

Exadata is Still Oracle

Richard Burns
Teradata Corporation

TERADATA®

THE BEST
DECISION
POSSIBLE™

Exadata is Still Oracle

Table of Contents

<i>Executive Summary</i>	2
<i>Introduction</i>	3
<i>Oracle Exadata Database Machine Overview</i>	3
<i>Exadata Storage Server</i>	3
<i>Oracle Exadata Database Machine</i>	6
<i>Summary of Oracle Exadata Architecture</i>	8
<i>Exadata for Data Warehousing – A Critique</i>	8
<i>Exadata is NOT Intelligent Storage</i>	9
<i>Exadata is NOT Shared Nothing</i>	10
<i>Exadata does NOT Enable High Concurrency</i>	11
<i>Exadata does NOT Support Active Data Warehousing</i>	12
<i>Exadata does NOT Provide Superior Query Performance</i>	13
<i>Exadata is Complex</i>	15
<i>Conclusion</i>	16

Executive Summary

At the Oracle Open World conference in September 2010, Oracle introduced the third version of its Exadata platform, mainly a hardware upgrade sporting the latest generation Sun Intel®-based servers. Since its introduction of Exadata in 2008, Oracle has blitzed the marketplace, extolling Exadata technology with great fanfare, but only referencing customers executing simple batch reports. Exadata may help Oracle to address basic 1980's reporting problems, but it does little to handle the complex workloads and analytics demanded from today's active data warehouses.

In our view, fundamental database issues remain with Oracle Exadata. Exadata still relies on the same Oracle shared disk architecture that was designed to optimize transaction processing performance but which is ill-suited to manage complex data warehouse tasks and active analytic workloads.

While Exadata improves Oracle's I/O performance, Exadata does not tackle Oracle's underlying performance and scalability problems with large-scale data warehousing that stem from its shared disk architectural foundation. Analysis shows that resource contention continues to limit Exadata I/O performance despite its increased I/O bandwidth. Many query operations remain unaffected by Exadata and are subject to the same resource sharing constraints as previously. These exhibit the same poor performance characteristics as they did before Exadata.

Exadata is Still Oracle

Despite Oracle's claims that Exadata solves its performance and scalability problems in data warehousing, a close examination of the Exadata architecture reveals how little Oracle has really changed, how few of Oracle's classic data warehousing performance issues are addressed by Exadata, and how complex Exadata really is.

In the end, Exadata delivers far less improvement in data warehouse performance than Oracle promises.

Introduction

In September 2010, at the Oracle Open World conference, Oracle introduced the third generation of its Exadata database platform. The new edition is primarily a hardware upgrade to the latest generation Sun Intel-based servers. Oracle Exadata has been successful upgrading existing Oracle operational databases and applications, but Exadata's success in data warehousing and business intelligence uses, especially for large-scale business-critical applications, is still unproven nearly three years after its introduction.

Since its introduction, Oracle has blitzed the marketplace, extolling Exadata technology with great fanfare, but only referencing customers executing simple batch reports. Exadata may adequately address basic 1980's reporting problems, but it does little to handle the complex workloads and analytics demanded from today's active data warehouses. In our view, fundamental database issues remain with Exadata.

Exadata still relies on the same Oracle shared disk architecture that was designed to optimize transaction processing performance but which is ill-suited to manage complex data warehouse tasks and active analytic workloads.

In this paper, we review the latest members of the Exadata product family, and conclude that the Oracle Exadata Database Machine, even with these upgrades, falls well short of meeting critical needs of large- and medium-scale data warehouse users.

The fundamental cause of Oracle's data warehousing limitations, in our analysis, remains Oracle's shared disk architecture. Oracle's shared disk approach, even in its Exadata incarnation, continues to be a poor match for the requirements of large-scale data warehousing. Exadata may well be the "world's best OLTP database," as Larry Ellison claims, but Oracle's fundamental dependence on a shared resource design limits its ability to overcome its shortcomings for the more data-intensive requirements of data warehousing.

Despite Oracle's claims that Exadata solves its performance and scalability problems in data warehousing, a close examination of the Exadata architecture reveals how little Oracle has really changed, how few of Oracle's classic data warehousing performance issues are addressed by Exadata, how narrow is the class of business intelligence queries that significantly benefits from Exadata, and how complex and costly Exadata really is.

In the end, Exadata provides far less improvement in data warehouse performance than Oracle promises.

Oracle Exadata Database Machine Overview

The Oracle Exadata Database Machine is a complete, pre-configured Oracle RAC database system that combines Oracle RAC database servers with new Exadata Storage Servers, all hosted on an Intel Xeon® hardware platform produced by Oracle's Sun division. The latest generation offers two distinct Exadata products. In the Oracle Exadata Database Machine X2-2 database servers contain two Intel Westmere six-core processors, while in the Oracle Exadata Database Machine X2-8 each server contains eight Intel Nehalem ten-core processors. According to Oracle, the X2-2 replaces Exadata V2 and is aimed at data warehousing while the X2-8 is a new product intended for operational database consolidation.

The Exadata Storage Servers provide an alternate storage sub-system for Oracle database systems; one that is designed to improve I/O performance and scalability for Oracle databases. Oracle uses the same Exadata Storage Server configuration for both the X2-2 and the X2-8 products.

Exadata Storage Server

The Oracle Exadata Storage Server provides special-purpose storage for Oracle databases. It replaces SAN or NAS storage systems from third-party vendors that have been used to provide Oracle shared

Exadata is Still Oracle

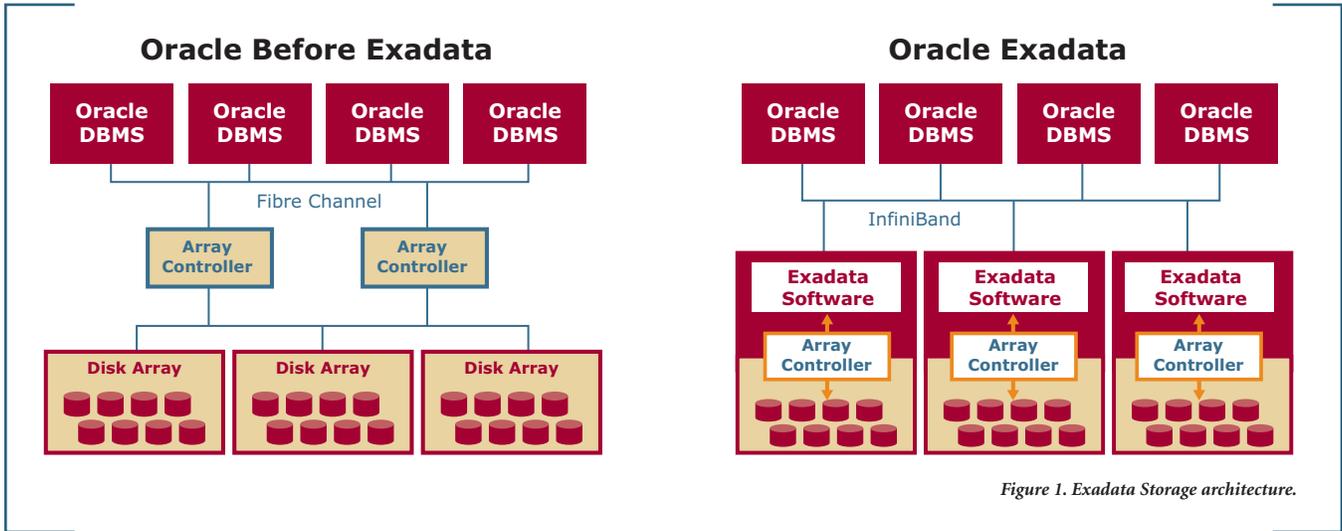


Figure 1. Exadata Storage architecture.

