IBM Power Systems

Technical University 2011

# Introduction to POWERHA SystemMirror for AIX Standard Edition



**Shawn Bodily**

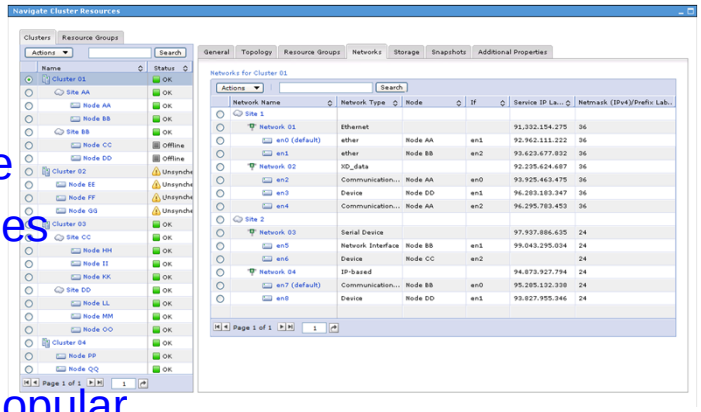
**Certified Consulting IT Specialist**

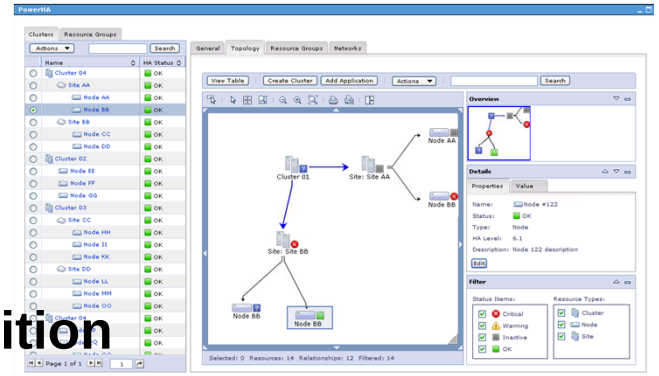
**sbodily@us.ibm.com**

# PowerHA SystemMirror Editions



- **PowerHA SystemMirror for AIX Standard Edition**
- Cluster management for the **data center**
  - Monitors, detects and reacts to events
  - Establishes a heartbeat between the systems
  - Enables automatic switch-over
- IBM shared storage clustering
  - Can enable near-continuous application service
  - Minimize impact of planned & unplanned outages
  - Ease of use for HA operations
- Smart Assists – application agents
  - Out of the box deployment for SAP and other popular applications
- Mature Product
  - 23 Major releases (averaging one a year)
  - Over 12,000 customers worldwide



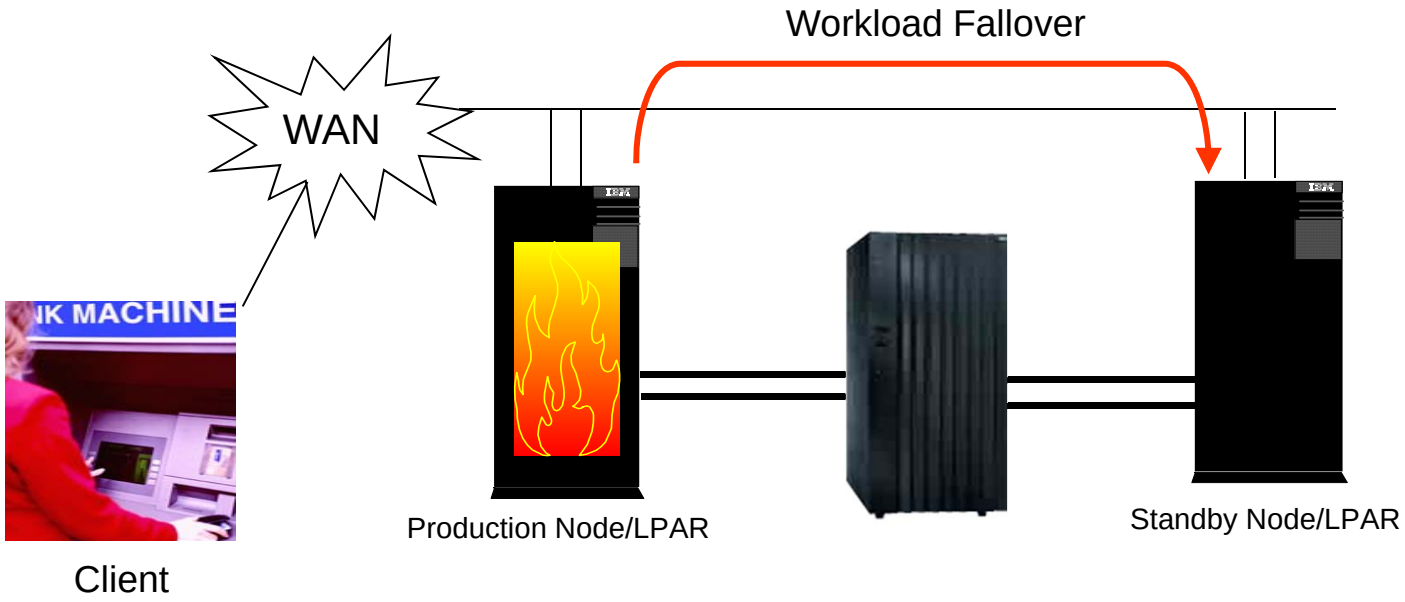


- **PowerHA SystemMirror for AIX Enterprise Edition**
- Cluster management for the **Enterprise**
  - Multi-site cluster management
  - Includes the Standard Edition function

# What is high availability?



- High availability characteristics:
- The reduction or elimination of downtime
- Solution may address planned  or unplanned down time
- Solution need not be fault tolerant but should be fault resistant
- **Solution should eliminate single points of failure (SPOF)**



# Eliminating single points of failure



| Cluster object | Eliminated as a single point of failure by: |
|---|---|
| Node | Using multiple nodes |
| Power source | Using multiple circuits or uninterruptible power supplies |
| Network adapter<br>Network | Using redundant network adapters<br>Using multiple networks to connect nodes |
| TCP/IP subsystem | Using non-IP networks to connect adjoining nodes and clients |
| Disk adapter<br>Disk | Using redundant disk adapter or multipath hardware<br>Using multiple disks with mirroring or raid |
| Application | Adding node for takeover; configuring application monitor |
| VIO server | Implementing dual VIO servers |
| Site | Adding an additional site |

**The fundamental goal of successful cluster design is the elimination of single points of failure (SPOF).**

# PowerHA and LPM Feature Comparison



| | PowerHA | LPM |
|---|---|---|
| Live OS/App move between physical frames* | | ✓ |
| Server Workload Management** | | ✓ |
| Energy Management** | | ✓ |
| Hardware Maintenance | ✓ | ✓ |
| Software Maintenance | ✓ | |
| Automated failover upon System Failure (OS or HW) | ✓ | |
| Automated failover upon HW failure | ✓ | |
| Automated failover upon App failure | ✓ | |
| Automated failover upon vg access loss | ✓ | |
| Automated failover upon any specified AIX error (via customized error notification of error report entry) | ✓ | |

*~ 2 seconds of total interruption time

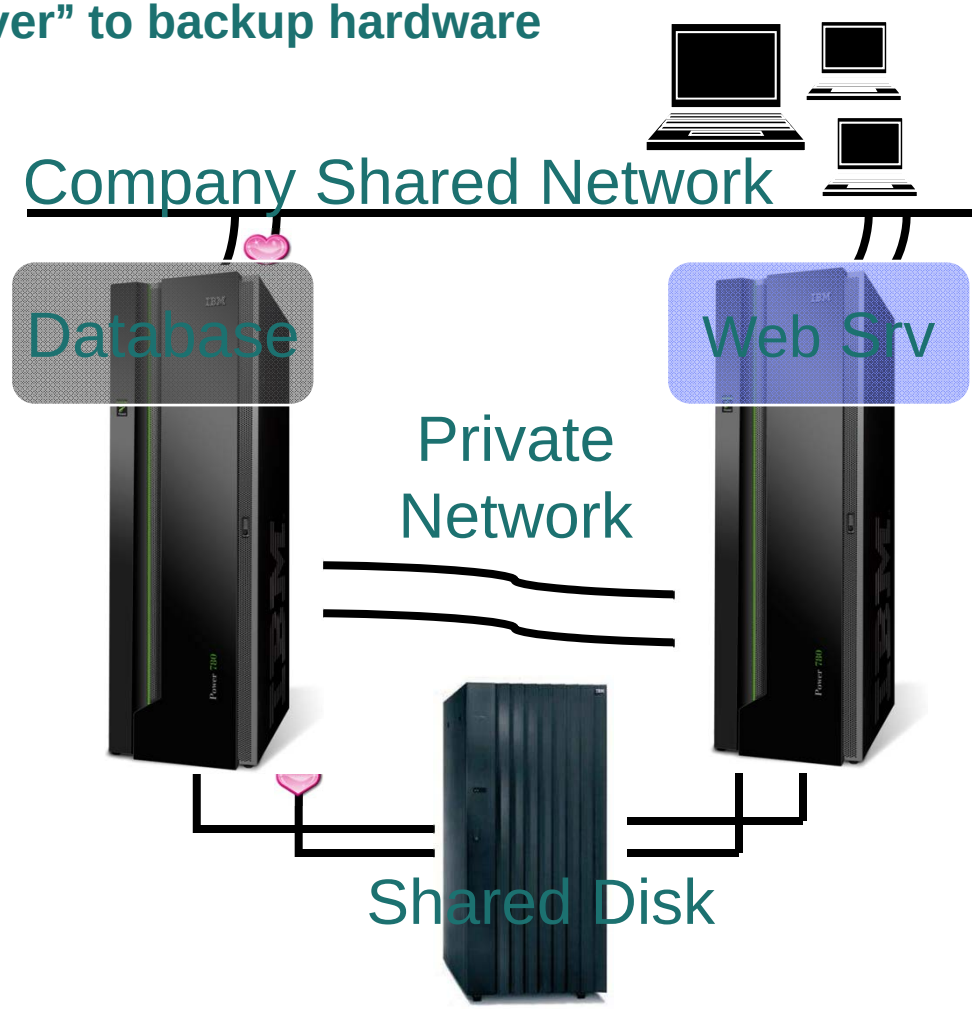
** Require free system resources on target system

