PETER BAER GALVIN

# Pete's all things Sun: the Exadata V2 architecture and why it matters

Peter Baer Galvin is the chief technologist for Corporate Technologies, a premier systems integrator and VAR (www.cptech.com). Before that, Peter was the systems manager for Brown University's Computer Science Department. He has written articles and columns for many publications and is co-author of the *Operating Systems Concepts* and *Applied Operating Systems Concepts* textbooks. As a consultant and trainer, Peter teaches tutorials and gives talks on security and system administration worldwide. Peter blogs at http://www.galvin.info and twitters as "PeterGalvin."

*pbg@cptech.com*

**EVEN BEFORE ORACLE CONSUMMATED** its purchase of Sun, the two companies announced their first combined product, the Exadata V2. The product was announced in October 2009 and started shipping in March of 2010. This "appliance" is designed to execute both OLTP and data warehouse operations, either individually or concurrently. It is based on Sun hardware and runs the Oracle Linux operating system, the Oracle 11GR2 database, RAC clustering, Oracle Exadata software, and, optionally, other software components.

The Exadata V2 is interesting on its own merits and provides an implementation of Sun and Oracle best practices for optimal database performance that is worth exploring. Exadata V2 becomes compelling to study when it is viewed as the first of probably many applications from the combined Oracle/Sun company.

In this column I will describe the various aspects of the Exadata V2 architecture from both points of view—database implementation best practice blueprint and first of a likely generation of Oracle/Sun appliances. There are several surprises along the way, including outstanding (claimed) performance, ease of migration, and, under some circumstances, low cost of adoption.

## Exadata Background

The first Exadata came out in September 2008. It was based on HP hardware, ran Oracle software, and lacked some of the innovations found in Exadata V2. In fact, its lower performance made it appropriate only for data warehousing (DW) uses, not OLTP, pitting it head-to-head against other DW appliances such as those from Teradata and Netezza. Oracle has not published information on how many Exadata sold, but there is a clue in their product decisions. Approximately 12 months after Exadata V1 was announced it was put out to pasture in favor of Exadata V2, and Oracle terminated all sales of Exadata V1 [1, 2].

Oracle seems to be greatly emphasizing V2, with a large announcement, advertising, and many informational events. As of Oracle's financial statements covering its third quarter (ending February 2010), Oracle expects to sell $100m worth of Exadata in its fourth quarter [3]. It appears that Oracle is very

serious about Exadata V2 and expects strong sales, but does the architecture support such exuberance?

## Exadata V2 Appliance

Fundamentally, Exadata V2 consists of several components in several "supported" configurations. It *is* an appliance in that it comes pre-built from the factory and only takes a few days for the on-site hardware and software configuration. It is also an appliance in that there are only certain configurations that are supported. Other custom configurations are possible, but a custom Exadata V2 has less support available than the standard configurations. That is, the components are supported, but a customer can no longer call the Oracle 800 number and say "my Exadata V2 is slow" and have Oracle handle the debugging and optimizing. Also in that lesser-support category are solutions for the Exadata V2 not blessed by Oracle. For instance, it is possible to use the InfiniBand interconnect infrastructure for data access or for backups, but that again makes a configuration "custom" (although it appears that Oracle might now be supporting InfiniBand use for backups).

## Hardware Architecture

There are two major hardware components to the Exadata V2. The Database Machine is the server node where the database runs, and it accesses data stored in the Storage Server. The servers are interconnected by a redundant 40Gb/sec InfiniBand network. Application access to the appliance is via the database servers' multiple, redundant, front-end 1Gb NICs.

Exadata V2 comes in "basic system," "quarter rack," "half rack," and "full rack" versions. Further, you can add up to seven full racks to any of the rack configurations to create a really, really big database server. The "basic system" is just a database server and a storage server, designed for development or QA rather than production use. A quarter rack of Exadata V2 includes two database servers and three storage servers. It can be upgraded to a half or full rack. The half-rack includes four DB servers and seven storage servers. The full rack includes eight DB servers and 14 storage servers.

Each database server is a Sun x4170 with 72GB of memory, dual 146GB SAS boot disks, hardware RAID, two four-port QDR (Quad Data Rate) InfiniBand ports, four Gigabit Ethernet ports, and one Ethernet ILOM management port. Each provides two quad core Intel Nehalem CPUs. The storage servers come in two flavors. One has SAS drives for maximum performance, while the other has SATA drives for maximum storage capacity. Both versions are based on the Sun x4275 server with two quad core Intel Nehalem CPUs, 24GB of memory, 12 drives, and four 96GB Sun Flash Accelerator F20 cards. Each also has hardware RAID, two four-port QDR InfiniBand HBAs, four Gigabit Ethernet ports, and one Ethernet ILOM management port. Each storage node has 384GB of "smart flash cache" storage capacity. The flash memory is not in the form of disk drives, but, rather, plugs into the PCI bus slots for improved throughput. The SAS nodes provide 7.2TB (twelve 600GB 15K RPM drives) of raw storage, while the SATA nodes provide 24TB (twelve 2TB 7.2K RPM drives).

