

The Best Thing Since Sliced Bread: Effective DRS and HA in production

Nitin Suri

Systems Engineer

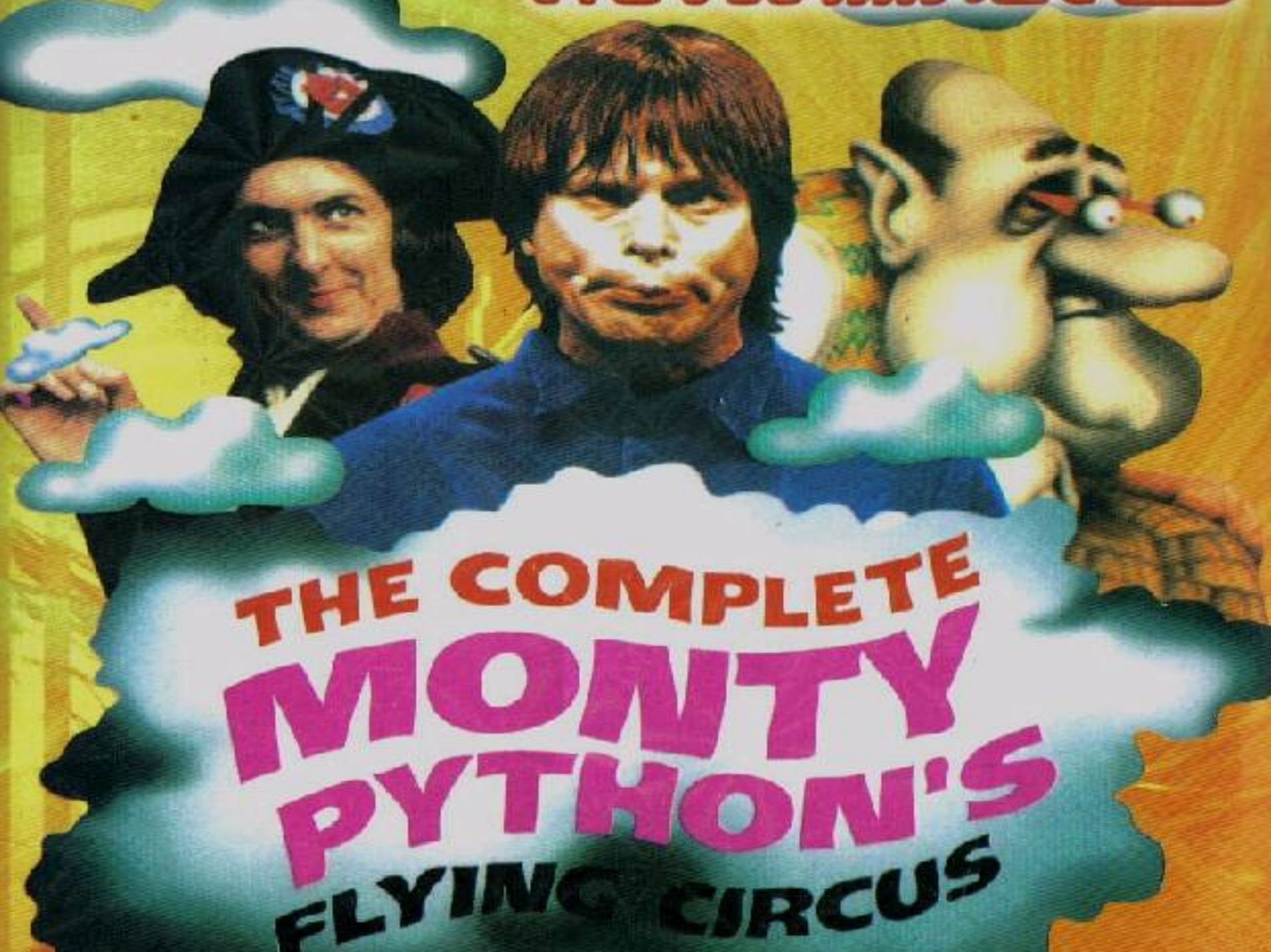


VMWORLD 2006

nitin

suri





**THE COMPLETE
MONTY
PYTHON'S
FLYING CIRCUS**



TomātoTomA^[b]to



VMWORLD 2006

2





SAFFORD
10.0 km

QUEENS
10.0 km

SKI SHOP
7 m

RESERVATIONS
47 m

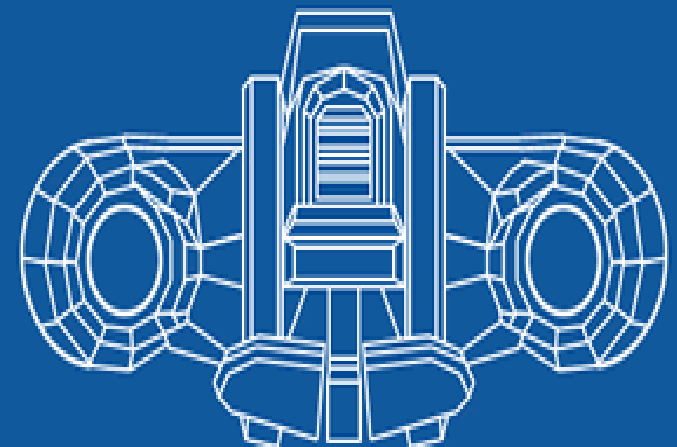
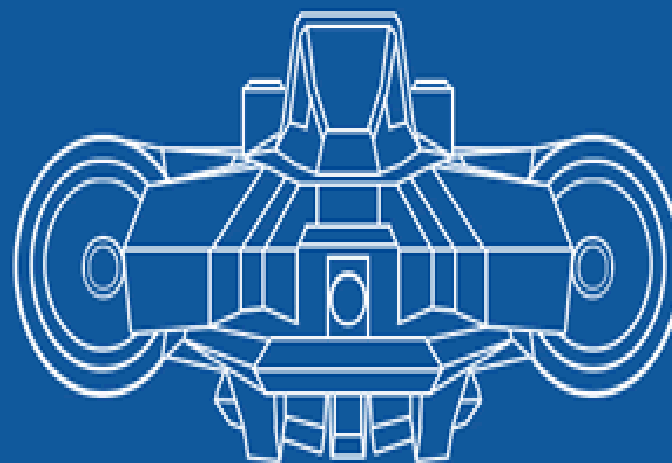
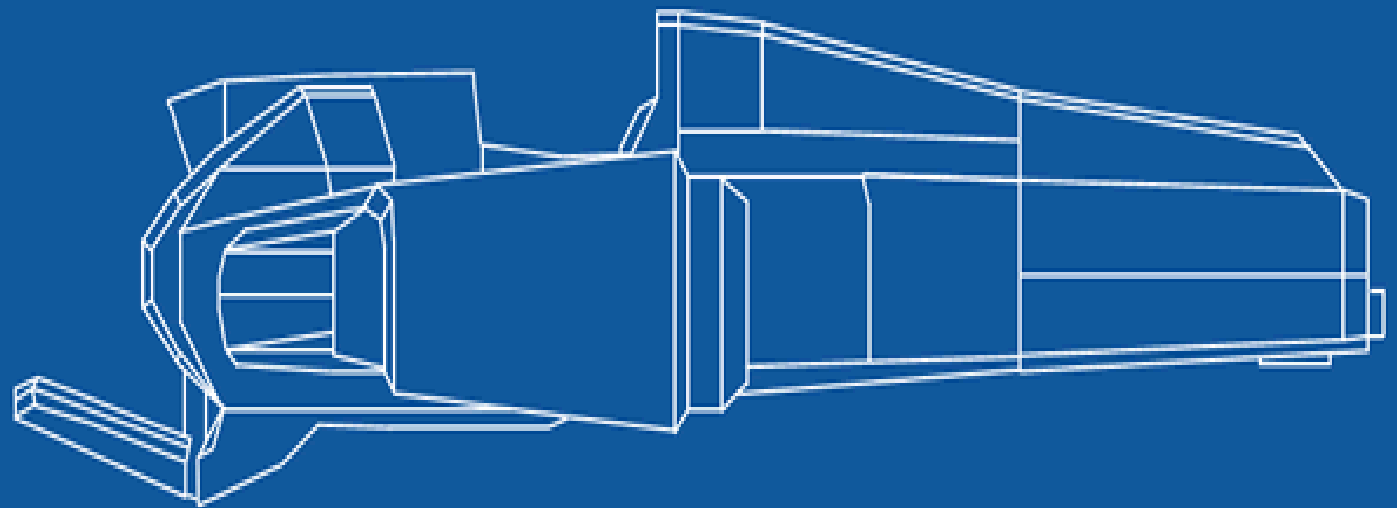
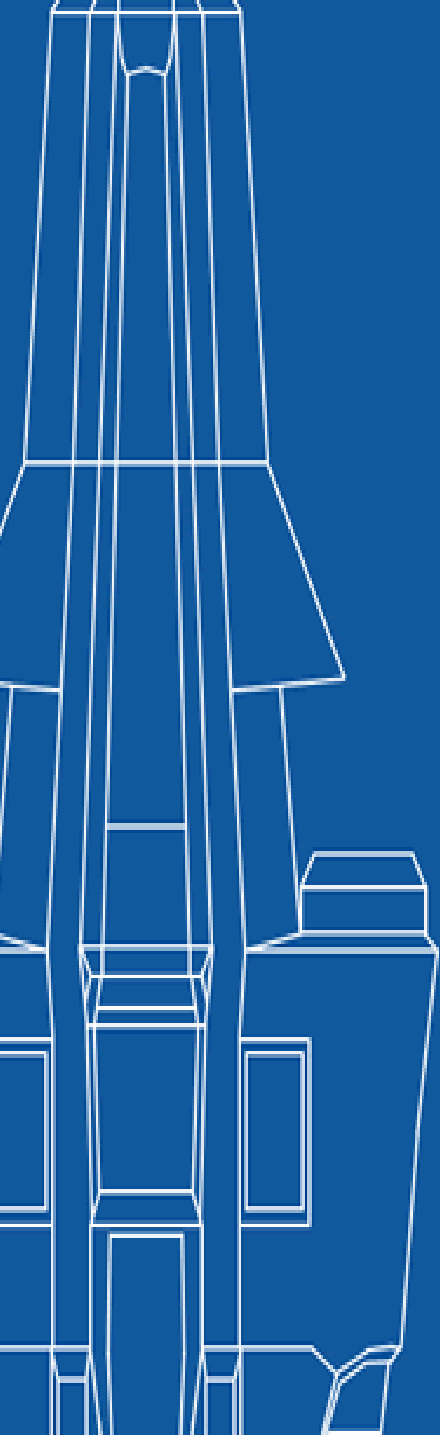
CERWINIA
10.0 km