# Services Outages



**PowerHA™ protects against service outages by detecting problems and quickly "failing over" to backup hardware**

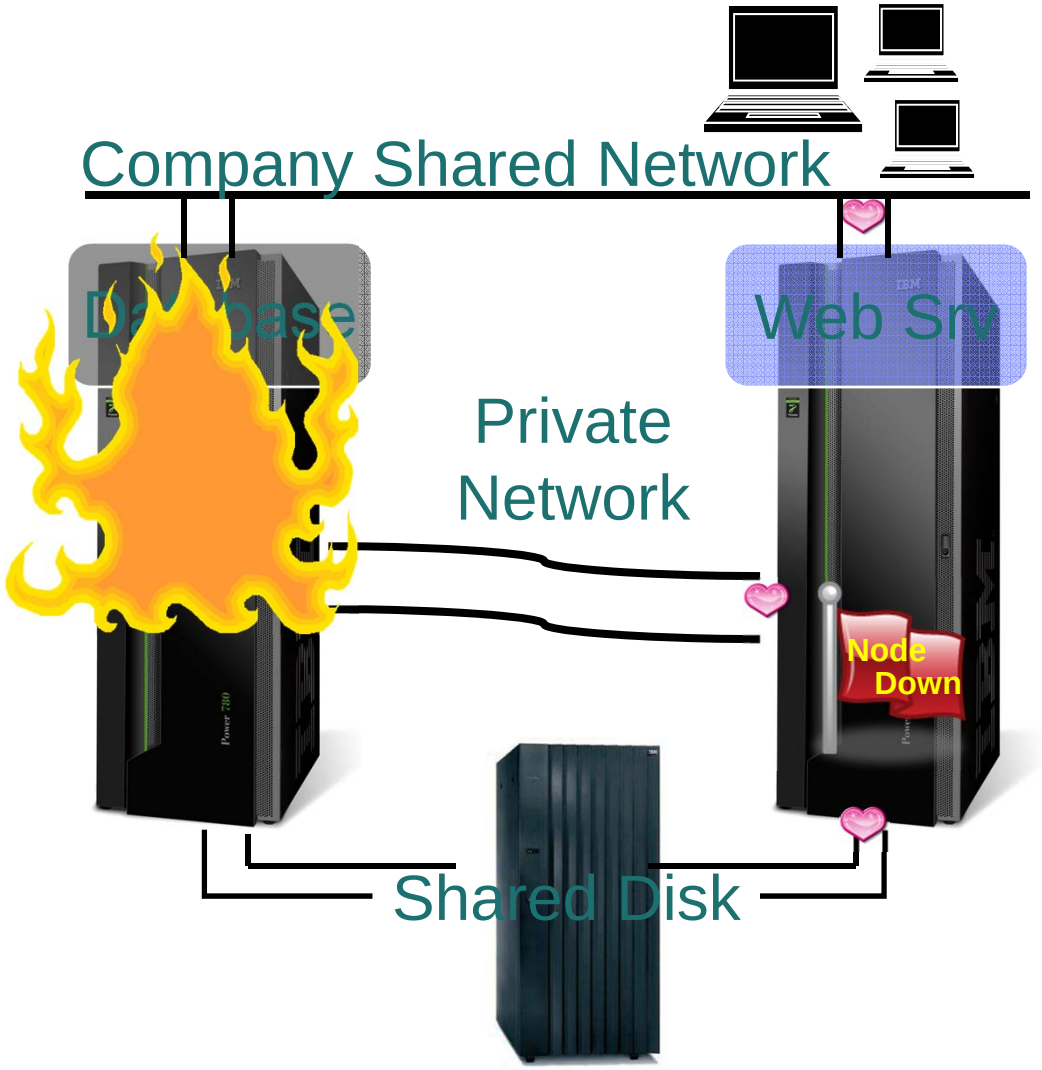
- Two nodes (A and B)
- Two networks
  - Private (internal) network
  - Public (shared) network
- Shared disk
  - All data in shared storage available to both nodes
- Critical applications
  - Database server
  - Web server
    - Dependent on DB



# Example Failure #1: Node failure



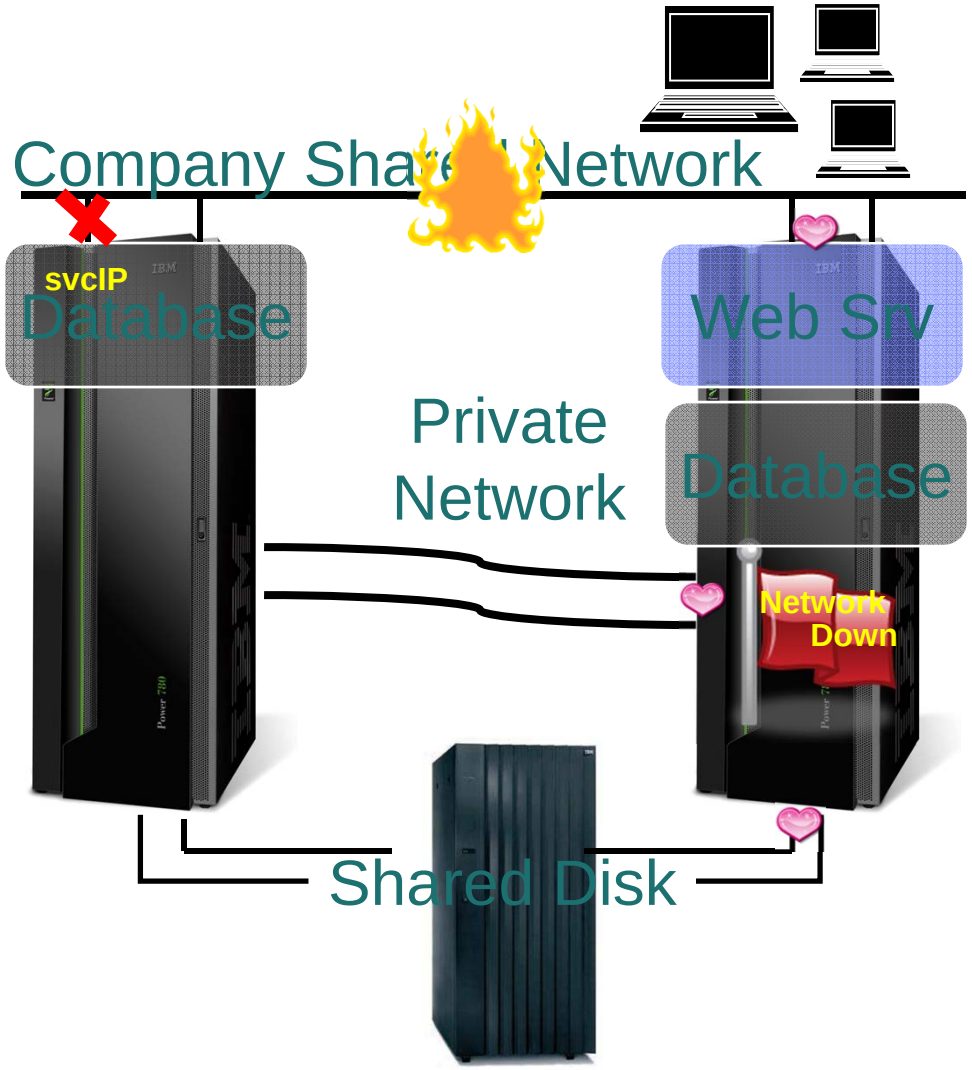
- Node A fails completely
- Node B detects the loss of Node A
- Node B starts up its own instance of the Database.
- Database is temporarily taken-over by Node B until Node A is brought back online
- This includes rootvg loss



