An Oracle White Paper
September 2010

Oracle Database Smart Flash Cache

ORACLE

Introduction

Oracle Database 11g Release 2 introduced a new database feature: Database Smart Flash Cache. This feature is available on Solaris and Oracle Enterprise Linux and allows customers to increase the effective size of the Oracle database buffer cache without adding more main memory to the system. For transaction-based workloads, Oracle database blocks are normally loaded into a dedicated shared memory area in main memory called the System Global Area (SGA). Database Smart Flash Cache allows the database buffer cache to be expanded beyond the SGA in main memory to a second level cache on flash memory. The Sun Storage F5100 Flash Array and the Sun Flash Accelerator F20 PCIe Card provide a natural fit for Oracle Database Smart Flash Cache and offer an excellent opportunity for end users to take advantage of this new functionality.

Online Transaction Processing (OLTP) environments are expected to benefit from this technology. It is appropriate for new Oracle Database deployments with IO-intensive workloads as well as existing environments with main memory constraints and IO-bound applications. Benefits can include both increased transaction throughput and improved application response times.

Oracle Database environments with the potential to make effective use of this technology include: workloads with repeated short transactions where many users access the same data, storage systems that exhibit intensive disk read activity, and systems under heavy main memory pressure that prevents more memory being allocated to the SGA buffer cache.

This feature is expected to primarily benefit read-mostly and read-only workloads. It is also supported in Real Application Cluster (RAC) environments – Database Smart Flash Cache can be applied to individual RAC nodes as required.

Oracle's Database Smart Flash Cache

Over many years the performance of spinning disk-based storage devices has not improved significantly. That situation has changed dramatically thanks to flash-based storage devices. For instance, 4 milliseconds can be considered a representative response time for small spinning-disk reads. Flash-based devices can deliver the same read in 0.4 milliseconds, an order of magnitude improvement in the response time. Oracle's new Database Smart Flash Cache feature leverages this I/O breakthrough offered by flash-based storage devices.

Oracle's Database Smart Flash Cache functions as a victim cache, which means that it stores clean (unmodified) database blocks that have been evicted from the SGA buffer cache to make room for other blocks. If a block selected for eviction has been modified, the dirty block is first written to disk by one of the database writer (DBWR) processes, then it is written to the Database Smart Flash Cache. Blocks can later be retrieved from the Database Smart Flash Cache and returned to the SGA as required. If a block cannot be found either in the SGA buffer cache or the Database Smart Flash Cache, it will be retrieved from disk.

Note that Exadata V2 Smart Flash Cache functions quite differently – it is a write-through cache that brings data into flash cache on read-write disk operations, whereas Database Smart Flash Cache is a read-only cache that brings data into flash cache on buffer cache replacement. Database Smart Flash Cache behaves in a similar way to the L2ARC used by the Oracle Solaris Zettabyte File System (ZFS).

How Does It Work?

The lifecycle of a block looks something like this:

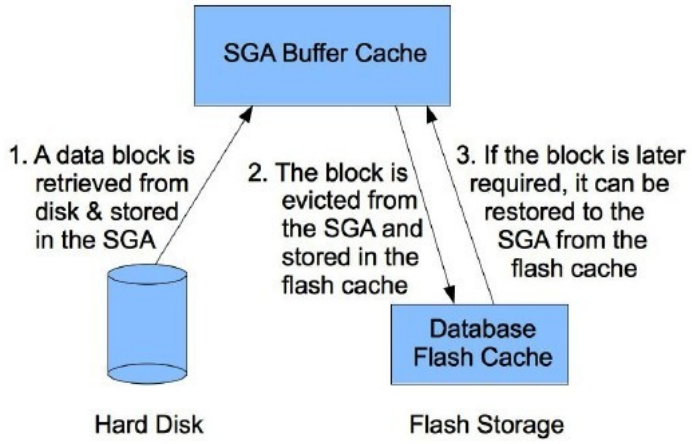

Figure 1: Lifecycle of a data block with Database Smart Flash Cache

Implementation

- The Database Smart Flash Cache feature introduces two new init.ora parameters: DB_FLASH_CACHE_FILE, which identifies the flash device
- DB_FLASH_CACHE_SIZE, which specifies the amount of flash storage available to the Database Smart Flash Cache

The DB_FLASH_CACHE_FILE parameter can point to a raw device (including a logical volume), a file, or an ASM disk group. Note that for raw devices, the partition being used should start at cylinder 1 rather than cylinder 0 (to avoid the disk's volume label).

