

# Demystifying HACMP

Dan Braden  
Advanced Technical Support  
Washington Systems Center

[dbraden@us.ibm.com](mailto:dbraden@us.ibm.com)

May 14, 1999

## Agenda

Availability environments and challenges

Idle standby concepts

- IP address takeover - IPAT

- Shared disk takeover

- Resource groups

Other clusters

Network availability

HACMP on the SP

FDDI, ATM and disk specifics

HACMP flavors

Implementation and skills

Other availability products

Documentation

## Availability environments & challenges

High availability vs. fault tolerance

Single points of failure - SPOF

Site disasters

24x7 vs. 8x5

Applications

- Post crash processing
- Application crashes
- *HACMP requires an application that doesn't need manual intervention after a system crash*

Backups

Concurrent HW/SW maintenance

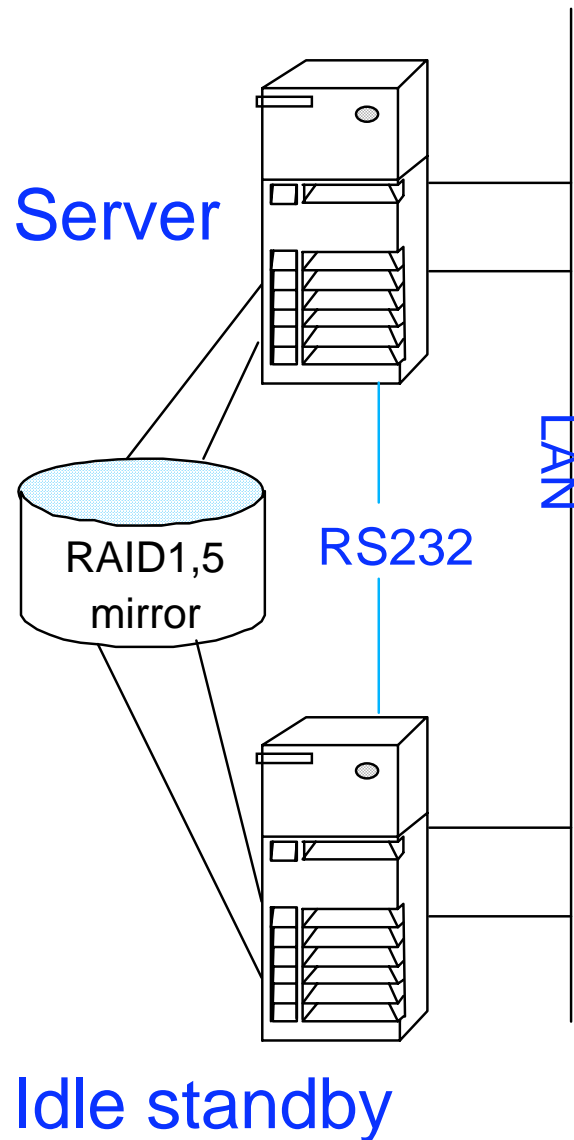
One should not confuse high availability with fault tolerance. The distinction is that with fault tolerant systems, hardware failures are transparent to applications and users, and some form of concurrent maintenance is generally provided. High availability solutions, such as HACMP, generally have short outages when hardware fails. Another distinction, is that fault tolerant systems generally aren't as open as high availability solutions; consequentially, fewer applications are available for fault tolerant systems. From a cost viewpoint, fault tolerant systems are generally much more expensive. Both high availability solutions and fault tolerant systems generally don't handle software failures, though other facilities exist that can restart applications (e.g. the SRC or use of /etc/inittab) or cause a node to fail when an application fails; thus, allowing the other node to start the application.

**A key assumption** of both fault tolerant and highly available systems is that the system is sufficiently reliable that one is highly likely to be able to repair or replace a broken part before something else breaks. With HACMP one must also design in no single point of failure (SPOF), as HACMP does not provide availability for power, disk, or components outside the cluster.

One of the challenges of managing high availability or fault tolerant systems is system maintenance. The operating environment, 8 hours a day 5 days a week to 24 hours a day 7 days a week, has a big impact on managing the system(s). The prior case has a long maintenance window available, the later none. One must ask how they will maintain the system with the available maintenance window. How will backups be performed? How will software changes be tested? How will configuration changes be tested? Customers with limited maintenance windows often have a test cluster used for development and testing prior to implementation in production.

One should note that HACMP requires that an application is capable of restarting without manual intervention after a system crash. Most major applications meet this requirement, and often those that don't can have the recovery steps written in a script.

# Idle standby concepts



## Clients access server via TCP/IP

- ASCII terminals attached via terminal server

## Heartbeats and failure detection

- Heartbeats: IP, RS232, tmscsi and tmssa
- Failures: adapter, network and host

## Disk takeover

## IP address takeover

## Application start/stop scripts

## Resource groups

- IP address, disk and start/stop scripts

## Failure notification

- Adapter, network and host
- AIX error log errors

## Fallover time

## False takeover

- Dead man switch

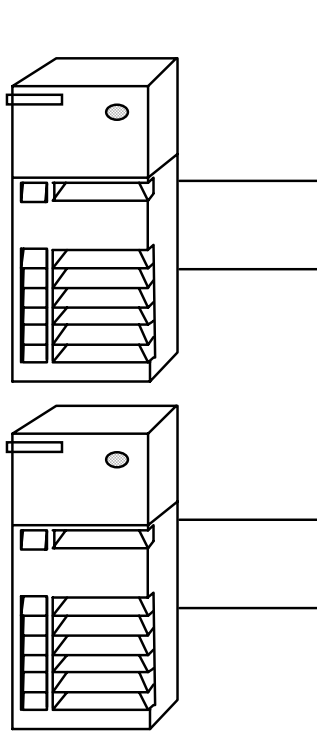
The main concepts of HACMP apply to a simple idle standby configuration. Generally, the hot standby node is used for development or other purposes until the server fails. HACMP clusters are characterized by: clients access the server via a network, highly available data resides on shared disk, and multiple heartbeat paths exist for failure detection. **Most flavors of HACMP only detect three kinds of failures: node, network and network adapter.** By default, HACMP doesn't do anything when the network fails. When nodes or adapters fail, their operations are moved to another node or adapter. This results in a disruption of short duration (30 sec. to several minutes depending upon a number of factors) and clients usually lose access to the application and must reconnect. Failures are detected via lack of heartbeats, and this results in an event - HACMP is event driven. HACMP doesn't handle disk or other failures, and one must include facilities to eliminate single points of failure, e.g. mirroring, RAID 1 or RAID 5 for disk, UPS for power, etc.

A resource group includes a highly available IP address, shared disk (usually), and application start and stop scripts. Resource groups move between nodes in a cluster.

Failure notification is very important, as HACMP is designed to handle only one failure which must be corrected before another failure occurs.

False takeover can occur if system and network performance are not planned and managed, so HACMP can send and receive heartbeats. HACMP uses a dead man switch facility to assure it can access the kernel in a reasonable time, thus, if the system CPU is overloaded, the dead man switch causes the system to crash which prevents multiple systems from fighting over the resources. The failure detection rate is configurable and can help prevent this condition, but performance monitoring and planning is very important.

