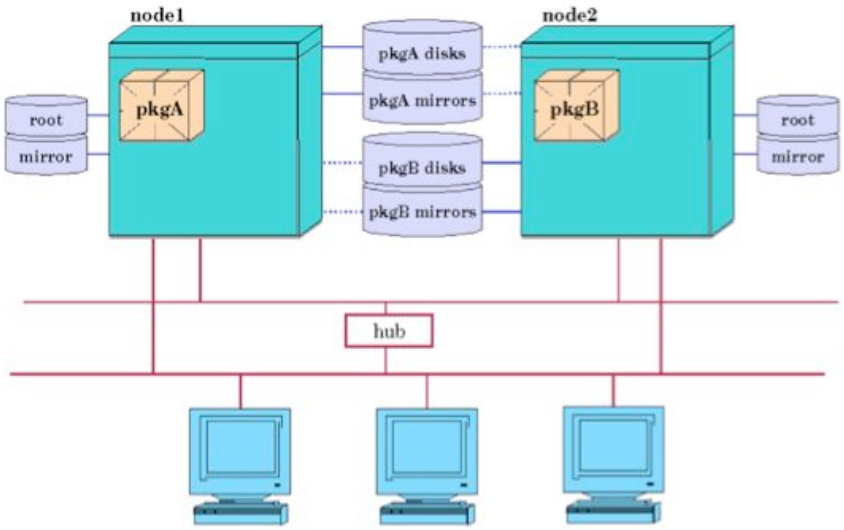
High Availability Systems (cluster)

High Availability Systems

- HA systems is a special configuration that allows application services to continue in spite of a SPoF.
- SPoF can be at software, CPU, disk, network or power supply level
- The basic idea is the complete redundancy of system elements.
- HA is a special layer added to OS. HA instructions run usually in priviliged mode
- The cost of the system increases!
- The complexity of the system increases!

Typical configuration



The basic idea: 2 coupled systems with the proper software to manage failures

High Availability Goals

- Elimination of both the planned or unplanned downtime
- Elimination of SPoF
- Fault Resilience, NOT fault tolerance. High availability is NOT Fault Tolerance

Fault tolerance: No service interruption. Everything is reduntant and working in a parallel way

High Availability: minimal service interruption. HA is less expensive. In many HA configuration backup resources are available for normal operation

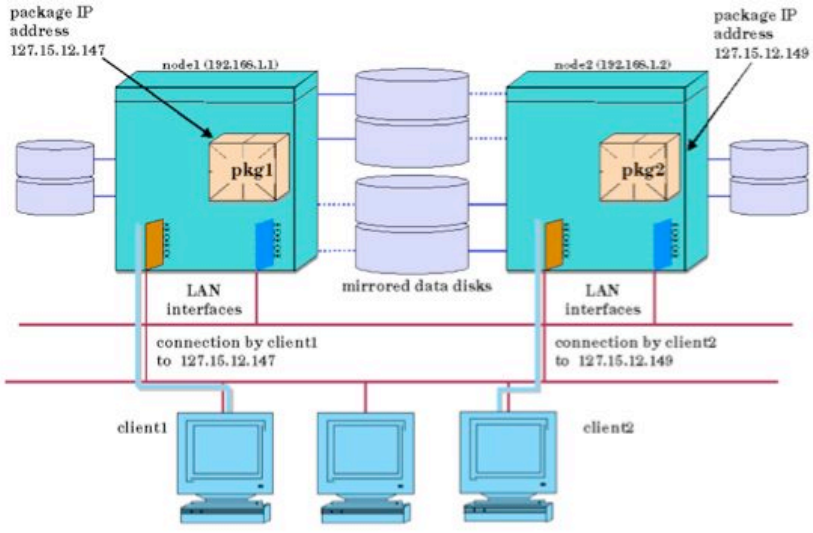
Planned downtime is 80% of the total downtime. It is typically due to maintenance, software or hardware upgrade

