An Oracle White Paper
June 2010

DELIVERING HIGH BANDWIDTH WITH SUN QFS SOFTWARE

Demands of High-Bandwidth Applications

For years digital data exchange simply involved sending small amounts of data, such as email, source code, or HTML pages. Today, the continuing network revolution has users distributed throughout the world, with data shared from Boston to Beijing and replicated across hundreds of locations and manipulated in real time around the clock. In addition, users are accessing a wide range of services that must move large amounts of data quickly.

For example, new internet services deliver movies in high definition format to tens of thousands of households concurrently. No longer limited to a few highly specialized scientific applications, high performance computing now generates immense data sets that must be efficiently shared among thousands of users. Databases are distributed geographically across widely separated datacenters to improve reliability, availability, and performance, and to preserve business continuity in the event of disaster. These demanding new requirements complicate system architectures, networks, and datacenter management.

Fast Access to Data

High-bandwidth applications have raised expectations. Users want to be able to take advantage of applications at any time and from any location. As a result, applications cannot wait hours or days to access information. Data must be available when needed, and with little latency, if demanding application needs are to be met.

Reliable, Available, and Survivable Data Storage

Businesses depend on data to make discoveries faster, deliver video to subscribers, complete transactions, and more. When systems are unavailable or fail, revenue often is lost. Business survival can be threatened if a catastrophe interrupts service and destroys information. Data must be stored redundantly, and replicated sufficiently, to preserve business continuity for every conceivable contingency.

High Concurrency With High Throughput

Streaming audio and video, real-time data streams such as Short Message Service (SMS), social networking, and voice over IP (VoIP), and download-based commerce all depend on the ability to service thousands of concurrent sessions with high throughput. In order for a new music release, video stream, or application program to be obtained when needed, data must be available to many consumers simultaneously.

Simplified Management

Thirty years ago, data processing equipment was expensive and datacenter labor was relatively cheap. Thanks to Moore's Law the opposite is now true. Cost-effective hardware and skilled labor are the norm. With labor costs dominating IT budgets, finding ways to simplify management and gain efficiency is key to saving money and enabling IT departments to concentrate on delivering quality of service.

Manage Data Independent Of Physical Device Constraints

Datacenters have access to broader data storage options than ever before. All kinds of devices, from flash technology-based storage to high-capacity disk and super-capacity tape drives, are connected through a myriad of networks. Even though data is stored on this broad range of different physical devices from a wide range of vendors, fast, transparent data access is essential to effective system and business operation.

The Sun QFS Shared File System

Sun QFS software is a shared SAN file system designed to solve file system performance bottlenecks by leveraging underlying disk technology and hardware. The high-performance, configurable Sun QFS shared file system helps to overcome traditional UNIX® file system shortcomings, such as lengthy file system checks after an unintended interruption, long file system generation times, and limitations in file system scaling due to a finite number and size of files. Key performance-related characteristics of the Sun QFS shared file system are summarized in the following sections.

- **Reduced latency and increased throughput.** The Sun QFS shared file system separates the logical structuring of data from physical volume constraints to make more efficient use of physical devices. Metadata that describes files can be maintained separately from files' data in order to eliminate disk contention and increase throughput. This file data capability can be combined with flash technology-based devices for even greater performance benefit.
- **Workload optimization**. Sun QFS software allocates file extents in Disk Allocation Units (DAUs), the file system equivalent of disk blocks. Variable-length DAUs allow substantially better matching of the workload to the file system for optimum performance.
- **Simplified management**. Sun QFS software includes embedded volume management functions, including disk striping and disk allocation. Eliminating the need for a separate layer of volume management software, Sun QFS software simplifies system administration while making efficient use of physical storage hardware for reliable, available systems.