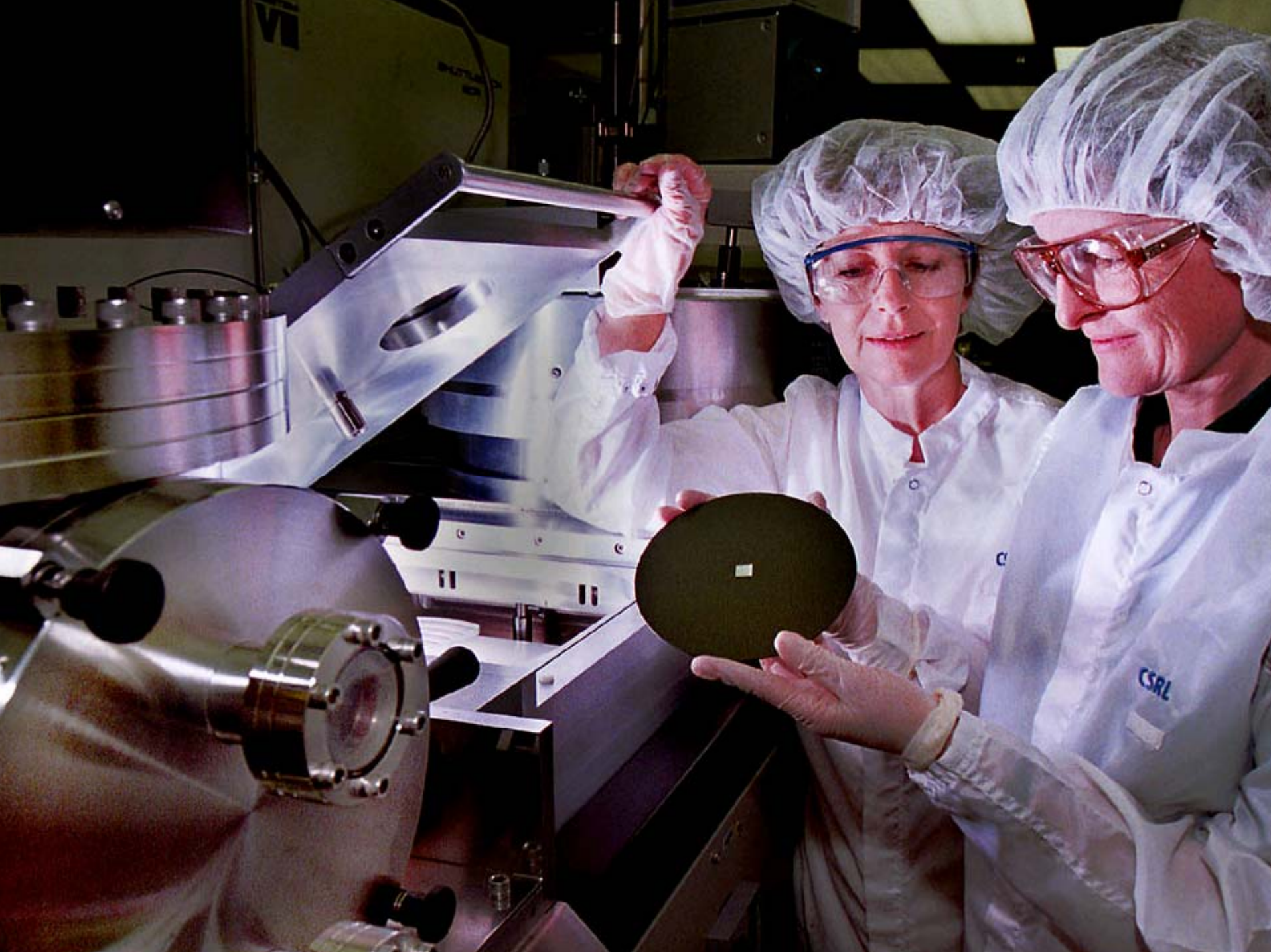
CURS
10.0 km

SKI DECK
10.0 km

MATT
10.0 km

agenda









HA

DRS



ASSUMES

Internals

Installation

Resource Management

NOT

Clustering Strategies

SAN and N/W tips

Clusters are ...

- A collection of ESX hosts and VMs
- Shared resources
- Shared management interface.
- Host's resources become part of the cluster's resources.
- Enable for DRS, HA, or both.

The screenshot shows the 'New Cluster Wizard' window in VMware. The title bar reads 'New Cluster Wizard'. The main heading is 'Cluster Features' with the subtitle 'What features do you want to enable for this cluster?'. On the left, a list of features is shown: 'Cluster Features', 'VMware DRS', 'VMware HA', and 'Ready to Complete'. The 'Cluster Features' section is selected. On the right, there is a 'Name' field containing 'Lab Cluster'. Below this, the 'Cluster Features' section contains instructions: 'Select the features you would like to use with this cluster. At least one option must be selected in order to create the cluster.' Two options are checked: 'VMware HA' and 'VMware DRS'. Descriptions for each are provided. At the bottom, there are buttons for 'Help', '< Back', 'Next >', and 'Cancel'.

New Cluster Wizard

Cluster Features
What features do you want to enable for this cluster?

Cluster Features
VMware DRS
VMware HA
Ready to Complete

Name
Lab Cluster

Cluster Features
Select the features you would like to use with this cluster. At least one option must be selected in order to create the cluster.

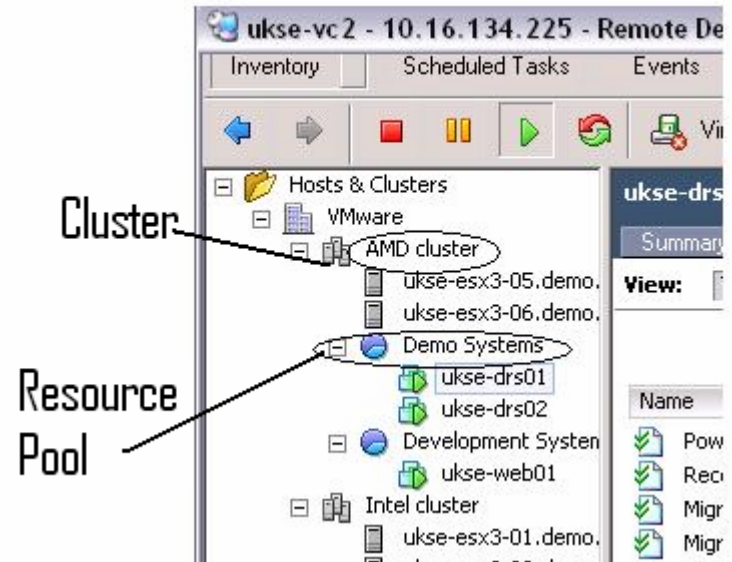
☒ **VMware HA**
VMware HA allows VirtualCenter to automatically migrate and restart virtual machines when a host fails.