Unplanned downtime: HW and SW failure

Terminology

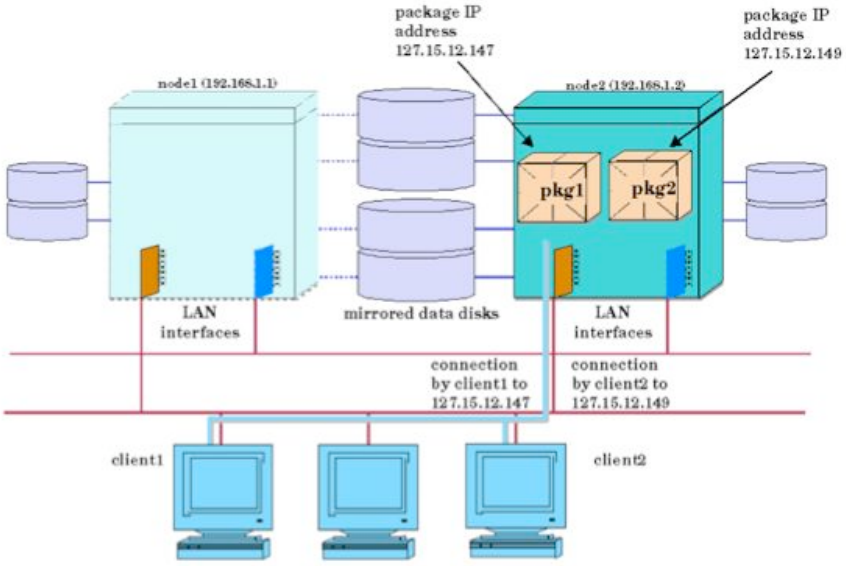
- Cluster: networked group of nodes
- Node: each single server or SPU (service processor unit)
- Package: a set of resources defining an application. A package is the basic unit of a cluster. A package has storage space, network addresses, users, application services..
- Active/Primary/Adoptive node: for each package a primary and one or more adoptive node must to be defined
- Failover: situation in which one or more nodes fail. Packages are switched according to the failover policy defined, A failover is followed by a cluster reformation process. When the cluster is reformed, the failback policy defines on which node the package will run in the new cluster.
- Heartbeat message:signal from each functioning node to transmit their status. (that they are “up”)
- Cluster Quorum: minimum number of node to form a cluster

Before a failover



IP addresses are associated with Packages, not with nodes. The users are connected to the package logical address. When the cluster is not working (for example halted by System Admin., server is accessible with the stationary node ip address (192.169.1.1 for node 1)

After a Failover



Problems after a failover: ip address and routing, users, application versions, permissions, patch levels, OS level, disks configuration (import/export/mounting), cluster reformation, failback and failover..

Package Failover Configurations

- No failover: a package can run only on a specific node.

Why?

Case of nodes with different performance

The package needs too much resources to run on others nodes. It may cause the other node to crash!

Performance (a kind of server affinity)

- Configured Node List: a list containing primary and adoptive nodes in order or priority. If no priority is defined, the node with the lowest number of package running has the highest priority

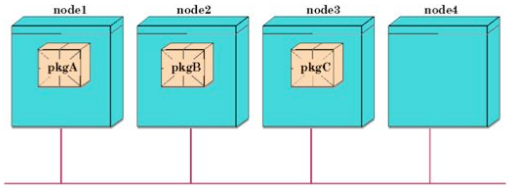
2 policies:

- Automatic rotating standby: the package is switched to the node with the minimum amount of package running
- Configured Node failover: the package is switched to the next node in the Configured Node List

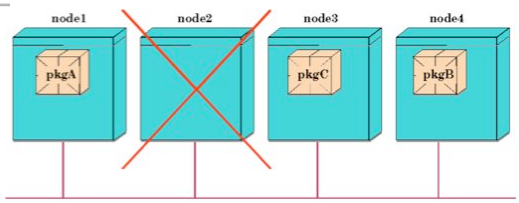
Failover Policy: examples

Package Name	NODE_NAME List
pkgA	node1, node2, node3, node4
pkgB	node2, node3, node4, node1
pkgC	node3, node4, node1, node2

