Spinnaker Networks, Inc.

Spinnaker Technology Overview

SPINNAKER NETWORKS

Overview

- Goals of Next Generation NAS systems
- Architectures
- Storage Applications
- Spinnaker's system

Next Generation NAS Features

- Online NAS scaling
  - add resources easily
    - network ports, storage, servers
  - efficiently scale to 100s of servers
  - no file name or mount changes
  - no disruption to concurrent accesses
- Online reconfiguration
  - change storage properties/location for data
  - move clients within the cluster

SPINNAKER
NETWORKS

Next Generation NAS Features

- Multiple site support
  - single management point for multiple sites
    - management updates queued during partitions
  - designed to work over WAN
    - localize data for LAN-speed access to data
    - predictable behavior in case of WAN link failure

Design Options

- Meta-data servers
- Distributed locking servers
- NAS switching
- Integrated NAS switching

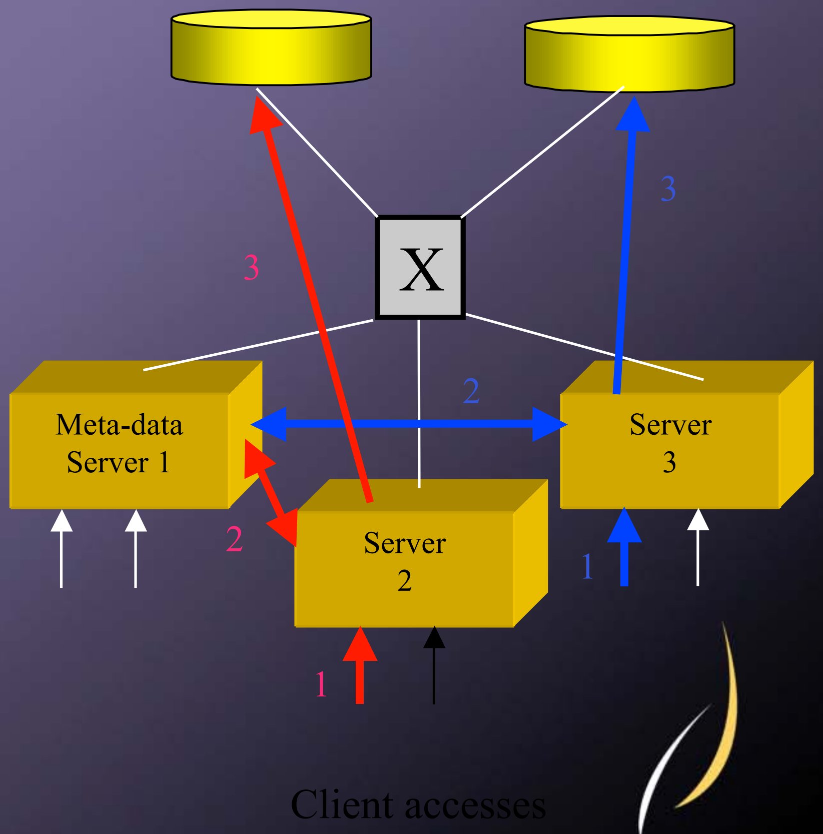
SPINNAKER
NETWORKS

Design Choices

- What is a meta-data server
  - all meta-data processed at one server
  - IO processing distributed among storage clients
    - reads and writes still coordinate with meta server
- What is distributed locking
  - each server obtains locks to process ops locally
    - locks obtained from locking server (or smart drives)
    - locks obtained for block allocation, inode allocation, directory modifications

Meta-data servers

- Meta-data server
  - perform dir ops
  - performs block allocation
  - coordinates R/W ops
- Locking server
  - lock server grants locks to others servers
  - may have to do revokes

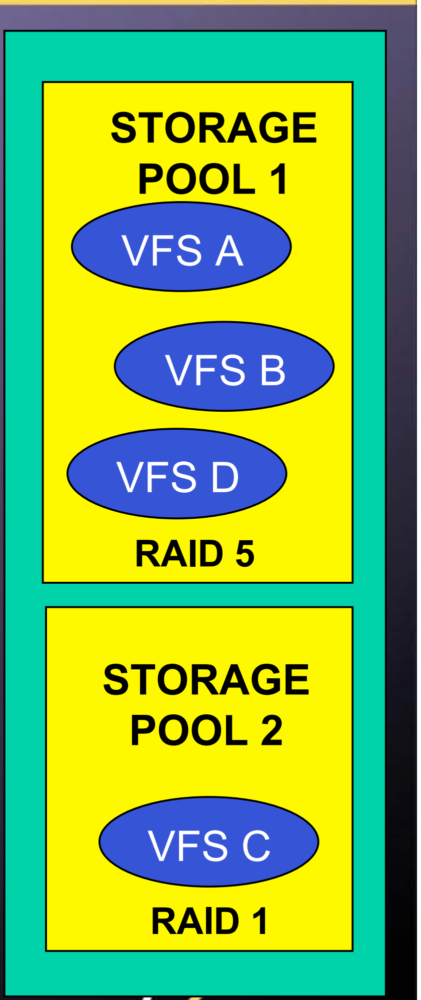


Design Choices

- What is NAS switching
  - analogous to SAN switching
  - lots of NAS service ports
  - divide export name space into subtrees (VFSes)
  - forward request to specific server
    - based on VFS
    - and based on geography (if more than one copy)
- Separate switch v.s. integrated switching
  - separate switch element or
  - integrated switch/file server pair

