An Oracle White Paper
June 2010

Accelerating Databases with Oracle's Sun Storage F5100 Flash Array Storage

ORACLE

Introduction 2
Overview of the Sun Storage F5100 Flash Array 3
Performance and Data Access Characteristics 3
Practical Implementation Considerations 4
Reference Architecture 5
Small Configuration Example 8
Large Configuration Example 9
System Performance Testing 10
Method 10
Client-Side New Order Response Time vs. New Order Rate 10
Oracle Tablespace I/O Read Service Time vs. I/O Rate 11
Scaling to Larger Systems 14
Read Service Time vs. Workload per Sun Flash Module 14
Index Read Service Time vs. Storage Workload 16
Conclusions 17
Appendix: System Tuning 18
Tuning for the Oracle Solaris OS 18
Storage Software and Configuration 20
Oracle Database Configuration 23
References 25

Introduction

Oracle's Sun Storage F5100 Flash Array storage is an innovation in storage technology, delivering the low latency and high I/O bandwidth characteristic of flash media to meet the storage requirements of modern databases.

This white paper shows how to apply the Sun Storage F5100 Flash Array as storage for database indexes in order to accelerate application performance. Storing indexes on flash storage can improve performance in two ways:

- Given a fixed workload, indexes on flash storage can reduce application response time
- Given a fixed desired response time, indexes on flash storage can increase the maximum supportable workload

This white paper includes practical examples of storage systems augmented with the Sun Storage F5100 Flash Array for index storage. The architecture described here offers a practical balance of performance and capacity for cost-constrained online transaction processing (OLTP) and very large database (VLDB) applications. The Sun Storage F5100 Flash Array can be added to these existing systems with few other changes. This article shows how the combination of flash and SAN disk technology can be applied to improve the performance of existing I/O-constrained systems.

Testing was performed using the Oracle Database 10g and Oracle Database 11g on the Oracle Solaris 10 operating system (OS) on both the x86/x64 and SPARC® platforms. These examples demonstrate application response time improvement of up to 50%, and application throughput improvement of up to 40%.

Overview of the Sun Storage F5100 Flash Array

The Sun Storage F5100 Flash Array is a serial attached SCSI (SAS) device based on serial ATA (SATA) flash devices called Oracle's Sun FlashFire Modules — a flash-based replacement for spinning disk drives. Sun Flash Modules overcome the performance and wear limitations of consumer-grade flash memory devices by providing:

- Capacitor-backed DRAM to support write processing
- A wear leveling algorithm to maximize the lifetime of the flash media

Performance and Data Access Characteristics

The Sun Storage F5100 Flash Array is an ideal building block for single-instance databases with I/O requirements that demand short data access time, high data access rates, and a workload dominated by read I/O operations. The Sun Storage F5100 Flash Array provides high-speed storage that can be seen as similar in function to simple "just a bunch of disk" (JBOD) trays with spinning disks. A single Sun Storage F5100 Flash Array storage device can support up to 80 Sun Flash Modules. Very high reliability, availability, capacity, workload, and response time requirements can be met by horizontal scaling across multiple Sun FlashFire Module devices and Sun Storage F5100 Flash Array systems.

When planning to support database applications with the Sun Storage F5100 Flash Array, important application considerations include requirements for physical I/O operations such as:

- Write service time
- Read service time
- Write rate
- Read rate

Important storage configuration choices related to these requirements include:

- Response time and throughput of the Sun Storage F5100 Flash Array Sun FlashFire Modules,
- Throughput and queuing limits of the host bus adapter (HBA) and HBA driver software

The white paper "Balancing System Cost and Data Value with Sun StorageTek Tiered Storage Systems"[1] provides a complete discussion of how to determine performance requirements for a database into architectural requirements of the database system. This discussion includes:

- How to translate database data access rate into physical drive I/O rate
- How to determine the most appropriate media I/O rate based on data access time requirements
- How to estimate the most appropriate media count based on reliability, availability, access time, access rate, and capacity requirements

Based on the theory presented in that white paper and Sun Storage F5100 Flash Array-specific engineering data, basic capacity planning for Oracle database applications can be based on the heuristics in Table 1.

TABLE 1.SUN STORAGE F5100 FLASH ARRAY PERFORMANCE HEURISTICS

PARAMETER	HEURISTIC VALUE
Typical small-block (8 KB) write service time per Sun FlashFire Module	1 to 3 ms over 100 to 4800 IOPS
Typical small-block (8 KB) read service time per Sun FlashFire Module	1 to 3 ms over 100 to 16000 IOPS
Typical large-block (1024 KB) write throughput per Sun FlashFire Module	50 MBPS
Typical large-block (1024 KB) read throughput per Sun Flash Module	260 MBPS
Supported maximum number of Sun Flash Modules per HBA for large-block sequential throughput	20
Supported maximum number of Sun Flash Modules HBA for small-block random throughput	20
Raw capacity per Sun Flash Module	24 GB

Practical Implementation Considerations

The dramatic improvements that can be realized through flash technology require that a system architect consider two important constraints:

- Physical access to a single Sun Flash Module is restricted to a single host system.
- Lightly-threaded write service times may be longer than traditional NVRAM-based storage systems.

Writes to redo log files stored in flash memory require more time than those of traditional NVRAM-based systems — a consequence of the write service time characteristics of flash devices. In database applications that are not sensitive to redo log write response time, the Sun Storage F5100 Flash Array can provide a high-bandwidth solution for storing the database redo log. However, applications that are constrained by redo log write response time, such as maintenance updates to dictionary tables, will benefit from using traditional NVRAM to store the redo log files.

Reference Architecture

This white paper defines a reference architecture for a database segregated into three components, as shown in Figure 1:

- Production table data files
- Production index files
- Flash recovery area

All table data files are contained in the production files (marked *P* in the figure), all index files are contained in the production index files (marked *P'*), and all recovery-related files are contained in the flash recovery area (FRA, marked *F* in the figure). The online redo log and control files are multiplexed over *P* and *F*. Although this example uses the Sun Storage F5100 Flash Array for storage of database indexes, the array may also be used to support any database data file that requires the low-latency and high bandwidth features of flash storage.

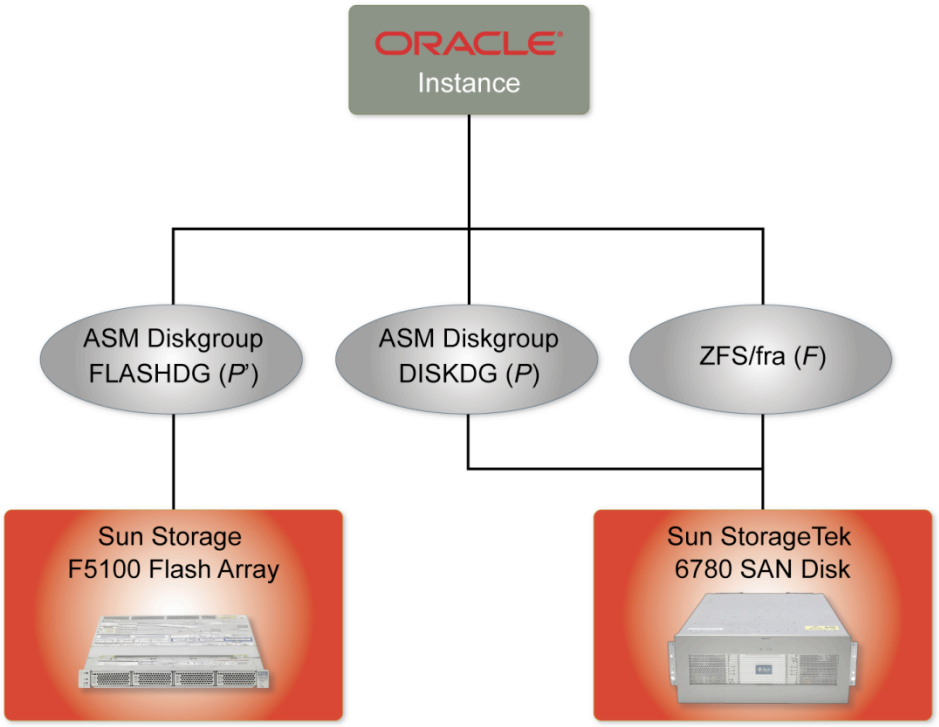

Figure 1.Reference architecture for using the Sun Storage F5100 Flash Array to store database indexes