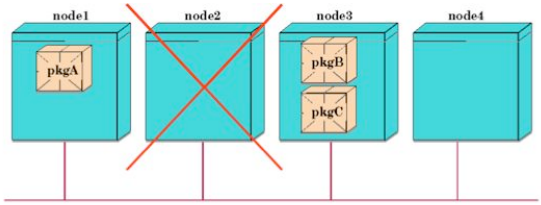
Before a Failover



After failover with automatic rotating standby



After failover with configured node



Failover policy: 2 policies

- Automatic: when a node with the highest priority re-joins the cluster, the package is switched
- Manual: the package is switched using command-line

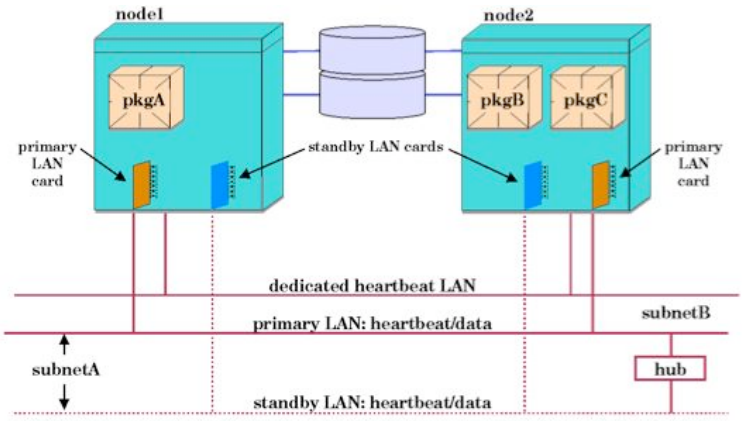
Make network components redundant

- Every subnet accessed by cluster must have redundant network interface
- Redundant cables needed
- Hub, bridge, switches, concentrators must be redundant

On each server:

- Primary interface: an interface that has an IP address associated to it
- Standby Interface: no IP address

Redundant Network Component: LAN

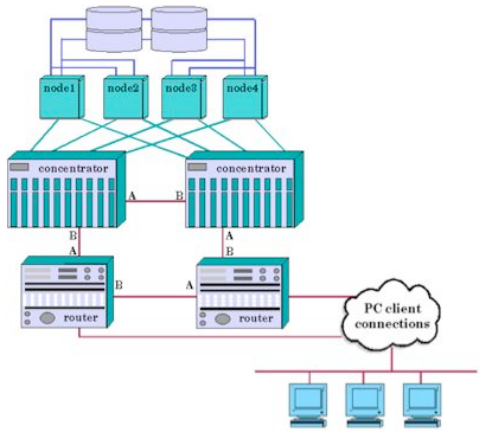


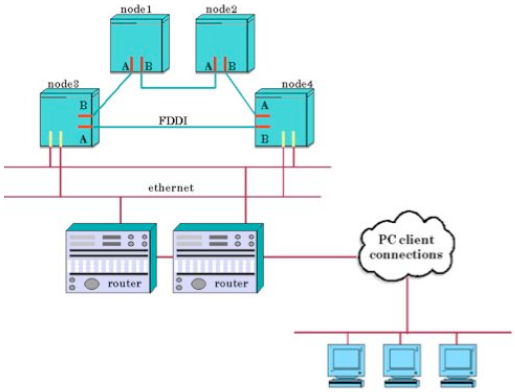
Subnet B only for heartbeat message. A single failure of the hub can be tolerated as well. Hub needed if subnet fails.

If the primary LAN of node1 fails => local failover and the standby card becomes the primary. Node 1 communicate with node 2 using hub.

Heartbeat line redundant!