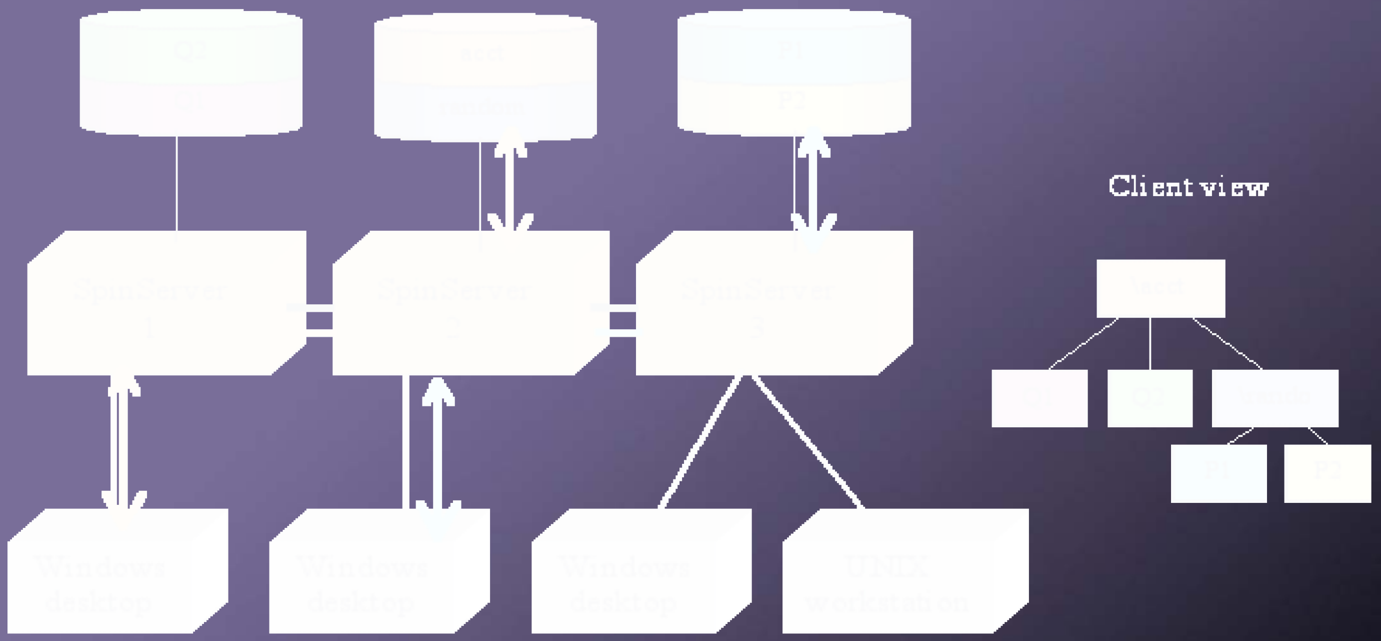
NAS switching

- Each storage pool is a collection of one or more RAID sets
- One or more VIRTUAL FILE SYSTEMS (VFS) created in each storage pool.
- Requests are switched based on VFS
- Each VFS may be placed in a storage pool with attributes to meet business requirements



SPINNAKER NETWORKS

NAS switching



Design Choices-Scaling

- Scale to 100s of machines
  - meta-data server is bottleneck
    - scale to 4-10 level, but every op seen by meta server
  - distributed locking server is bottleneck
    - most operations will actively communicate with lock server
      - write ops must get dir locks, allocation locks
      - even read ops must synchronize with delete, write
    - caching locks does not help much
      - 2nd level caches are fairly ineffective
      - may hurt, in the presence of lock revokes
    - scales similarly to meta-data server
  - suited for reading/writing large blocks of large files
  - NAS switches scale similar to switched networks

Design Choices-Failure Modes

- Failure isolation
  - Meta-data and distributed locking servers
    - can lose entire cluster with one bad server
    - must repair entire cluster file system together
      - repair (“fsck”) time doesn’t scale with cluster size
  - Switched NAS
    - each server’s file system controlled by one server
    - limits damage an errant server can do