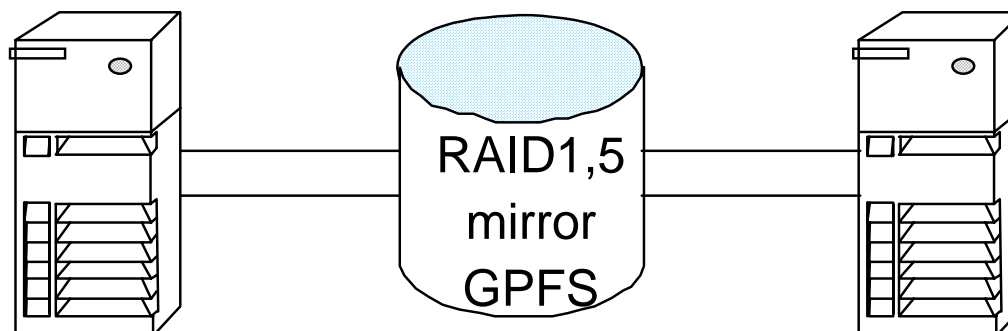
## IP address takeover (IPAT)



	At system boot	With HACMP running	After adapter failure	After host failure	Adapter Type
	1.1.1.2	<b>1.1.1.10</b>	na	na	Boot/ <b>Service</b>
	2.2.2.2	2.2.2.2	<b>1.1.1.10</b>	na	Standby
	1.1.1.3	1.1.1.3	na	1.1.1.3	Service
	2.2.2.3	2.2.2.3	1.1.1.3	<b>1.1.1.10</b>	Standby

- Two logical IP networks, Netmask 255.255.255.0
- One physical network
- Clients always access **1.1.1.10**
- MAC address takeover or ARP cache update is also needed

# Shared disk takeover



## On host A:

1. Configure disks - `cfgmgr`
2. Create VGs - `mkvg`
3. Create LVs/FSs  
`mklv crfs mklvcopy`
4. Set `autovaryon` to no - `chvg`

## On host B:

5. Configure disks - `cfgmgr`  
Check PVIDs
6. Import VG - `importvg`
7. Set `autovaryon` to no

- VG varied on to one node at a time
- SCSI reserve restricts access to one node
- HACMP varies on the VG
- Disk names may be different on each system
- LVM changes
  - CSPOC - HACMP 4.2+
  - Lazy update at HACMP 4.2+
  - Better than lazy update with `bos.rte.lvm 4.2.1.4 +`
- GPFS setup is via RVSDs



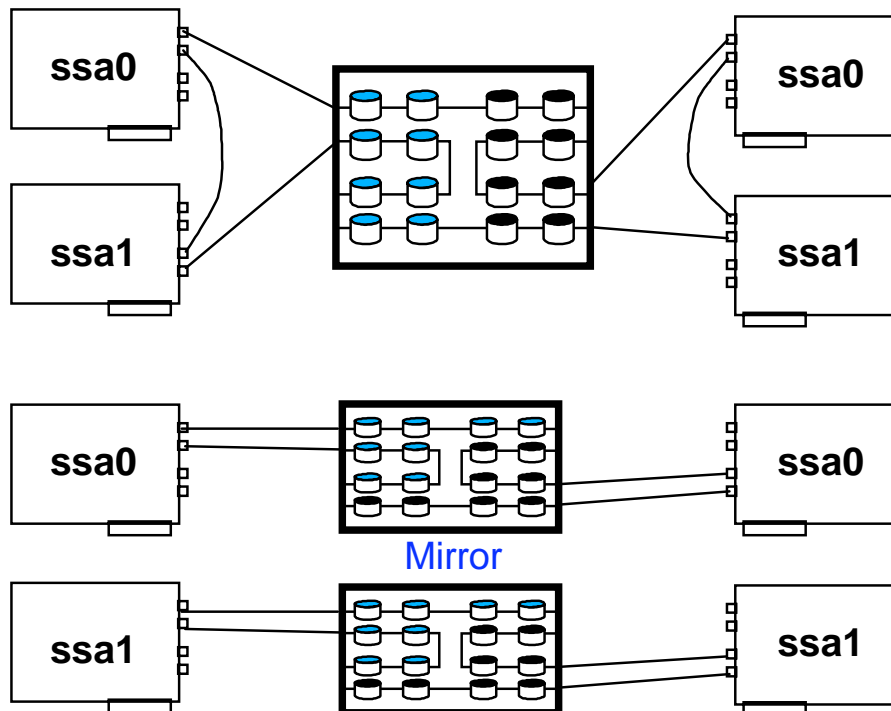
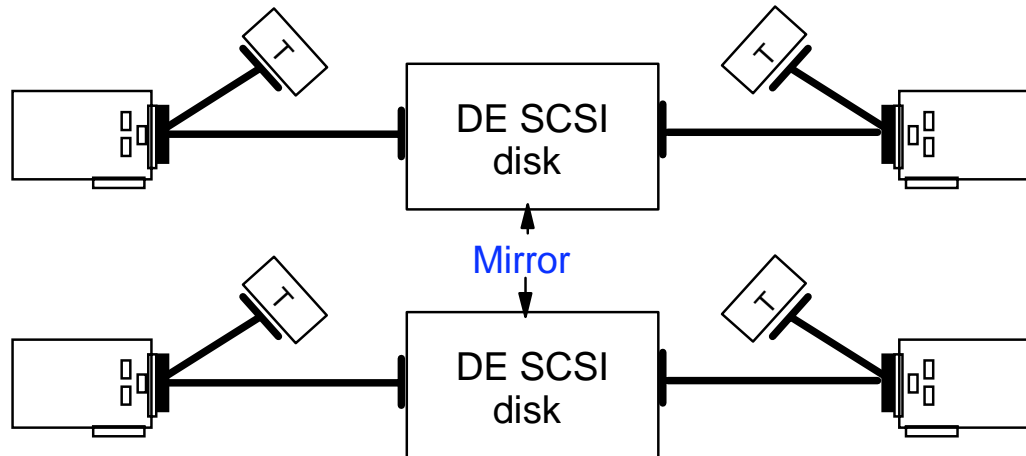
Disk setup and maintenance require skills beyond that used for a single system. First, one must build in no SPOF for the design: typically using mirroring, RAID 1 or RAID 5, and either separate adapters for each copy of the data or dual adapters for access to the disks (with appropriate software support) to handle adapter failure. One must also consider device and LVM issues (like different disk names on each system), and keep the ODM on each system up to date after changes to the volume group (VG): e.g. after replacing a disk, adding a LV to the VG or putting a PVID on a disk.

One must also understand the SCSI reserve (which also applies to SSA) which keeps multiple systems from accessing the same disk at the same time. When a VG is varied on, the SCSI reserve is set (the exception being concurrent VGs). It is released when the VG is varied off, the power to the disk is cycled, or when HACMP issues a low level SCSI command to release the reserve. Shutting down a system normally doesn't varyoff the VG.

Managing changes to LVM has improved with new releases. Previously, one had to stop the cluster for a short time to export/import changed VGs on other nodes. Next came the "lazy update" which used a timestamp to determine if a VG needed to be exported/imported at fallover. AIX LVM was then changed to provide a "better than lazy update" which did not require an export/import on nodes sharing the VG. CSPOC has been improved to use the better than lazy update so that most VG changes can be handled cluster wide for both standard and concurrent VGs.