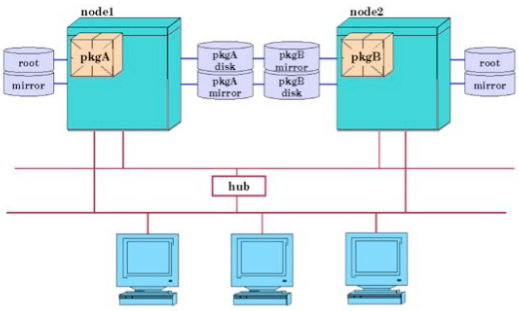
Redundant Network Component: Fiber Optic

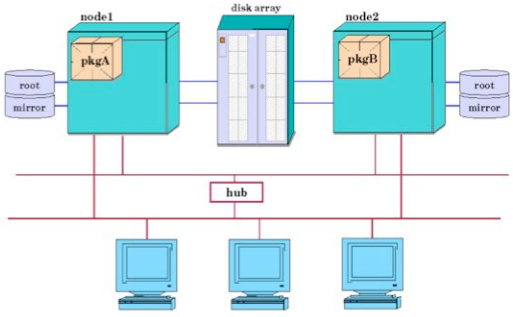


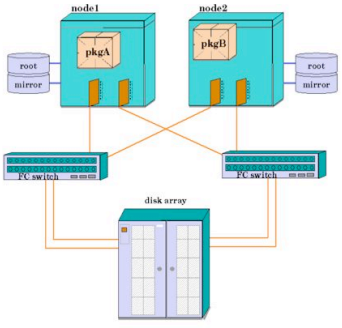


Star or ring configuration. Router Configured to send package in both direction. Redundant cables, concentrators, concentrators ports, routers, routers ports. By Using a serial line for heartbeat I can add more redundancy for the heartbeat

Storage Configuration







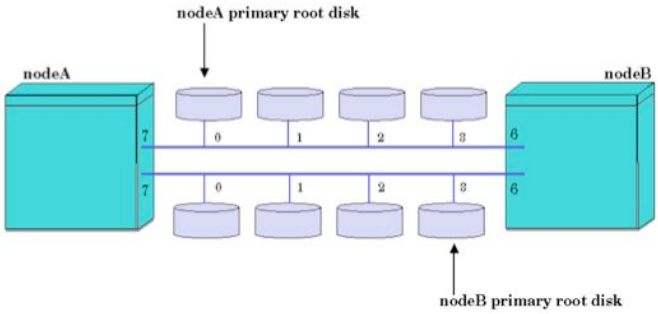
3 Configurations:

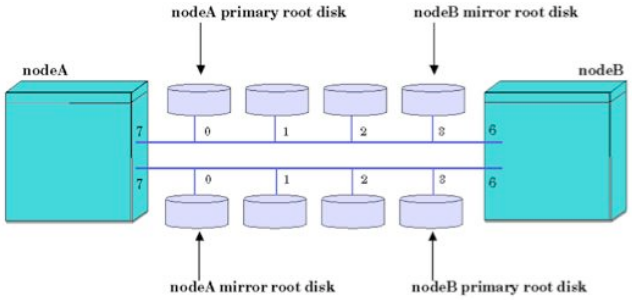
1. Shared Disk (SCSI..)
2. Disk Array (RAID, full redundancy, Business Copy)
3. Storage Area Network (SAN)

Storage Configuration: boot

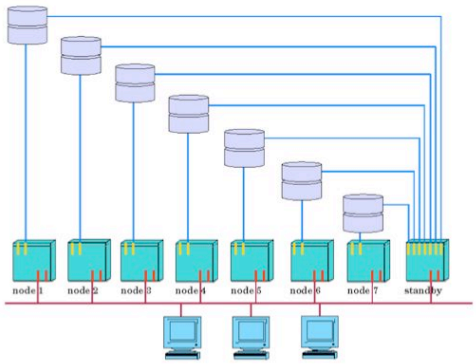
Using only SCSI disks there is the limitation that 2 servers cannot boot at the same time from the same SCSI bus. Primary root disks must be on different SCSI bus.

