High Performance QFS with Solid State Disk Metadata Storage

Bryan Banister *San Diego Supercomputing Center*
Jamon Bowen *Texas Memory Systems*

Abstract

SUN's QFS file system provides a powerful global file system across a distributed storage area network (SAN). The file system provides excellent bandwidth for multiple users to access a shared file system. QFS uses an architecture where the file system metadata is stored on a specific storage device. The metadata storage device can often become the bottleneck for critical file system maintenance operations, such as backups, and can impede the first opening of files under a heavy concurrent user load. The San Diego Supercomputer Center deployed the SAM-QFS file system to serve as a high performance archive and data collection resource, currently serving over 500 Terabytes of scientific data. While the large disk cache provided excellent transfer rates and ample storage capacity, the metadata device became completely overwhelmed while servicing simultaneous user requests and administrative tasks. In an attempt to resolve their metadata performance problems, the San Diego Super Computer Center tested a RAM based solid state disk from Texas Memory Systems for the metadata storage. This change resulted in a huge decrease in the completion time for file system operations across a range of activities. On one of SDSC's largest file systems, the file system metadata backup operation decreased from 21 hours and 40 minutes to 34 minutes.

Table of Contents

Problem

The San Diego Supercomputer Center is a world leader in using, innovating, and providing information technology to enable advances and new discovery in science and engineering. Focusing on data-oriented and computational science and engineering applications, San Diego Supercomputer Center serves as an international resource for data cyberinfrastructure through the provision of software, hardware, and human resources in multi-disciplinary science and engineering, and serves as a leadership

national cyberinfrastructure center to the National Science Foundation (NSF) and broader communities.

In order to support the explosive growth in data intensive computing, in which applications will generate 10's to 100's of Terabytes of data in a single run, San Diego Supercomputer Center deployed a large SAM-QFS archival system.

SUN's QFS file system allows multiple servers to attach directly to shared disk storage over a storage area network (SAN) and view all of the storage as through a single file system. This allows the bandwidth of data transfers to the file system to exceed what any one server could provide. Allowing multiple servers to all have access to a shared file system requires more complex management of the file system than a single server case does. Access to files and directories in a file system are mapped using metadata to decode the physical location of a file on the shared disk storage and other information. This metadata is a small component of a file system; however, it is referenced heavily and experiences a disproportionately large amount of IO than its size would suggest.

The SAM-QFS file system at SDSC provides a 130 Terabyte shared disk cache file system, which acts as a fast front end to the massive tape archive and is accessible directly from each of the large high performance computer (HPC) clusters at San Diego Supercomputer Center. Utilizing the current generation of tape drive technology from STK, the archive has a maximum capacity of over six Petabytes. The following logical diagram depicts the connectivity of these systems to the SAM-QFS archive, with arrows showing the storage access paths used from San Diego Supercomputer Center's current compute systems.