- **Non-disruptive file system management.** Sun QFS software allows file systems to be created, tuned, expanded, or reduced on running systems, removing one of the complications that makes high availability difficult to achieve.
- **Support for high-speed direct I/O.** Historically, file systems imposed real performance penalties compared to the raw performance delivered by physical devices. The Sun QFS shared file system provides a separate direct I/O data path that allows a properly configured system to do high-throughput I/O. This permits Sun QFS to deliver I/O throughput at a rate close to the physical hardware.
- **Simultaneous access.** The Q-Write feature of the Sun QFS software enables simultaneous reads and writes to the same file from different threads by disabling the POSIX write-lock mechanism, removing a bottleneck common to conventional file systems. This benefits databases and multithreaded applications by allowing multiple execution threads to update a single file in parallel. Applications using the Q-Write feature must be designed to control multiple threads writing to the same file to ensure data integrity is not compromised.
- **Shared access.** Fibre Channel technology increases connectivity by allowing multiple servers to connect to the same disk drive or subsystem through multiported disk arrays or Fibre Channel switches. While UNIX file systems can share the same hardware by splitting up disk subsystems, multiple servers cannot share the same physical disk area. Files on these disks can only be shared via methods such as NFS or FTP. With Sun QFS software functionality, any Sun QFS file system can be configured to be mounted directly on multiple systems simultaneously and shared by several servers running the Oracle Solaris operating system. One system is the metadata server for a file system and all other systems mount the same file system as clients.

Streaming Performance in Sun QFS Software

One of the strengths of the Sun QFS shared file system is supporting high-capacity I/O for large files and very large databases, required, for example, by video streaming applications.

- **Data throughput.** Systems must deliver data to each user fast enough to maintain uninterrupted high-quality video at the user's TV, computer, or mobile device. High-definition video requires delivering 8 MB/sec for each data stream.
- **Isochrony.** *Isochrony* is the property of data delivery with a low variance in the inter-arrival time of data blocks or packets — that is, data is delivered at an even, predictable rate. Poor isochrony can be compensated to some extent by keeping larger data buffers at the receiving system. Doing so increases the price per set-top device. In an on-demand video system, that small incremental price increase must be multiplied by thousands or millions of set-top devices. Traditional UNIX file systems have great difficulty with these sorts of streaming loads, because data tends over time to become fragmented and distributed randomly over the

physical medium. In contrast, the Sun QFS shared file system is well-suited for intensive streaming workloads.

Streaming Benchmarks

Oracle performed a series of benchmark tests to confirm the performance of the Sun QFS shared file system under streaming workloads. The experimental system configuration is shown in Figure 1 and detailed in Table 1. A total of 17 Sun Fire X4150 servers from Oracle were used, with one server storing the QFS file system metadata, and the remaining 16 servers acting as data hosts. All Sun QFS servers were connected to eight Sun StorageTek 6140 disk arrays from Oracle via Fibre Channel connections. Each disk array was configured as two logical units, providing 16 logical units, or one per host. Test runs were performed using a simple shell script that was launched on each Sun QFS server by an `rsh(1)` invocation from a separate workstation.

Each experiment was repeated with one to 16 hosts, with throughput measured for each run and an aggregate total computed. Each test run transferred 1,496 MB of data in 8 MB blocks. The results show scalability from 1 to 16 hosts for each specific experimental configuration.