A full rack of Exadata V2 therefore provides 5.3TB of Smart Flash Cache and 100TB of SAS or 336TB of SATA (or a mix of the two) raw storage. The storage is by default configured to be striped within a storage server and mirrored between storage servers (via ASM, Oracle's Automated Storage Management software), yielding 28TB SAS or 100TB SATA of usable storage. According to Oracle, a full rack of the Exadata V2 appliance provides 21GB/

sec of disk I/O, 50GB/sec of flash I/O, and one million I/Os per second aggregate. Oracle initially advertised benchmark results, but those numbers were unaudited and Oracle was forced to stop publicizing them until the audit could be completed. In the meantime we have only Oracle's promise that the full rack can perform millions of database transactions per minute and tens of millions of queries per minute.

## Software Architecture

The blueprint of how to architect an optimal database server is certainly thought-provoking. Rather than take the path of some vendors of configuring flash disks as a fast tier of storage, Oracle/Sun is continuing to forge a role for flash as a new tier of cache. By definition, a new storage tier would be faster and more expensive than the lower tier while being slower and lower-cost than the upper tier. That is exactly the case for flash memory. Such a change in storage hierarchy hasn't taken place in dozens of years, so expect a profound effect on price/performance as flash is integrated into hardware, operating systems, and applications. Within Exadata V2, the flash memory is used as an LRU cache by default (with optimizations to avoid sequential I/O from evacuating the cache). However, the software allows for tuning of the use of the cache, including pinning tables into flash (in essence treating it like fast disks) and preventing tables from being cached in flash. This approach makes a lot of sense and should be a model for other vendors (or customers) to follow.

Other aspects of the Exadata V2 architecture also align with fundamental best practices. Oracle RAC provides clustering for performance and availability. Database instances need not be RAC-clustered, but by default they are. Intel Nehalem CPUs provide great performance in small footprint servers. Rather than have a central storage unit (and central I/O bottleneck), the storage is grid-shaped to match the database server grid architecture. The database server talks to the storage server using iDB (Intelligent Database protocol) and offloads operations to the storage server using that protocol. iDB is itself based on RDSv3 (the industry-standard Reliable Datagram Sockets protocol). The interconnect between the nodes is low-latency, high-throughput InfiniBand. The Exadata V2 runs Oracle's Enterprise Linux (essentially Red Hat Enterprise Linux but supported by Oracle), which has a lot of RDP optimization built in.

Added to these standard components are some Exadata V2–only ones. Hybrid Columnar Compression (HCC) is a new compression technology. It avoids the problems of other compression methods, in essence allowing data to be compressed and still used for OLTP transactions rather than being limited to just data warehouse operations [4]. This form of compression is especially important considering that data warehouses can use much more storage than OLTP databases and that the Exadata V2 has fairly low storage capacity options. If the HCC technology is as effective as Oracle claims, in many cases compressing data 10x and sometimes even 50x, then this compression can provide effectively very large storage capacities. Further, the HCC pertains to flash memory as well as on-disk, potentially greatly increasing the flash capacity. Also included is "Exadata Storage Server Software." This new package configures, manages, and monitors the storage nodes (including flash memory use). One final Exadata V2-only feature is "Exadata Smart Scan" [5]. This optimization offloads certain operations to the storage servers, allowing them to perform operations on their stored data and return only the results, rather than returning bulk data to the database servers for

them to sift through. Operations that can be offloaded (and are by default) include table scans, join filtering, backups, and tablespace creation.

