CXFS
Clustered File System from SGI

April 1999

R. Kent Koeninger
Strategic Technologist, SGI Software
kentk@sgi.com
www.sgi.com

File Systems Technology Briefing

UNIX (Irix)

Applications

CXFS

XFS

XVM

FC driver

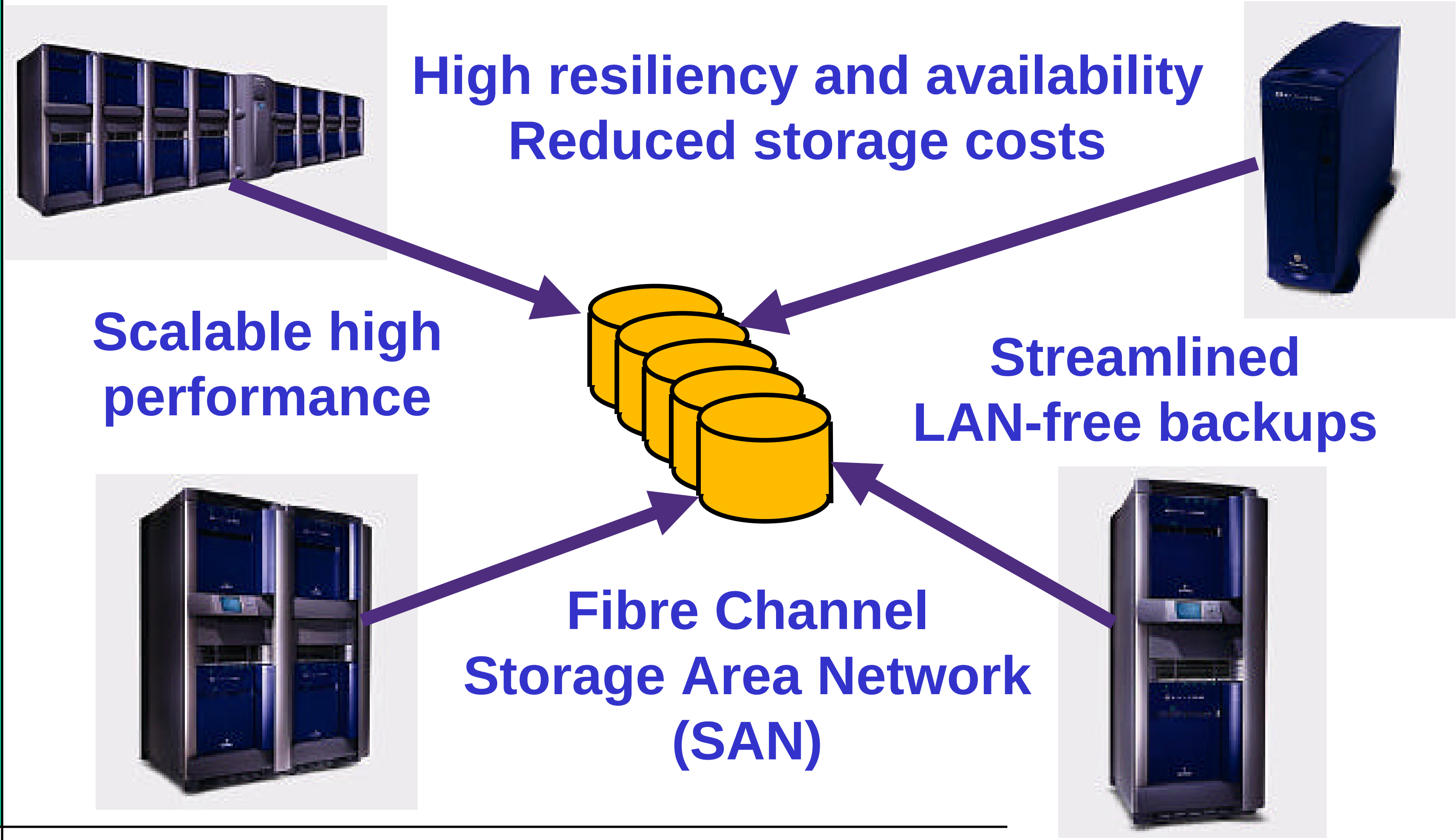
- **Clustered file system features: CXFS**
- **File System features: XFS**
- **Volume management: XVM**