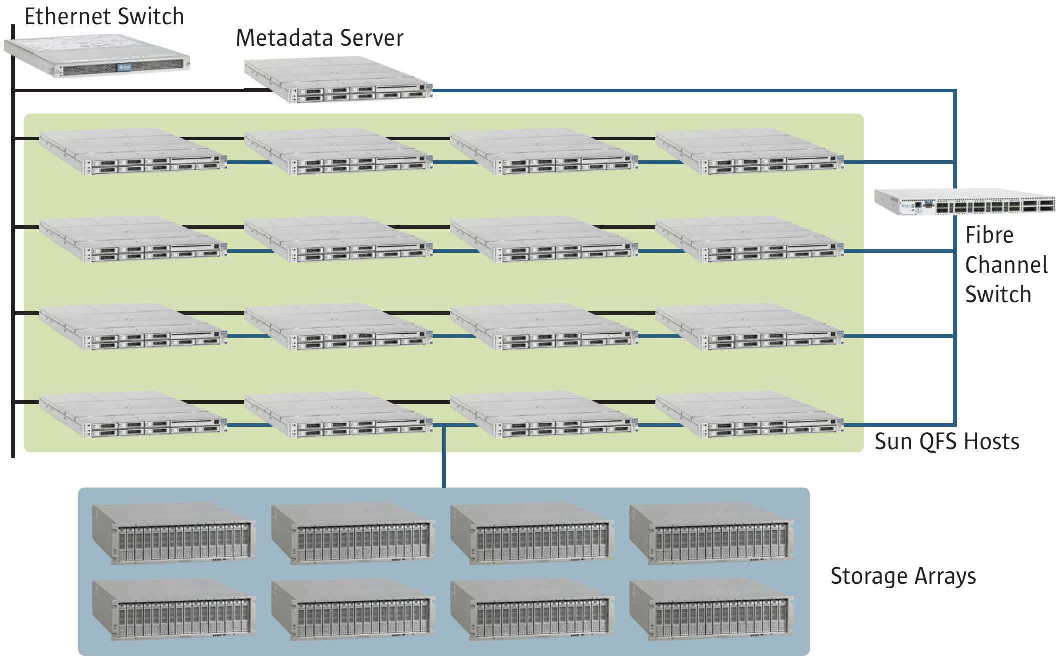

Figure 1. Sixteen QFS hosts and one QFS metadata server were configured with eight Sun StorageTek 6140 disk arrays in the test system.

Test runs were performed that varied the number of servers used in each run. In each test case, each simulated user session transferred 1,496 MB of data as 187 8MB blocks. The number of user sessions and servers were varied among test runs in order to test system scalability.

TABLE 1.TEST SYSTEM CONFIGURATION

	SYSTEMS AND COMPONENTS	CONFIGURATION
SERVERS	• Sun Fire X4150 servers (17)	• Four CPUs running at 3.3 GHz • 4 GB memory
STORAGE	• Sun StorageTek 6140 arrays (8) with firmware revision 7.50.13.10	• RAID-5, 8 drives with 7 data drives, 1 parity drive • 2 LUNs per disk array, 16 LUNs total • 512 KB segment size • Read-ahead enabled • Write cache enabled • Write cache with replication disabled
NETWORKING	• Qlogic SANbox 9000 Stackable Chassis Switch (1) with firmware revision 6.2.1.3.0 • Sun StorageTek 4 Gbit Fibre Channel PCIe HBA (1) • Extreme Networks Summit Z450a 48t switch (1)	• 64 ports • 4 Gbit/sec Fibre Channel switch • 1 port • 4 Gbit/sec Fibre Channel HBAs per host, located in slot 2 • 48 ports • 1 Gbit/sec Ethernet network interface

Comparing Streaming Performance with Raw I/O

A system architect must balance two apparently incompatible needs in a streaming video system: effective management of extremely large data sets with highly concurrent, complex workloads, and stringent quality of service (QoS) requirements with high throughput. Streaming applications in particular live and die on consistent high throughput while transferring large files. This test compared the data throughput for Sun QFS software with data throughput in a configuration that writes to and reads from the physical device directly using standard write and read system calls.

Comparison measurements were taken with data transferred to one logical disk on each server, varying the number of servers from 1 to 16. As Figure 2 shows, throughput in deployments utilizing Sun QFS software scales linearly as the number of paired hosts and logical units increases. Of particular interest is the fact that Sun QFS software throughput is effectively indistinguishable from raw device throughput. These results indicate that the advanced file system services provided by the Sun QFS software have little or no impact on performance. The only exception is initial writes. In this case extents must be allocated. Write scalability is nearly linear, although somewhat slower than raw writes. This test was performed with a default disk allocation unit (DAU) size of 64 KB. The implications of differing DAU sizes are explored in the next section.

