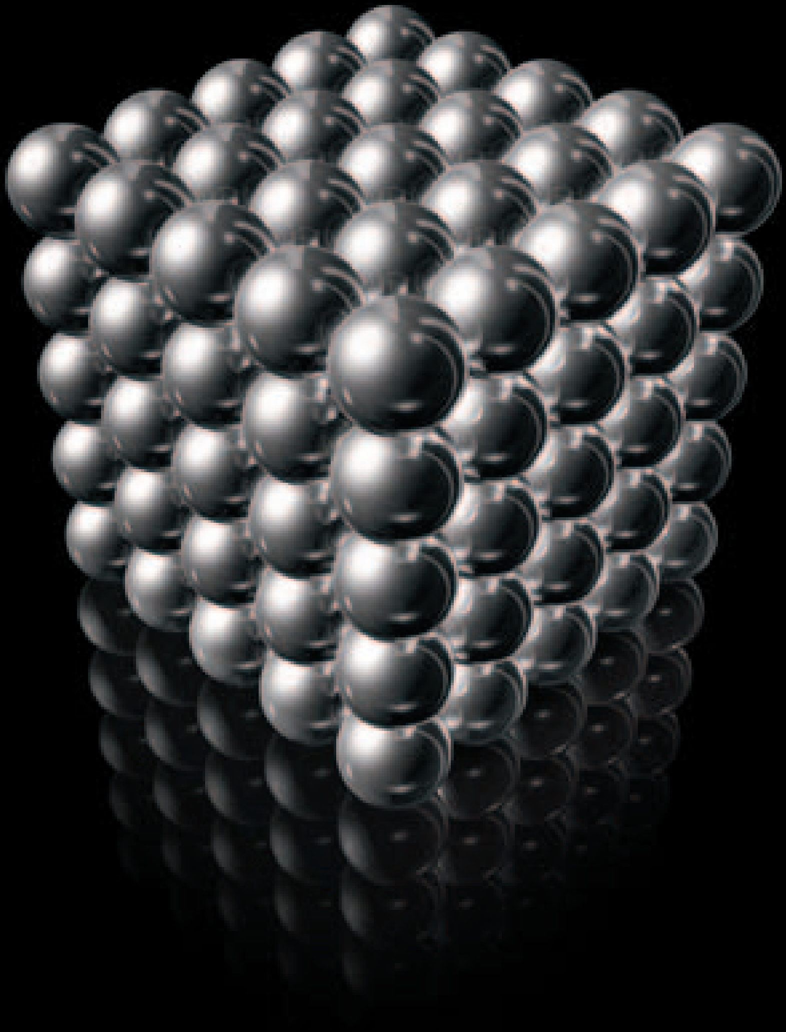
# Xsan Deployment and Tuning Guide

Guidelines and examples for configuring
Xsan storage area network volumes



Apple
1 Infinite Loop
Cupertino, CA 95014-2084
408-996-1010
www.apple.com

Simultaneously published in the United States and Canada.

019-0242/09-02-05

# Contents

# About This Guide

## This guide shows you how to choose the best Xsan setup and configuration options for your users and applications.

The performance and availability of an Xsan volume depend on a variety of parameters, including how you organize available storage, how you configure the SAN's Fibre Channel and Ethernet networks, and how you set basic file system parameters such as block size and stripe breadth. This guide goes beyond the basic instructions in the *Xsan Administrator's Guide* to help you choose options that result in the best performance from your Xsan volumes. The guide includes:

- Guidelines for configuration choices
- Instructions for using the Xsan Tuner application to measure SAN and volume performance
- Sample configurations for network attached storage, high-performance computing, and video production

## Using This Guide

- For general tips and guidelines, see Chapter 1.
- For help using the Xsan Tuner application, see Chapter 2.
- For sample deployments, see Chapter 3.

## For More Information

The *Xsan Administrator's Guide* contains basic instructions for setting up Xsan volumes along with information about managing Xsan volumes, including problem-solving tips and command-line alternatives for common tasks.

You can find the guide:

- On the Xsan Installer disc
- In the folder /Library/Documentation/Xsan on any computer where Xsan is installed
- At www.apple.com/server/documentation

You can also check the Xsan web pages at www.apple.com/xsan.

## Notation Conventions

The following conventions are used in this book wherever shell commands or other command-line items are described.

| Notation | Indicates |
|---|---|
| `monospaced font` | A command or other terminal text |
| `$` | A shell prompt |
| `[text_in_brackets]` | An optional parameter |
| `(one\|other)` | Alternative parameters (type one or the other) |
| `underlined` | A parameter you must replace with a value |
| `[...]` | A parameter that may be repeated |
| `<anglebrackets>` | A displayed value that depends on your SAN configuration |

# Setup and Tuning Guidelines

# 1

## This chapter offers guidelines for making configuration choices that can affect SAN performance.

How quickly SAN clients can transfer data to and from Xsan volumes depends on a variety of factors, including:

- The configuration of and load on the SAN's Ethernet network
- The layout and performance of the SAN's Fibre Channel network
- Settings for the Xserve RAID systems that provide LUNs
- The organization of Xsan volumes and storage pools
- Xsan file system settings

## Setting Up the Ethernet TCP/IP Network

Ethernet connections are used in several ways in an Xsan storage area network:

- Xsan clients and controllers use Ethernet to exchange volume metadata.
- Xsan clients can use Ethernet for access to networks outside the SAN (campus or corporate intranet or the Internet).
- Xsan controllers can use Ethernet connections for remote management.
- Xserve RAID systems can use Ethernet connections for system management.
- Fibre Channel switches can use Ethernet connections for switch management.

You have two basic options:

- Use one Ethernet network for all traffic. This is the less expensive option, but is also less secure and might not provide the best possible performance.
- Use two separate networks; one for metadata and another for all other IP traffic. This configuration is slightly more expensive (requiring two Ethernet adapters for each computer) but offers greater security and better performance because routine network traffic doesn't interfere with SAN volume metadata traffic.

### Set Up a Private Metadata Network

Non-SAN-related Ethernet traffic can interfere with the exchange of metadata among Xsan controllers and clients. For example, using the same connection for both Xsan metadata exchange and Internet access can slow file system performance. Similarly, using the same Ethernet network to connect client computers to directory services and SAN metadata can affect SAN performance.

If SAN performance is critical for your users or applications, keep all extraneous traffic off the network that clients and controllers use to exchange metadata. For best SAN performance, set up a private Ethernet TCP/IP network for the exclusive use of Xsan clients and controllers. For other types of network traffic, including Internet access, Xserve RAID and Fibre Channel switch management, remote SAN management, or directory services, connect each client or controller to a second, private SAN Ethernet network using a second network adapter.

### Use Switches Instead of Hubs

Ethernet switches generally offer better performance than hubs. Use switches, not hubs, in the SAN Ethernet network.

## Setting Up the Fibre Channel Network

Xsan uses Fibre Channel connections to:

- Transfer user data directly between clients and data storage pools
- Transfer metadata between controllers and metadata storage pools

### Verify Base Fibre Channel Performance

Because the devices connected to a Fibre Channel network automatically adjust their speed to match the slowest device on the fabric, it is important to check that all connections in the fabric are operating at 2 GB/s.

**To check Fibre Channel connection performance:**

- Use the management software provided with your Fibre Channel switches to test the performance of your Fibre Channel fabric.

### If Your Fibre Channel Fabric Is Running Slower Than Expected

The following paragraphs list things you can check if your Fibre Channel fabric is not running at the expected 2 GB/s.

#### Check Cables

One faulty cable in a fabric can slow the entire network. Check all cables to make sure they are capable of full transmission speed. Use your switch management software to isolate the faulty cable by checking the performance of specific connections.

### Use Qualified Transceivers in Matching Pairs

Check with the manufacturers of the devices you are connecting to your fabric to be sure that the transceivers (GBICs) you are using are qualified for use with their devices.

Also, use identical transceivers (same manufacturer and model number) on both ends of each cable. Mismatched optical transceivers (even if they are both separately qualified for use with your devices) can cause Fibre Channel communication errors and degrade SAN performance.

