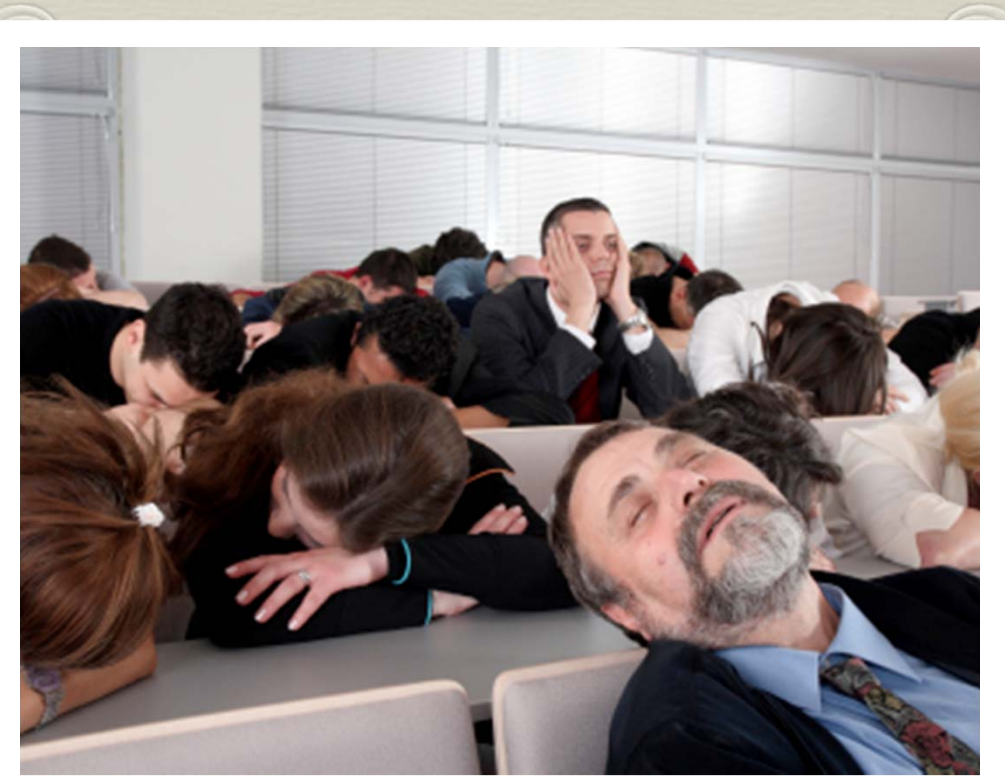
Highly-Available Distributed Storage

UF HPC Center
Research Computing
University of Florida



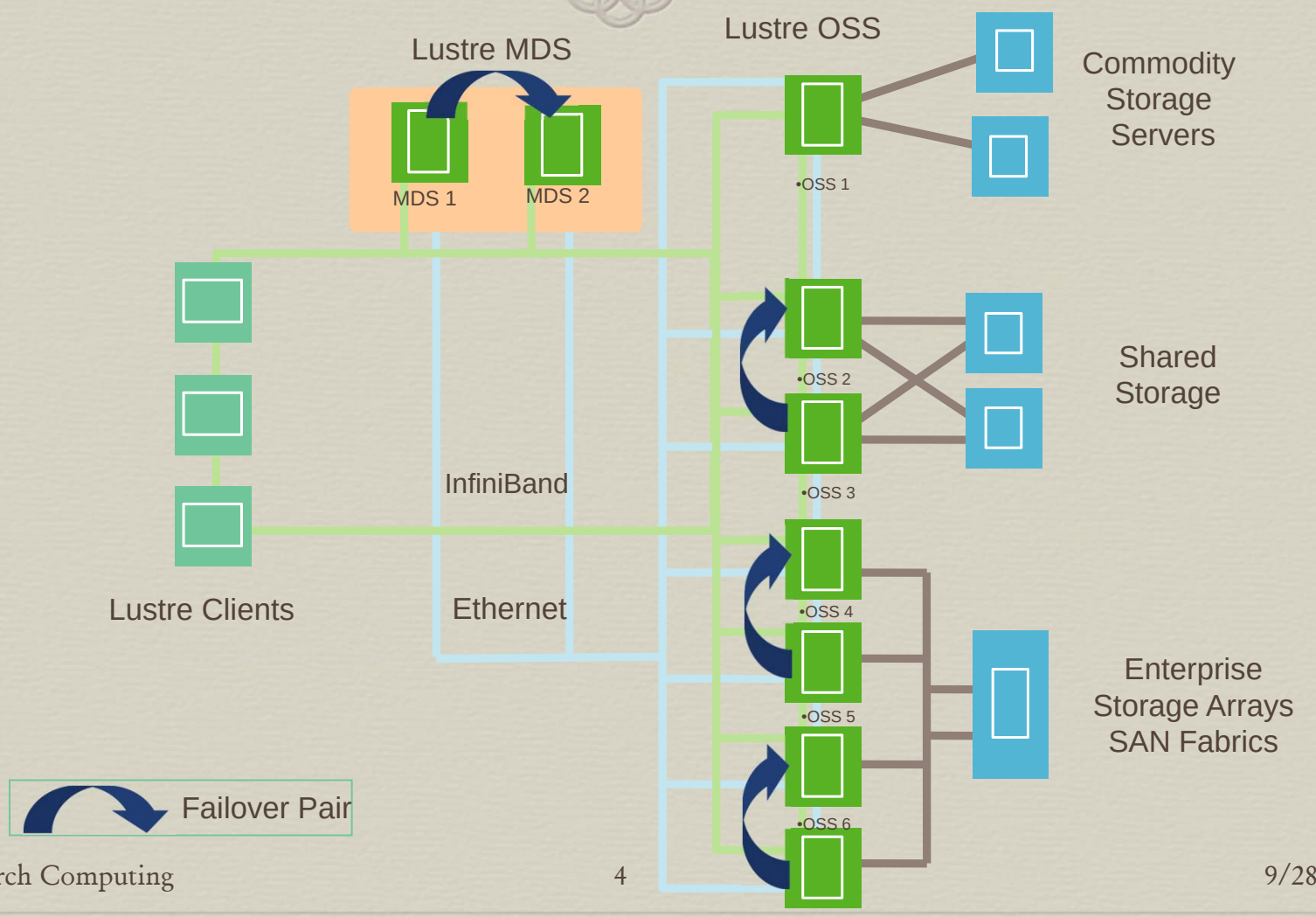
Storage is Boring

Slow, troublesome, albatross around the neck of high-performance computing

HA Storage

- **Previous Implementation**
  - Scalable (Lustre-based)
  - High-Performance
  - Affordable (very cost effective)
  - Relatively Reliable
- **Still want to improve**
- **Provide a better experience for users**

HA Storage


Lustre MDS
Lustre OSS
MDS 1
MDS 2
•OSS 1
•OSS 2
•OSS 3
•OSS 4
•OSS 5
•OSS 6
Commodity Storage Servers
Shared Storage
Enterprise Storage Arrays SAN Fabrics
InfiniBand
Ethernet
Lustre Clients
Failover Pair

HA Storage

- **Traditional Approach**
  - Multiple Servers
  - External Storage Chassis
  - Dual-Active RAID Controllers
  - Dual-Attached Fibre Channel or SAS
  - Dual-Homed Drives
  - HA Cluster Software
- **Costs Quickly Mount ($2800 - $3500 TB)**

HA Storage

- **Our Approach**
  - Integrated Server + Storage Chassis
    - Lower Cost
    - Higher Density
  - **Internal PCIe RAID Controllers**
    - Lower Cost
    - As good or better performance
  - **Commodity SATA Disk Drives (Enterprise)**
  - **HA Cluster Software**

HA Storage

- **Problem**