storage in the past. Exadata Storage Servers, or cells, replace the disk arrays and controllers, while an InfiniBand® network provides connectivity to the Oracle RAC database servers. From the viewpoint of the Oracle database, Exadata storage is treated the same as the SAN or NAS storage subsystem it replaces. In other words, from the perspective of the Oracle database, Exadata is simply another storage device. It is managed by Oracle’s Automated Storage Manager in the same way as any other storage device in an Oracle database system (See Figure 1.).

Exadata Storage Cells are based on standard Intel Xeon processors. The same Exadata Storage infrastructure is used for both X2-2 and X2-8 products. Exadata servers employ Intel’s latest generation processor, code named *Westmere*. In X2-based systems, Exadata cells are built on Sun x4270 M2 servers that contain dual six-core Xeon L5640 processors running at 2.26 GHz, with 24GB of memory per cell and 12 internal disks for data storage connected through a PCI-mounted disk array controller. Two disk configurations are available, one containing 12 high-speed

(15K RPM) 600GB SAS disks, called the high-performance option, and another with 12 slower (7200 RPM), high capacity 3TB drives, referred to as the high capacity option.

The 600GB SAS storage option provides up to 7.2TB of spinning storage per server and provides up to 1.8 GBps of data bandwidth to the Oracle database. The 3TB SAS storage option offers up to 36TB of spinning storage and up to 1 GBps of data bandwidth (See Figure 2.). Oracle strongly recommends the 600GB disk storage option for users concerned about query performance and reliability.

Each Exadata cell also contains 384GB of flash storage implemented via four PCI-e cards that hold 96GB of flash storage each.

Storage Type	Storage Capacity	User Data Volume	Data Bandwidth
600GB SAS	7.2TB	2TB	1.8 GBps
3TB SAS	36TB	7TB	1 GBps

Figure 2. Oracle Exadata Storage Server capacity.

Exadata is Still Oracle

The Exadata flash configuration is unchanged from the previous Exadata V2 system. By eliminating the mechanical delays inherent in rotating disk technology, Exadata Flash Storage supports higher random I/O throughput for a portion of the data. And it is primarily used as cache space for data aging out of Oracle memory-based buffer cache.

The purpose of Exadata is to improve I/O performance for both OLTP and data intensive business intelligence applications. Exadata achieves improved I/O performance in four ways by:

- > Employing a high-speed Infiniband network between every Exadata cell and each Oracle database server.
- > Expanding the bandwidth of the storage network as Exadata cells are added.
- > Filtering only data of interest to executing queries in Exadata processors to reduce data volume before transmitting data from Exadata cells to the Oracle database servers.

- > Adding flash storage to improve random I/O throughput for some of the data.

For many years, Oracle has advocated configuring data warehouse I/O subsystems to deliver high sustained data rates, and scaling I/O bandwidth to maintain that rate as the system grows. Oracle's best practices, reflected in its Oracle Optimized Warehouse reference architectures, specify I/O configurations designed to meet these objectives. Leading Oracle customers have been deploying large data warehouse systems that achieve these goals within Oracle's shared disk architecture with very large, high bandwidth, and quite expensive storage subsystems. So supporting parallel I/O across a high bandwidth, scalable storage network is not a new concept for Oracle. By delivering a hardware-based solution, Exadata simply builds the storage network into its architecture to guarantee adequate I/O performance.

The principal innovation of Exadata is to migrate some query processing steps to the Exadata cells. With Exadata, Oracle query processing is split into two stages, running on separate groups of servers connected to one another by an Infiniband network (See Figure 3.). In the first stage, the Exadata software retrieves data into Exadata cells, performs column projections and row restrictions based on the SELECT list and the WHERE clause predicates of the query, decompresses data and reassembles rows as needed, and returns filtered row sets across the network to the Oracle database server or Real Application Cluster (RAC), where the second stage occurs. In the second stage, the Oracle database performs the remaining query operations, which may include sorting, group by, aggregation, and analytic operations, and returns the answer set to the requestor.

The Oracle Exadata Storage Server is connected to the Oracle RAC system via dual Infiniband links, each running at 40 Gbps. Multiple Exadata Storage Servers can be linked together to provide a scalable data access layer for the Oracle DBMS, though since Exadata V2, this approach appears to be discouraged by Oracle. An Exadata Storage Server can contain data for multiple Oracle databases, and individual Oracle databases can be deployed across clusters of Exadata Storage Servers.

Oracle Automated Storage Management provides software mirroring for data

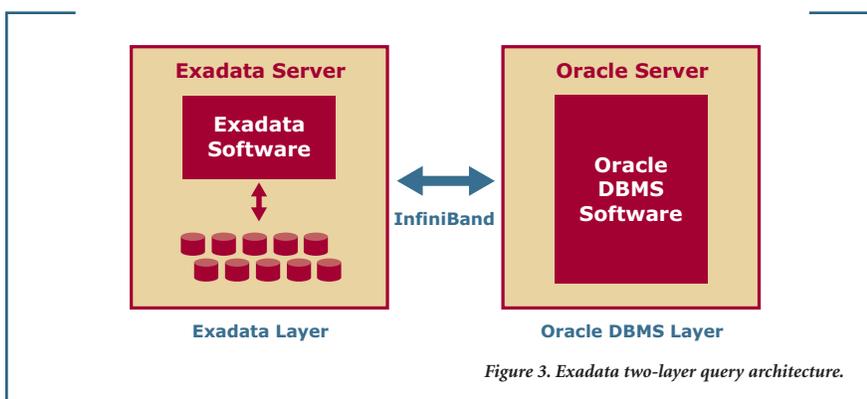


Figure 3. Exadata two-layer query architecture.

Exadata is Still Oracle

protection. The Exadata Storage Server comes with Oracle Enterprise Linux, Oracle Exadata Storage software, and management software already installed. Each Exadata Storage Server is a self-contained server, holding database data and running Exadata Storage software.

Exadata is currently available only for Oracle Enterprise Linux installations. Oracle systems deploying Exadata Storage Servers are required to run Oracle 11g, Release 2 (11gR2) or higher.

Oracle Exadata Database Machine

The primary Exadata product however, is a pre-configured, appliance-like database machine that integrates Exadata Storage Servers and Oracle RAC Database Servers in a single system connected via a high-speed Infiniband communication fabric. In the latest generation, Oracle offers two distinct Exadata products, the X2-2 and the X2-8.