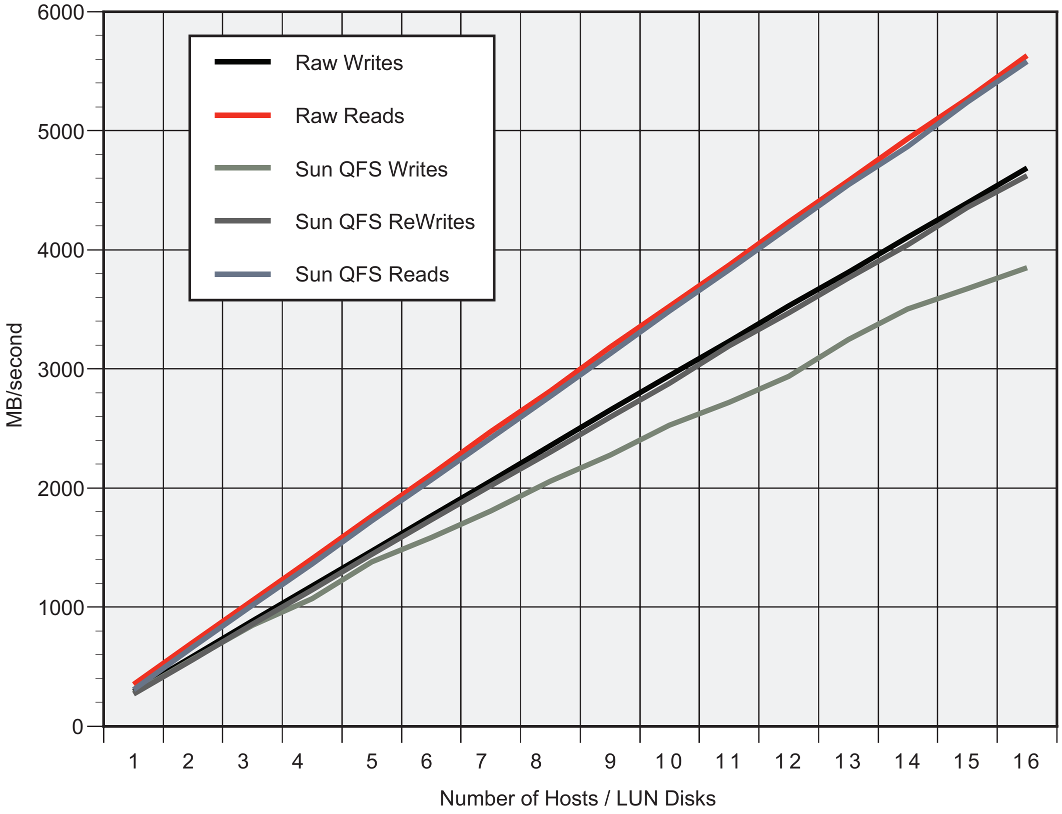

Figure 2. Sun QFS streaming throughput scales linearly and is effectively identical to raw throughput.

Effect of Increasing the Disk Allocation Unit Size

Sun QFS software allocates files in units of a DAU. A DAU is a logical block size managed by the Sun QFS software independent of the physical block size of the storage hardware. By choosing different DAU sizes, system administrators can tune the performance of the Sun QFS shared file system to differing workloads. By default the DAU size is 64 KB. By setting the DAU size to 8 MB, the additional load associated with updating metadata and allocating extents is decreased since the number of metadata updates and allocations is reduced by a factor of 128. As a result, initial write performance is increased and closely approximates raw device writes, as shown in Figure 3.

Changing the DAU size is a simple method of tuning the system to a different workload. This example demonstrates that even simple tuning can make a significant improvement in the performance of the Sun QFS file system.