An alternative method to provide a highly available filesystem is to use GPFS implemented with RVSDs. In this case HACMP does not handle disk takeover when a VSD server node fails. Instead, PSSP's RVSD code does. A GPFS server node can also exist in an HACMP cluster, but the VSD disks are handled by PSSP.

# Disk alternatives



## SCSI

- Two busses for no SPOF
- Y cables, to disconnect hosts for maintenance
- SCSI IDs on adapters, 5 & 6
- Remove internal terminators
- Hot pluggable disk?
- Mirroring or RAID needed
- Position on bus not important

## SSA

- 2 adapters/host/loop or
- 1 adapter/host/loop and mirror across loops
- Position on loop affects performance
- Dual paths to disk with loop
- Bypass circuits keep loop intact

## DE SCSI adapters

Description	Feature Codes	Type Label	System Bus	SCSI Buses	Connectors
SCSI 2 DE High Performance External IO Controller	2420	4-2	MCA	1	Ext. high density SCSI 50
SCSI 2 DE Fast Wide Adapter/A	2413 2416 2417	4-6	MCA	2	Int. SE card edge 50 Int. SE high density 68 Ext. DE bus 68
Enhanced SCSI 2 DE Fast Wide Adapter/A	2412 2418 2419	4-C	MCA	2	Int. SE card edge 50 Int. SE high density 68 Ext. DE bus 68
SCSI 2 DE Fast Wide Adapter	2409 6209	4-B	PCI	1	Ext. high density micro D 68
Ultra SCSI DE Adapter	6207	4-L	PCI	1	Ext. high density micro D shell 68

Some hardware products support attaching multiple adapters in the same system to the same disk; e.g. SSA, the 7135 and the VSS (but not to two systems for the VSS). These products include software support for handling configuration of the same disk on multiple adapters and I/O routing through the appropriate adapter. Without this capability, to eliminate SPOFs, we have two options: mirror across separate adapters and turn LVM quorum off for the VG, or promote an adapter failure to a node failure so the standby node takes over the resource group. Using the second option leaves the SCSI bus as a SPOF. For SSA, an adapter failure (or break in the loop) doesn't prevent a host with another adapter on the loop from accessing the disk. SSA also has an advantage over SCSI in the SCSI does not support connecting/disconnecting/powering up/powering down devices on a SCSI bus while data is being transferred on the SCSI bus - though unlikely, it is possible to corrupt data being transferred on the bus during such an event (the exception being hot-pluggable SCSI disks). Another SSA advantage is the ability to put the hosts in separate buildings with the SSA fiber extenders.

Alternative disk subsystems include the 2105 SCSI VSS (though we either mirror across RAID 5 arrays or live with a SCSI bus as a SPOF), 7137 SCSI RAID array (typically configured with two 7137s in RAID 0 mode and mirrored across each), 7135 (using RAID 5 or RAID 1 with two separate SCSI attachments), 7131-105 SCSI (typically mirrored across separate units), or 7204 SCSI disks (typically mirrored across separate units). The SCSI disk subsystems, other than the 7135 typically use separate units on separate SCSI busses, mirroring across the units.

Usually differential ended (DE) SCSI is used due to the cable lengths involved, as opposed to single ended (SE) which requires shorter cable lengths, though SE is used with the 7027-HSC disk enclosures which fits in a rack, but total cable lengths are limited, so the hosts typically need to be in the same rack or an adjacent rack. DE SCSI also provides improved noise tolerance over SE.

While the disk is shown between the systems, for SCSI it doesn't matter where on the bus the systems or disk is, just that they are on the shared SCSI bus, so for SPs, generally the cabling goes from node to node then to the disk.

## SSA Adapters

FC	6214*	6216	6217*	6218*	6215	6219	6225
Type Label	4-D	4-G	4-I	4-J	4-N	4-M	4-P
Bus type	MCA	MCA	MCA	PCI	PCI	MCA	PCI
RAID support	none	none	RAID 5	RAID 5	RAID 5	RAID 5	RAID 5,0
Muti-attach JBOD	2-way	8-way	1-way	1-way	8-way	8-way	8-way
RAID 5	n/a	n/a	1-way	1-way	2-way	2-way	2-way
Intermix with	6214 6216	6216 6214	none	none	6215 6219 6225	6215 6219 6225	6215 6219 6225
Fast Write Cache**	none	none	none	none	4 MB	4 MB	32 MB
Read Cache	none	none	8 MB	8 MB	32 MB	32 MB	64 MB
IO/s Non-RAID	3000	3000	3000	3000	3000	3000	10,000
IO/s RAID 5 ***	n/a	n/a	3000	3000	3000	3000	4700
IO/s RAID 5 ****	n/a	n/a	1000	1000	1000	1000	2700
MB/s non-RAID	35 r/w	35 r/w	35 r/w	35 r/w	35 r/w	35 r/w	90 r/w
MB/s RAID 5	n/a	n/a	29r 7w	29r 7w	29r 13w	29r 13w	55 ****

\* *Withdrawn*

\*\* *Fast Write Cache is optional, not supported on multi-initiator loops*

\*\*\* *Cache hits (100% reads)*

\*\*\*\* *Non-cache hits (70%/30% r/w)*

r/w *Read / Write*

*Boot Support: 6214, 6216, 6217, 6219, 6225  
Certain RS/6000 models only (varies by adapter)*