Oracle Exadata Database Machine X2-2

The Oracle Exadata Database Machine X2-2 is a hardware refresh of the previous Exadata V2 system. It replaces V2's Nehalem quad-core processors with

Westmere six-core processors in both the Oracle database and Exadata servers. And it increases the memory on the Oracle database nodes. No software changes are added, and the Exadata storage devices remain the same as in Exadata V2. Oracle is targeting the X2-2 product at data warehousing applications.

The Oracle Exadata Database Machine X2-2 is available in several configuration sizes. A full cabinet Oracle Exadata Database Machine includes the following components packaged in a standard 19-inch rack (42U):

- > Eight Oracle RAC database servers running Oracle 11gR2 software on Sun x4170 M2 servers with dual Intel Xeon X5675 quad-core processors running at 3.06 GHz and 96GB of memory, optionally expandable to 144GB
- > Fourteen Exadata cells configured with 384GB of PCI mounted Flash Cache and either 12 x 600GB SAS or 12 x 3TB SAS disks
- > Three Sun Infiniband switches for scalable inter-processor communications
- > Ethernet network for client communications

The Oracle Exadata Database Machine X2-2 delivers an Optimized Warehouse configuration designed to deliver a peak I/O bandwidth slightly greater than the 260MB per second per processor core that Oracle advocates for best system performance.

Each Oracle RAC database server comes with Oracle Database 11gR2 Enterprise Edition pre-loaded and includes software components that Oracle strongly recommends for large-scale data warehousing:

- > Real Application Clusters
- > Partitioning
- > Hybrid Columnar Compression
- > High availability options
- > Enterprise Manager Diagnostics and Tuning Pack

Each Oracle Exadata Database Machine X2-2 full cabinet can hold up to 100TB of spinning disk with 600GB SAS disks, and up to 504TB with 3TB SAS drives (See Figure 4.). Oracle claims that up to eight cabinets can be connected to expand system capacity before additional external Infiniband switches must be added. Field experience to date however, has rarely spotted even a two-cabinet Oracle Exadata Database Machine devoted to a single database. Multi-cabinet Exadata configurations appear mainly for hosting multiple databases on dedicated servers in a consolidated hardware cluster.

Storage Type	Storage Capacity	User Data Volume	Data Bandwidth
600GB SAS	100TB	28TB	25 GBps
3TB SAS	504TB	150TB	14 GBps

Figure 4. Oracle Exadata Database Machine capacity.

Exadata is Still Oracle

Oracle Exadata Database Machine X2-8

The Oracle Exadata Database Machine X2-8 is a new Exadata product based on the Sun x4800 server. In recent product offerings, Oracle has stressed the price-performance benefits of clusters of smaller servers like the Exadata X2-2 configuration and previous Exadata generations. The X2-8 represents a return to Oracle configurations based on large processor count SMP servers. It allows Oracle to offer more processing power and more memory on smaller RACs, thus reducing the scalability



Figure 5. Oracle Exadata Database Machine.

and reliability problems that have confronted large RAC environments. Large SMP servers still command a price premium, however, and the X2-8 is no exception. The X2-8 hardware costs 50% more per cabinet than the equivalently powered X2-2 system. Software and maintenance costs are also higher due to the database license requirements for the higher core count on the X2-8 system.

The Oracle Exadata Database Machine X2-8 starts at a full cabinet configuration, containing two Sun x4800 servers, and is available in full cabinet increments. The X2-8 combines the Sun 4800 Oracle database servers with the same Exadata storage infrastructure as the X2-2 system. A full cabinet Exadata X2-8 system contains these components:

- > Two Oracle RAC database servers running Oracle 11gR2 software on Sun x4800 servers (5U) with eight Intel Xeon X7560 ten-core processors running at 2.26 GHz and 2TB of memory
- > Fourteen Exadata cells configured with 384GB of PCI mounted Flash Cache and either 12 x 600GB SAS or 12 x 3TB SAS disks
- > Sun Infiniband switches for scalable inter-processor communications
- > Ethernet network for client communications

A full cabinet Exadata X2-8 system contains 160 cores and 4TB of memory, and has roughly the same processing power as a full cabinet X2-2 system. The X2-8 has the same Exadata storage as the

X2-2 product (See Figure 4.), and Oracle claims that up to eight X2-8 cabinets can be linked together. Like the X2-2, the X2-8 product comes with Oracle Database 11gR2 preloaded.

Oracle's target market for the Exadata X2-8 product is operational database consolidation, so is not expected to be widely used for data warehousing. As a result, our analysis will focus primarily on the X2-2 product.

Hybrid Columnar Compression

On the software front, Oracle has added Hybrid Columnar Compression (HCC), an Oracle 11gR2 feature exclusively for Exadata systems. HCC works by taking rows in adjacent data blocks, called a compression group, vertically partitioning them by column, compressing each column partition, and storing the resulting compressed column partitions side-by-side in one or more data blocks as needed.

While Exadata SmartScan can operate on compressed data, subsequent query processing requires decompression and reassembly of rows. This typically requires multiple I/Os, perhaps as many as the number of blocks in a compression group, and significant processing power to decompress requested columns and reassemble column values into rows.

HCC offers two compression modes – query and archive. Query mode is prescribed for active data, with archive mode reserved for dormant data. Archive mode yields a higher compression ratio at much higher cost to compress and decompress.

Exadata is Still Oracle

Oracle claims that HCC in query mode provides up to 10X compression with low impact on query performance. Like any compression technique, user mileage will vary, but skepticism is certainly warranted. For real-world data warehouses it seems reasonable to expect that fitting ten pounds of data in a one-pound bag will be rare. What we have seen to date is roughly half that compression at best.

By shrinking data volume, HCC reduces the I/O necessary to process data-intensive queries. In many cases however, the cost of decompression and row reassembly offsets the I/O savings, resulting in little, if any, performance benefit to either CPU time or query response time. This may be part of the motivation for Oracle's decision to limit availability of the HCC feature to Exadata systems, which has ample processor power in the Exadata layer. Consequently, the main benefit of HCC is likely to be storage savings rather than improved query performance.

The Exadata HCC feature has other costs that limit its use. HCC can only be applied at bulk load time. SQL inserts are not eligible for compression using HCC. Updates to HCC compressed data cause the compression unit to revert to row organization, either in uncompressed format, or compressed using other, less aggressive Oracle compression options, some of which are optional, extra cost features. HCC also effectively disables row level locking within the compression unit by acquiring the same locking level on all rows in the compression unit as the most stringent lock request. Because of these

limitations, Oracle recommends HCC for use with static data only. In other words, the HCC feature is not applicable to actively updated data warehouses required for operational business intelligence.

Summary of Oracle Exadata Architecture

Compared to the previous Exadata generation, in the X2-2 product Oracle has modestly upgraded the hardware per cabinet, including more powerful Intel processors and more memory in the database tier. Oracle has also introduced a new "fat SMP" Exadata configuration, the X2-8, which appears to be targeted at operational database consolidation rather than data warehousing.

