Why Exadata Is the Best Platform for Oracle Database In-Memory

ORACLE WHITE PAPER | MAY 2015

Introduction

Oracle Database In-Memory transparently accelerates analytic queries by orders of magnitude, enabling real-time business decisions. Oracle Database In-Memory uses a "dual-format" architecture that enables data to be maintained in both row format and a pure in-memory columnar format. This columnar data can be scanned very fast by taking advantage of SIMD vector processing[1] and In-Memory Storage Indexes[2]. With Oracle Database In-Memory it is possible to scan billions of rows per processor core per second purely in-memory. It is now feasible for businesses to run real-time analytics on their critical business data without impacting the performance of their existing systems.

With the benefits of Oracle Database In-Memory, does it matter what platform you run your database on? Yes, the Oracle Exadata Database Machine (Exadata) has been the preferred platform for running Oracle Database since its release in 2008, and it provides distinct advantages for running Oracle Database In-Memory as well. The following are the top 10 most important advantages that Exadata brings to Oracle Database In-Memory:

- Exadata efficiently scales Oracle Database In-Memory
- In-Memory fault tolerance
- Exceed DRAM limits and transparently scale across Memory, Flash and Disk
- In-Memory Aggregation optimization can be offloaded to Exadata storage cells
- Exadata provides high storage bandwidth to quickly populate the Oracle Database In-Memory column store
- Parallel execution NUMA support
- Exadata is Oracle's Database In-Memory development platform
- Elastic configurations enable custom configurations so that you only pay for what you need
- Use of Oracle Trusted Partitions can reduce software licensing costs
- Exadata is a database consolidation platform and Oracle Database In-Memory further enables consolidation

In this paper we will examine each of these points and explain in detail why Exadata is the best platform for running Oracle Database In-Memory.

[1] Single Instruction processing Multiple Data values allow evaluating an array of column values together in a single CPU instruction.

[2] In-Memory Storage Indexes allow data pruning to occur based on filter predicates supplied in a SQL statement.

Exadata efficiently scales Oracle Database In-Memory

Exadata uses a scale-out architecture for both database servers and storage servers. The Exadata configuration carefully balances CPU, I/O and network throughput to avoid bottlenecks. As an Exadata system grows, database CPUs, storage, and networking are added in a balanced fashion ensuring scalability without bottlenecks. This scale-out architecture can accommodate any size workload and allows seamless expansion from small to extremely large configurations while avoiding performance bottlenecks and single points of failure. This is very important for Oracle Database In-Memory as it ensures the In-Memory column store (IM column store) can scale out across multiple nodes due to increased parallelism and very low latency interconnect messaging.

In a Real Application Clusters (RAC) environment, objects with the INMEMORY attribute specified can be distributed across the cluster by rowid range, by partition or by subpartition. By default, Oracle decides the best way to distribute the object given the type of partitioning used, if any. Alternatively this can be overridden by using the `DISTRIBUTE` sub-clause.

In-Memory fault tolerance

Given the shared nothing architecture of the IM column store in a RAC environment, some applications may require a fault-tolerant option. On Exadata it is possible to mirror the data populated into the IM column store by specifying the `DUPLICATE` subclause of the `INMEMORY` attribute. This means that each In-Memory Compression Unit (IMCU) populated into the IM column store will have a mirrored copy placed on one of the other nodes in the RAC cluster. Mirroring the IMCUs provides in-memory fault tolerance as it ensures data is still accessible via the IM column store even if a node goes down. It also improves performance, as queries can access both the primary and the backup copy of the IMCU at any time.

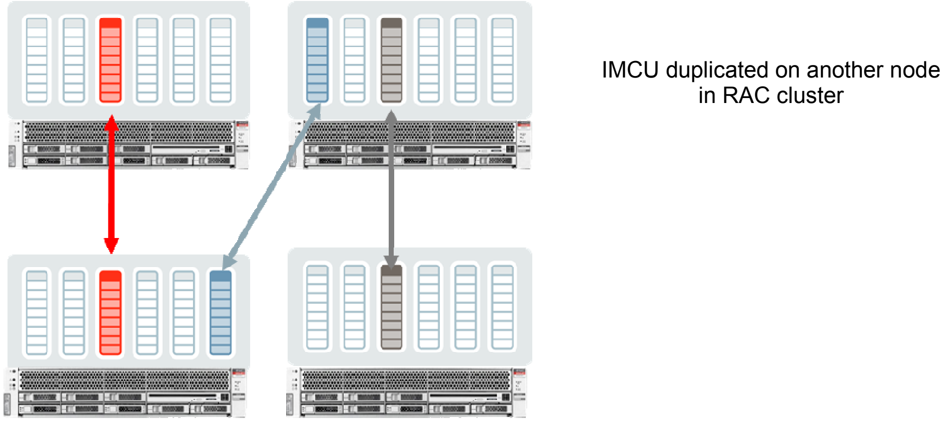

Figure 1. Objects in the IM column store on an Exadata Database Machine can be mirrored to improve fault tolerance

Should a RAC node go down and remain down for some time, the only impact will be the re-mirroring of the primary IMCUs located on that node. Only if a second node were to go down and remain down for some time would the data have to be redistributed.

