# Building a Highly Available TruCluster Server Configuration

**Greg Yates**
Consultant, Tru64 UNIX Support
Hewlett Packard

HP WORLD 2003
Solutions and Technology Conference & Expo

# Topics

- **Background Concepts**
- Configuration Examples
- Cluster Management Operations
- Network Connections
- Storage
- Cluster Interconnect
- Member Systems and Quorum

# Background Concepts

- Topics
  - Reliability Terms and Concepts
  - Non-Component Related Reliability Factors
  - Definitions from TruCluster 5.X

# Uptime, Availability and "9"s

- The amount of time that the system is available for its intended use by the customer
- A typical well maintained industry server can achieve an uptime of 99.0%
- When advertised, are the following factors being included:  system maintenance down time, application down time?

# Uptime, Availability and "9"s (cont)



- Biggest drawbacks to showing a certain uptime or availability
  - Difficult to analytically determine based on a design
  - Dependent on many factors other than system itself, i.e., environment, staff, procedures etc.
- Also consider data integrity
  - Wouldn't you rather have data integrity than uptime (given that choice)?

# Points of Failure

- Analyze the component diagram of the system and determine effects of component failures
- “Single Point of Failure”, where the failure of a single component would cause the entire system/cluster to be unable to deliver all of its services
- Can miss a lot:
  - Likelihood (probability) that a given component will fail
  - Whether a failover and resulting offline time of the application will occur
  - Different levels of difficulty to repair after a failure
  - Cluster software as an inescapable single point of failure

# Mean Time Between Failure (MTBF)



- The average amount of time before a component will fail
- Computing System MTBFs requires a database of component failure times, usually based on historical evidence
  - Not always easy to come by
- Component MTBF is not a direct relationship to system uptime
  - Some failures are not single points of failure
  - Redundancy and failover capability means more components, more component means statistically more failures will occur – but (hopefully) higher system uptime

# Disaster Tolerance

- Special requirement
  - Resilience from major destructive events that could completely destroy or make in-operable all components of a system/cluster at once - a collection of machines/equipment in close proximity
    - Bomb blast, Tornado, Floods, etc.
- Typically requires maintaining multiple sites with replicated equipment at a desired geographical distance from each other with a target minimum down time to switch from the active to the remote site

# Serviceability

- System features to allow quick or on-line servicing, which minimizes down time
- Without specific configuration examples or statistical data, often used as an indicator or potential for high uptime (A.K.A. D.H.Brown-like studies)
  - Unfortunately only an indicator, not hard data about the reliability and uptime of a system

# Non-Component Related Reliability Factors

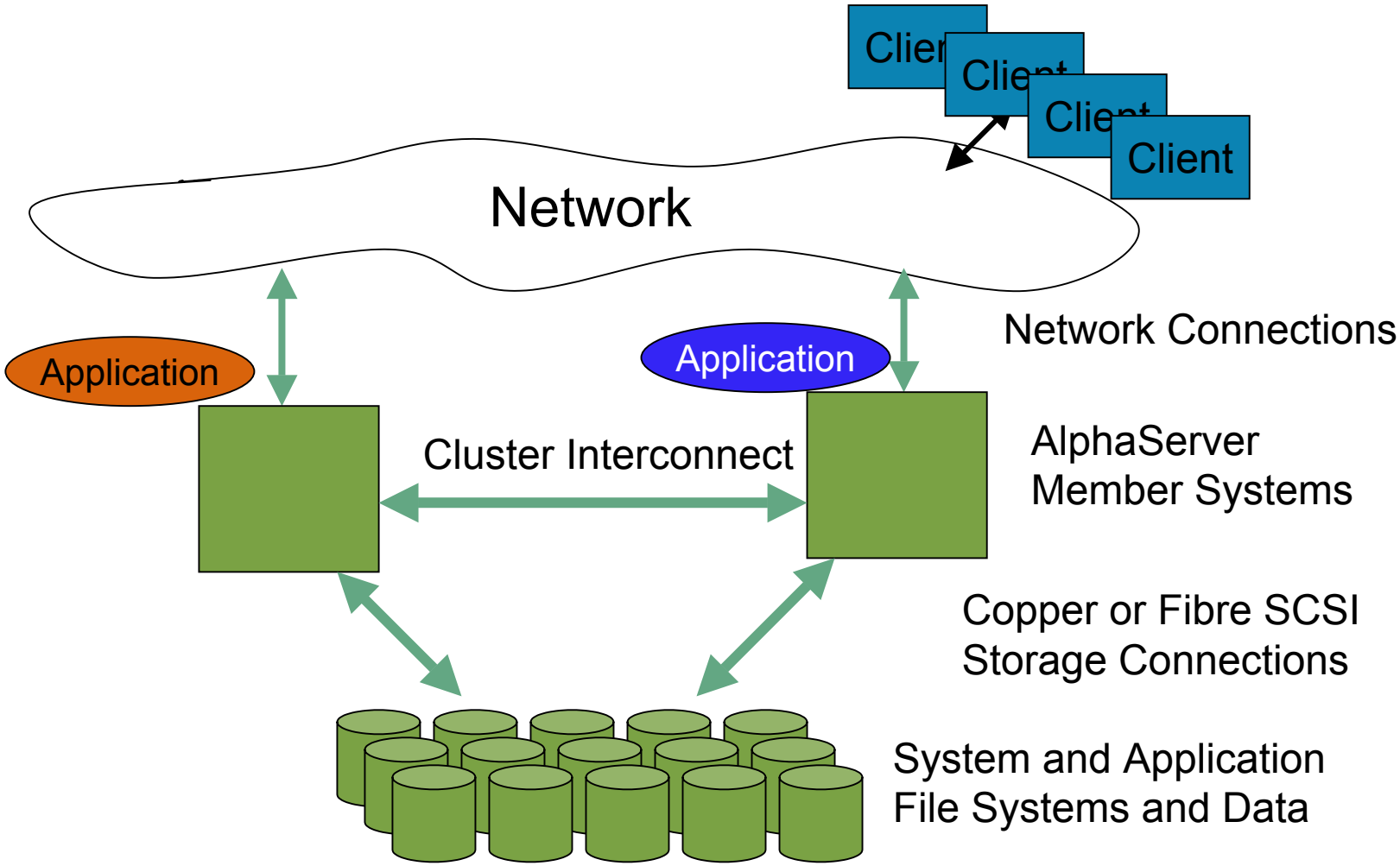


- **Many factors effect the reliability and availability of a computing system independent of the hardware and software being used:**
  - Design Methodology
  - Change Management
  - Monitoring
  - Staff Expertise (Services)
  - Avoid the Bleeding Edge
  - Acceptance Testing
  - Non-Production Sandbox
  - Maintenance
  - Disaster Recovery Preparations
- Same HW/SW configuration at two different sites can have very different uptime

# Topics

- Background Concepts
- **Configuration Examples**
- Cluster Management Operations
- Network Connections
- Storage
- Cluster Interconnect
- Member Systems and Quorum

# A TCS 5.X Cluster



# Definitions from TCS 5.X Clustering



- **Members**: AlphaServer Systems running the Tru64/TCS software - Tru64 TCS 5.1B supports 8 members in the cluster
- **Cluster Interconnect**: Dedicated hardware communication channel used between cluster members to coordinate activities and transfer and share data - TCS 5.1B supports Memory Channel, Fast and Gigabit Ethernet for the cluster interconnect
- **Storage**: Hardware or software RAID storage entities used to store system, cluster and application data - TCS 5.1B supports UltraSCSI III (Copper) or Fibre Channel, LSM and StorageWorks RAID Arrays

# Definitions from TCS 5.X Clustering (cont)



- **Network Connections**: Standard network communication channels that connect the cluster members to the outside world and allow client systems to access applications/services provided by the cluster
- **Application**: Some cluster services may be “built-in” or “canned” with the system software such as DHCP serving or NFS serving. Other services provided to the outside world are based on third-party applications such as Oracle’s Oracle Server
- **Clients**: Computer Systems and people outside the cluster that access it’s services via network connections

# Failover and the Virtual Server Model



- Much of the availability aspect of clusters is built on the technique of Failover – moving an application and data access from one cluster member to another
  - Single instance (need failover)
  - Multi instance (don't need failover)
  - Cluster aware (don't need failover)
  - CFS (automatic)
- Goodness – Failover means that individual cluster components can fail and the cluster will still be able to provide applications/services

# Failover and the Virtual Server Model (cont)



- Failover will mean the temporary suspension or pause in (single instance) application activity
  - Can vary from momentary suspension of I/O to complete application stop and restart
  - Redundancy that can prevent failover from happening at all, improves uptime
  - Multi-instance Applications are usually less effected – no restart
- Must be careful (or at least aware) of the implications of loading the combined systems (or redundant components) with more load than the surviving system (component) can handle at the same performance levels
  - Member Systems
  - Multi-Pathed HBA
  - NIC Link Aggregation (more later)

# Failover and the Virtual Server Model (cont)

- Some customers/sites are fine with degraded performance with a failure – but not all