☒ **VMware DRS**
VMware DRS enables VirtualCenter to manage hosts as an aggregate pool of resources. Cluster resources can be carved up into smaller resource pools for users, groups, and virtual machines.
VMware DRS also enables VirtualCenter to manage the assignment of virtual machines to hosts automatically, suggesting placement when virtual machines are powered on, and migrating running virtual machines to balance load and enforce resource allocation policies.

Help < Back Next > Cancel

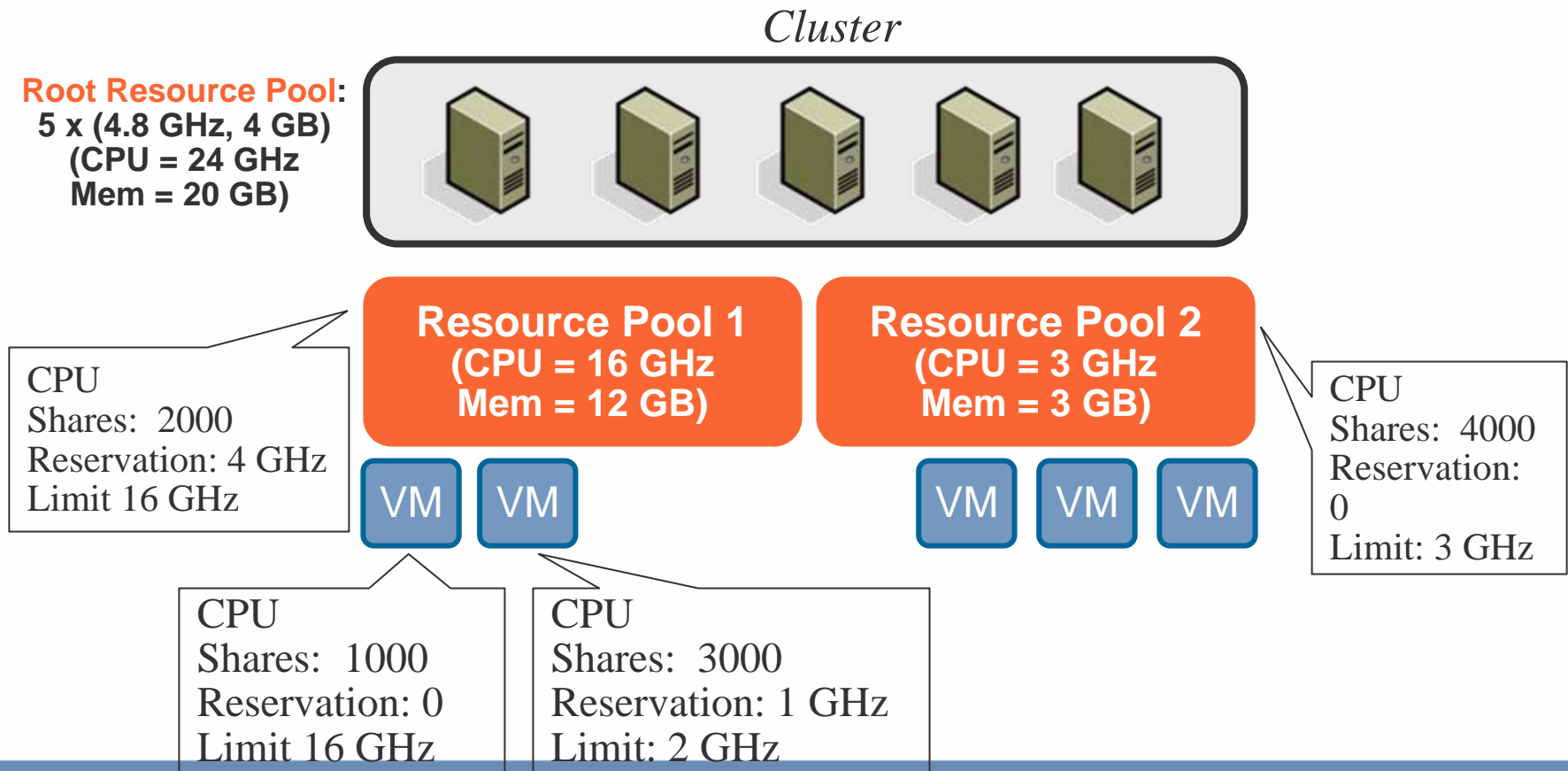
Resource Pools

- Pool of CPU and memory for VMs.
- Used to distribute the computing resources in standalone hosts and clusters
- Attributes: Shares, reservations, limits and expandable reservation
- Achieve hierarchy and resource isolation of multiple resource pools in the same cluster.
 - Departments, functions, teams, projects, clients etc.



Resource pools in a DRS cluster

- To subdivide the computing resources in a cluster



DRS

- Improves resource utilization across all hosts and resource pools.
- Configured for manual, partially or fully automated control
- Input: Resource usage information for hosts and VMs.
- Output: Recommendations for virtual machine placement

DRS Recommendations

- Recommendations based on :
 - Enforcing resource policies accurately
 - Reservation- Guarantee of resources (*at least*)
 - Limits- Upper Bound (*not more than*)
 - Shares – Relative Priority (iff system is overcommitted)
 - Load Balancing VM Loads
 - Balance average CPU and memory loads.
 - Host enters maintenance.
 - Positional/Placement Constraints
 - Affinity/Anti-affinity rules (Domain controllers on different hosts)
 - Vmotion Compatibility (CPU type, LAN, SAN connectivity)

What does DRS do? 1

■ Initial placement

- When you first power on a virtual machine in the cluster
 - Output is recommended host list (prioritized list)
 - Manual Mode
 - Recommends placement. Admin chooses.
 - Partially Automated Mode and Fully Automated:
 - Places VM on host. No recommendations.

What does DRS do? 2

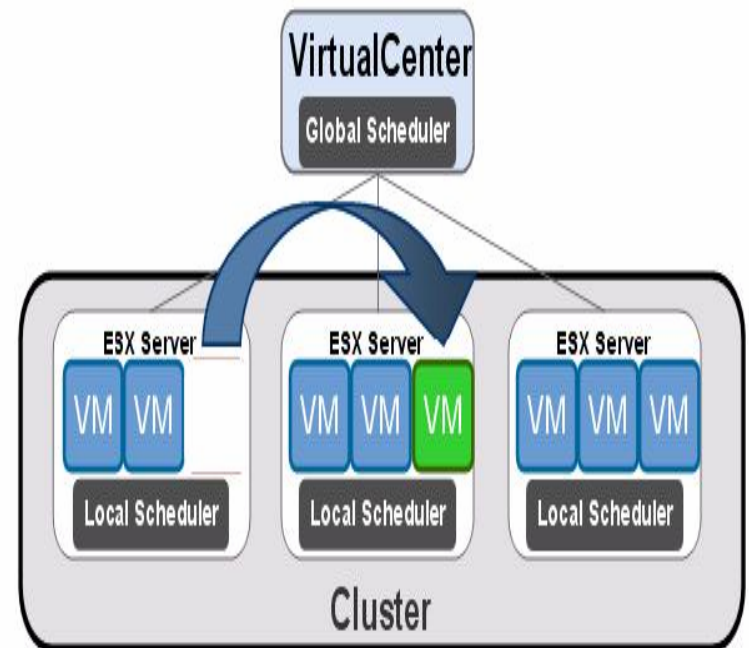
■ Runtime operations

➤ Load Balancing

- Manual Mode and Partially Automated Mode
 - Recommends placement. Admin chooses.
- Fully Automated:
 - Migrates VMs as required. No recommendations.

DRS Architecture – Under the Covers

- Management of all cluster resources
- **Scheduling Levels**
 - **Local:** Within host (determines which processor to run a VM on).
 - **Global :** Within cluster (determines which host to place VMs on).
- **DRS Module**
 - Invoked by VC Server (vpxd)
 - Each instance manages cluster of hosts
 - Invoked
 - Regularly (default 5 mins)
 - Upon event (host addition, removal)
 - Output is recommended actions



A Star is Born – The DRS Star Rating

- Star ratings indicate reduction in cluster imbalance.
 - E.g In "aggressive mode", a slight imbalance triggers VMotion
- Balance CPU and memory load
- Mandatory recommendations for violations (affinity, anti-affinity rules, host evacuation to put a host in maintenance mode).

