Exadata High Capacity Storage Server X7-2 or Exadata Database Machine X7-8
	Storage Server Cloud Scale Software Update	The Storage Server Cloud Scale Software Update feature introduces a brand new cloud-scale software update process for storage servers. You point the storage servers to a software store. The storage servers download new software in the background. You can schedule the preferred time of software update. Storage servers automatically upgrade the Oracle Exadata System Software in a rolling fashion while keeping the databases online. A single software repository can be used for hundreds of storage servers. This feature provides simpler and faster software updates for Cloud and On-Premise customers.	Exadata Storage 18c and higher

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
	Performance improvements for Storage Server Software Updates	Updating Oracle Exadata Storage Server Software now takes significantly less time. By optimizing internal processing even further, the cell update process is now up to 5 times faster compared to previous releases. Even though most Exadata patching occurs with the application online, this enhancement dramatically reduces the patching window.	Oracle Exadata Storage Server Software release 12.1.2.3.0 and higher
	Performance improvements for Exadata DB Server Software Updates	Database server software update process now takes significantly less time than before and is up to 40% faster compared to previous releases. This helps reduce the cost and effort required to update the software on database server.	Exadata Storage 18c and higher
	Flash and Disk Life Cycle Management Alerts	Monitors ASM rebalance operations due to disk failure and replacement. Management Server sends an alert when a rebalance operation completes successfully or encounters an error. Simplify status management.	Oracle Database release 12.1.0.2 BP4 and later, Oracle Exadata Storage Server Software release 12.1.2.1.0 and higher.
	Cell Alert Summary	Oracle Exadata Storage Server Software periodically sends out an e-mail summary of all open alerts on Exadata Cells. The open alerts e-mail message provides a concise summary of all open issues on a cell.	Oracle Exadata Storage Server Software release 11.2.3.3.0 and higher
	LED Notification for Storage Server Disk Removal	When a storage server disk needs to be removed, a blue LED light is displayed on the server. The blue light makes it easier to determine which server disk needs maintenance.	Oracle Exadata Storage Server Software release 11.2.3.2.0 and higher
	Drop Hard Disk for Replacement	Simple command for an administrator to remove hard disk from Exadata cell. The command checks to ensure that the grid disks on that hard disk can be safely taken offline from ASM without causing a disk group force dismount. If it is successful, service LED on the disk will be turned on for easy replacement.	Oracle Exadata Storage Server Software release 11.2.3.3.0 and higher
	Drop BBU for Replacement	Simple command for an administrator to initiate an online BBU (battery backup unit) replacement. The command changes the controller to write-through caching and ensures that no data loss can occur when the BBU is replaced in case of a power loss.	Exadata X3 and X4 generations only. Exadata X5s disk controller HBAs come with 1 GB supercap-backed write cache instead of BBU.
	Minimize or eliminates false disk failures	I/Os are automatically redirected to healthy drives. The targeted unhealthy disk is power cycled. If the drive returns to normal status, then it will be re-enabled and resynchronized. If the drive continues to fail after being power cycled, then it will be dropped. Eliminates false-positive disk failures and helps preserve data redundancy, reduce operational management and avoids drop rebalance.	X5 Storage or higher since power cycle support required in chassis Only relevant for High Capacity Hard Disks and Extreme Flash SSDs
	Exadata AWR and Active Report	The Exadata Flash Cache Performance Statistics sections have been enhanced in the AWR report: 1) Added support for Columnar Flash Cache and Keep Cache. 2) Added a section on Flash Cache Performance Summary to summarize Exadata storage cell statistics along with database statistics.	Oracle Exadata Storage Server Software release 12.1.2.2.0 and higher Oracle Database release 12.1.0.2 Bundle Patch 11 and later,

AREA	FEATURE	HA BENFITS	DEPENDENTCIES
		The Exadata Flash Log Statistics section in the AWR report now includes statistics for first writes to disk and flash.I/Os are automatically redirected to healthy drives. The targeted unhealthy disk is power cycled. If the drive returns to normal status, then it will be re-enabled and resynchronized. If the drive continues to fail after being power cycled, then it will be dropped. Eliminates false-positive disk failures and helps preserve data redundancy, reduce operational management and avoids drop rebalance.	X5 Storage or higher since power cycle support required in chassis Only relevant for High Capacity Hard Disks and Extreme Flash SSDs