  - All storage is internal to each chassis
  - No way for one server to take over the storage of the other server in the event of a server failure
  - Without dual-ported storage and RAID controllers how can one server take over the other's storage?
- **Solution**
  - InfiniBand
  - SCSI Remote/RDMA Protocol (SRP)

HA Storage

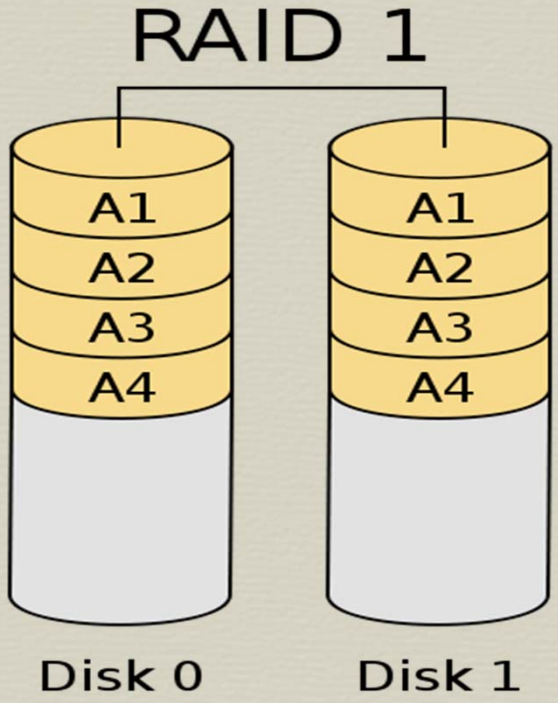
- **InfiniBand**
  - Low-latency, high-bandwidth interconnect
  - Used natively for distributed memory applications (MPI)
  - Encapsulation layer for other protocols (IP, SCSI, FC, etc.)
- **SCSI Remote Protocol (SRP)**
  - Think of it as SCSI over IB
  - Provides a host with block-level access to storage devices in another host.
  - Via SRP host A can see host B's drives and vice-versa

HA Storage

- Host A can see host B's storage and host B can see host A's storage but there's a catch…
- If host A fails completely, host B still won't be able to access host A's storage since host A will be down and all the storage is internal.
- So SRP/IB doesn't solve the whole problem.
- But… what if host B had a local copy of Host A's storage and vice-versa (pictures coming – stay tuned).
- Think of a RAID-1 mirror, where the mirrored volume is comprised of one local drive and one **remote** (via SRP) drive