Migration Recommendations

Priority	Virtual Machine	Reason	Source Host	Target Host	CPU Load	Memory Load
★★★★★	test06	Satisfy anti-affinity rule	vcuiqa010.en...	vcuiqa012.en...	source: 6019 MHz, target: 24 MHz	source: 2 GB, target: 511 MB
★★★★★	test01	Balance average CPU loads	vcuiqa010.en...	vcuiqa012.en...	source: 6019 MHz, target: 24 MHz	source: 2 GB, target: 511 MB
★★★★★	test02	Balance average CPU loads	vcuiqa010.en...	vcuiqa012.en...	source: 6019 MHz, target: 24 MHz	source: 2 GB, target: 511 MB
★★	test08	Balance average CPU loads	vcuiqa010.en...	vcuiqa012.en...	source: 6019 MHz, target: 24 MHz	source: 2 GB, target: 511 MB


Apply Migration Recommendation

Migration Threshold

- Specify which recommendations are automatically applied
- Level 1 - smallest number of migrations.
- Level 5 – Will never see it

☒ Fully automated

Virtual machines will be automatically placed onto hosts when powered on, and will be automatically migrated to attain best use for resources.

Migration threshold: Conservative  Aggressive

Apply recommendations with four or more stars.
This will apply recommendations that promise a significant improvement in the cluster's load balance.

Level	Recommendation applied
1 – Most conservative	With 5 stars only
2 – Moderately conservative	4 or more stars
3 – Midpoint (default)	3 or more stars
4 – Moderately aggressive	2 or more stars
5 – Aggressive	1 or more stars

Creating a DRS Cluster

- Implicit Resource Pool
- Requirement
 - Consistent network labels (case sensitive!)
 - DRS requires shared storage
 - VMotion Requirements
- Add hosts to cluster
 - Can keep host's resource pool hierarchy .
 - Concept of modes (Normal, Maintenance and transition)

Remove Host from a Cluster

- For manual or partially automated
 - Displays recommendations
- For fully automated
 - Migrates running VMs to different hosts.
- Put host into *Maintenance Mode*.
 - Restricted operations:
 - New VMs can NOT be powered on
 - No VMs migrated to this host.
 - Admin shuts down running VMs.
 - Shut down the host.
- Note: If other hosts fail, no VMs are failed over to a host in maintenance mode

DRS Cluster Settings

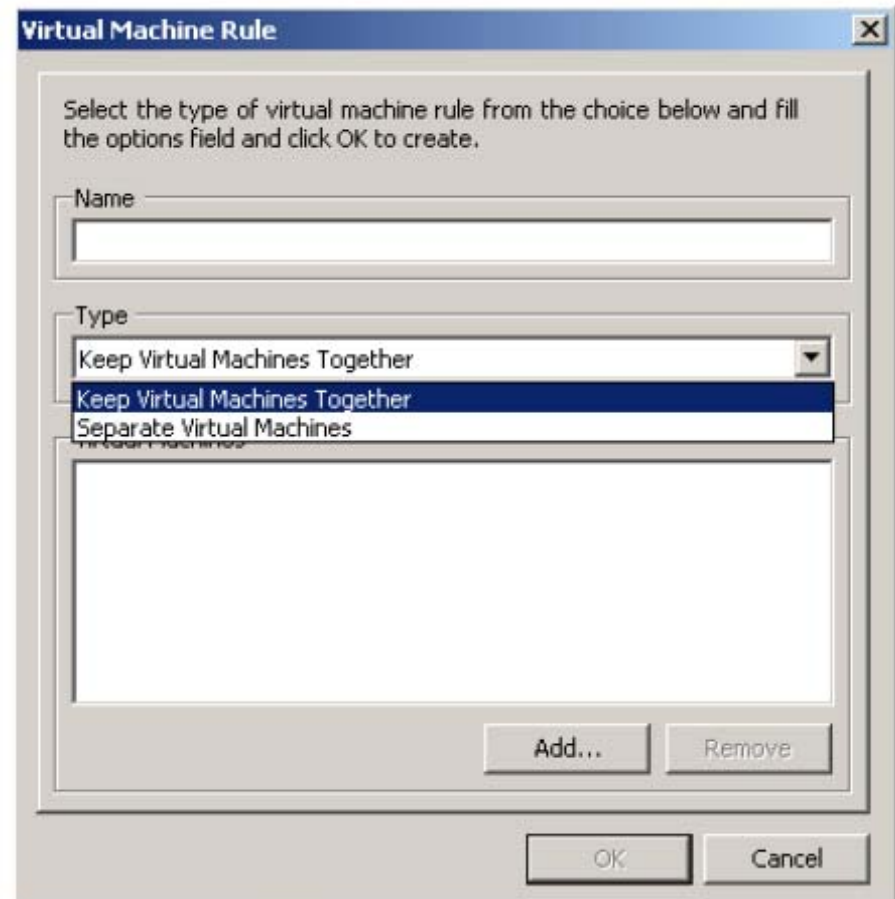
■ Placement Constraints

➤ Affinity

- Improve performance to take advantage of locality of resources.
- e.g. 3 Tier arch using internal networking

➤ Anti-affinity

- For load balancing and availability
- eg 2 dbs NOT on the same host



DRS Cluster Settings (cont)

- Automation Level
 - Over ride option (VM level rather than cluster level)
 - *Disable* option
 - No migrations
 - No recommendations.
 - Tip:
 - For critical applications :
Use manual mode on a per VM basis.

The screenshot shows the VMware DRS Virtual Machine Options configuration page. On the left, a navigation pane lists 'General', 'VMware HA', 'Virtual Machine Options', 'VMware DRS', 'Rules', and 'Virtual Machine Options' (highlighted). The main content area has a title 'Use this page to set individual automation mode options for virtual m the cluster.' and a dropdown menu 'Virtual Machine or Automation Level contains:'. Below this is a table with two columns: 'Virtual Machine' and 'Automation Level'. The table lists several VMs: ProdTemplate, Prod04-2, Dev03, Test04-1, Prod03-2 (highlighted), TestNATRouter, Prod03-1, Dev01, and Dev02. The 'Automation Level' for Prod03-2 is 'Default (Fully Automated)', and a dropdown menu is open for it, showing options: 'Fully Automated', 'Manual', 'Partially Automated' (highlighted), 'Default (Fully Automated)', and 'Disabled'.

Virtual Machine	Automation Level
ProdTemplate	Default (Fully Automated)
Prod04-2	Default (Fully Automated)
Dev03	Default (Fully Automated)
Test04-1	Default (Fully Automated)
Prod03-2	Default (Fully Automated) ▼
TestNATRouter	Fully Automated
Prod03-1	Manual
Dev01	Partially Automated
Dev02	Default (Fully Automated)
	Disabled

Tips

- Expandable Reservation
 - When to use
 - To ensure flexible allocation of resources.
 - Can be used at a nested level.
 - When not to use
 - To limit the amount of resources available for reservation in a pool.
 - e.g. hosting companies, internal bill-able users
- Keep an eye on CPU/Memory for hosts
 - Adjust DRS settings if necessary
 - Set to manual in order to see how things operate
 - Use DRS and Resource Pool for a number of hosts and NOT “per host”

DRS Cluster Usage Monitoring

- Through VC
- Colour coded scheme:
 - No colour – All resource constraints are satisfied.
 - Yellow (Overcommitted) – Some resource constraints are not satisfied.
 - Over commitment or hosts gone down.
 - Solution: Increase resources or decrease reservations.
 - Red – Cluster is invalid or internally inconsistent
 - DRS violation or HA violation (usually an admin bypassed VC and made resource pool changes on a host).
 - Current failover capacity is lesser than configured failover capacity.

DRS Failures

■ VMotion Errors

- VM is in a cluster relationship (e.g. MSCS) with another VM
- VM has a CPU affinity to run on physical CPUs
 - Fix: Remove the VMs from the cluster, fix the affinity settings, move it back into the cluster.
- VM has an active connection to CD-ROM/floppy disk with a local image mounted or an internal virtual switch

■ VMotion Warnings:

- VM is configured to an internal virtual switch but is not connected to it.
- VM is configured to access a local CD-ROM or floppy but is not connected to it

DRS – Can't power on a VM

- Can't power on a VM even when the reservation is less than the unreserved capacity of its parent resource pool?
- A host must have sufficient
 - Aggregate unreserved capacity.
 - Per-core capacity

Disable: set the per-host configuration option
Cpu/VMAdmitCheckPerVcpuMin to 0 (not recommended)

Best Practices—Tips and Tricks 1

- Plan for Nested Resource Pools effectively.
- Max Hosts supported for DRS and HA per cluster in VC2.0 : 16
- Use Fixed Reservation for parent resource pool.