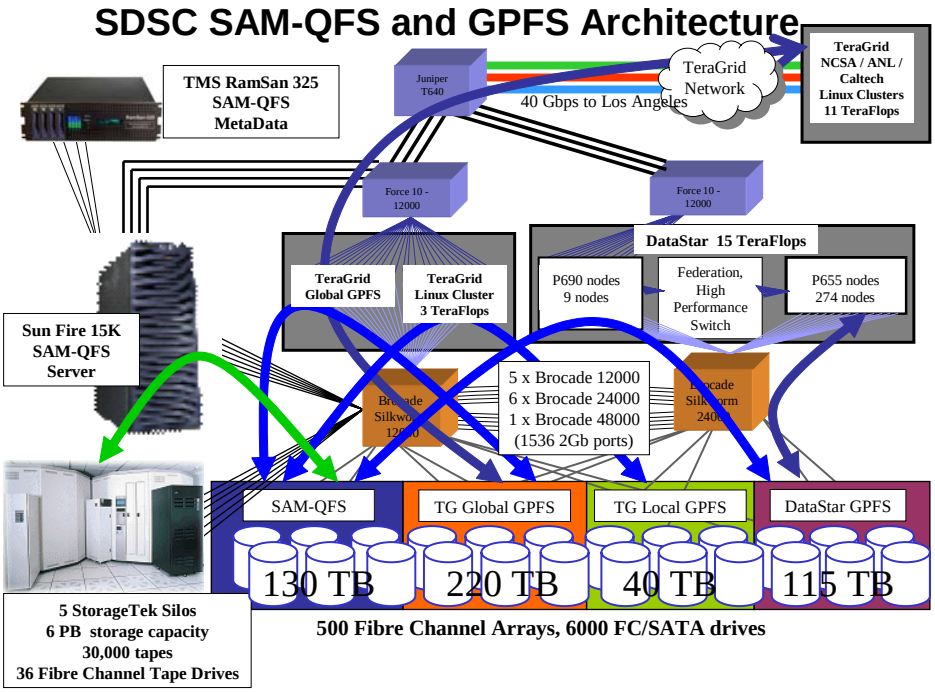

Figure 1 - SDSC SAM_QFS and GPFS Architecture

With direct access to the shared disk cache over the SAN and the ability to “pin” data to the shared disk cache, multiple applications can access data in the archive directly just as they would a normal high performance file system. This access method has lead to increased I/O load on the metadata device as it handles these requests along with administrative tasks such as metadata backups and data archiving to tape.

Solution

In QFS, the metadata is isolated onto a metadata device to allow a storage system that can address this heavier IO load. The glossary of “Sun StorEdge™ QFS and Sun StorEdge™ SAM-FS File System Administration Guide” [1] defines a metadata device as follows:

> **metadata device** A separate device (for example, a solid-state disk or mirrored device) upon which Sun StorEdge QFS file system metadata is stored. Separating the file data from the metadata can increase performance. In the `mcf` file, a metadata device is declared as an `mm` device within an `ma` file system.

In keeping with this definition, San Diego Supercomputer Center tested the RamSan-325 solid state disks for a QFS file system where the metadata device had been identified as the bottleneck for the file system.

The RamSan 325 Solid State Disk

The RamSan-325 system from Texas Memory Systems is a Fibre Channel attached storage device with DDR memory as the primary storage medium. The system provides low latency access to the storage space for any access pattern (read/writes, small/large block, random/sequential, burst/sustained). Using DDR memory allows sub millisecond response time even under heavy IOPS load (>100,000 IOPS) [2]. The RamSan-325 has seen wide application as a performance enhancement to Oracle™ and MS SQL™ databases by placing the performance critical sections of the database onto the RamSan.

Figure 2 - The RamSan-325 Solid State Disk

The storage system incorporates redundant arrays of batteries and disks to create a non-volatile storage device. Prior to system startup, the contents of an internal RAID are read into memory, and upon power loss or system shutdown the contents of memory are written to the RAID before the system powers off.