# Example Failure #2: Loss of network connection



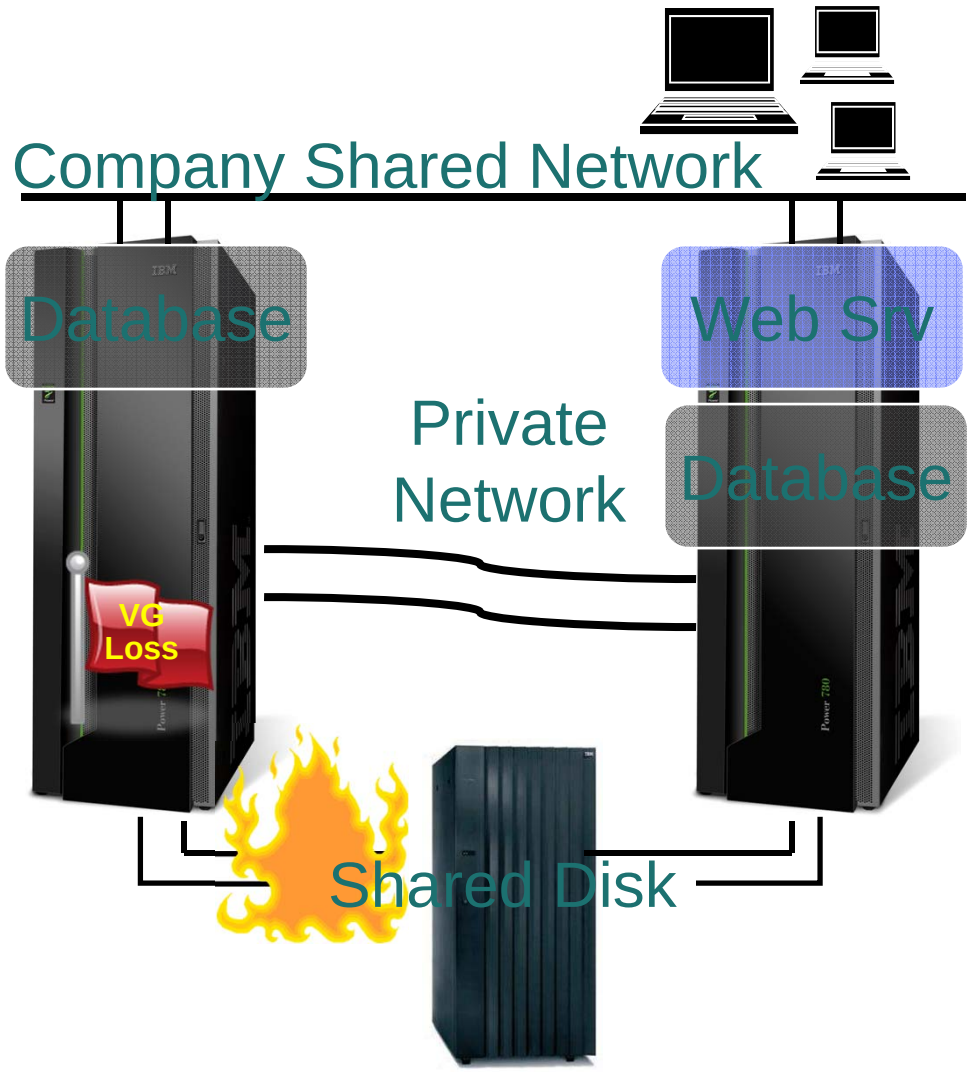
- Node A loses a NIC
- Because of NIC redundancy, the service IP swaps locally
- Operations continue normally while problem is resolved
- If total public network connectivity was lost a fallover could occur



# Example Failure #3: Loss of shared storage access



- Node A loses access to shared data storage
- After missing I/Os, AIX marks disks missing
- AIX error report logs "LVM_SA_QUOR_CLOSE" on shared data volume group
- PowerHA automatic AIX error notifications traps on that error and performs a resource group move



# Rootvg System Event

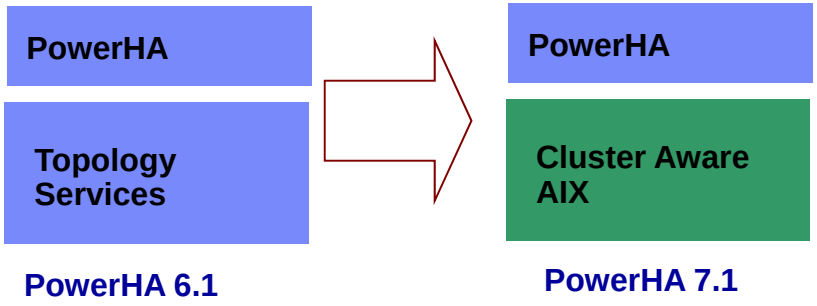


- New kernel level monitoring (7.1)
- Monitors the loss of rootvg
- Defaults response is to log event and reboot causing fallover to occur
- Smitty sysmirror->Custom Cluster Configuration->Events->System Events

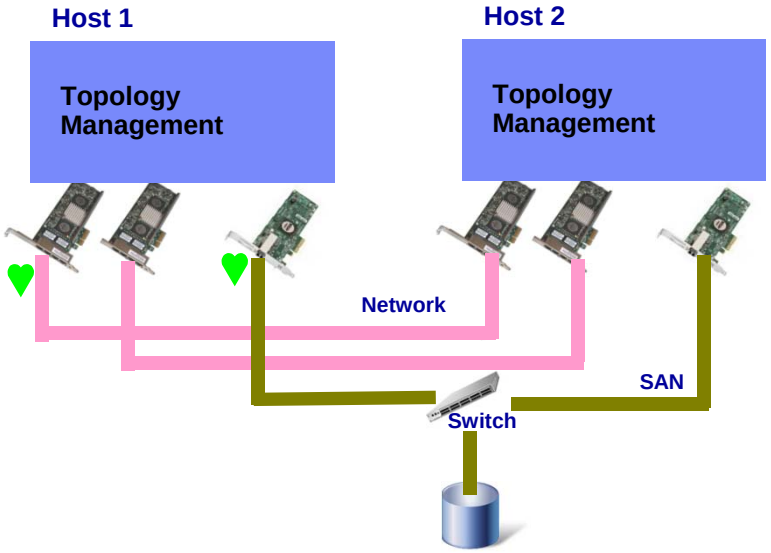
```
                        Change/Show Event Response

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                  [Entry Fields]
* Event Name                                     ROOTVG     +
* Response                                       Log event and reboot  +
* Active                                         Yes        +
```

# Cluster Aware AIX: Topology Management (1 of 2)

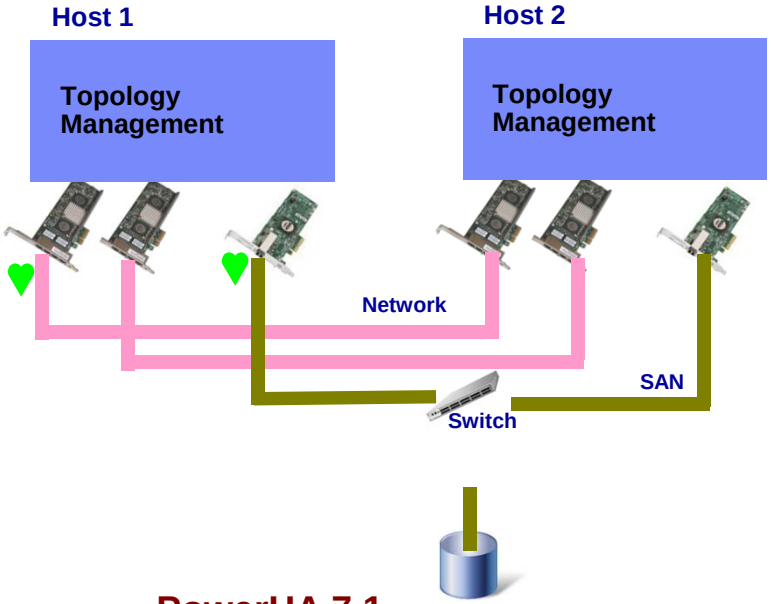





- By default mutli-channel based Topology management
- Use of high speed SAN links for cluster communication
- Kernel based health management



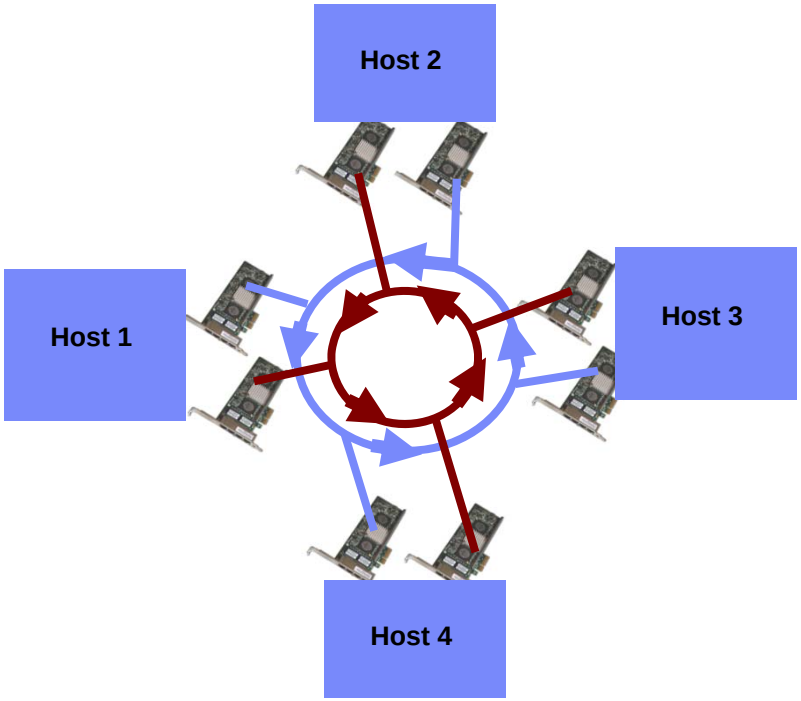

PowerHA 6.1

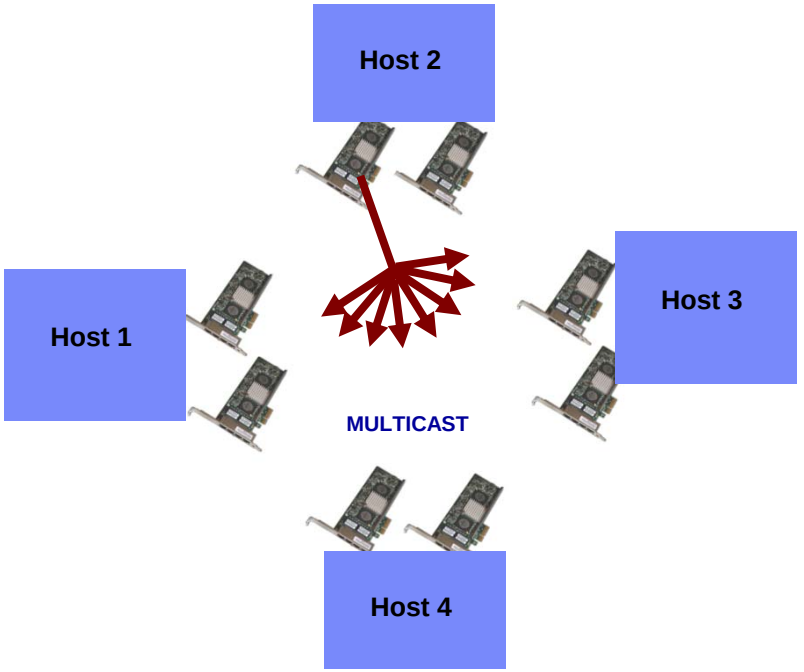



PowerHA 7.1

# Cluster Aware AIX: Topology Management (2 of 2)






## PowerHA 6.1

- Heartbeat Rings: detailed protocol
  - Leader, Successor, Mayor etc
  - Difficult to add/delete nodes
- Requires IP aliases management in the subnet