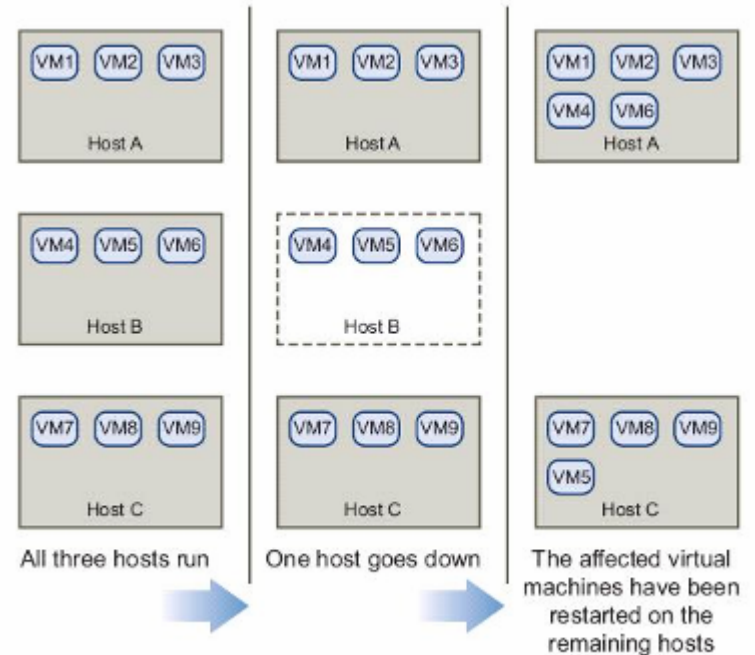
Best Practices—Tips and Tricks 2

- Track DRS and be satisfied with it.
- If DRS does make strong recommendations, please use them.
- Enable Automation
 - But base it on your comfort level –environment, experience etc.
 - Let DRS autonomously manage most VMs.
 - If one is extremely conservative use manual mode for critical VMs.
 - For non-critical VMs use automated.

VMware HA

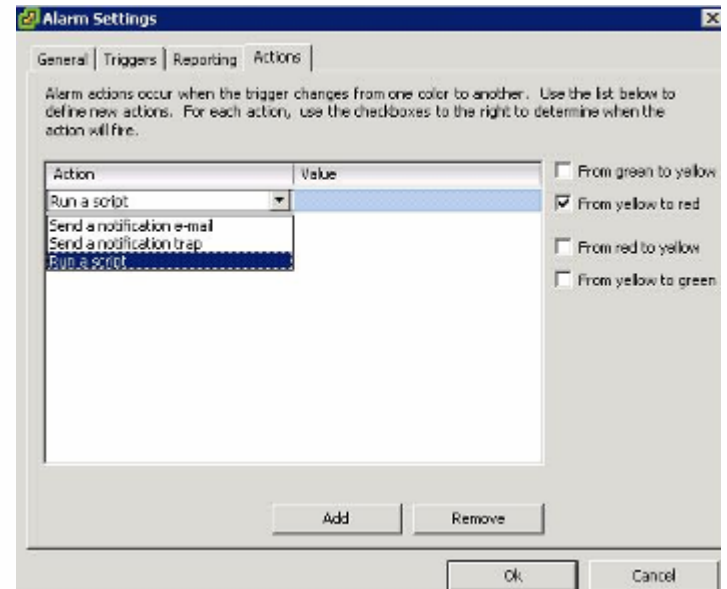
VMware HA

- A cluster enabled for HA monitors for host failure.
- Provides high availability to VMs
- Automatic failover on a cluster of ESX Server hosts.
- Customizable behaviour for individual VMs.



Properties

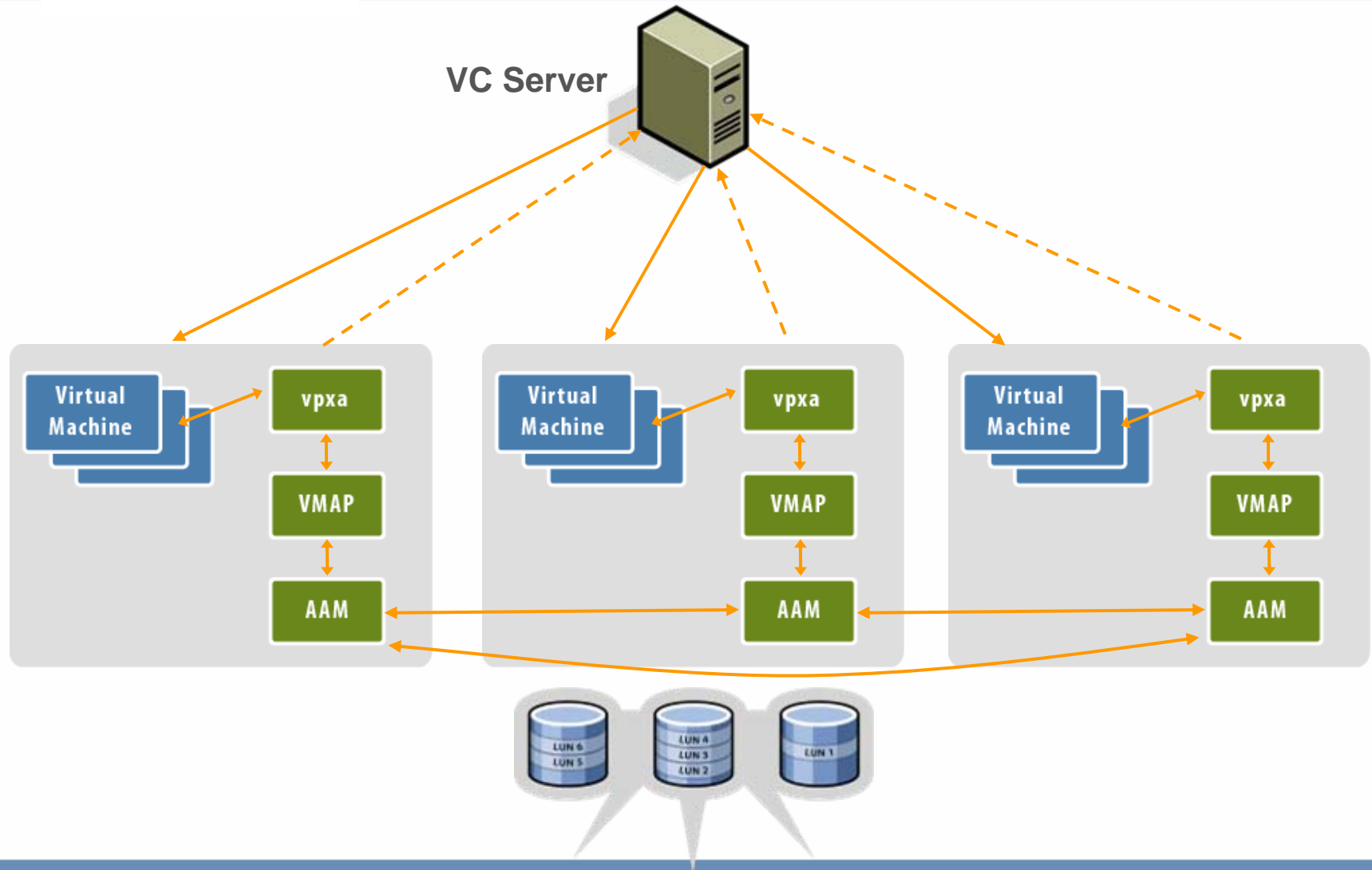
- Downtime will be minimal but non-zero.
- Does not use VMotion!
- VMware HA does not manage individual VM failure
 - Instead use VC alarms and scripts
 - monitor VM heartbeat



Pre-requisites

- Able to power-on a VM from all hosts within the cluster
 - Access to common resources (shared storage, VM network)
- Host should be configured for DNS
 - DNS resolution of all hosts within cluster

Architecture



Automated Availability Manager

- Runs in the service console when an HA cluster is created.
- Maintains an in-memory database of active nodes in the cluster
- Uses heartbeats to co-ordinate the active and passive nodes.
- Very high dependency upon fully functional host name resolution.
 - Check `/etc/hosts` and `/etc/resolv.conf`.
 - Log files in the service console in `/opt/LGTOaam512/`

Primary & Secondary Hosts

Primary Hosts

- Avoid overhead- one host acts as controller (first host).
- Interprets Rules , Provides Redundancy
- Initiates failover actions
- New host has to communicate with a primary host to complete configuration
- If primary host goes down, HA automatically promotes another host to primary status.