SPINNAKER
NETWORKS

Design Choices-Storage Classes

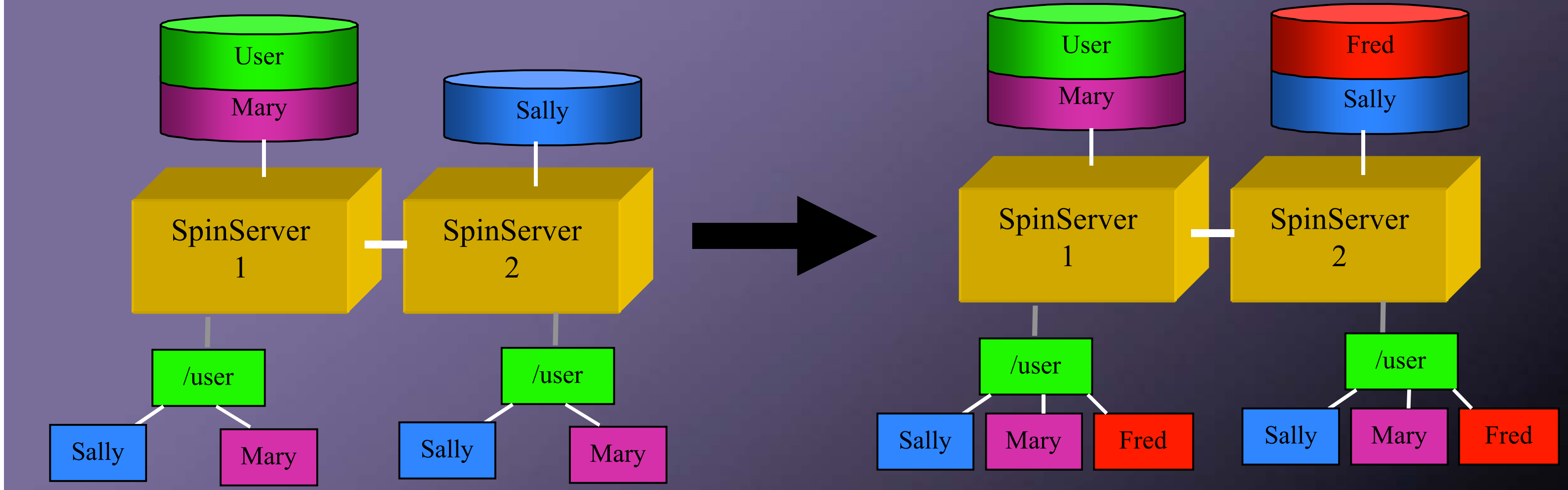
- Storage service classes are important
  - manage performance, cost, data redundancy
- Meta-data & distributed locking servers
  - scales with a large homogeneous pool of blocks
    - does not help with managing service classes
  - service classes require additional layer of abstraction
- VFSes provide ideal foundation for service classes
  - storage pools are defined by service classes
  - VFSes inherit service class from containing pool
  - move VFS to new storage pool as desired class changes

Storage Applications

- Large user community
- Branch office / WAN support
- Parallel data processing
- Storage Server Class Management
- Databases
- NAS consolidation

Large User Community

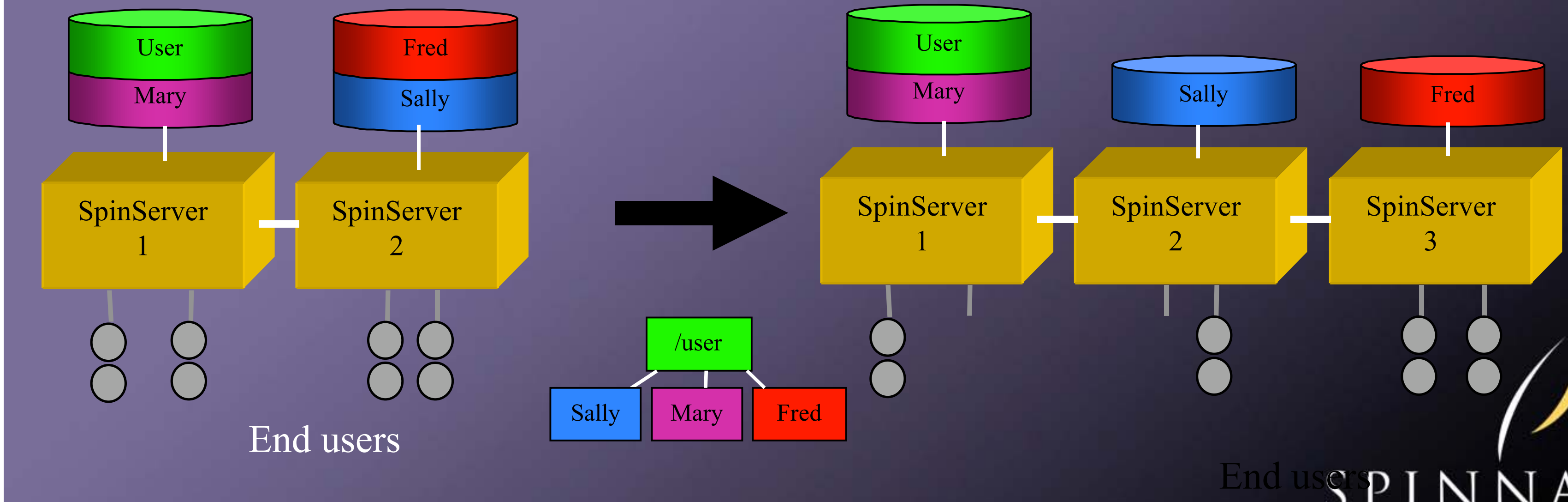
- Add users with their own VFSes
  - each gets own quota and minimum reservation
  - users added to servers with free space
  - server location independent of file name



