



Tiered Storage and Sun Java™ System Messaging Server

White Paper
June 2009

Abstract

Service providers face unrelenting pressure to increase storage for user mailboxes. This paper explains how using production products, such as Sun Communications Suite software, the Solaris ZFS™ file system, and a Sun Storage array, can cut storage costs by half, while performance (measured by messages per second) can be increased by more than 50 percent in environments with millions of users. Extensive load testing has shown that this single-rack solution provides a cost-effective message store for two million users. It is easily replicated and highly modular.

Table of Contents

Introduction—The Market Environment	3
Goals and Requirements	4
Performance.....	4
Scalability	5
Cost	5
Backup	5
Availability	5
Java System Messaging Server Message Store.....	6
The Message Store.....	6
User Files	7
System Files.....	7
Solaris ZFS File System	8
Sun StorageTek™ Storage Array.....	9
Message Store Reference Solution	9
Future Directions	11
Summary.....	12
More Information	12
Sun Java Communication Suite	12
Message Store Architecture:	12
Sun StorageTek Storage Array	12
Solaris ZFS File System	12

Introduction—The Market Environment

Service providers and other organizations that offer e-mail to large numbers of users are facing enormous pressure to increase storage, and for many reasons. It is a pressing challenge for large enterprises, universities, or any organization with a large number of e-mail users. The ongoing growth in both volume and size of e-mail messages is staggering. In the U. S. alone, more than 200 million messages are sent per day, and that number is expected to double within four years.¹ It is estimated that the daily average size of e-mail sent and received by an average corporate user will increase from 16.4 megabytes (MB) in 2006 to 21.4 MB in 2010—an increase of 25 percent over four years.² In addition, attachment size has increased significantly, as people share documents, presentations, pictures, videos, and more.

A further reason driving demand for increased storage is that users demand access to their e-mail from more than one device or location, through services like Web mail and IMAP. Also, corporate retention policies and government regulations (such as Sarbanes-Oxley) are encouraging organizations to keep e-mail for longer periods of time.

Finally, there is significant competitive pressure. One major e-mail service grew from a 4 MB quota in 1997 to 100 MB in 2004, 1 GB in 2005, and ultimately to unlimited free mailbox storage in 2007. In contrast to traditional Internet service provider (ISP) e-mail policies, Web service providers Google, Yahoo, and Windows Live offer e-mail storage and retention that is effectively unlimited.

With significant upward demand on e-mail storage in the foreseeable future, straight-line scaling of storage infrastructures is economically unfeasible. Large e-mail operations, including service providers, universities, businesses, and governments, must find ways to reduce the cost per gigabyte of short- and long-term storage.

¹ <http://www.radicati.com/?p=1484> "Companies finding ways to cope with large volumes of electronic messages," The Radicati Group, August 23, 2008

² <http://windowsitpro.com/whitepapers/index.cfm?fuseaction=ShowWP&wpid=50e5f2e5-1539-4211-97d4-846fbbd3aaa4&feed=rss> "Understand the Business Reasons for Email Storage Management and Taking Control of Your Email," Windows IT Pro, 2009

Goals and Requirements

When creating a cost-effective message store, there are several goals. The solution must be able to handle a large-scale production workload, at a price per gigabyte that is significantly less than existing solutions, while leaving room for future growth. Also, the solution must be highly available and able to be backed up in hours, not days.

These goals are especially challenging because message stores hold large amounts of highly available data with demanding performance requirements. E-mail is a very I/O-intensive application, where transaction speed is critical to customer satisfaction. When increasing storage capacity, service providers must ensure that e-mail services maintain acceptable performance levels under all load levels and conditions, such as message I/O, backup, and replication.

For message stores, a key measure of storage performance is Input-Output Operations per Second (IOPS). Message stores are among the highest IOPS applications that exist, because a single message store can generate 15 or more IOPS per message delivered. These are typically many small, random writes. Most performance issues with message stores arise from insufficient disk I/O capacity. For this reason, customers have kept all portions of message stores on their highest-performing, most expensive disks.

Traditionally, e-mail message stores meet performance demands by using high-end fibre channel (FC) and 15K RPM drives—typically among the most expensive storage components—because this is architecturally simpler and less risky. Using fast drives and interfaces minimizes associated bottlenecks. However, this is not a cost-effective approach, especially when increasing mailbox quotas by as much as 100 times.