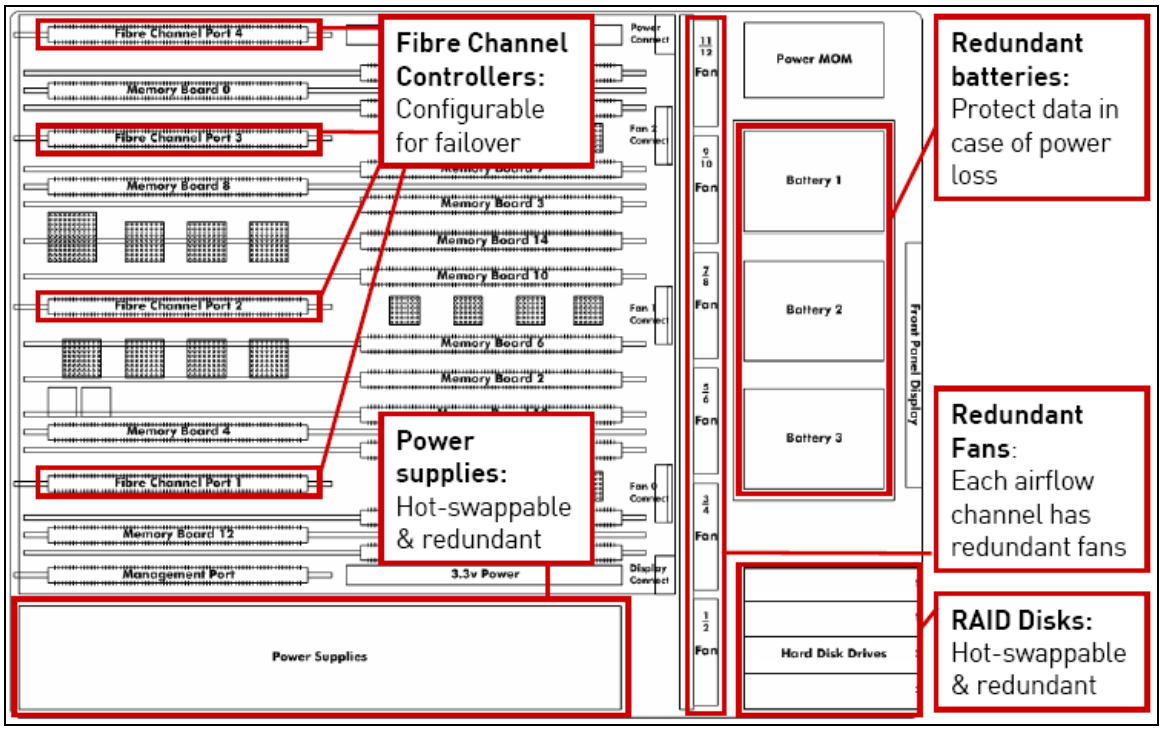

Figure 3 - RamSan-325 Architecture

Results

SAM-QFS Administrative Utility Differences

The archival system needs to support the regular use of administrative commands along with regular user access, metadata backups, etc. One typical utility that administrators use is the Sfind command, similar to the UNIX find command, which searches through a file system for files and directories with chosen properties. The Sfind command allows administrators to search for files that have certain archival properties, such as all files stored on a specific tape. Since this information is contained in the file system metadata, it generates significant load on the metadata device. The following snapshot shows the iostat output during an Sfind activity on the RamSan-325.

```
SAM-QFS Sfind Operation: to determine number of files
Calculated 16k byte IO for reads.
~4740 max IOPS at 100% busy per LUN.

Sfind to determine number of files
                 extended device statistics
  r/s    w/s   kr/s   kw/s wait actv wsvc_t asvc_t  %w  %b device
  0.0    0.0    0.0    0.0  0.0  0.0    0.0    0.0   0   0 c8t20010020C2032485d0s0
471.8   26.6 2146.4   13.3  0.0  0.1    0.0    0.2   1  10 c9t21010020C2032485d1s0
  0.0    0.0    0.0    0.0  0.0  0.0    0.0    0.0   0   0 c10t20020020C2032485d2s0
474.0   30.0 2152.8   15.0  0.0  0.1    0.0    0.2   1  10 c11t21020020C2032485d3s0

                 extended device statistics
  r/s    w/s   kr/s   kw/s wait actv wsvc_t asvc_t  %w  %b device
 60.4    0.0  954.3    0.0  0.0  0.0    0.0    0.3   0   2 c8t20010020C2032485d0s0
 59.0    0.0  943.9    0.0  0.0  0.0    0.0    0.3   0   2 c9t21010020C2032485d1s0
 60.2    0.0  954.3    0.0  0.0  0.0    0.0    0.3   0   2 c10t20020020C2032485d2s0
 59.4    0.0  950.2    0.0  0.0  0.0    0.0    0.3   0   2 c11t21020020C2032485d3s0
-------------------------------------------------------
```

Note that the `asvc_t`, the average service time, of active transfers are under 0.3 milliseconds. This latency improvement provided significant difference in the performance of this operation.

SAM-QFS Filesystem Dump Differences

File system metadata backups are critical to the survival of the user data held in the archive. San Diego Supercomputer Center uses this backup data to restore files upon user request and to restore operation of the archive in the event of a catastrophic storage or software failure. With the current metadata storage, metadata backups take close to 24 hours to complete. SDSC set a goal of performing metadata backups four or more times a day, providing better data integrity to the archive data. IOSTAT outputs from the existing metadata storage showed a constant 100% busy statistic for the metadata devices, clearly showing the bottleneck imposed by the disk performance.

The following snapshot displays the disk and system performance during a metadata backup with the RamSan-325.

```
us sy wt id
1 10 21 68
                  extended device statistics
  r/s    w/s   kr/s   kw/s wait actv wsvc_t asvc_t  %w  %b device
143.4    0.4 2258.0    0.4  0.0  0.0    0.0    0.3   0   4 c8t20010020C2032485d0s0
140.8    0.0 2242.8    0.0  0.0  0.0    0.0    0.3   0   4 c9t21010020C2032485d1s0
144.0    0.0 2260.4    0.0  0.0  0.0    0.0    0.3   0   4 c10t20020020C2032485d2s0
140.8    0.0 2244.4    0.0  0.0  0.0    0.0    0.3   0   5 c11t21020020C2032485d3s0
```

The percent busy of the RamSan-325 metadata devices, at 4 %, shows the how lightly this operation taxed the storage system.

The impact of the RamSan-325 on the backup operation is shown by the difference in the completion times of metadata dumps on the current metadata storage vs. the RamSan-325 as the metadata storage. The following snapshots compare the metadata dump output of two file systems with the same metadata. Notice the start and end times for the backups.