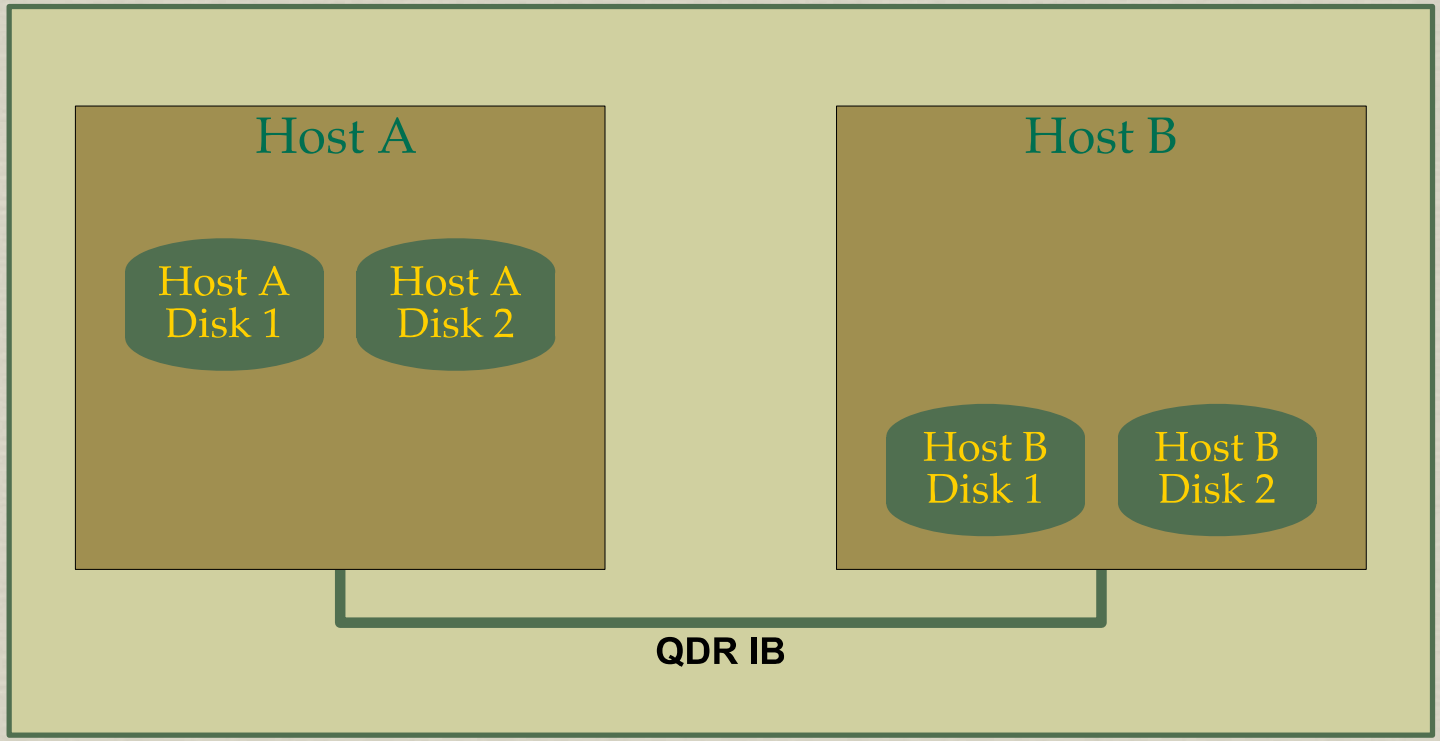
HA Storage

- **Mirrored (RAID-1) Volumes**
  - Two (or more) drives
  - Data is kept consistent across both/all drives
  - Writes are duplicated to each disk
  - Reads can take place from either/all disk(s)