Any message store design must meet minimum criteria in performance, cost, backup, and availability when provisioned as a large-scale service provider deployment that includes millions of subscribers and several terabytes of existing message data.

Performance

While several elements comprise message store performance, messages per second is a reliable, top-level metric. This includes writes, reads, and deletions. While many service providers see sustained throughput at less than 400 messages per second, more headroom is always a desirable goal. Note that improving the messages-per-second rate can also reduce the number of required servers and storage, lowering overall costs.

Scalability

The solution must be viable for two million users, and be easily replicatable to expand beyond that.

Cost

Key contributors to message store costs include high-performance architectures based on FC technologies. When combined with volume management and cluster software licensing fees, total storage costs can approach \$5.00 per gigabyte. Any solution must be able to handle the current production workload as well as a future, larger, production workload, at a price point that is significantly lower than the current implementation.

Backup

To protect subscriber data, message stores must be routinely, simultaneously, and completely backed up with a production workload, in 12 hours or less.

Availability

Service continuity and rapid recovery are essential to a service provider's e-mail service. Any message store solution must include failover capabilities that provide graceful degradation in the presence of hardware or software faults, while allowing continued operation.

Java System Messaging Server Message Store

A team of engineers from Sun's storage and messaging server groups designed, built, and extensively tested a message store configuration that exceeds cost, performance, backup, and availability goals, as outlined previously. The solution is modular, in that it fits within a single rack, and multiple racks can be used to accommodate large numbers of users. Many factors contribute to this solution, including:

- Sun Java™ System Messaging Server uses a highly modular design, enabling those functions that require high-performance to run on separate subsystems from those that do not. The solution enables a mix of lower-cost Serial Advanced Technology Attachment (SATA) and high-performance FC storage technologies. In addition, the fact that it is a single-copy message store means there is only one copy of any e-mail message stored, no matter how many users store the same message in their inboxes.
- Solaris ZFS file system delivers several cost and performance benefits. It also provides comprehensive volume management, availability, and backup functionality without licensing fees.
- A Sun storage array was chosen because it is designed to handle large data sets and delivers market-leading performance for primary storage requirements. Service providers can mix SATA and FC drives to design cost-effective tiered storage environments with centralized management.

The Java System Messaging Server message store configuration employs these technologies to provide a highly available storage environment that increases performance by 50 percent over previous implementations, while at the same time reducing cost per gigabyte by the same number. See Table 1.

	Capacity	IOPS	Cost/IOPS	Cost/GB
SATA 7.2K	1,000 GB	100	\$22.00	\$2.20
FC 15K	300 GB	300	\$8.50	\$8.50

Table 1: SATA drives offer lower performance at much lower cost per GB.

The Message Store

A dedicated data store, the message store in the Sun Java System Messaging Server enables the delivery, retrieval, and manipulation of Internet mail messages. It works with Post Office Protocol Version 3 (POP3) and Internet Messaging Access Protocol 4 (IMAP4) client access servers to provide flexible and easy access to messaging. Connected to a Web mail server, it also provides messaging capabilities to Sun Convergence and Sun Java System Communications Express (also part of the Java Communications Suite) in a Web browser.

The message store uses a hybrid design that combines an indexed database to store message header information, and flat text files to store message content. Each of these components has a separate file structure, so they can be kept on different storage subsystems. This hybrid design creates a high-performance, highly scalable message store that is robust and efficient. Two general areas in the Java System Messaging Server's message store handle user files and system files.

User Files

A single-copy message store maintains only one copy of each message per partition. When it receives a message addressed to multiple users, a group, or distribution list, it adds a reference to the message in each user's inbox. This avoids saving a unique copy of the message for each user, unnecessarily duplicating data. (Note that individual message status flags, such as seen, read, answered, deleted, and so on, are maintained in each user's folder.)

This single-copy approach helps contain the message store's overall size. Even as the average size of both messages and attachments continues to increase, only a single copy is stored, no matter how many subscribers have it in their inboxes.