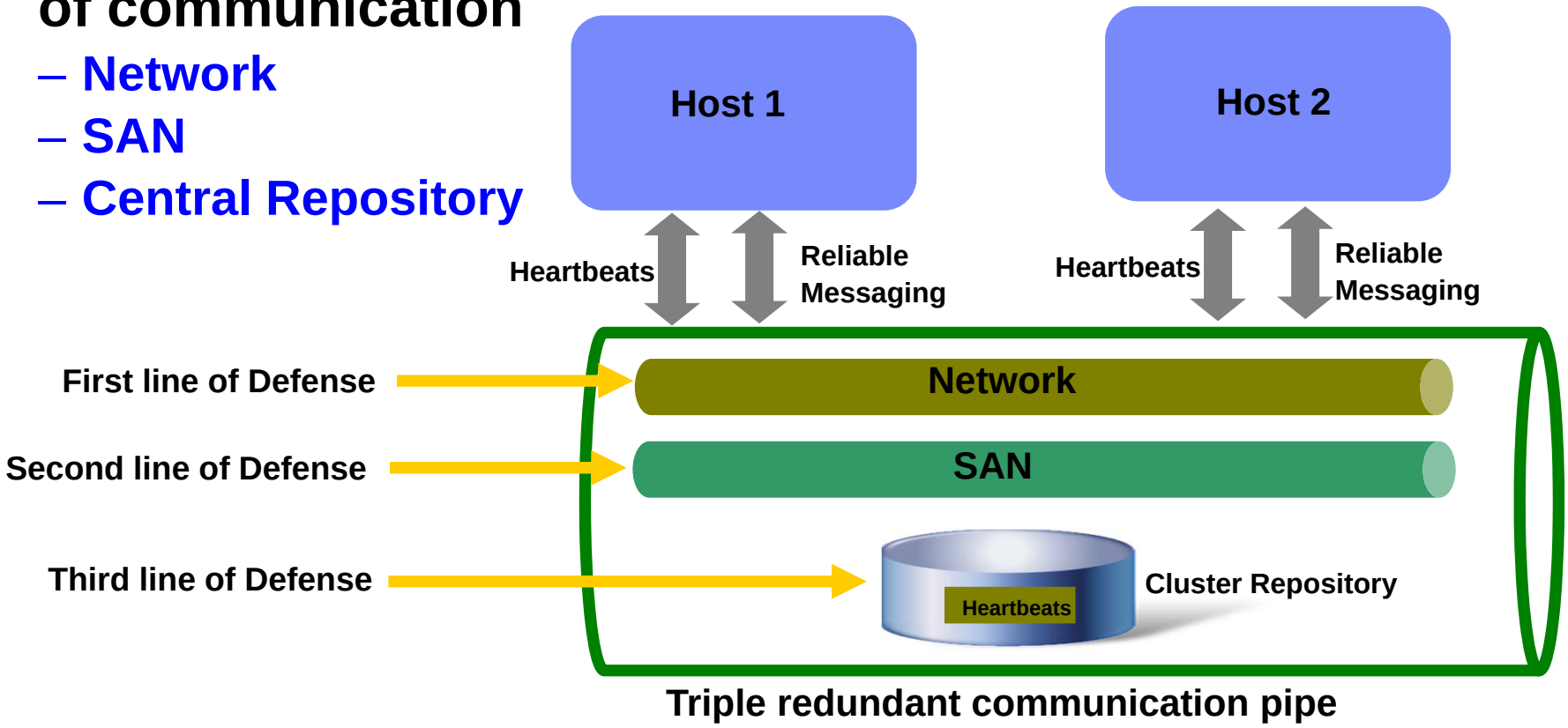
## PowerHA 7.1

- Multicast based protocol
  - Discover and use as many adapters as possible
  - Use network and SAN as needed
  - Adapt to the environment: delay, subnet etc
- Kernel based cluster message handling

# Default Multi Channel Health Management

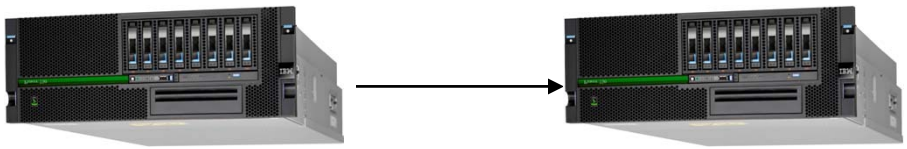


- **Minimal Setup**
- **Multiple channels of communication**
  - **Network**
  - **SAN**
  - **Central Repository**

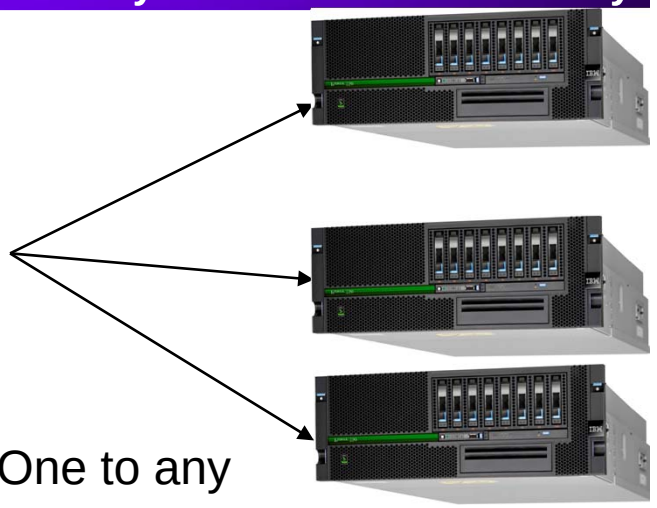


# Failover possibilities

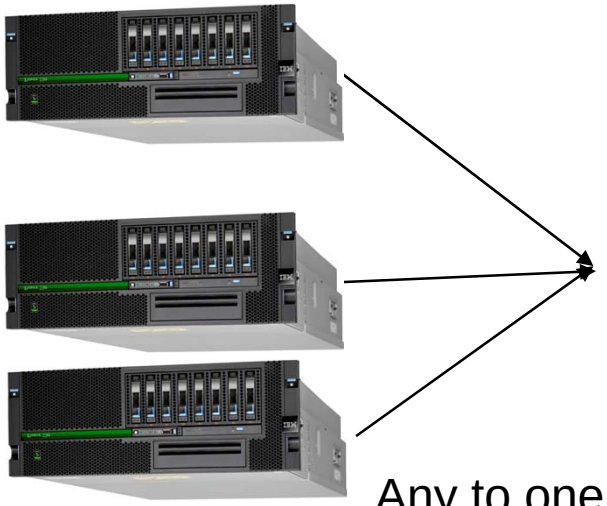




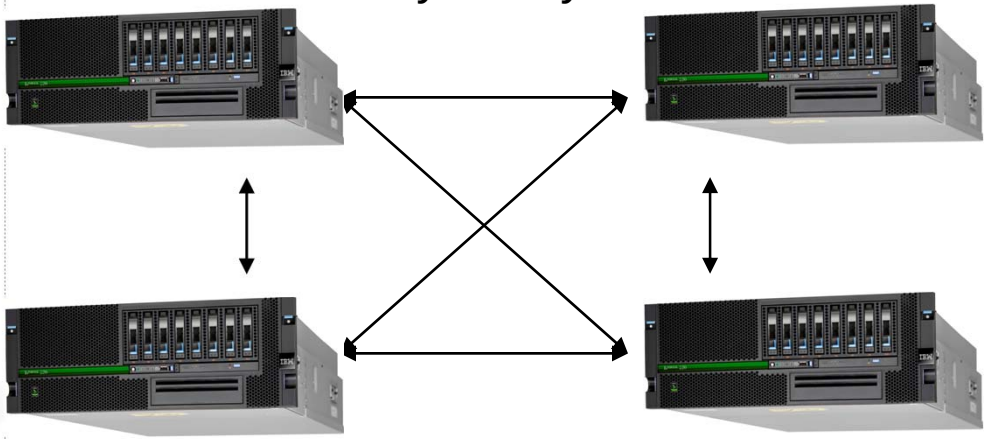
One to one



One to any



Any to one



Any to any

# Common Resources to make highly available



## Service IP Address(es)

- The IP Addresses that users/client apps will use for production
- This can be one or multiple addresses
- Not limited to the number of interfaces when utilizing aliasing

## Application (Server)

- Application(s) desired to be controlled/protect by POWERHA
- Many cases can be user provided start/stop script
- May take advantage of pre-packaged application Smart Assists.

