CXFS, SGI’s Clustered File System

Thomas Feil

Dir. Marketing Storage Solutions EMEA

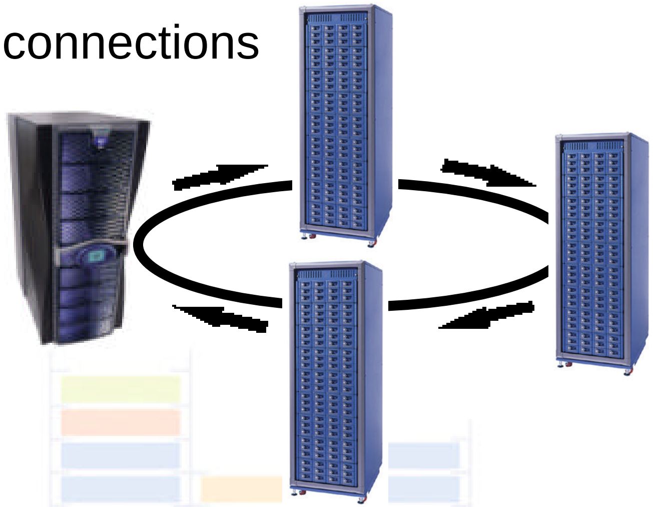
Agenda

- **Introduction**
  - What is a Storage Area Network aka SAN?
    - Fibre Channel Technologies and Topoligies
    - The Fibre Channel Fabric
- **CXFS, Delivering on the Promise**
  - CXFS Overview
  - CXFS Concepts
  - CXFS Performance
- **CXFS, Serving Advanced Environments**
  - High Availability
  - HSM
  - NFS, SAMBA

What is a SAN?

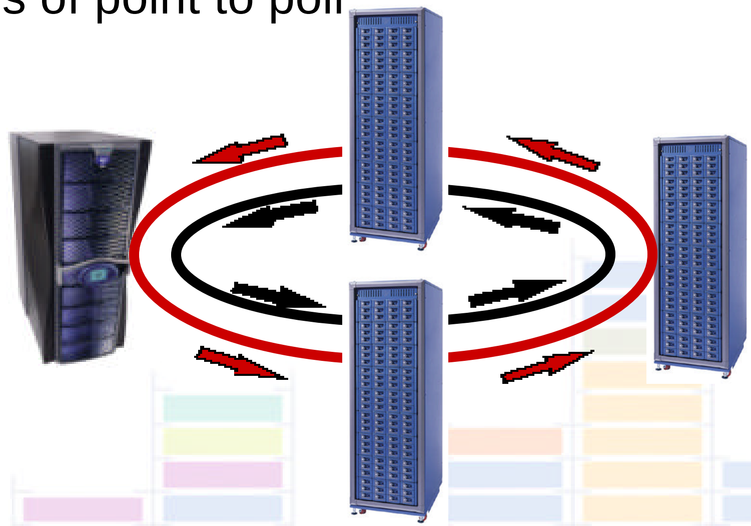
- Fibre Channel Topologies: Loops

- Each port arbitrates for access to the loop
- Ports that lose the arbitration act as repeaters
- Hubs make a loop look like a series of point to point connections



Single Loop

Data flows around the loop, passed from one device to another



Dual Loop

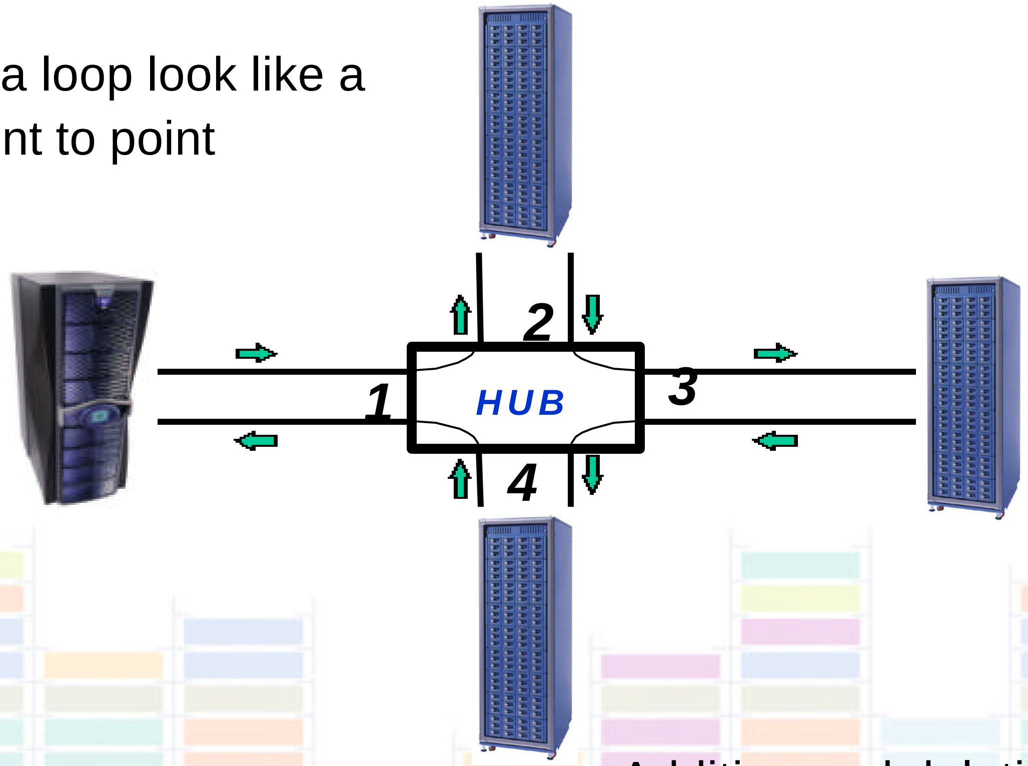
Some data flows through one lo while other data flows through the second loop

What is a SAN?

- Fibre Channel Topologies:

Hubs

Hubs make a loop look like a series of point to point connections.