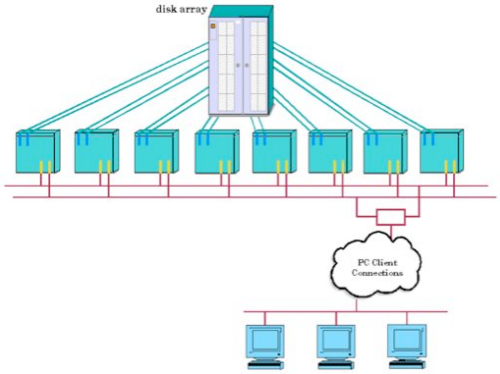
The mirrored image of the primary root disk can be on the same bus of another node primary root disk. The possibility of a failure is very low: the primary node must fail, and node b and a rebooted at the same time!





Multi-server Storage Configuration





- 7 nodes, one standby
- 8-node cluster using disk-array

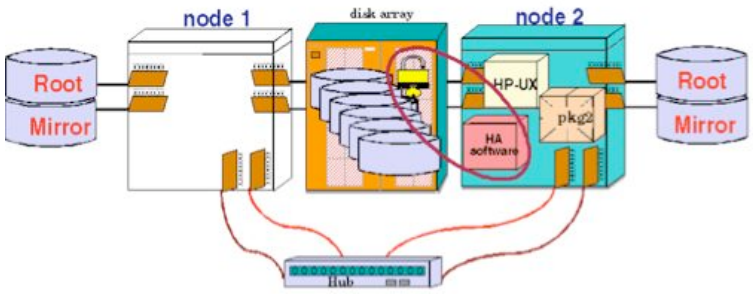
The problem of cluster reformation

- Re-formation occurs
- when one node reboots (planned or failure)
- when the cluster fails
- one inactive node enters the cluster (user's command)
- a node is halted
- the heartbeat is lost.

The problem of quorum

- During reformation, a group of node tries to form a cluster. If they don't have the quorum, they cannot start the cluster.
- What if the nodes, during reformation, are divided in two groups with the same number of elements? 2 clusters running?

Cluster Quorum



- Always a strict majority needed.
- In case of an even number of nodes, a tie-breaker is used, usually a cluster lock disk.

A cluster lock disk is a special area on 1 LVM disk located in a volume group that is shareable. When a node obtains the lock disk this area is marked.

Single or Dual lock disk. A single can be a SPoF (example of the power circuit). A dual disk is not a redundant lock disk, you need both to run the cluster.

The disk must be active when there is a cluster reformation!

Alternative: Quorum Server outside the cluster instead of a disk.

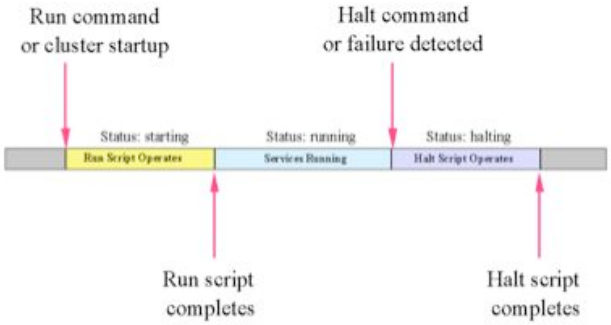
How the cluster works

- We mainly refer to HP-UX Service Guard
- We have to analyze these fundamental component
- Package Manager
- Network Manager
- Disk Manager
- Command line interface

Package Manager

- It runs, halts, switches packages
- Monitor the status of nodes/packages
- Failover/Failback policy + Node list used to manage the switching
- Each package has a start and halt script associated to it.
- Start script is executed by the cluster Package Manager in case of:
  1. Cluster Reformation (in case of a rejoin, only if automatic package fail back is active)
  2. After a package switch (due to a failure of the package on the node)
  3. If the command cmpkgstart is launched
- Halt script is executed by the cluster Package Manager in case of:
  1. Node failure, if fast failover is set to false (otherwise package processes are killed)
  2. If the command cmpkghalt is launched

Package Time Line



1. Before the control script starts
2. During script execution
3. While package's services are running
4. When a service (or network, or disk) fails
5. During halt execution script
6. When the node is halted with a Command
7. When the node fails