## Shared Storage

- Volume Groups
- Logical Volumes
- JFS
- NFS

# Other Standard Resource Options



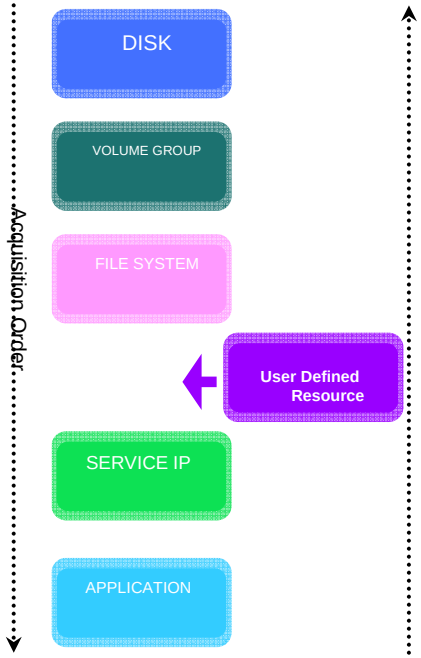
```
Filesystems                                              ALL
Filesystems Consistency Check                            fsck
Filesystems Recovery Method                              parallel
Filesystems/Directories to be exported (NFSv3)
Filesystems/Directories to be exported (NFSv4)
Filesystems to be NFS mounted
Network For NFS Mount
Filesystem/Directory for NFSv4 Stable Storage
Volume Groups
Concurrent Volume Groups
Use forced varyon for volume groups, if necessary        false
Disks
GMVG Replicated Resources
GMD Replicated Resources
PPRC Replicated Resources
SVC PPRC Replicated Resources
EMC SRDF® Replicated Resources
Hitachi TrueCopy® Replicated Resources
Generic XD Replicated Resources
AIX Connections Services
AIX Fast Connect Services
Shared Tape Resources
Application Servers
Highly Available Communication Links
Primary Workload Manager Class
Secondary Workload Manager Class
Delayed Fallback Timer
Miscellaneous Data
Automatically Import Volume Groups                       false
Inactive Takeover
SSA Disk Fencing                                         false
Filesystems mounted before IP configured                 true
WPAR Name
User Defined Resources                                  [ ]
```

Resources in Red are Enterprise Edition Specific

# User Defined Resources (7.1)



- **Presently, A user can introduce a new resource type by creating a Application Server.**
  - PowerHA requires a start/stop/monitor for that resource
- **However, PowerHA follows a strict pre-known order to handles the resources.**
  - Volume Groups will be handled first
  - Application Servers will be handled at the last.
- **This approach is not flexible for end users**
- **What's new in the next release**
  - Introducing a concept of 'User Defined Resource Types' where, user is allowed to develop a bundle which includes attributes/verifications/order for PowerHA etc.
  - Framework accepts Methods to verify/start/stop/monitor/cleanup/restart the user defined resource
  - A xml file can be supplied as input which will be having definition for the user defined resource type
  - PowerHA allows to create instances of the user defined resource type and these instances can be added into RG as resources.



# Custom Resource Groups



## Startup Preferences

- Online On Home Node Only (cascading) - (OHNO)
- Online on First Available Node (rotating or cascading w/inactive takeover) - (OFAN)
- Online On All Available Nodes (concurrent) - (OAAN)
- Startup Distribution

## Fallover Preferences

- Fallover To Next Priority Node In The List - (FOHP)
- Fallover Using Dynamic Node Priority - (FDNP)
- Bring Offline (On Error Node Only) - (BOEN)

## Fallback Preferences

- Fallback To Higher Priority Node - (FBHP)
- Never Fallback - (NFB)

# Additional Granular Resource Group Options



**Resource Group Dependencies**

- **Parent/Child Relationships**
  - Great for Multi-Tier environments
- **Location Dependencies**
  - **Online on Same Node**
    - All resource groups must be online on the same node
  - **Online on Different Nodes**
    - All resource groups must be online on different nodes
  - **Online on Same Site**
    - All resource groups must be online on the same site
- **Specific Order processing (7.1)**
  - STARTAFTER
  - STOPAFTER

**Resource Group Priorities (Different Node Dep.)**

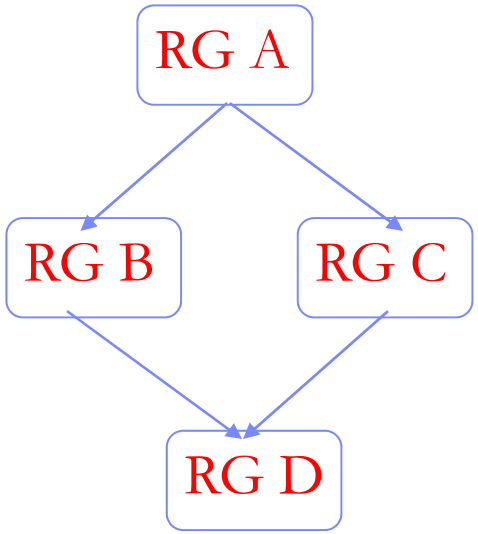
- Low
- Intermediate
- High

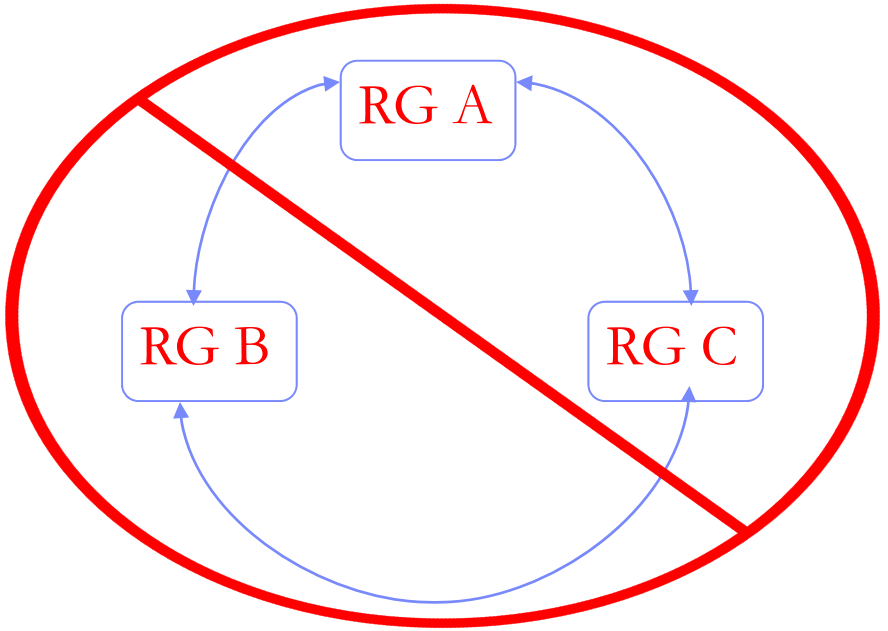
# Resource Group Dependencies



- The maximum depth of the dependency tree is three levels, but any resource group can be in a dependency relationship with any number of other resource groups

- Circular dependencies are not supported, and are prevented during configuration time



# Online on Different Node Priorities

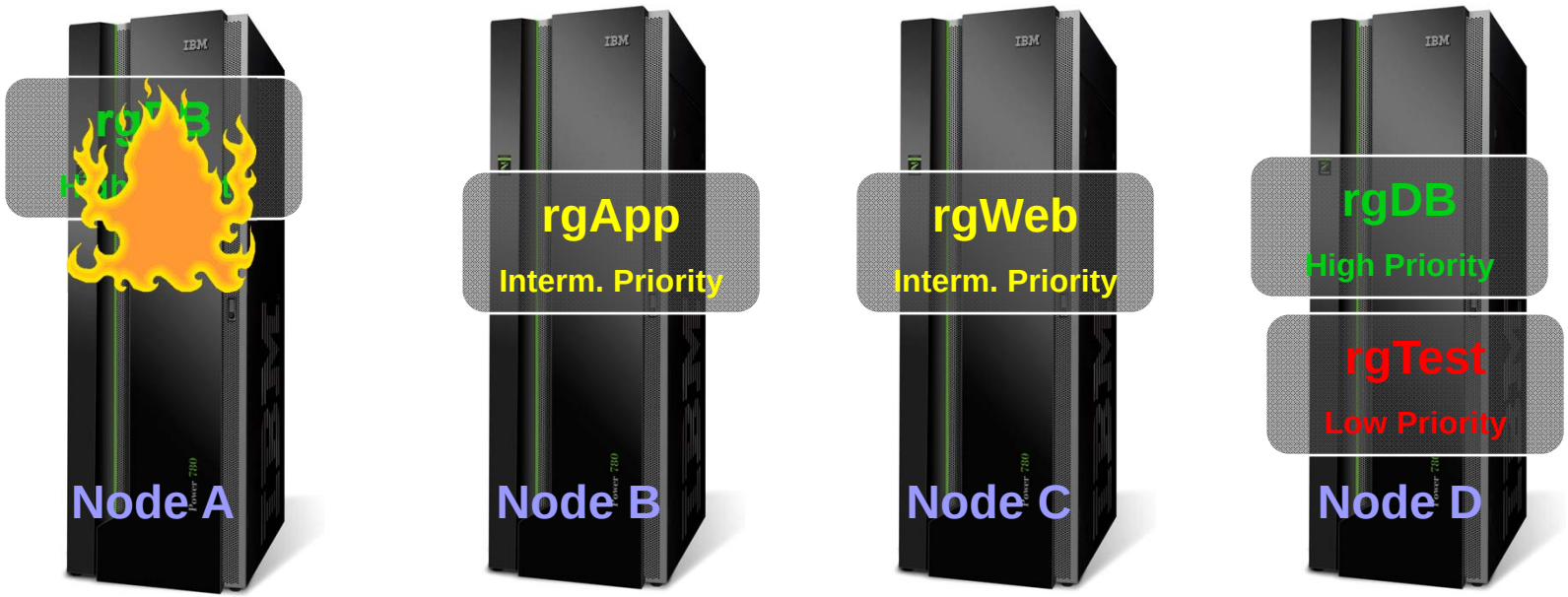


- You can assign High, Intermediate, and Low priority to each resource group
- Higher priority resource groups take precedence over lower priority groups at startup, fallover, and fallback
- **High** priority groups can force **Intermediate** and **Low** priority groups to move or go offline
- **Intermediate** priority groups can force **Low** priority groups to move or go offline
- **Low** priority groups cannot force any other groups to move or go offline
- Groups of the same priority cannot force each other to move or go offline
- RGs with the same priority cannot come ONLINE (startup) on the same node
- RGs with the same priority do not cause one another to be moved from the node after a fallover or fallback