sgi™

Clustered File Systems

CXFS

CXFS — Clustered SAN File System



High resiliency and availability
Reduced storage costs

Scalable high performance

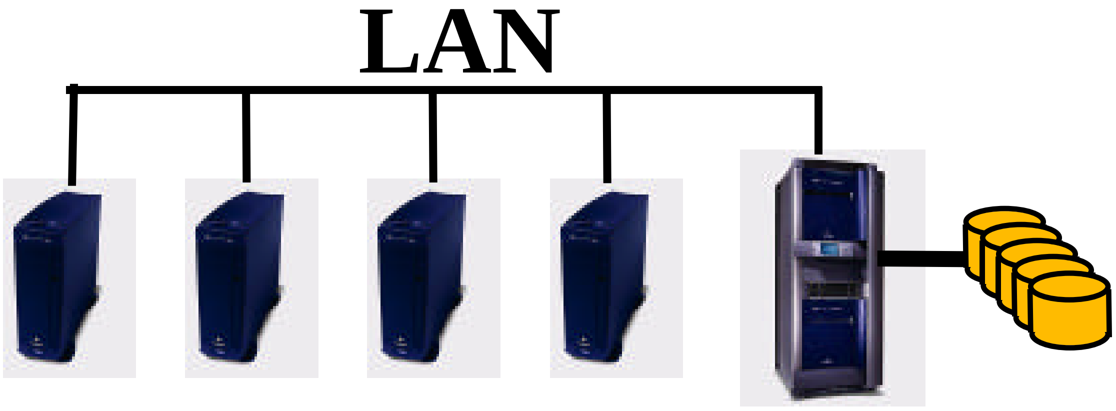
Streamlined LAN-free backups

Fibre Channel
Storage Area Network
(SAN)

CXFS: Clustered XFS

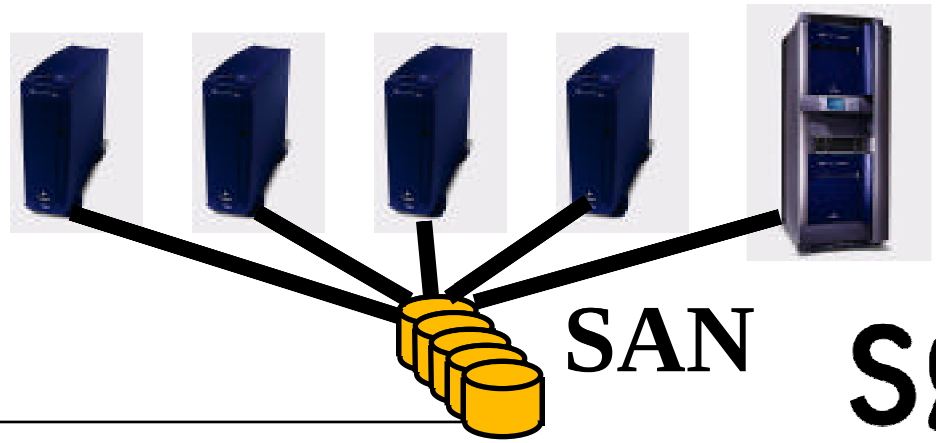
- **Clustered XFS (CXFS) attributes:**
  - A **shareable high-performance** XFS file system
    - Shared among multiple IRIX nodes in a cluster
    - Near-local file system performance.
      - Direct data channels between disks and nodes.
  - A **resilient** file system
    - Failure of a node in the cluster does not prevent access to the disks from other nodes
  - A **convenient** interface
    - Users see standard Unix filesystems
      - Single System View (SSV)
      - Coherent distributed buffers

Comparing LANs and SANs



LAN: Data path through server (Bottleneck, Single point of failure)

SAN: Data path direct to disk (Resilient scalable performance)