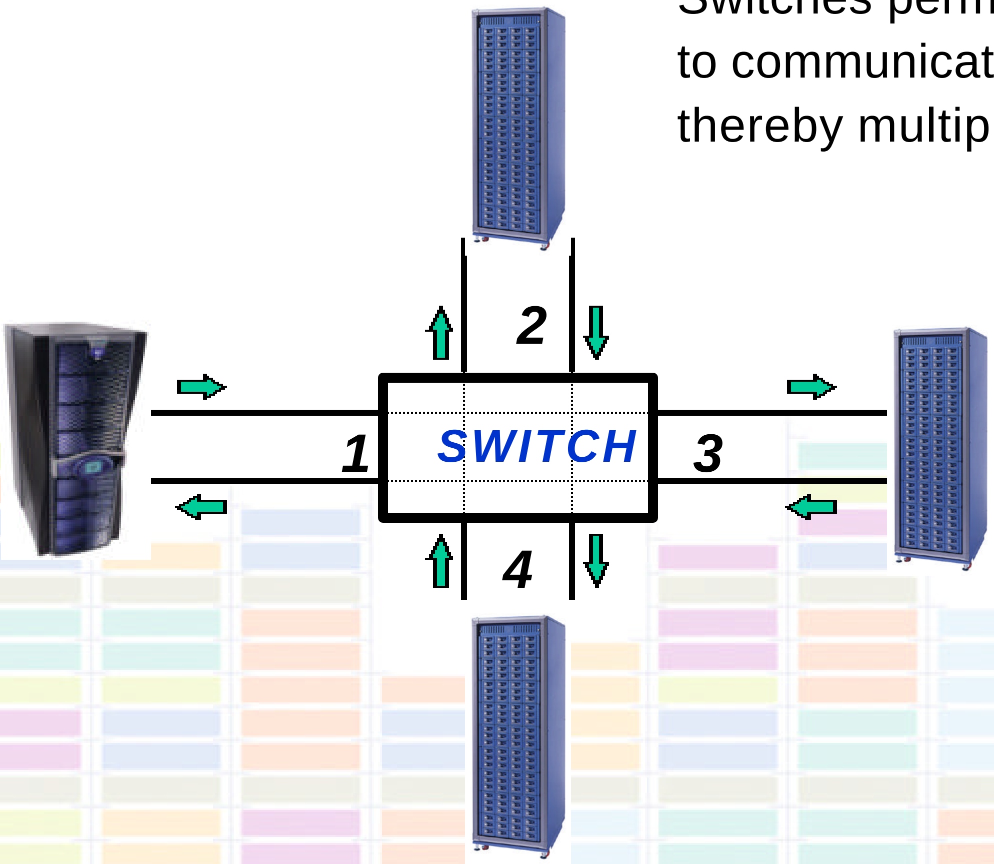


Addition and deletion of nodes is simple and non-disruptive to information flow.

What is a SAN?

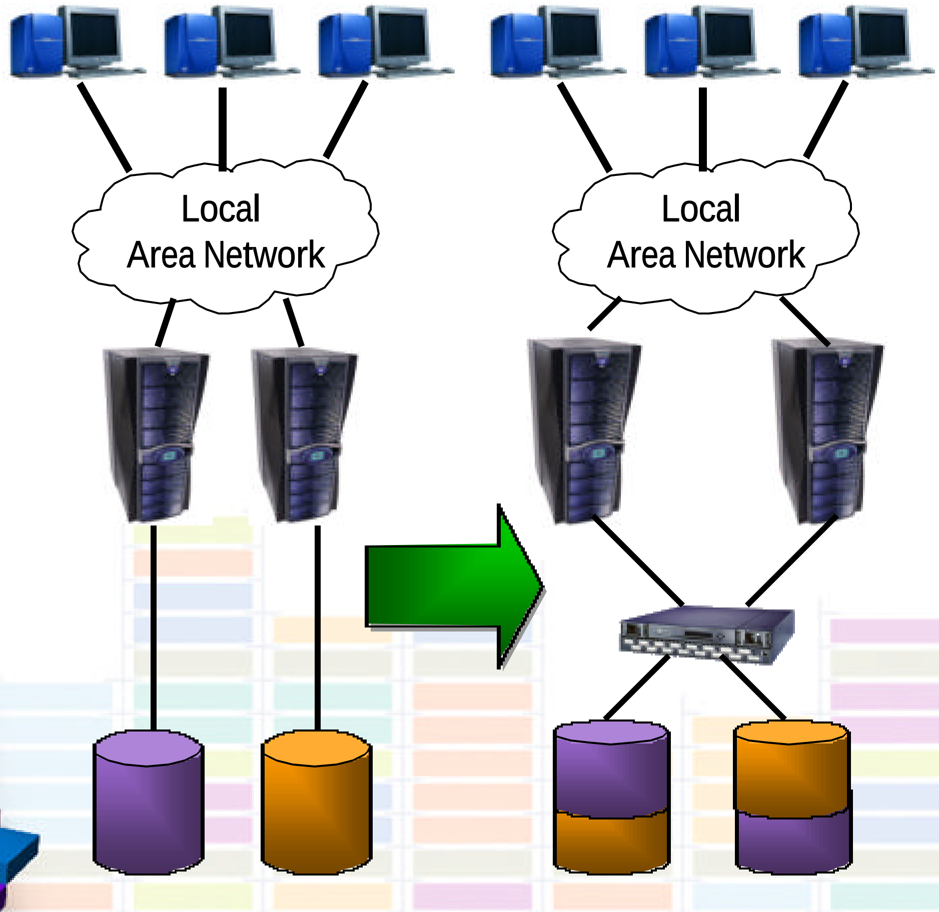
- Fibre Channel Topologies: Switches

Switches permit multiple devices to communicate at 100 MB/s, thereby multiplying bandwidth.



What is a SAN?

- From Direct-Attach to SAN-Attach

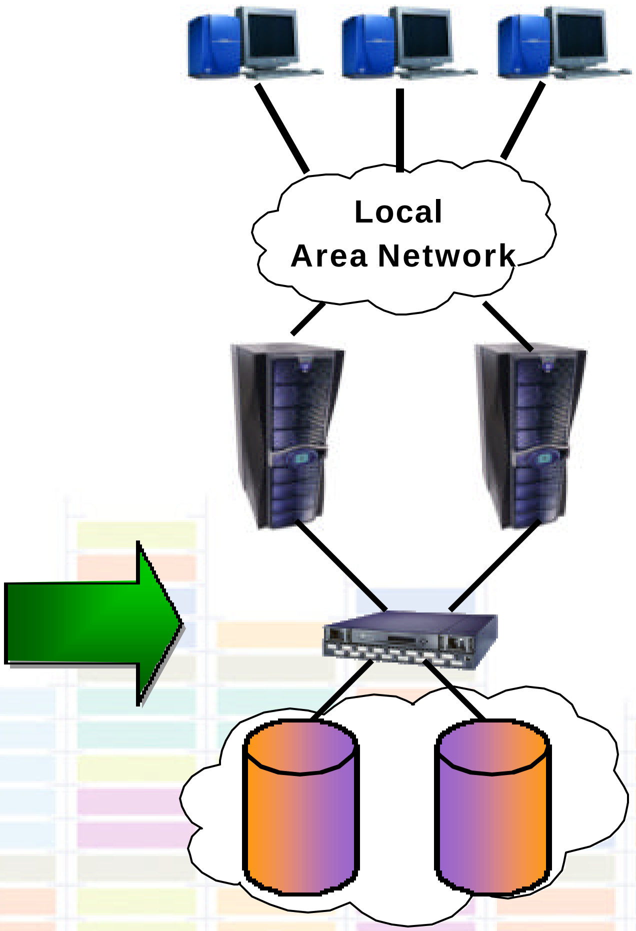


Promises:

- Centralized management
- Storage consolidation
- High availability and fault tolerance
- Modular scalability
- Shared infrastructure
- High bandwidth

What is a SAN?