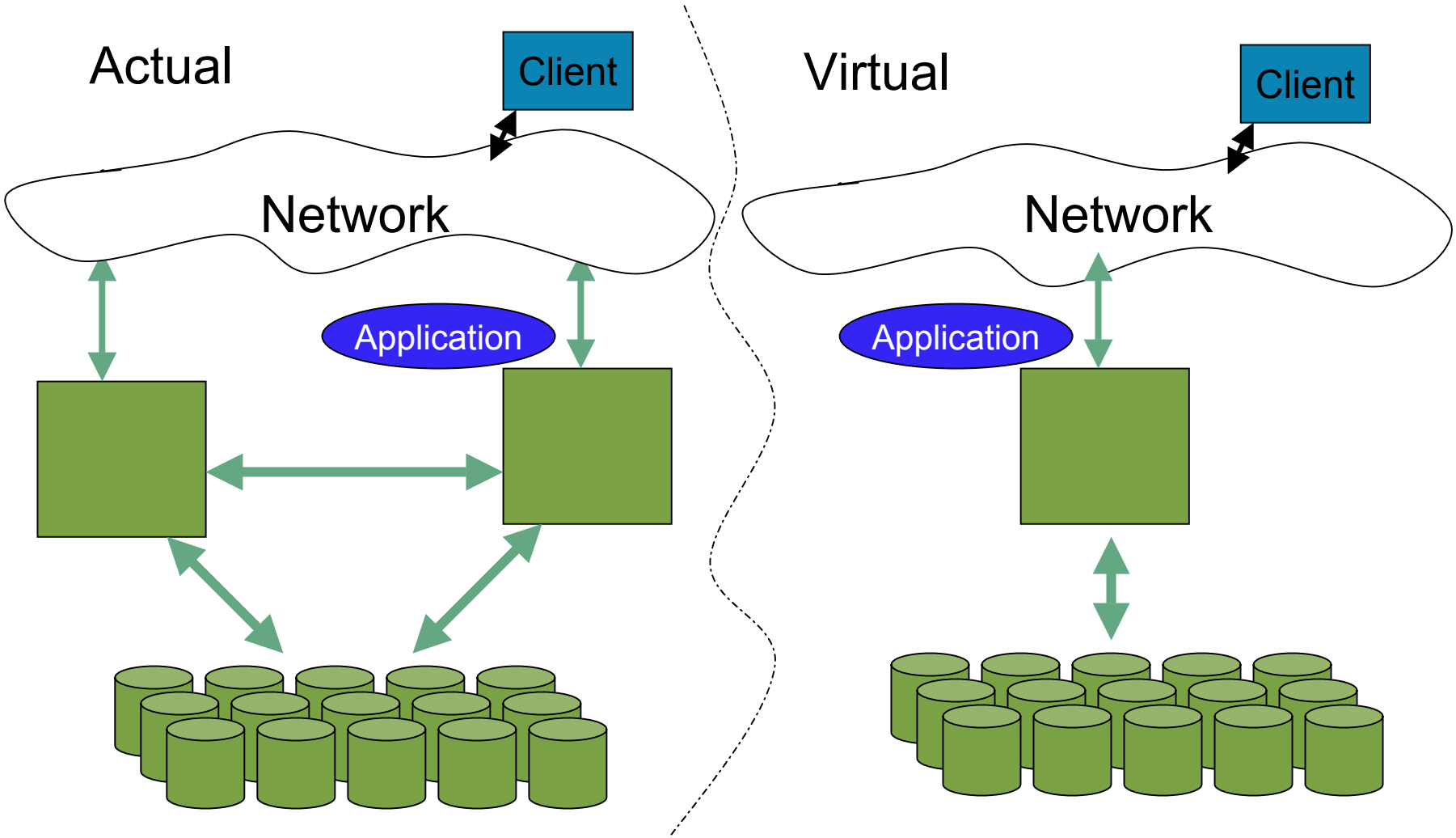
# Non-Single Points of Failure in TCS 5.1B

- Loss of any of the following may cause individual member systems to fail and cause a failover (application offline), but would not take the entire cluster offline:
  - Member boot disks and swap
  - The cluster interconnect to a member (assuming suitable quorum configuration)
  - Member system power supplies

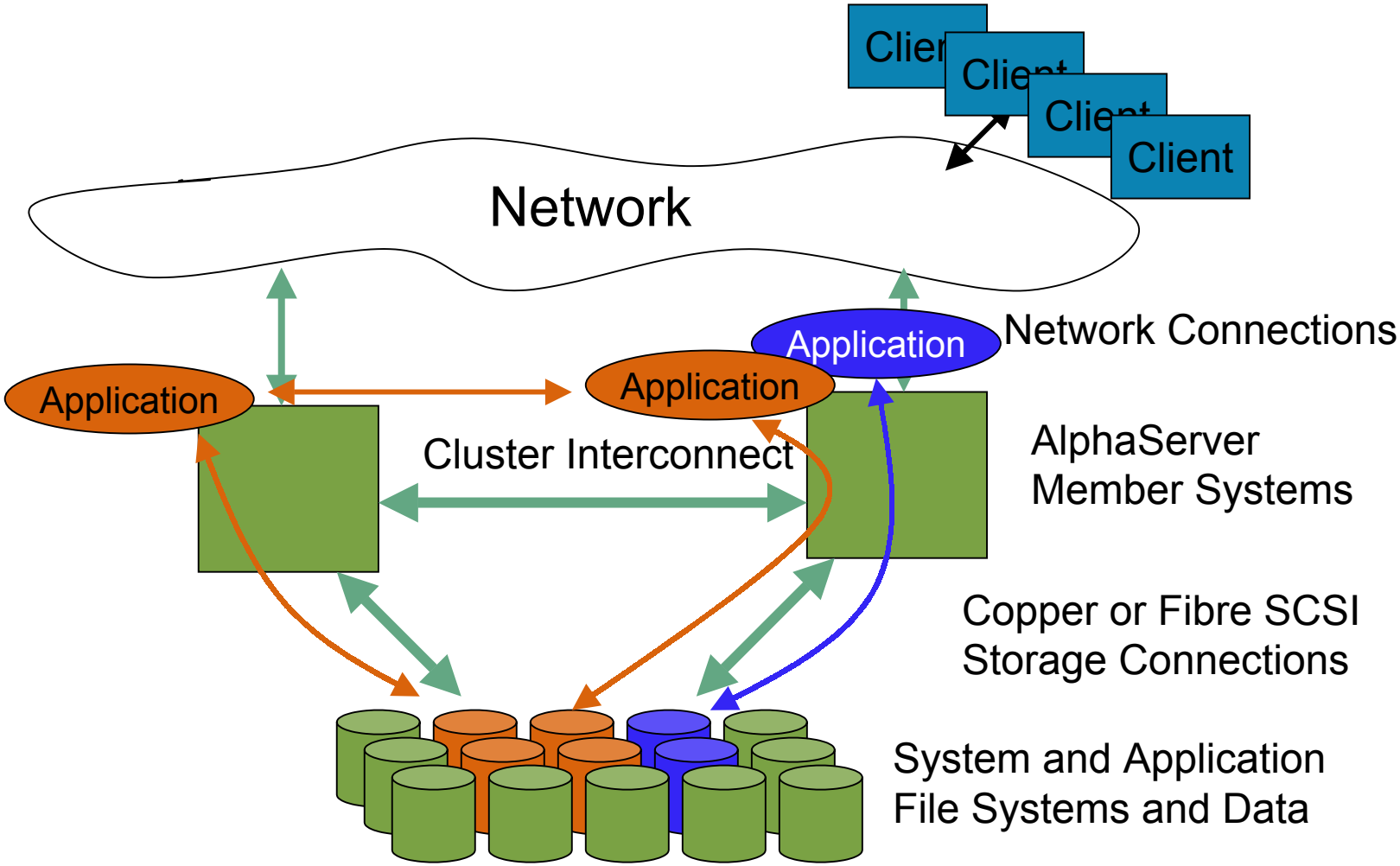
# Non-Single Points of Failure in TCS 5.1B (cont)

- Similarly, loss of any of the following would cause degraded performance to perhaps unacceptable levels, but would not cause a member to leave the cluster or a failover:
  - The single connection to storage (HBA) of a member (assuming the file systems are on disks on shared buses)
  - The single network connection (NIC) of a member
  - A NIC in a LAG (or NetRAIN) interface