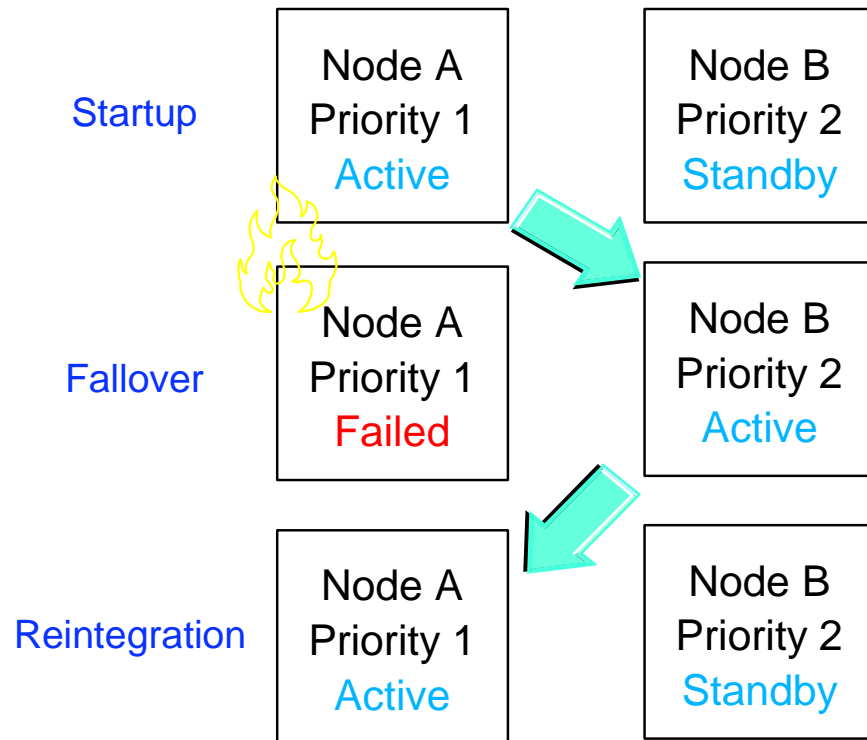
*AIX 4.1.4 + (varies by adapter)*

*No booting from RAID arrays*

*Target Mode SSA - 6215, 6219 & 6225 only*

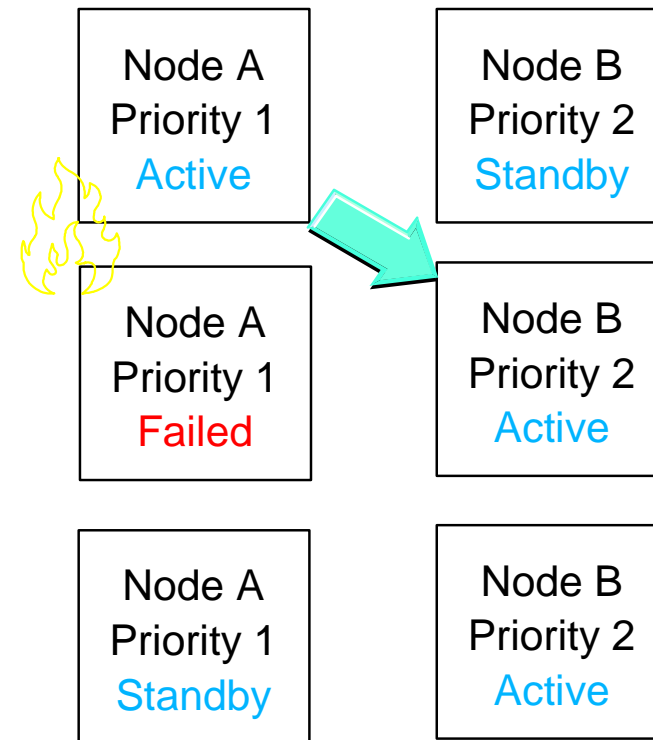
# Resource groups

## Cascading resource groups



- Good for dissimilar systems
- Disruption at reintegration
- Reintegration can be planned  
No backup node until then
- IPAT to standby adapters

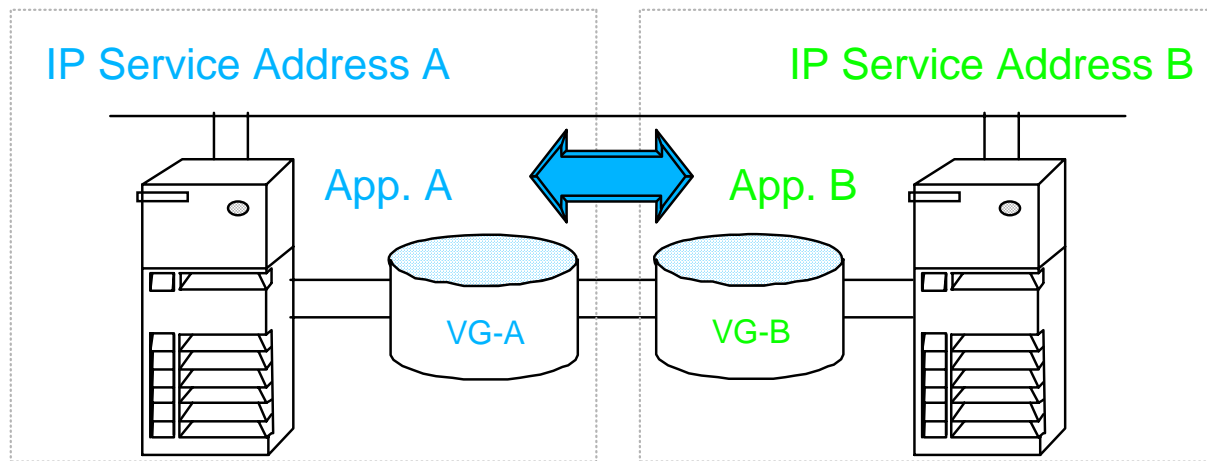
## Rotating resource groups



- No disruption at reintegration
- Priority chain can be changed with dynamic reconfiguration feature
- IPAT to boot/service adapters

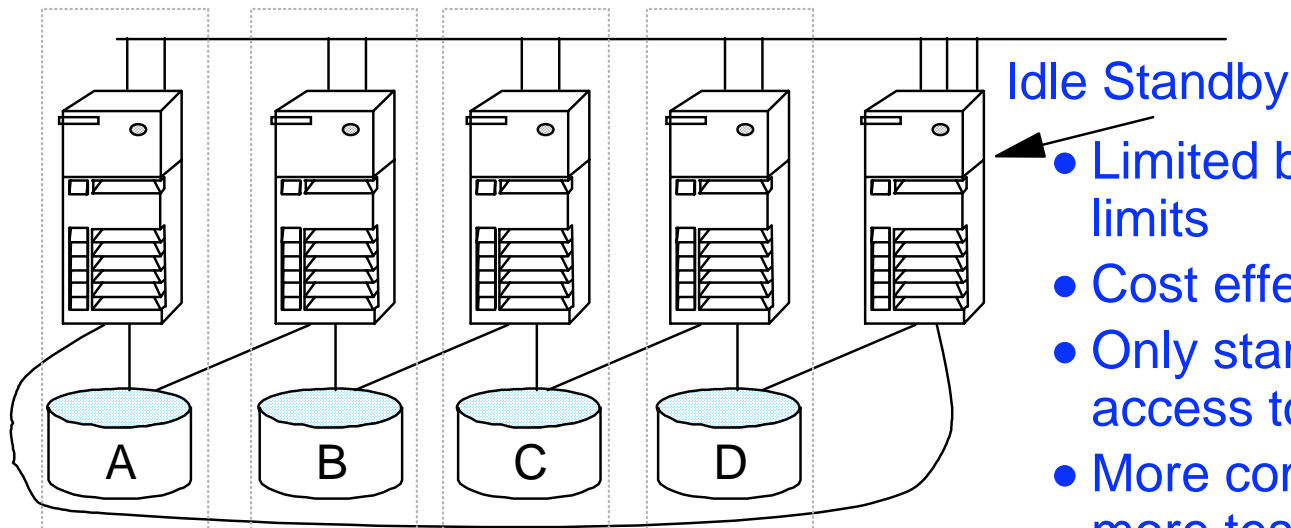
## Other clusters

### Mutual takeover



- Two workloads
- Beware of dead man switch

### N+1 clusters



- Limited by disk connection limits
- Cost effective
- Only standby nodes need access to the disk
- More complicated scripts and more testing needed

Rotating and cascading clusters behave differently at failover and at reintegration. On a cluster with a rotating resource group, the highly available IP address will move to either a boot adapter or a service adapter which is not part of a resource group on the standby node. Cascading clusters move the highly available IP address to a standby adapter. At reintegration (i.e. when HACMP is started on the failed server) cascading resource groups move to the node with the highest priority while rotating resource groups stay where they are. For configuring systems, a minimum of two adapters per physical network (on which a highly available IP address will exist) should exist to distinguish between network and network adapter failures (i.e. a boot/service and standby adapter). For systems which may control multiple resource groups, configure enough standby adapters for cascading resource groups and enough boot/service (a service adapter which is not part of a resource group) adapters for rotating resource groups. An exception is the SP switch which uses IP aliasing and for which only one adapter is supported per node (more on the SP later).