- From sharing infrastructure to sharing data!

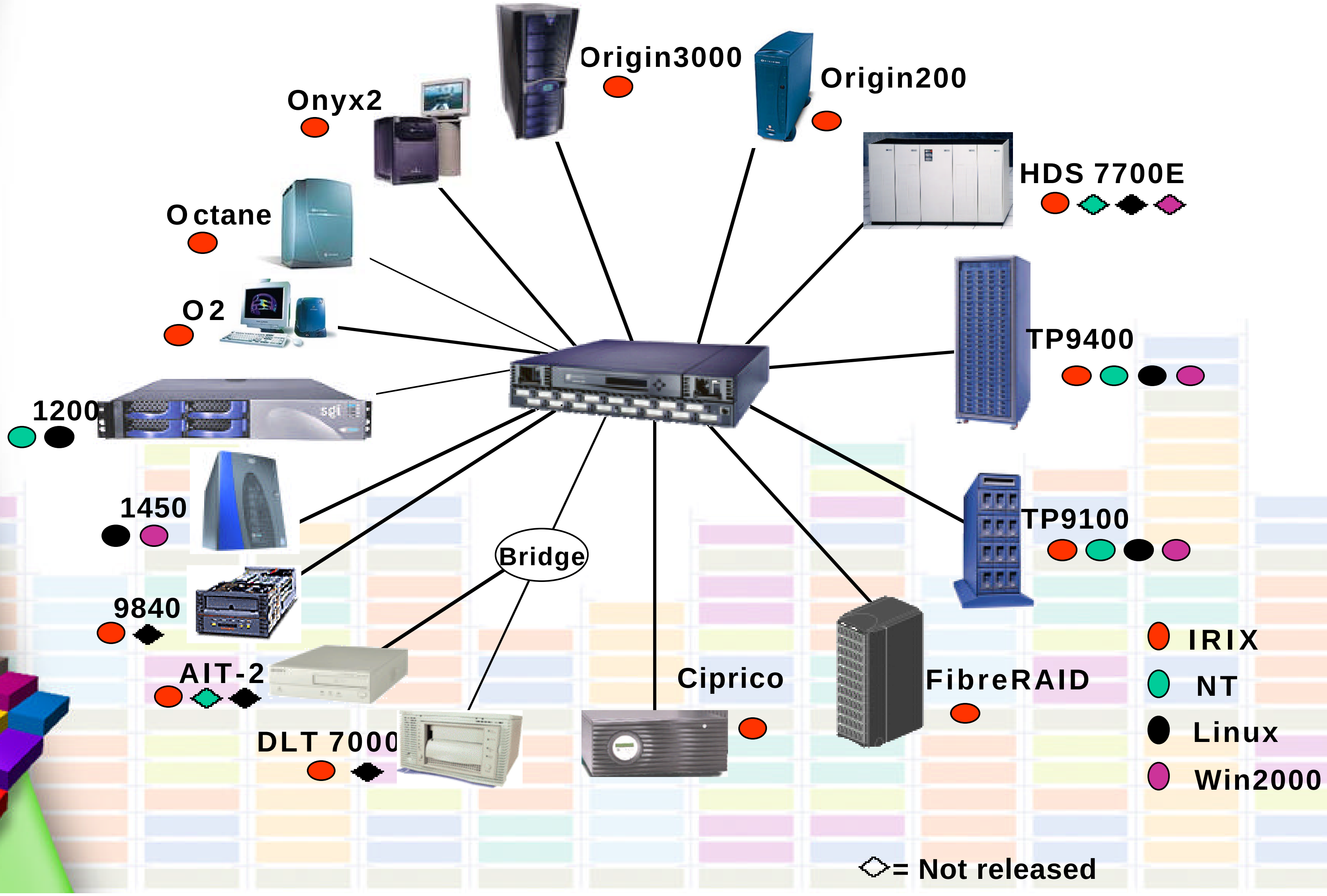


Benefits

- True value of SAN is sharing data among san attached systems => seamless access, no copy, ftp etc.
- With transparent filesystem, *lan-free backup* with already deployed applications
- Centralized management of data not just infrastructure
- Flexible scalability

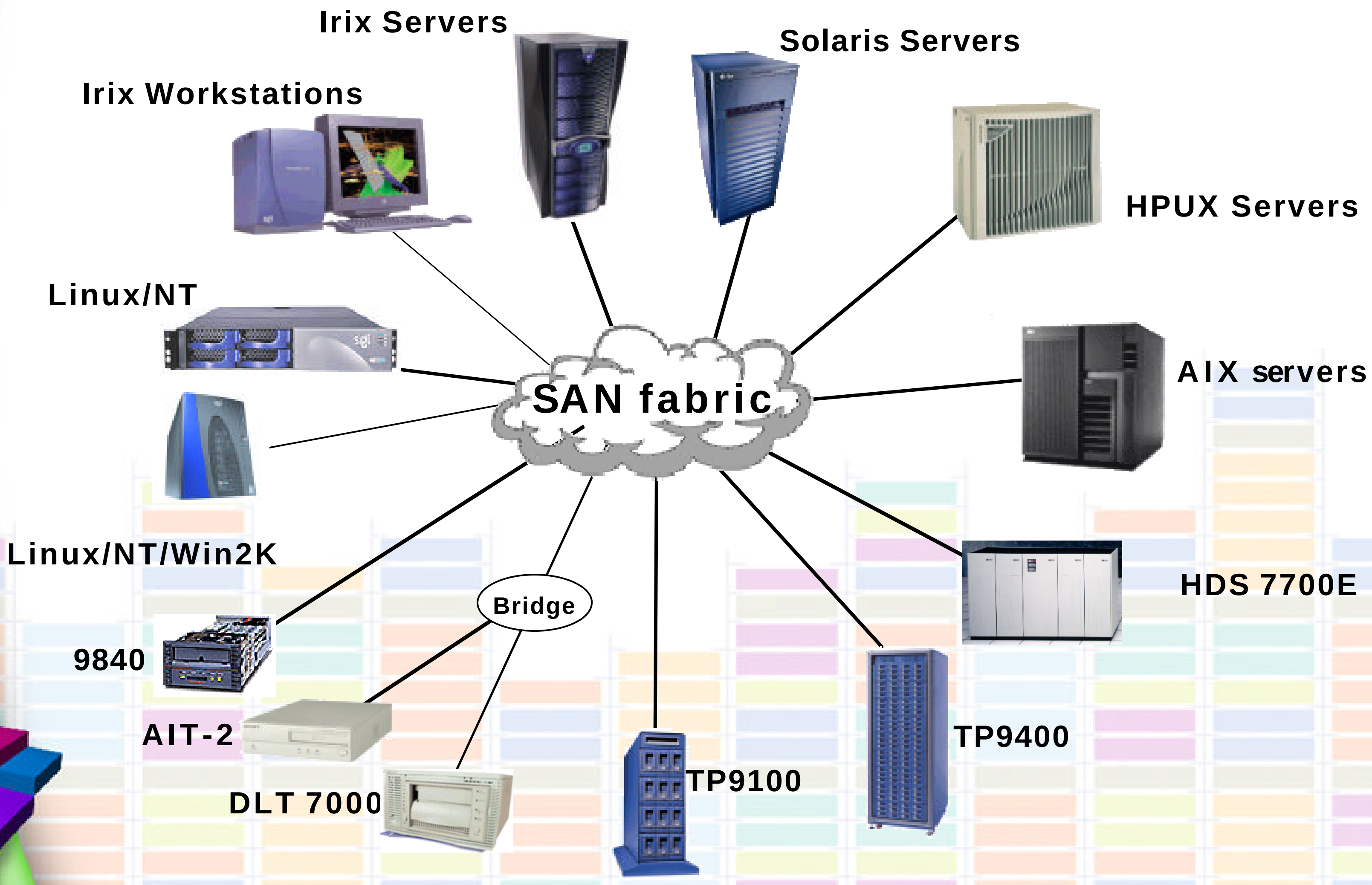
What is a SAN?