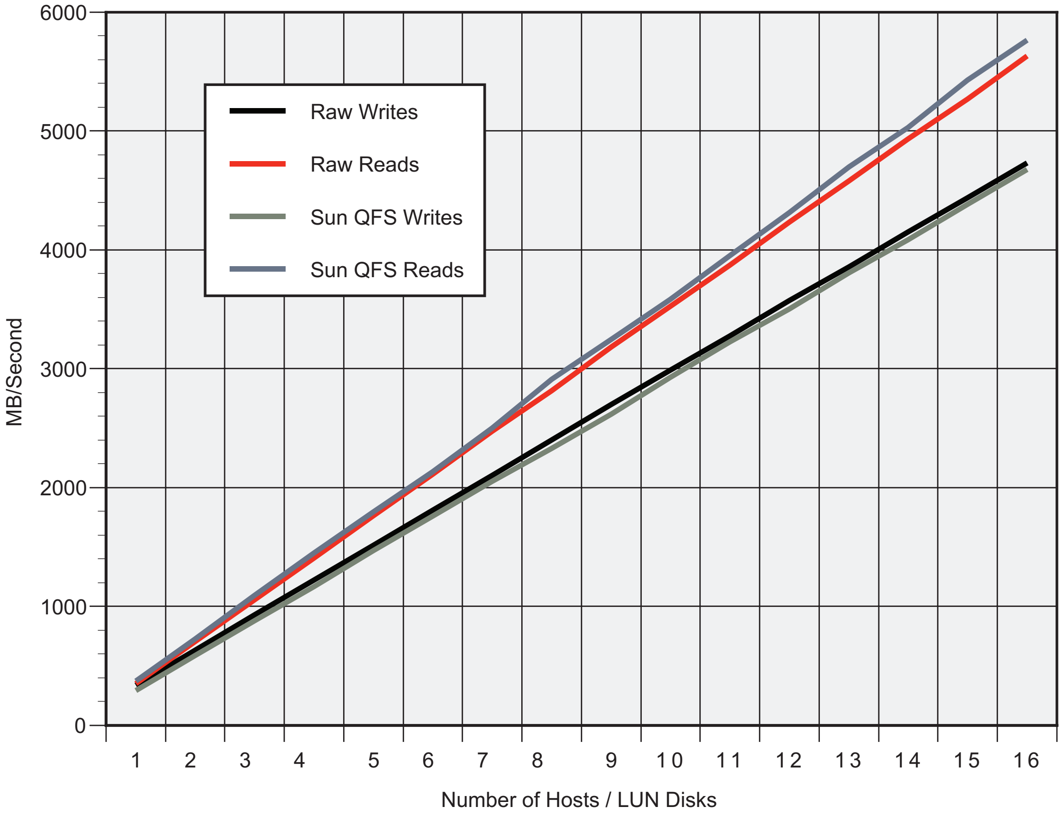

Figure 3. The effect of increasing DAU size is to increase write performance until it closely approximates raw I/O performance.

Effect of Sun QFS Buffer Caching

Sun QFS software can achieve even higher throughput rates in workloads where data is reused by using the QFS buffer cache. Doing so allows file data to be cached in system memory, reducing the number of times physical disk I/O must be performed to satisfy a request. As shown in Figure 4, the effect of using the buffer cache in appropriate workloads can be quite dramatic. Data throughput can be increased by a factor of five while preserving linear scalability across the number of hosts.

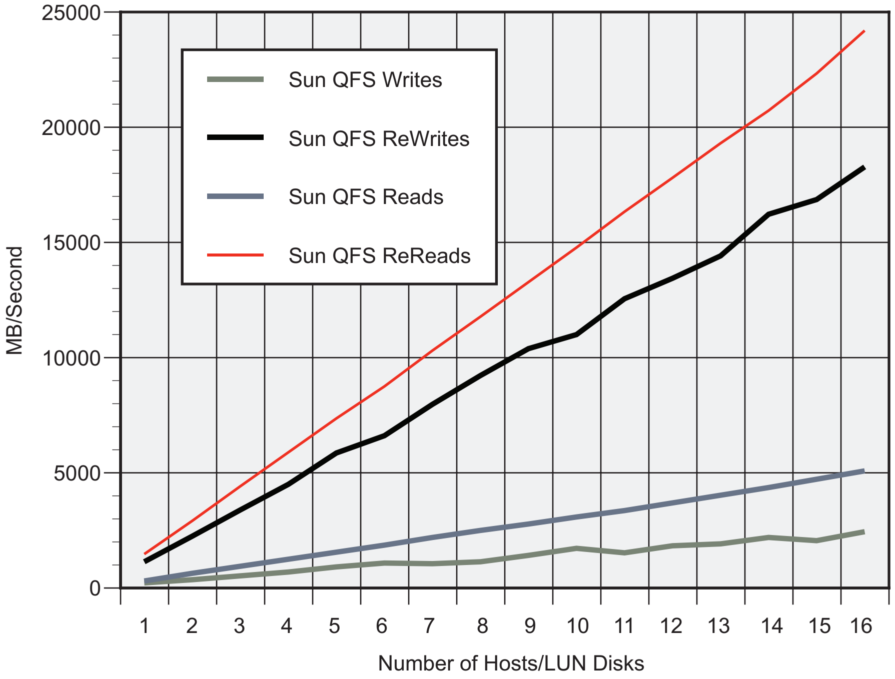

Figure 4. Using buffer caching when data is reused increases throughput dramatically.