Oracle claims only 20 percent improvement in disk I/O performance, the limiting factor in scan performance. The latest generation includes no upgrade to the Smart Flash Storage system, so like previous Exadata systems, the flash device throughput, although nominally higher than disk, is limited by the rate at which the processors can ingest data to be approximately the same rate as disk throughput, and thus offers no additional throughput advantages.

As a result, the data warehouse performance improvements Oracle claims with Exadata X2-2 come almost entirely from the use of Intel Westmere processors, and demonstrates minimal added value from the database software itself. With Exadata X2, Oracle continues to throw even more hardware to address limitations of their database software. As the next section

shows, the additional hardware does not solve the fundamental problem Oracle has with data intensive analytic workloads.

Exadata for Data Warehousing – A Critique

With Exadata, Oracle claims to deliver superior performance for data warehouse workloads at a lower price, lower even than inexpensive data warehouse appliances from Teradata Corporation or Netezza. According to Oracle, this is possible because Exadata combines intelligent data filtering and fast, scalable data movement with the sophisticated Oracle database. The combination allows Oracle to deliver "the world's fastest database machine," according to Ellison.

However, a closer look at the Exadata architecture leads to a very different conclusion. Our analysis indicates that:

- > Exadata does not put query intelligence closer to the storage than competitive products such as the Teradata® Database.
- > Exadata is not a shared nothing system like the Teradata Database so it continues to wrestle with scalability issues caused by resource contention in its shared resource environment.
- > Exadata does not enable high concurrency due to its shared resource architecture, a problem the addition of flash storage does little to help.
- > Exadata does not support active data warehousing because its key query performance optimizations are ineffective when forced to handle concurrent updates.

Exadata is Still Oracle

- > Exadata does not provide superior query performance because it only addresses a small part of the query performance issues Oracle faces, and that part shrinks as queries grow more complex.
- > Exadata is complex because it adds multiple layers of query processing requiring significant hardware resources to attempt to work around its architectural limitations for data-intensive decision support workloads.
- > Exadata is expensive to purchase and operate.

The bottom line? The Oracle Exadata Database Machine is simply throwing hardware – lots of hardware – at what is fundamentally a software problem.

Our analysis shows that the Teradata Data Warehouse Appliance, which addresses the same data warehouse appliance market segment as Exadata, outperforms the Oracle Exadata Database Machine with less hardware at a lower price. Further, our analysis shows that Exadata cannot achieve the active data warehouse capabilities of the Teradata Active Enterprise Data Warehouse to enable enterprise data integration, fully accessible by thousands of concurrent users performing complex analytics, standard reports, ad-hoc queries, and continuous updates.

The next sections examine Exadata’s architecture to illustrate its capabilities and limitations.

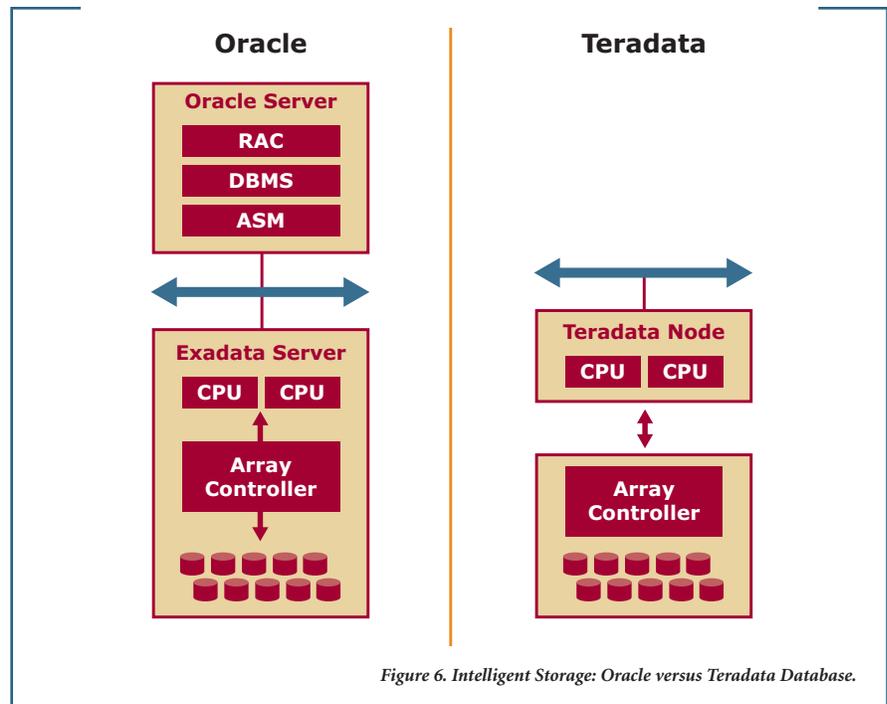


Figure 6. Intelligent Storage: Oracle versus Teradata Database.

Exadata is NOT Intelligent Storage

Under the covers, an Exadata cell is a standard Intel-based server, just like a Teradata node, and accesses data in the same way (See Figure 6.). Both read database data from a disk array into server memory to process it. In Exadata, the disks and array controller are contained within the server, while for the Teradata platform, the disk subsystem is in a separate enclosure – a simple packaging distinction. In fact, because a Teradata node has more storage network connections, it can deliver data to the processors more than two times faster than Exadata does.

Unlike Teradata Database, which performs all query processing within the nodes, Exadata cells only perform initial column

projections from the select list, and row restrictions using the WHERE clause predicates. Exadata can also perform some fact table row restrictions based on joined dimensions (for star schema joins). The Exadata cells transmit the resulting row sets to the Oracle database server, where all other query operations are executed, just as they have always been. This includes operations such as aggregation, sorting, analytic operations, data transformations, updates, temporary table creation and processing, locking and concurrency control, as well as row and column restrictions too complex for “SmartScan” processing in Exadata. Even simple reporting queries usually require other query operations such as aggregation, and all of these are performed in the database server, not the Exadata server.

Exadata is Still Oracle

In an Exadata environment, the storage path between disk and database server may well be shorter than in a classic large-scale Oracle SAN environment with multiple layers of storage switches. On the other hand, while Exadata offloads I/O operations and initial filtering from the database server, compared to Teradata Database, it introduces latency that lengthens query response time even for the simplest queries, because it requires data to be transmitted between the Exadata cell and the Oracle database server during query execution.

To put it simply, Teradata Database puts all query intelligence as close to the data as Exadata's initial data filtering operations. In addition, those query operations performed in the Oracle database server are now further away from the data than the comparable operations in Teradata Database.

Exadata is NOT Shared Nothing

Exadata improves parallel I/O performance and speeds up data retrieval compared to previous Oracle versions, but Exadata does not magically transform Oracle from a shared disk into a shared-nothing architecture – one that minimizes the number of shared system resources to reduce time spent waiting for other processes to finish using them. While superficially Exadata cells run independently of each other – just as disk arrays never interact with one another – every Oracle database process still must have access to all database data. Remember, to the Oracle database, Exadata is simply another storage device.