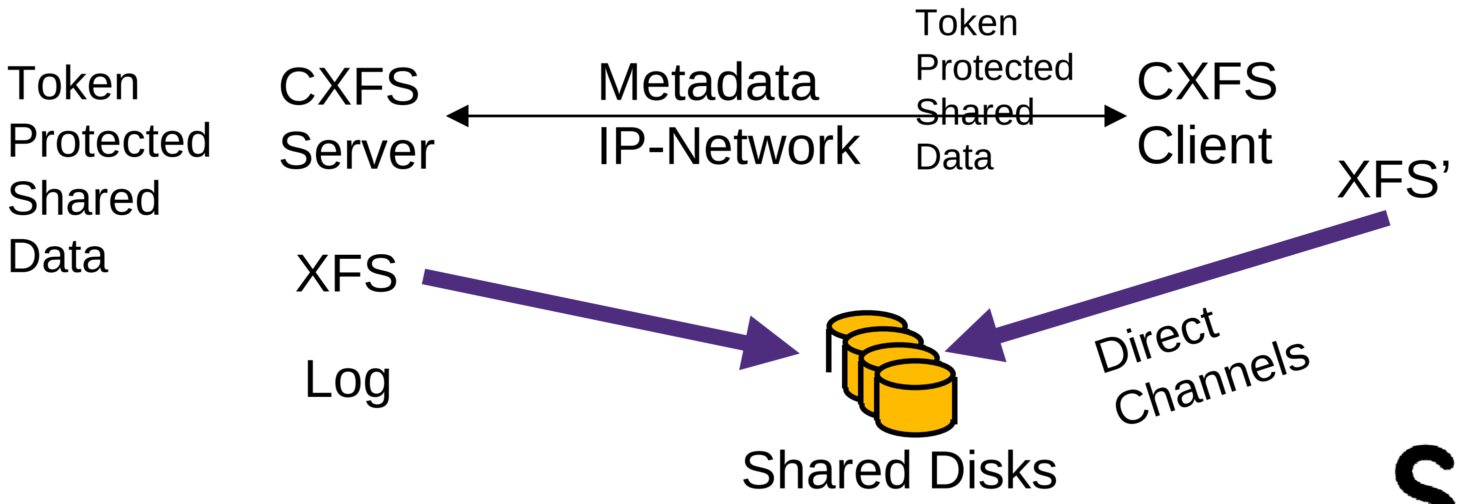


CXFS Server Node

Coherent
System
Data Buffers

CXFS Client Node

Coherent
System
Data Buffers



Token
Protected
Shared
Data

CXFS
Server

XFS

Log

Metadata
IP-Network

Token
Protected
Shared
Data

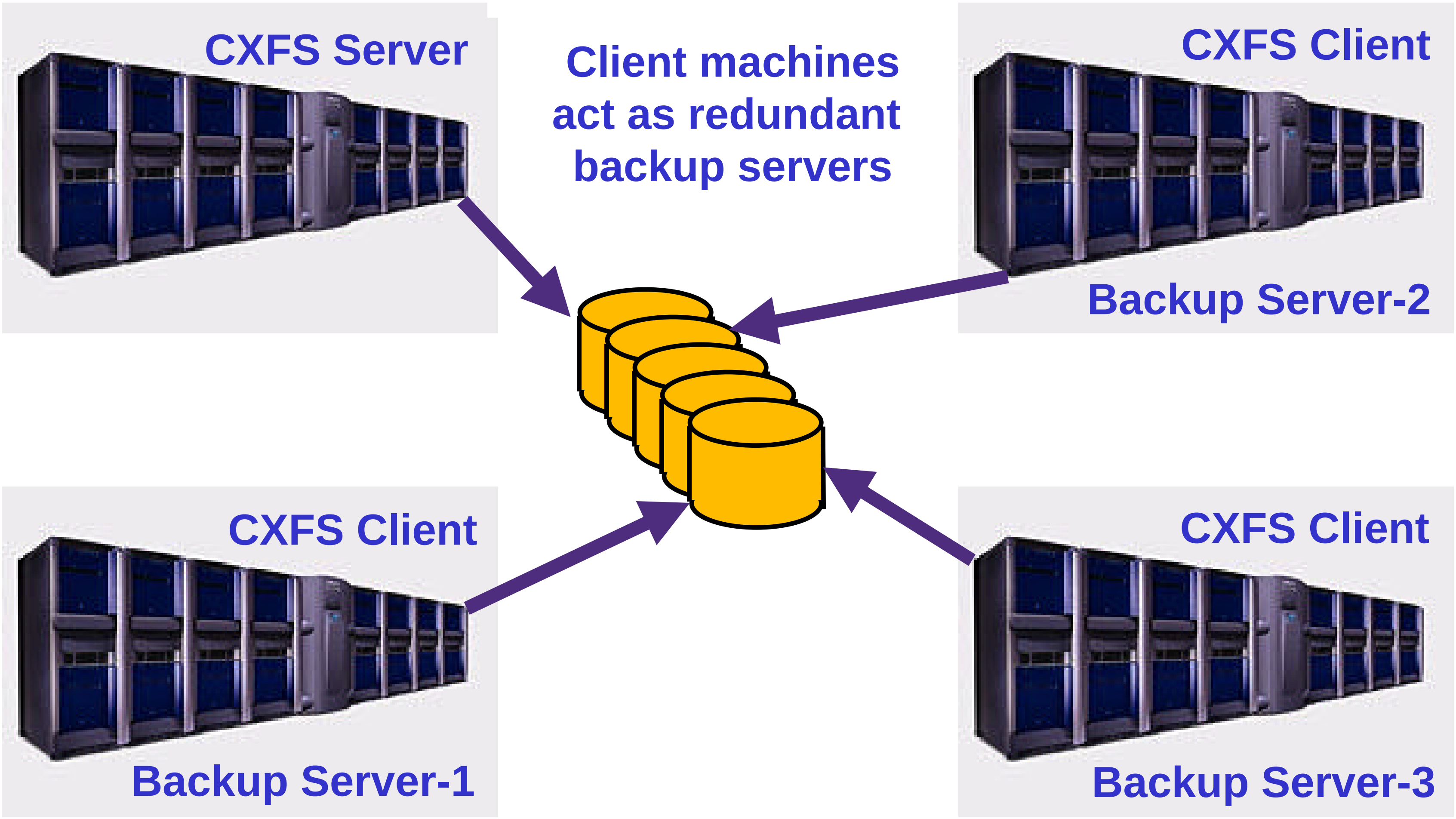
CXFS
Client

XFS’