# Failover and the Virtual Server Model



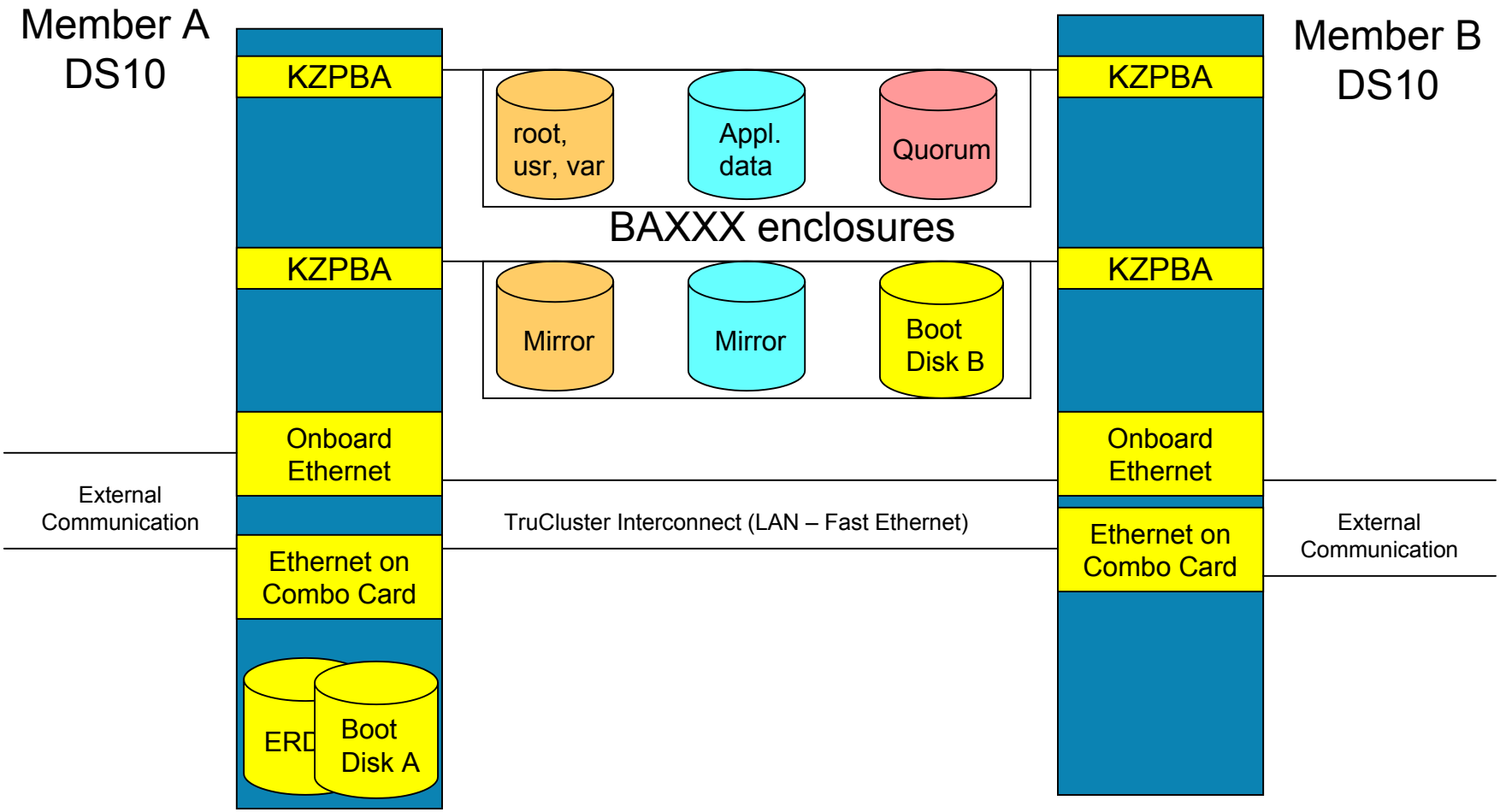
# Single and Multi-Instance Applications



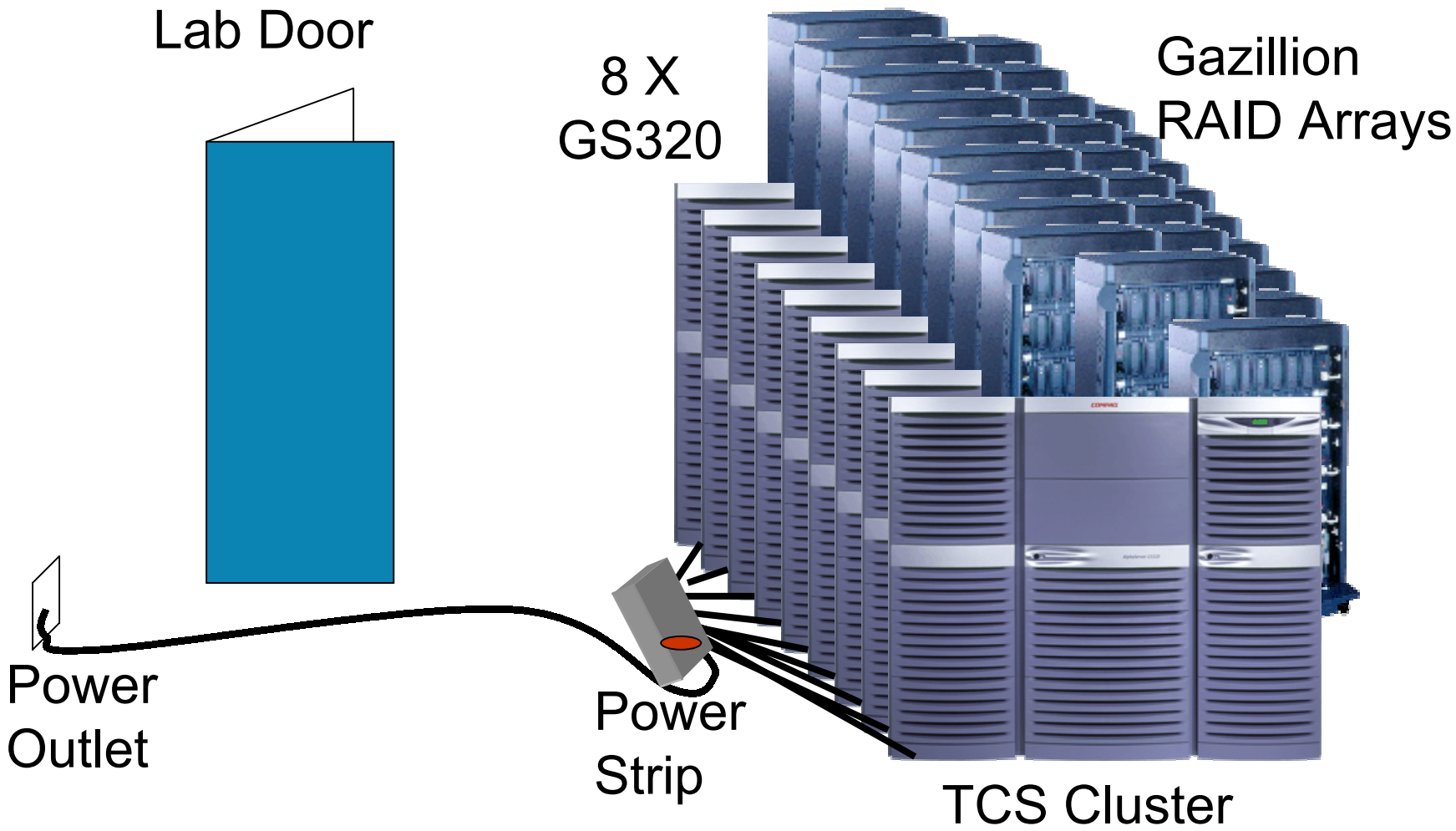
# Examples "from Scratch" - Resources

- Topics
  - Lowest Cost , No Single Point of Failure Cluster
  - Maximum Redundancy Cluster Configuration
  - Most Expensive Single Point of Failure Cluster

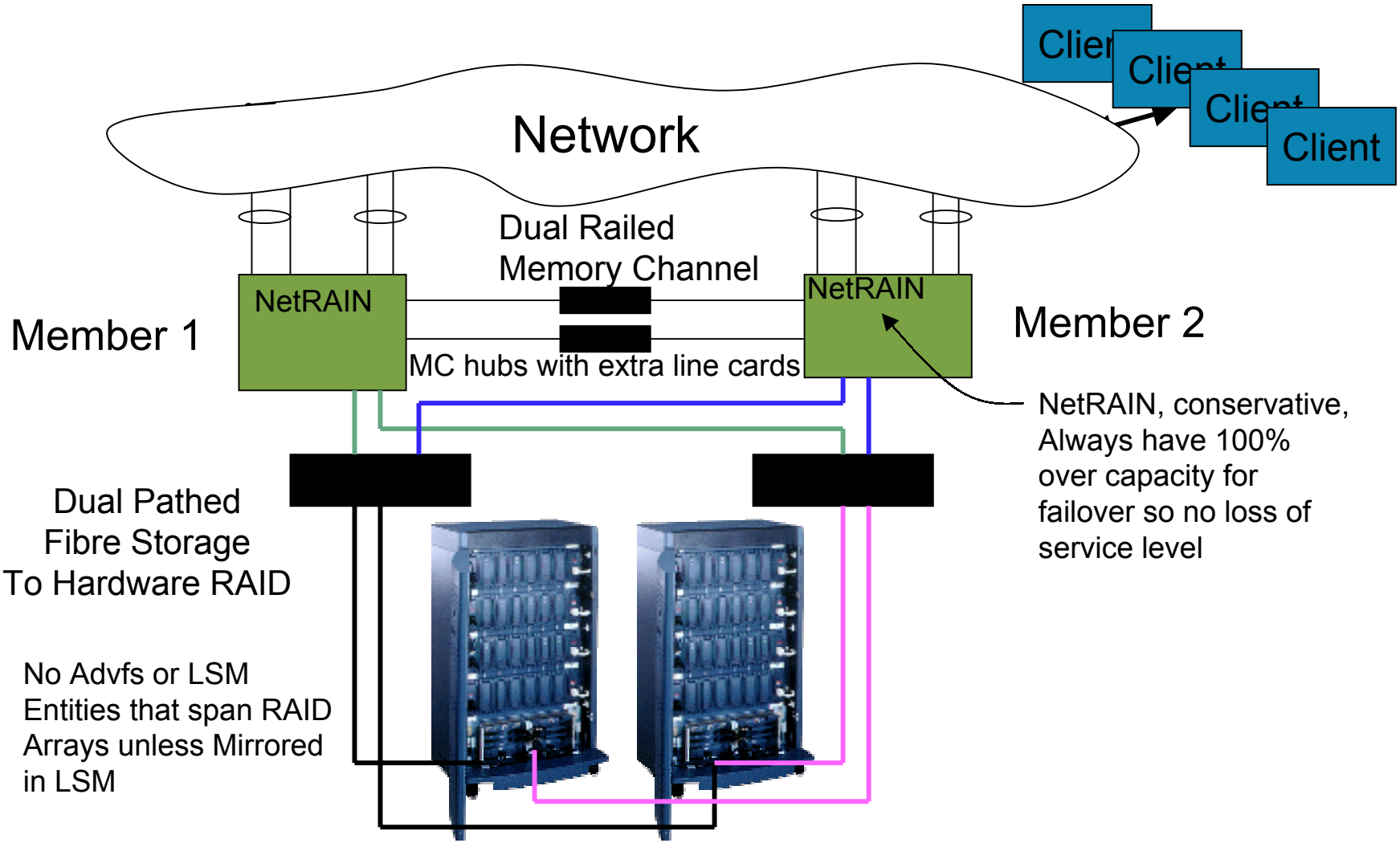
# Lowest Cost, No Single Point of Failure Cluster



