An Oracle White Paper
October 2009

Exadata Smart Flash Cache and the Sun Oracle Database Machine

ORACLE

Exadata Smart Flash Cache

The Sun Oracle Database Machine delivers extreme performance and scalability for all database applications including Online Transaction Processing (OLTP), Data Warehousing (DW) and consolidation of mixed database workloads. One of the key enablers of this is the Exadata Smart Flash Cache and the intelligent Oracle Exadata Storage Server Software that drives it. Exadata Smart Flash Cache intelligently caches data from the Oracle Database replacing slow mechanical I/O operations to disk with very rapid flash memory operations. Exadata Smart Flash Cache is the fundamental technology of the Sun Oracle Database Machine Full Rack that enables the processing of up to 1 million random I/O operations per second (IOPS), and the scanning of data within Exadata storage at up to 50 GB/second.

Oracle Database 11g: The First Flash Optimized Database

Oracle's Exadata Smart Flash Cache feature is unique. It is not a disk replacement – software intelligence determines how and when to use the Flash storage, and how best to incorporate Flash into the database as part of a coordinated data caching strategy. Scale out Exadata storage enables the benefits of flash performance to be delivered all the way to the application. Traditional storage arrays have many internal and network bottlenecks that prevent realizing the benefits of flash. Flash can be added to storage arrays, but they can not deliver much of the potential performance to applications.

The Sun Oracle Database Machine delivers 50 GB/sec of bandwidth from flash in Exadata V2. This is drastically higher than other solutions, and that is on uncompressed data. Combine this bandwidth with Exadata and Oracle Database compression and offload processing and the effective bandwidth is much higher. Traditional storage arrays can not keep up with the bandwidth of disks. That was the main problem solved by Exadata V1. Most storage arrays are bottlenecked at small single digit GB/sec of bandwidth. Adding flash to these systems will not improve their bandwidth. It will only increase the severity of the bottleneck. Traditional storage arrays can not deliver the benefits of flash to a data warehouse. The same applies to batch and reporting in OLTP systems.

Traditional storage arrays were better at servicing random IOPS issued by simple disks. They kept up with the random IOPS because there were fewer numbers of them. When flash, which is orders of magnitude faster than regular disks, is added in to the system, they are once again bottlenecked. For example a high-end storage subsystem can only run about 120,000 IOPS. Oracle is achieving 1 million IOPS at the database level. That means 1 million IOPS in flash, through the storage servers, across the network, and into the database servers. Storage arrays are already bottlenecked, and adding flash just exposes the bottlenecks even more.

Oracle is using flash PCIe cards in Exadata V2 – not flash disks. While it is easy to add flash disks into an existing storage subsystem without having to change anything else, the potential of the technology is not realized. Disk controllers and directors were never designed to keep up with the performance that flash disks enable. By using flash PCIe cards, Oracle's solution doesn't have a slow disk controller limiting flash performance. Exadata storage delivers close to 1GB/sec of throughput from each flash card, scales that performance linearly across the 4 cards in every Exadata Storage Server. Traditional storage arrays do not allow flash cards to be added to the system. Their architecture would need to be redesigned to avoid the disk controller limitations.

Oracle has implemented a smart flash cache directly in the Sun Oracle Exadata Storage Server. The Exadata Smart Flash Cache allows frequently accessed data to be kept in very fast flash storage while most of the data is kept in very cost effective disk storage. This happens automatically without the user having to take any action. It is the ultimate Information Lifecycle Management (ILM) solution. The Oracle flash cache is smart because it knows when to avoid trying to cache data that will never be reused or will not fit in the cache. Also, Oracle allows the user to provide directives at the database table, index and segment level to ensure that specific application data is kept in flash. Tables can be moved in and out of flash with a simple command, without the need to move the table to different tablespaces, files or LUNs like you would have to do with traditional storage with flash disks.

Oracle flash technology is tightly integrated into our end-to-end architecture. It is not a bolt-on accelerator that the user has to manually manage and optimize. In Exadata V2, Oracle introduced a new generation of compression technologies called Exadata Hybrid Columnar Compression that enables much more compression than was ever possible

before. This has a large number of benefits including greatly reducing the cost of storing large amounts of data, and increasing the speed at which data can be scanned. It also is very synergistic with our new flash technology. By compressing data by a factor of ten times or more, Oracle fits ten times more data into flash. This means that our flash becomes a lot more effective than the same flash capacity in any other product.