■ 4 Host Limit

- Max number of host failures = Maximum number of primaries.
- AAM doesn't allow more than 4 hosts
 - Amount of overhead rises significantly (for each additional primary)

Order of VM restart

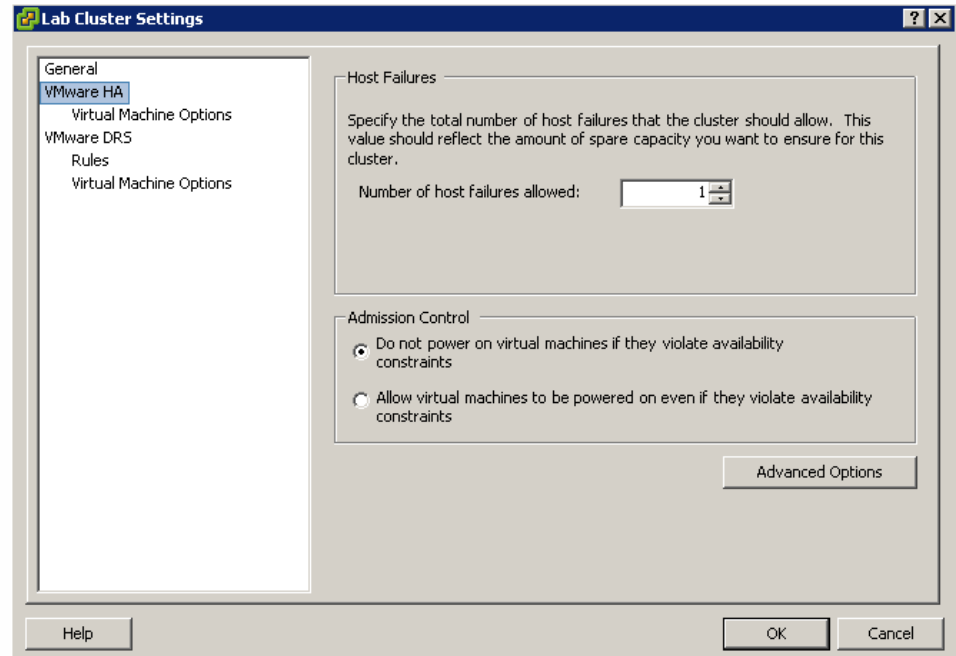
- Restart only powered on VMs that are HA-enabled.
- VMs with higher priority will get started first.
- Which Host?
 - Host Failover Sequence Algorithm of HA
 - In VC2 , goes alphabetically through host list, picks the first host that has enough capacity to accommodate a VM.
 - In VC2.1, picks the host with the most unreserved capacity.

Planning and Tips

- For planning purposes, consider:
 - Number of hosts for which you want to guarantee failover
 - Plan failure capacity for different scenarios(1 ,2,3,4 host failures)
 - Each host has some CPU and memory to power on VM
 - Each VM must be guaranteed its CPU and memory requirements
 - Prioritise your VMs – Should you power on a VM? What order? Best practice.
 - Ensure all servers are placed into DNS to create HA clusters.

HA Cluster Configuration

- Two parameters
 - Host failures
 - Redundant capacity (1 to 4)
 - Admission Control (Fairness)
 - Uptime or resource fairness?
 - When to or not to power on a VM.



HA –Special Situations

■ Power off host

- HA restarts any virtual machines running on that host on a different host.

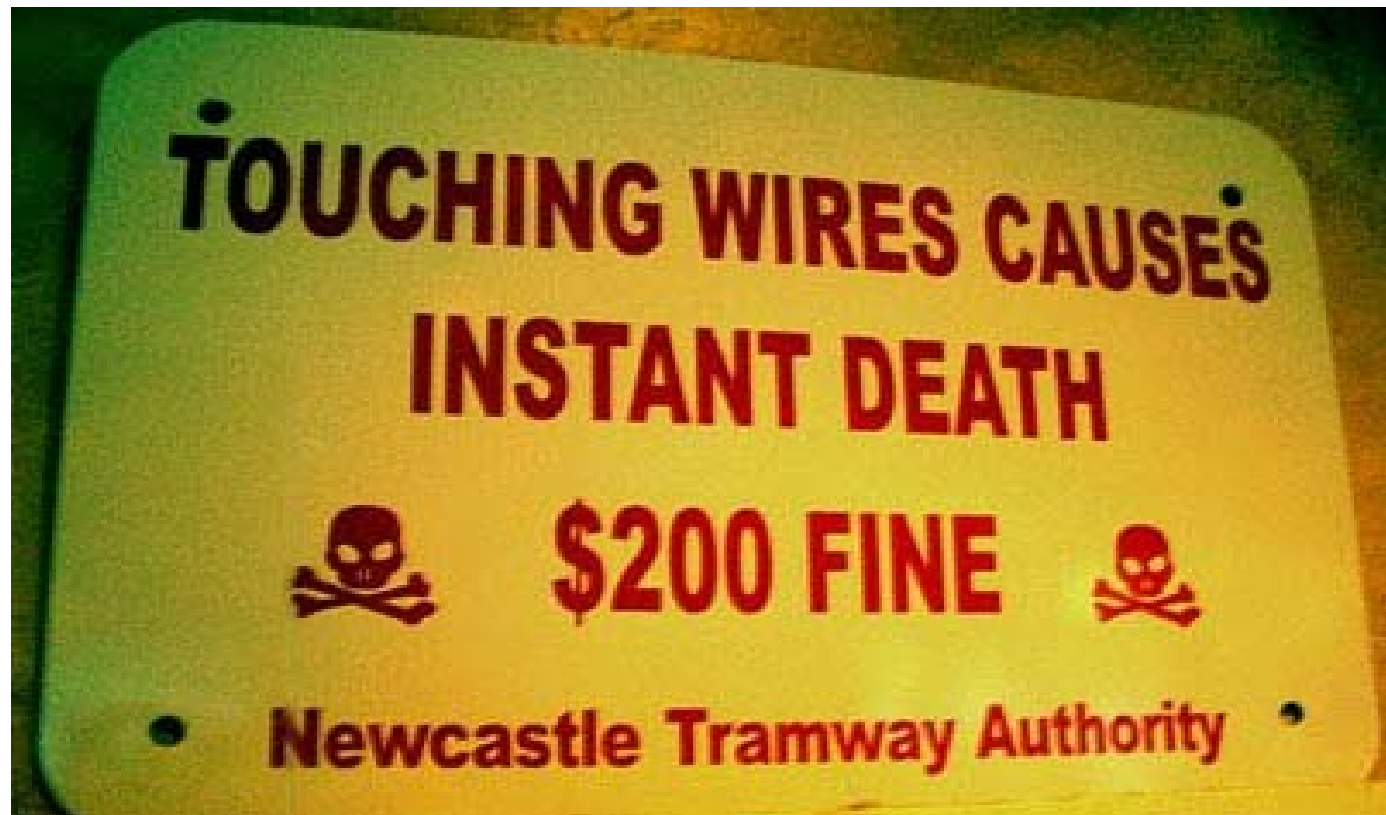
■ Current failover capacity does not match configured failover capacity (Red Cluster)

- More hosts failed than anticipated.
- Fails over VMs with higher priorities first

- Note: In host failure, HA does not failover any VMs to a host in maintenance mode

HA-- Troubleshooting

- IP Connectivity
- DNS resolution
- Ensure storage and networks are visible throughout the cluster.
- No user should manage the hosts by bypassing VC and tweaking resource reservations.
 - Causes state to go to red
- Check logs:
 - /opt/LGTOaam512/log/*
 - /opt/LGTOaam512/vmsupport/*



The HA Error Top 3 Countdown

HA Configuration Error Number 3

- All configuration tasks fail and the following message appears: “Could not find a primary host to configure DAS on”
 - A host that was removed from the VC inventory still appears on the internal list of hosts for this cluster.
 - Workaround: create a new cluster and add hosts to this new cluster.
 - Resolved in ESX Server 3.0.1 and VC 2.0.1.

HA Configuration Error Number 2

- “Configuring HA failed” or “while using HA, the vm did not failover”.
 - Size of Fully Qualified Domain Name (FQDN) or short host name.

- Workaround:

If the host short name is more than 29 characters, change the HOSTNAME entry in /etc/sysconfig/network to the shorter name.

If using an FQDN that is greater than 29 characters:

- Change the FQDN to less than or equal to 29 characters.
- Remove the existing cluster.
- Create a new cluster.
- Add all the hosts back to the cluster.