# Highest Cost Single Point of Failure Cluster



# Typical Mid-size configuration



# Topics

- Background Concepts
- Configuration Examples
- **Cluster Management Operations**
- Network Connections
- Storage
- Cluster Interconnect
- Member Systems and Quorum

# Cluster Management Operations

- Topics
  - Recovering Cluster Disks and File Systems
  - Recovering the Whole Enchilada
  - Patches

# Recovering Individual Disks and File Systems

- Recovery – Are you ready for any of the following?
  - Failed cluster file system
    - Cluster Admin Guide Chapter 11
  - Failed member boot disk
    - Cluster Admin Guide Chapter 11
  - Failed quorum disk
    - Cluster Admin Guide Chapter 4

# Losing the Whole Cluster

- What do you boot when the cluster is blown away?
- First Issue - Booting
  - Installation Media – CD, Install Image on Disk or Network (RIS)
  - Original Tru64 OS installation (Emergency Repair Disk)
- Next Issue - Hardware Naming (device special files)
  - Booting system will create device naming based on hardware discovery
    - Potentially (likely) different than the HWMGR database that would be restored from the backup

# Deleting and Adding Members

- (Re) Installations
  - Deleting a member from a cluster that is non-operational and later adding back in with a minimum number of configuration steps

# Upgrades and Patch Kits

- Rolling Upgrades and Patch Kits
  - Good news
    - Entire cluster never goes out of service
  - Bad news
    - Every member must be re-booted at least twice
    - End-to-end procedure time can be longer than *hoped*

# Upgrades and Patch Kits (cont)

- Non-rolling upgrades now in V5.1 PK5 and later
  - Good news
    - Patch install is very quick (even with more than two members)
  - Bad news
    - Cluster is unavailable during patch installation
    - Cannot be done for version upgrades (installupdate)

# Topics

- Background Concepts
- Configuration Examples
- Cluster Management Operations
- **Network Connections**
- Storage
- Cluster Interconnect
- Member Systems and Quorum

# Public Networks

- Topics
  - Minimum Requirements and Advanced Options
  - NetRAIN
  - Link Aggregation

# Minimum Requirements

- Members' interfaces need not be symmetrical, but for availability need some level of replication
- Physical Network Interfaces
- Logical Network Interfaces
  - NetRAIN – Active/Passive redundancy from multiple NICs
  - LAG – Active/Active redundancy from multiple NICs
    - Common switch

# NetRAIN

- Redundant Array of Independent Network Adaptors
  - A logical network interface made from multiple physical network interfaces
  - One active physical link, additional paths are in standby
  - Requires a switch or redundant switches for NSPOF in a CI

# Link Aggregation (LAG) or Trunking

- A logical network interface made from multiple physical network interfaces
  - Active/active with load balancing
  - Both inbound and outbound
  - Two or more of the same type of network interface (Ethernet) dedicated to a single server **or switch**. The interfaces must all be of the same speed and operate in full duplex mode

    Common switch means you can't avoid a single point of failure
  - LAG is not supported for LAN Cluster Interconnect

# Effects of Failures

- Effect of failed network interfaces
  - EVM (Event Management) event occurs if monitored with Network Interface Failure Finder (NIFF)
    - If a member of a NetRAIN or LAG interface, transitions to remaining physical members of the set

# Topics

- Background Concepts
- Configuration Examples
- Cluster Management Operations
- Network Connections
- **Storage**
- Cluster Interconnect
- Member Systems and Quorum

# Storage

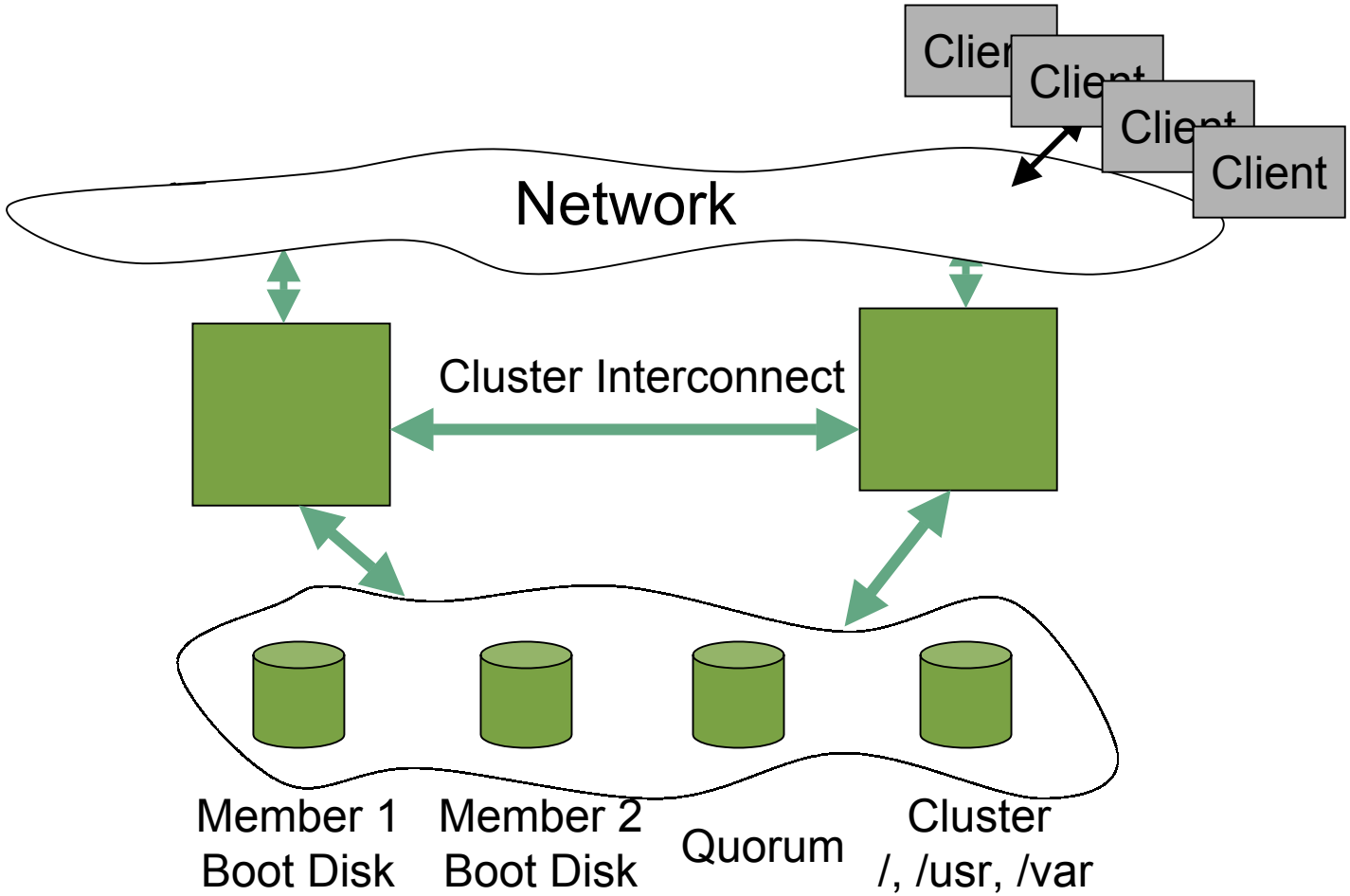
- Topics
  - Ground Rules: Required Logical Disks
  - Disks – Availability Requirements
  - Disks – Other Issues

# Ground Rules: Required Logical Disks



- Independent of Hardware or Software RAID, a TCS 5.X cluster requires the following logical disks:
  - **Cluster "system" Disk**: Disk(s) for cluster "/", "/usr" and "/var"
    - Can be multiple disks – this is just the minimum
    - Minimum Size ~2.5GB
  - **Member Boot Disk(s)**: Contains member's boot partition, swap space and small Connection Manager (CNX) partition. One per host
    - Any additional space on this disk should not be utilized for any other purpose.
    - Minimum Size ~513 MB
  - **Quorum Disk**: Highly recommended for two node clusters, suggested for 4, 6 and 8 node clusters. Not suggested for any other node count
    - Includes a partition for the Connection Manager (CNX).
    - Additional space of this disk cannot be utilized for any other purpose.
    - Minimum Size: 1 MB