Before control script start

- The package can have one or more nodes eligible to run it
- A node is selected by the packet manager (only node in the node list and compliant to failover policy)
- The node must have all the resources available
- Resource are: subnet network, storage, monitored services on which the package is dependent.
- Each package has a start script associated to it (customizable by the Sys. Administrator). It contains all the information needed for starting the package. It must be present on all the nodes where the package is allowed to run.
- When the node is selected, the script is launched

During Start Script Execution

The Start Script for the package contains the information in order to:

1. Volume Group Activation
2. Mount file systems
3. Assign package IP address to LAN card on the node
4. Execution of customer-defined run commands
5. Start each package service
6. Start package resources needed by the package that were specially marked for deferred startup

During Start Script Execution (2)

Exit Codes:

0 - Package Started normally

1 - Abnormal exit. The package can't start on that node and there is no other node available. The package is disabled

2 - alternative exit (Restart exit). There was an error, but the package is allowed to start on another node

Timeout - The script failed to launch the package before the timeout

While services are running

- Package Manager will monitor all the resources needed by the package
- Resources started by the start script are monitored using PID of the processes started by the package.
- Subnets (monitored by Network Manager)
- Configured resources on which the package depends (started maybe by other packages)
- Each service can be restarted X times before declaring it failed and begin the failover procedure

When a service, Subnet or resource fails

- A service fails after X restart fail
- A subnet fails and there is no standbys
- A node fails

In case of failure, halt script is executed (graceful failover), but it is possible to set up a forced failover (halt script not executed). In this case, all the package-dependent processes are killed. If there are processes hanging, the package will not be able to start on other node.

You use force failover to reduce the downtime

If the package is enabled to run on another node, the failover procedure is started.

When a package is halted with a command

- HP UX has *cmhaltpkg* to stop a package by command line
- The halt script is always executed

During Halt Script

1. Halts any deferred resources that had been started earlier
2. Halts all package services
3. Executes any customer-defined halt commands
4. Removes package IP address from the LAN card node
5. Unmounts file systems
6. Deactivates volume groups

Exit code like the start script. After a TIMEOUT, package manager will kill all package processes

The Network Manager

- Detect faliures and Recover from network card and cable failures

Stationary and floating IP addresses.

Each node has at least one IP address for each active network interface (stationary IP). When the node is not part of the cluster stationary IP are used. They are NOT associated with packages but with physical node.

IP address associated with packages are called floating IP addresses. They are used only when the package is running. They are used on the primary LAN if available. DNS should be configured to associate package name to floating IP address.

Stationary and Floating IP address can both switch to a local standby card

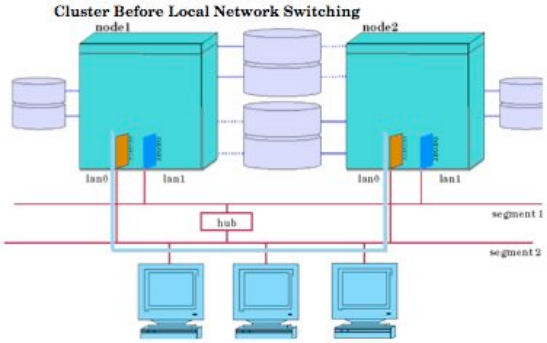
Only Floating IP address can be taken over by a LAN interface on an adoptive node.

The Network Manager

- The network Manager monitors the network interfaces with a polling strategy.
- One interface on one node is assigned to be the poller. One puller for each bridged subnet.
- The poller will poll the other primary and standby interfaces on the same bridged whether they are still healthy
- Usually standby card are used for polling (less traffic for active interfaces)
- The poller simply sends packages periodically and keeps the number of packets received by each LAN in the bridged net. If the count of packets for a specified interface doesn't increase during a configured time, the network is declared down.

The Network Manager: local failover.

The standby LAN must not have an IP address configured



TCP/IP connections are NOT lost.

For IPv4, Ethernet, Token Ring and FDDI ARP protocol is used. The operative system sends out an unsolicited ARP to notify remote systems of update the addressing mapping between MAC address and IP address.