SPINNAKER
NETWORKS

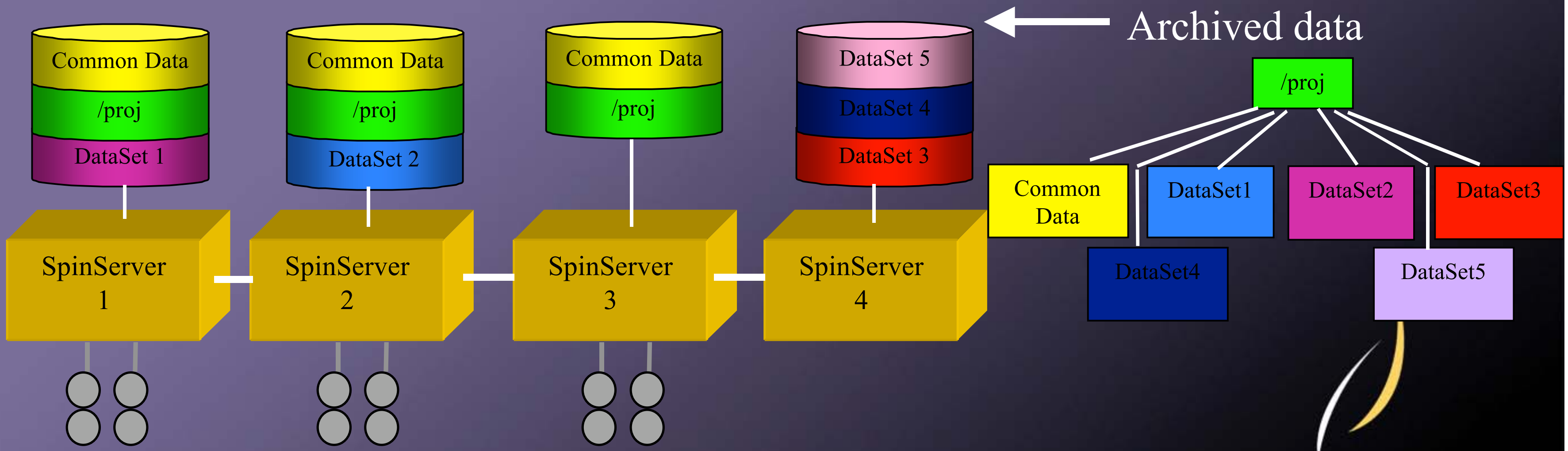
Easy Capacity Scaling

- Add new servers and capacity
  - relocate online data without user visible changes
  - relocate clients transparently between physical interfaces
  - Online relocation happens without client disruption
- Saves costs of working around outages



Integrate Low Cost Storage

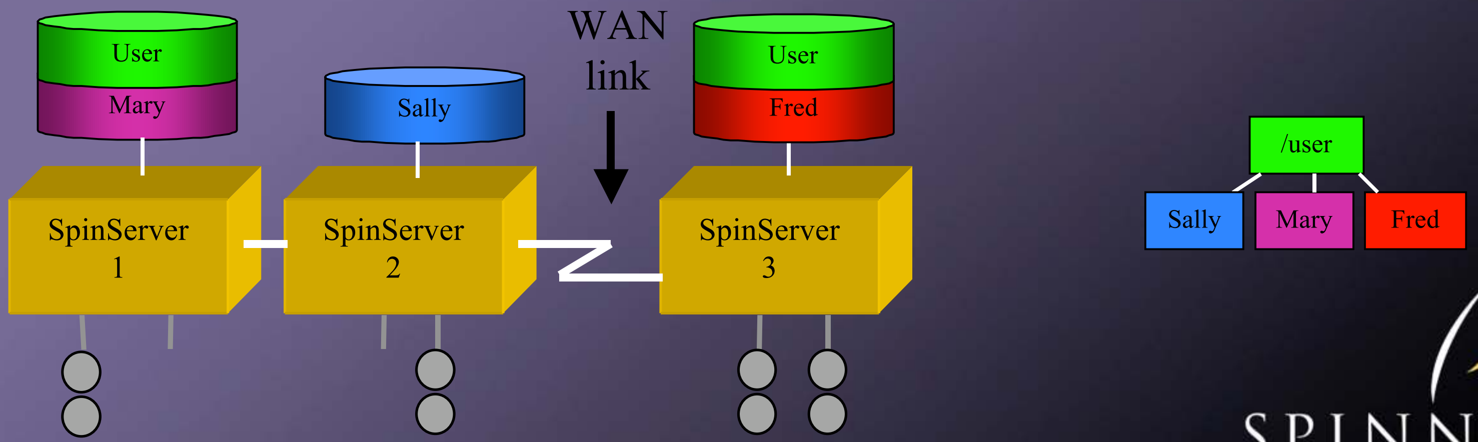
- Online archiving to inexpensive storage
    - active projects reside in high performance storage pools
    - inactive projects moved to lower cost storage pools
    - no name space changes, no disruption of accesses
- Saves $ through appropriate use of expensive and cheap storage