Distribution of data on Exadata storage is managed by Oracle's Automatic Storage Manager (ASM). By default, ASM stripes each Oracle data partition across all available disks on every Exadata cell (See Figure 7.). This produces thin stripes on all disks for every partition. Oracle calls this allocation policy *Stripe and Mirror Everywhere*, or SAME, and ASM automatically implements the SAME data allocation policy. So in fact, not only can each query process read from all disks, under SAME data allocation, all query processes will read from all disks.

Oracle advocates the SAME allocation policy for data warehousing because it believes that in its shared disk environment this policy optimizes access performance across diverse access patterns to different tables. While it's possible to control data allocation manually, as the number of tables grows, the complexity of specifying data placement manually becomes quickly

unmanageable. For these reasons, Oracle users are likely to rely on ASM's automated data allocation strategy.

By contrast, Teradata Database assigns every data partition to a distinct set of disks, and each data partition is owned by a separate database process. At query execution time, each Teradata process, called an Access Module Processor (AMP), reads data from its own disks into its own memory, without contending with other AMPs for resources.

In Oracle, as a consequence of SAME policy, every parallel slave or worker running part of a parallel query will request data from all Exadata cells. This means that even within a single parallel query, the individual query worker processes may all request data from the same disks concurrently. This produces potential contention on each disk for disk head location and I/O bandwidth. So even within a single parallel query, I/O resource contention is likely.

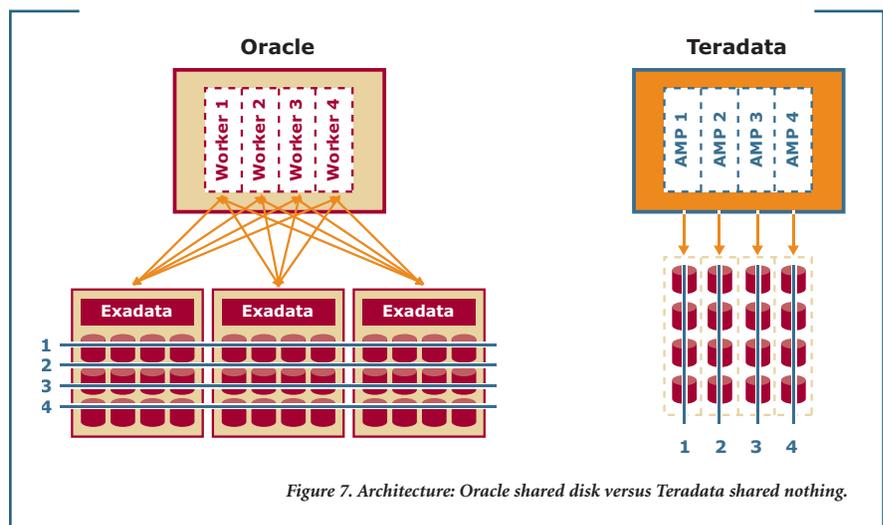


Figure 7. Architecture: Oracle shared disk versus Teradata shared nothing.

Exadata is Still Oracle

So while Exadata certainly increases the I/O parallelism of Oracle, its shared disk architectural foundation remains unchanged. Exadata does nothing to reduce or eliminate the structural contention for shared resources that fundamentally limits the scalability – of data, users, and workload – of Oracle data warehouses.

Exadata does NOT Enable High Concurrency

Increasing the number of users accessing the data warehouse requires the ability to handle an expanding volume of concurrent queries and to service the rapidly growing I/O demand this entails. The enterprise-class SAS disks used by Exadata, rotating at 15K RPM, are capable of delivering data to requestors at about 120 MBps. To maximize I/O throughput, Exadata reads data off disk in large chunks. Exadata defaults to 4MB data blocks. At a 4MB block size, 30 concurrent I/Os saturate a drive, even without allowances for seek time.

Assuming an optimal data allocation using the SAME policy described above, it takes

very few concurrent parallel queries to fully consume the data bandwidth available in an Exadata system. Consider four concurrent queries running eight-way parallel. The eight parallel slave processes of the first query will read concurrently from all the Exadata disks. When the eight parallel slaves belonging to the second query read data, they will also access all the same disks. In the best case, only three of them can begin their I/O operations before some thread encounters an I/O Wait condition (See Figure 8.).

Additional I/O requests from the parallel query slaves of other concurrent queries will be forced to queue, waiting for the I/O requests from queries 1-3 to complete. While these I/O requests are queued, forward progress on the requesting queries is stalled, and processors for both the Exadata servers and the database servers may sit idle for longer periods, waiting for I/O. Higher query concurrency exacerbates the problem, yielding longer and longer I/O queues and less efficient processor utiliza-

tion, affecting both throughput and query response time. In a large enterprise data warehouse environment, more than 1,000 concurrent queries may be active. Despite the large number of processors in both the Exadata cells and the database servers, an Oracle Exadata system will be quickly overwhelmed by workload of such magnitude.

Caching of database pages only provides modest mitigation of the concurrent I/O bottleneck intrinsic to Exadata systems for two principal reasons. First, the memory allocated to buffer cache is such a small fraction of the storage volume, typically 1/5th of one percent. A basic sales reporting application, for example, needs to service numerous concurrent requests from managers of different stores, for distinct sales data that would generate rapid buffer flushes. Second, in a large-scale enterprise data warehouse covering a number of subject areas, highly diverse queries yielding unpredictable access patterns predominate, limiting the benefits of caching. Even with caching, the Exadata

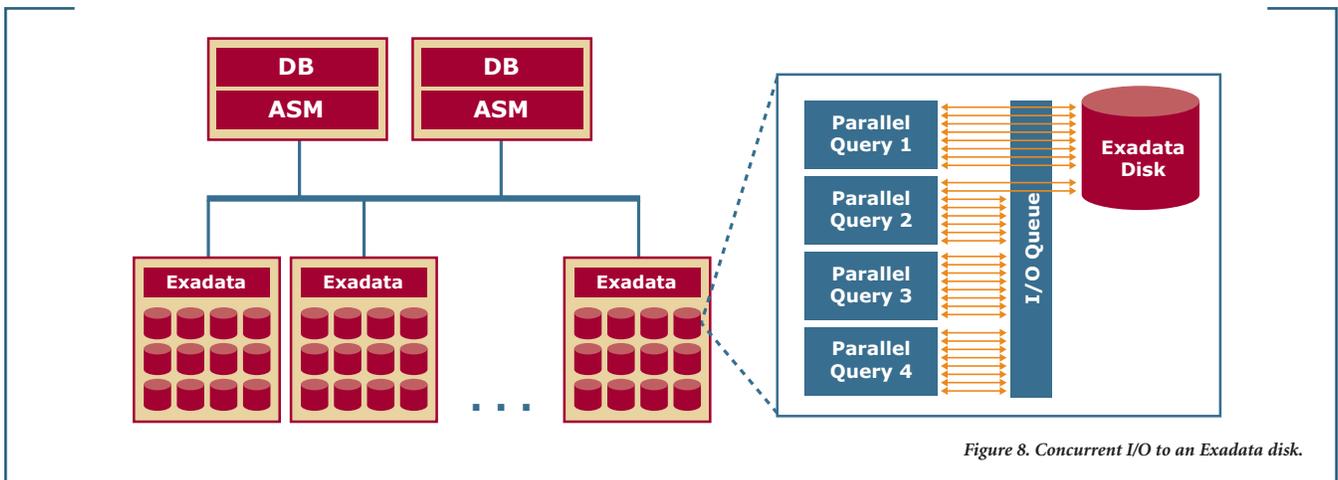


Figure 8. Concurrent I/O to an Exadata disk.