# Two-Node Cluster Disk Requirements



# Disks – Availability Requirements



- Which Logical Disks are Single Points of Failure?
  - Cluster "system" Disk – yes
  - Member Boot Disk(s) – no, but will lose a member and potentially cause failover
  - Quorum Disk – no
- RAID Requirements
  - Cluster "system" Disk(s) – yes, Software (LSM) or Hardware RAID in 5.1B
  - Member Boot Disk(s) – highly desired, Hardware RAID only
  - Quorum Disk – desired, Hardware RAID Only
    - Typically, if H/W RAID is available, protect everything

# Disks – Availability Requirements (cont)



- Bus/Fabric Requirements
  - A Disk or RAIDSet is only as available as the bus/fibre that connects the host to it
    - Multi-path configuration is desired in the same way RAID is

      Not just multi-path, multi-fabric (no common switches or hubs)
  - Special note, two-node cluster, If single busing for low cost consideration,
    - Do not place a member boot disk and quorum disk on the same bus – the bus becomes a single point of failure

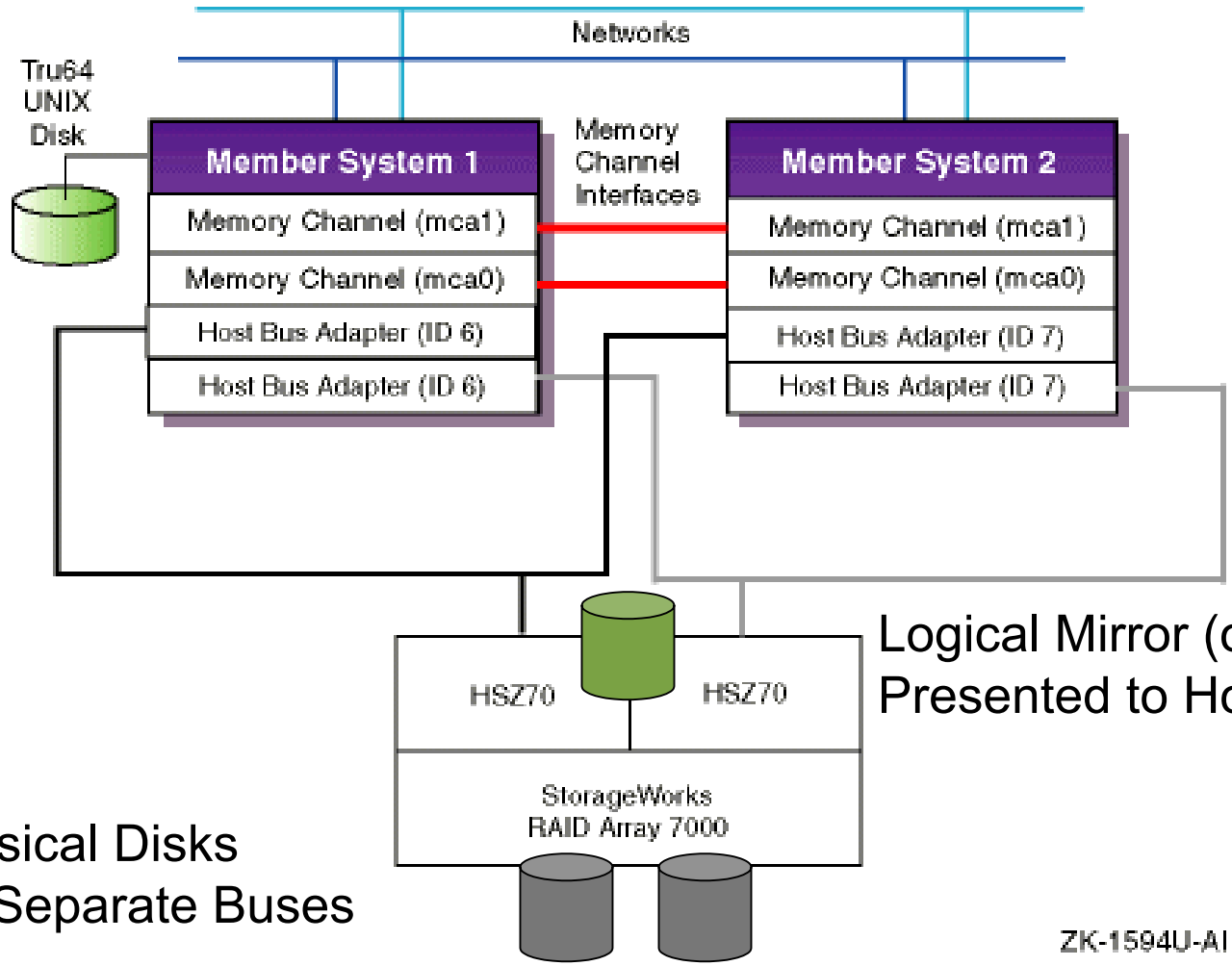
# Disks – Sharing Requirements



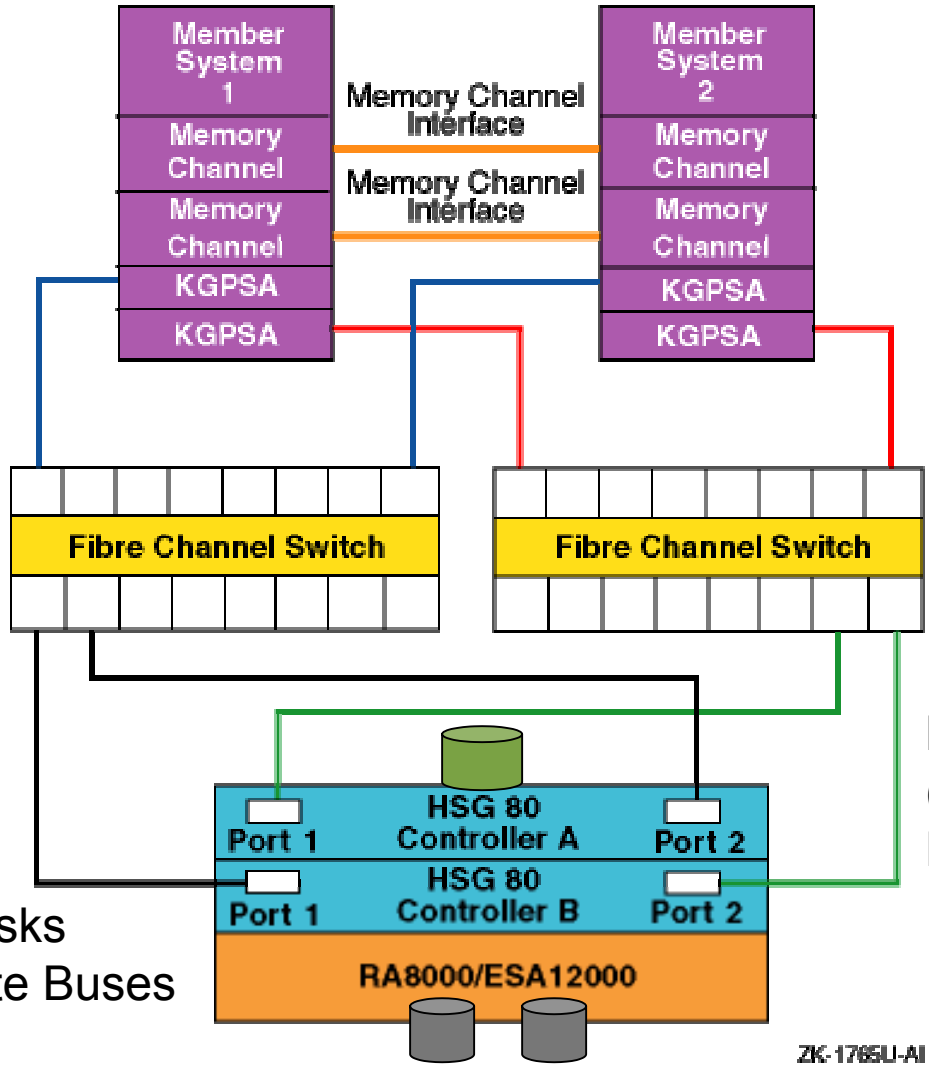
- Sharing Requirements
  - Not so much an issue in Fabrics – attach every one to the fabric
  - Member Boot Disk(s) – conceptually the member that uses it
    - In practice clu_add_member, clu_bdmgr force disks to be on the shared bus

      No clu_bdmgr – migrate command to move a member's boot disk from the shared bus to a private drive
    - Also benefits of crash dump recovery and the availability of the member
  - Quorum – must be reachable from all members