IEEE 802.3 doesn't have rearp function, so TCP/IP connections lost

The Network Manager

- IPv6 uses NDP protocol to discover neighbors. NDP does:

1. Determine the link-layer address of neighbors on the same link and quickly purge cached values invalid
2. Find neighboring routers willing to forward packet
3. Actively keep track of which neighbors are reachable, and which are not, and detect changed link-layer address
4. Search for alternate functioning routers willing when a path to a route fails

○

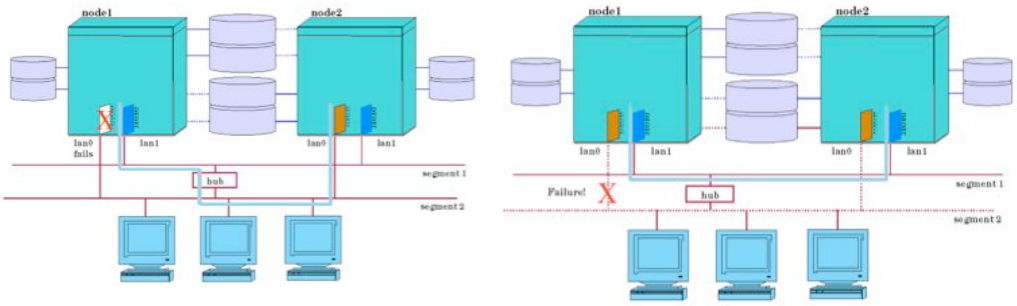
TCP/IP connections are not lost.

IP packages are lost, but they will be resend.

UDP-based applications don’t resend the package, but they should be prepared for it.

Local switch is not supported between LAN of different types at the moment

The Network Manager: Remote Failover



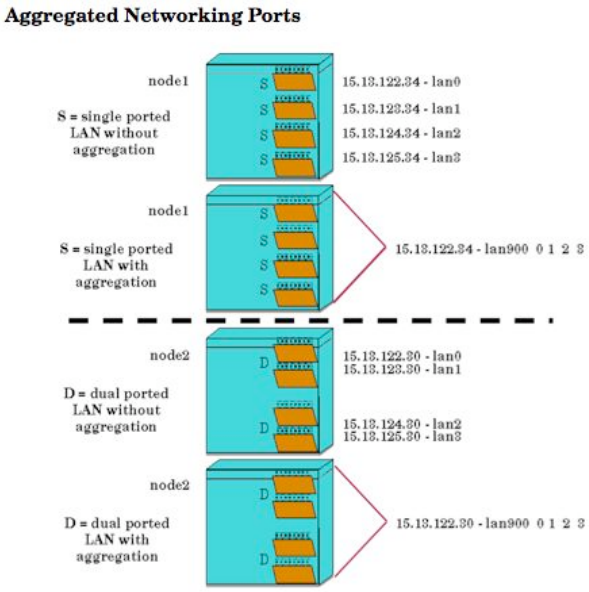
LOCAL REMOTE

- The standby LAN must not have IP address associated to it

With a remote switching :

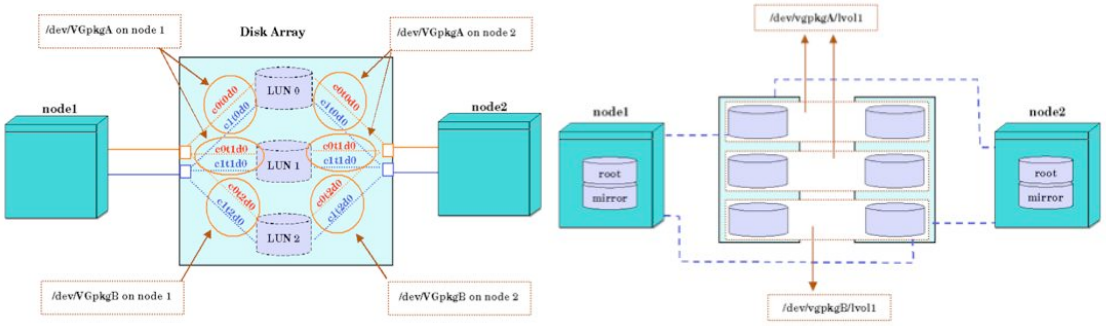
- TCP/IP connections are lost. Applications must connect again.
- LAN of the same type
- ARP broadcast sent to notify the change of MAC address (ARP request with no reply, receiver and sender are the same)