Exadata is Still Oracle

architecture suffers from a concurrent I/O bottleneck that limits its ability to adequately support even modest concurrent query levels.

Oracle Flash Cache and Concurrency

Oracle Flash Cache, added in Exadata V2, partially mitigates, but does not solve Oracle's concurrency problem for data warehouse queries. Flash Cache provides a second level cache to hold data flushed out of Oracle's main-memory-resident Buffer Cache. While the Flash Cache is larger than the Buffer Cache, it is still smaller than the active data in a busy enterprise data warehouse. While data can be pinned in Flash Cache, diverse access patterns in a large-scale data warehouse lower the cache hit ratio. In addition, data accessed via a table scan, the preferred access method in Exadata, occupies a separate memory area outside the buffer cache, so as not to flush the cache too rapidly. Thus, it's not available for storing in Flash Cache. Also, updates to cached data require that the

Flash Cache be refreshed from disk. This undercuts any concurrency benefits in the continuous update environment common for active data warehouses. In other words, Oracle Flash Cache benefits OLTP applications, for which it was designed, but has limited benefit for data warehousing.

Exadata does NOT Support Active Data Warehousing

Enterprise data warehouses that support operational business intelligence applications demand intraday updates. The latency between events and the ability of the business to respond intelligently to them is continually shrinking. Even single application data marts and special purpose analytic data warehouses are increasingly requiring online updates to respond rapidly to recent events. Exadata's performance benefits are seriously compromised in an online or active update environment.

To ensure data integrity, databases must give each query a consistent view of the

data it needs. Typically this is accomplished by ensuring that a query only sees the state or contents of the database at the time it begins execution. Oracle uses a multi-version concurrency control (MVCC) mechanism to manage a logically consistent view of data for every query.

Oracle maintains multiple versions of updated data blocks, each representing its data contents at a different point as updates to the block occur. Oracle distinguishes different versions of a data block using a System Change Number (SCN), which is an Oracle system-wide value that increments with every transaction. An SCN represents the state of the database at a particular point. Data blocks contain the SCN of the transaction that last updated them. Queries are assured of obtaining a consistent view of the data by only accessing versions of data blocks with SCNs equal to or less than the current SCN at the time the query begins (See Figure 9.).

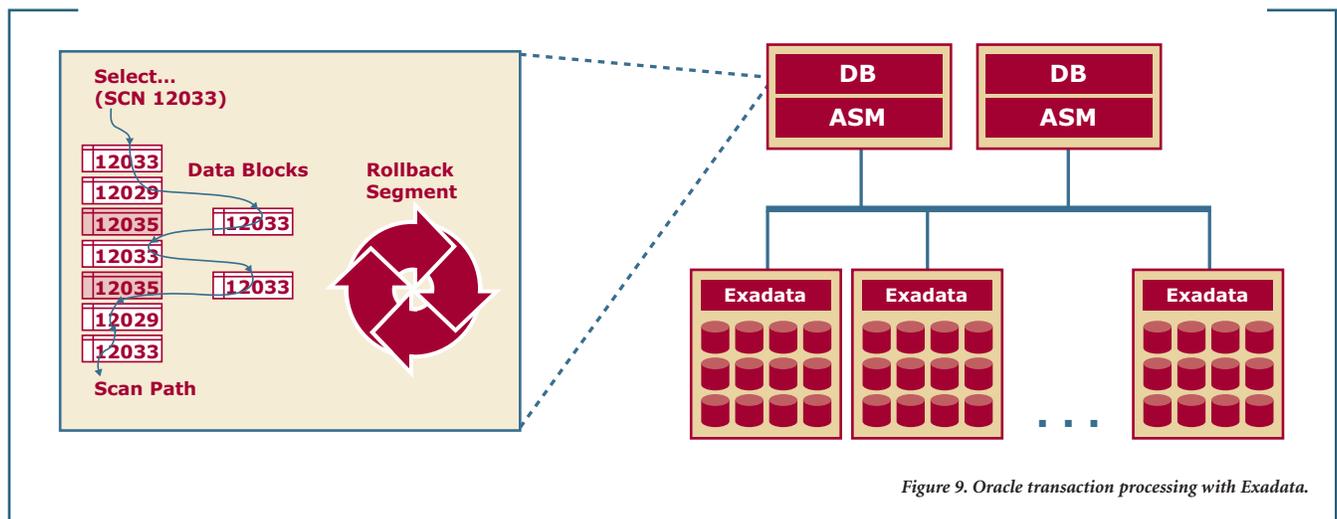


Figure 9. Oracle transaction processing with Exadata.

Exadata is Still Oracle

The software algorithms that maintain this logically consistent view of data for every query run in the Oracle database layer, not in the Exadata layer. Since it relies on shared database structures, such as Oracle rollback segments, the database buffer cache, and global lock data structures, for access to previous versions of data blocks (with lower SCNs), this logic can only be run from a software layer that has a global view of the Oracle instance.

Exadata filtering occurs prior to SCN checking. For tables or partitions being actively updated, Exadata simply does not know whether the version of each data block it reads is the correct version for the query requesting it, so the Exadata run-time system will disable SmartScan filtering in these cases. Exadata simply passes unfiltered blocks to the database server, which performs the version check to determine whether or not it has the right version of the block before applying column projections and WHERE clause restrictions. In some cases, Oracle may also need to do additional I/O to retrieve the correct version of the data block from the Oracle Rollback Segment. In these increasingly common active update cases, the benefits of Exadata SmartScan are eliminated.