# Dual Pathing to an HSZ Controller



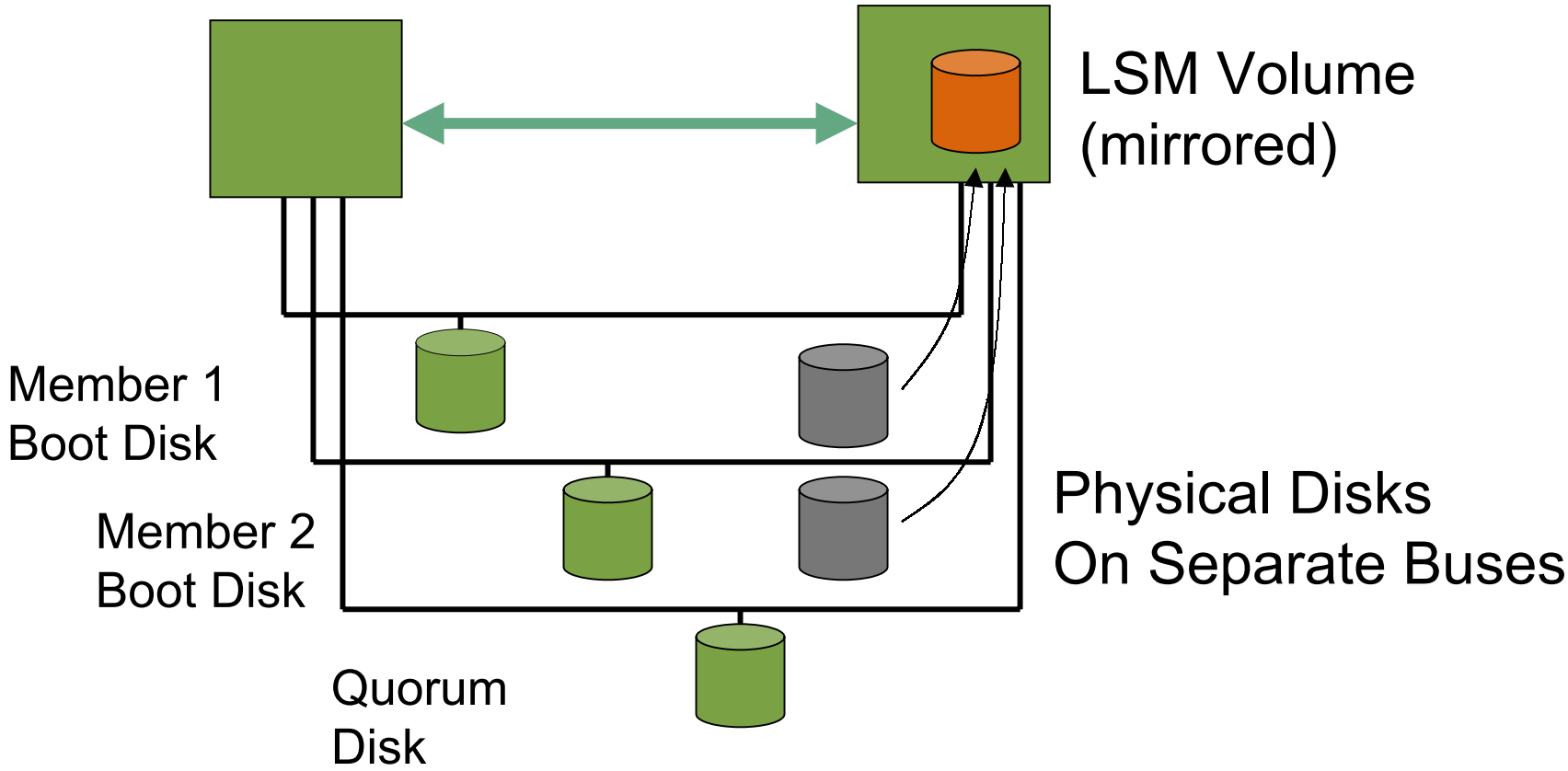
# Dual Pathing to an HSG Controller



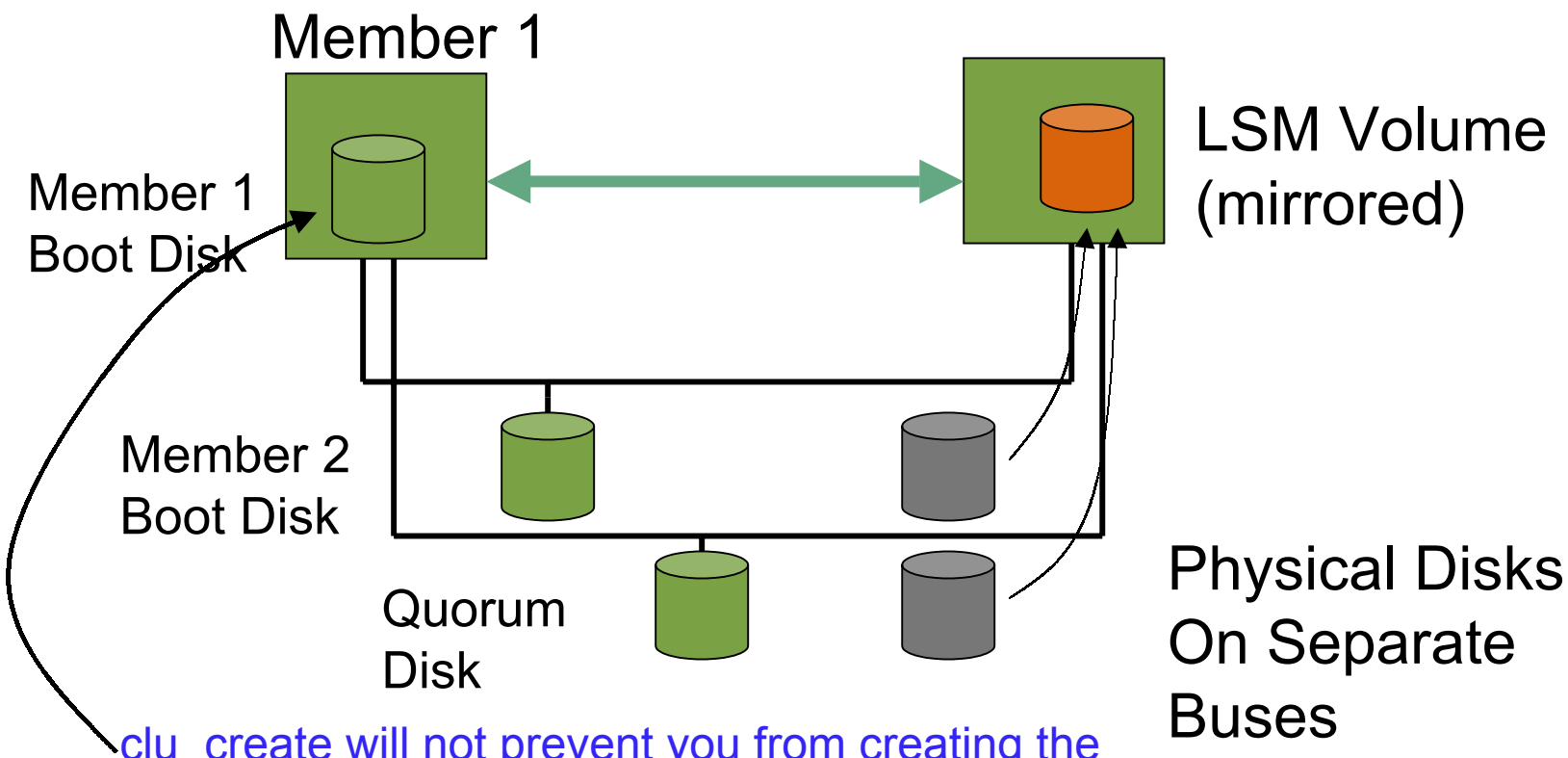

Logical Mirror
(or RAIDSet)
Presented to Host

Physical Disks
On Separate Buses

# Separate Buses for Non-HW RAID configs – 3 Buses!



# Separate Buses for Non-HW RAID configs – 2 Buses




clu_create will not prevent you from creating the member boot disk on an internal drive, but
If this drive fails, can not use clu_bdmgr to restore
Will need to rebuild cluster from scratch (clu_create)

# Disks – Other Issues

- Separation of System from Site and Application Data
  - Be able to restore system file systems, add and remove members etc. without disturbing application/data
    - Application, member specific data – careful of putting it in the system …/cluster/member/memberX/… directories

      removed with a clu_member_delete

# Some Advfs and LSM considerations



- Keep AdvFS domains to single volume, single fileset,
  - The more volumes in a domain, the higher the failure rate – any volume failure hurts the entire domain
- Be aware the LSM's automatic placement of priv areas may be too smart for itself - all disk group priv areas could end up on the same raid array
  - Sees multiple paths to all disks, doesn't figure out WWID to identify whether separate HSx pairs
  - Best to either manually confirm placement or manually place priv areas

# Topics

- Background Concepts
- Configuration Examples
- Cluster Management Operations
- Network Connections
- Storage
- **Cluster Interconnect**
- Member Systems and Quorum

# Cluster Interconnect

- Topics
  - Role and Ground Rules
  - Implementation Options
  - Looking at Memory Channel
  - Looking at LAN Based Interconnects

# Role and Ground Rules

- Required dedicated private communication path between cluster members for communication, synchronization, data
- Must be one rail (cluster interconnect path) connecting all members as a minimum
  - Can be made dual redundant for high availability

# Role and Ground Rules (cont)

- What happens when interconnect fails?
  - One Rail of Two - Individual Connections/Adaptors
    - Transparent Failover to Other Rail
  - Entire interconnect
    - Two-node cluster with quorum disk

      One node will get quorum disk and continue
    - Three or more node cluster

      If no members can communicate then no member will get more than its 1 vote and quorum will be lost
  - Interconnect performance may degrade (or become overloaded)