# Example: Online on Different Nodes



- rgDB, rgApp, rgWeb, rgTest
  - Non-concurrent
  - rgDB nodelist: nodeA, nodeD, nodeC, nodeB
  - rgApp nodelist: nodeA, nodeB, nodeD, nodeC
  - rgWeb nodelist: nodeA, nodeB, nodeC, nodeD
  - rgTest nodelist: nodeD, nodeC, nodeB, nodeA
- rgDB has High priority, rgApp has Intermediate priority, rgWeb and rgTest have Low priority



# Dynamic Node Priority - DNP



- Fallover candidate node chosen by available resources
  - Free CPU
  - Paging Space
  - Disk I/O

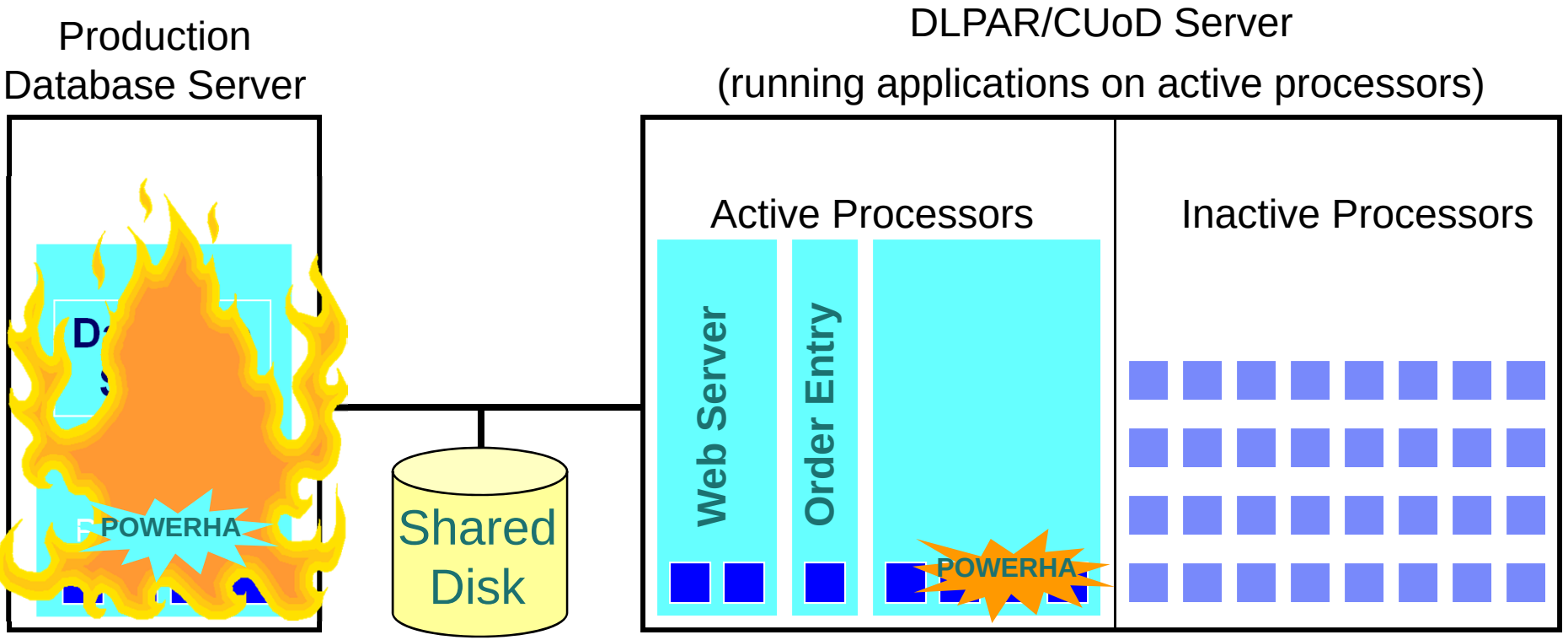
- Adaptive Fallover (7.1)
  - user supplies a script which can dictate the failover behavior.
    - The supplied script will be executed on all nodes by PowerHA to know whether the given node can be used as host for failing over RG
    - Since it is a script, no of checks can be performed by user to let PowerHA know whether the node can be used as a host on a dynamic basis.

- The return code of a user-defined script determines the destination node:
  - cl_lowest_nonzero_udscript_rc
  - cl_highest_udscript_rc

# DLPAR/CoD configuration



- POWERHA on the primary machine detects the failure
- Running in a partition on another server, POWERHA grows the backup partition, activates the required inactive processors and restarts application



# Dynamic Node Priority new Adaptive Fallover



- **Present PowerHA support static/dynamic failover policies**
  - Static: New failover node will be the next node from the node participation list
  - Dynamic Node Priority : allows the failover node to be having certain degree of free CPU/memory/IO activity

- This is not sufficient in some specific cases like SAP Enqueue replication, where RG for Enqueue server need to failover where Enqueue Replication is running**.**
- **What's new in the next release** (7.1)
  - Introducing a new dynamic failover policy where, it asks user to supply a script which can dictate the failover behavior.
    - The supplied script will be executed on all nodes by PowerHA to know whether the given node can be used as host for failing over RG
    - Since it is a script, no of checks can be performed by user to let PowerHA know whether the node can be used as a host on a dynamic basis.

- The DNP feature is enhanced to support two more policies.
- The return code of a user-defined script is used in determining the destination node:
  - cl_lowest_nonzero_udscript_rc
  - cl_highest_udscript_rc

# Application Monitoring



- **POWERHA can monitor applications in one of two ways:**
  - *Process Monitor* – determines the death of a process
  - *Custom Monitor* – monitors health of the application using a monitor method you provide

- **Decisions upon failure**
  - **Restart –** Can establish a number of restarts to restart locally. After a specified restart count, if app continues to fail you can escalate to a fallover.
    - **Notify –** Send email notification
    - **Fallover –** Move application and associated resource group to next candidate node.

- **Suspend/Resume Application Monitoring at anytime.**

# PowerHA File Collections



- **Management feature to simplify keeping common files consistent among cluster nodes.**
- **Allows one or more files to be kept in sync throughout the cluster.**
  - **Can use wildcard filenames**
  - **Can specify an entire directory**

- **Completely automatic and supports all regular files.**
- **Meant for typical configuration files.**
- **Files can be synchronized in three ways:**
  - Manually - using SMIT
  - During cluster Verification and Synchronization
  - Automatically - upon a change in the file.

- **PowerHA provides two default File Collections**
  - Configuration_Files
  - HACMP_Files

- **The original files are backed up under /var/hacmp/filebackup, just one backup copy is maintained, with full path and name.**

# Two-Node Configuration Assistant



- Two-Node Configuration Assistant uses existing PowerHA configuration discovery to further simplify configuration.
- Use it to set up high availability for a single-application cluster. (Hot Standby Configuration)
- SMIT and Java-GUI interfaces are provided.
- Uses File Collections and auto-corrective actions
- All done by answering 5 easy questions*

*Pre-reqs to use this feature are:

PowerHA must be installed and communication daemon running
IP addresses must be assigned to interfaces (and in /etc/hosts)
Shared storage must be available to both nodes
Volume Group must be defined to at least one of the two nodes
Application server scripts must exist on at least primary node

.

# PowerHA can configure a cluster in five questions



1. What is the address of the backup node?
2. What is the name of the application?
3. What script PowerHA should use to start it?
4. What script PowerHA should use to stop it?
5. What is the service IP label that clients will use to access the application?

PowerHA Version 6.1 menu shown below

```
                    Two-Node Cluster Configuration Assistant

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Primary / Local Node                                jordan                +
* Communication Path to Takeover Node                 []                    +
* Application Server Name                             [jordan_app_01]
* Application Server Start Script                     []
* Application Server Stop Script                      []
* Service IP Label                                    []                    +
  Netmask(IPv4)/Prefix Length(IPv6)                   []
* Resource Group Name                                 [jordan_rg_01]
* Cluster Name                                        [jordan_cluster]
```

Recorded demo

http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS2453

# Service Alias Distribution Policies



*Resource level location policy* have 7 total options including:

**Collocation** - all service labels will be on the same physical resource.

**Collocation with Persistent Labels** - all service labels will be on the same interface as the persistent IP.

**Collocation with Source -** all service labels will be on the same physical resource utilizing source service IP

**Anti-collocation** - all resources of this type will be allocated on the first physical resource which is not already serving (or serving the least number of) a resource of the same type.

- This is identical to the the default in previous versions.