POST DEPLOYMENT – EXADATA MAA CONFIGURATION

The following sections provide references to complementary Exadata MAA practices:

1. Exadata Holistic Health Check: Refer to Oracle Exadata Database Machine EXAchk or HealthCheck (Doc ID 1070954.1).
2. Overview and Follow Up of Features/Solutions: Oracle Exadata Database Machine: Maximum Availability Architecture Presentation and corresponding HA documentation https://docs.oracle.com/en/database/oracle/oracle-database/19/high-availability.html
3. Exadata Database Consolidation Best Practices: Best Practices For Database Consolidation On Oracle Exadata Database Machine
4. Exadata VM Practices: Oracle Exadata Database Machine: KVM Virtualization Best Practices for RoCE/PMEM-Based Systems or Oracle Exadata and OVM - Best Practices
5. Exadata Software Updates Practices: Oracle Exadata Software Planned Maintenance
6. Exadata Maintenance Guide or Exadata docs: https://docs.oracle.com/en/engineered-systems/exadata-database-machine/books.html
7. Backup and Restore Practices: Oracle Exadata Database Machine Backup and Restore Configuration and Operational Best Practices
8. Generic Exadata MAA white papers: https://www.oracle.com/database/technologies/high-availability/exadata-maa-best-practices.html
9. Cloud MAA papers: https://www.oracle.com/database/technologies/high-availability/oracle-cloud-maa.html
10. Generic MAA papers including application failover, Active Data Guard, GoldenGate, Migration practices: https://www.oracle.com/database/technologies/high-availability/oracle-database-maa-best-practices.html

OPERATIONAL BEST PRACTICES FOR EXADATA MAA

The following operational best practices are required for a successful Exadata implementation and documented in 6 Operational Prerequisites to Maximizing Availability . Key elements are highlighted below.

- Document your high availability and performance service-level agreements (SLAs) and create an outage/solution matrix that maps to your service level agreements.

  Understanding the impact to the business and the resulting cost of downtime and data loss is fundamental to establishing Recovery Time Objectives (RTO) and Recovery Point Objectives (RPO). RTO measures your tolerance for downtime while RPO measures your tolerance for data loss. It is also likely that RTO and RPO will be different for different classes of outages. For example, server and disk failures usually have RTO/RPO of zero. A complete site failure may have larger RTO/RPO as well as less stringent performance SLAs. This is due to managing the trade-off between the potential frequency of an outage occurring and the cost or complexity of implementing HA/DR.
- Validate HA and Performance SLAs. Perform simple database node, database instance and database failure testing to validate the expected HA response including all automatic, automated, or manual repair solutions. Ensure that the application RTO and RPO requirements are met. Ensure that application performance is acceptable under different scenarios of component failures. For example, does the application continue to meet performance SLAs after node failure, Exadata Storage cell failure, and Data Guard role transition?
- Periodically (e.g. at least once a year) upgrade Exadata and database software as recommended in My Oracle Support Note 888828.1.

  Exadata will be delivered and deployed with the then current recommended HA software and system components. Once deployed it is necessary to periodically run exachk and refer to the Exadata software maintenance best practices section of the MAA scorecard to evaluate if your existing Exadata software is within recommended range. The software maintenance checks within exachk will alert you to any critical software issues that may be relevant to your environment (be sure to download the latest version of exachk before running).

  Between exachk releases a new Exadata critical issue that requires prompt attention may be identified, resolved, and information about the issue published in My Oracle Support Note 1270094.1. To receive proactive notification of newly published Alerts for Exadata critical issues from My Oracle Support, configure Hot Topics E-Mail for product Oracle Exadata Storage Server Software.

- Pre-production validation and testing of software patches is one of the most effective ways to maintain stability. The high-level steps are:
  - Review the patch and upgrade documentation.
  - Evaluate any rolling upgrade opportunities in order to minimize or eliminate planned downtime.
  - Evaluate whether the patch qualifies for Standby-First Patching, described in My Oracle Support Note 1265700.1.
  - Validate the application in a test environment and ensure the change meets or exceeds your functionality, performance, and availability requirements. Automate the procedure and be sure to also document and test a fallback procedure.
  - If applicable, perform final pre-production validation of all changes on a Data Guard standby database before applying them to a production system.
  - Apply the change in your production environment.
- Execute the Exadata MAA health check (exachk), as described in My Oracle Support Note 1070954.1. Before and after each software patch, before and after any database upgrade, or minimally every month, download the latest release of exachk and run it in your test and production environments to detect any environment and configuration issues. Checks include verifying the software and hardware and warning if any existing or new MAA, Oracle RAC, or Exadata hardware and software configuration best practices need to be implemented. An MAA score card has been added with MAA configuration checks and best practices. An upgrade module has been added to proactively detect any configuration issues pre and post database upgrade.
- Execute Data Guard role transitions and validate restore and recovery operations. Periodically execute Application and Data Guard switchovers to fully validate all role transition procedures. We recommend conducting role transition testing a minimum of once per quarter.
- Configure Exadata monitoring and Automatic Service Request[2]. Incorporate monitoring best practices as described in Enterprise Manager MAA OTN website.