If additional fault tolerance is desired, it is possible to populate an object into the IM column store on each node in the cluster by specifying the `DUPLICATE ALL` sub-clause of the `INMEMORY` attribute. This will provide the highest level of redundancy and provide linear scalability, as queries will be able to execute completely within a single node.

The DUPLICATE ALL option may also be useful to co-locate joins between large distributed fact tables and smaller dimension tables. By specifying the DUPLICATE ALL option on the smaller dimension tables a full copy of these tables will be populated into the IM column store on each node. In the example in figure 2, when a query joins a partition of the sales table to one or more of the dimension tables all of the data required for the join will be in the local node, avoiding having to fetch data across nodes to complete the join.

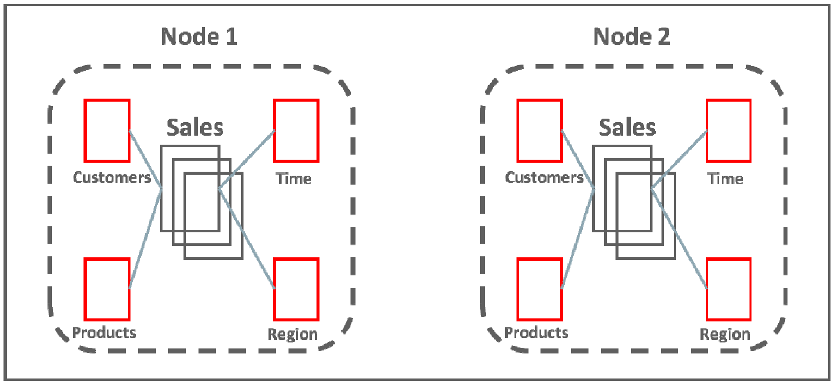

Figure 2. Distributed fact table with duplicated dimension tables

Exceed DRAM limits and transparently scale across Memory, Flash and Disk

With Exadata, your application can make use of all storage tiers (memory, flash, & disk) without having to be aware of where the data resides or suffer suboptimal performance when not all of the data resides in-memory in the IM column store. On Exadata data can reside in the IM column store, in the database buffer cache, in Flash storage or on disk storage and your application never needs to be aware of data location because Oracle Database can seamlessly access that data.

When data resides on Exadata storage servers, Exadata’s Smart Flash Cache feature can dramatically accelerate Oracle Database processing by speeding I/O operations. The Flash provides intelligent caching of database objects to avoid physical disk I/O. Exadata storage provides an advanced compression technology, Hybrid Columnar Compression (HCC), that typically provides 10x level of data compression and boosts the effective data transfer by an order of magnitude.

This means that all data access, and not just data that has been populated into the IM column store, will be as efficient as possible.

Exadata also includes Smart Scan, a unique technology that offloads data-intensive SQL operations into the Oracle Exadata storage servers. This is similar to and complements Oracle Database In-Memory processing by enabling seamless and efficient access to data on any storage tier. By pushing SQL processing to the Exadata storage servers when data is not in the IM column store, data filtering and processing occurs immediately and in parallel across all storage servers as data is read from disk. Exadata Smart Scan reduces database server CPU consumption and greatly reduces the amount of data moved between storage and database servers. This enables scaling and efficient SQL processing across all storage tiers whether data resides in the IM column store, on flash storage or on disk storage.

In-Memory Aggregation optimization can be offloaded to Exadata storage cells

With the introduction of Oracle Database In-Memory comes the new In-Memory Aggregation optimization, or Vector Group By feature. In-Memory Aggregation (IMA) provides new SQL execution operations that accelerate the performance of a wide range of analytic queries against star and similar schemas. These include the KEY VECTOR USE and VECTOR GROUP BY operations which enable the use of a vector transformation plan that minimizes the amount of data that must flow through the execution plan. This minimizes the amount of CPU used as compared to alternative plans.

The result of this is that IMA can transform joins to KEY VECTOR filters on the fact table and aggregate data in a single pass while lowering CPU use. This is extremely fast when the entire table resides in the IM column store, but what if the entire table doesn't fit into the IM column store? On Exadata when tables are accessed and they have not been populated in the IM column store, IMA is enhanced by the ability to offload the KEY VECTOR USE operation to Exadata storage servers. This might occur when the table is partitioned and only the most recent partitions are loaded into the IM column store and the other partitions are on disk. The offload capability distributes key vector processing across Exadata storage servers and minimizes the volume of data that must be returned to the database nodes.

Exadata provides high storage bandwidth to quickly populate the Oracle Database In-Memory column store

When data is initially populated into the IM column store it is read directly from disk in its row format, pivoted 90 degrees to create columns and then compressed. The faster you can read the data, the faster you can complete the population process. Exadata storage offers outstanding IO performance ensuring the data population process is not I/O bound.

The population process is conducted by a set of background worker processes. These worker processes can operate in parallel to populate the IM column store as fast as data can be read off of disk and CPUs can process that data. This is where the high I/O performance and CPU resources of Exadata come into play to make the population of the IM column store as fast as possible. The number of background worker processes can also be controlled to take further advantage of Exadata's scalability.