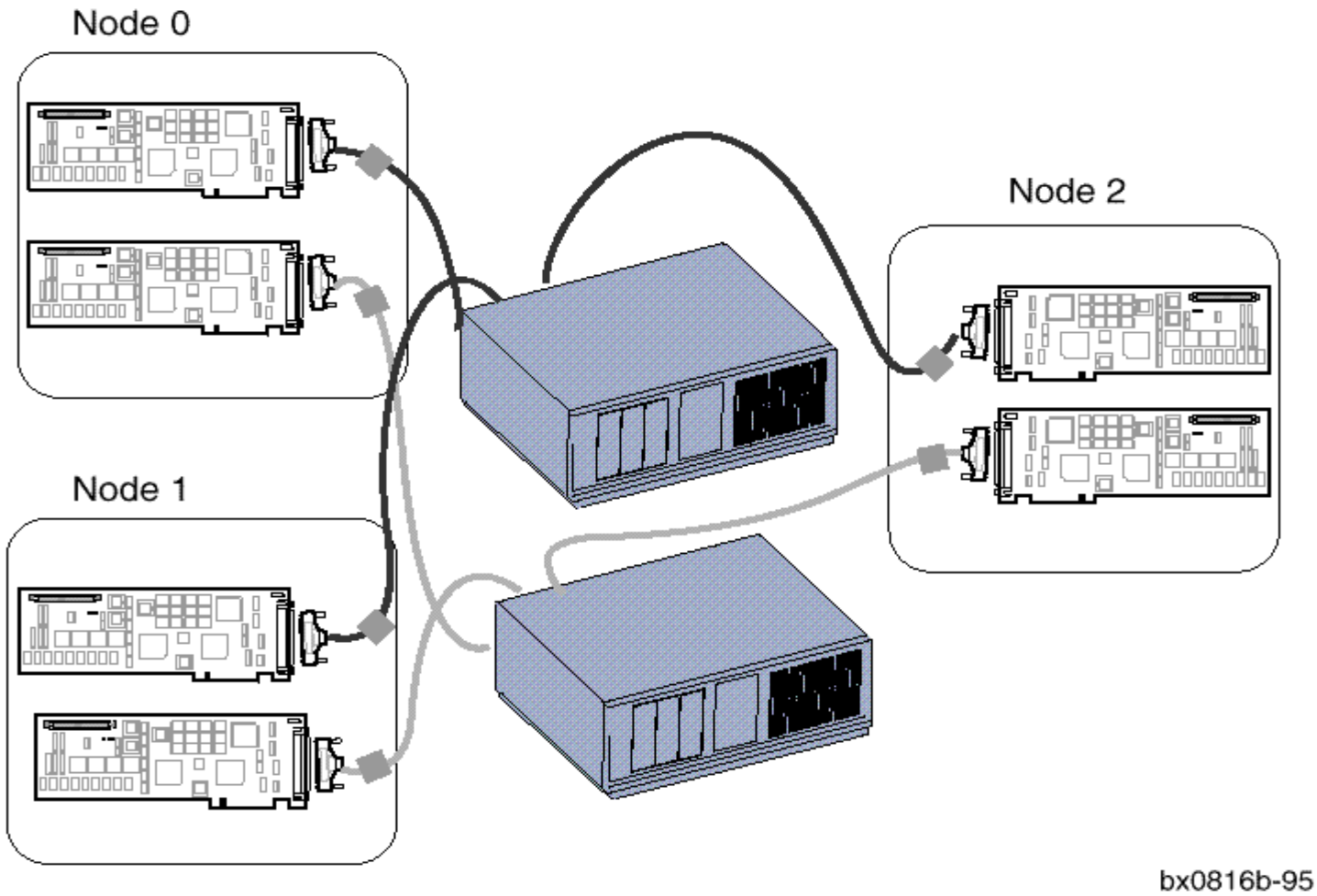
# Implementation Options

| Type | Band-width | Distance | Cost | Slots | Active/ Passive Standby | APIs | Release First Supported |
|---|---|---|---|---|---|---|---|
| Memory Channel | 100 M**bytes**/s<br>2.9 micro second latency<br>(Effective - 50 Mbytes/sec) | 20m<br>(6000 m with Fibre Link) | High | 1 rail /<br>PCI<br>(5V slot) | Passive Standby | TCP/IP Sockets<br>MCI API<br>HPTC MPI | 5.0A |
| Fast Ethernet | 100 M**bits**/s<br>(Effective - 12 Mbytes/sec) | 412m | Low | 4 rails /PCI<br>But single HW card | Passive Standby<br>(NetRAIN) | TCP/IP Sockets | 5.1A |
| Gigabit Ethernet | ~70MBytes Effective | 1600m – SPOF<br>1000m – NetRain | Med | PCI<br>(66mhz for best perf) | Passive Standby<br>(NetRAIN) | TCP/IP Sockets | 5.1A + |

# Looking at Memory Channel

- Hub/Cable state changes seen immediately as hardware failures by cluster software
  - Plan ahead for any maintenance or upgrade that would require disconnecting or modifying memory channel hardware
  - Use a hub even if a two node cluster (hub also makes growth to additional nodes easier)

# Memory Channel – Dual Rail Configuration



# Looking at LAN Based Interconnects

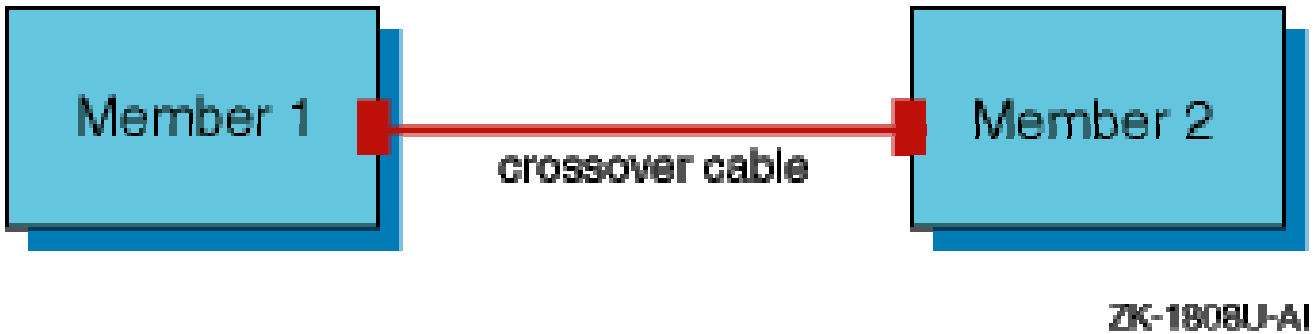


- 3 configurations supported, with 3 hops maximum
  - Single cross-over cable rail between two members
  - Single switch/hub connecting up to eight members
  - Dual switch with switch interlinks
- Private
- Aggregation not supported – only NetRAIN
- No long distance configurations – yet
- Cross over cables for single rail configs only – otherwise switches/hubs
- Be careful of switch/hub configuration for no-single point of failure, i.e., use a separate switch for each rail
  - Can also cross link switches for greater fault tolerance

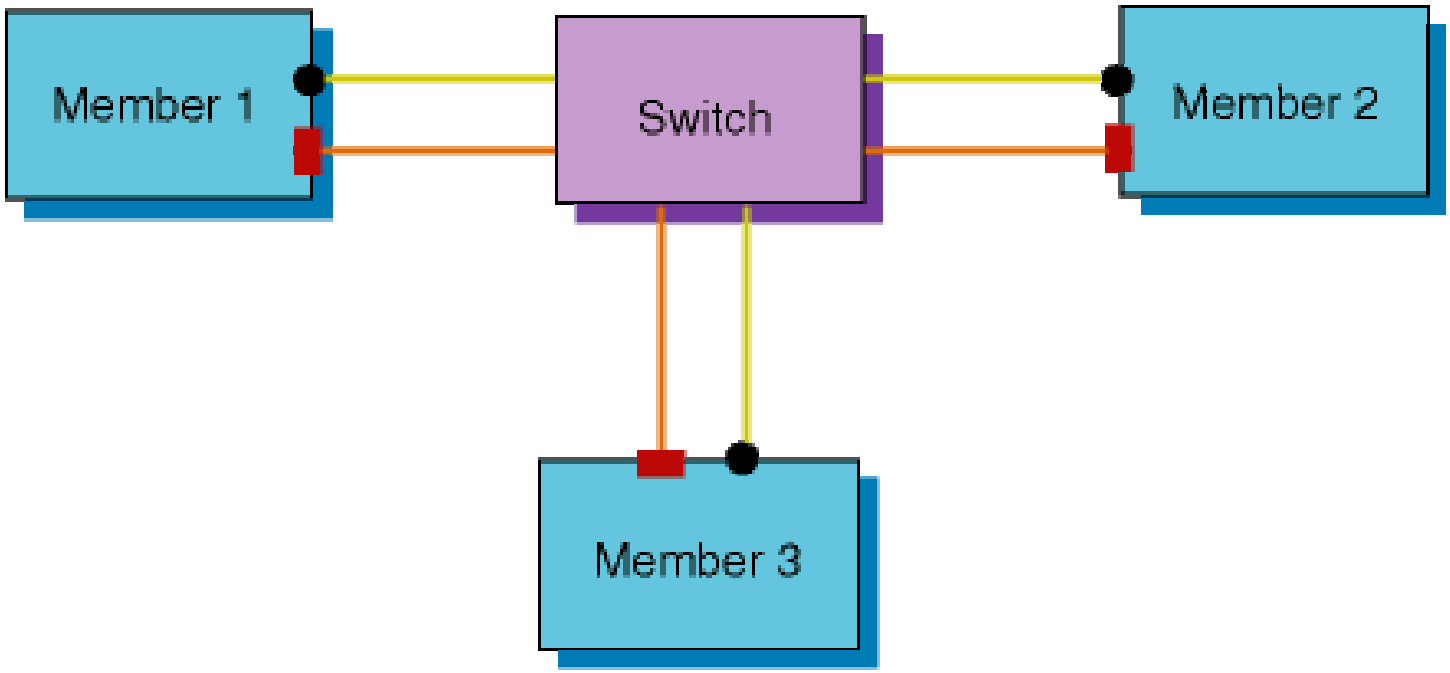
# Looking at LAN Based Interconnects (cont)