Here are some examples of these variables in use in an init.ora file:

Identifying the file location of the Database Smart Flash Cache (which could also be a symlink to a raw device):

```
db_flash_cache_file = "/export/home/oracle/lffile_raw”
```

Indicating that the Database Smart Flash Cache is managed by an ASM instance:

```
db_flash_cache_file = "+dg1/lffile_asm"
```

Specifying the size of the Database Smart Flash Cache (in this case 50 Gbytes):

```
db_flash_cache_size = 50G
```

A typical sizing of the Database Smart Flash Cache is two to ten times the size of SGA memory buffers. Note that header information is stored in the SGA for each flash cache buffer - 100 bytes per buffer in exclusive mode, 200 bytes per buffer in RAC mode - so the number of available SGA buffers is reduced as the Database Smart Flash Cache size increases, and the SGA size should be increased accordingly.

The implementation of Database Smart Flash Cache currently only allows a single flash device to be specified. For this reason some form of volume management is necessary when using either the Sun Storage F5100 Flash Array or the Sun Flash Accelerator F20 PCIe Card. Tests based on a Sun Storage F5100 Flash Array have shown Oracle Database Automatic Storage Management (ASM) to be a better choice than a traditional volume manager.

Note that the Database Smart Flash Cache is not shared across RAC nodes; it is private to each RAC instance. For that reason, each RAC instance must define its own local flash cache file path. When the flash cache is managed by ASM, a separate diskgroup is needed for the flash cache of each instance.

Test Results

To characterize the Database Smart Flash Cache feature, a number of tests were run using a synthetic Online Transaction Processing (OLTP) workload on a Sun SPARC Enterprise M5000 with a Sun Storage F5100 Flash Array attached. The workload simulated an environment where multiple concurrent DB users submitted transactions against a database. Under heavy load, the large number of

concurrent transactions generated a significant number of Input/Output operations (I/Os) to the storage sub-system. The workload included only read-only transactions and therefore only issued reads to the database.

The graph in Figure 2 below shows test results when the workload was disk-bound (meaning that application throughput was limited by the number of I/Os that could be processed by the disk subsystem). As Database Smart Flash Cache was introduced and increased, throughput increased steadily up to a point where it flattened off. This "knee" in the throughput curve roughly corresponded to the I/Os per second (IOPS) limit for our particular flash configuration (more controllers connected to the Sun Storage F5100 Flash Array would have been needed to drive more IOPS). Had more IOPS been available, it might have been possible to drive the throughput higher.

In the test environment, 8 Gbytes was set aside for the SGA buffer cache, and the dataset size totaled 72 Gbytes. Consequently, approximately 65 Gbytes of flash cache were needed to cache the data entirely (assuming that about 1 Gbyte of the SGA buffer cache was consumed by flash cache headers, and that 7 Gbytes were therefore available in the SGA buffer cache for data blocks).

The graph shows that physical disk reads gradually diminished as the Database Smart Flash Cache size increased. Flash writes (which are needed to populate the Database Smart Flash Cache buffers) also diminish with more flash - with 50 Gbytes of flash storage, the majority of heavily-used buffers are already located in either the SGA or the Database Smart Flash Cache.

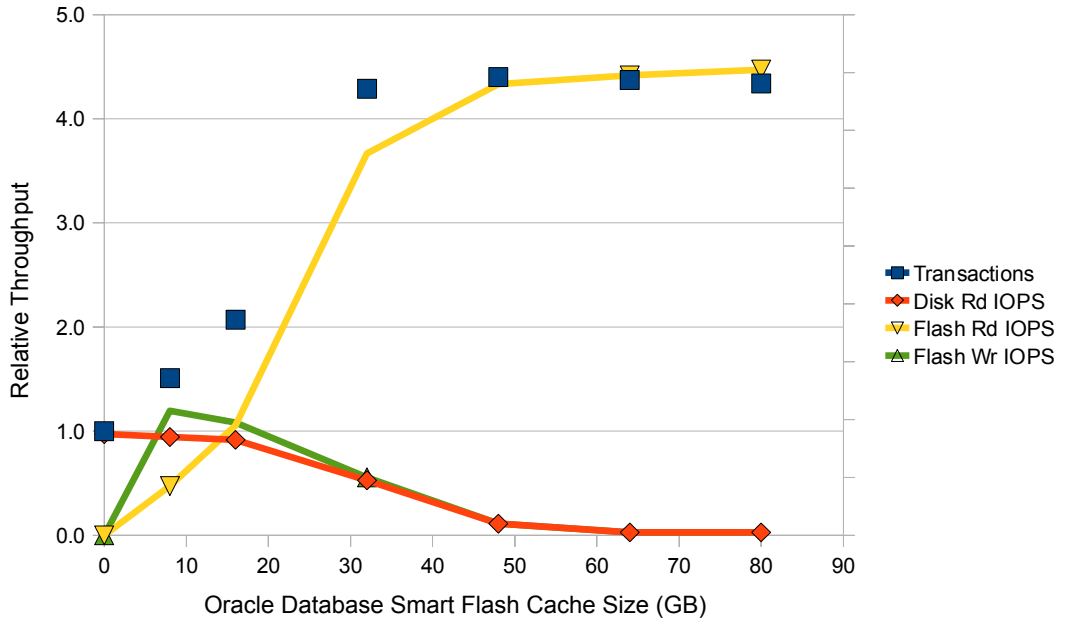

Figure 2: Relationship between Throughput and Disk and Flash I/Os

The final graph in Figure 3 below shows the transaction response times associated with the benchmark tests described above. In each case, the use of Database Smart Flash Cache resulted in a significant