Direct
Channels

Shared Disks

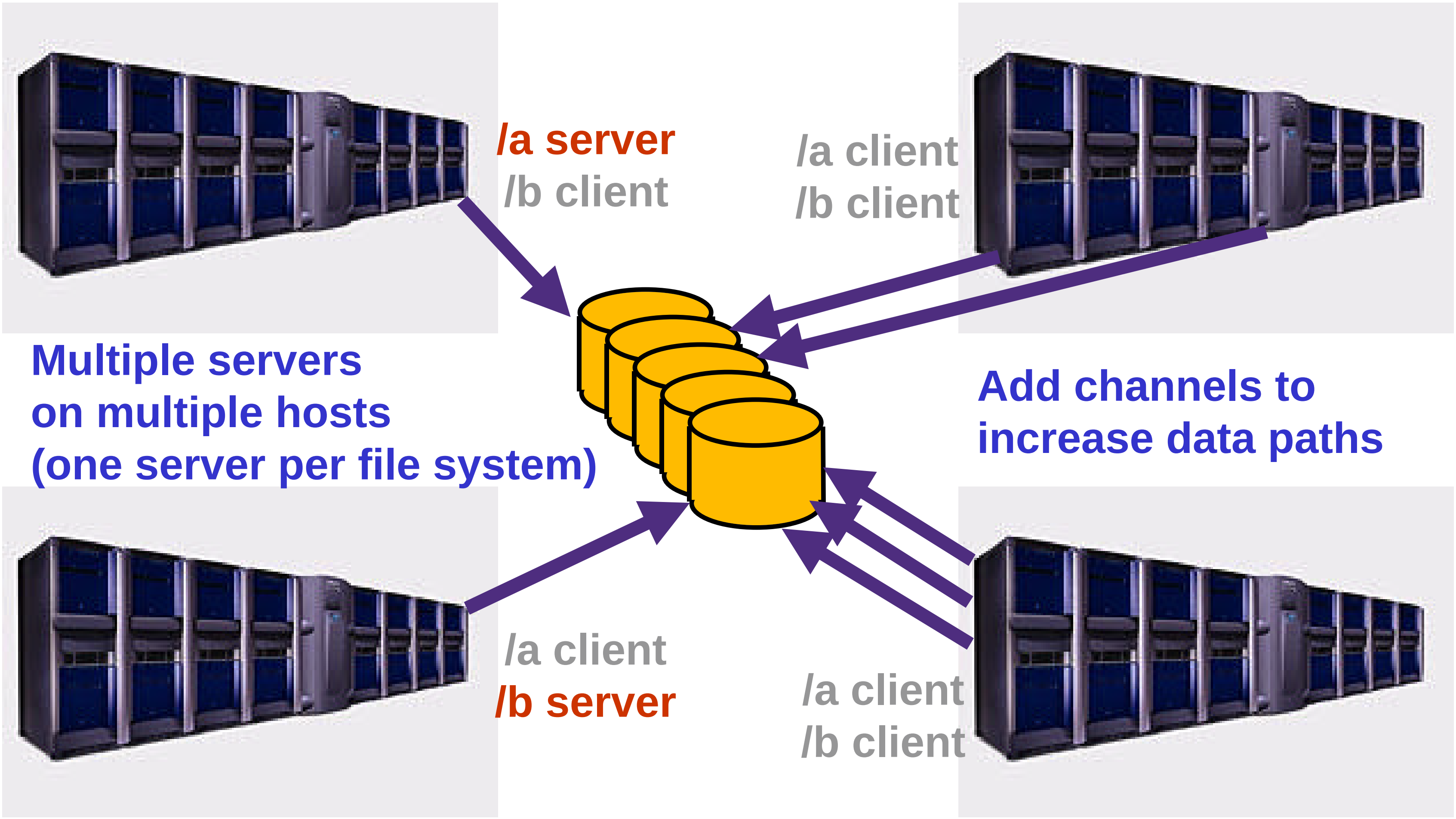
Fully Resilient - High Availability



CXFS Interface and Performance

- **Interface is the same as multiple processes reading and writing shared files on an SMP**
  - Same open, read, write, create, delete, lock-range, etc.
- **Multiple clients can share files at local file speeds**
  - Processes on the same host reading and writing (buffered)
  - Processes on multiple hosts reading (buffered)
  - Processes on multiple hosts reading and writing, using direct-access IO (non-buffered)
- **Transactions slower than with local-files:**
  - Shared writes flush distributed buffers related to that file
  - Metadata transactions (file creation and size changes)