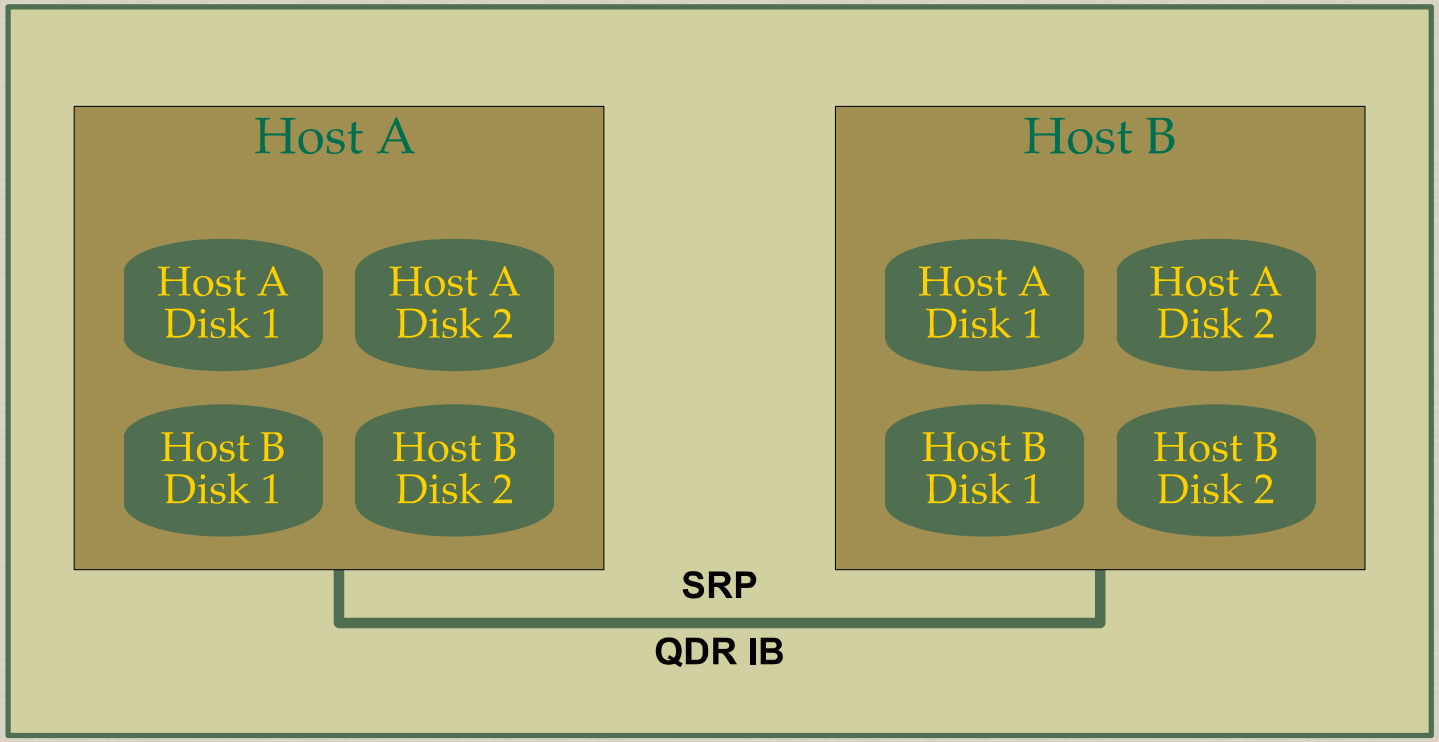
Remote Mirrors

- Not Possible?