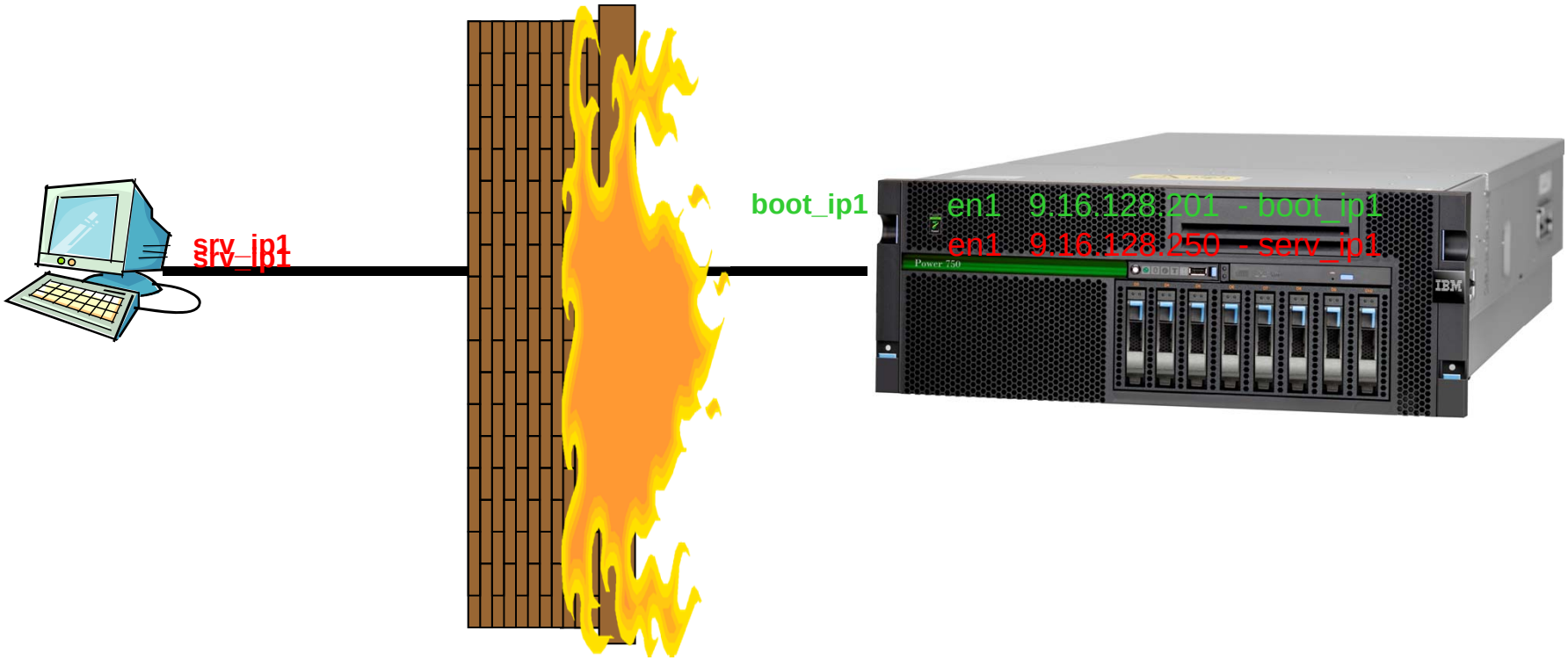
**Anti-collocation with Source -** Same as above plus source service IP.

**Anti-collocation with Persistent Labels** - service labels will almost never be on the same interface as the persistent IP, that is, service will occupy a different interface as long as one is available, but if no other is available then they will occupy the same interface