Importance of a Test Environment

2 http://www.oracle.com/us/support/auto-service-request/index.html

Investment in sufficient test system infrastructure is essential to Exadata MAA. The benefits and trade-offs of various strategies for deploying test systems for Exadata are described in Table 2.

TABLE 2. TRADEOFFS FOR DIFFERENT TEST AND QA ENVIRONMENTS

TEST ENVIRONMENT	BENEFITS AND TRADEOFFS
Full Replica of the Production Exadata	Validate all patches and software changes. Validate all functional tests. Full performance validation at production scale Full HA validation especially if the replica includes the standby system.
Standby Exadata	Validate most patches and software changes. Validate all functional tests. Full performance validation if using Data Guard Snapshot Standby but this can extend recovery time if a failover is required. Role transition validation. Resource management and scheduling is required.
Shared Exadata	Validate most patches and software changes. Validate all functional tests. This environment may be suitable for performance testing if enough system resources can be allocated to mimic production. Typically, however, a subset of production system resources, compromising performance testing/validation. Resource scheduling is required.
Smaller Exadata system or Exadata with Exadata Snapshots	Validate all patches and software changes. Validate all functional tests. No performance testing at production scale. Limited full-scale high availability evaluations. Exadata snapshots are extremely storage efficient.
Older Exadata system	Validate most patches and software changes. Limited firmware patching test. Validate all functional tests unless limited by some new hardware feature Limited production scale performance tests. Limited full-scale high availability evaluations.
Non-Exadata system	Validate database and grid infrastructure software and patches only. Validate database generic functional tests. Limited testing of Exadata specific software features (e.g., HCC, IORM, Storage Index, etc.) Very limited production scale performance tests Limited high availability evaluations.

CONCLUSION

Exadata MAA is an integrated solution that provides the highest performing and most available platform for Oracle Database. This technical whitepaper has highlighted the HA capabilities that are delivered pre-configured with every Exadata Database Machine along with post-delivery configuration and operational best practices used by administrators to realize the full benefits of Exadata MAA.

APPENDIX 1: EXADATA MAA OUTAGE AND SOLUTION MATRIX

Unplanned Outages

The following outage and solution matrix in Table 3 is an example of extensive high availability testing that Oracle conducts. The MAA recommended solution is provided for each type of outage along with the expected application recovery time (RTO) assuming sufficient system resources are still available to meet your application's performance SLAs and the application has been configured to transparently fail over to an available service.. To evaluate operational readiness and evaluate if your application's performance SLAs are met, Oracle recommends simulating the key faults (e.g. instance failure, node failure, logical failures, hangs and complete database failure to validate DR) while running a real-world (using Real Application Testing and Database Replay) workload on an Exadata MAA test system. The priority column reflects suggested testing priority based on combination of probability of occurrence, importance of operational readiness, customer testing importance (not Oracle testing priority). Most outages should incur zero database downtime and a minimal application brownout for any connections. If comparing with different hardware or storage vendor, inject the same equivalent fault and repeat the same workload for both environments. For real world examples of how Exadata achieves end-to-end application availability and near zero brownout for various hardware and software outages, refer to this Exadata MAA video (http://vimeo.com/esgmedia/exadata-maa-tests) or the latest Exadata MAA presentation (https://www.oracle.com/a/tech/docs/exadata-maa.pdf) .

Whether you are deploying manual or automatic failover, evaluate end-to-end application failover time or brownout in addition to understanding the impact that individual components have on database availability. Refer to Continuous Availability - Application Checklist for Continuous Service for MAA Solutions to enable applications to minimize impact.

If there are sufficient system resources after an unplanned planned outage, the application impact can be very low as indicated by the table below.