The Oracle Database 10g or 11g instance communicates with Oracle Automatic Storage Management (ASM) and Oracle Solaris Zettabyte File System (ZFS) for access to production and recovery data (Table 2). Oracle ASM provides mirrored data protection (normal redundancy) for index files stored on the Sun Storage F5100 Flash Array. This storage target is designed to deliver ultra-low (1-3 ms)[1]

data access time for single-block reads at extreme access density (greater than 30 IOPS/GB) for a small subset of critical data that is needed early-and-often by important transactions.

TABLE 2.SOFTWARE CONFIGURATION

SOFTWARE CONFIGURATION	DATA LAYOUT
• Oracle Solaris 10 OS (update 7) • Oracle ASM with Oracle Database 10g/11g • Oracle ASM normal redundancy (Flash) • Oracle ASM external redundancy (SAN) • Oracle Solaris ZFS dynamic striping (SAN)	• FLASHDG: Indexes • DISKDG: Online redo, table data • Oracle ASM normal redundancy (Flash) • Oracle Solaris ZFS: online redo, FRA, archived redo

As part of this reference architecture, Oracle ASM implements striping (external redundancy) over hardware-mirrored data protection for database table data files stored on the Sun StorageTek 6780 storage array. This storage target is designed to deliver low data access times (5-15 ms) at high data access rates (1-3 IOPS/GB) for mission-critical production data and log files. Oracle Solaris ZFS provides a scalable and feature-rich storage solution to help ensure efficient space utilization and enterprise-class data protection to completely protect the critical recovery components of the Oracle database, including online redo log, archived redo log, backup sets, and FRA.

Small Configuration Example

Figure 2 shows a small example implementation of this reference architecture. The tested configuration was built using Oracle's Sun Fire X4600 M2 server with eight quad-core AMD Opteron™ processors and 32 GB of RAM to host the database server and OLTP application. The Sun Fire X4600 M2 server runs Oracle Solaris 10 OS (Update 7) and the Oracle Database 10g (version 10.2.0.1). Access to index files on the Sun Storage F5100 Flash Array is provided by four Sun StorageTek SAS HBAs from Oracle over four 4-lane SAS-1 channels. Access to the production data files and log files are through four 8-drive RAID5 logical units (LUNs) and access to the recovery files is through two 8-drive RAID5 LUNs on a Sun StorageTek 6540 storage array from Oracle, via four 4 Gb Sun StorageTek Fibre Channel (FC) HBAs from Oracle. In this example, all disk drives in the 6540 were 73 GB capacity, 15 000 RPM, 2 Gb Fibre Channel devices. (Note that the Sun StorageTek 6540 storage array was used in testing for convenience; Oracle's Sun StorageTek 6780 is the current recommended storage array.)

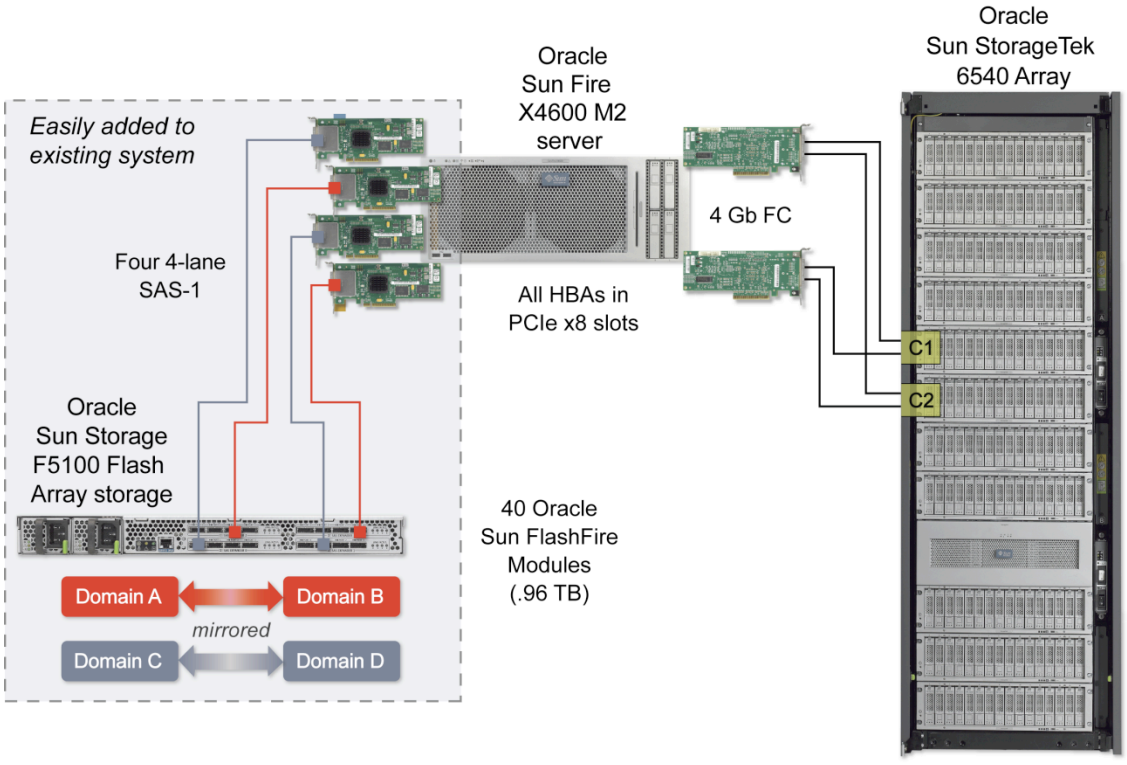

Figure 2.The Sun Storage F5100 Flash Array can be added to a legacy database deployment in a modular fashion.

Large Configuration Example

Figure 3 shows a large example implementation of this reference architecture. The tested configuration was built using a Sun SPARC Enterprise® T5440 server with up to 256 threads and 128 GB of RAM to host the database server and OLTP application. The Sun SPARC Enterprise T5440 server ran the Oracle Solaris 10 OS (Update 7) and the Oracle Database 10g (version 10.2.0.4). Access to index files on the Sun Storage F5100 Flash Array was through four Sun StorageTek SAS HBAs over four 4-lane SAS-1 channels. Access to the production table data and redo log files was through six 16-drive RAID 1+0 LUNs, and access to recovery files was through four 16-drive RAID5 LUNs on the Sun StorageTek 6780 storage array, via two 8 Gb Sun StorageTek Fibre Channel HBAs. All disk drives in the storage array were 450 GB capacity, 15 000 RPM, 4 Gb Fibre Channel devices.