- Full SGI Support - Today



What is a SAN?

- Full SGI Support - Future



Agenda

- **Introduction**
  - What is a Storage Area Network aka SAN?
    - Fibre Channel Technologies and Topoligies
    - The Fibre Channel Fabric

➡ **CXFS, Delivering on the Promise**
  - CXFS Overview
  - CXFS Concepts
  - CXFS Performance

- **CXFS, Serving Advanced Environments**
  - High Availability
  - HSM
  - NFS, SAMBA

CXFS Overview

- Based on XFS, A World-Class Filesystem

Reliable

- Log/Journal
- Field proven

Fast

- Fast metadata speeds
- High bandwidths
- High transaction rates

Scalable

- Full 64 bit support
- Dynamic allocation of metadata space
- Scalable structures and algorithms

Open source version available for Linux from http://oss.sgi.com

CXFS Overview

- XFS Reliability

Field proven

- Run for years on thousands of IRIX systems.
- Part of IRIX since 1994
  - Released as part of IRIX 5.3

Log/Journal

- XFS designed around log
- No UNIX *fsck* is needed
- Recovery time is independent of filesystem size
  - Depends on system activity levels

Usually, recovery completes in under a second

CXFS Overview

- XFS Speeds

Fast metadata speeds

- B-Trees everywhere (Nearly all lists of metadata information)
  - Directory contents
  - Metadata free lists
  - Extent lists within file

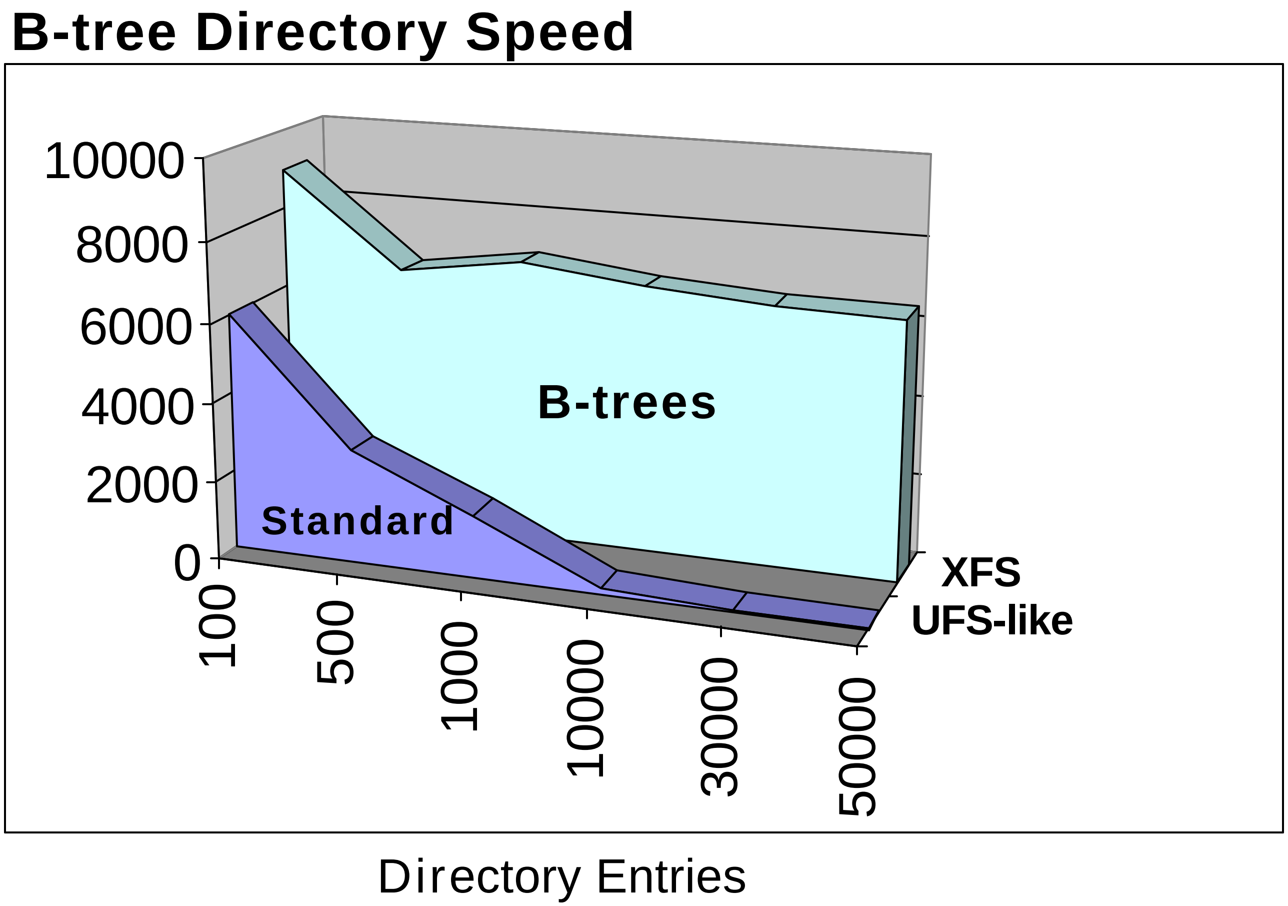
High bandwidths on SGI Origin 2000

- 7.32 GB/s on one filesystem (32p O2000, 897 FC disks)
- > 4 GB/s to one file (same Origin, 704 FC disks)
- Large extents (4 KB to 4 GB)
- Request parallelism
- Read ahead/Write behind