Mutual takeover clusters allow both machines to be used for production; however, when one machine is running both workloads, one risks a system crash from the dead man switch if the CPU becomes overloaded. Many customers use the standby system for development work, and when it takes over the production workload, development work is stopped until the primary server is repaired.

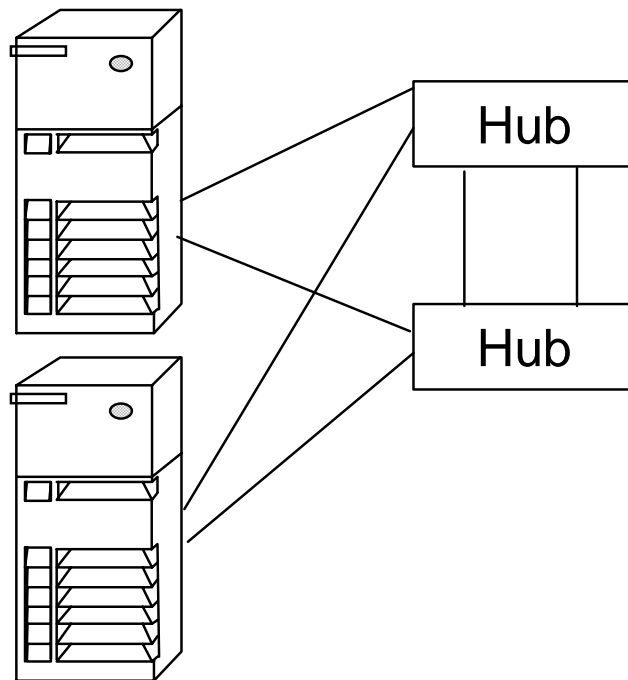
N+1 clusters use the machine hardware in a cost effective manner, with a N:1 versus a 1:1 ratio of servers to standby systems. The size of the cluster is constrained by disk connection limitations: currently 8 nodes can share SSA disk. Shown is a 4+1 cluster with disks connected to all nodes. One can construct a larger cluster with one hot standby sharing disk with each server (not shown) or by using highly available disk provided outside the cluster (e.g. GPFS or other applications using RVSD). Two downsides of larger clusters are the additional testing necessary, and the greater probability of two servers failing prior to fixing the first failure.

Note that for non-IP heartbeat paths in multi-host clusters, HACMP requires only heartbeats to neighboring nodes; thus, for RS-232 two ports/node are needed (ports used for accessing the service processor are not supported for HACMP)

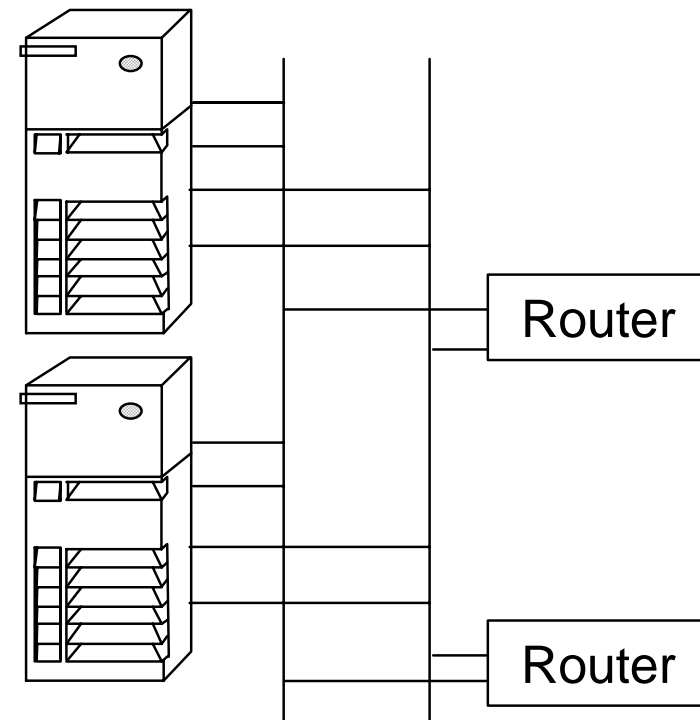


# Network availability

## Dual homing

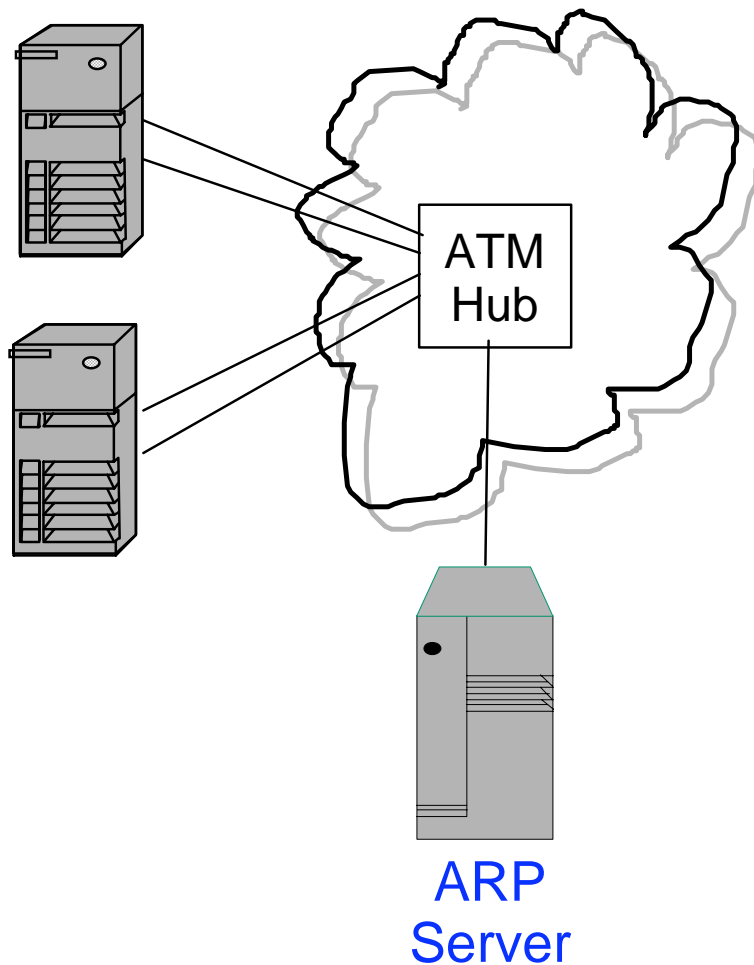


## Dual networks



- Eliminates hubs as SPOF
- Use dual homing and dual networks in a cluster backbone with intelligent routers to provide network availability
- Where are clients attached?
- Routing is trickier

# FDDI and ATM



- **FDDI**
  - Dual Attach System adapters (DAS)
    - Two slots
    - Two cables to HUB(s)
  - Single Attach System adapter (SAS)
    - One cable
  - Both have one IP address
  - SAS adapters recommended due to slot usage
- **ATM**
  - ARP server could be a SPOF
  - Broadcast not supported
  - clinfo updates ARP cache by pinging
  - LAN emulation or classical IP supported