CXFS Scalability



Multiple servers
on multiple hosts
(one server per file system)

Add channels to
increase data paths

CXFS Scalability

- **Software supports up to 64 clients or servers per cluster**
  - Fabric prices will tend to limit the host count to less-than 64
- **Multiple CXFS servers**
  - One per file system
- **Normal local-host buffering for near local-file performance**
  - Except when files are used for shared-reads-writes
    - Coherence maintained on a per I/O basis using tokens
- **Files accessed exclusively locally on CXFS server see local XFS metadata performance (bypasses CXFS path)**
- **CXFS supports High-Availability (HA) environments with full fail-over capabilities**
- **CXFS sits on top of XFS: Fast XFS features**

Heterogeneous CXFS

- **IRIX servers and clients in second half of 1999**
  - IRIX-XFS/XVM performance and file-system features
- **Clients for Windows NT, Linux and other major UNIX system in 2000**
  - Performance and features may be limited by particular OS interfaces
- **Servers for Linux and possibly other OSes to follow**

File Systems

XFS

sgi™

XFS: A World-Class File System

- **Speed**
  - Fast metadata speed
  - High bandwidths
  - High transaction rates
  - Guaranteed-rate IO and real-time file systems
- **Reliability**
  - Mature log-based file system
- **Scalability**
  - 64 bit: 9 million terabytes
- **Flexibility**
  - Dynamic allocation of metadata space

Fast Metadata Transactions

- **Efficient log-based transactions**
- **Rapid recovery from system interruptions**
  - Avoids FSCK (many minutes on other file systems)
  - Sub-second file-system recovery times
- **Efficient metadata techniques**
  - Structured for fast searches
  - Rapid space allocation techniques

XFS Metadata Performance

- **Fast crash recovery**
  - Log based: No fsck
- **Supports extremely large file systems**
  - 64 bits and scalable structures
- **Supports large sparse files**
  - Full 64 bit direct addressing
- **Supports large contiguous files**
  - Efficient search algorithms and data structures
- **Supports large directories**
  - Efficient B-trees
- **Supports large numbers of files**
  - Dynamic allocation of inode space

XFS: Reliable and Quick Recovery

- **Database log technology used for file system meta-data management**
  - –No UNIX ***fsck*** is needed
- **High file system integrity**
- **Recovery time is independent of system size**
  - –Depends on system activity levels
  - –Generally recovery requires only a few seconds
- **Very high file system performance:**
  - –Log implemented with advanced techniques that use fewer I/O operations than standard UNIX

Efficient Physical IO

- **Avoids unnecessary writes**
  - Asynchronous buffering
    - Delay writes as long as possible

- **Contiguous allocation of disk space**
  - Delay allocation of disk space by delaying writes
    - Avoids fragmentation
    - Tends to allocate large contiguous segments

- **Well orchestrated data paths and buffer**
  - Through volume manager on operating system