SPINNAKER NETWORKS

Branch Office

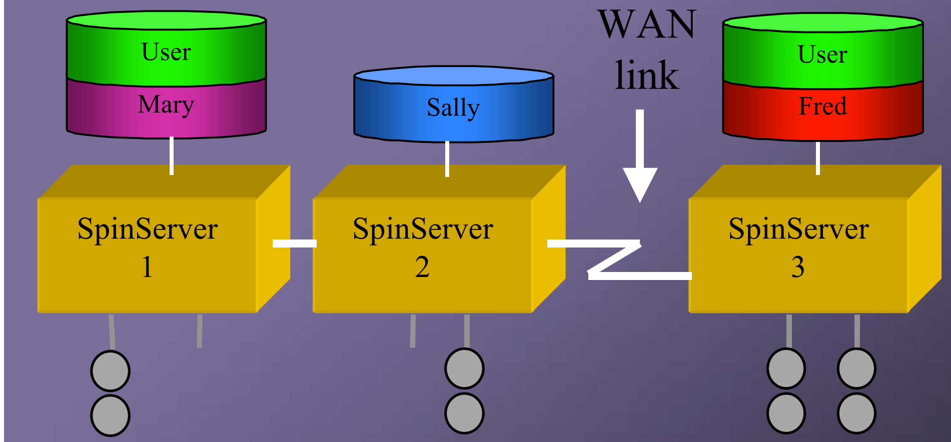
- Single shared name space
  - branch office has local data for high speed access
  - convenient for mobile users

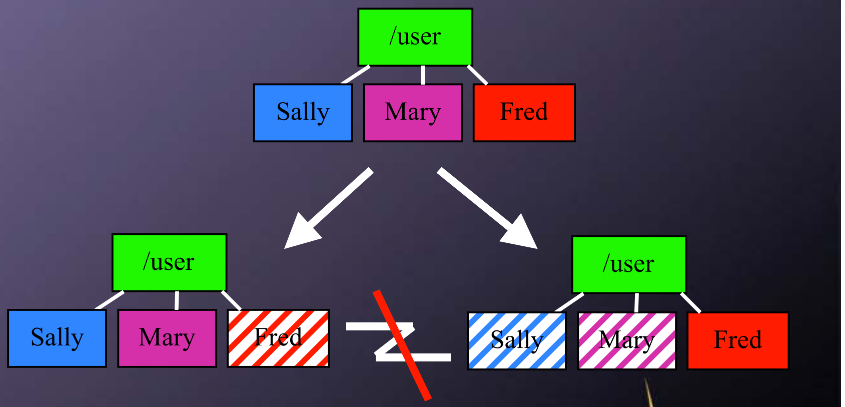


SPINNAKER
NETWORKS

Branch Office

- System divides during WAN link failures
  - top directory (/user) replicated for speed, availability
  - branch office has uninterrupted access to local data
  - management requests queued for branch office