TABLE 3. UNPLANNED OUTAGE/SOLUTION MATRIX

OUTAGE SCOPE	FAULT INJECTION PROCESS	EXADATA MAA	PRIORITY
site failure		Seconds to 5 minutes[3] Database Failover with a Standby Database Complete Site Failover Application Failover	LOW BUT WORTH TESTING FOR DR READINESS
clusterwide failure or production Exadata Database Machine failure		Seconds to 5 minutes Database Failover with a Standby Database Complete Site Failover Application Failover	LOW BUT WORTH TESTING FOR DR READINESS
computer failure (node) or RAC database node failure (simulating the impact of hardware failure, RAC node evictions, reboots or motherboard failure)	1. Unplug or forcefully power off database node 2. Wait 30 seconds or more 3. Restore power and power up database node, if needed 4. Wait for database node to be fully up	Small application downtime for cluster detection, cluster reconfiguration and instance recovery. For Exadata, cluster detection can be as low as 2 seconds. Managed automatically by Oracle RAC Recovery for Unscheduled Outages	HIGH

3 Recovery time indicated applies to database and existing connection failover. Network connection changes and other site-specific failover activities may lengthen overall recovery time.

OUTAGE SCOPE	FAULT INJECTION PROCESS	EXADATA MAA	PRIORITY
Database Instance failure or RAC database instance failure	Kill -11 PMON background process or shutdown abort the target instance	Small application downtime for affected connections. For the affected connections of the failed instance, brownout will consist of cluster reconfiguration (1sec) and instance recovery which is significantly faster on Exadata with Exadata write back flash cache. No database downtime[4]. Managed automatically by Oracle RAC Recovery for Unscheduled Outages	HIGH
Exadata Storage Server failure (simulating a storage head failure)	1. Unplug or forcefully power off storage cell 2. Wait longer than ASM disk repair timer	Small application impact with sub-second cell storage delay using our InfiniBand fabric fast detection mechanism.	LOW
Exadata disk pull and then push	1. Pull disk out Wait 10 seconds or more 2. Plug the same disk drive back in the same slot	Zero application brownout with Exadata write-back flash cache. Exadata and Oracle ASM tolerate storage failures and quickly redirect I/O to mirror(s) with minimum service level impact. Oracle can distinguish between a user pulling a good disk and a true failed disk. For a disk pull and push, a simple ASM resynchronization is done of the delta changes.	LOW
Exadata disk failure	Use the simulation commands: 1. alter physicaldisk <disk controller:disk slot #> simulate failuretype=fail 2. wait 1 minute 3. alter physicaldisk <disk controller:disk slot #> simulate failuretype=none	A true disk failure results in an immediate drop of the failed disk and subsequent ASM rebalance with no service level impact. Starting in Exadata cell 11.2.3.2.0 or higher, a blue LED light will indicate when the failed disk can be replaced.	HIGH
Exadata flash disk or flash DOM failure	1. Cannot physically pull the flash disk. Simulation command: alter physicaldisk <physicaldisk name of flash module> simulate failuretype=fail 2. Wait 1 minute 3. End Simulation command: alter physicaldisk <physicaldisk name of flash module> simulate failuretype=none	Small application impact with write back flash cache and fast repair of stale data.	MEDIUM

4Database is still available, but portion of application connected to failed system is temporarily affected.

OUTAGE SCOPE	FAULT INJECTION PROCESS	EXADATA MAA	PRIORITY
Power failure or PDU failure or loss of power source or supply to any computer or Exadata cell storage server	1. Pull power to one of the PDU	No application brownout due to redundant power failure.	LOW
human error		< 30 minutes[5] Recovering from Human Error	HIGH
hangs or slow down		See *Oracle Database High Availability Overview* documentation for solutions for unplanned downtime and for Application Failover	HIGH

5 Recovery times from human errors depend primarily on detection time. If it takes seconds to detect a malicious DML or DLL transaction, then it typically only requires seconds to flash back the appropriate transactions, if properly rehearsed. Referential or integrity constraints must be considered.

CONNECT WITH US

Call +1.800.ORACLE1 or visit oracle.com.
Outside North America, find your local office at oracle.com/contact.

blogs.oracle.com facebook.com/oracle twitter.com/oracle

Deploying Oracle Maximum Availability Architecture with Exadata Database Machine
July, 2020

Author: Lawrence To
Contributing Authors: Michael Nowak, Glen Hawkins