### Check Fibre Channel Switch Port Configuration

The Request for State Change Notifications (RSCN) that is generated when a client on the SAN restarts can cause dropped frames in video streams to other clients.

To avoid interrupting SAN traffic to other clients if one client restarts, check your Fibre Channel switch documentation to see if you can configure the switch to suppress RSCNs on initiator ports. (On Qlogic switches, for example, this feature is called I/O StreamGuard.)

### Connect Devices to Specific Blades

If your Fibre Channel switch is based on a blade architecture, you might be able to improve performance by:

- Connecting pairs of devices that routinely exchange large volumes of data to the same blade in the switch
- Distributing loads across multiple blades instead of concentrating all of the load on one or two blades

# Configuring Xserve RAID Systems

Follow these guidelines when you set up your Xserve RAID systems for use as Xsan LUNs.

## Install the Latest Firmware

To be sure you get the best performance and reliability from your Xserve RAID systems, be sure to install the latest available firmware.

**To check for firmware updates:**

- Visit www.apple.com/support/xserve/raid/

## Connecting Xserve RAID Systems to an Ethernet Network

For best performance, don’t connect Xserve RAID controller Ethernet management ports to the SAN’s metadata network. Connect the ports to a separate Ethernet network.

## Choosing RAID Levels for LUNs

Use RAID 1 for metadata LUNs and RAID 5 for data LUNs.

### Use RAID 1 for Metadata LUNs

RAID 1 (mirroring) can give slightly better performance than the default RAID 5 scheme for the small, two-drive metadata LUNs that Xsan uses to store volume information. A single drive is almost always adequate for storing the primary volume metadata (10 GB of metadata space is enough for approximately 10 million files). The second, mirror drive protects you against metadata loss.

### Use RAID 5 for Data LUNs

Xserve RAID systems are optimized for excellent performance and data redundancy using a RAID 5 scheme. (RAID 5 stripes data across the available drives and also distributes parity data across the drives.) Xserve RAID systems ship already configured as RAID 5 LUNs. RAID 0 (striping with no parity) might give slightly better write performance but provides no data recovery protection, so RAID 5 is always a better choice for LUNs used to store user data.

## 7-Drive LUN vs 6-Drive LUN With Hot Spare

For best performance, use full 7-drive RAID 5 LUNs. Even if a drive fails, the degraded 6-drive array will continue to provide excellent performance. Keep in mind, however, that the array is unprotected against the loss (however unlikely) of a second drive until someone replaces the original faulty drive.

If you can't afford to have the LUN operating in a degraded state until someone replaces the faulty drive, you can configure your Xserve RAID systems as 6-drive RAID 5 arrays and use the seventh drive as a hot spare. Data on the faulty drive is reconstructed automatically without human intervention.

## Create One LUN per Xserve RAID Controller

For high performance data sets, create only one LUN on each Xserve RAID controller (one array on each side of the system). Xserve RAID systems ship with one RAID 5 array on each controller.

### Working With LUNs Larger Than 2 Terabytes

The capacity of an Xserve RAID array can exceed 2 terabytes (TB) if the system contains large drive modules. However, Xsan can't use a LUN that is larger than 2 TB. If you set up your Xserve RAID systems as one array per controller, as suggested above, you can't take advantage of the array capacity beyond 2 TB. To use as much available space as possible, you can move drive modules to other controllers or slice a large array into two smaller (less than 2 TB) LUNs. Slicing an array might, however, slow SAN performance.

***Note:*** For the best possible SAN performance, don't slice an array to create multiple LUNs on a single controller.

## Adjusting Xserve RAID Fibre Channel Settings

There are several Xserve RAID settings that can affect the Fibre Channel performance of the device and the SAN as a whole.

### Fibre Channel Speed

Be sure the Fibre Channel connection is set to operate at 2 GB/s.

### Fibre Channel Topology

To add an Xserve RAID system to a Fibre Channel fabric, set the topology to Automatic.

### Disable Hard Loop ID

Don't enable hard loop IDs for Xserve RAID systems in a Fibre Channel fabric.

**To adjust Xserve RAID Fibre Channel settings:**

- Open RAID Admin, choose a system, click Settings, and enter the management password for the system. Then click Fibre Channel.

## Adjusting Xserve RAID Performance Settings

Xserve RAID performance settings, affecting parameters such as drive caching, controller caching, and read prefetching, can have a significant effect on Xsan volume performance. Follow these guidelines.

### Enable Drive Cache

In addition to the caching performed by the Xserve RAID controller, each drive in an array can perform its own caching at the drive level to improve performance.

*Important:* If you enable drive cache for an Xserve RAID set, be sure that the system is connected to a UPS. Otherwise, you could lose cached data if the power fails.

**To enable drive cache for an Xserve RAID array:**

- Open the RAID Admin application, select the RAID system, and click Settings. Then click Performance and enable Drive Cache for the array.

### Enable Controller Write Cache

Without RAID controller write caching, a request to write data to the associated LUN is not considered finished until the data has been completely written to the physical disks that make up the array. Only then can the next write request be processed. (This is sometimes called "write-through caching.")

When the RAID controller write cache is enabled, a request to write data is considered finished as soon as the data is in the cache. This is sometimes called "write-back caching." Write requests are processed more quickly because the file system only needs to write to the fast cache memory and doesn't need to wait for the slower disk drives.

Always be sure to enable write caching on controllers that support metadata storage pools.

Although some large write requests might benefit from caching, often they do not. By placing a volume's metadata storage pool on a controller separate from the data storage pools, you can configure the metadata controller to use caching and the data controller to run without caching.

When the file system is relying on caching in this way, you must guarantee that data in the cache isn't lost before it is actually written to disk. Data that has been written to disk is safe if the power fails, but data in a cache is not. So, to be sure that a power failure can't cause the loss of cached data, protect your Xserve RAID systems with controller backup batteries or an uninterruptable power supply (UPS).

*Important:* If you enable Controller Write Cache on an Xserve RAID system, be sure that the system includes controller backup batteries and, preferably, is connected to a UPS.

**To enable Xserve RAID write cache:**

- Open the RAID Admin application, select the RAID system, and click Settings. Then click Performance and enable Write Cache for each controller.

### Set Read Prefetch to 8 Stripes

Read prefetch is a technique that improves file system read performance in cases where data is being read sequentially, as in the case of audio or video streaming, for example. When read prefetch is enabled, the controller assumes that a read request for a particular block of data will be followed by requests for subsequent, adjacent data blocks. To prepare for these requests, the controller reads not only the requested data, but also the following data, and stores it in cache memory. Then, if the data is actually requested, it is retrieved from the fast cache instead of from the slow disk drives.

Read prefetch is always enabled on Xserve RAID systems, though you can adjust the amount of data that is read. If you're using other RAID systems, check the documentation to find out how to enable read prefetch.

**To adjust the Xserve RAID read prefetch size:**

- Open the RAID Admin application, select the RAID system, and click Settings. Then click Performance and select a Read Prefetch size for each controller.

The default of 8 stripes is best for most applications.

## Estimating Base Xserve RAID Throughput

To estimate how many Xserve RAID systems you need to support specific throughput requirements, you can assume that one Xserve RAID with 14 drives set up as two RAID 5 arrays can handle a minimum of 160 MB of data per second (80 MB/s per RAID controller). This value is applicable to video streaming applications; other applications might achieve higher data rates.

Overall performance is also affected by SAN latency; see "About SAN Write Latency" on page 15.

# Configuring the Xsan File System

The following paragraphs summarize information you should consider when using Xsan Admin to configure your Xsan volumes.

## Organizing LUNs, Storage Pools, and Volumes

As you combine LUNs into storage pools and add storage pools to a volume, try to:

- Keep metadata in a storage pool on a separate RAID controller
- Add an even number of LUNs to each storage pool

### Separate Metadata From User Data

To prevent user data transactions from interfering with metadata transactions, create a separate storage pool for metadata and journal data and assign that pool to a separate LUN and controller.