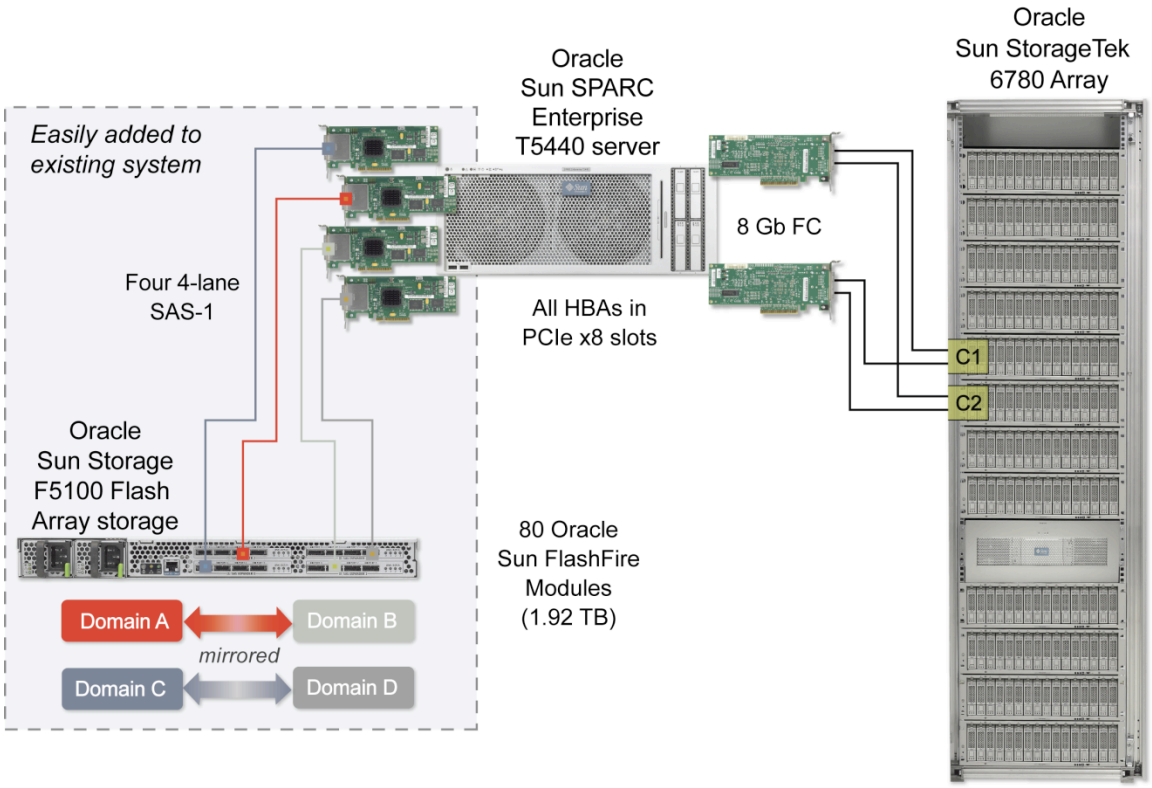

Figure 3.A large example implementation of the reference architecture

The small and large example implementations were both configured based on the architecture shown in Figure 1. Oracle ASM managed storage for production data files (indexes, tables, and online redo), and ZFS managed storage for recovery information (FRA, archived redo, online redo). Oracle ASM mirroring over Sun Storage F5100 Flash Array modules protected the index files (P′), and Oracle ASM striping over hardware RAID protected the database table data files (P) and copy 1 of the online redo log stream. Sun ZFS striping over hardware RAID protected the database flash recovery area (F), archived redo logs, and the second copy of the online redo log stream.

System Performance Testing

This section presents the results of a database study that compared the performance of storage architectures taking advantage of Sun Storage F5100 Flash Arrays for indexes with storage architectures based only on traditional SAN disk.

This study compared:

- Client-side new order service time versus new order rate
- Oracle instance measured (STATSPACK) tablespace I/O read service time versus I/O rate

Method

The assessment method employed in this article sought to determine the service time verses workload characteristics of the Sun Storage F5100 Flash Array in the context of an Oracle database supporting an online transaction processing (OLTP) application. The workload used in the test process was ramped from lightly loaded through application-level saturation. In the case of the Sun Storage F5100 Flash Array, throughput to store was limited by system bottlenecks strongly influenced by lock contention and read service times of database table data files stored on spinning disk. Although the test process accurately measures storage service time at fixed workloads, due to system-level performance constraints the test process did not measure the maximum throughput the Sun Storage F5100 Flash Array can support.

The test application implements new order, order status, payment, and stock level transactions to a 1 TB database. The Oracle Database 10g RDBMS hosts the database application data. In order to generate capacity planning information valid over a broad range of conditions, no specific application or database tuning was done to the system. Consequently, the throughput data represented in this paper represents a conservative estimate for system performance, and application-specific tuning could result in throughput increases.

Client-Side New Order Response Time vs. New Order Rate

Measuring changes in application service time as a function of application workload for systems based on hybrid flash/disk technology compared to traditional disk-only technology shows the kinds of gains realized at the application level for applications constrained by service time from the storage devices. In this example the average transaction executed eight writes and 25 reads. The reads are further distributed with 50% executed against indexes and 50% executed against tables.

Figure 4 shows the results of a small OLTP test system. In all cases new order service time was significantly lower for the system with index files stored on flash devices compared to the system with all data files stored on traditional disk. At a workload of 2500 new order transactions per minute (TPM), the hybrid disk/flash system delivers transaction service times of 0.2 sec compared to 0.4 sec in a disk-only solution. This 50% improvement in service time is a direct result of dropping I/O read service time for the indexes to 1 ms in case of the Sun Storage F5100 Flash Array compared to 15 ms in the case of the disk-only solution. At a fixed new order service time of 0.4 sec the maximum

supportable workload of the hybrid disk/flash system improves 36% to 3400 TPM compared to 2500 TPM in the disk-only system.