Exadata Smart Flash Cache Hardware

Each Sun Oracle Exadata Storage Server (Exadata cell) comes preconfigured with Exadata Smart Flash Cache. This low-latency solid state flash storage is packaged on a PCIe card (Sun Flash Accelerator F20) and each card has 96 GB of flash. Four flash cards are installed in each Exadata cell bringing the total flash storage per Exadata cell to 384 GB. Using PCIe based flash devices makes the full performance of the flash available. Using flash disk technology would limit the performance since flash disks would be connected to a relatively slow disk controller significantly reducing the IOPS and bandwidth that could be delivered. Each cell is capable of performing up to 75,000 Flash IOPS, end-to-end through the database and Exadata Storage Server, which is more than 20X the IOPs from regular disks.

Sun Flash Accelerator F20 PCIe Card

The performance the cache provides in the Sun Oracle Database Machine is multiplied by each Exadata cell in the Database Machine and is shown in the following table.

SUN ORACLE DATABASE MACHINE FULL RACK	SUN ORACLE DATABASE MACHINE HALF RACK	SUN ORACLE DATABASE MACHINE QUARTER RACK	SUN ORACLE DATABASE MACHINE BASIC SYSTEM
5.3 TB Exadata Smart Flash Cache	2.6 TB Exadata Smart Flash Cache	1.1 TB Exadata Smart Flash Cache	384 GB Exadata Smart Flash Cache
Up to 50 GB/second of uncompressed Flash data bandwidth	Up to 25 GB/second of uncompressed Flash data bandwidth	Up to 11 GB/second of uncompressed Flash data bandwidth	Up to 3.6 GB/second of uncompressed Flash data bandwidth
Up to 500 GB/second of compressed Flash data bandwidth	Up to 250 GB/second of compressed Flash data bandwidth	Up to 110 GB/second of compressed Flash data bandwidth	Up to 36 GB/second of compressed Flash data bandwidth
Up to 50,000 SAS or 20,000 SATA disk IOPS	Up to 25,000 SAS or 10,000 SATA Disk IOPS	Up to 10,800 SAS or 4,300 SATA Disk IOPS	Up to 3,600 SAS or 1,440 SATA Disk IOPS
Up to 1,000,000 Flash IOPS	Up to 500,000 Flash IOPS	Up to 225,000 Flash IOPS	Up to 75,000 Flash IOPS

With 5.3 TB of flash storage in a Full Rack configuration, the Oracle Database can perform up to 1 million IOPS (of 8K database blocks). In many cases the entire database will reside in the cache providing performance without the mechanical limits of standard disk technology. When multiple Full Racks are joined, via the provided InfiniBand fabric, hundreds of terabytes to petabytes of data can be stored in a single Database Machine. For each Full Rack added to the configuration 5.3 TB of flash storage is included scaling capacity and performance in lock step.

Exadata Storage Server Software

Exadata Smart Flash Cache provides an automated caching mechanism for frequently-accessed data in the Sun Oracle Database Machine. It is a write-through cache which can service extremely large numbers of random IOPS and enables OLTP applications to be deployed on the Database Machine.

Automated Management Of The Exadata Smart Flash Cache

The Oracle Database and Exadata Storage Server Software work closely together to cache frequently accessed data. When the database sends a read or write request to Sun Oracle Exadata Storage Server, it includes additional information in the request about whether the data is likely to be read again and therefore whether it should be cached. Based on the information the database sends the Exadata Storage Server Software intelligently decides which data will be re-read, and is worth caching, and those operations that would just waste cache. Random reads

against tables and indexes are likely to have subsequent reads and normally will be cached and have their data delivered from the flash cache. Scans, or sequentially reading tables, generally would not be cached since sequentially accessed data is unlikely to be subsequently followed by reads of the same data. Write operations are written through to the disk and staged back to cache if the software determines they are likely to be subsequently re-read.

Knowing what not to cache is of great importance to realize the performance potential of the cache. For example, when writing redo, backups or to a mirrored copy of a block, the software avoids caching these blocks. Since these blocks will not be re-read in the near term there is no reason to devote valuable cache space to these objects or blocks. Only the Oracle Database and Exadata Storage Server software has this visibility and understands the nature of all the I/O operations taking place on the system. Having the visibility through the complete I/O stack allows optimized use of the Exadata Smart Flash Cache to store only the most frequently accessed data.

All of this functionality occurs automatically without customer configuration or tuning and in most cases is the best use of the Exadata Smart Flash Cache.

User Management Of The Exadata Smart Flash Cache

There are two techniques provided to manually use and manage the cache. The first enables the pinning of objects in the flash cache. The second supports the creation of logical disks out of the flash for the permanent placement of objects on flash disks.

Pinning Objects In The Flash Cache

Preferential treatment over which database objects are cached is also provided with the Exadata Smart Flash Cache. For example, objects can be pinned in the cache and always be cached, or an object can be identified as one which should never be cached. This control is provided by the new storage clause attribute, CELL_FLASH_CACHE, which can be assigned to a database table, index, partition and LOB column.