High transaction rates: 92,423 IOPS

CXFS Overview

- XFS Speeds



CXFS Overview

- XFS Speeds

Full 64 bit support

- Large Filesystem
  - – 18,446,744,073,709,551,615 = 2^{64}-1 = 18 million TB
- Large Files
  - – 9,223,372,036,854,775,807 = 2^{63}-1 = 9 million TB

Dynamic allocation of metadata space

- Unlimited number of files

Scalable structures and algorithms (B-Trees)

Performance is not an issue with large numbers of files and directories

CXFS: Clustered XFS

CXFS Attributes:

- **Shareable high-performance XFS file system**
  - Shared among multiple IRIX nodes in a cluster
  - Near-local file system performance.
  - Direct data channels between disks and nodes.
- **Resilient File System (highly available)**
  - Failure of a node in the cluster does not prevent access to the disks from other nodes
- **Convenient Interface**
  - Users see standard Unix File Systems
- **Single System View (SSV)**

CXFS Concepts

- The Metadata Model

Metadata

- The data about a file, including:
  - size, inode, create/modify times, and permissions

Metadata server node (a.k.a. CXFS server)

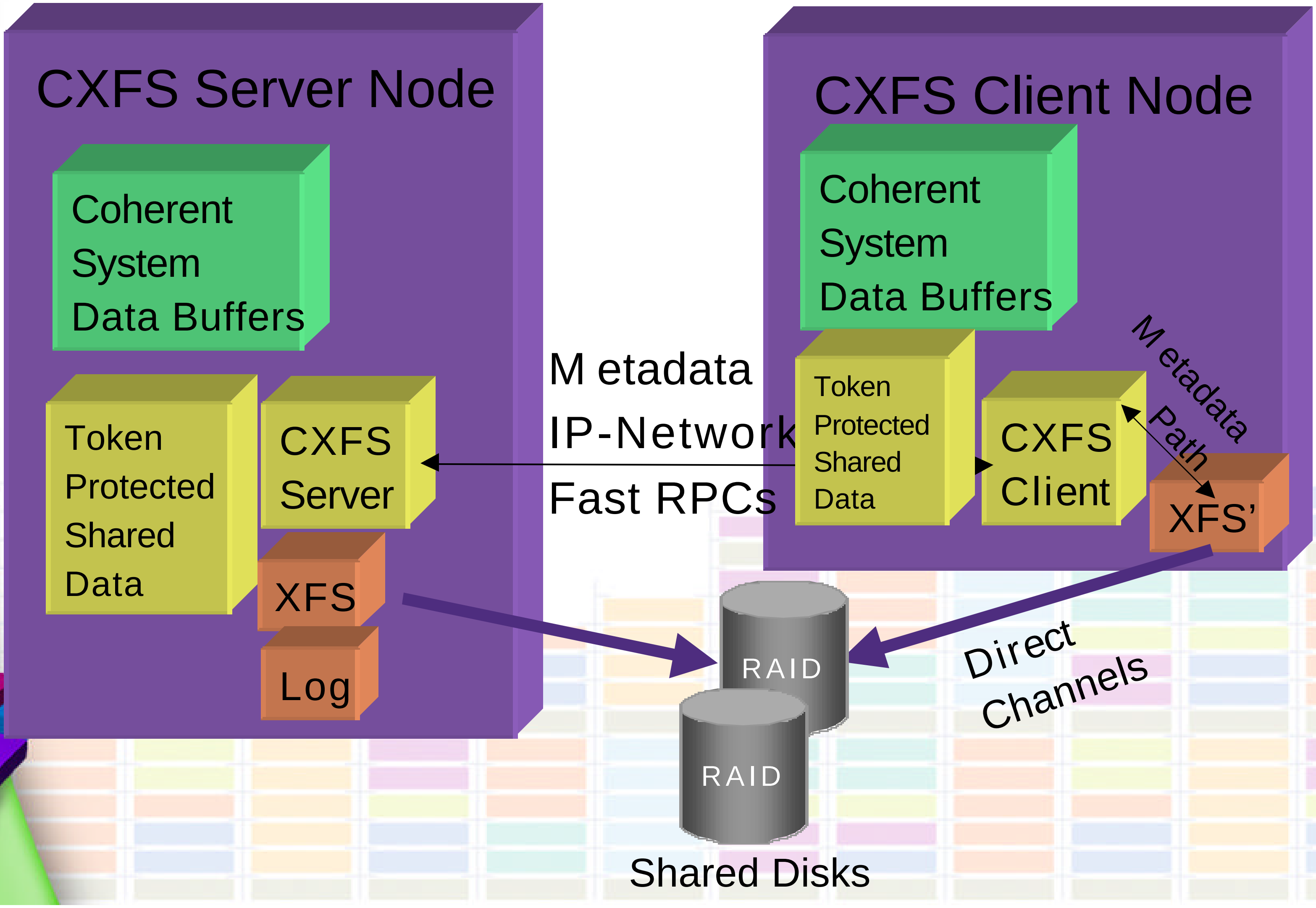
- One machine in the cluster that is responsible for controlling the metadata of files. It also plays “traffic cop” to control access to the file.
  - Backup metadata servers designated for fail-over
  - No single point of failure

Metadata client node (a.k.a. CXFS client)

- A machine in the cluster that is not the metadata server.
  - Must obtain permission from metadata server before accessing the file.

CXFS Concepts

- The M etadata M odel



CXFS Concepts

- Fast and Efficient Metadata

- Fast-asynchronous XFS metadata transactions in server

- Customized RPC mechanism
  - maximize communication speed among clients and the metadata server

Some other shared-file systems use NFS communication to read and write the metadata. This slows access to data

CXFS Concepts

- Full POSIX Filesystem API Support

Efficient buffering of metadata in clients

- Metadata is buffered in the clients
- Reread metadata if the file size or position changes