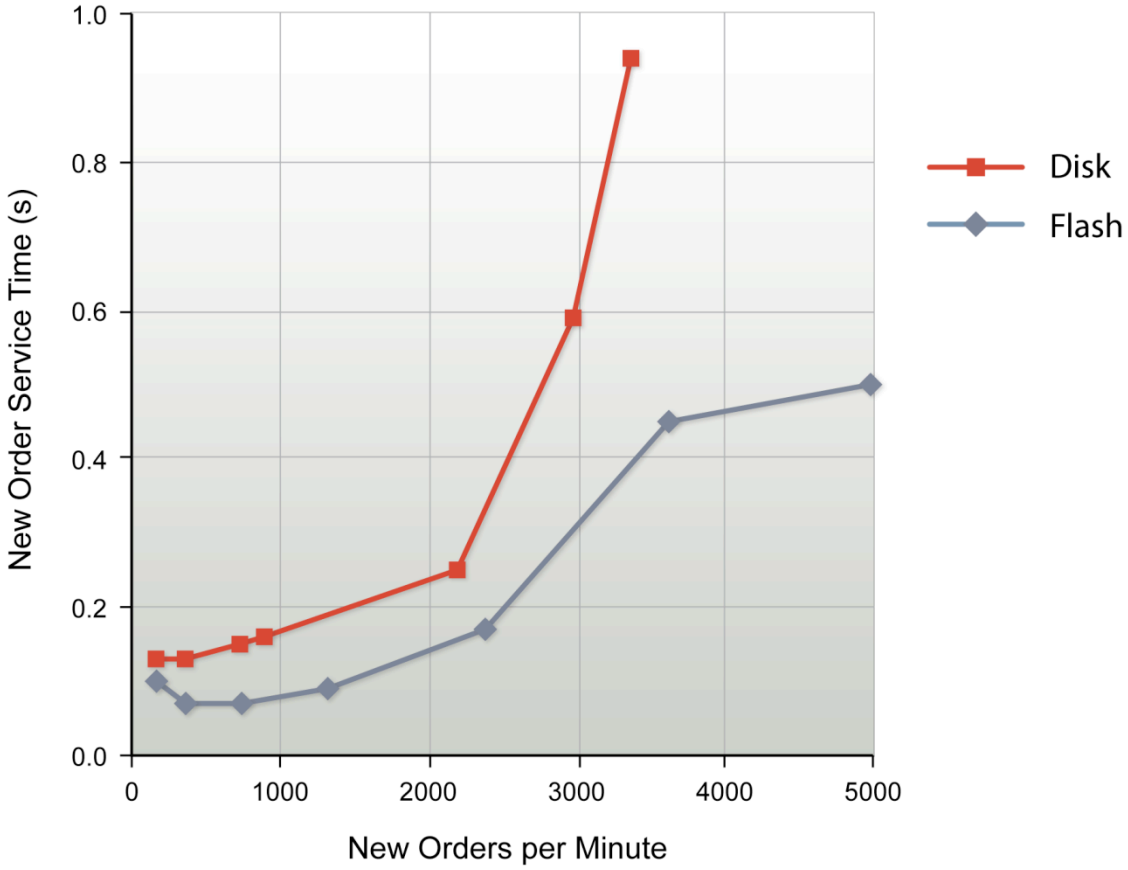

Figure 4.New order services times were reduced by 50% by using flash vs. disk.

This improvement in throughput is realized because more I/O can be processed in the same amount of time in the hybrid disk/flash system compared to a disk-only system.

Oracle Tablespace I/O Read Service Time vs. I/O Rate

Read service time from media (db file sequential read) is the leading wait event influencing transaction response time for the OLTP test system. In the case of the hybrid flash/disk architecture, this average includes data from two different populations: reads from flash and reads from disk. In the test system used to generate this illustration, 40% of the read I/O came from the indexes, with 60% of the read I/O coming from the table data and the I/O service time for reads coming from the data files.

Adding a Sun Storage F5100 Flash Array to off-load index I/O processing from an existing disk system to flash-based storage improves system performance in two ways:

- Read service time from index files is reduced dramatically
- Read service times for data files is reduced noticeably

In the case of the index files, compared to a modestly loaded 15 000 RPM disk drive, service time drops from 15 ms to 1 ms — an improvement of more than 90%. In the case of the table data files, because index processing has been moved to the flash device there is less workload for the disk to support, so the spinning disk can get the remaining work done more quickly.

Figure 5 shows the Oracle-instance reported db file sequential read wait event verses the total front-end I/O rate. The front-end I/O rate is defined as the sum of the physical reads, physical writes, and transactions executed per second. The service time is defined as the average I/O service time over all tablespaces, including the data and indexes. In the case of the test application, where about 50% of the I/O is executed against the indexes and 50% of the I/O is executed against the data, the average service time is approximately the average of the service time to each tablespace. In the case of migrating from spinning disk to Sun Storage F5100 Flash Array technology, the nearly ten times reduction in service time to the index effectively halves the average service time for the system. In the lightly-loaded case, average read service time drops from 6 ms to 3 ms, and as the disk begins to saturate, average read service time drops from 12 ms to 6 ms.

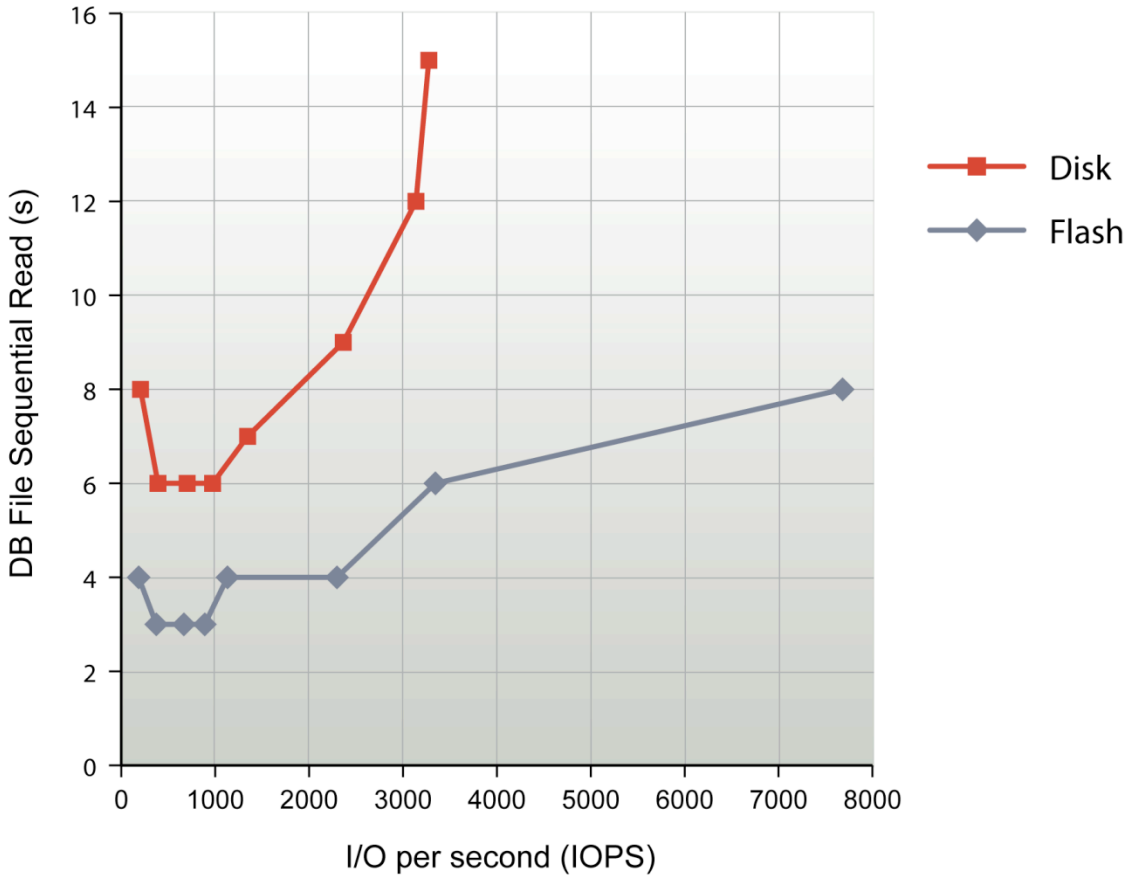

Figure 5.Small-system results: Oracle IOPS, Disk vs. Flash

Table 3 and Table 4 further show that the root cause of the 50% reduction in I/O service time at 3000 IOPS is a result of two effects:

- Index read service time drops 93% from 15 ms to 1 ms
- Table data read service time drops 20% from 15 ms to 12 ms.