## HACMP on the SP

- IPAT on the switch

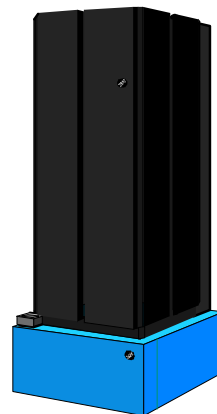
- IP addresses aliased on interface
  - Base IP address not used by HACMP
  - Nodes must be in the same partition
  - Switch adapter failure promoted to node failure
  - SP handles Eprimary failure

- Administrative ethernet

- Used for heartbeats only
  - Switched ethernet can be used

- HACWS

- Do you need it?
  - Backup CWS cannot:
    - Update passwords
    - Add/Change users



- VSDs and RVSDs

- Provide scaling and availability for disk
  - Up to 128 nodes
  - Oracle Parallel Server and GPFS only
  - RVSD servers generally outside cluster
  - RVSDs work like disk takeover

- HACMP ES

- Kerberos or .rhosts
  - PSSP provides event detection and heartbeats for HACMP ES 4.2.2
  - RSCT provides event detection and heartbeats for HACMP ES 4.3
  - Scaleable
  - Up to 32 nodes in a cluster
  - Works differently than HACMP classic

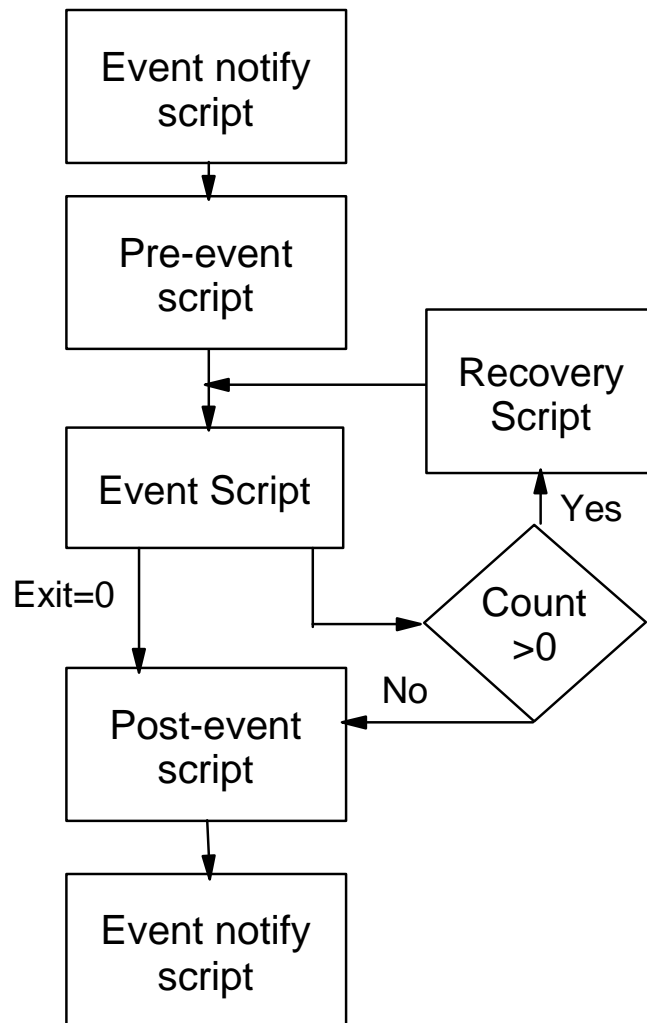
The SP has unique characteristics affecting HACMP, and software products tailored to it. The SP only allows one connection to the switch per node. So IP aliasing has been implemented and is used for making IP addresses highly available. The base switch IP address is not used by HACMP and remains on the adapter. But then we make additional IP addresses available, simultaneously, on the adapter (this is IP aliasing - providing more than one IP address on an adapter). The adapter itself remains a SPOF, but even if the switch network is not used for a highly available IP address, the network is used for a heartbeat path.

The administrative ethernet adapters are only used as a heartbeat path. Generally clients access a highly available IP address via another network. If multiple frames are involved, and routers are used between the administrative ethernets on each frame (or sets of frames), HACMP ES 4.3 has a global network feature that allows heartbeats to flow between these different LAN segments.

HACWS provides a highly available control workstation (CWS); however, the backup CWS is not functionally equivalent to the original CWS due to Kerberos limitations. One cannot manage Kerberos keys, add SP users, change user configurations or passwords. Shared disk is required for /spdata, and a special RS232 Y cable is needed to attach the CWS's to each frame. Environments that need HACWS might include those requiring Kerberos tickets or submitting parallel jobs.

RVSDs provide highly available disk independently from HACMP. Currently, Oracle Parallel Server (OPS) and GPFS are the only applications using this facility.

# Events and notification



- All but event script are user written
- 30+ predefined events/subevents
  - node\_up
  - node\_down
  - node\_down\_local
  - network\_down
  - start\_server
  - swap\_adapter
  - get\_disk\_vg\_fs
  - acquire\_service\_addr
- Parameters passed to event scripts
- AIX error notification setup via smit
  - Run once per error entry in error log
  - You write the script
  - List all possible errors with `# errpt -t`

# Cluster management

- **Cluster single point of control - CSPOC**
  - Add/delete/change users/groups/LVs/VGs for cluster from one node
  - Start & stop HACMP on all nodes
- **Dynamic reconfiguration**
  - Add, delete or modify (during cluster operation)
    - nodes, networks, network adapters, scripts and resource groups
  - Changes take place automatically and resource groups might move
- **HA View**
  - Monitor cluster from NetView
- **clinfo daemon and clstat**
  - clstat shows cluster status
  - Source provided to run on clients
  - Clients can react to cluster changes
- **User files and passwords**
  - rdist
  - NIS
  - Supper and AMD on SP

## HACMP flavors and versions

- **HACMP classic (HAS)**  
4.3 and 4.2.2 available at AIX 4.3.2 and 4.1.5 respectively
- **HACMP ES 4.2.1**  
PSSP provides event detection and requires AIX 4.2.1 and PSSP 2.3
- **HACMP ES 4.3**  
RSCT provides event detection and requires AIX 4.3.2 and PSSP 3.1 for SP nodes
- **HANFS**  
HACMP for NFS 2 node cluster only  
Maintains NFS locks at fallover  
Clients maintain access during fallover with a delay
- **Concurrent access**  
Scales CPU - up to 8 systems access shared disk concurrently  
Oracle Parallel Server on LAN (not SP)  
Cluster lock manager API  
Clients connect to any running node
- **HAGEO**  
Mirror disk between sites  
High bandwidth connection required  
Requires HACMP
- **Hardware support varies by release and AIX level**  
See <http://hacmp.aix.dfw.ibm.com> and announcement letters

The five flavors of HACMP provide different features for different environments. The classic HACMP (also called HAS) is available at three different versions, and HACMP ES is available in two different versions. The direction HACMP is taking, is towards using a unified cluster event detection using RSCT. In moving towards this goal, we've had three sets of cluster management processes: classic's clstrmgr, PSSP's daemons, and now RSCT. HACMP classic still uses the clstrmgr daemon, as does HANFS and concurrent access (CRM). ES has moved from the PSSP daemons to RSCT in the two versions available. Support for various versions of these flavors have their own AIX and possibly PSSP requirements. Generally, the newer versions have support for the latest features and hardware. Generally, (except for HACMP ES 4.2.2), the HAS, ES, CRM, and HANFS options are orderable as a feature code. Clusters must all run the same level and flavor of HACMP (except temporarily when upgrading from a prior release).