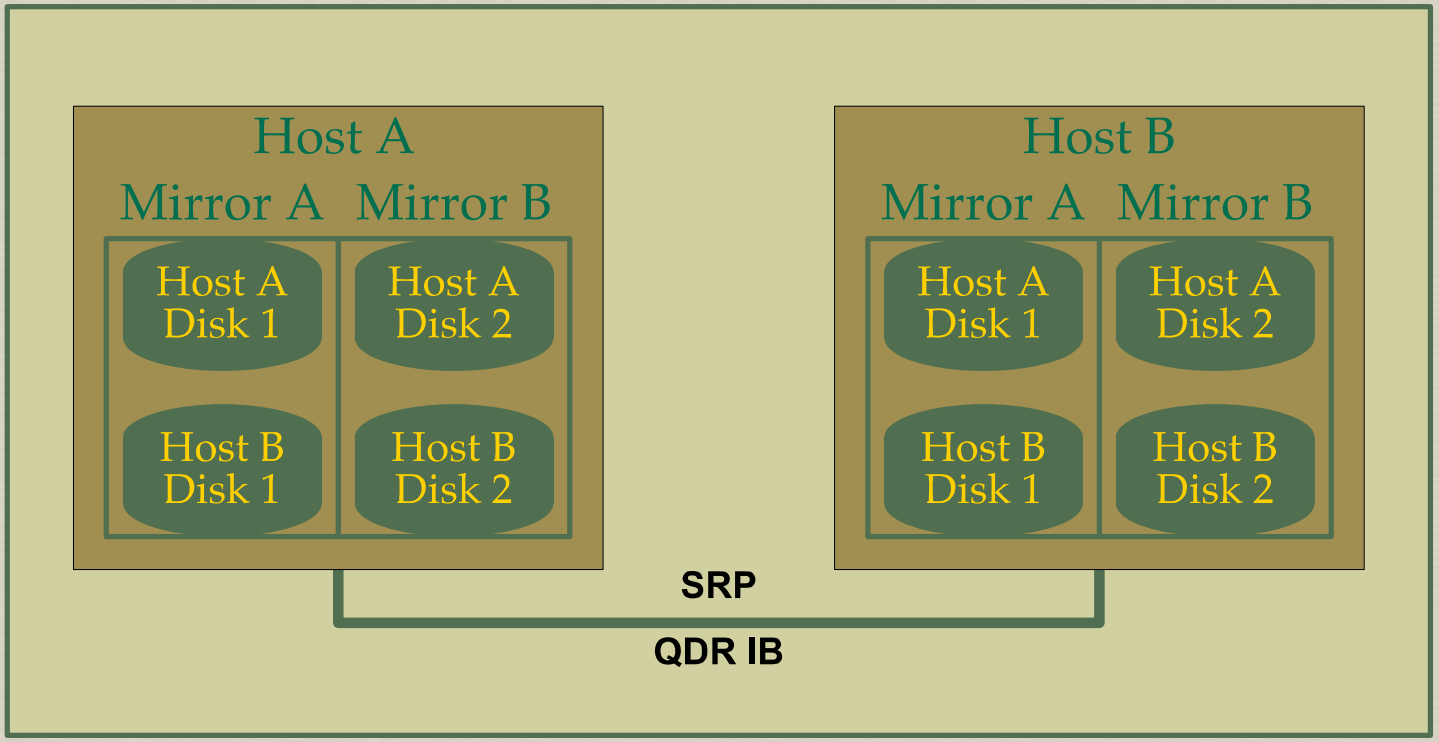
Remote Mirrors

- Remote targets exposed via SRP



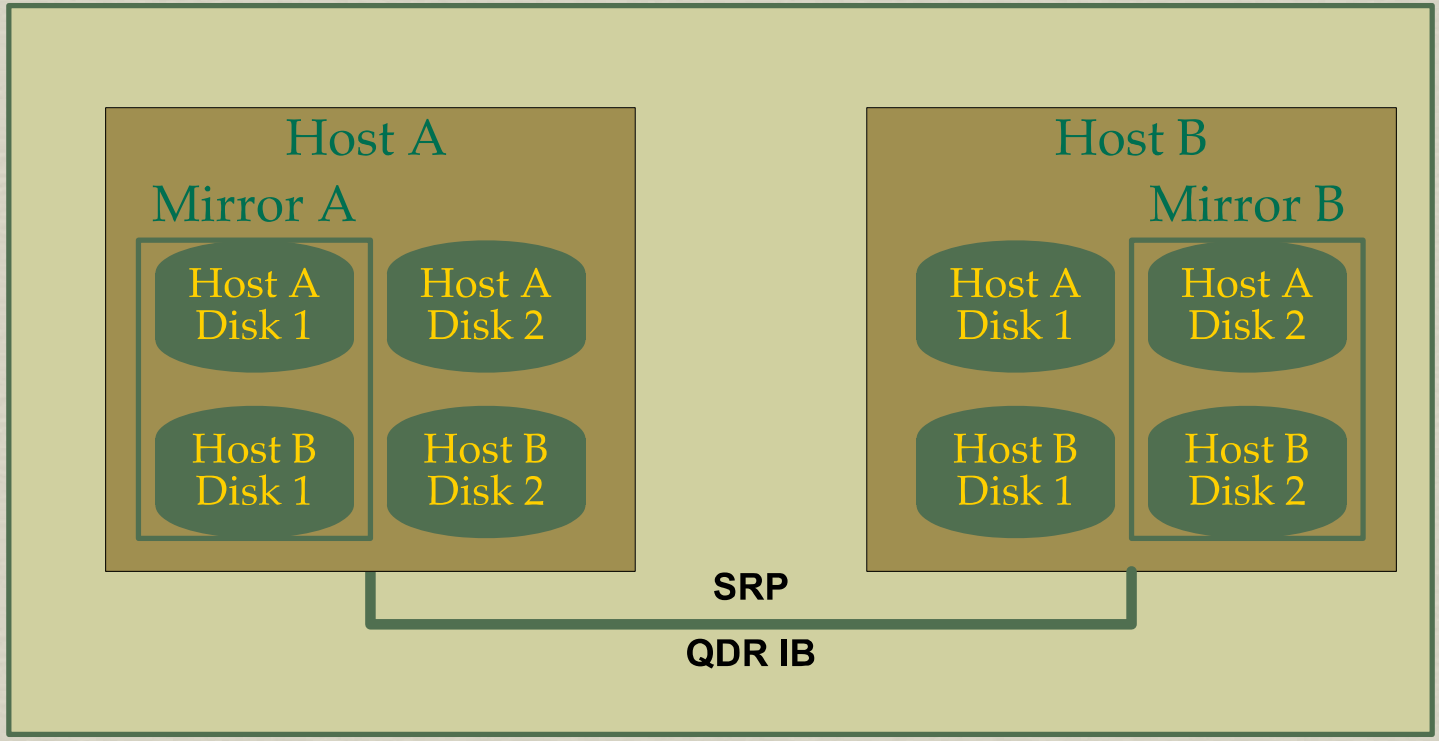
Remote Mirrors

- Mirroring Possibilities



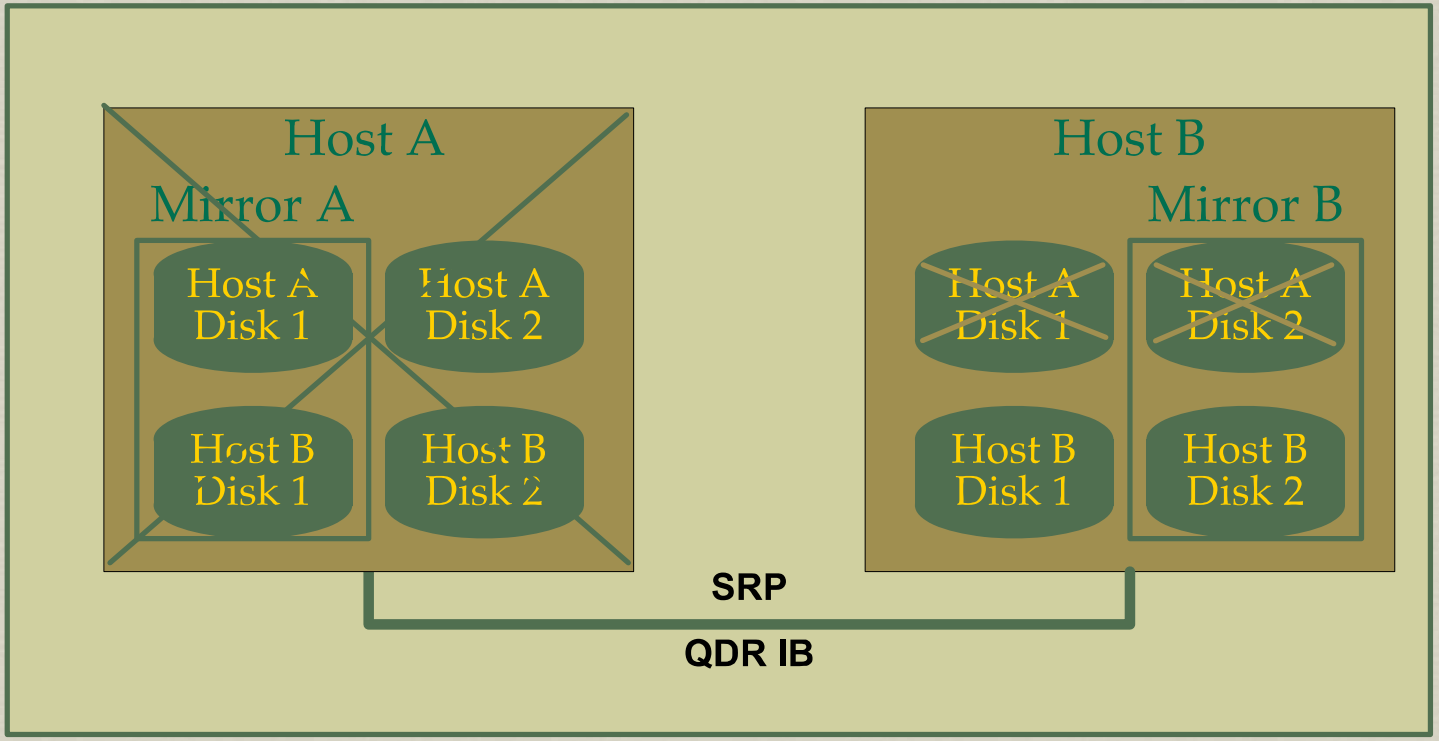
Remote Mirrors

- Normal Operating Conditions



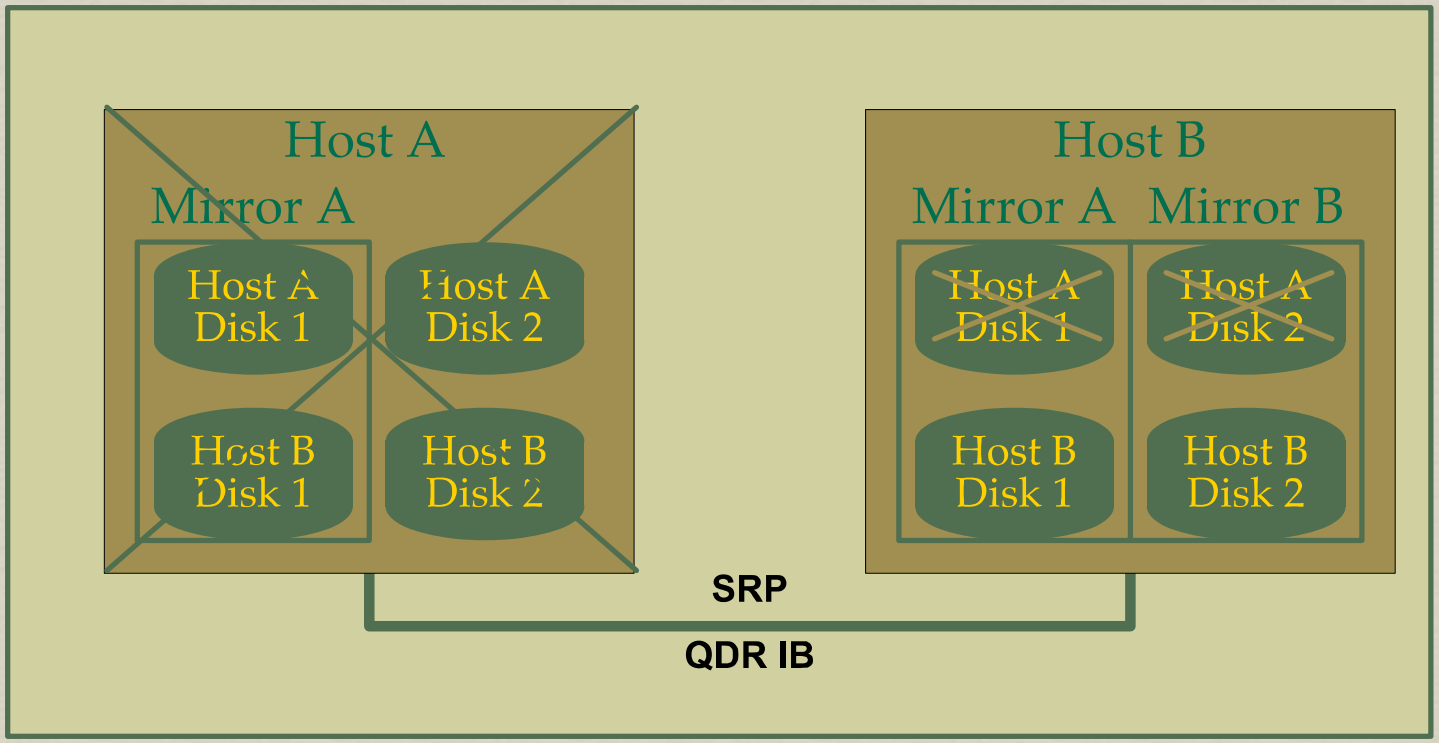
Remote Mirrors

- Host A is down



Remote Mirrors

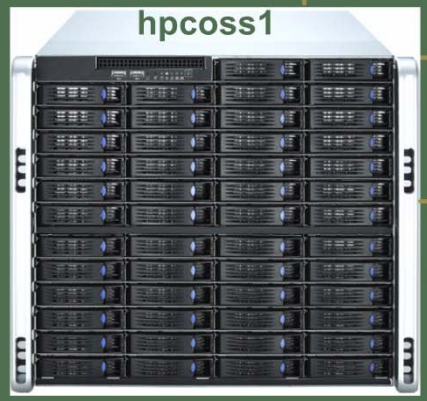
- Degraded mirrors on host B



SRP Mirrored Lustre HA OSS Pair

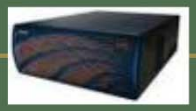
h1c1v0 (local)
h1c2v0 (local)
h1c3v0 (local)
h1c4v0 (local
h2c1v0 (remote)
h2c2v0 (remote)
h2c3v0 (remote)
h2c4v0 (remote)

h1c1v1 (remote)
h1c2v1 (remote)
h1c3v1 (remote)
h1c4v1 (remote)
h2c1v1 (local)
h2c2v1 (local)
h2c3v1 (local)
h2c4v1 (local)