TABLE 3.COMPARISON OF INDEX AND DATA READ SERVICE TIMES FOR A DISK-ONLY SYSTEM

TABLESPACE	READS	READS/SEC	RD(MS)	BLKS/RD	WRITES	WRITES/SEC	WAITS	WT(MS)
DATA	2173063	1117	16	1	1061755	546	3981	88
INDEXES	2400802	1234	15	1	413055	212	1580	11
UNDOTBS1	4153	2	1	1	18894	10	113	9
SYSAUX	1240	1	8	1	490	0	0	0
SYSTEM	173	0	13	5	34	0	31	7
TEMP	33	0	17	1	0	0	0	0
TOOLS	22	0	7	1	3	0	18	1

The 93% drop in index read service time comes from the reduction from spinning disk service times to flash service times, and the 20% drop in reads from the spinning disk results from reducing the I/O rate the disk is operating at by removing the index related I/O.

TABLE 4.COMPARISON OF INDEX AND DATA READ SERVICE TIMES FOR A HYBRID DISK/FLASH SYSTEM

TABLESPACE	READS	READS/SEC	RD(MS)	BLKS/RD	WRITES	WRITES/SEC	WAITS	WT(MS)
DATA	2101945	1089	12	1	1140742	591	4794	45
INDEXES	2409691	1248	1	1	441524	229	222	13
UNDOTBS1	10044	5	0	1	28418	15	140	1
SYSAUX	105	0	4	1	316	0	0	0
SYSTEM	127	0	12	7	29	0	9	4
TEMP	44	0	9	1	0	0	0	0
TOOLS	18	0	7	1	4	0	29	105

Scaling to Larger Systems

The test results presented so far represent a relatively small OLTP system based on the Oracle RDBMS. Further testing explored scalability, to evaluate whether these improvements extend to larger, higher capacity configurations.

Read Service Time vs. Workload per Sun Flash Module

A scalability test pushed the I/O rate to about 750 IOPS per Sun Flash Module with the I/O to the indexes tablespace. The results (see Figure 6 and Figure 7) show that for up to 750 IOPS per Sun Flash Module, the flash module consistently returns read I/O requests in 1–2 ms.

For comparison, the figures also include similar data taken for spinning disk. The comparison highlights the substantial improvement in response time and throughput as a result of including the Sun Storage F5100 Flash Array: an 85% reduction in I/O service time (2 ms vs 15 ms) at a 400% increase in throughput (750 IOPS vs 150 IOPS).

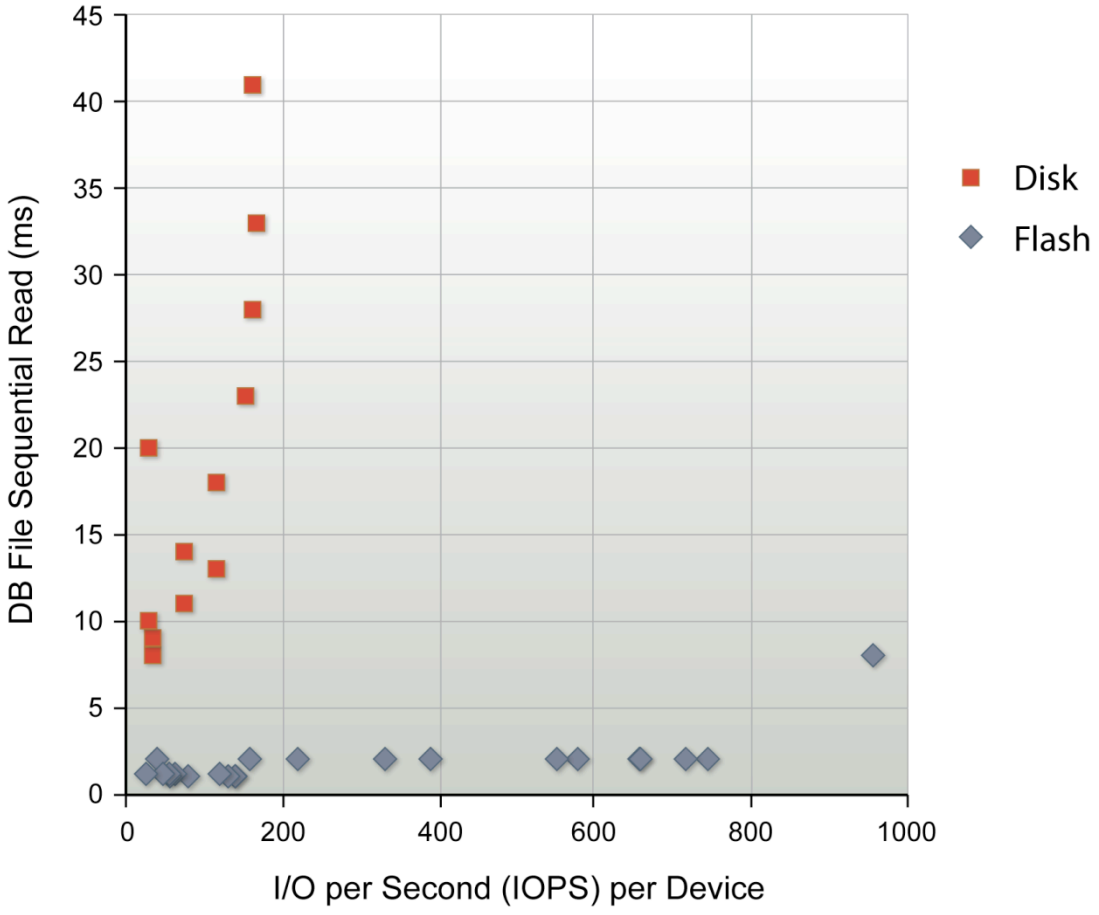

Figure 6.The Sun Storage F5100 Flash Array substantially improves response time.

These data show that adding the Sun Storage F5100 Flash Array for database index files is a practical method by which system planners can improve throughput and response time dramatically. Figure 7 uses the same data as Figure 6, but changes the scale to clarify the dramatic performance advantage of flash memory.