`/archive/science` used the current storage:

```
dumping /archive/science
Start time: Sep  1 2005 00:00:00
samfsdump statistics:
    Files:                  25432851
    Directories:   913735
    Symbolic links:2510
    Resource files:17
    File segments: 0
    File archives: 25417162
    Damaged files: 461
    Files with data:        0
    File  warnings:15686
    Errors:                 2
    Unprocessed dirs:       2
    File data bytes:        0
End time: Sep  1 2005 21:40:41
```

`/archive/ssdtest` used the RamSan-325:

```
dumping /archive/ssdtest
Start time: Sep  2 2005 00:00:00
samfsdump statistics:
    Files:                  24230756
    Directories:   828434
    Symbolic links:256
    Resource files:0
    File segments: 0
    File archives: 24171533
    Damaged files: 59662
    Files with data:        0
    File  warnings:0
    Errors:                 2
    Unprocessed dirs:       2
    File data bytes:        0
End time: Sep  2 2005 00:34:03
```

The RamSan-325's performance greatly exceeded San Diego Supercomputer Center's performance requirements, reducing the metadata backup time by 21 hours!

Raw Performance Results at SDSC (vdbench)

Part of the testing performed at SDSC included characterizing the RamSan-325's performance in the server/HBA environment at SDSC. This testing was done to ensure that the RamSan-325 system could support increasing demands as the file system changed. To accomplish this testing, the Sun StorEdge™ vdbench tool was used. Vdbench is a command line utility that collects a wide variety of performance metrics for a storage system [3]. In this case, it was used to test the maximum IOPS load that the server and storage system (RamSan-325) could support in this setup.

The following charts display collections of the test results. During the testing, the system printed the following message during the heavy IOPS loads:

```
16:45:06.063 *
16:45:06.063 * Warning: average processor utilization 93.09%
16:45:06.064 * Any processor utilization over 80% could mean that your system
16:45:06.064 * does not have enough cycles to run the highest i/o rate possible
16:45:06.065 * Maybe running 'java vdblite' may give you some more cycles
16:45:06.066 *
```

This indicated that the server CPU was the IOPS bottleneck during the testing, and that with higher server resources a higher IOPS could be achieved. The results show that the RamSan-325 has ample performance remaining to support the file system as it grows.