SRP Mirror Traffic
QDR IB

SDR IB



IPoIB

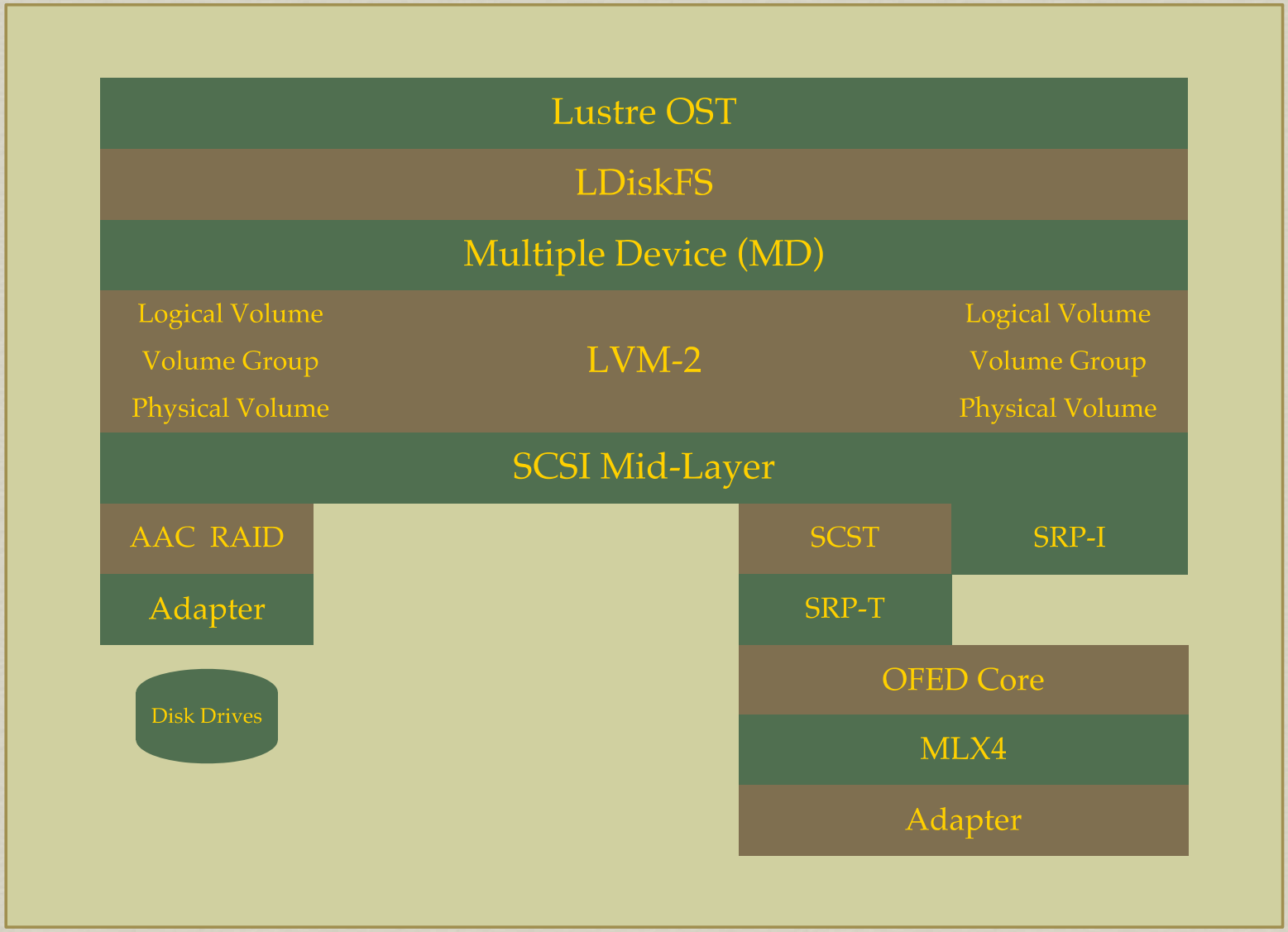
SDR IB

Each OSS:

Chenbro RM91250
SuperMicro X8DAH
2 x Intel E5620
24 GB RAM
2 x QDR IB
1 x SDR IB
4 x Adaptec 51245
RAID-6 (4+2)
8 x 8TB LUNs

OST0000 => MD100 => h1c1v0 + h2c1v0
OST0001 => MD101 => h1c2v0 + h2c2v0
OST0002 => MD102 => h1c3v0 + h2c3v0
OST0003 => MD103 => h1c4v0 + h2c4v0

OST0004 => MD104 => h1c1v1 + h2c1v1
OST0005 => MD105 => h1c2v1 + h2c2v1
OST0006 => MD106 => h1c3v1 + h2c3v1
OST0007 => MD107 => h1c4v1 + h2c4v1


Lustre OST
LDiskFS
Multiple Device (MD)
Logical Volume
Volume Group
Physical Volume
LVM-2
Logical Volume
Volume Group
Physical Volume
SCSI Mid-Layer
AAC RAID
Adapter
Disk Drives
SCST
SRP-I
SRP-T
OFED Core
MLX4
Adapter

HA Software

- **High-Availability Software (Open Source)**
  - Corosync
  - Pacemaker
- **Corosync**
  - Membership
  - Messaging
- **Pacemaker**
  - Resource monitoring and management framework
  - Extensible via Resource agent templates
  - Policy Engine

HA Software

- **Pacemaker Resources**
  - Highly-available services
  - IP Addresses, disk volumes, http servers, DNS, File systems, etc.
- **Pacemaker Resource Agents**
  - Typically BASH shell scripts
  - Conform to certain conventions (API)
  - Know how to stop, start, monitor, and validate a particular resource within the **Pacemaker** framework