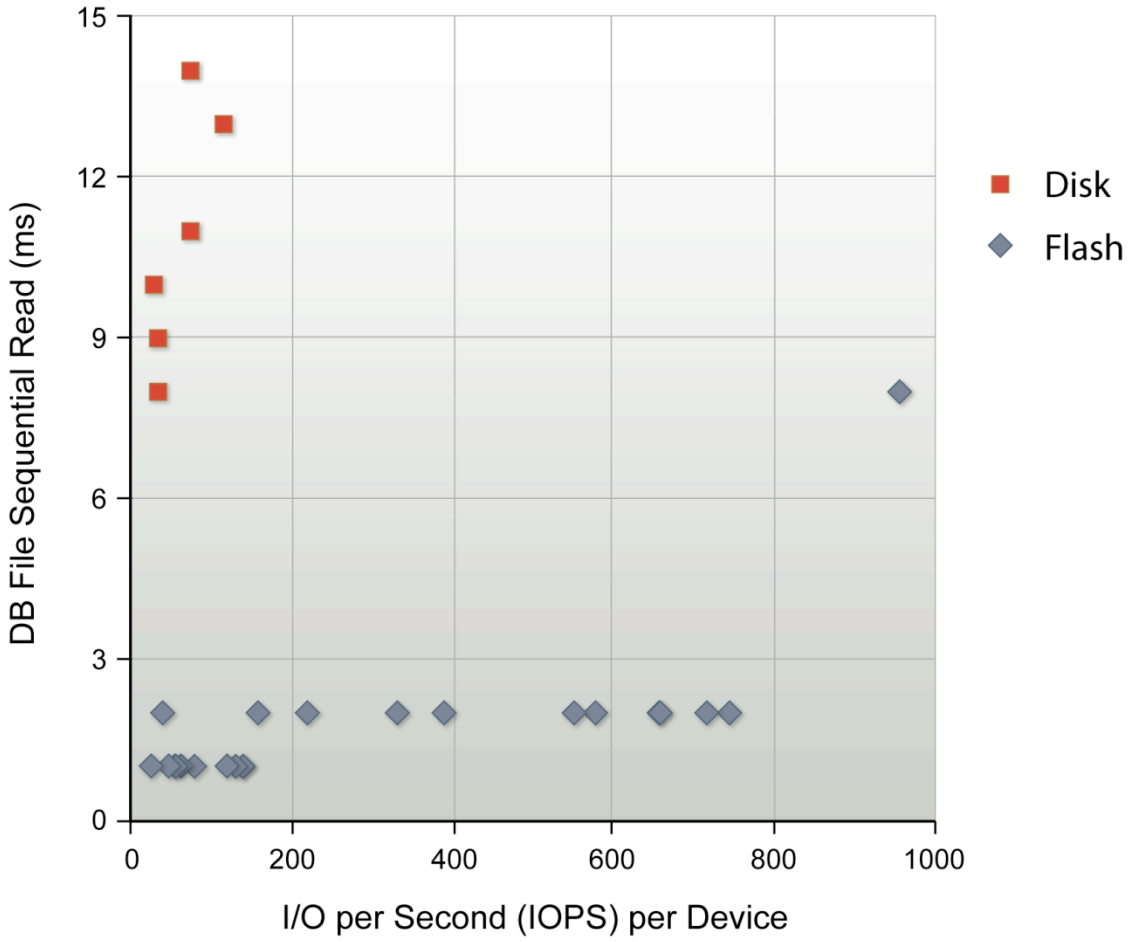

Figure 7.A close-up of flash array reads compared to disk reads shows much greater performance.

Index Read Service Time vs. Storage Workload

An important measurement taken during the large-system test scalability study compared realized index read service time from the Sun Storage F5100 Flash Array and traditional spinning disk. Figure 8 shows index read service time as a function of the total front-end I/O rate (index+table) for an OLTP system running from 5000-50,000 IOPS.

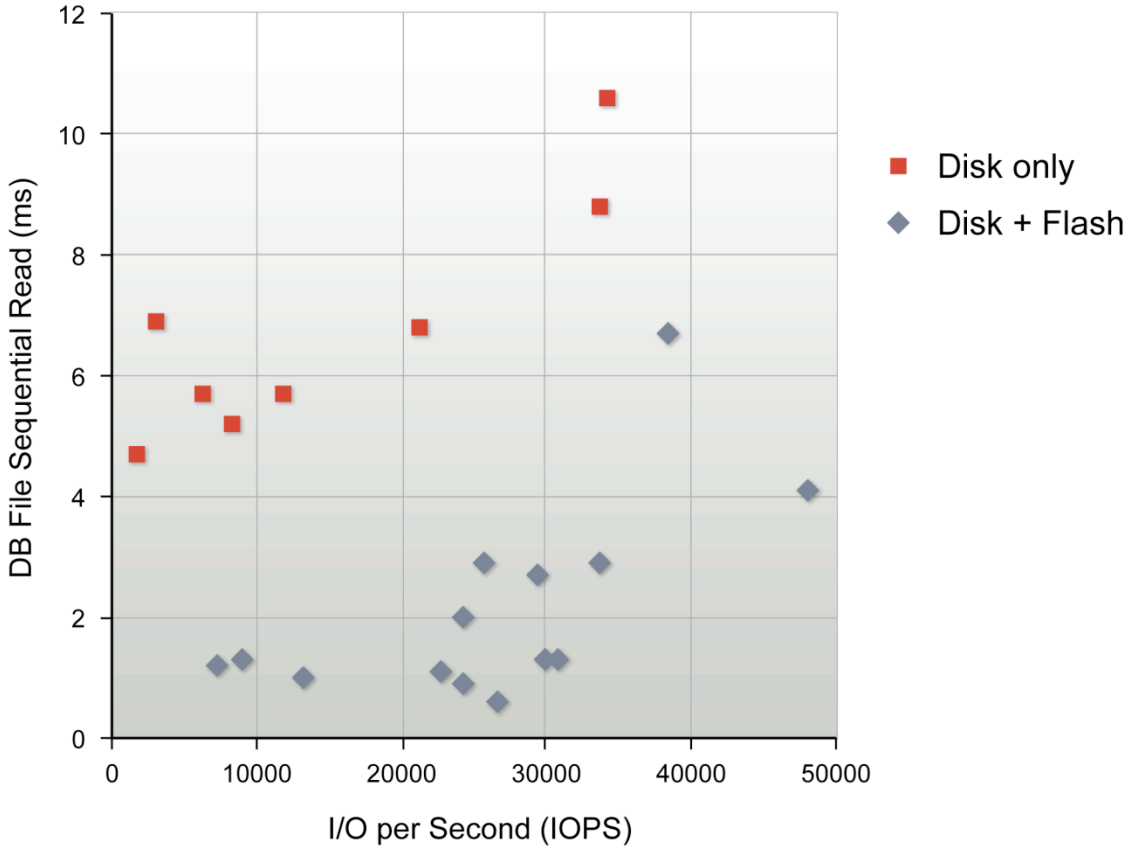

Figure 8.Combined disk and flash significantly improves index read service time over a broad range of I/O rates.

In this example, index read service time from the Sun Storage F5100 Flash Array measures 1-3 ms compared to 6-10 ms for the spinning disk drive. Performance gains of a 70% reduction in index I/O service time were achieved in this test case.

Conclusions

The data collected in this study highlight five important points:

- At a fixed rate of 2500 new order transactions per minute (TPM), the new order service time dropped from 0.4 sec to 0.2 sec (a reduction of 50%) when indexes were moved to the Sun Storage F5100 Flash Array.
- At a fixed new order service time of 0.4 sec, maximum throughput increases from 2500 TPM to 3400 TPM (an increase of 36%).
- Average read service time for all I/O (data + indexes) dropped 50% when indexes were stored on flash.
- Read service times of 2 ms per Sun Flash Module can be realized at workloads up to 750 IOPS per Sun Flash Module, yielding an access density of over 30 IOPS/GB.
- When the workload was scaled out to 10,000-30,000 IOPS, I/O service times for indexes dropped 60-70% when the indexes were moved to the Sun Storage F5100 Flash Array compared to traditional SAN storage.