User files in the message store are ideally suited for lower cost, higher capacity SATA (7.2K RPM) drives because:

- They place less intensive I/O demands on the message store file system. The majority of messages are read once, perhaps filed, and never accessed again.
- They typically require significantly more storage than the message store database and indexes.

System Files

Information on the entire message store is kept in a Berkeley database format for faster access. In addition to the messages themselves, the message store maintains an index, a cache of message header information, and other frequently used data to enable rapid retrieval of mailbox information by clients. The database and indexes require fast storage response times.

- System files in the message store are best suited for high-performance (15K) FC drives.
- Typically, the storage capacity requirements for system files are many times smaller than for user files.

Note that complete information on the Java System Messaging Server's message store is available online.

Solaris ZFS™ File System

A key enabling technology of this message store configuration is the Solaris zetta-byte file system (ZFS) file system. ZFS presents a pooled storage model that completely eliminates the concept of volumes and the associated problems of partitions, provisioning, wasted bandwidth, and stranded storage. Solaris ZFS provides many benefits, including superior data integrity, performance, and administration/volume management. The Solaris ZFS file system is used on Sun Open Storage products, including Sun Fire 4500 and Sun Fire 4540 servers and Sun Storage 7000 Unified Storage Systems.

All ZFS operations are copy-on-write transactions, so the on-disk state is always valid. Every block is check-summed to prevent silent data corruption, and the data is self-healing in replicated (mirrored or RAID) configurations. If one copy is damaged, ZFS detects it and uses another copy to repair it.

Also, ZFS enables the use of inexpensive disks because it provides disk scrubbing. Like error correcting code (ECC) memory scrubbing, the idea is to read all data to detect latent errors while they are still correctable. A scrub traverses the entire storage pool to read every copy of every block, validate it against a 256-bit checksum, and repair it if necessary, all while the storage pool is live and in use.

To save space, ZFS provides unlimited constant-time snapshots and clones. A snapshot is a point-in-time copy of a file system, while a clone is a writable copy of a snapshot. A clone is an extremely space-efficient way to store many copies of mostly-shared data such as workspaces, software installations, and diskless clients.

ZFS backup and restore are powered by snapshots. Any snapshot can generate a full backup, and any pair of snapshots can generate an incremental backup. Incremental backups are so efficient that they can be used for remote replication; for example, to transmit an incremental update every 10 seconds. For the this message store configuration, ZFS compression is used in the backup subsystem.

There are no arbitrary limits in ZFS. Users can store as many files as they want; full 64-bit file offsets, unlimited links, directory entries, snapshots, and so on.

As a way to enhance performance, ZFS provides built-in compression. In addition to reducing space usage by two to three times, compression also reduces the amount of I/O by the same ratio. For this reason, enabling compression actually makes some workloads go faster.

Complete information on the Solaris ZFS file system is available online.

Sun StorageTek™ Storage Array

Sun StorageTek arrays are an ideal component for the Java System Messaging Server message store, because they offer performance, expandability, availability, and flexibility. The solution requires a storage array that can offer both high-performance drives, and low-cost, high-capacity drives within a single rack.

Sun StorageTek arrays allow the intermixing of FC and SATA drives to create a tiered storage environment that matches storage cost to application and data performance requirements. Expansion trays can be used with FC drives for high-performance data, while moving less-demanding data to less-expensive expansion trays loaded with SATA drives.

Sun StorageTek arrays include many availability and reliability features. The arrays offer non-disruptive addition of capacity and volumes, RAID and segment size migration, plus switched technology with point-to-point connections. All components, from disk drives to midplane, are hot swappable. Using hot spares in every storage tray of the StorageTek array supports high availability. Every array controller has two power supplies, each with its own battery backup system providing redundant power for continuous uptime.

Message Store Reference Solution

Deploying a high-performance message store requires significant tuning and optimization of hardware and software systems. The engineering team created this configuration by modelling it after a national service provider's environment, including provisioning for millions of users and tens of millions of stored messages per store. Performance loads representing a typical day were applied, including:

- Millions of reads, writes, and deletes.
- Backups completed in 12 hours or less with hourly snapshots, while the message store was under load.
- “5-nines” uptime.