There are three values the CELL_FLASH_CACHE attribute can be set to. DEFAULT specifies the cache used for a DEFAULT object is automatically managed as described in the previous section. NONE specifies that the object will never be cached. KEEP specifies the object should be kept in cache.

For example, the following command could be used to pin the table CUSTOMERS in Exadata Smart Flash Cache

```
ALTER TABLE customers STORAGE (CELL_FLASH_CACHE KEEP)
```

This storage attribute can also be specified when the table is created.

The Sun Oracle Exadata Storage Server will cache data for the CUSTOMERS table more aggressively and will try keeping this data in Exadata Smart Flash Cache longer than cached data for other tables. In the normal case where the CUSTOMERS table is spread across many Sun Oracle Exadata Storage Servers, each Exadata cell will cache its part of the table in its own Exadata Smart Flash Cache. Generally there should be more flash cache available than the objects KEEP is specified for. This leads to the table being completely cached over time.

While the default behavior for sequential scans is to bypass the flash cache, this is not the case when KEEP is specified. If KEEP has been specified for an object, and it is accessed via an offloaded Smart Scan, the object is kept in and scanned from cache. Another advantage of the flash cache is that when an object that is kept in the cache is scanned, the Exadata software will simultaneously read the data from both flash and disk to get a higher aggregate scan rate than is possible from either source independently.

Creating Flash Disks Out Of The Flash Cache

When an Exadata cell is installed, by default, all the flash is assigned to be used as flash cache and user data is automatically cached using the default caching behavior. Optionally, a portion of the cache can be reserved and used as logical flash disks. These flash disks are treated like any Exadata cell disk in the Exadata cell except they actually reside and are stored as non-volatile disks in the cache. For each Exadata cell the space reserved for flash disks is allocated across sixteen (16) cell disks – 4 cell disks per flash card. Grid disks are created on these flash-based cell disks and the grid disks are assigned to an Automatic Storage Management (ASM) diskgroup. The best practice would be to reserve the same amount of flash on each Exadata cell for flash disks and have the ASM diskgroup spread evenly across the Exadata cells in the configuration just as you would do for regular Exadata grid disks. This will evenly distribute the flash I/O load across the Exadata cells and flash.

These high-performance logical flash disks be used to store frequently accessed data. To use them requires advance planning to ensure adequate space is reserved for the tablespaces stored on them. In addition, backup of the data on the flash disks must be done in case media recovery is required, just as it would be done for data stored on conventional disks. This option is primarily useful for highly write intensive workloads where the disk write rate is higher than the disks can keep up with.

Mission Critical Availability Of The Exadata Smart Flash Cache

The hardware used for the Exadata Smart Flash Cache is very reliable but all hardware is subject to failure. Spreading the flash cache across 4 PCIe cards mitigates some of this risk. If there is a failure of one of the flash cards the Exadata Storage Server Software automatically detects the loss of the card and takes the failed portion of the flash cache offline. During this process the Exadata cell continues to operate and serve data from the remaining cache. All the data is persistently stored and protected on the regular disks in the system so no data will be lost or

corrupted. So while performance might be reduced because there is less flash cache available to service I/O requests, the system keeps running without interruption or data loss. This allows replacement of the failed flash card to be deferred until a convenient time at which the Exadata cell can be taken offline and flash card replaced. After the card is replaced the Exadata Storage Server Software automatically detects the presence of the new card and automatically starts using the additional flash cache.

If logical flash disks have been placed in the flash and one of the flash PCIe cards fails, again the impact of the failure is minimized by the Exadata Storage Server Software and ASM. If a flash card failure occurs, the 4 flash cell disks on the failed flash card are automatically taken offline and I/O to those disks are serviced from the mirrored extents stored in flash on other Exadata cells. Eventually an ASM re-balance would occur to re-silver the data across the unaffected flash disks. Once the faulty card is replaced, the flash disks will automatically be added back into the ASM diskgroup and a re-balance will be performed reestablishing the normal configuration.

Conclusion

The Exadata Smart Flash Cache is the power behind the OLTP functionality of the Sun Oracle Database machine. The Exadata Smart Flash Cache delivers unprecedented IOPS for the most demanding database applications. The Exadata Smart Flash Cache can more than double the scan rate for data in warehouse or reporting applications. By knowing what data to cache and how to automatically manage the cache, the Oracle Database, with the Exadata Smart Flash Cache, is the first and only flash enabled database.

White Paper Title
October 2009
Author: Ron Weiss
Contributing Authors: Caroline Johnston, Juan Loaiza, Lawrence McIntosh, Mahesh Subramaniam, Kodi Umamageswaran

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

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

0109