End user response time savings for a specific implementation depend on how much time the user's transaction spends waiting on I/O, and how much time can be saved by storing some or all of the application data on the Sun Storage F5100 Flash Array. In the OLTP example shown previously, a typical transaction executes 16 I/O operations, and roughly one half of those operations are index reads. The combined I/O stack to the traditional disk configuration services I/O operations in 10 ms, while the combined I/O stack to the Sun Storage F5100 Flash Array services I/O operations in 3 ms. In the case where all I/O comes from disk, the total time spent on physical I/O is 160 ms, while in the case of a mixed configuration using the Sun Storage F5100 Flash Array for indexes and SAN disk for data files, the total time spent waiting on I/O drops 35% to 104 ms.

The reference architectures and supporting test data provide basic guidelines for designing and implementing database storage systems based on the Sun Storage F5100 Flash Array and traditional SAN storage. The data show examples of the application and storage performance gains that can be achieved, as well as important configuration details required to get the most out of the Sun Storage F5100 Flash Array with the Oracle Solaris OS. The system throughput can be scaled beyond the results shown in this paper by horizontally scaling storage components with additional Sun Storage F5100 Flash Arrays and Sun HBA using the guidelines and data presented in this paper.

Appendix: System Tuning

The storage response time and throughput limits of flash-based storage technologies can dramatically exceed those of spinning disk. Consequently, legacy device driver and storage software may need to be tuned to extract maximum storage throughput. Critical driver and storage software configurations include the number of queue tags available to send I/O requests to Sun Flash Modules. Likewise, optimizations in the Sun Flash Module require I/O block alignment on 4 KB boundaries for minimum response time and maximum throughput. By ensuring appropriate queuing and alignment of I/O to the Sun Flash Module devices, the system administrator can realize maximum benefit from the Sun Storage F5100 Flash Array.

Tuning for the Oracle Solaris OS

There are two important tuning steps to help ensure that Oracle Solaris runs well with the Sun Storage F5100 Flash Array:

- Provision the partitions of the Sun Storage F5100 Flash Array flash modules to begin storing Oracle ASM data on a 4 KB boundary.
- Update the Oracle Solaris `sd` driver to take advantage of the queue depth and non-volatile storage available with the Sun Storage F5100 Flash Array.

Compared to traditional disk devices, the Sun Storage F5100 Flash Array uses a larger page size (4 KB vs 512 B), and this requires the administrator to update the default configuration of the disk label. 4 KB I/O alignment can be preserved on an Oracle Solaris instance running on the SPARC architecture by starting the partition on a cylinder that is a multiple of 8: 0, 8, 16, 24, 32, and so on.

Due to the x86 BIOS present on x86/x64 systems running Oracle Solaris, the first cylinder is hidden from the operating system. As a result, cylinder 0 from the perspective of the operating system is actually cylinder 1 from the perspective of the Sun Flash Module. Preserving 4 KB I/O alignment can be accomplished by starting on cylinder 7 and adding multiples of 8 to cylinder 7: 7, 15, 23, 31, and so on. The "Appendix: Configuring Disk Layout" on page 23 details the discussion of disk layout and provides alternate methods to achieve correct I/O alignment on the Flash Module.

The example system shown in this paper used slice 6 from the default disk label available with the Oracle Solaris update 7 for the SPARC architecture. The output of the Oracle Solaris *format(1M)* utility displaying the partition table is shown below.

```
partition> p
Current partition table (original):
Total disk cylinders available: 23435 + 2 (reserved cylinders)
Part      Tag    Flag     Cylinders         Size            Blocks
  0       root    wm       0 -   127      128.00MB    (128/0/0)     262144
  1       swap    wu     128 -   255      128.00MB    (128/0/0)     262144
  2     backup    wu       0 - 23434       22.89GB    (23435/0/0) 47994880
  3 unassigned    wm       0                  0         (0/0/0)            0
  4 unassigned    wm       0                  0         (0/0/0)            0
  5 unassigned    wm       0                  0         (0/0/0)            0
  6        usr    wm     256 - 23434       22.64GB    (23179/0/0) 47470592
  7 unassigned    wm       0                  0         (0/0/0)            0
partition>
```

After aligning the device partition on a 4 KB boundary, the next tuning step is to make sure that the file system or volume manager maintains that alignment for the data it stores. The default configuration options for Oracle ASM maintain alignment provided the first cylinder is aligned. In the case of SPARC systems running Oracle Solaris, cylinder 0 is reserved for the disk label, so Oracle ASM disk groups must start on cylinder 8, 16, 24, 32, and so on. In the case of x86/x64 systems running Oracle Solaris, Oracle ASM may start on partition 7, 15, 23, and so on.

The last critical system tuning step specific to the Sun Storage F5100 Flash Array is to update the Oracle Solaris `sd` driver to account for the device queue depth and non-volatile cache of the Sun Flash Modules. The tuning can be accomplished with the following entries in the `/kernel/drv/sd.conf` file:

```
sd-config-list = "ATA     MARVELL SD88SA02","throttle-max:24, throttle-min:1,
cache-nonvolatile:true";
```

Note — The `sd` driver configuration is sensitive to spaces in the command line. Between ATA and MARVELL there should be five spaces, and between MARVELL and SD88SA02 there should be a single space.

In addition to the Sun Storage F5100 Flash Array specific tuning, the Oracle Solaris operating system used for the system shown in this paper was updated to support the Oracle RDBMS software and SAN storage devices with the following tuning parameters set in the `/etc/system` file:

```
set noexec_user_stack = 1
set noexec_user_stack_log = 1
set sdd:sdd_max_throttle=256
set maxphys = 8388608
set sd:sd_max_throttle=256
set semsys:seminfo_semmni=100
set semsys:seminfo_semmsl=256
set semsys:seminfo_semvmx=32767
set shmsys:shminfo_shmmax=4294967296
set shmsys:shminfo_shmmni=100
```