The recommended metadata LUN consists of only two drives (see "Use RAID 1 for Metadata LUNs" on page 10). To avoid wasting additional drives in the half of the Xserve RAID system that contains the metadata LUN, you can do either of the following:

- Move the spare drives to another system
- Use the drives to create a second LUN where you store files that are seldom accessed

It's also possible to create separate storage pools for metadata and journal data, although Xsan Admin only lets you segregate the two together. To create separate pools for metadata and journal, you must work directly with the configuration file for the volume.

For more information on working directly with the configuration files, see the command-line appendix of the *Xsan Administrator's Guide* or the `cvfs_config` man page. You can also look at the example configuration files in

```
/Library/Filesystems/Xsan/Examples
```

### Set Up an Even Number of LUNs

Storage pools consisting of an even number of LUNs outperform pools consisting of an odd number of LUNs.

## Volume and Storage Pool Settings

To determine the best settings for your Xsan volume, you might need to try several combinations, testing each with the Xsan Tuner application and comparing the results.

### Choosing a Volume Block Allocation Size

In general, smaller file system block sizes are best in cases where there are many small, random reads and writes, as when a volume is used for home directories or general file sharing. In cases such as these, the default 4 KB block size is best.

If, however, the workflow supported by the volume consists mostly of sequential reads or writes, as is the case for audio or video streaming or capture, you can get better performance with a larger block size. Try a 64 KB block size in such cases.

### Choosing a Storage Pool Stripe Breadth

The Mac OS X (or Mac OS X Server) operating system, which handles file data transfers for Xsan, performs 1 megabyte (MB) data transfers. As a result, Xsan gets maximum efficiency from the operating system when it transfers blocks of data that are a multiple of 1 MB.

At the other end of the transfer, the LUN also works well when it receives 1 MB of data at a time. So, when data is written to a storage pool, you get the best performance if 1 MB of data is written to each LUN in the storage pool The amount of data Xsan writes to a LUN is determined by the product of two values you specify when you set up a volume:

- The volume's block allocation size (in kilobytes)
- The stripe breadth of the storage pools that make up the volume (in number of allocation blocks)

transfer size = block size x stripe breadth

For example, the default Xsan block size of 4 KB combines with the default storage pool stripe breadth of 256 blocks to produce a transfer size of 1 MB. If you increase the block size to 64 KB, for example, to suit data streaming, set the stripe breadth to 16 blocks, so the product of the two remains 1 MB.

### Choosing a Storage Pool Multipath Method

You can increase performance by attaching each client to the SAN using two Fibre Channel cables. Xsan can take advantage of multiple Fibre Channel connections to a client to increase transfer rates. The multipath method you specify for a storage pool determines how Xsan uses more than one Fibre Channel connection between a client and a storage pool.

| Multipath method | Description |
|---|---|
| Rotate | Xsan alternates transfers among the available client Fibre Channel connections. In this case, each transfer uses a different connection than the transfer before it, so it can be started before the preceding transfer finishes. |
| Static | Each LUN in the storage pool is assigned to a client Fibre Channel connection when the volume is mounted. |

If a client has two Fibre Channel connections to the SAN, you can increase transfer speeds between the client and a storage pool by setting the storage pool's multipath method to Rotate.

## About SAN Write Latency

The waiting period from the time of the write request until the notification that the request has been processed is called "latency," and is an important measure of volume performance. Xsan constantly monitors file system write latency and records hourly summaries in the cvlog log file.

In general, peak metadata throughput begins to suffer when the average latency exceeds 500 microseconds.

Xsan writes hourly summaries of metadata write latency into each volume's log file. Scan the log for entries that contain the text "PIO HiPriWr SUMMARY".

**To check file system metadata write latency:**

- Open Xsan Admin, select the volume in the SAN Components list, and click Logs. Choose Volume Log from the Show pop-up menu and the controller you want to examine from the On pop-up menu. Then type PIO in the filter field and press Return.

  The entry lists the average, minimum, and maximum metadata write latencies (in microseconds) for the reporting period. For example:

```
[0802 15:20:30] (Debug) PIO HiPriWr SUMMARY LUN0 sysavg/350 sysmin/333
     sysmax/367
```

**To generate a new latency summary:**

You can use the command line to generate a new latency summary on demand instead of waiting for the next hourly update.

- Open Terminal and type

```
$ sudo cvadmin -F volume -e 'debug 0x01000000'
```

where volume is the name of the volume.

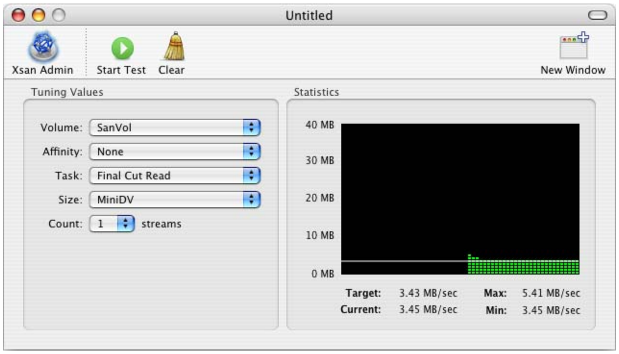
To see the new summary entry from the command line, type

```
$ tail -100 /Library/Filesystems/Xsan/data/volume/log/cvlog
```

# Using the Xsan Tuner Application 2

## This chapter shows how to use the Xsan Tuner application to test the performance of Xsan volumes.

You can use Xsan Tuner to test the data and video transfer capabilities of your storage area network and its Xsan volumes. Xsan Tuner can simulate both standard UNIX reads and writes and Final Cut Pro video reads and writes for a variety of common video formats. Use Xsan Tuner to see if your SAN can handle planned workloads before you put it into production use.



Volume: SanVol
Affinity: None
Task: Final Cut Read
Size: MiniDV
Count: 1 streams

**Target:** 3.43 MB/sec **Max:** 5.41 MB/sec
**Current:** 3.45 MB/sec **Min:** 3.45 MB/sec

## Where to Get Xsan Tuner

The Xsan Tuner application is available by following the link in the Knowledge Base article that describes it. Go to the Xsan support website at www.apple.com/support/xsan and search for Xsan Tuner.

## Installing Xsan Tuner

There is no installer for Xsan Tuner; simply copy the application to the hard disk of a computer on the SAN.

***Note:*** You can use the Xsan Tuner application only on computers running Mac OS X version 10.4 or later or Mac OS X Server version 10.4 or later.

## Starting Xsan Tuner

Double-click the Xsan Tuner icon to start the application.

# About the Tests

Xsan Tuner can perform four basic types of tests:

- Standard UNIX file reads
- Standard UNIX file writes
- Final Cut Pro video stream reads
- Final Cut Pro video stream writes

## The UNIX Read and Write Tests

The UNIX Read and UNIX Write tests perform low-level sequential I/O using the standard BSD read and write calls. You can use this test to simulate the type of load your SAN experiences when users or applications open, save, or copy files.

## The Final Cut Pro Read and Write Tests

The Final Cut Pro Read and Final Cut Pro Write tests simulate the load imposed on the SAN by streams of video. You can use these tests to find out how many video streams of particular video formats the SAN can support without dropping frames.

### You Can Test Without Installing Final Cut Pro

Xsan Tuner includes the codecs needed to perform Final Cut Pro reads and writes, so you can test the performance of Final Cut Pro video processing on client computers without actually installing the Final Cut Pro software on all of the clients.

### Test Video Streams Don't Include Audio

The video streams generated by the Final Cut Pro read and write tests do not include any audio content.

### Test Video Streams Are Not Staggered

When you choose to simulate more than one video stream on a client, Xsan Tuner launches all of the streams simultaneously. You can't simulate staggered streaming using Xsan Tuner.

### Supported Video Formats

Xsan Tuner can test Final Cut Pro streaming performance for a variety of standard-definition (SD) and high-definition (HD) video formats.