## Analysis

The blueprint provided by the Exadata V2 is certainly solid. There are a few surprises and conclusions that are worth noting as well:

- Existing Oracle licenses are transferable to Exadata (including Oracle DB, RAC, and Partitioning). That can greatly reduce the cost of an Exadata that is being used for database consolidation, for example.
- The Exadata looks to be an excellent consolidation engine. Included with the Exadata software are resource management tools that can, for example, give some databases resource priority over others. These tools also allow the use of the flash storage to be fine-tuned, pinning specific tables into flash or letting Oracle use the flash as an extended cache.
- The Exadata V2 is designed to be able to perform OLTP and data warehouse transactions concurrently. If a single system can be used both ways, consider the implications compared to stand-alone, separate data warehouse solutions. Normally data must be extracted from the OLTP system, copied to the DW system, imported there, and then processed. The extraction and copying are overhead, on both the OLTP and DW systems. And any reports or queries on the DW system are performed against "stale data," data from the time the extraction started. Now consider being able to do DW operations against live, current OLTP data. And according to the performance numbers published by Oracle, those operations could run much faster than on most DW systems. That speed could result in completing more complex reports, allow more ad hoc queries, and so on. Such a change could be a fundamental advantage to DW consumers (finance and senior management, for example).
- Consider the cost of Oracle database software licenses. Now consider the hardware on which they run. Increasing the performance of that software gains your site more database performance at the same database license cost. The Exadata V2 is optimized to run OLTP and data warehouses very quickly. The resource management software included with the Exadata and its use as a consolidation engine probably lead to the appliance running with more databases using more resources and with less reserved headroom than having a non-Exadata database environment. That means that, for a given number of Oracle database licenses, your site would get more database performance.
- As mentioned above, customization of the pre-defined Exadata V2 configuration is allowed. For example, if your business required fewer database engines and more storage, it would be possible to get such a configuration from Oracle. Also, some sites might want to use the included Infiniband interconnect for fast backup of the data. However, the support model for custom configurations is likely to be different from the pre-defined ones. At the moment, even splitting a full rack of Exadata V2 into two racks (to prevent the rack from being a single point of failure) is a custom configuration.
- You can't build your own Exadata V2 system. Even though the hardware components of Exadata V2 are off-the-shelf Sun servers and networking, there is "magic sauce" in the Exadata. The Exadata storage software manages the storage nodes; the Exadata servers offload storage-centric operations to the storage nodes (again increasing the database performance you get with those Oracle licenses); and "Hybrid Columnar Compression," a new method for compressing columns of data while still making them available for OLTP access, is an Exadata V2–only feature. Following the Oracle/Sun

best practices and blueprints and using the same hardware components, could lead to something similar to the Exadata V2 in terms of features and performance, but the lack of those features means that it will not match the features and performance of Exadata V2. Further, the appliances are delivered pre-configured, and once the delivery team hands it over to the customer (a few days' effort), Oracle 11GR2, RAC clustering, InfiniBand configuration, and storage layout are all in place and performance is pre-tuned. Getting that to work from scratch on a build-your-own database server can add many weeks (and some risks) to a project. That should be considered when a datacenter is considering the buy-vs.-build decision.

The choice of SPARC vs. x86 and Linux vs. Solaris is difficult to interpret. Exadata V1 was based on the same technologies, Intel and Linux, so the shortest path to a new release was staying with them. And Oracle has stated that they are enthusiastic about SPARC and Solaris and plan to invest more in them than Sun did, so I don't believe there are any hidden messages in these selections.

## Conclusions and the Future

Until real-life field use experience with Exadata V2 appliances is gained, it is difficult to determine what the future will hold. I believe the Exadata V2 will be a successful offering from Oracle/Sun, given the compelling architecture and claimed break-through performance. Exadata V2 does seem to be a great marriage of Oracle and Sun and bodes well for future combined-technology products. They will certainly be worth considering, and if the price, performance, and feature set of each one make sense for a given datacenter, the datacenter could also gain from deployment cost savings and the simpler support allowed by calling one vendor rather than many. When I think of the potential cross-pollination of features between Oracle and Sun, the excitement increases. Just imagining Sun 7000-style DTrace-based analytics being applied to other types of applications (databases or virtualization, for example) makes me hopeful for the marriage of Oracle and Sun. Time, as always, will be the final arbiter of whether this is the reality or whether some lesser one comes to pass. In the meantime, watch the blogs for breaking information [6, 7, 8, 9].

### REFERENCES


[1] http://www.oracle.com/us/products/database/exadata/index.htm.

[2] http://www.reuters.com/article/idUSTRE58E80D20090915.

[3] http://www.theregister.co.uk/2010/03/26/oracle_q3_f2010_numbers/.

[4] http://www.oracle.com/technology/products/bi/db/exadata/pdf/ehcc_twp.pdf.

[5] http://www.oracle.com/technology/products/bi/db/exadata/pdf/exadata-technical-whitepaper.pdf.

[6] http://kevinclosson.wordpress.com/.

[7] http://structureddata.org/.

[8] http://blogs.oracle.com/databaseinsider/.

[9] http://ctistrategy.com/.