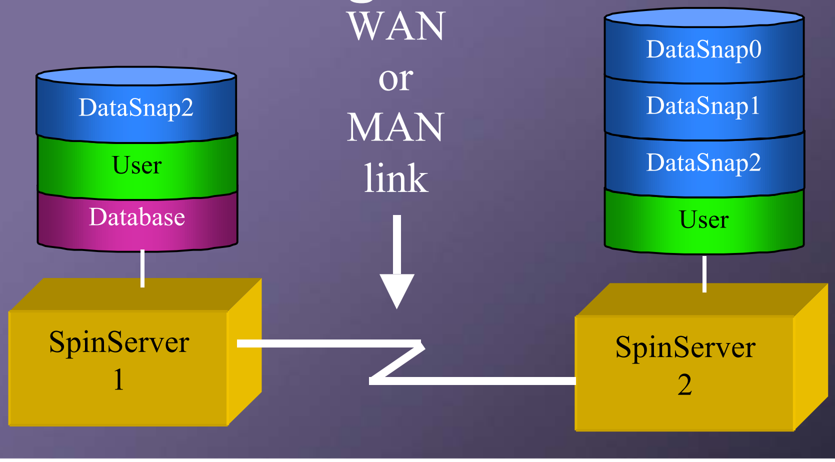




SPINNAKER
NETWORKS

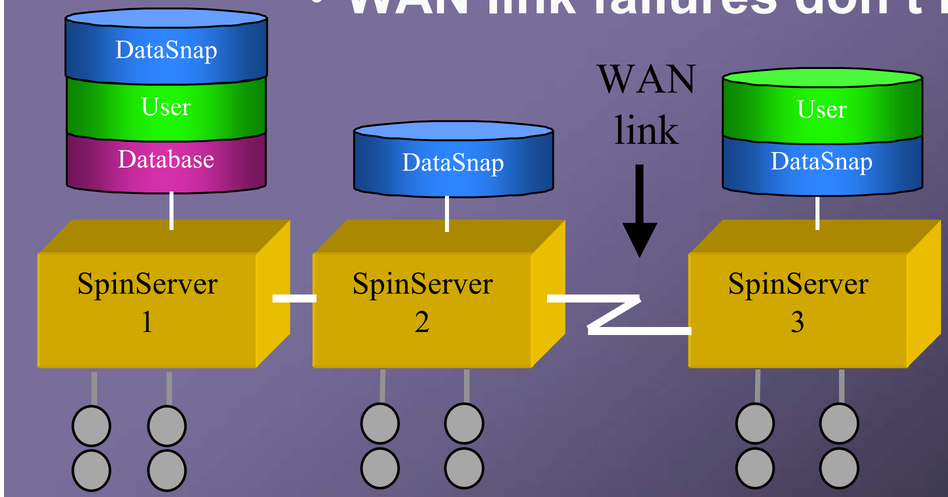
Database Operations

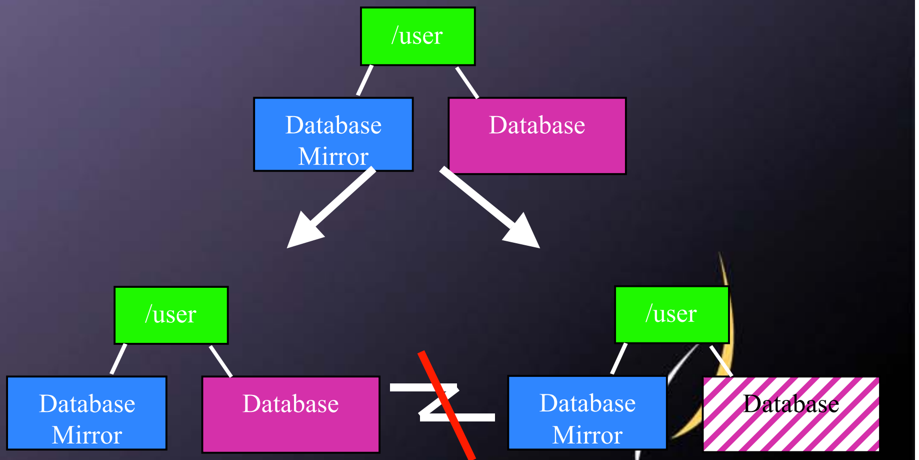
- Database operation with copies for disaster recovery
  - Database operations performed at server 1
    - mirrored every 5 minutes to snapshot for application failure recovery
    - multiple snapshots kept at remote site
      - only differences between snapshots are stored
- Database backup versions inexpensively stored
  - and storage online makes recovery easy



Database Operations

- High performance database for many readers
  - Updates made only to primary database
    - snapshots mirrored to remote sites every 5 minutes minute
  - Database readers access mirror
    - local mirrors allow scaling to many thousands of database users
    - WAN link failures don’t interfere with database readers