Using well-known, highly proven hardware and software components, the message store was designed, configured, thoroughly tested, and refined to fit on a single data center rack. (See Figure 1.) Working with full-scale parameters, the solution was tuned to exceed price, performance, availability, and backup goals on a storage platform containing low-cost SATA drives and a smaller number of high-performance FC drives—all running the Solaris ZFS file system.

Different Message Store components have separate file structures.

Inexpensive SATA drives used for majority of storage requirements: messages, attachments, and backup. As average message size increases, storage happens here.

High-performance FC drives used for small more demanding database and indexes.

Entire Message Store managed by ZFS file system.

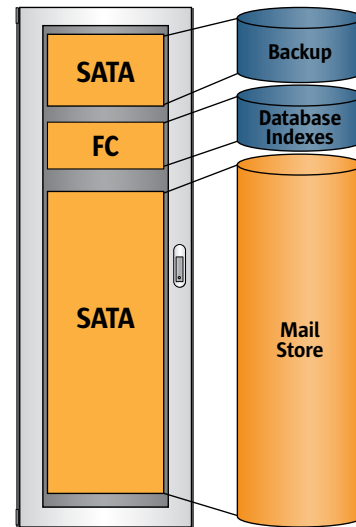


Figure 1: A high-performance Java System Messaging Server message store is a modular, single-rack solution that improves throughput while dramatically reducing storage costs.

Cost per gigabyte was reduced by 50 percent:

- Far fewer FC drives are needed. SATA drives cost less per gigabyte to buy. Because they spin at slower speeds, they use less electricity and are therefore less expensive to operate.
- Single-copy message store means that only one copy of the message is stored, regardless of the number of recipients.
- Licensing fees for clustering and volume management software are eliminated.
- ZFS compression reduces backup storage requirements.

Performance was increased by 50 percent:

- Overall design much more performant than traditional, FC-based solution.
- Performance improvements were the result of tuning, ZFS, and the StorageTek 6540 server.

Backup performance:

- Accomplished within 12-hour window, with hourly snapshots, under full load.
- ZFS compression reduced network traffic associated with backups.

Future Directions

While the Java System Messaging Server has successfully met its goals of a 50-percent reduction in storage costs and a 50-percent improvement in messages-per-second throughput, the team believes there is significant room for improvement. Much of this belief relies on the Solaris ZFS file system. Among its many capabilities, ZFS can transparently optimize the use of different types of storage devices. This has been demonstrated here—both SATA and FC drives are used in a solution that is less expensive and more performant. Solid-state drives (SSDs) based on flash technology are now coming to market. While these drives are more expensive per gigabyte, they offer tens or hundreds of times the IOPS performance of even the fastest hard disk drives. SSDs offer the potential to raise the performance level even more, while simultaneously lowering the overall cost of deploying a large messaging store.

Summary

By taking advantage of the inherent capabilities of a Sun StorageTek storage array, Java System Messaging Server, and Solaris ZFS file system, a single-rack solution can be provisioned that meets service level requirements for millions of users at dramatically lower costs. The Java System Messaging Server's messaging store handily exceeds performance requirements, doing so while simultaneously protecting user data through hourly snapshots and backups that are completed within 12-hour windows. The overall design provides a foundation to further reduce costs by leveraging new products and technologies from Sun and other industry providers.

Detailed information about the Java System Messaging Server message store is available by contacting your Sun representative.

More Information

Sun Java Communication Suite

sun.com/comms

wikis.sun.com/display/CommSuite/Messaging+Server+and+Tiered+Storage+Overview

wikis.sun.com/display/CommSuite/Messaging+Server+Message+Store+on+a+Sun+StorageTek+6540+Array+Case+Study (internal)

Message Store Architecture:

wikis.sun.com/pages/viewpage.action?pageId=52727016

Sun StorageTek Storage Array

www.sun.com/storage/disk_systems/

Solaris ZFS File System

www.sun.com/software/solaris/zfs_learning_center.jsp



Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA Phone 1-650-960-1300 or 1-800-555-9SUN (9786) Web sun.com

© 2009 Sun Microsystems, Inc. All rights reserved. Sun, Sun Microsystems, the Sun logo, Solaris, OpenSolaris, ZFS, and Java are trademarks or registered trademarks of Sun Microsystems, Inc. or its subsidiaries in the United States and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the US and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc. Printed in USA 06/2009