Oracle Database In-Memory will also repopulate IMCUs when the number of stale entries in an IMCU reaches a staleness threshold. Again, with Exadata's high I/O performance this can occur in the background with no noticeable effect on application performance.

Exadata provides a very fast Interconnect with special protocols to speed up Oracle Database In-Memory scale-out

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

Exadata this high-speed transfer and large bandwidth for messaging between IM column stores keeps the IM column stores transactionally consistent and in sync with each other. This enhances scale out for distributed objects as well as objects that have been duplicated.

Parallel execution NUMA support

Today's multi-socket processors employ memory architectures that allow a process on one socket to access memory that is connected to another socket. This is referred to as NUMA, or Non-uniform Memory Access (NUMA). In a NUMA system, CPU & Memory resources are divided into multiple logical nodes, typically based on CPU socket. On the Exadata x4-8 there are 8 CPU sockets and therefore 8 NUMA nodes[3]. Each NUMA node can access both local and remote memory resources but the local memory access will be a lot faster (2-3X faster).

Oracle Database In-Memory is NUMA aware and can take advantage of NUMA on Exadata. When the IM column store is allocated on a NUMA system it is divided into stripes with a different stripe being placed on each NUMA node. During the population of the IM column store we assign a NUMA node id to each IMCU in the IM column store. When we read data from the IM column store we co-locate the reader process on the NUMA node where the IMCU being read is located. This allows us to greatly reduce the number of remote reads and ensures the fastest access to all data populated in-memory.

Exadata is Oracle's Database In-Memory development platform

Exadata is the development platform for Oracle Database In-Memory. Thus, IM issues are discovered and fixed on Exadata first. Exadata is also the primary platform for Oracle Database testing, HA best practices validation, integration and support. The same reasons it is the best platform for Oracle Database apply to Oracle Database In-Memory.

When Exadata Bundle Patches are released they include fixes for both Exadata and Oracle Database In-Memory. This provides more frequent updates than normal Critical Patch Updates (CPU) or Patch Set Updates (PSU), and they are created specifically for Exadata and Oracle Database In-Memory and have been tested and certified to work together.

Elastic configurations enable custom configurations so that you only pay for what you need

Oracle recognizes a practice in the industry to pay for server usage based on the number of CPUs that are actually turned on – the "Capacity on Demand (CoD)," or "Pay as You Grow" models. With CoD, Oracle allows customers to license software for only the number of cores that are activated when the server is installed. Exadata elastic configurations, introduced in Exadata X5, enable custom combinations of database and storage servers that are tailored to specific workloads. With Oracle Database In-Memory if you need more database servers for additional CPU and memory but not additional storage capacity you can create that configuration and only purchase the Exadata servers needed for your workload.

[3] NUMA Node: a block of memory and the CPUs, I/O, etc. physically on the same bus as the memory

Use of Oracle Trusted Partitions can reduce software licensing costs

On Exadata the use of Oracle Trusted Partitions is allowed. This means that you can run Oracle VM Server (OVM) as a means to limit the number of Oracle software licenses required. When Exadata VMs run Oracle Database In-Memory you only pay for the Oracle Database In-Memory option for the vCPUs that are actually used within the virtual machine.

Exadata is a database consolidation platform and Oracle Database In-Memory further enables consolidation opportunities

Database consolidation is one of the major strategies that organizations use to achieve greater efficiencies in their operations. Increasing the utilization of hardware resources while reducing administrative costs are primary goals of consolidation projects. Exadata is optimized for Oracle Data Warehouse and OLTP database workloads, and its balanced database server and storage grid infrastructure make it an ideal platform for database consolidation. Exadata is a modern architecture featuring scale-out industry-standard database servers, scale-out intelligent storage servers, and an extremely high speed InfiniBand internal fabric that connects all servers and storage. In many ways Oracle Database In-Memory “completes” Exadata by applying in-memory performance techniques that are similar to those that are used by Exadata on disk and flash. Consolidation on Exadata allows customers to simultaneously optimize performance and cost by placing specific objects in-memory and then use the extreme performance flash and high capacity disk to increase consolidation capacity for all other data. The result is a solution that gives the speed of DRAM, the IOPs of flash, and the cost effectiveness of disk.

Conclusion

Exadata is the best platform for running Oracle Database and the Database In-Memory option. Database In-Memory takes full advantage of Exadata's unique hardware features, enabling better performance than any other hardware platform. These features include a very fast interconnect enabling IM fault tolerance and scale-out, high storage bandwidth and IOPs enabling fast IM column store population, seamless access to all storage tiers and the running of mixed workload environments. Exadata is also an excellent consolidation platform with support for Oracle Trusted Partitions to limit the number of Oracle software licenses to just those that are needed and elastic configurations so that you only configure the hardware you need. All of this along with Oracle's commitment to ensuring that all hardware and software components are pre-configured, pre-tuned and pre-tested to work seamlessly together for the best possible performance and reliability in the industry make Exadata the best platform for running Oracle Database and the Oracle Database In-Memory option.

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

Why Exadata Is the Best Platform for
Oracle Database In-Memory
May 2015