The CXFS application programmer interface (API) is POSIX compliant

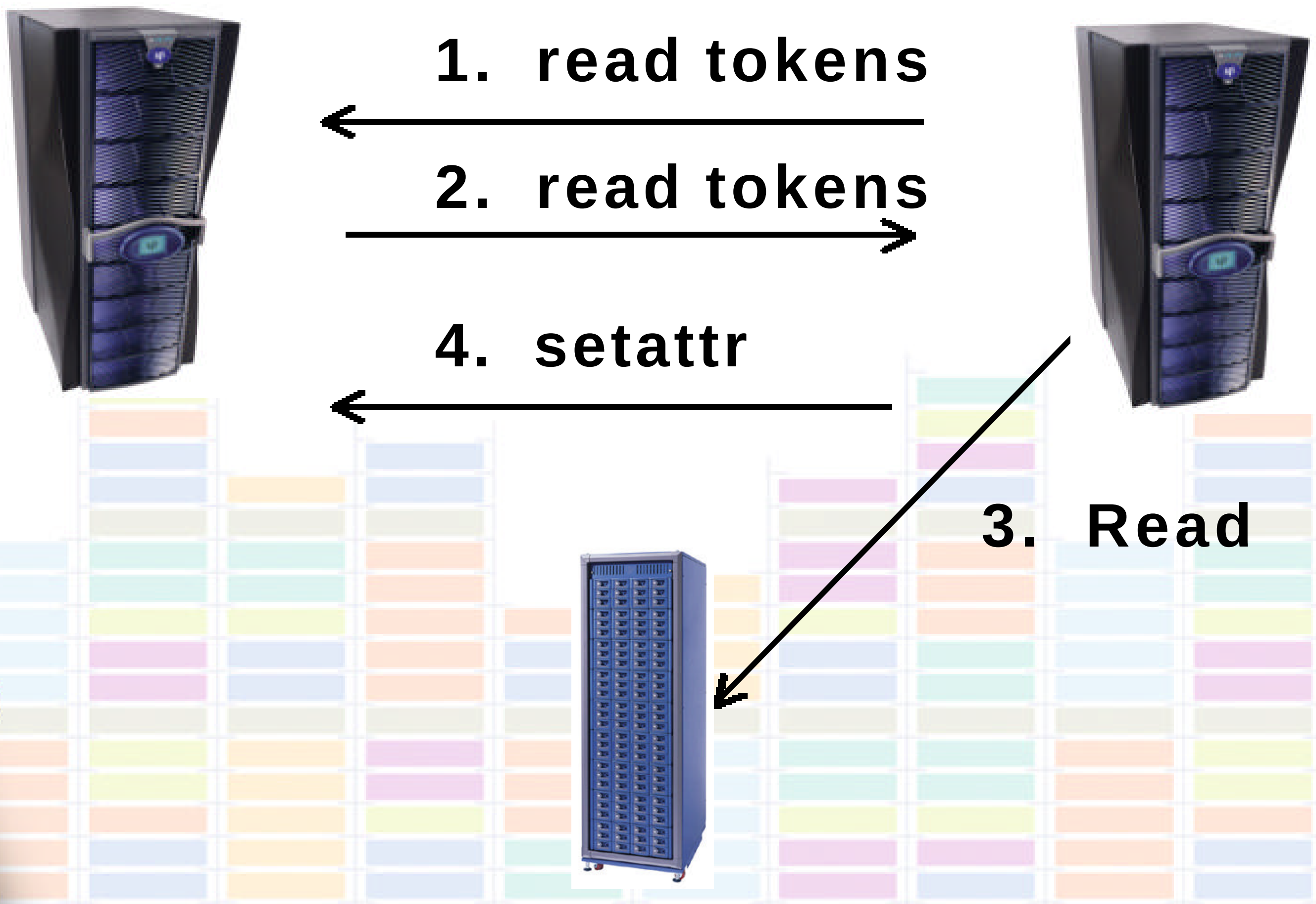
- Fully coherent buffering, as if a single system
  - Writes flush cashes on other nodes
- Compliant with POSIX file system calls
  - Including advisory record locking

No special record-locking libraries required

- For example: NFS supplies a separate non-POSIX record-locking library, which is not needed with CXFS.