Storage Software and Configuration

Oracle ASM stores the index files on a mirrored (normal redundancy) disk group, FLASHDG, and table data files and online redo logs on a SAN attached disk array on a striped disk group (external redundancy), DISKDG. The following SQL*PLUS command created FLASHDG:

```
create diskgroup flashdg normal redundancy
disk '/dev/rdsk/c2t10d0s6',
      '/dev/rdsk/c2t11d0s6',
      '/dev/rdsk/c2t12d0s6',
      '/dev/rdsk/c2t13d0s6',
      '/dev/rdsk/c2t14d0s6',
      '/dev/rdsk/c2t15d0s6',
      '/dev/rdsk/c2t16d0s6',
      '/dev/rdsk/c2t17d0s6',
      '/dev/rdsk/c2t18d0s6',
      '/dev/rdsk/c2t19d0s6',
      '/dev/rdsk/c2t1d0s6',
      '/dev/rdsk/c2t20d0s6',
      '/dev/rdsk/c2t2d0s6',
      '/dev/rdsk/c2t3d0s6',
      '/dev/rdsk/c2t4d0s6',
      '/dev/rdsk/c2t5d0s6',
      '/dev/rdsk/c2t6d0s6',
      '/dev/rdsk/c2t7d0s6',
      '/dev/rdsk/c2t8d0s6',
      '/dev/rdsk/c2t9d0s6',
```

```
        '/dev/rdsk/c5t0d0s6',
        '/dev/rdsk/c5t10d0s6',
        '/dev/rdsk/c5t11d0s6',
        '/dev/rdsk/c5t12d0s6',
        '/dev/rdsk/c5t13d0s6',
        '/dev/rdsk/c5t14d0s6',
        '/dev/rdsk/c5t15d0s6',
        '/dev/rdsk/c5t16d0s6',
        '/dev/rdsk/c5t17d0s6',
        '/dev/rdsk/c5t18d0s6',
        '/dev/rdsk/c5t19d0s6',
        '/dev/rdsk/c5t1d0s6',
        '/dev/rdsk/c5t20d0s6',
        '/dev/rdsk/c5t2d0s6',
        '/dev/rdsk/c5t4d0s6',
        '/dev/rdsk/c5t5d0s6',
        '/dev/rdsk/c5t6d0s6',
        '/dev/rdsk/c5t7d0s6',
        '/dev/rdsk/c5t8d0s6',
        '/dev/rdsk/c5t9d0s6',
        '/dev/rdsk/c8t0d0s6',
        '/dev/rdsk/c8t10d0s6',
        '/dev/rdsk/c8t11d0s6',
        '/dev/rdsk/c8t12d0s6',
        '/dev/rdsk/c8t13d0s6',
        '/dev/rdsk/c8t14d0s6',
        '/dev/rdsk/c8t15d0s6',
        '/dev/rdsk/c8t16d0s6',
        '/dev/rdsk/c8t17d0s6',
        '/dev/rdsk/c8t18d0s6',
        '/dev/rdsk/c8t19d0s6',
        '/dev/rdsk/c8t1d0s6',
        '/dev/rdsk/c8t20d0s6',
        '/dev/rdsk/c8t2d0s6',
        '/dev/rdsk/c8t4d0s6',
        '/dev/rdsk/c8t5d0s6',
        '/dev/rdsk/c8t6d0s6',
        '/dev/rdsk/c8t7d0s6',
        '/dev/rdsk/c8t8d0s6',
        '/dev/rdsk/c8t9d0s6',
        '/dev/rdsk/c11t0d0s6',
        '/dev/rdsk/c11t10d0s6',
```

```
      '/dev/rdsk/c11t11d0s6',
      '/dev/rdsk/c11t12d0s6',
      '/dev/rdsk/c11t13d0s6',
      '/dev/rdsk/c11t14d0s6',
      '/dev/rdsk/c11t15d0s6',
      '/dev/rdsk/c11t16d0s6',
      '/dev/rdsk/c11t17d0s6',
      '/dev/rdsk/c11t18d0s6',
      '/dev/rdsk/c11t19d0s6',
      '/dev/rdsk/c11t1d0s6',
      '/dev/rdsk/c11t20d0s6',
      '/dev/rdsk/c11t2d0s6',
      '/dev/rdsk/c11t4d0s6',
      '/dev/rdsk/c11t5d0s6',
      '/dev/rdsk/c11t6d0s6',
      '/dev/rdsk/c11t7d0s6',
      '/dev/rdsk/c11t8d0s6',
      '/dev/rdsk/c11t9d0s6';
```

The ASM disk group hosting the table data files and online redo log stream, DISKDG, was created with the following SQL*PLUS command:

```
create diskgroup diskdg external redundancy
disk
'/dev/rdsk/c14t600A0B8000475B0C0000396D4A14407Cd0s0',
'/dev/rdsk/c14t600A0B8000475B0C000039714A144157d0s0',
'/dev/rdsk/c14t600A0B8000475B0C000039754A14426Dd0s0',
'/dev/rdsk/c14t600A0B8000475BAC00003B5C4A143FA6d0s0',
'/dev/rdsk/c14t600A0B8000475BAC00003B604A14407Dd0s0',
'/dev/rdsk/c14t600A0B8000475BAC00003B644A14416Ed0s0';
```

The ZFS pool and file system hosting the flash recovery area was created with the following Oracle Solaris commands:

```
# zpool create frapool c14t600A0B8000475B0C000039794A144410d0\
                       c14t600A0B8000475B0C0000397D4A1445D5d0\
                       c14t600A0B8000475BAC00003B684A1442C9d0\
                       c14t600A0B8000475BAC00003B6C4A144483d0
# zfs create frapool/arch
```

Oracle Database Configuration

The test database was configured to emulate a storage-bound OLTP system. As previously described, index files are stored in the FLASHDG disk group, online redo logs and database table data are stored in the DISKDG disk group, and the flash recovery area and archived redo logs are stored by ZFS. The complete server parameter file is shown below.

```
log_archive_dest_1='LOCATION=/frapool/arch/bench'
log_archive_format=%t_%s_%r.dbf
db_block_size=8192
db_cache_size=8192M
db_file_multiblock_read_count=128
cursor_sharing=force
open_cursors=300
db_domain=""
db_name=bench
background_dump_dest=/export/home/oracle/admin/bench/bdump
core_dump_dest=/export/home/oracle/admin/bench/cdump
user_dump_dest=/export/home/oracle/admin/bench/udump
db_files=1024
db_recovery_file_dest=/frapool/fra/bench
db_recovery_file_dest_size=214748364800
job_queue_processes=10
compatible=10.2.0.1.0
java_pool_size=0
large_pool_size=0
shared_pool_size=2048m
processes=4096
sessions=4131
log_buffer=104857600
audit_file_dest=/export/home/oracle/admin/bench/adump
remote_login_passwordfile=EXCLUSIVE
pga_aggregate_target=1707081728
undo_management=AUTO
undo_tablespace=UNDOTBS1
control_files=/oradata/bench/control0.ctl,/frapool/fra/bench/
control1.ctl
db_block_checksum=false
```

```
SQL> create tablespace indexes
   2 datafile '+FLASHDG/bench/indexes000.dbf' size 128M reuse autoextend on
   3 extent management local segment space management auto;
SQL> alter tablespace indexes add datafile '+FLASHDG/bench/indexes001.dbf' size 128M
   2 reuse autoextend on;
SQL> alter tablespace indexes add datafile '+FLASHDG/bench/indexes002.dbf' size 128M
   2 reuse autoextend on;
...
SQL> alter tablespace indexes add datafile '+FLASHDG/bench/indexes039.dbf' size 128M
   2 reuse autoextend on;
SQL> create tablespace data datafile '+DISKDG/bench/data000.dbf' size 128M
   2 reuse autoextend on extent management local segment space management auto;
SQL> alter tablespace data add datafile '+DISKDG/bench/data001.dbf' size 128M
   2 reuse autoextend on;
```

References

REFERENCES	
WHITE PAPERS	
Balancing System Cost and Data Value With Sun StorageTek Tiered Storage Systems	http://wikis.sun.com/display/BluePrints/Balancing+System+Cost+and+Data+Value+With+Sun+StorageTek+Tiered+Storage+Systems
Deploying Hybrid Storage Pools With Sun Flash Technology and the Solaris ZFS File System	http://wikis.sun.com/display/BluePrints/Deploying+Hybrid+Storage+Pools+With+Sun+Flash+Technology+and+the+Solaris+ZFS+File+System
Configuring Sun Storage 7000 Unified Storage Systems Systems for Oracle Databases	http://wikis.sun.com/display/BluePrints/Configuring+Sun+Storage+7000+Unified+Storage+Systems+for+Oracle+Databases

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

Accelerating Databases with the Sun Storage
F5100 Flash Array Storage
June 2010 PN:cnt0000000 Rev. 1

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.