The following table lists the transfer rate that Xsan Tuner tries to sustain for each stream of the indicated video format.

| Video type | Format | Minimum data rate per stream |
|---|---|---|
| Standard Definition | MiniDV | 3.43 MB/s |
| | DVCAM | 3.43 MB/s |
| | DVCPRO | 3.43 MB/s |
| | DVCPRO 50 | 6.87 MB/s |
| | Uncompressed SD 8-bit | 20.02 MB/s |
| | Uncompressed SD 10-bit | 26.7 MB/s |
| Compressed High Definition | DVCPRO HD | 5.49 MB/s |
| Uncompressed High Definition | 720p 24 fps | 42.19 MB/s |
| | 720p 30 fps | 52.73 MB/s |
| | 720p 60 fps | 105.47 MB/s |
| | 1080 24p 8-bit | 94.92 MB/s |
| | 1080 24p 10-bit | 126.56 MB/s |
| | 1080i 8-bit | 118.65 MB/s |
| | 1080i 10-bit | 158.2 MB/s |

***Note:*** These are the target data rates used by the Xsan Tuner application based on QuickTime use of codecs, and may vary slightly from other published data rates for the same video formats.

## About the Test Files

To perform a UNIX or Final Cut Pro read test, Xsan Tuner must first create sample source files. The files are created in a folder named Xsan Tuner at the root level of the volume you are testing.

The first time you run a read test, you'll see a status bar as Xsan Tuner creates the necessary files.

***Note:*** The creation of the test files Xsan Tuner uses for a test can significantly affect the results of other tests in progress. Xsan Tuner displays a dialog to inform you when it is creating test files. If you are testing using multiple clients simultaneously, wait until all the test files are created before you start a test on any client.

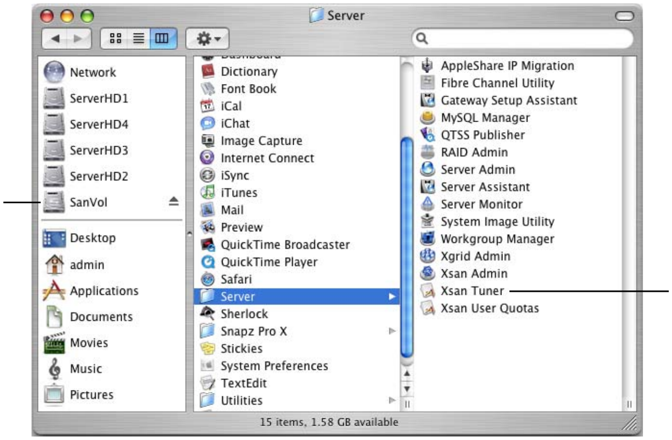
Xsan Tuner deletes its test files when you quit the application.

## Performing a Test

To test the performance of a SAN volume, you need to:

- Prepare the client computers
- Choose test type and related settings
- Start the test

### Step 1: Prepare the Client Computer



Mount the Xsan volume on the client . . .

. . . and install the Xsan Tuner application.

1 Make sure that the Xsan volume you want to test is mounted on the client.

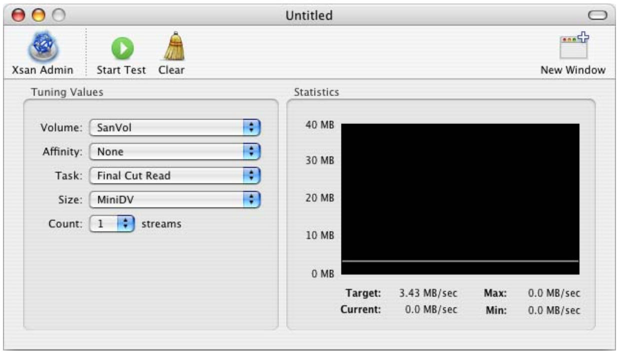
If the volume is not mounted, use Xsan Admin to mount it.

2 Copy the Xsan Tuner application to the client you want to test.

*Note:* You can use the Xsan Tuner application only on computers running Mac OS X version 10.4 or Mac OS X Server version 10.4 Tiger or later. For information on moving your Xsan environment to Tiger, see the *Xsan Migration Guide,* available at www.apple.com/server/documentation.

3 To test multiple clients on the SAN simultaneously, repeat steps 1 and 2 for each client.

### Step 2: Choose Test Type and Settings



Volume: SanVol
Affinity: None
Task: Final Cut Read
Size: MiniDV
Count: 1 streams

Target: 3.43 MB/sec
Current: 0.0 MB/sec
Max: 0.0 MB/sec
Min: 0.0 MB/sec

1 Open Xsan Tuner on the client computer.

2 Choose the volume you want to test from the Volume pop-up menu.

*Note:* You can use Xsan Tuner to test any volume mounted on the client, not just Xsan volumes.

3 To test a specific storage pool in an Xsan volume, select the pool's affinity name from the Affinity pop-up menu.

4 Choose a test type from the Task pop-up menu.

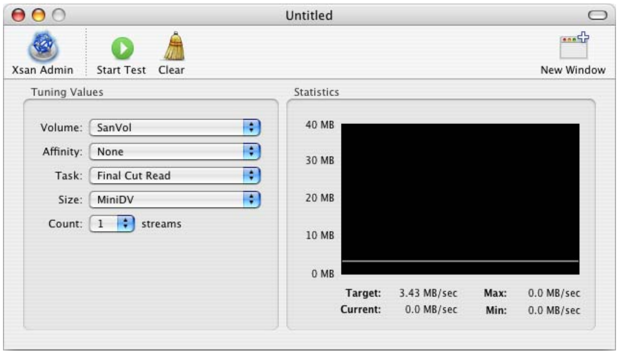
5 If you have chosen a Final Cut Pro test, choose a video format from the Size pop-up menu.

6 To simulate more than one stream of the chosen data type, choose from the Count pop-up menu.

7 If you expect to perform the same test with the same parameters at another time, you can choose File > Save to save the test window as you have configured it, and then open it again at a later time.

8 If you are testing several clients simultaneously, repeat these steps on the other clients.

### Step 3: Run the Test



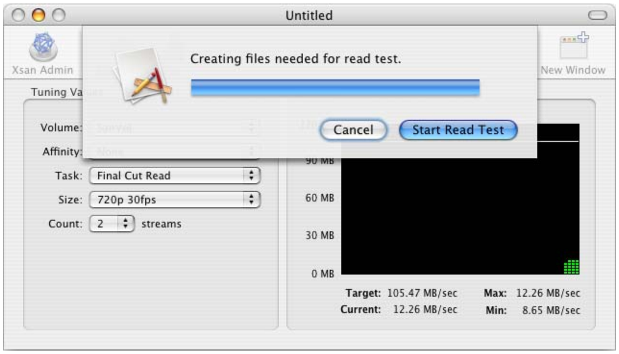
Volume: SanVol
Affinity: None
Task: Final Cut Read
Size: MiniDV
Count: 1 streams

Target: 3.43 MB/sec Max: 0.0 MB/sec
Current: 0.0 MB/sec Min: 0.0 MB/sec

1 When you have selected the volume to test and the test parameters, click Start Test.

If you are performing a write test, the test begins immediately.

If you are performing a read test, Xsan Tuner first creates readable test files.



Creating files needed for read test.

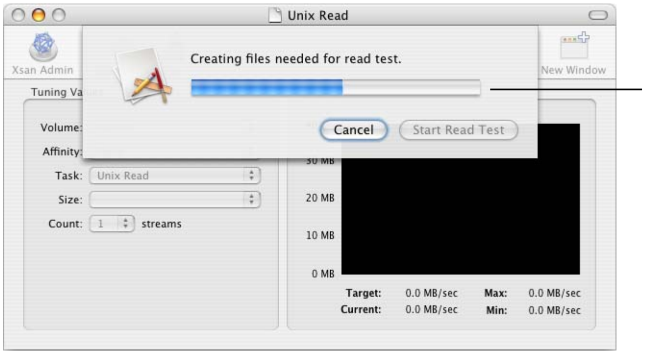
Volume: SanVol
Affinity: None
Task: Final Cut Read
Size: 720p 30fps
Count: 2 streams

Target: 105.47 MB/sec Max: 12.26 MB/sec
Current: 12.26 MB/sec Min: 8.65 MB/sec

2 If this is the first time you've requested a read test, wait until Xsan Tuner creates its test files, and then click Start Read Test.