HA Configuration Error Number 1

■ HA Configuration Fails

➤ Check DNS, FQDN

➤ You have just added a new host to the cluster

- Check /opt/LGTOaam512/log/aam_config_util_addnode.log
- /var/log/vmware/vpx/vpxa.log
- In VC, right click on the host that shows the HA problem and click reconfigure for HA.
- Were all the hosts responding?
- If not –new host cannot communicate with any of the primary hosts.
- Solution:
 - Disconnect all the hosts that are not responding before you can add the new host.
 - The new host becomes the first primary host.
 - When the other hosts become available again, their HA service is reconfigured.

HA Miscellaneous Issues 1

- “HA service doesn’t start up” or “HA error on the host, after booting successfully after host failure”
 - Due to problem rejoining the rest of the cluster
 - Fix: use the ReconfigureHA task on the host to correctly configure the HA service on the host.
- Deploying a VM from a template into a cluster, the updated BIOS settings are lost.
 - Affects HA/DRS clusters but not DRS-only clusters.
 - Fix: Change the BIOS setting in the affected deployed VM.

HA Miscellaneous Issues 2

- ESX Server is running slowly.
- Top command shows vmware-hostd using a lot of resources
- vmware-hostd Uses a Lot of CPU or Has Generated a Core Dump
- Why
 - DRS / HA cluster that contains VMs originally built on an ESX Server 2.x host.
- Fix: To adjust resource allocation for vmware-hostd:
 - For each VM, right click it and go to Edit Properties.
 - Select Resources.
 - Ensure you "touch" each CPU and memory resource:
 - Restart hostd: type: `service mgmt-vmware restart`
 - Set resources to the settings you want.

Clusters and VC Server Failure

- **HA**

- HA clusters continue to work
- Restart VMs on other hosts.

- **DRS**

- The hosts in DRS clusters continue running using available resources
- Note: There are no recommendations for resource optimization.

- If changes to hosts or VMs are made using a VI Client connected to an ESX Server host while the VC Server is unavailable, those changes DO take effect.

HA Best Practices

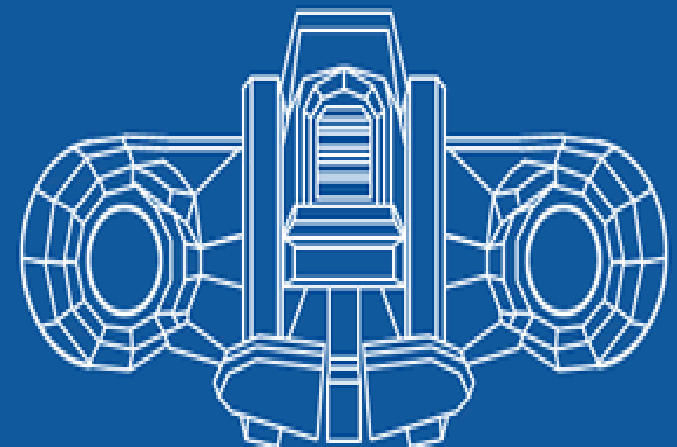
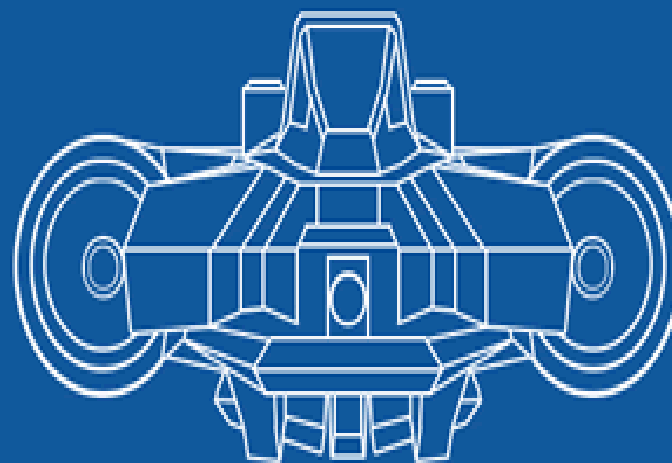
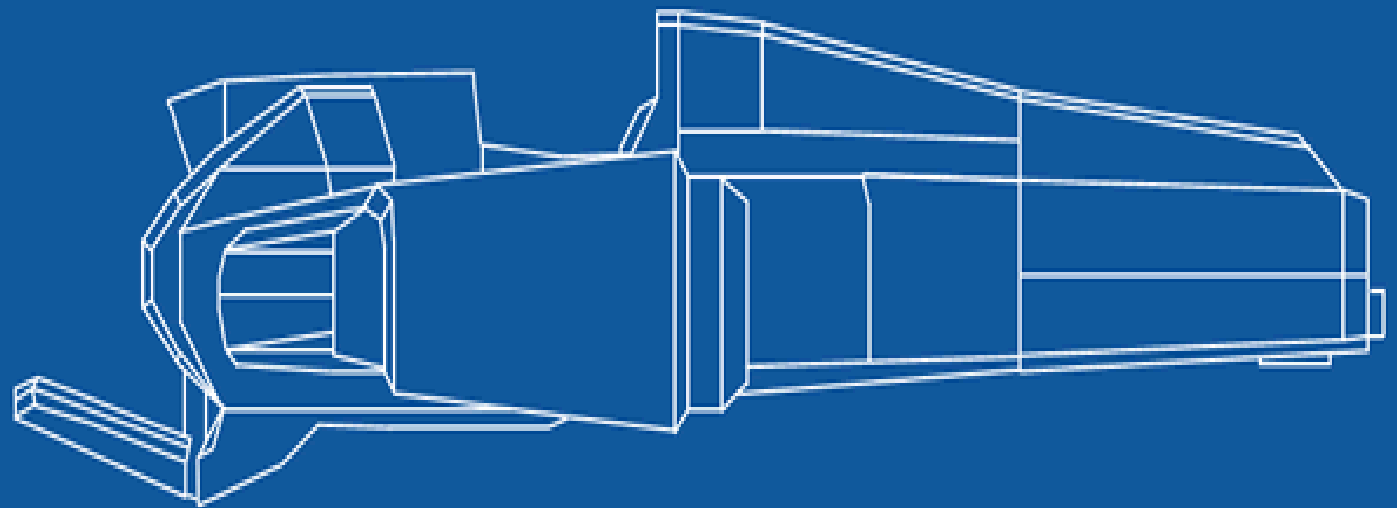
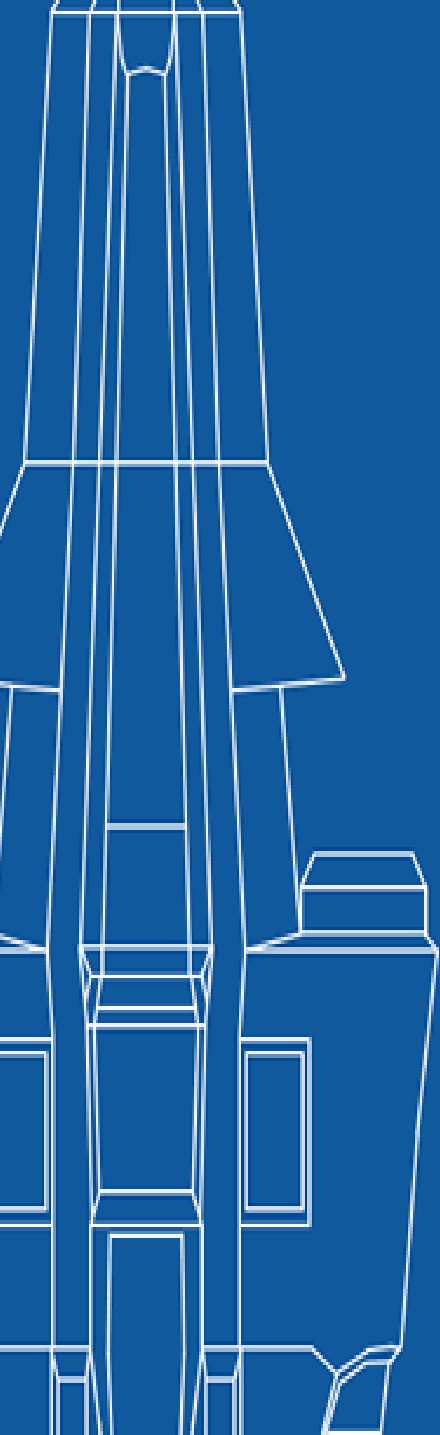
- Most issues are due to DNS issues.
 - Limit of 29 chars for the host name and DNS suffix
 - Enter Fully Qualified Domain Name
 - Each host in an HA cluster must be able to resolve the host name and IP address of all other hosts in the cluster.
 - Set up DNS on each host
 - Recommended: Edit the /etc/hosts file to provide redundancy in case DNS lookups fail (documentation discourages it).
 - Proper way-- Edit the nsswitch.conf file and change the hosts line to read: "hosts: dns files".