CXFS Concepts

- Read Metadata Flow



Metadata Server

Metadata Client

1. read tokens

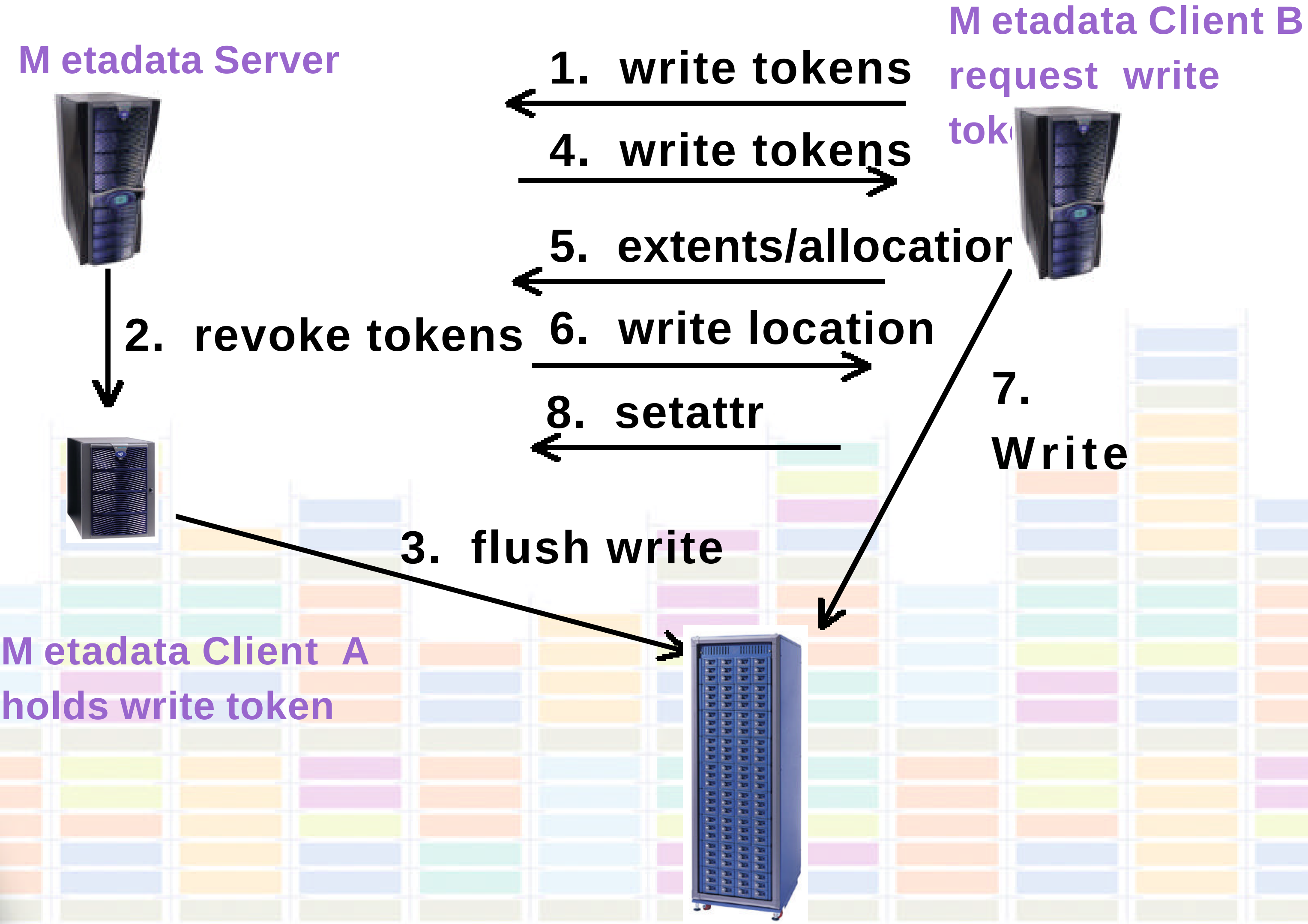
2. read tokens

3. Read

4. setattr

CXFS Concepts

- Write Metadata Flow



Metadata Server

Metadata Client B
request write
tok

1. write tokens

4. write tokens

5. extents/allocatior

6. write location

8. setattr

2. revoke tokens

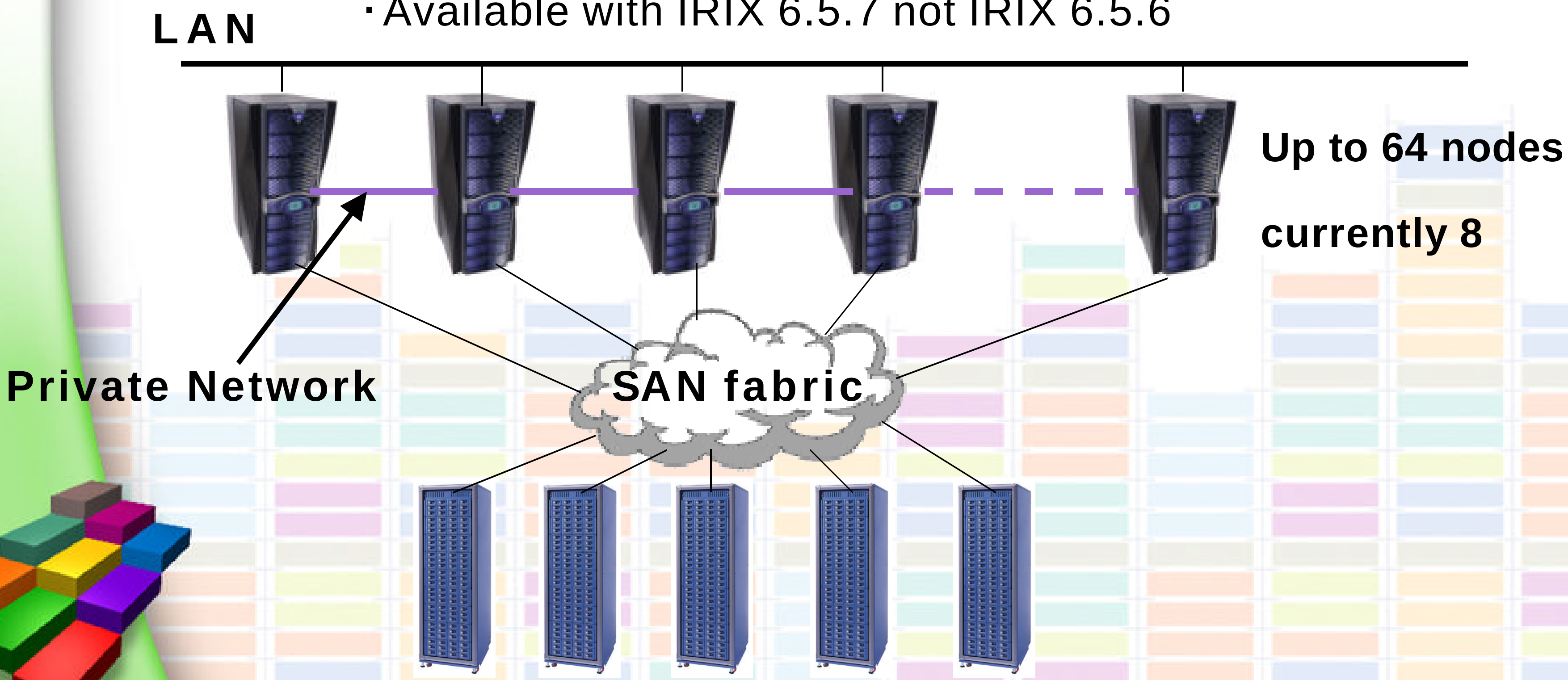
7.
Write

3. flush write

Metadata Client A
holds write token

CXFS Resilience and Performance

- **Single server manages metadata**
  - Backup metadata servers designated for fail-over
    - No single point of failure
    - Available with IRIX 6.5.7 not IRIX 6.5.6



CXFS Performance

- Optimal

- **When there are many:**
  - reads from and writes to a file that is opened by only one process
  - Reads from and writes to a file where all processes with that file open reside on the same host
  - Reads from a file where multiple processes on multiple hosts read the same file
  - Reads from and writes to a file using direct-access I/O for multiple processes on multiple hosts

CXFS Performance

- Not Optimal

- **Multiple processes on multiple hosts that are reading and writing the same file using buffered I/O**
  - direct-access I/O (e.g. databases) are okay
- **When there will be many metadata operations such as:**
  - Opening and closing files
  - Changing file sizes (usually extending a file)
  - Creating and deleting files
  - Searching directories

Real Life Benchmarks showed there is

almost no measureable difference

between XFS and CXFS read/write performance

CXFS Scalability

- **Supports up to 64 clients or servers per cluster**
  - IRIX 6.5.6 supports 8 clients
- **Multiple metadata servers can exist in a cluster**
  - One per file system
- **Files accessed exclusively locally on CXFS metadata server see local XFS metadata performance**

CXFS Summary (1/2)

- Supports guaranteed-rate IO and real-time file systems
  - For real-time and digital media applications
  - NOT on IRIX 6.5.9
- Fast recovery times: No fsck
- Avoids unnecessary writes by delaying writes as long as possible
- Contiguous allocation of disk space to avoid fragmentation
- 9 Peta Byte File System Size
  - If historical trends continue, will last 60+ years

CXFS Summary (2/2)

- **Fast directory searches**

- **Sparse file support**
  - Holes allowed in files for large direct-access addressing

- **DMAPI for Hierarchical File Systems (HFS)**
  - Interfaces to SGI's Data Migration Facility (DMF) and third-party HSMs: Veritas, FileServ, ADSM
  - Available on IRIX 6.5.8