*Important:* If you are testing multiple clients, don't start a test on any client until the test files have been created on all clients. The process of creating the test files for one client significantly affects test results on other clients.

The following screen captures show how the creation of the test file for a read test can influence the results of a separate test already in progress.

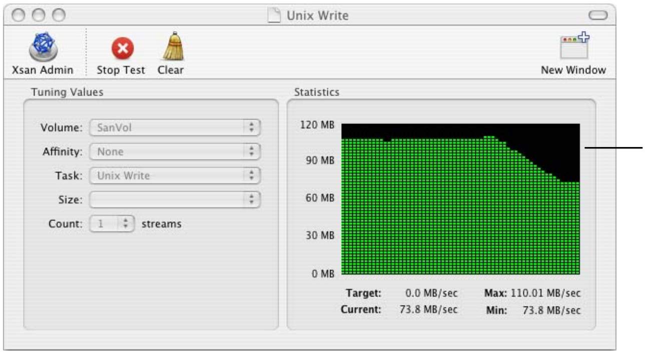


Creating files needed for read test.

Volume:
Affinity:
Task: Unix Read
Size:
Count: 1 streams

Target: 0.0 MB/sec Max: 0.0 MB/sec
Current: 0.0 MB/sec Min: 0.0 MB/sec

Creating test files for a read test . . .



Volume: SanVol
Affinity: None
Task: Unix Write
Size:
Count: 1 streams

Target: 0.0 MB/sec Max: 110.01 MB/sec
Current: 73.8 MB/sec Min: 73.8 MB/sec

. . . can have significant effect on another test in progress.

## Interpreting Test Results

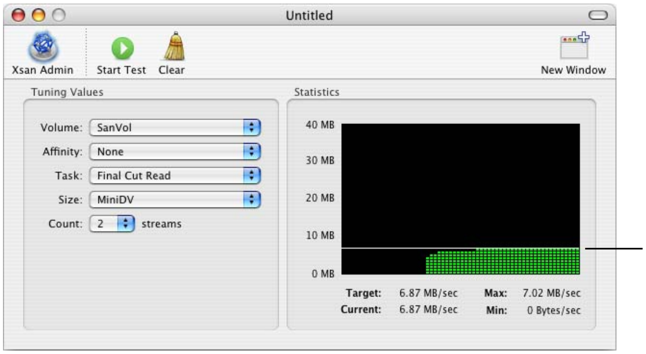
Xsan Tuner is designed to estimate SAN performance in terms of the number of streams a client should be able to read or write (for video) or the raw data transfer rate (for UNIX reads and writes). In many cases, a client doing actual work might get better results than indicated by the test. The value reported by Xsan Tuner can be considered a reliable minimum value.

### Understanding UNIX Test Results

The UNIX test results can be understood directly as the rate at which data is moving over the SAN in response to client read and write requests.

## Understanding Final Cut Pro Test Results

The Final Cut Pro Read and Final Cut Pro Write tests performed by Xsan Tuner compare the data rates possible through a specific Final Cut Pro codec against the throughput needed to stream that data without dropping frames.

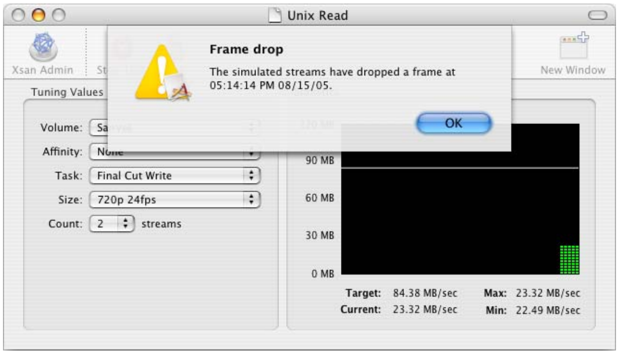


Volume: SanVol
Affinity: None
Task: Final Cut Read
Size: MiniDV
Count: 2 streams

Target: 6.87 MB/sec Max: 7.02 MB/sec
Current: 6.87 MB/sec Min: 0 Bytes/sec

The white line marks the minimum required data rate for the specified streams.

A frame drop message, shown below, indicates that the required throughput can't be sustained by the SAN in its current configuration:



**Frame drop**

The simulated streams have dropped a frame at 05:14:14 PM 08/15/05.

Volume: Sa
Affinity: None
Task: Final Cut Write
Size: 720p 24fps
Count: 2 streams

Target: 84.38 MB/sec Max: 23.32 MB/sec
Current: 23.32 MB/sec Min: 22.49 MB/sec

### Adjusting the Result for Audio Streams

Because the video streams generated by the Xsan Tuner Final Cut Pro read and write tests do not include audio content, you should adjust the test results downward if your clients will be working with audio.

If the client computer you are testing will be processing both audio and video, subtract one stream from the Xsan Tuner results to estimate how many video streams can be supported.

For example, if Xsan Tuner shows that the SAN can support four video streams for a client, the actual number of sustainable streams is more likely to be three if the streams include audio.

### Adjusting the Result for Staggered Streams

When you test a client for multiple streams, Xsan Tuner launches all of the streams simultaneously. Xsan Tuner does not simulate staggered streaming. If you configure Final Cut Pro to take advantage of staggered streams for your routing work, you might see a two- or three-stream improvement in actual use over the number of streams reported by Xsan Tuner.

# Deployment Examples

# 3

## This chapter shows examples of how to set up Xsan volumes in a storage area network.

Read this chapter to see how you can use Xsan to set up volumes on a storage area network to provide these different types of storage:

- Network attached storage
- Computational cluster
- Video production

## Which Example Should You Follow?

The three deployment examples address different performance needs.

### Minimal Hardware and Standard Performance

The network attached storage example shows how to set up SAN volumes using a minimum of hardware and network connections.

### Moderate Performance

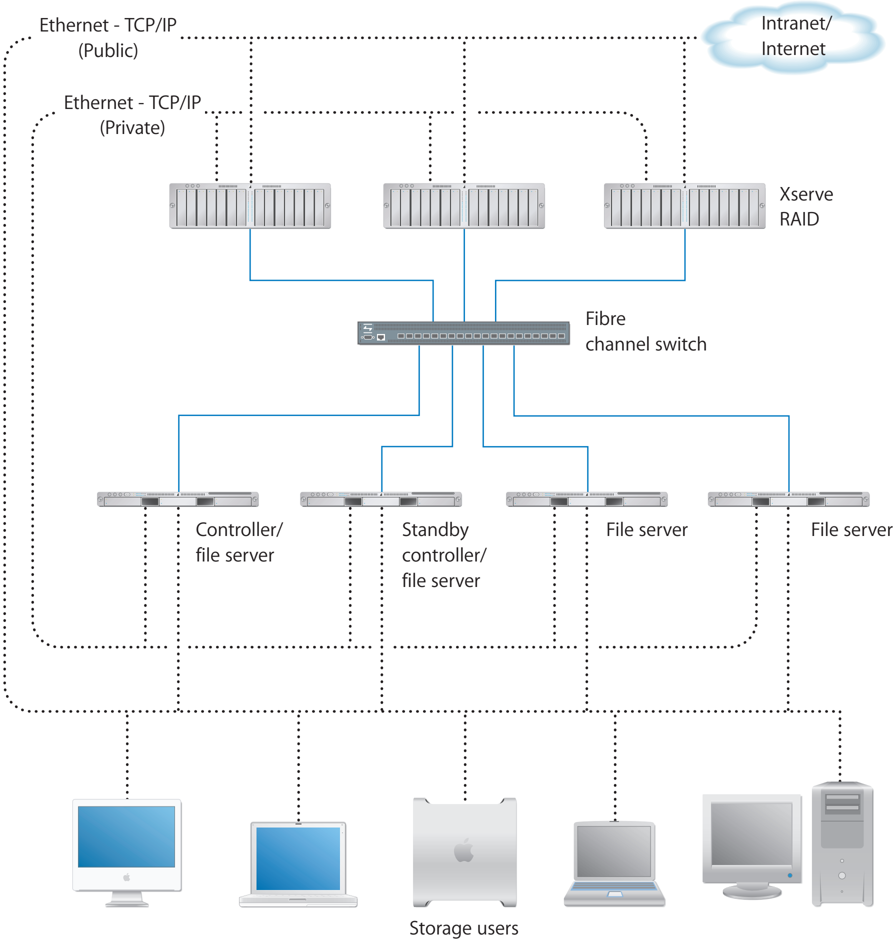
The computational cluster example shows how to achieve moderate performance.