response time improvement. Where the workload was significantly constrained by disk I/O and adequate Database Smart Flash Cache was assigned, the gains were very significant with response times improving almost 5 times. These results demonstrate that the vastly superior read and write performance of the Sun Storage F5100 Flash Array can lead directly to improved transaction response times for disk-constrained applications.

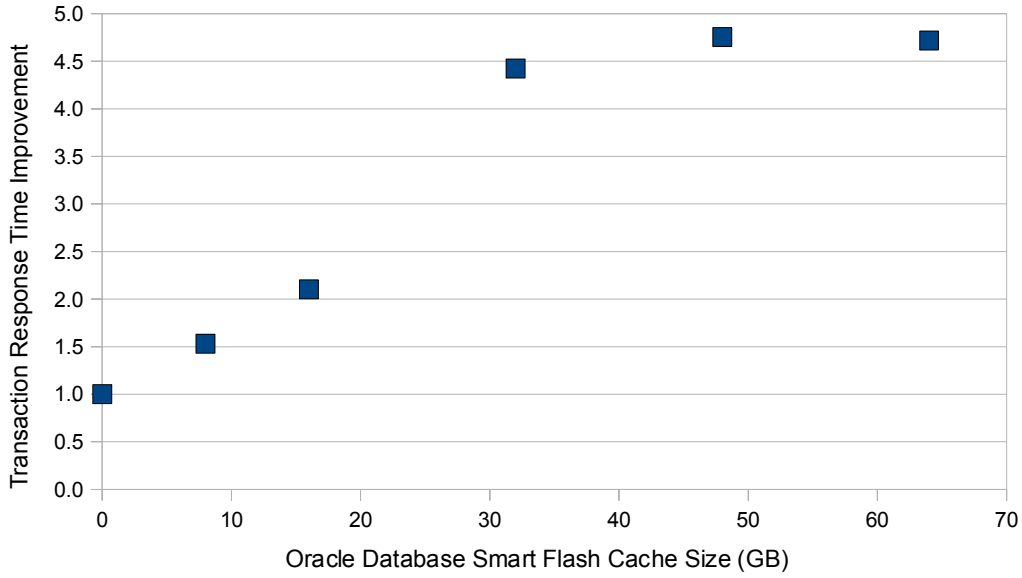

Figure 3: Response Time Improvement with Increasing Database Smart Flash Cache Size

Note that benefits may vary for write-intensive workloads, and improvements are likely to be less significant. Further investigations are underway to quantify the benefits in such environments; the results will be shared in a future whitepaper.

Conclusion

Oracle's Database Smart Flash Cache can make a big difference to application throughput for workloads that are disk bound, particularly read-mostly or read-only workloads. Note that where an unlimited amount of main memory is available, the best result will always be achieved by caching disk blocks in the SGA buffer cache.

The following guidelines offer an indication of where to expect gains from Database Smart Flash Cache:

- The greater the number of disk IOPS done by a workload, the bigger the benefit from Database Smart Flash Cache
- The benefit increases up to the point where the knee of the throughput curve is reached, after which increasing the size of flash cache offers no further benefit (the combined cache size

may already exceed the size of the working set for the application)

- Disk-bound workloads are likely to see significant improvements in transaction response times
- Read-mostly and read-only workloads are likely to see the greatest benefit; a pilot study is the best way to determine the potential benefit of Database Smart Flash Cache for write-intensive workloads
- Although Database Smart Flash Cache reduces physical disk reads, it also benefits writes because it reduces the load on the disks caused by reads, therefore leaving more IOPs available for writes

Available at no additional cost, Database Smart Flash Cache on Oracle Solaris and Oracle Enterprise Linux has the potential to offer considerable benefit to users of Oracle Database 11g Release 2 with disk-bound read-mostly or read-only workloads, through the simple addition of flash storage such as the Sun Storage F5100 Flash Array or the Sun Flash Accelerator F20 PCIe Card.

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

Oracle Database Smart Flash Cache
September 2010
Authors:
Allan Packer
Kevin Jernigan
Senthil Ramanujam

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

SOFTWARE. HARDWARE. COMPLETE.