Concurrent access uses concurrent LVM, such that a volume group is active on up to 8 systems; consequentially, to preserve data integrity, applications writing to the disk must communicate to the other systems and get a cluster wide lock - this requires writing an application using a lock manager API. Only Oracle Parallel Server (OPS) has been written using this API. This is different than OPS on the SP which uses VSDs to provide cluster wide disk I/O services, and uses an Oracle written lock manager.

HANFS provides a way to make an NFS server highly available, at a lower cost than using HACMP. In addition, it preserves locks at fallover, and due to it's stateless protocol, clients only see a delay at fallover. It doesn't have all of classic HACMP's bells and whistles though, such as snapshots, dynamic reconfiguration, CSPOC, and others; consequentially is only appropriate for NFS servers.

HAGEO provides the ability to mirror disk across sites; consequentially requires a high bandwidth link between them (generally greater than a T3 link or multiple T3 links). Implementation takes typically two months. Alternatives include application data replication, or remotely connected disk (e.g., SSA fiber).

HACMP ES also provides the ability to react to other types of events, and can enhance availability due to application failures or other conditions (e.g. CPU utilization high or a nearly full file system ), but this requires additional programming.



## HACMP flavors and features

	HACMP ES 4.2.2	HACMP ES 4.3	HAS (classic)	HANFS	CRM
Systems support	SP only	SP &/or RS6000	SP &/or RS6000	SP &/or RS6000	SP &/or RS6000
AIX pre-req.	4.2.1	4.3.2	Varies by release	Varies by release	Varies by release
PSSP pre-req.	2.3	3.1	N/A <sup>1</sup>	N/A <sup>1</sup>	N/A <sup>1</sup>
Dynamic reconfiguration	No	4.3 +	4.2 +	No	4.2 +
Event emulation	Yes	Yes	4.2.2 +	No	4.2.2 +
CSPOC	Yes	Yes	4.2 +	Yes	Yes
Global networks	No	Yes	No	No	No
FDDI MAC takeover	No	Yes	4.3 +	4.3 +	4.3 +
Target mode SSA	No	Yes - IX85491	4.2.2 +	4.2.2 +	4.2.2 +
Protocol over serial networks	No	No	Yes	Yes	Yes
Max. systems per cluster	16	32	8	2	8
ATM LANE <sup>2</sup>	No	Yes	4.3 +	4.3	4.3
Kerberos support <sup>3</sup>	Yes	Yes	4.2.1 +	4.2.2 +	4.2.1 +

1. HACMP use is independent of PSSP for these features
2. ATM ARP server is a SPOF
3. Kerberos is not provided for non-SP environments

This foil shows the various features and AIX/PSSP prerequisites for the five flavors of HACMP.

Dynamic reconfiguration provides the ability to move resource groups on the fly without stopping HACMP on a node. Event emulation provides the ability to emulate events (as far as the event doesn't depend upon the results of a command). CSPOC was introduced at HACMP 4.2 and has been enhanced at 4.3 to ensure UIDs and GIDs are unique cluster wide when adding users or groups, allows extendvg, reducevg, mklv, rmlv, splitlvcopy, mirrorvg, unmirrorvg and chlv to work in cluster wide and in concurrent mode, and has a faster syncvg for concurrent mode.

The global network feature is targeted for SPs with multiple administrative ethernet segments with multiple frames, such that failure of the other network(s) connecting the nodes will result in network down events rather than partitioned clusters.

A note on positioning HAS vs. ES: Target mode SSA, target mode SCSI, and RS232 connections between machines are HACMP "serial" networks providing heartbeats, but not "protocol" messages in which the ES version daemons vote amongst themselves to agree on what's happening in the cluster. Customers which have user written network down post event scripts (e.g. to stop the application until the network is up) and only have one IP network, should use HAS. Customers with multiple IP networks or no post event scripts for network down events should use ES. Migration between versions with different underlying cluster technology (RSCT, PSSP, or clstrmgr) is more disruptive has has a higher risk of an outage than migration between versions with the same underlying cluster technology.

IPAT is supported for both classical IP and LAN emulation in ATM environments. The ATM ARP server will remain as a SPOF. Also, while two ethernet or token ring adapters can be emulated on a single ATM adapter, a minimum of two ATM adapters should be ordered per system for HACMP to work correctly. Finally, while Kerberos is provided by PSSP, it is not provided for standalone RS/6000s by IBM, and while use of Kerberos is an option, /.rhosts can be used instead.

## Implementation & skills

### Skills

- HACMP
- Application
- Shell script
  - Most of HACMP is written using scripts
  - Application start/stop scripts
  - Failure notification scripts
- Network
- LVM

### Implementation

- Planning session:
  - DBA, network, system and HACMP
  - Run by HACMP skilled person
- What hardware?
  - New
  - Rent/lease/borrow
  - Production system
    - Large testing windows needed
- Install HW, AIX, HACMP
  - Install application
  - Write application start/stop scripts
  - Setup notification
  - TEST** failover for all failure types
  - Recover from all failure types
  - Test backup and restore
  - Document recovery/mgmt. processes
- Two weeks or more

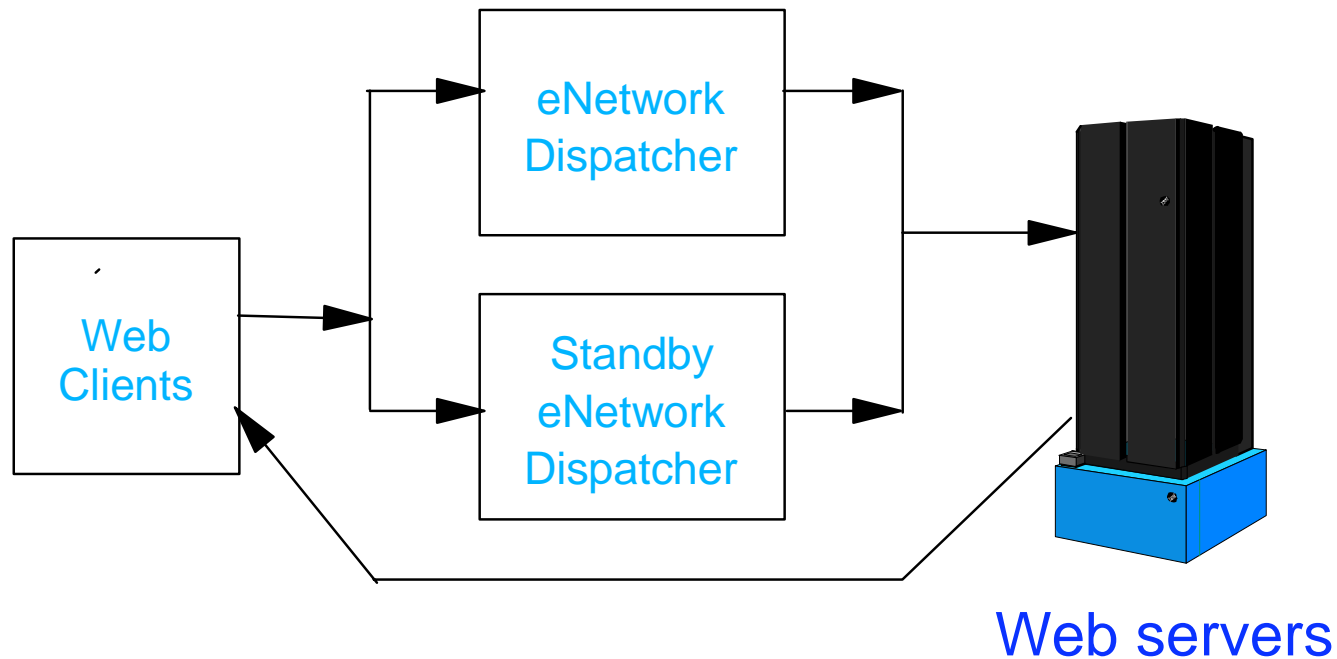
Implementing HACMP typically involves more than one person: the system administrators, a skilled HACMP resource, the application specialists/administrators, and the network administrators. Everyone needs to understand how HACMP works. Change control is more important than for standard RS/6000 implementations. If the network administrator isn't aware of what's happening, he may make changes to the network which stop HACMP from working correctly. Performance planning and monitoring are also important for preventing dead man switch crashes. Shell scripts need to be written to start and stop the application(s).