Improving I/O parallelism may yield large improvement in I/O performance. However, filtering, which can easily offer a 100 times or greater reduction in data volume, has a much larger effect on query performance. While Exadata still performs

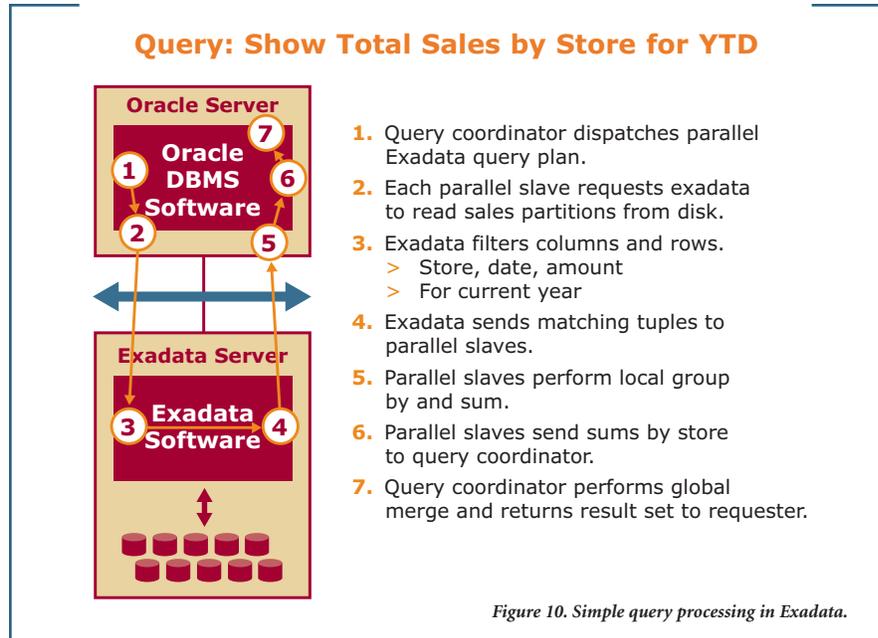


Figure 10. Simple query processing in Exadata.

parallel I/O for the query, the largest benefit provided by Exadata, early filtering of columns and rows meeting the query specification, which may drastically reduce query data volume, is not useful for tables or partitions being actively updated. For active data warehouse or operational business intelligence environments, these are precisely the tables and partitions you want to benefit fully from all performance enhancement that Exadata provides.

Exadata does NOT Provide Superior Query Performance

Query performance involves all SQL operations, from scanning data, to sorting and grouping, to complex OLAP operators. Any query, even a relatively simple one, involves several operations unaffected by Exadata. For example, a single table

reporting query will perform column and row filtering in Exadata, but grouping, local aggregation, and global merging of aggregated results all happens in the Oracle database layer (See Figure 10.). The overall performance of this query depends on the efficiency of both the operations executed by Exadata and the operations performed in the Oracle database layer.

As queries become more complex, Exadata continues to perform the needed filtering, but even more of the query operations, and a greater share of the resources consumed, take place in the Oracle database layer. A more complex query, for example, may need to perform one or more non-partition-wise joins and complex OLAP operations, all of which will happen in the database layer

Exadata is Still Oracle

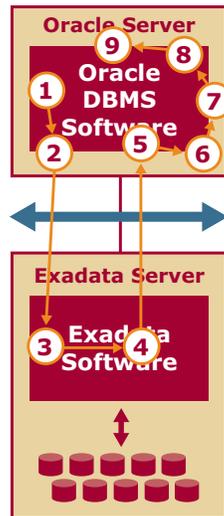
(See Figure 11.). In general, as the complexity of queries grows, the work performed by Exadata shrinks as a proportion of the total work of the query.

Thus, Exadata has the biggest performance impact on the simplest data warehouses, such as single application data marts with few tables arranged in a simple star schema. Conversely, Exadata exhibits a significantly diminishing benefit with more complex data warehouse environments, those with complex multi-subject schemas, running complex queries and ad-hoc data exploration in a mixed workload environment that includes online updating and operational business intelligence applications. In other words, as query and data complexity grow, the benefits of Exadata diminish.

The standard performance difficulties that Oracle data warehouses typically encounter as the scale of data, users, and workload increases, none of which is addressed by Exadata, grow to dominate the performance profile of the data warehouse (See Figure 12.). The inconsistent performance patterns that plague Oracle data warehouses remain. For large-scale EDW workloads, the work performed post-scan, and therefore out of Exadata's scope, is likely to overwhelm any performance contribution provided by Exadata.

In summary, Exadata benefits appear to be limited to a small fraction of the work a medium- to large-scale data warehouse needs to be able to perform both efficiently and scalably. As a whole, data warehousing spans a wide variety of query characteristics that can be categorized into five stages of data warehouse maturity.

Query: Show Sales by Store by Customer Age, Sex, Income



1. Query coordinator dispatches parallel Exadata query plan.
2. Each parallel slave requests exadata to read sales and customer data from disk.
3. Exadata filters columns and rows.
 - > Store, date, amount for current year
 - > Customer demographics
4. Exadata sends matching tuples to parallel slaves.
5. Parallel slaves redistribute customer data for join.
6. Parallel slaves perform join.
7. Parallel slaves perform local group by and sum.
8. Parallel slaves send sums by store to query coordinator.
9. Query coordinator performs global merge and returns result set to requester.

Figure 11. More complex query processing in Exadata.

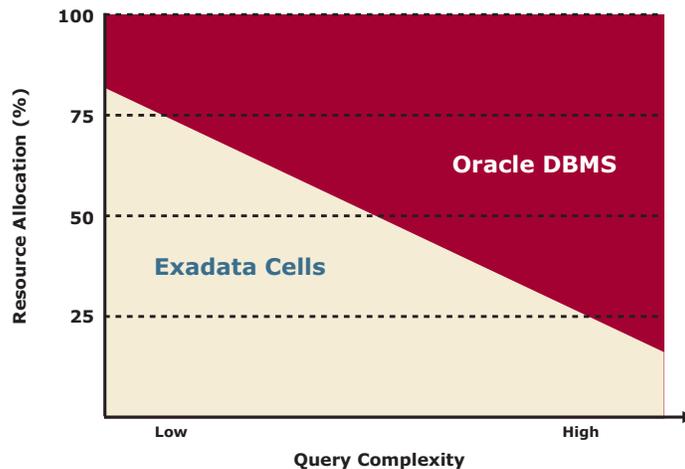


Figure 12. Exadata impact on query performance.

As data warehousing practice matures over time, higher stages, involving more complex business questions and more

demanding technical requirements, play an ever larger role in the mix of work performed on the data warehouse.

Exadata is Still Oracle

Compared to pre-Exadata Oracle technology, Exadata offers the largest impact on Oracle performance for entry-level data marts with simple schemas and few users. Exadata I/O improvements are seriously compromised at higher levels of query concurrency, however, and virtually non-existent at concurrency levels common in large-scale EDWs. The early data reduction provided by Exadata-level filtering of rows and columns is turned off in the face of concurrent updates to the tables or partitions being queried, a basic requirement of operational intelligence applications, such as customer service or impact analysis of marketing promotions. Users should expect that for most of the work performed in mature data warehouses, Exadata will show no better than a neutral performance impact (See Figure 13.).

Unlike Exadata, Teradata's unified query execution supports very high levels of parallelism for all query operations; its shared nothing model has allowed Teradata to demonstrate excellent scalability characteristics across multiple dimensions – data, users, and workload – in the same database. And its query execution architecture is not constrained in the face of active updates. These characteristics are among many that have enabled Teradata to claim numerous customers who are achieving excellent performance results at every stage of data warehouse maturity.

Exadata is Complex

Exadata achieves its limited benefits at the high cost of increased architectural complexity. A single cabinet Oracle Exadata Database Machine X2-2 contains an eight-way RAC system running eight

independent instances of Oracle 11gR2 that must all share data. All of the challenges of managing workloads and shared data access in a RAC data warehousing environment remain. There are very few Oracle customers using eight-way RAC environments for data warehousing. And with two cabinets of Oracle's Database Machine X2-2, that jumps to a 16-way RAC system.