### Best Possible Performance

The video production example shows how to achieve the best possible performance.

## Example 1: Network Attached Storage

Network attached storage (NAS) is commonly used to provide centralized, manageable storage for clients on the Internet or on a private intranet. A typical NAS appliance contains storage devices and a controller that gives clients access to the storage using a network file system such as NFS, AFP, or SMB/CIFS.



You can use a combination of Xserve and Xserve RAID to provide NAS and, in addition, provide other server services that NAS appliances can't offer. Adding Xsan lets you set up a NAS service that is easy to expand, both in terms of storage capacity and file system services, without impact to clients.

In a NAS configuration, the real clients of the storage are not connected to the SAN, but instead connect to the storage server over a TCP/IP network. The storage server itself is the true Xsan client computer and, in the case of a single server computer, is also the SAN controller. The server provides file services using Mac OS X Server AFP, SMB, or NFS services to other computers on the network.

## Objectives of This Configuration

This example shows one way to set up a SAN to meet the following requirements:

- Provide 100 GB of network-accessible storage for each of 80 users
- Support users running Mac OS X, Windows, or UNIX
- Storage must be always available

## Deployment Decisions

The following paragraphs answer the planning questions listed in the *Xsan Administrator's Guide* in a way that satisfies the objectives of the network attached storage deployment example.

### How much storage is needed?

100GB of storage for each of 80 users requires at least 8 TB of user disk space. An Xserve RAID with a full 14 400 GB drive modules can provide approximately 5.6 TB of total storage. However, to provide redundancy, we'll set aside one drive on each controller (on each half of an Xserve RAID system) as a spare, and organize the remaining 6 drives as RAID 5 arrays. Using RAID 5 means that the equivalent of one drive (distributed across all 6 drives) out of each array is dedicated to RAID parity. So, each array on each Xserve RAID controller has the equivalent of 5 drives available to store 2 TB of user files, or a total of 4 TB per Xserve RAID. At 100 GB per user, each Xserve RAID can support 40 users. Three Xserve RAID systems, configured as RAID 5 arrays with one spare drive per unit, will provide sufficient space and leave an extra array for metadata and journal data.

### What storage arrangement makes the most sense for user workflow?

In this example, users don't connect to the SAN directly, so they don't see how storage in the SAN is organized. Instead, they connect to share points offered by file services on the file servers. The true Xsan clients are the file servers themselves. The intent of the NAS configuration is to offer large amounts of storage without imposing any particular organization (users can do that themselves), so it suffices to set up large volumes with a top-level folder for each user.

### How is the storage presented to users?

When users connect to the server share point, they see a list of folders, one for each user.

### What levels of performance do your users require?

The primary constraint on file access speed is the Ethernet network that connects users to the file servers. By comparison, the speed of the SAN and its storage devices is not an issue. However, if you host many users, the performance of the file servers is also important. So, this NAS setup includes multiple servers.

### How important is availability?

Users expect access to their files at all times, so high availability is critical. To address this, the NAS example includes a standby metadata controller that can take over if the primary controller is unresponsive.

### What are the requirements for security?

Security is important in an environment where many users are sharing the same storage. In this NAS example, security can be achieved by assigning appropriate access privileges to users on the file servers. Because only the file servers have direct access to the SAN, no special security setup is needed for the SAN itself.

### What RAID schemes should be used for the RAID arrays?

Because availability and recoverability are important, the RAID arrays are set up to use a RAID 5 scheme. Xserve RAID systems are optimized for high performance using RAID 5.

### Which storage pools make up each volume?

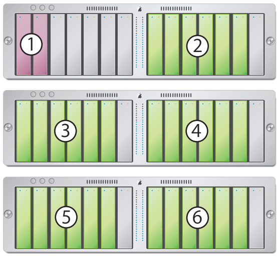
The volume consists of just two storage pools, one for metadata and the other, consisting of 5 LUNs, for user data.

### How are individual volumes organized?

Again, because the primary goal here is to provide general-purpose storage, no special organization is imposed on the available storage.

### Which LUNs go in each storage pool?

All the storage is being provided by similar Xserve RAID systems configured as RAID 5 arrays, so differences in LUN performance are not an issue. We'll create two arrays per Xserve RAID system for a total of 6 LUNs, and then assign 5 LUNs to the user data storage pool. 6-drive arrays permit a hot standby drive for unattended protection during long computation sessions.



| | | |
|---|---|---|
| ① | 2-disk RAID 1 | Metadata pool |
| ② | 6-disk RAID 5 | User data pool |
| ③ | 6-disk RAID 5 | |
| ④ | 6-disk RAID 5 | |
| ⑤ | 6-disk RAID 5 | |
| ⑥ | 6-disk RAID 5 | |

**Which clients should be able to access each volume?**
The real SAN clients in this example are the file servers that users connect to. Each of the servers should have access to the SAN volume for maximum availability.

**Which computers will act as controllers?**
To avoid dedicated controllers, one of the file servers will act as the metadata controller and a second file server will be configured as a standby controller.

**Do you need standby controllers?**
High availability is a high priority, so a second file server is also configured as a standby controller so it can take over the role of metadata controller if necessary.

**Do you want to use controllers as clients also?**
The controllers in this example are the real clients of the SAN, so each is set up to act as both client and controller.

**Where do you want to store file system metadata and journal data?**
Data in this example is delivered to users over Ethernet, so the performance of the SAN itself is not as critical as when clients are connected directly to the Fibre Channel network. However, to ensure good volume performance, we'll create a separate storage pool for metadata and journal data.

**What allocation strategy?**
Data in this volume is stored in a single storage pool of five LUNs. Because there is only one user data storage pool, allocation strategy is not an issue, so we'll accept the default allocation strategy: Round Robin.

**What block size and stripe breadth should we use?**
Because shared network storage typically involves lots of files and many small, randomly positioned reads and writes, we'll use the relatively small default file system block size of 4 KB together with a 256-block stripe breadth for the storage pool. This should provide good space utilization.

# Example 2: Storage for Computational Clusters

You can also use Xsan with Xserve and Xserve RAID systems to provide manageable storage for a high performance computational cluster.

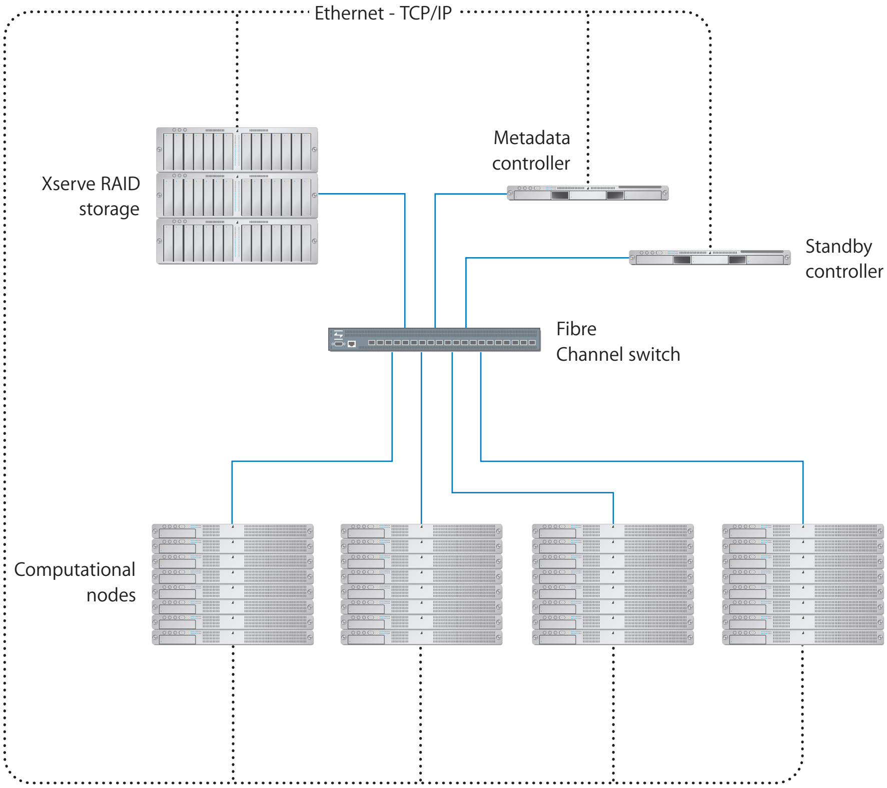
## Requirements for This Example

This example shows one way to set up a SAN to meet the following requirements:

- Provide computational nodes with access to any of three existing 600 GB data sets and space to write similarly large result files
- Minimize file read and write times
- Protect the integrity of the data sets

## Small Computational Cluster Setup

Here's one way to set up a SAN to support a small computational cluster (one with 64 or fewer devices attached to the SAN).