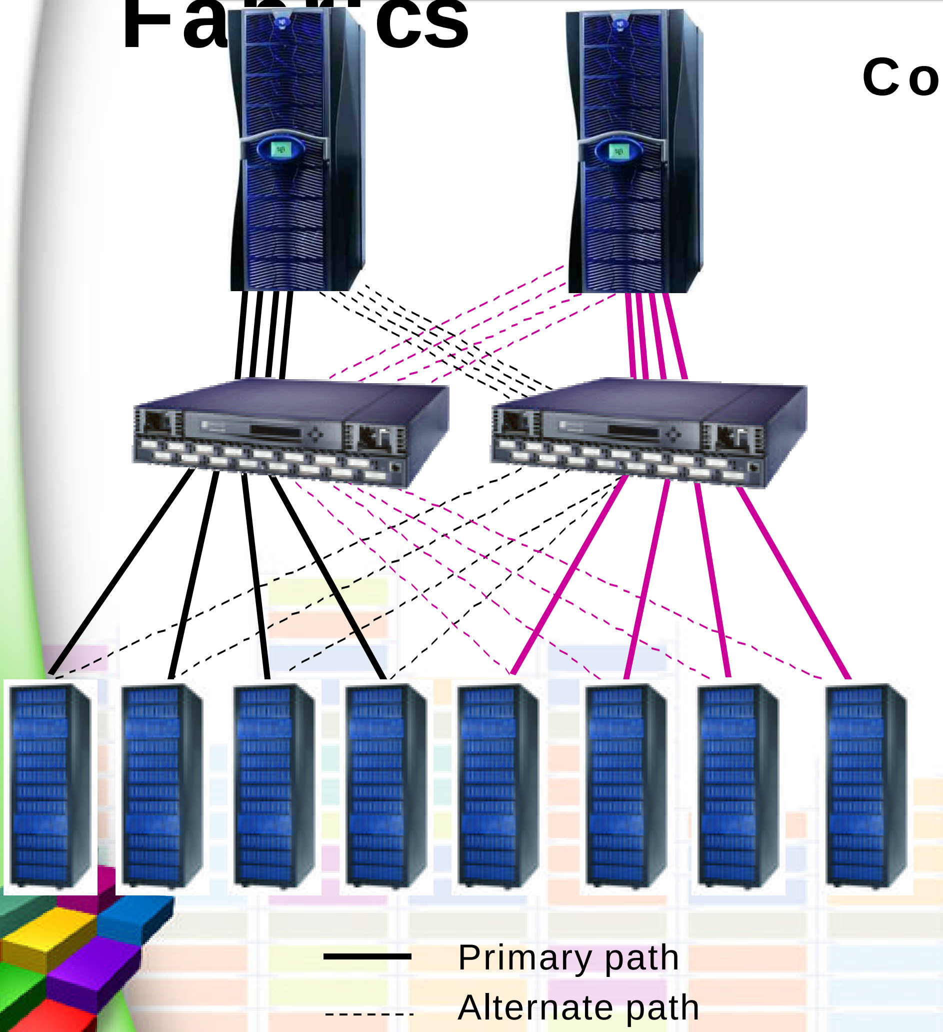
Agenda

- **Introduction**
  - What is a Storage Area Network aka SAN?
    - Fibre Channel Technologies and Topoligies
    - The Fibre Channel Fabric

- **CXFS, Delivering on the Promise**
  - CXFS Overview
  - CXFS Concepts
  - CXFS Performance

➡ **CXFS, Serving Advanced Environments**
  - High Availability
  - HSM

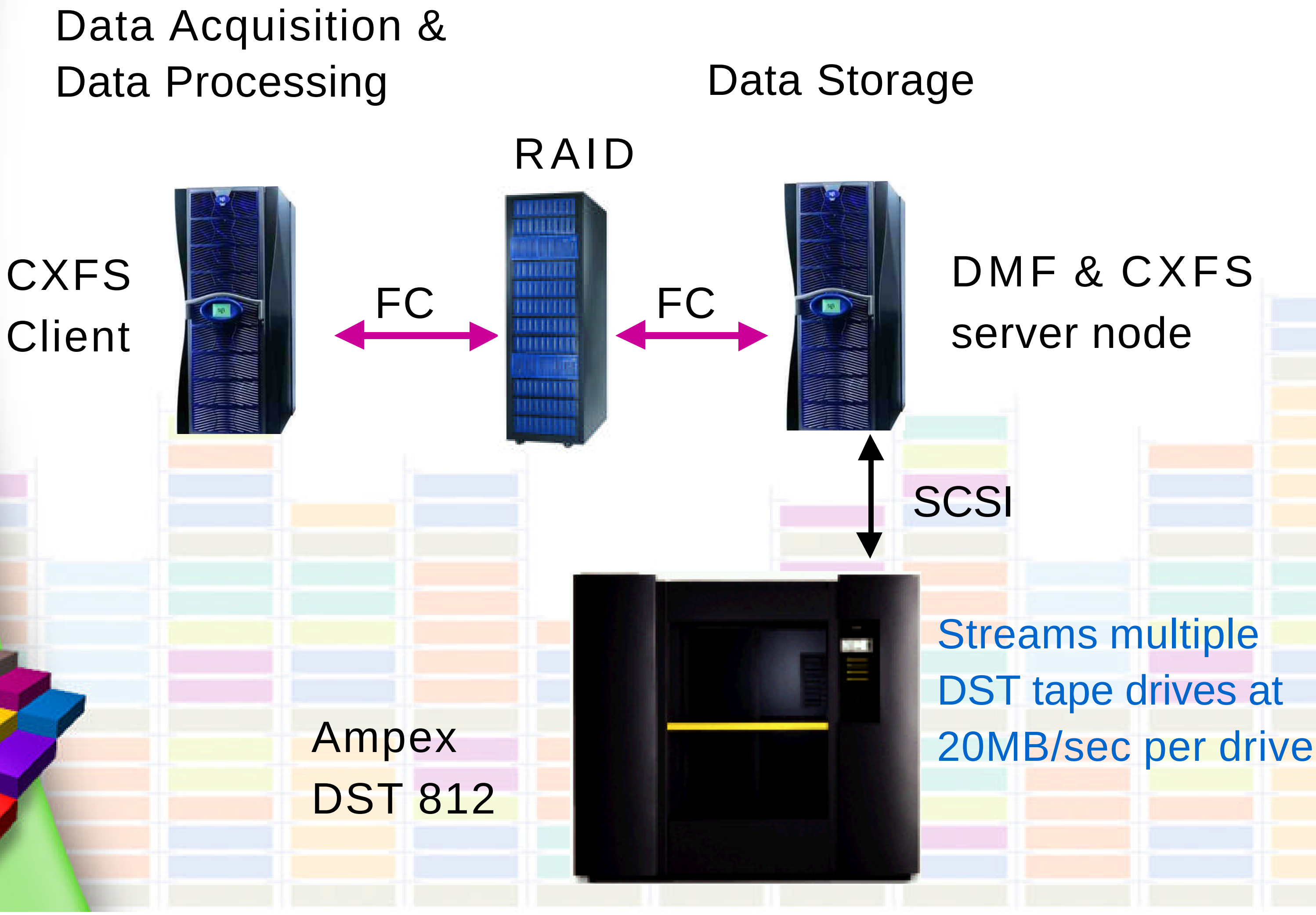
High Availability 16 Port Fabrics



Configuration attributes

- Each Origin system has 4 primary paths and 4 alternate paths to Fibre Channel RAID storage
- Each alternate path is via a separate HBA , switch fabric and storage controller
- Each system has access to any storage
  - in a failover situation
  - for backup of data
  - for CXFS access to shared data

CXFS/DMF Example



Data Acquisition &
Data Processing

Data Storage

RAID

CXFS
Client

FC

FC

DMF & CXFS
server node

SCSI

Ampex
DST 812

Streams multiple
DST tape drives at
20MB/sec per drive

sgi™

sgi™

One step ahead