A typical implementation starts with ensuring that everyone knows how the cluster works, and is typically covered in a planning session run by a skilled HACMP resource. Then the parties work together to setup and test the cluster. Failures are initiated under a load (pulling network cables, powering down systems, pulling disk drives, etc.) and one checks that notification and failover work as planned. Then recovery is performed (and the recovery procedure is documented) so that the cluster is returned to a fully redundant state. Backup/restore of the systems and data should also be tested.

Implementation on a system in production is usually not recommended, as testing windows are needed, typically a minimum of 4 hours, which will disrupt production. A better approach is to setup on additional equipment and test until it works correctly, then use a cutover approach. A cutover might entail hooking up the production disk to the cluster, or using a backup/restore of the systems and/or data, or changing the IP addresses used on the production and cutover systems.

The time to implement HACMP is typically two weeks, but more complicated clusters and applications, or multiple applications that interact with each other will take longer. The length of time is also dependent upon the skills of the parties involved.

# SecureWay Network Dispatcher



- Inbound traffic goes through dispatcher nodes
- Outbound traffic goes directly to client
- Uses it's own high availability code
- Multiple web servers offer  
Availability  
Load balancing
- Works with eNetwork Firewall

## Site disaster recovery

- Remotely attached disk
  - SSA fiber extenders (up to 10 km) with HACMP
- Application data replication
  - Keep an up to date copy at a remote system over WAN
  - Function shipping
- HAGEO
  - Mirrors LVs across systems over WAN
  - I/O shipping
  - Asynchronous mode - 1 system/site
  - Synchronous mode - 1 cluster/site
- Periodic copies of local data to remote disk over WAN
- Keep copies of backup tapes at remote site

The SecureWay eNetwork Dispatcher product is like some other products in that it has its own high availability function written into the application, and in such situations, using the application availability code is usually preferable to using HACMP. This product also provides availability for read-only web servers. If the web servers write to a database, the database could be made highly available outside the web server cluster using HACMP.

Site disaster recovery options include using remotely attached disk: each site containing half the disk with mirroring between sites. With SSA fiber, the disk appears to be locally attached. In case the primary site fails, the hot standby node takes over the application using HACMP. Since HACMP requires one physical network for IPAT, LAN bridges (not routers) may be needed between sites.

Many applications, primarily databases, provide a replication feature whereby data is kept consistent between sites. This uses "function shipping" in that the SQL transaction is sent across the WAN. HAGEO uses "I/O shipping" where writes to a geo-mirrored LV are sent across the WAN - usually requiring more bandwidth between sites than function shipping requires. These slow down the application due to latencies and increased overhead (sending the I/O or update via TCP/IP to the remote site and waiting for the completion acknowledgement).

HAGEO offers asynchronous and synchronous options. The synchronous option ensures both local and remote writes are complete before returning control to the application, the asynchronous option only waits for the local write to complete. The asynchronous option only uses one server/site. The synchronous option allows a cluster at each site, though the remote site can be a single system. If you use the asynchronous option, at site failover, the remote site will be missing transactions (make sure this is acceptable before using it). HAGEO implementations take approximately two months.

- **HACMP Manuals**

Concepts & Facilities, Planning Guide, Installation Guide, Troubleshooting Guide, HANFS for AIX, Programming Client Applications, Group Services Programming Guide, Event Management Programming Guide, Programming Locking Applications, Master Index

- **Redbooks**

High Availability on the RS/6000 Family	SG24-4551
Implementing High Availability on RISC/6000 SP	SG24-4742
RS/6000 SP High Availability Infrastructure	SG24-4338
HACMP Enhanced Scalability	SG24-2081
HACMP/6000 Customization Examples	SG24-4498
High Availability Strategies for AIX	GG24-3684
HACMP/6000 Mode 3 Implementation	GG24-3685
An HACMP Cookbook	SG24-4553

- **Web sites**

[www.rs6000.ibm.com/software/Apps/hacmp](http://www.rs6000.ibm.com/software/Apps/hacmp)  
[hacmp.aix.dfw.ibm.com](http://hacmp.aix.dfw.ibm.com) (IBM intranet)  
[www.pok.ibm.com/afs/aix/www/pps/cgi-bin/bobgens/phoenix/phoenix\\_indepth.html](http://www.pok.ibm.com/afs/aix/www/pps/cgi-bin/bobgens/phoenix/phoenix_indepth.html)  
[dscrs6k.aix.dfw.ibm.com](http://dscrs6k.aix.dfw.ibm.com) (IBM intranet, choose Hints & Tips)

- **Consultant's report**

[w3.rs6000.ibm.com/mktmat/dl/rshacmp.html](http://w3.rs6000.ibm.com/mktmat/dl/rshacmp.html) - HACMP ranked #1 by DH Brown

- **In Search of Clusters - Gregory Pfister SR23-8294**



# Hardware support



as of 5/14/99

Hardware Software	Announce Letters	HACMP 4.2.2 AIX 4.1.5	HACMP 4.2.2 AIX 4.2.1	HACMP 4.2.2 AIX 4.3.2	HAMCP 4.3 AIX 4.3.2
VSS 2105-B09 2105-100	198-129	IX81408 IX69008	IX81408 IX63204	IX81408	Yes
7133-D40/T40	198-280	Yes	Yes	Yes	Yes
7013-S7A 7015-S7A 7017-S7A	198-247 198-248 198-238	No	No	Yes	Yes
43P 7043-150 7043-260	198-240 198-239	No	Yes	Yes	Yes
9076-50H 9076-55H	198-246	No	No	PSSP 3.1	PSSP 3.1
Gigabit Ethernet FC 2969	198-243	No	No	Yes	Yes
PCI Token Ring FC 2920	198-092	Yes	Yes	Yes	Yes
8 Port Async. FC 2943	197-276	No	Yes	Yes	Yes
128 Port Async. FC 2944	197-276	No	Yes	Yes	Yes
10/100 PCI Enet FC 2968	197-276	Yes	Yes	Yes	Yes