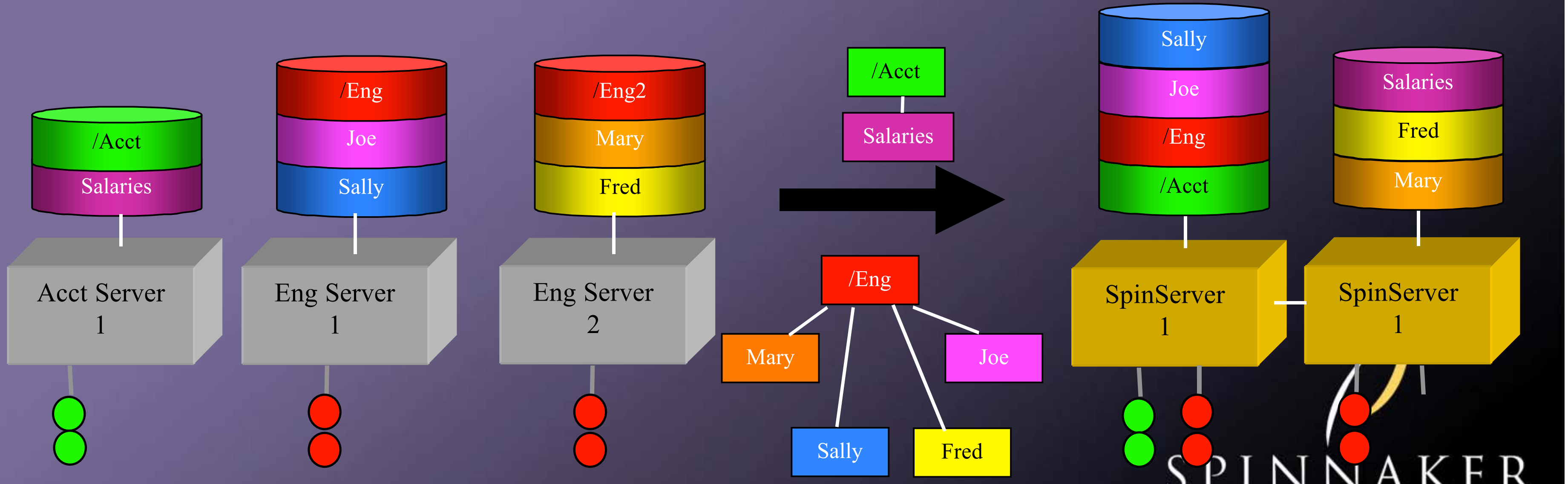




SPINNAKER
NETWORKS

NAS Consolidation

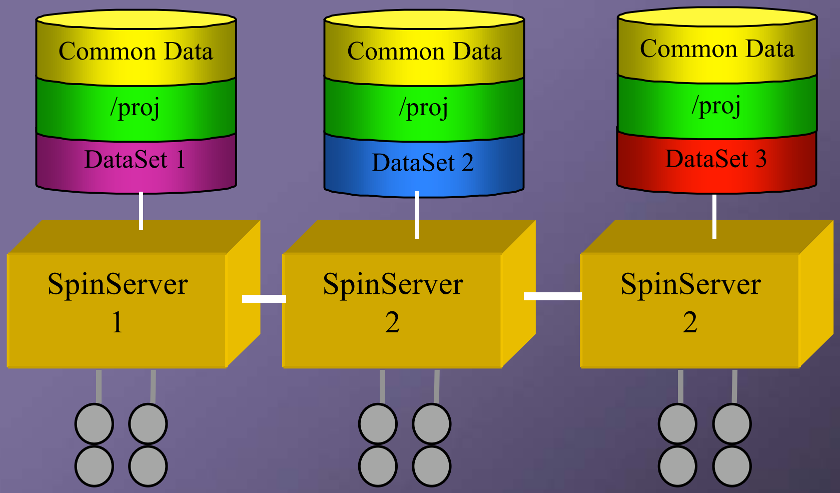
- Virtual servers
  - provides "firewall-like" security functionality
  - each has its own VFSes and IP subnet
    - with no sharing between virtual servers
- Multiple departments efficiently share same servers



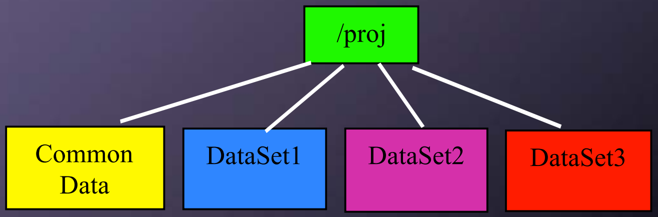
SPINNAKER NETWORKS

Parallel Data Processing

- High aggregate bandwidth to data
  - total bandwidth to compute servers scales linearly
  - rebalance data online as desired
  - without updating 10,000 compute servers' config



Compute servers

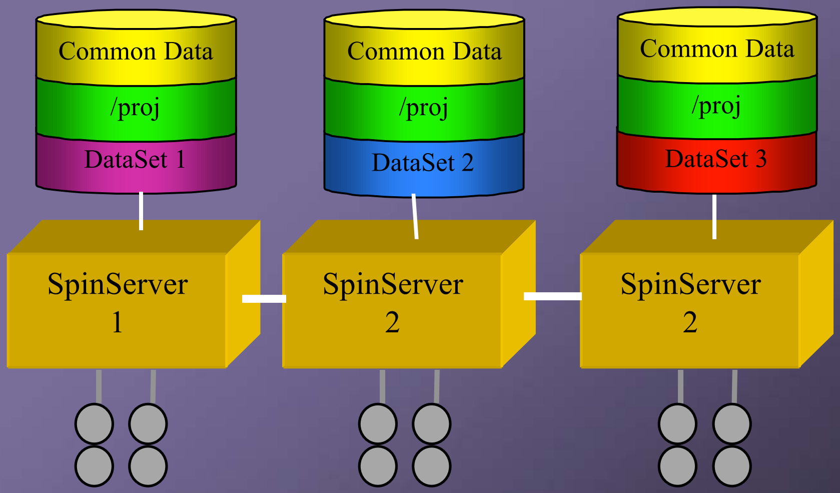


View from any compute server

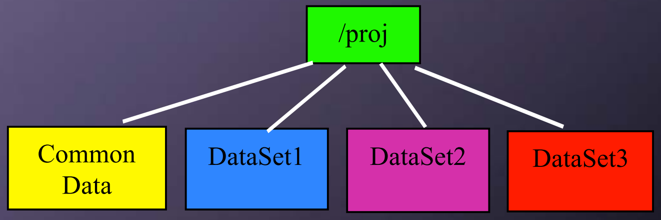
SPINNAKER
NETWORKS

Parallel Data Processing

- High bandwidth for reads to common data
  - mirror heavily read data to multiple servers
  - servers load balance among multiple copies