Automatic Port Aggregation



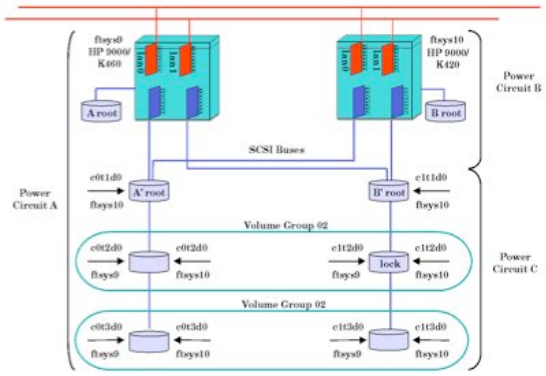
- More physical interface aggregated with one IP address.
- Multi-ported network interface can be aggregated as well.
- Fault tolerance at IP level a failure of a single port is transparent. No local or remote failover until there is at least one interface working in the aggregation.

Disk Manager



- Multiple path to disk. In a Disk Array physical location of disk sis managed by the Disk Array. Hard to identify from OS.
- File systems and Volume Group configured on all the adoptive node! That means definition of VG, Physical Disk, Mount Point, Permissions, VG mode (exclusive, shared, concurrent).
- All this information are inserted in the package control file.

Setting up a cluster



- Performance, cost, fault tolerance to be considered…

Commands

- Many commands are used to manage the cluster:
- Monitor Cluster and package status
- Modify cluster failback/failover policy
- Commit the configuration
- Distribute configuration (ASCII file)
- Halt/start Cluster
- Halt, Start, Switch Package
- Test network connections
- Monitor services

Example: monitor the cluster with cmviewcl -v


```
CLUSTER      STATUS
example      up
  NODE         STATUS       STATE
  ftsys9       up           running

    Network_Parameters:
    INTERFACE    STATUS       PATH         NAME
    PRIMARY      up           56/36.1      lan0
    STANDBY      up           60/6         lan1

    PACKAGE      STATUS       STATE        AUTO_RUN     NODE
    pkg1         up           running      enabled      ftsys9

      Policy_Parameters:
      POLICY_NAME      CONFIGURED_VALUE
      Failover         configured_node
      Failback         manual

      Script_Parameters:
      ITEM         STATUS       MAX_RESTARTS   RESTARTS        NAME
      Service      up                      0          0        service1
      Subnet       up                      0          0        15.13.168.0

      Node_Switching_Parameters:
      NODE_TYPE    STATUS       SWITCHING    NAME
      Primary      up           enabled      ftsys9       (current)
      Alternate    up           enabled      ftsys10

 NODE         STATUS       STATE
  ftsys10      up           running

    Network_Parameters:
    INTERFACE    STATUS       PATH         NAME
    PRIMARY      up           28.1         lan0
    STANDBY      up           32.1         lan1

    PACKAGE      STATUS       STATE        AUTO_RUN     NODE
    pkg2         up           running      enabled      ftsys10

      Policy_Parameters:
      POLICY_NAME      CONFIGURED_VALUE
      Failover         configured_node
      Failback         manual

      Script_Parameters:
      ITEM         STATUS       MAX_RESTARTS   RESTARTS        NAME
      Service      up                      0          0        service2
      Subnet       up                      0          0        15.13.168.0

      Node_Switching_Parameters:
      NODE_TYPE    STATUS       SWITCHING    NAME
      Primary      up           enabled      ftsys10      (current)
      Alternate    up           enabled      ftsys9
```

Output of the command cmviewcl. The name of the cluster is example, it has 2 nodes running, each of them with one package running.