Scalability: Room to Grow

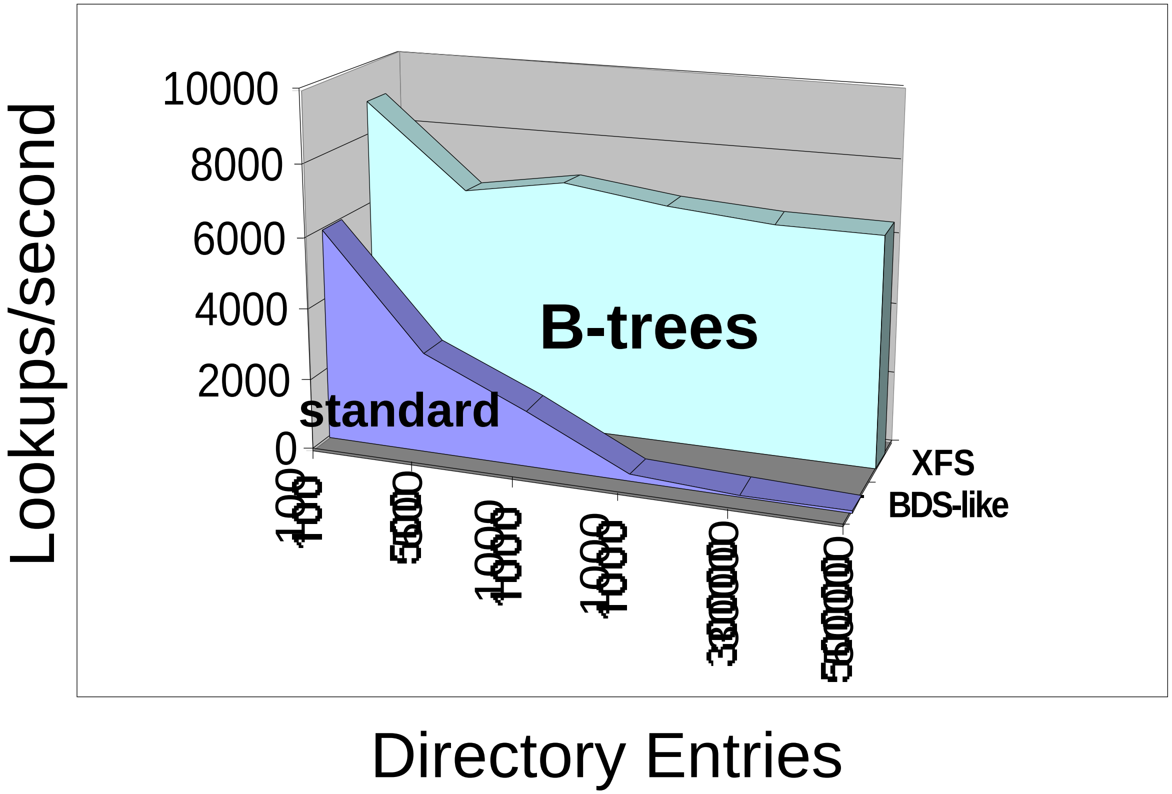
- **Disk drive growth remains exponential**
  - Recently 1.7 x per year
  - Historically 10x every 10 years
- **XFS' 64 bit address space exceeds even this projected exponential growth far into the future**
  - 2^{63}-1 = 9 million terabytes = 9 exabytes
- **The price of the storage hardware and the channel capacity of the hosts are likely to be the limiting factors for growth, not XFS**

Scalable Performance

- **Peruse large file systems rapidly**
  - B-tree structures and other sophisticated techniques
- **Supports huge file systems**
  - Large amounts of data
    - Huge numbers of files
    - Huge files
  - Large numbers of disks
  - Large file systems
    - Striped, mirrored, and concatenated file systems

sgi™

XFS B-tree Directory Speed



- XFS supported 1 M entries at 66 lookups/second

XFS Data Bandwidth

- **XFS delivered near raw I/O performance on the largest disk configuration we have been able to test**
  - **Over 4 Gbytes/second (read and write)**

- **Configuration:**
  - 88 Fibre Channel loops, 8 disks per loop: 704 disks
  - One process—one file descriptor
    with parallel asynchronous I/O to a real-time file system
  - 32 processor Origin2000 system