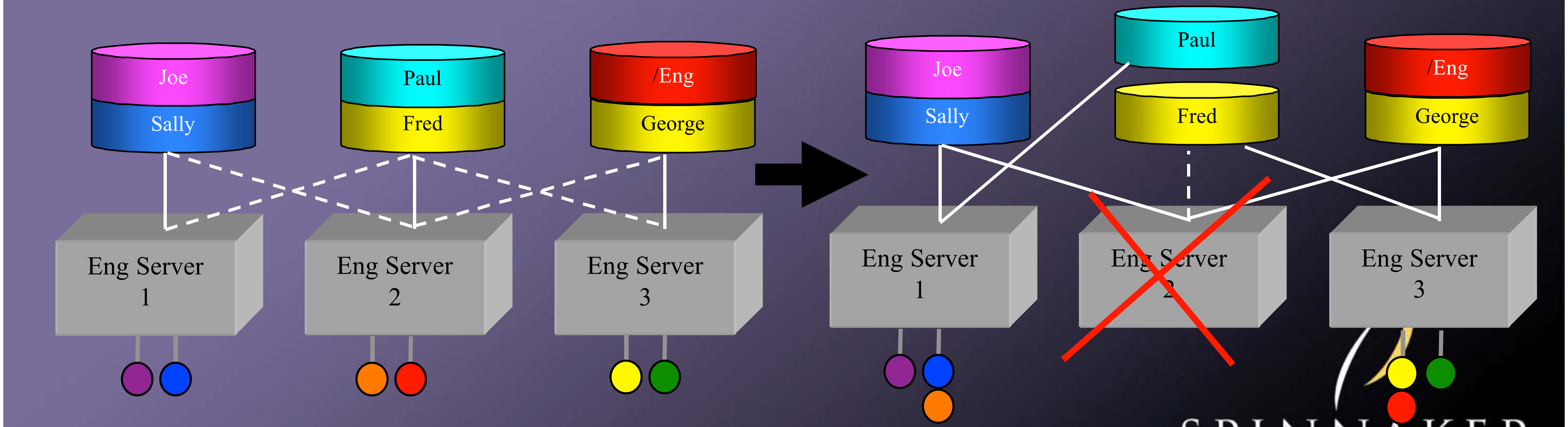
Compute servers



View from any compute server

Highly Available System

- Inexpensive redundant systems
  - server’s storage pools fail-over independently
  - example: each server gets 50% more load during failure
  - fail-over penalty decreasese with more storage pools

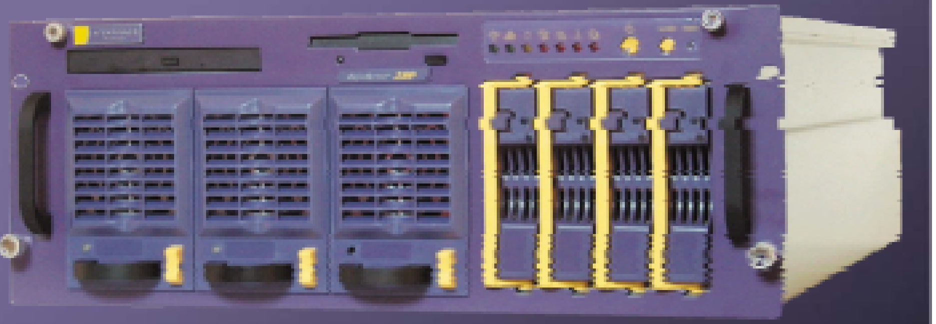


SPINNAKER NETWORKS

Spinnaker's System

SPINNAKER
NETWORKS

SpinServer 3300



File Protocols

NFS for Unix and CIFS for Windows file systems
ftp (rel 2.0)

SpinServer 3000 Series Capacities

2 x 1GigE data ports and 2 x 1GigE cluster ports
4 FCAL loops ñ 22TB total capacity
4 GB cache

SpinStor Disk Array

15 disk drives per array, 3U height
Dual RAID controllers
36GB, 73GB, 146GB drive capacities

SpinCluster

Up to 512 systems per cluster with standard GigE switch interconnect
Virtual Servers spanning multiple physical SpinServers
Dynamic, on-line data movement (rel 2.0)
Up to 11 PB of disk storage
Near linear performance scaling

SpinServer 3000 High Availability and Business Continuance

Redundant power supply, fans
Redundant network, cluster and FC connections
Mirrored SpinFS drives in SpinServer
High Availability, 1+1 failover, failback (rel 2.0)
Asynchronous data mirroring, SpinMirror (rel 2.0)

Management

NDMP backup support
SNMP integration with industry storage management
CIM compatible
SpinServer CLI and GUI

SPINNAKER
NETWORKS

Performance – Single stream

- Single stream read/write, 9K MTU
  - 94 MB/sec read
  - 99 MB/sec write
  - with file sizes larger than cache
    - scheduling disk IO operations

Performance – Spec SFS97_R1

- One server
  - 31876 ops/sec NFSv3 over UDP
  - 23363 ops/sec NFSv3 over TCP
- 6 servers
  - 131930 ops/sec NFSv3 over UDP
  - 117538 ops/sec NFSv3 over TCP
- Spec SFS V3.0 (http://www.spec.org)

Summary

- NextGen NAS Systems not all the same
  - how large do you expect to scale
  - how expensive is downtime to accommodate
    - how important is online management
  - how important is manageable storage Class of Service
    - does your data's access pattern change over time
  - consider mix of ops from your application
    - mostly large reads and writes, or more typical NFS traffic
- Various architectures handle the above differently

Spinnaker Networks, Inc.

Thank you

SPINNAKER
NETWORKS