## Planning Decisions

The following paragraphs show how the planning questions in Chapter 2 are answered for this small computational cluster example.

### How much storage is needed?

Three Xserve RAID systems, one for metadata and the 1.8 TB of input data and two others for up to 8 TB of results, are adequate.

### What organization of storage makes the most sense for user workflow?

The "user" in this example is an application processing a large data set, so no special organization is needed to help humans navigate or manage the storage hierarchy. Root-level access to the large data sets is adequate.

### What levels of performance do users require?

Performance is an important goal in this example. So, all of the computational nodes are attached directly to the Fibre Channel network.

### How important is availability?

Ensured availability of the file system during long, unattended computational sessions is important, so we'll include a standby controller.

### What are the requirements for security?

Security is satisfied by the fact that the SAN is physically restricted to the cluster itself. There is no access from outside.

### What RAID schemes should be used for the RAID arrays?

Xserve RAID systems are set up to provide RAID 5 arrays for a balance of performance and recoverability.

### How many volumes are needed?

This example uses a single volume with two affinities, one for input data and the other for results.

### Which storage pools make up each volume?

For this example, the single volume consists of three storage pools:

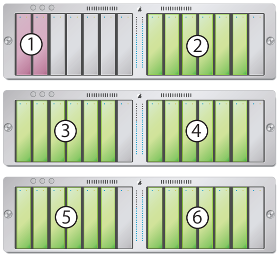
- 1 pool for metadata
- 1 pool with affinity for source data sets
- 1 pool with affinity for results

### How are individual volumes organized?

Because the volumes are used by an application and not by human users, we'll store the data sets at the root level of the volumes.

### Which LUNs go in each storage pool?

For metadata, we use a single 2-drive RAID 1 array. For the data pools, we can use the RAID 5 LUNs available straight out of the box on the Xserve RAID systems:  one on the system that also supports the metadata and two on each of the other Xserve RAID systems for a total of 5.



| | | |
|---|---|---|
| ① | 2-disk RAID 1 | Metadata pool |
| ② | 6-disk RAID 5 | In data pool w/affinity for source |
| ③ | 6-disk RAID 5 | Out data pool w/affinity for results |
| ④ | 6-disk RAID 5 | |
| ⑤ | 6-disk RAID 5 | |
| ⑥ | 6-disk RAID 5 | |

### Which clients should be able to access each volume?

All clients (processing nodes) have access to the data volume.

### Which computers will act as controllers?

There is a dedicated metadata controller.

### Do you need standby controllers?

Yes, one.

### Do you want to use controllers as clients also?

The metadata controllers are dedicated to that task and will not function as clients.

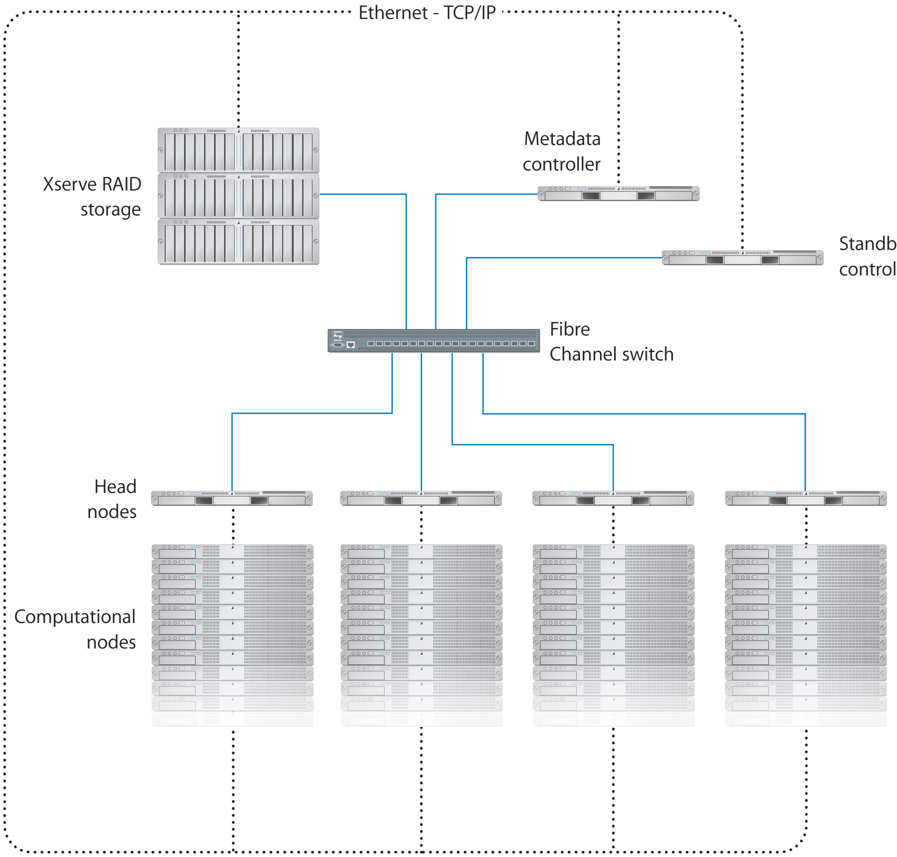
### Where do you want to store file system metadata and journal data?

For the sake of performance, metadata and journal data are stored on a separate LUN dedicated for that purpose.

## A Larger Computational Cluster

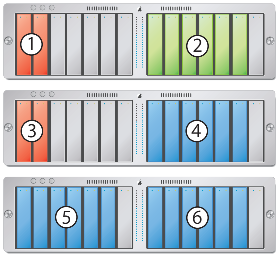
A cluster consisting of more than 64 nodes requires a different configuration (similar to NAS) because Xsan is limited to 64 Fibre Channel devices (controllers, clients, and storage devices) in a SAN.

To handle more than 64 devices, you can set up some of the processors to act as "head nodes" that are connected to the SAN and reshare the SAN data with the other processors over Ethernet.