**Anti-collocation with Persistent Labels and Source** - Same as above plus source service IP.

# Source Service IP Distribution Policy Example

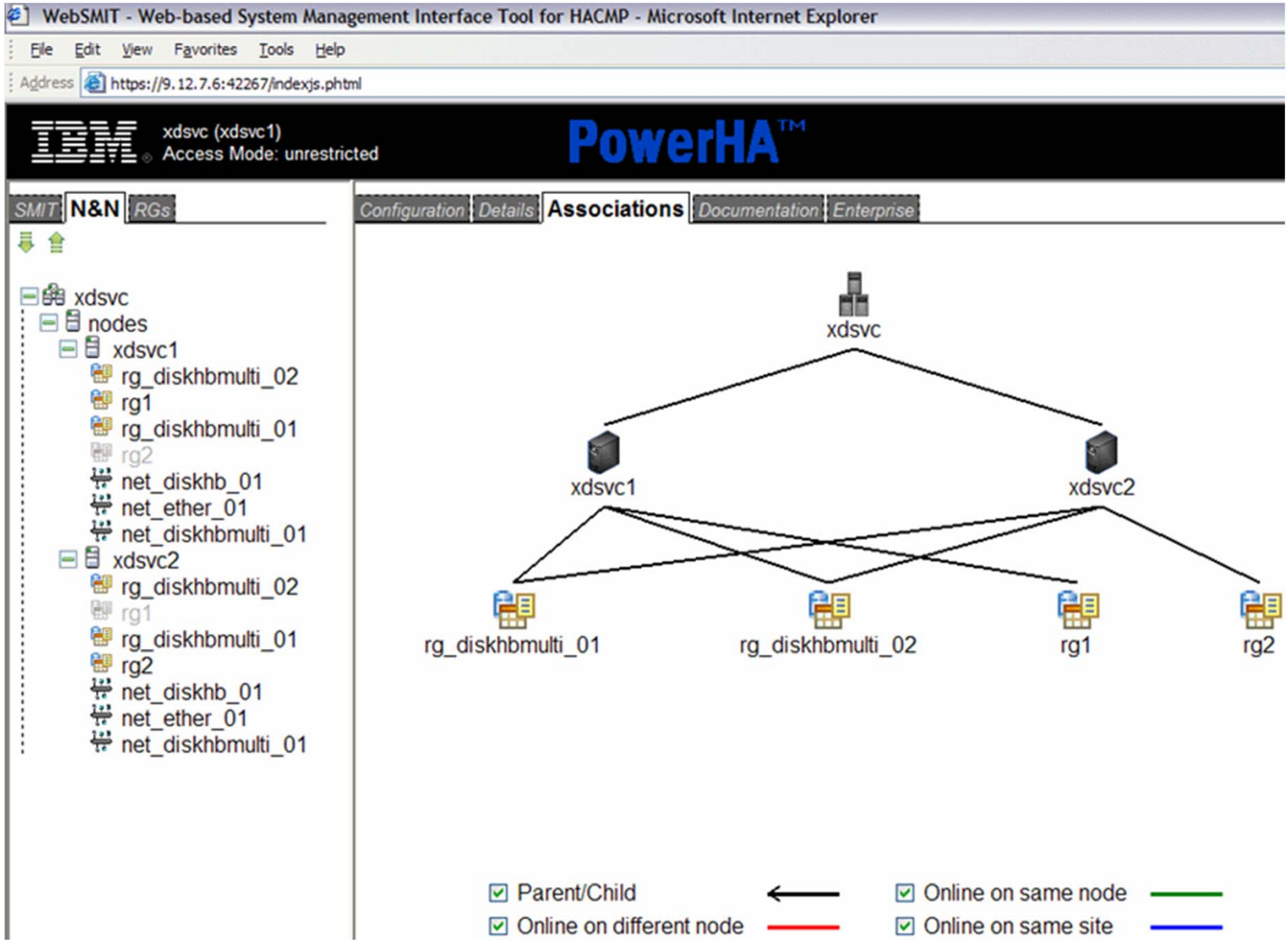





**Firewall rule allows srv_ip1**

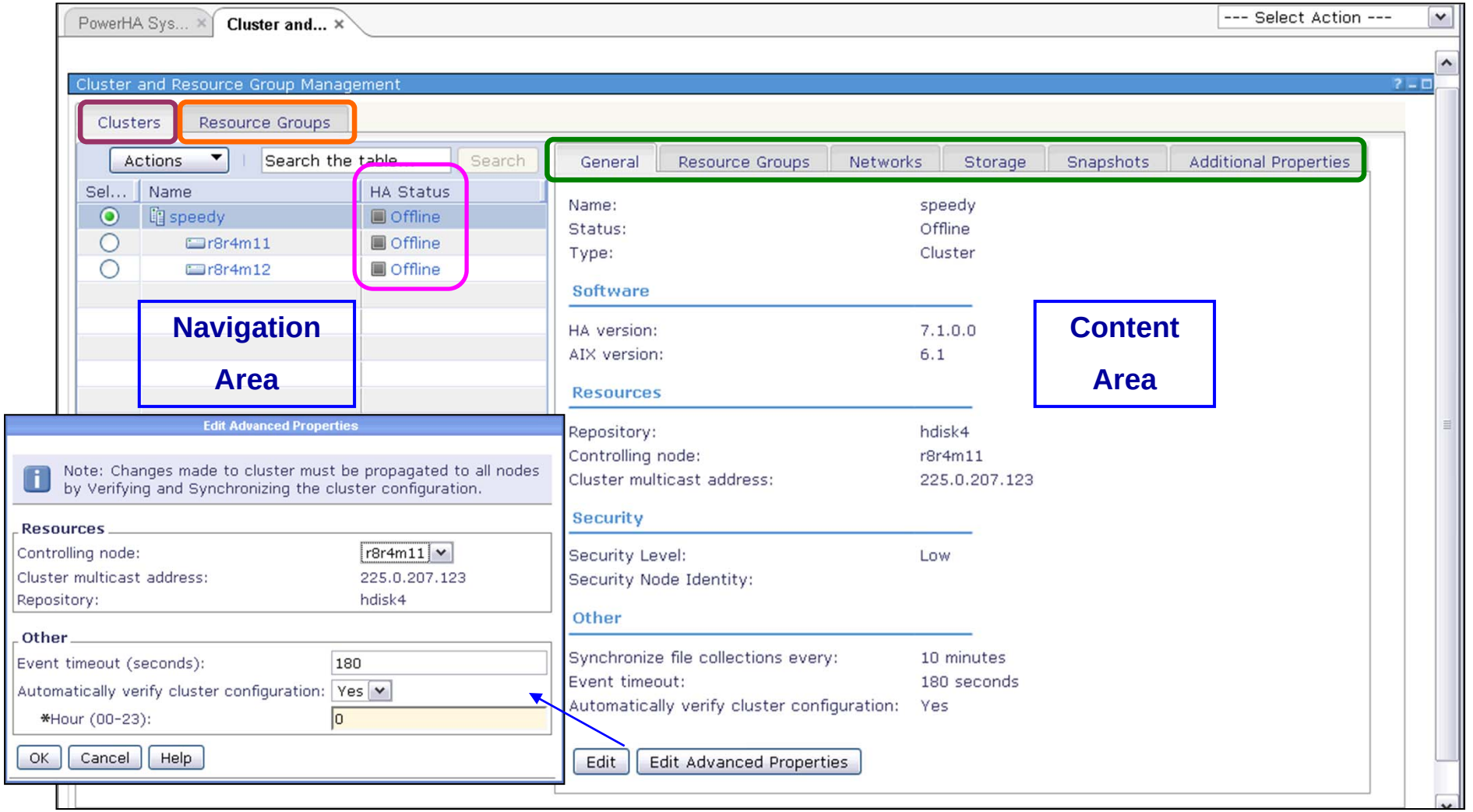
# WebSMIT




WebSMIT - Web-based System Management Interface Tool for HACMP - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address https://9.12.7.6:42267/indexjs.phtml
xdsvc (xdsvc1)
Access Mode: unrestricted
PowerHA™
SMIT N&N RGs
Configuration Details Associations Documentation Enterprise
xdsvc
nodes
xdsvc1
rg_diskhbmulti_02
rg1
rg_diskhbmulti_01
rg2
net_diskhb_01
net_ether_01
net_diskhbmulti_01
xdsvc2
rg_diskhbmulti_02
rg1
rg_diskhbmulti_01
rg2
net_diskhb_01
net_ether_01
net_diskhbmulti_01
xdsvc
xdsvc1
xdsvc2
rg_diskhbmulti_01
rg_diskhbmulti_02
rg1
rg2
Parent/Child
Online on different node
Online on same node
Online on same site

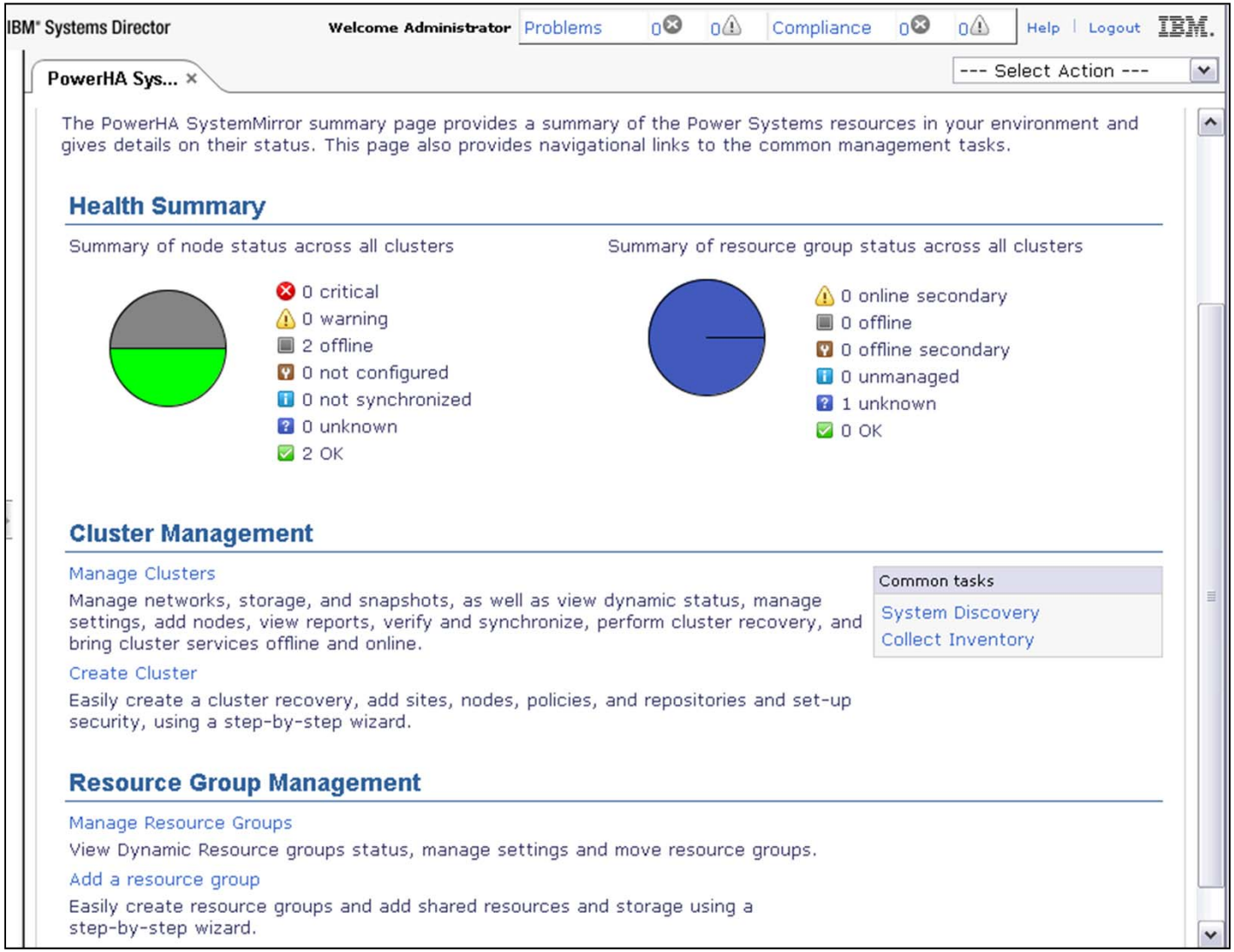
# Systems Director Plug-in – Management View





# System Director Plug-in – Summary Panel





# Clmgr – cluster command line (7.1)



- Supported actions
  - add
  - delete
  - manage
  - modify
  - move
  - offline
  - online
  - query
  - recover
  - sync
  - view
- Supported object classes
  - cluster
  - site
  - node
  - interface
  - network
  - resource_group
  - service_ip
  - persistent_ip
  - application_controller
  - application_monitor
  - dependency
  - file_collection
  - fallback_timer
  - volume_group *(incomplete coverage)*
  - logical_volume *(incomplete coverage)*
  - file_system *(incomplete coverage)*
  - physical_volume *(incomplete coverage)*
  - method *(incomplete coverage)*
  - report
  - snapshot
  - tape

# Clcmd – cluster distributed commands (7.1)



- /usr/sbin/clcmd
- Provided by CAA
- Distributes command to all cluster nodes
  - Ease of use
  - Reminiscent of dsh from SP2 days
- Example
  - clcmd cat /etc/cluster/rhosts

```
NODE jessica.dfw.ibm.com
------------------------------
jessica
jordan
------------------------------
NODE jordan.dfw.ibm.com
------------------------------
jessica
jordan
```

# Smart Assists – List of all Smart Assists



| | SystemMirror 7.1.0 | SystemMirror 7.1.1 |
|---|---|---|
| DB2 Enterprise Edition | 9.5 | 9.7 |
| WAS | 6.1 | 6.1 |
| WAS N/D | 6.1 | 6.1 |
| HTTP Server | 6.1 | 6.1 |
| TSM | 6.1 | 6.2 |
| TDS | 5.2 | 6.3 |
| Filenet | 4.5.1 | 4.5.1 |
| Lotus Domino Server | 8.5.1 | 8.5.1 |
| Oracle Database | 11g r1 | 11g r1 |
| Oracle Application Server | 10g R2 | 10g R2 |
| SAP | SAP ERP netweaver 2004s | SAP SCM 7.0 with Netweaver 7.0 EHP1 for FVT |
| | | SAP SCM 7.0 with Netweaver 7.0 EHP2 for SVT |
| - MaxDB | | v7.6 |
| - Oracle | 10g R2 | 10g R2 |
| - DB2 | | 9.7 |
| MQSeries | | 7.0.1.5 |
| AIX Print Server | | AIX 6.1 |
| AIX DHCP | | AIX 6.1 |
| AIX DNS | | AIX 6.1 |

# POWERHA Cluster Test Tool



- The Cluster Test Tool reduces implementation costs by simplifying validation of cluster functionality.
- It reduces support costs by automating testing of an POWERHA cluster to ensure correct behavior in the event of a real cluster failure.
- The Cluster Test Tool executes a *test pla*n, which consists of a series of individual tests.
- Tests are carried out in sequence and the results are analyzed by the test tool.
- Administrators may define a custom test plan or use the automated test procedure.
- Test results and other important data are collected in the test tool's log file.

# Additional information and Training



PowerHA v7.1 New Features Training Course (AQ100)

http://tinyurl.com/6dvghl3

PowerHA SystemMirror 7.1 for AIX (Redbook)

http://www.redbooks.ibm.com/abstracts/sg247845.html?Open

Follow me on Twitter:

http://twitter.com/#!/POWERHAguy