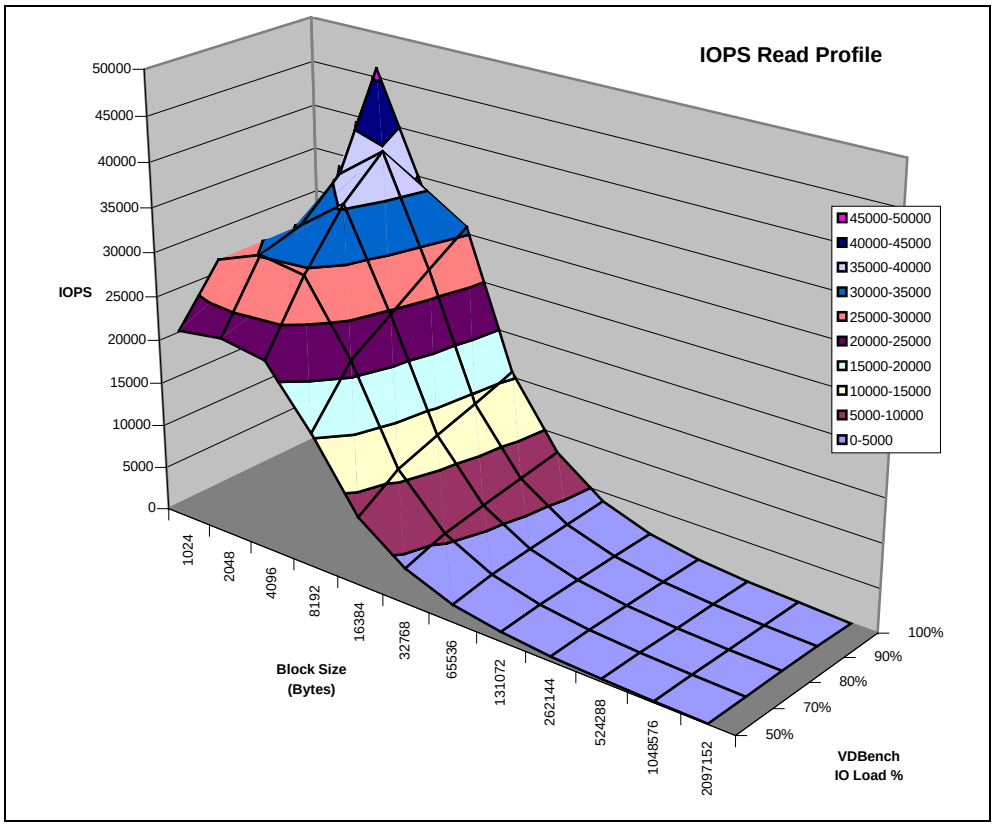

Figure 4 – Four Disk VDBench Read Load Profile

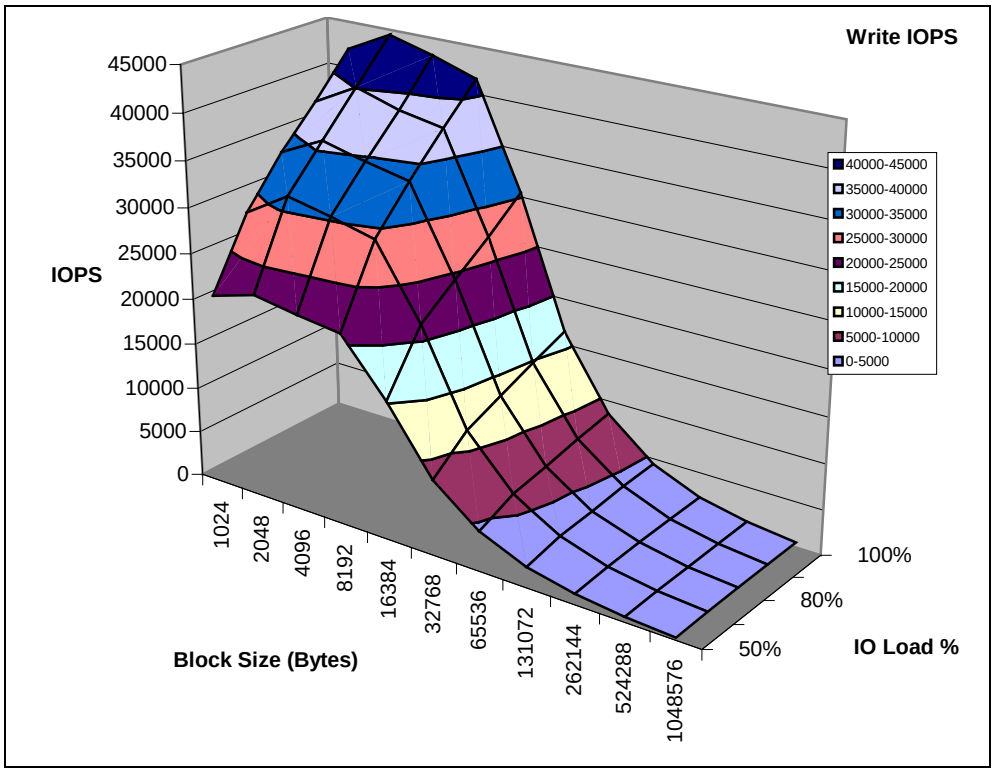

Figure 5 - Four Disk VDBench Write Load Profile

Summary

After extensive evaluation, the TMS RamSan-325 solid state disk system clearly meets the needs of the San Diego Supercomputer Center's SAM-QFS Archival system. Metadata transactions generate heavy IOPS loads that are unsuitable for most traditional RAID storage systems. The RamSan-325's tremendous IOPS capability ensures that the metadata device will not bind the archive and provides performance capacity for the archive's future growth. Given this metadata performance enhancement, San Diego Supercomputer Center will be able to improve data integrity in the file system by conducting more frequent metadata backups. Finally, archive users will enjoy the quick response to metadata queries from the archive, making their interaction with the archive more efficient and productive.

References

[1] "Sun StorEdge™ QFS and Sun StorEdge™ SAM-FS File System Administration Guide," Sun Microsystems, June 2004

[2] Storage Performance Council SPC-1 Benchmark, www.storageperformance.org

[3] A. Scharmer, S. Johnson, "Storage System Bottlenecks and Their Solutions," Sun Microsystems, May 2005

[4] "Sun StorEdge™ QFS and Sun StorEdge SAM-FS Software Installation and Configuration Guide," Sun Microsystems, June 2004

[5] "Solaris RamSan Benchmarking and Tuning Guide," Texas Memory Systems, 2004