In addition, there is nothing in Exadata to address the complex challenges of using parallelism in an Oracle environment. Now we add to that the challenges of balancing workload and resource utilization across a separate architectural layer. Plus new analysis and tuning will be needed to determine appropriate I/O block sizes, I/O workload types, and other factors, to use the Exadata storage layer effectively and

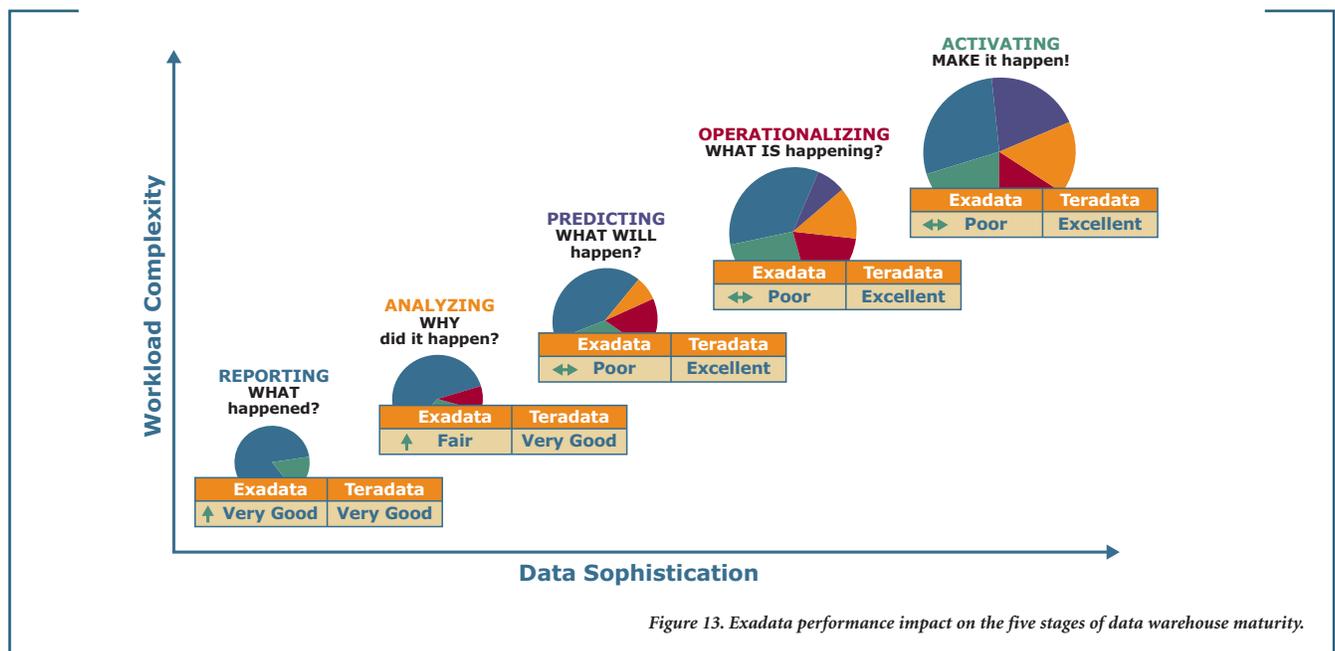


Figure 13. Exadata performance impact on the five stages of data warehouse maturity.

Exadata is Still Oracle

efficiently. Clearly, Oracle does not gain the ease of management, virtualization of computing resources, and linear scalability that has been, and continues to be, the hallmark of Teradata systems. In fact, they have created a system that is even more difficult for DBAs to manage.

Exadata's design involves a two-layer query execution engine that requires substantial processing power. To achieve modest improvements in peak scan performance and user data capacity, the Oracle Exadata Database Machine requires far more computing resources than Teradata Database requires.

The Teradata Data Warehouse Appliance, on the other hand, manages significantly higher levels of parallelism and balanced resource utilization that deliver much more performance per server than the comparable Exadata appliance product.

Requiring far more hardware resources to support a two-tiered query engine, Exadata is inherently more complex. Balancing work across the two layers adds new challenges to the already extensive DBA tuning workload. It has more components to maintain. More components mean higher failure rates. Both imply higher DBA labor requirements. The large hardware requirement of the Exadata implementation is also more expensive to acquire and has much higher power and cooling requirements, so it's also more expensive to operate.

Such large disparity in hardware resources suggests that Oracle is simply throwing hardware at what is fundamentally a database software problem. At its core, Oracle is a shared disk system and has always struggled to achieve consistently high levels of performance for data intensive analytic databases. Exadata offers some improvement in I/O performance compared to earlier Oracle versions, but does not fundamentally alter this architectural constraint. As a result, despite a large hardware commitment, Exadata is likely to suffer similar performance limitations in large-scale data warehouse applications to pre-Exadata versions of Oracle.

Conclusion

Exadata technology and the Oracle Exadata Database Machine address a serious problem that Oracle faces with data warehouse processing. Data warehouses are much larger and far more data intensive than the standard transaction processing applications for which Oracle has a well-deserved reputation. Data warehouse systems must be able to service requests for large data volumes efficiently and be able to scale effortlessly to accommodate expanding workloads – tasks that have historically exposed the limits of Oracle's shared disk architecture. With Exadata, Oracle combines higher bandwidth, more scalable networks with smarter software that allows early filtering to reduce query data volume. The Oracle Exadata Database Machine packages this into a pre-configured system to avoid the

poor user configurations that commonly lead to Oracle performance problems.

As a result, Exadata-based systems offer higher data throughput than most previous Oracle versions achieved. While in this sense Exadata is a better Oracle, it is far from the groundbreaking innovation that Oracle claims. Exadata does not tackle Oracle's underlying performance and scalability problems with large-scale data warehousing that stem from its shared disk architectural foundation. Analysis shows that resource contention continues to limit Exadata I/O performance despite its increased I/O bandwidth and parallelism. Many query operations are untouched by Exadata, are subject to the same resource sharing constraints, and so continue to exhibit the same performance characteristics as before Exadata. As data warehouse systems grow in scale and complexity, the part of the performance problem addressed by Exadata shrinks in size and importance.

Whatever performance improvements that Exadata achieves come at the cost of increased software complexity due to the introduction of a two-level query architecture and significant additional hardware. Exadata throws a lot of hardware at what is essentially a database software limitation handling data intensive analytic workloads. The resulting configuration is more costly to acquire, more expensive to operate. In the final analysis, Exadata delivers far less than promised.

The Best Decision Possible is a trademark, and Teradata and the Teradata logo are registered trademarks of Teradata Corporation and/or its affiliates in the U.S. or worldwide. Teradata continually improves products as new technologies and components become available. Teradata, therefore, reserves the right to change specifications without prior notice. All features, functions, and operations described herein may not be marketed in all parts of the world. Consult your Teradata representative or Teradata.com for more information.

Copyright © 2012 by Teradata Corporation All Rights Reserved. Produced in U.S.A.