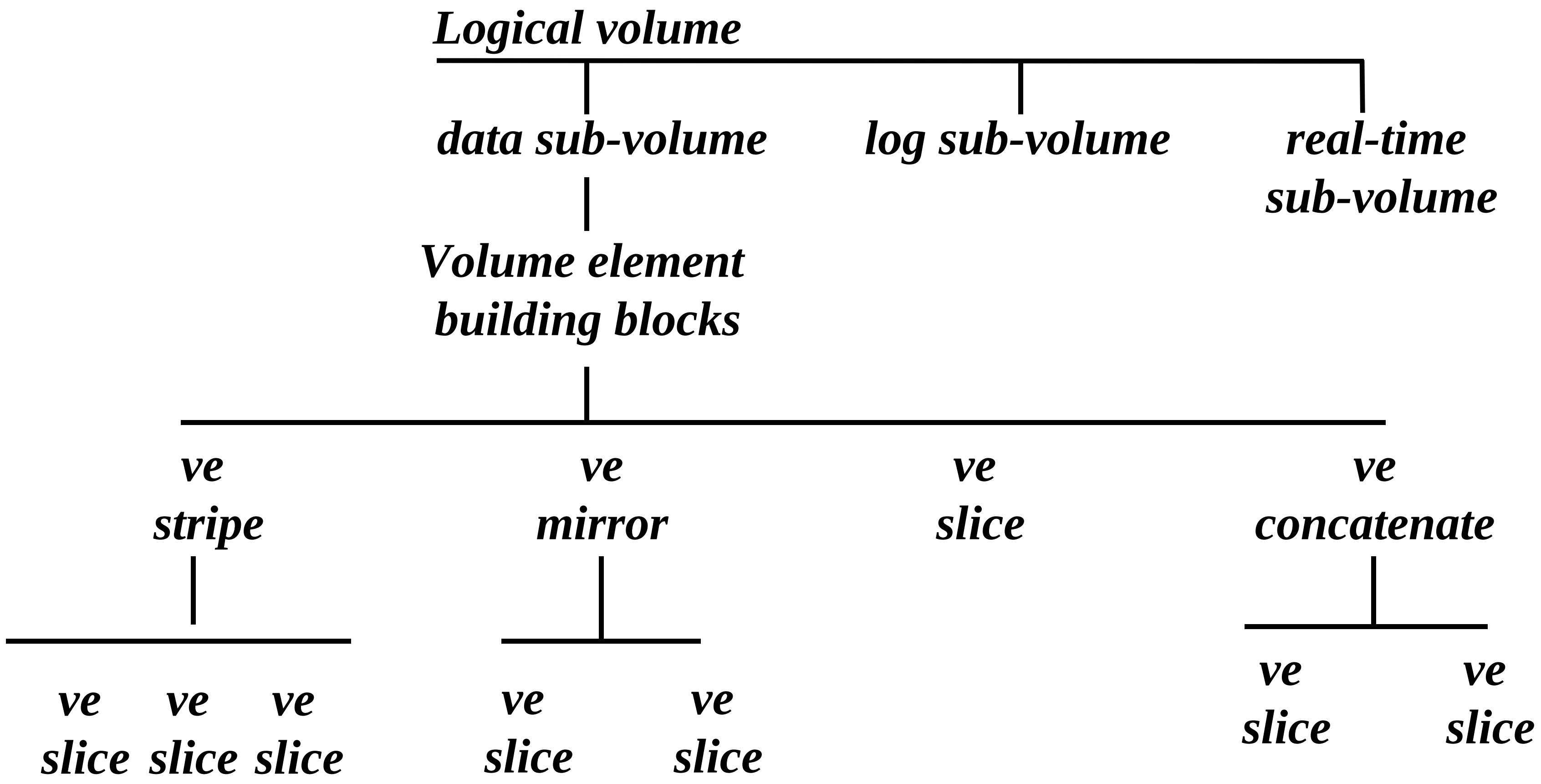
Other XFS Features

- **Guaranteed ratio IO (GRIO)**
  - Important for guaranteeing bandwidth for real-time and digital media applications
- **Optimizations for real-time files**
- **Sparse file support**
  - Holes allowed in files for large direct-access addressing
- **Parallel space allocation increases speed**
  - Fastest minute-sort and fastest terabyte-sort benchmarks
- **DMAPI for hierarchical file systems (HFS)**
  - Interfaces to SGI’s Data Migration Facility (DMF) and third-party HSMs: Veritas, FileServ, ADSM

Volume Management

XVM

XVM: Volume Management



Logical volume

data sub-volume

log sub-volume

real-time sub-volume

Volume element building blocks

ve stripe

ve mirror

ve slice

ve concatenate

ve slice *ve slice* *ve slice*

ve slice *ve slice*

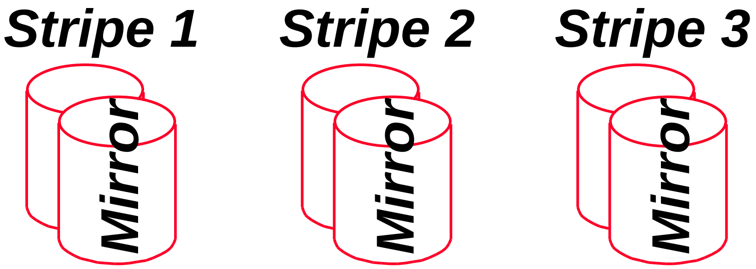
ve slice *ve slice*

XVM - Volume Management

- **Striping, mirroring, and concatenation of volume elements**
  - Flexible combinations of mirroring and striping
- **Thousands of disks: E,g,. 64K stripe width**
  - Practically unlimited
- **Self identifying volumes**
- **Subvolumes separate data, log, and real-time information**
- **On-line configuration changes**
- **Clustering support (multi-host volume sharing)**

XVM Flexible Combinations

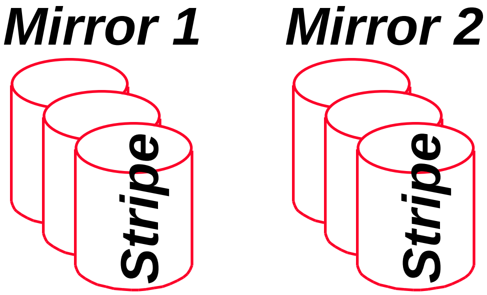
Striped Mirrors



Stripe 1 Stripe 2 Stripe 3

Mirror Mirror Mirror

Mirrored Stripes



Mirror 1 Mirror 2

Stripe Stripe

XVM Performance

- **Performance measured on XVM predecessor: XLV**
  - With modifications to XLV
- **Same hardware configuration as in previous XFS performance slide**
  - 88 Fibre Channel loops, 704 disks, 32 PE O2K
- **Near raw I/O disk speed**
  - Over 4 Gbytes/second (read and write)