- Can you really make do with 100Mbits/sec?
  - Be careful of CFS, Application and CLUA Traffic
    - Especially in the face of HBA and NIC failures which otherwise are not single points of failure
    - LAN Configuration in documentation includes sections on tips in this area

# LAN Configuration – Crossover Cable

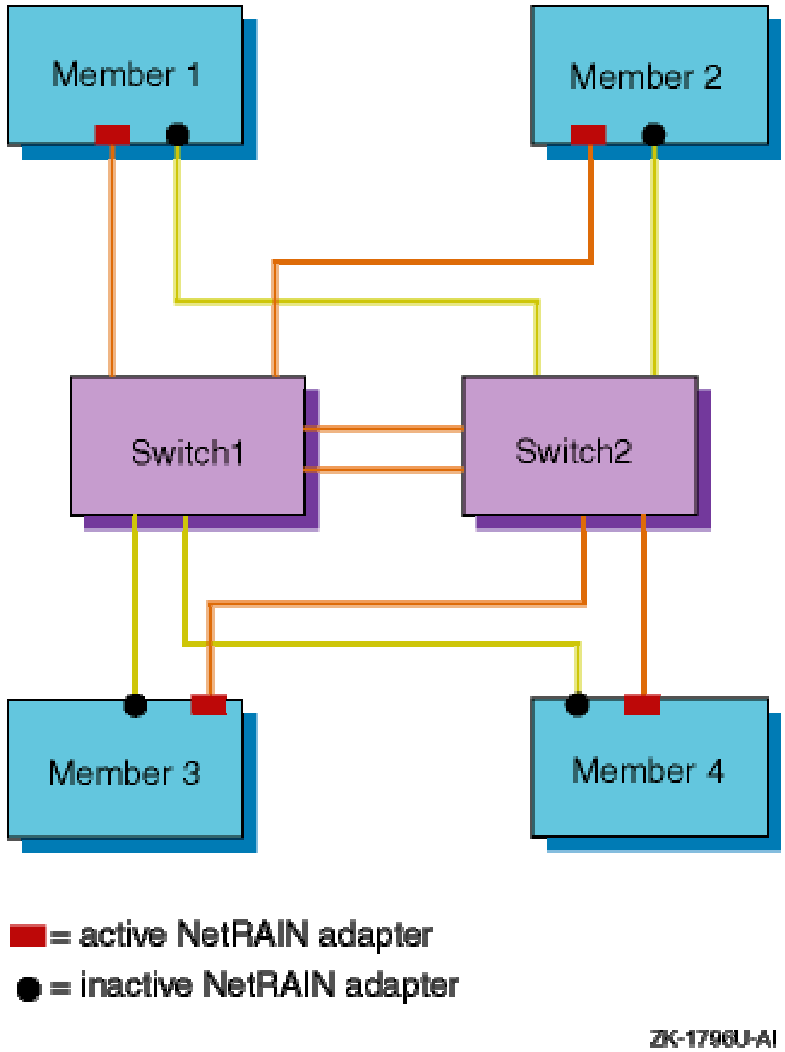


# LAN Configuration – Single Switch



# LAN Configuration – Dual Switches

Warning: No spanning tree protocol except the links between the switches.



# Topics

- Background Concepts
- Configuration Examples
- Cluster Management Operations
- Network Connections
- Storage
- Cluster Interconnect
- **Member Systems and Quorum**

# Member Systems and Quorum

- Topics
  - Characteristics
  - Compromising A Member's Participation
  - Dynamic Node Removal and Addition
  - Quorum

# Member Characteristics

- **Node Count** in Cluster: TCS 5.1B: **8 nodes**
- **Model**: AlphaServer that supports the required cluster interconnect, storage and network adaptors – just about any system from DS10 on up
- Sierra Clusters (modified TruCluster) can have up to **1024 nodes** (with some tweaks can go higher)
  - Of the Top 10 Supercomputer sites (per http://www.top500.org), 4 are HP, of those, 3 are Sierra (Tru)Clusters

# Member Characteristics

- **Memory**: 192 MB
  - Member systems that run out of free memory are not the most reliable cluster members
- **CPU**: 1 (minimum)
  - Member systems that run out of free CPU cycles are not the most reliable cluster members
- **Homogeneity**: Not a requirement
  - But can add to the complexity which can lead to problems with miss-matched loads etc. (for example after failover)

# Compromising A Member's Participation



- Member Systems must be well maintained
  - Hardware Config – UPS, Environmental, Partitions, Patch kits, Firmware, Monitoring, etc.
- Overloaded members (from a performance perspective) do not make good members
  - Cluster software designed and tested around the assumption that most members work and respond as expected or fail completely
  - Overloaded systems can become intermittently unresponsive.
- Be careful of
  - Too heavy a load
  - Unbalanced load (CFS, cluster alias, CAA)
  - Unbalanced Hardware (heterogeneous clusters)

# Dynamic Node Removal and Addition



- For various serviceability operations it is important to be able to dynamically add and remove member systems
- `clu_add_member` or `clu_delete_member`
  - Do not require that the targeted member system be a current operational member of the cluster
  - Do not require that the system be physically present!
  - This is a very good thing, as long as one member of the cluster is operational, you can delete everyone else out no matter how broken they are

# Dynamic Node Removal and Addition

- `clu_delete_member` deletes the following data
  - Member boot disk (if it can reach it, otherwise ignores) - important because of sysconfigtab!
  - …/cluster/members/memberX/… directories
    - **So, keep your member-specific application data elsewhere for easy removal and reintroduction of individual members**
  - The member from /.rhosts
  - The ics ip information from /etc/hosts

# Quorum – Definition and Purpose



- A distributed Cluster Software component, Connection Manager (CNX), maintains and enforces **cluster membership** using a “quorum” (voting) algorithm
- CNX uses the **cluster interconnect**, and for some configurations a **quorum disk** to make membership decisions
- Goal – to insure no “split-brain” or “partitioned” clusters
  - Avoid two or more different groups of systems within the cluster erroneously thinking themselves, each independent clusters
    - Yet sharing (and corrupting) shared data
    - Like a civil war without knowing the opposing side exists

# Quorum – Terms and Algorithm



- Quorum Algorithm
  - Each member system gets a logical vote
    - In TCS members can have a vote of 1 or 0
    - The total of all member's votes (assuming they are present) is the "**Expected Votes**"
    - To form a cluster and operate a subset of cluster systems must get a **majority (quorum) of the expected votes**
  - Example, 3-node cluster, each member has vote of 1
    - Requires at least two systems (votes = 2 out of expected of 3) for the members to operate as a cluster
  - Quorum Votes = round_down ((Expected Votes+2)/2)

# Members Only - The Numbers without a Quorum Disk

| Size of Cluster | | Expected Votes | Number of members to form a Cluster | Percentage of members to form a Cluster |
|---|---|---|---|---|
| 1 | | 1 | 1 | 100% |
| 2 | | 2 | 2 | 100% |
| 3 | | 3 | 2 | 66% |
| 4 | | 4 | 3 | 75% |
| 5 | | 5 | 3 | 60% |
| 6 | | 6 | 4 | 66% |
| 7 | | 7 | 4 | 57% |
| 8 | | 8 | 5 | 62% |

# Quorum – Why and How a "Quorum Disk" (cont)

- For even-numbered-node-count clusters
  - Increases availability
  - Acts as a "tie-breaker" (in case of cluster partition)
- Loss of quorum disk itself, in an otherwise normal odd-node count cluster
  - Is not a single point of failure because the cluster will still have enough votes from members
  - Does not cause any performance degradation of anything

# Members and Quorum Disk - The Numbers



| Size of Cluster | Quorum Disk / Quorum Disk Votes | Expected Votes | Number of members to form a Cluster (assuming quorum disk ok) | Percentage of members to form a Cluster |
|---|---|---|---|---|
| 1 | No | 1 | 1 | 100% |
| 2 | 1 | 3 | 1 | 50% |
| 3 | No | 3 | 2 | 66% |
| 4 | 1 | 5 | 2 | 50% |
| 5 | No | 5 | 3 | 60% |
| 6 | 1 | 7 | 3 | 50% |
| 7 | No | 7 | 4 | 57% |
| 8 | 1 | 9 | 4 | 50% |

# Acknowledgements and Resources

Original Presentation by Brad Nichols (Tru64 UNIX Engineering)

*TruCluster Server Handbook*, Digital Press

http://h30097.www3.hp.com/cluster/

gry@hp.com


Interex, Encompass and HP bring you a powerful new HP World.