## Volume Configuration

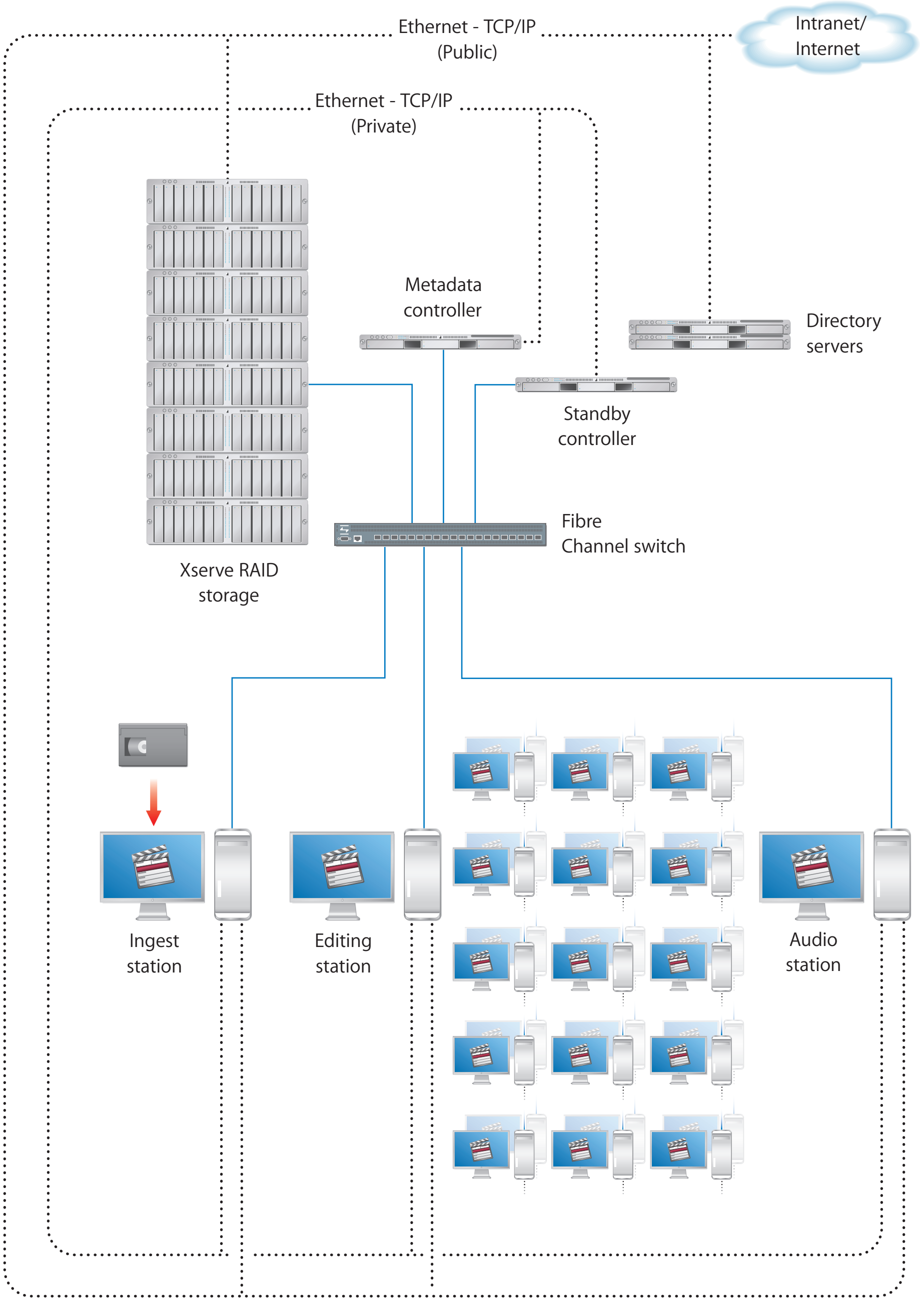
In this example, input data and output data are stored on separate volumes to increase performance. For data protection, however, we use 6-drive arrays with a hot spare in each array.



| | | |
|---|---|---|
| ① | 2-disk RAID 1 | Metadata pool for source volume |
| ② | 2-disk RAID 1 | In data pool for source volume |
| ③ | 6-disk RAID 5 | Metadata pool for results volume |
| ④ | 6-disk RAID 5 | Out data pool for results volume |
| ⑤ | 6-disk RAID 5 | |
| ⑥ | 6-disk RAID 5 | |

## Example 3: Storage for Video or Film Production Group

This example shows how to use Xsan to set up a storage area network to support a group of video editors.



The example is based on these assumptions:

- 16 editing workstations working with up to 4 streams of DVCPRO 50
- 1 editing workstation working with 2 streams of 8-bit 1080i

## Deployment Decisions

The following paragraphs answer the planning questions listed in the *Xsan Administrator's Guide* in a way that satisfies the objectives of the video storage deployment example.

### How much storage is needed?

In this example we are dealing with video and want to avoid dropped frames. So, deciding on the required number of Xserve RAID systems is guided more by the resulting SAN throughput than by raw storage space. In this example, we need to support 16 editing stations, each of which might be working with 4 simultaneous streams of DVCPRO 50 video. As shown in the table on page 19, each stream of DVCPRO 50 video requires a data rate of 7.7 MB per second.

4 streams x 7.7 MB/s per stream = 31 MB/s per station

16 stations x 31 MB/s per station = 496 MB/s total bandwidth required

A single Xserve RAID system (2 controllers with 7 drives each) can provide approximately 160 MB/s of throughput in an Xsan volume. Four Xserve RAID systems can provide 640 MB/s throughput, which should be able to handle the 496 MB/s requirement with overhead for audio and some additional cushion. So, the DV storage volume consists of four Xserve RAID systems for data and one half of a system for the metadata storage pool.

For the high-definition 1080i work, we need to provide one editing station with two streams:

2 streams x 120 MB/s per stream = 240 MB/s

Two Xserve RAID systems can provide the necessary bandwidth with a cushion. We'll incorporate these systems into a second volume for the high-definition work.

Add one Xserve RAID system to support the two metadata storage pools and another for storing audio files, and a total of eight systems should provide the needed storage and throughput.

### What levels of performance do your users require?

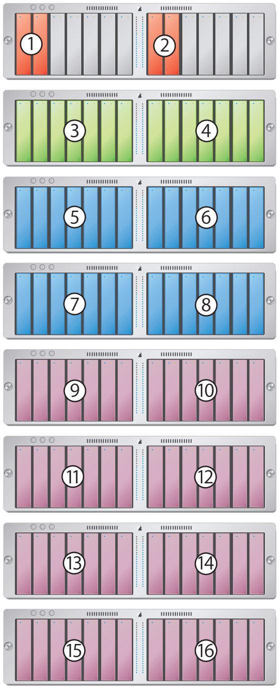
Performance is critical in this example. The primary concern is avoiding dropped video frames. We were guided by this requirement when we selected the number of Xserve RAID systems for the configuration. To make sure we get the best performance out of the file system itself, we'll keep metadata separate from other network traffic (such as Internet access and directory transactions) by using two Ethernet networks and by storing the metadata for each volume on its own storage pool and RAID controller.

### How important is availability?

High availability is important, so the configuration includes a standby metadata controller. All the data LUNs are set up as 6-drive RAID 5 arrays with a hot spare drive on each controller.

### Which storage pools make up each volume?

Each volume consists of just two storage pools, one for metadata and the other for user data. Data LUNs are composed of 7-disk RAID 5 arrays for maximum performance.



| | | |
|---|---|---|
| ① | 2-disk RAID 1 | Metadata pool, DV online volume |
| ② | 2-disk RAID 1 | Metadata pool, DV storage volume |
| ③ | 7-disk RAID 5 | Audio pool, DV online volume |
| ④ | 7-disk RAID 5 | |
| ⑤ | 7-disk RAID 5 | DV pool, DV online volume |
| ⑥ | 7-disk RAID 5 | |
| ⑦ | 7-disk RAID 5 | |
| ⑧ | 7-disk RAID 5 | |
| ⑨ | 7-disk RAID 5 | DV pool, DV storage volume |
| ⑩ | 7-disk RAID 5 | |
| ⑪ | 7-disk RAID 5 | |
| ⑫ | 7-disk RAID 5 | |
| ⑬ | 7-disk RAID 5 | |
| ⑭ | 7-disk RAID 5 | |
| ⑮ | 7-disk RAID 5 | |
| ⑯ | 7-disk RAID 5 | |

### Where do you want to store file system metadata and journal data?

Because we want the best possible performance from the volumes, we store the metadata storage pool for the two volumes on separate RAID controllers.

**What allocation strategy?**
Data in each volume is stored in a single storage pool. Because there is only one user data storage pool, allocation strategy is not an issue, so we'll accept the default allocation strategy:  Round Robin.

**What block size and stripe breadth should we use?**
Because video streams involve mostly sequential reads and writes, the video storage pools can benefit from a larger file system block size of 64 KB together with a 16-block stripe breadth. This should provide good space utilization.