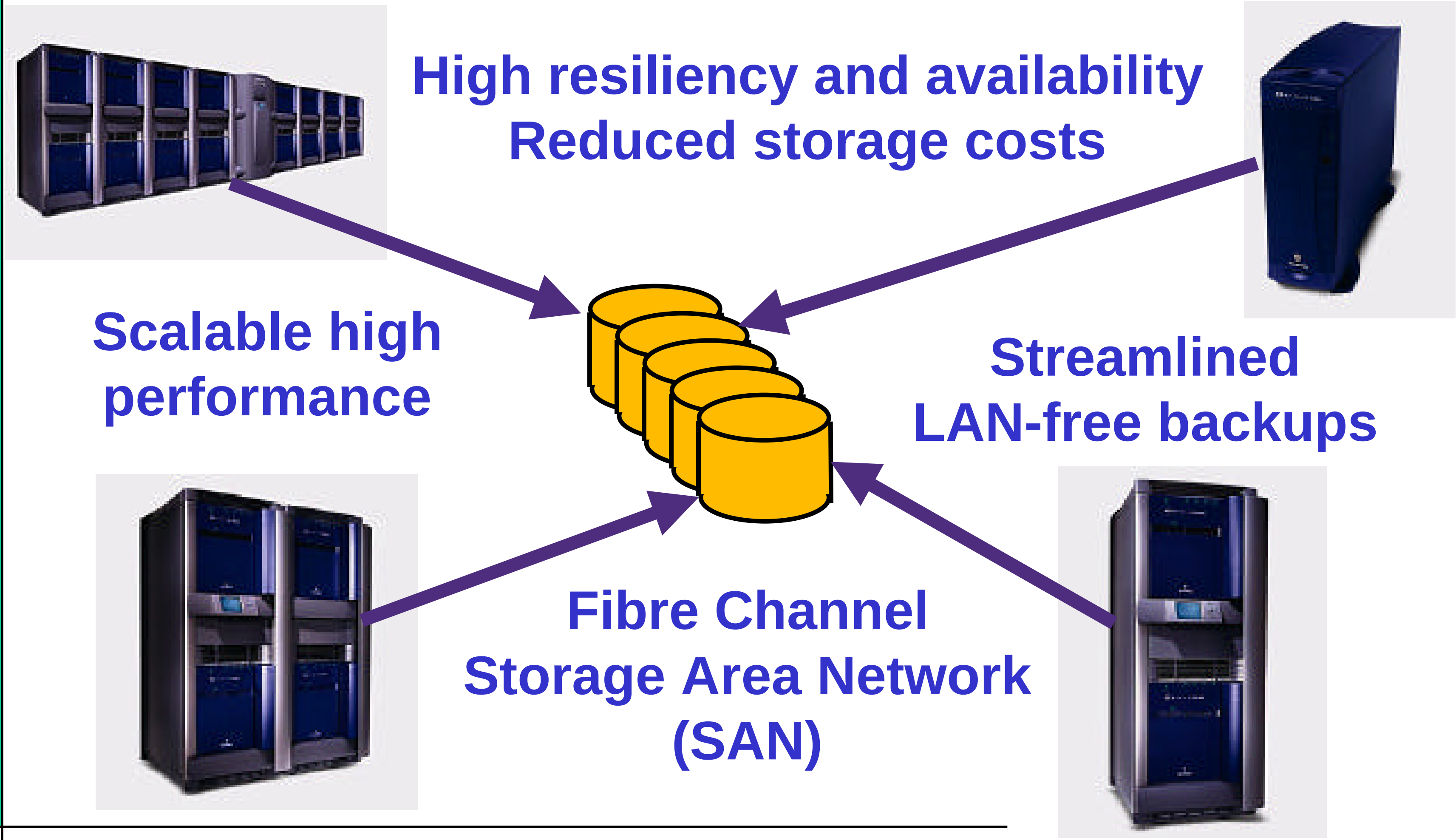
Summary

- **CXFS is the highest-performance shared-file system**
  - With full resilience (High Availability)

- **XFS and XVM are fastest and most scalable file-system and storage management technologies available**
  - High bandwidth, fast metadata, fast recovery, flexible, huge address space, huge volume capacity, feature rich

sgi™

CXFS — Clustered SAN File System



High resiliency and availability
Reduced storage costs

Scalable high performance

Streamlined LAN-free backups

Fibre Channel
Storage Area Network
(SAN)