Real-World Experiences with Sun QFS Software

The Sun QFS shared file system is used in many large-scale applications, particularly those with large databases and file systems or large files that must be shared by many users.

Vancouver Olympics: 10,000 Media Outlets and No Down Time

Every four years, a new media giant is built, runs for just two weeks, and ends — the Olympic Games. During those two weeks, the IT infrastructure of the Games must provide content to 10,000 media outlets, reaching an audience of three billion people, with no down time and no interruptions.

The Vancouver Olympic Committee (VANOC) chose Oracle platforms and the Sun QFS shared file system on which to build this transitory media empire. VANOC implemented the full IT infrastructure in two years, and has already conducted more than $100 million in online sales. The Olympic Games began in February 2010, with all essential services at 15 Games locations, from ticketing to real-time results, using the Sun QFS software. By using Oracle systems and software, VANOC was able to meet a very aggressive deployment schedule, sustain extreme performance demands, and did so with substantial energy savings. For more information, please see: http://www.sun.com/featured-articles/2009-1121/feature/index.jsp

Major League Baseball: Keeping the Fans Happy

Major League Baseball has had an internet presence for a long time. That presence has grown to include MLB.com, as well as sites for 30 major league clubs, minor league clubs, the Major League Baseball Player Association, and the National Baseball Hall of Fame and Museum. The site also provides real-time streaming video and audio, along with box scores and statistics for nearly 2,500 games per year, and experiences nearly a billion annual visitors.

A very large part of this internet presence is storing, managing, and archiving digital audio and video content. Major League Baseball maintains two datacenters with more than 100 Sun Fire servers along with Sun StorageTek disk and tape systems that share and stream the data from a Sun QFS shared file system.

Using Oracle platforms and the Sun QFS shared file system, Major League Baseball has been able to offer up to 15 live games daily and more than 6,000 audio-streamed games during the season. It has delivered more than one billion minutes of streaming media and over 2,430 full-length games to over one billion visitors, with no application downtime in two years of operation. For more information, please see:
http://www.sun.com/customers/service/mlbam.xml

HBO: Five Nines Availability for HD Video Delivery

Media outlets are experiencing radical changes as they transition from analog to digital audio and video. A major provider of video content, HBO has two video networks, delivers video on demand to thousands of cable systems, and produces hundreds of hours of original programming every year. With all this content, HBO was drowning in analog video tapes. In 2003, HBO decided to move all their standard and high-definition video to a fully digital repository based on Oracle platforms.

HBO replaced their video tape libraries with Sun Fire servers and Sun StorageTek 9990 storage arrays, storing the data using the Sun QFS shared file system. Doing so eliminated over 80% of HBO's video tape equipment and reduced maintenance and labor costs, while delivering HBO's global content with 0.99999 availability. For further information, see:
http://www.sun.com/customers/pdf/hbo.pdf

For More Information

To read more about the Sun QFS shared file system, see the URLs listed below.

TABLE 2.WEB SITES FOR MORE INFORMATION.

DESCRIPTION	URL
SUN QFS SOFTWARE	http://sun.com/storage/management_software/data_management/qfs/
SUN STORAGE ARCHIVE MANAGER AND QFS FILE SYSTEM INFORMATION CENTER	http://wikis.sun.com/display/SAMQFSDocs/Home
ORACLE STORAGE WHITE PAPERS	http://sun.com/storage/white-papers/
SUN QFS SHARED FILE SYSTEM SUN STORAGE ARCHIVE MANAGER	http://www.sun.com/storage/white-papers/qfs-samfs.pdf

Delivering High Bandwidth
With Sun QFS Software
June 2010

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

This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.