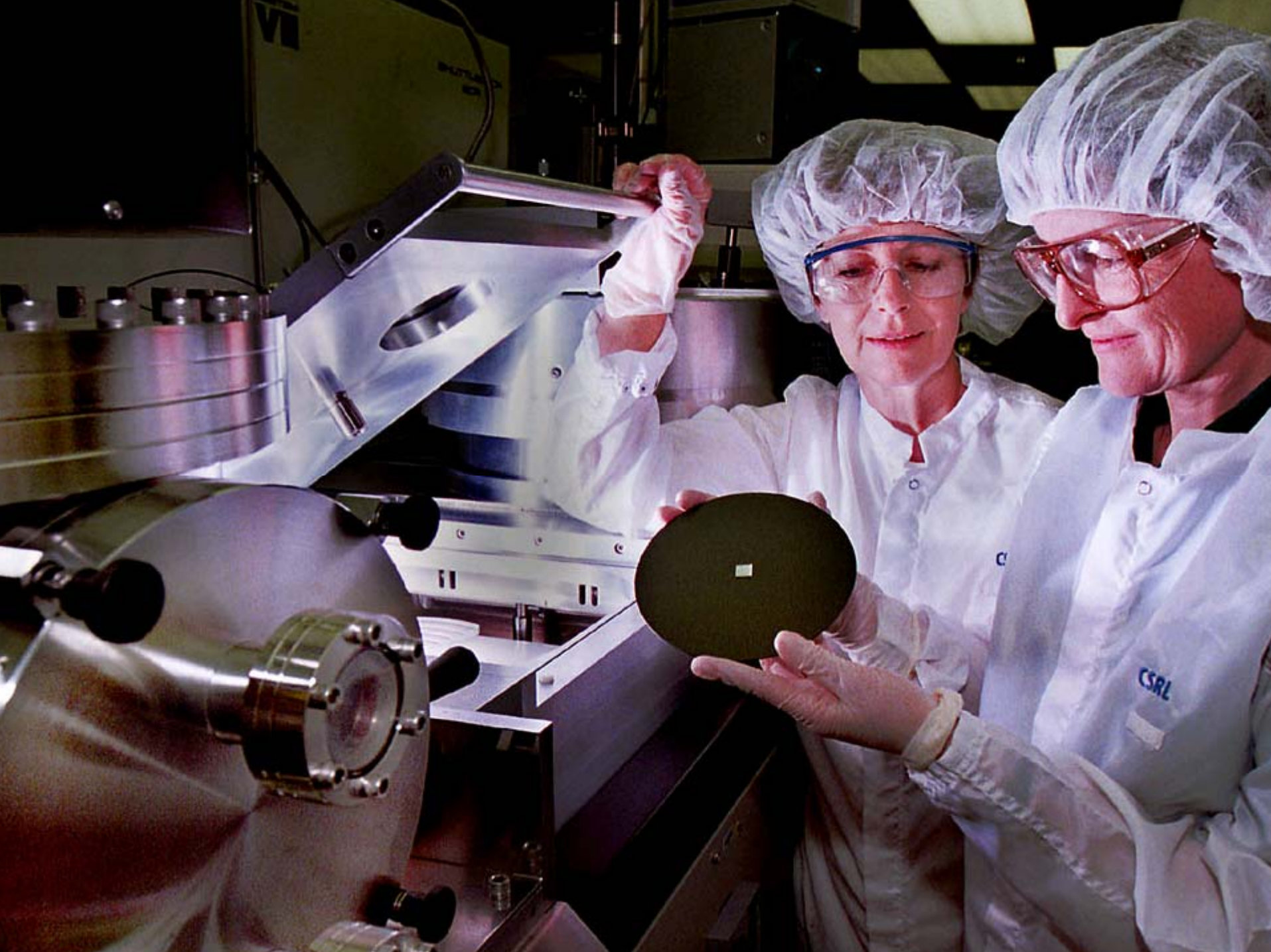
HA Trouble Shooting – The Usual Suspects

- Check for DNS
 - Make sure you can resolve the short hostname (without domain name) of each ESX host from each other ESX host in the cluster.
- Check length of Fully Qualified Domain Name (FQDN)
 - FQDN too long - make sure the fully qualified domain name of all hosts is less than 29 characters.
- Check entries in /etc/hosts and /etc/resolv.conf
 - Put hostname.FQDN there as well
 - Ping each host, from each host, by hostname, by FQDN.
 - Ping the VC Server from each host
- Check Log files in /opt/LGTOaam512/log
 - Files to look out for:
 - aam_config_util_listprimaries.log – Shows the primary hosts
 - aam_config_util_listnodes.log

The Importance of Planning

- Plan effectively
- Assess workloads and their requirements
 - Profile CPU/Memory/IO on existing platforms
 - CPU/Memory/IO Mixture has to be balanced









Conclusion

- Understand your business and user requirements
 - Assess workloads of your applications
 - Understand and know clearly where you want to go with it
- HA / DRS are great enablers for your business environment.
- Understand HA / DRS
 - Strengths
 - Limitations
 - The trade-offs involved

Most importantly, have FUN on the way!!!!

Acknowledgements and thanks to ..

Simon Clews, Abheek Anand, Puneet Zaroo, Sriya Santhanam, David Day, Martyn Storey.

No animals were harmed in the making of this presentation

Presentation Download

Please remember to complete your
session evaluation form
and return it to the room monitors
as you exit the session

The presentation for this session can be downloaded at
<http://www.vmware.com/vmtn/vmworld/sessions/>

Enter the following to download (case-sensitive):

Username: cbv_rep
Password: cbvfor9v9r

Some or all of the features in this document may be representative of feature areas under development. Feature commitments must not be included in contracts, purchase orders, or sales agreements of any kind. Technical feasibility and market demand will affect final delivery.

VMWORLD 2006



EXTRA SLIDES

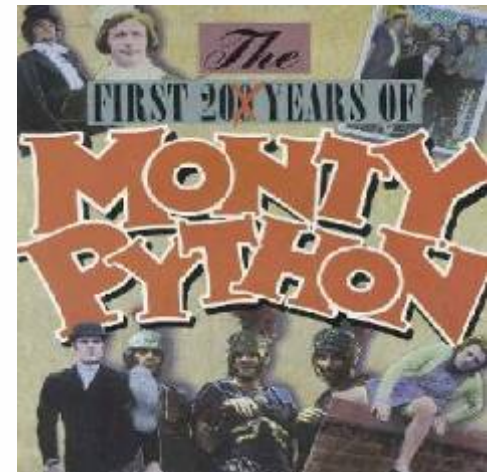
Split Brain Condition

- Scenario: One node of the cluster loses contact with the cluster.
- Each node has a node isolation verification address that it tries to ping to determine if it is connected to the network or isolated.
 - This verification address is the default gateway for the service console interface.
 - To change the address change `das.isolationaddress` parameter
 - If the host can't ping its own isolation address, then it knows it is isolated from the network, instead of thinking the other nodes have crashed.
 - If the node is isolated, the user can determine whether to leave the VM on or not (powering off releases the lock on disks).
 - By default, virtual machines are shut down on the isolated host in case of a host isolation incident.
- If a virtual machine continues to run on the isolated host, VMFS disk locking prevents it from being powered on elsewhere.
- VMware HA waits 15 seconds before deciding that a host is isolated
- VMware HA does not do a clean shutdown of the VM

HA Caveats

■ Incorrect Host Network Isolation

- Host network isolation detection occurs within 15 seconds .
- After failure is detected, all VMs are failed over to other hosts.
- If the network connection is restored before 12 seconds have elapsed, other hosts in the cluster do not treat this as a host failure, and the virtual machines remain powered on (on their original host).
- After 12 seconds, the clustering service on the isolated host shuts down and the virtual machines are powered off.
- Note: If the network connection is restored shortly after 12 seconds, the virtual machines are not started on other hosts because the original host is not considered to be isolated. Thus, the virtual machines are powered off but not failed over.



DRS Modes and Placement

Modes	Manual	Partially Automated	Fully Automated
Placement			
Initial Placement	VC displays a list of recommended hosts, with more suitable hosts higher on the list.	VC places the VM on the appropriate host	VC places the VM on the appropriate host
	Administrator is free to choose any host	Administrator does not choose	Administrator does not choose
VM Migration	VC does not take automatic actions to balance resources	VC does not take automatic actions to balance resources	VC places VMs that join the cluster on appropriate hosts and migrates running virtual machines between hosts as needed
	Summary page indicates that migration recommendations are available	Summary page indicates that migration recommendations are available	No migration recommendations are made
	Migration page displays recommendations	Migration page displays recommendations	VI Client's Migration Tab displays History of migration