HA Resources

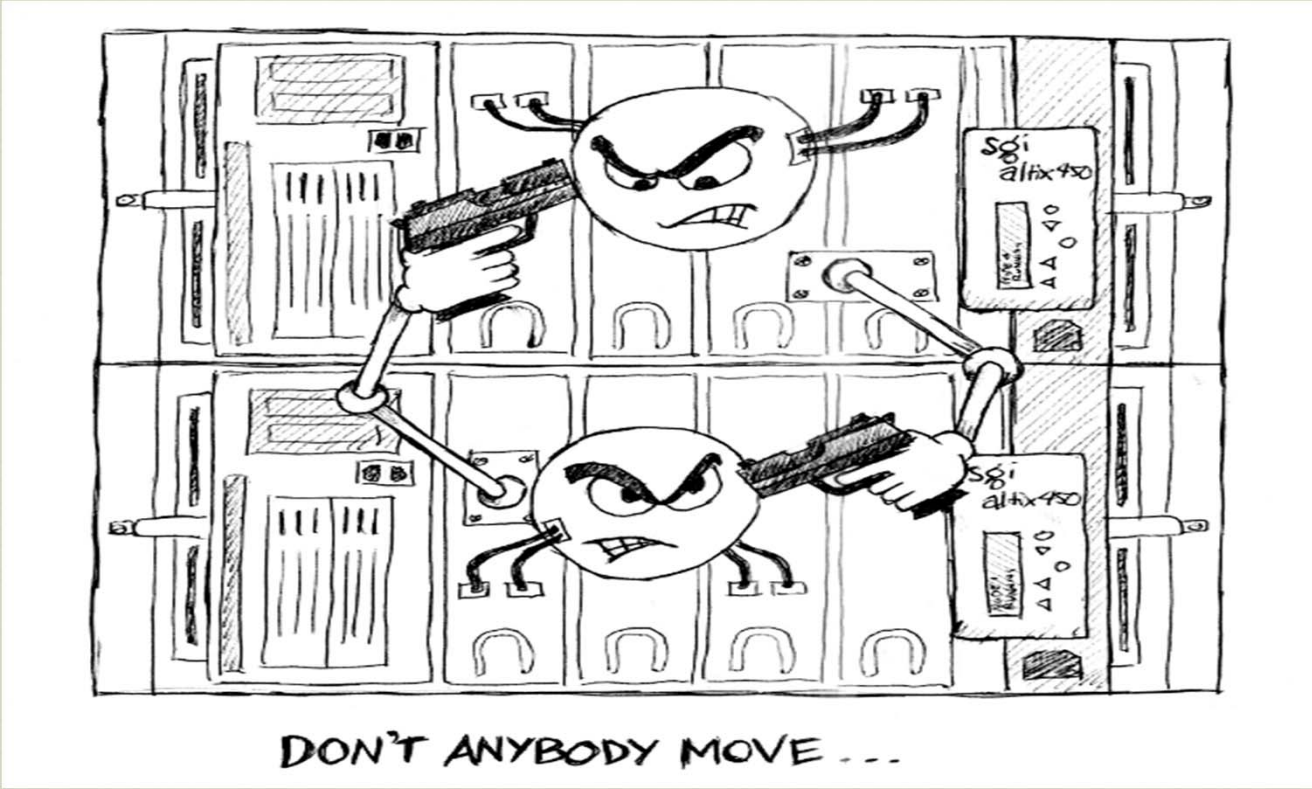
- **What are our HA resources?**
  - Mirrored disk volumes
  - Lustre File System instances
- **Disk Volume + File System = Lustre OST/MDT**
  - OST = *Object Storage Target* (many)
  - *MDT = MetaData Target* (one)
- **Host A & Host B: Failover Pair**
  - Mirrored volumes can be assembled on either host
  - File system mounted where ever mirror is assembled
  - Hence, a specific OST or MDT can reside on either server.

HA Resources

- **In Practice**
  - MDT
    - Active-Standby Servers
    - Only one MDT (currently supported)
  - OSTs
    - Active-Active
    - 4 OSTs per Server (normally)
    - 8 OSTs per server (degraded)

HA Storage

- **Split Brain Syndrome**



HA Storage

- **Not just Highly Available**
- **Also high-performance**
  - **4 PCI-E RAID Controllers per Server**
    - 2 RAID-6 (4+2) Logical Disk per Controller
    - 8 Logical Disks per Server (4 local, 4 remote)
    - 490 MB/sec per Logical Disk
    - 650 MB/sec per Controller (parity limited)
  - **Three IB Interfaces per Server**
    - IB Clients (QDR, Dedicated)
    - IPoIB Clients (SDR, Dedicated)
    - SRP Mirror Traffic (QDR, Dedicated)

HA Storage

- **High-Performance (continued)**
  - **Per Server Throughput**
    - 1.1 GB/sec per server (writes – as seen by clients)
    - 1.7 GB/sec per server (reads – as seen by clients)
  - **Actual server throughput is 2x for writing (mirrors!)**
  - **That's 2.2 GB/s per Server**
  - **85% of the 2.6 GB/s for the raw storage**

HA Storage

- **Keeping your data safe**
  - Mirrors enable failover
  - Provide a second copy of the data
  - Each Mirror
    - Hardware RAID
    - RAID-6 (4+2), two copies of parity data
  - Servers protected by UPS
    - Orderly shutdown of servers in the event of a sudden power outage.
    - 3+1 Redundant power supplies each to a different UPS.

HA Storage

- Thank you!
- Questions?