Avoiding the 16 Biggest DA & DRS Configuration Mistakes

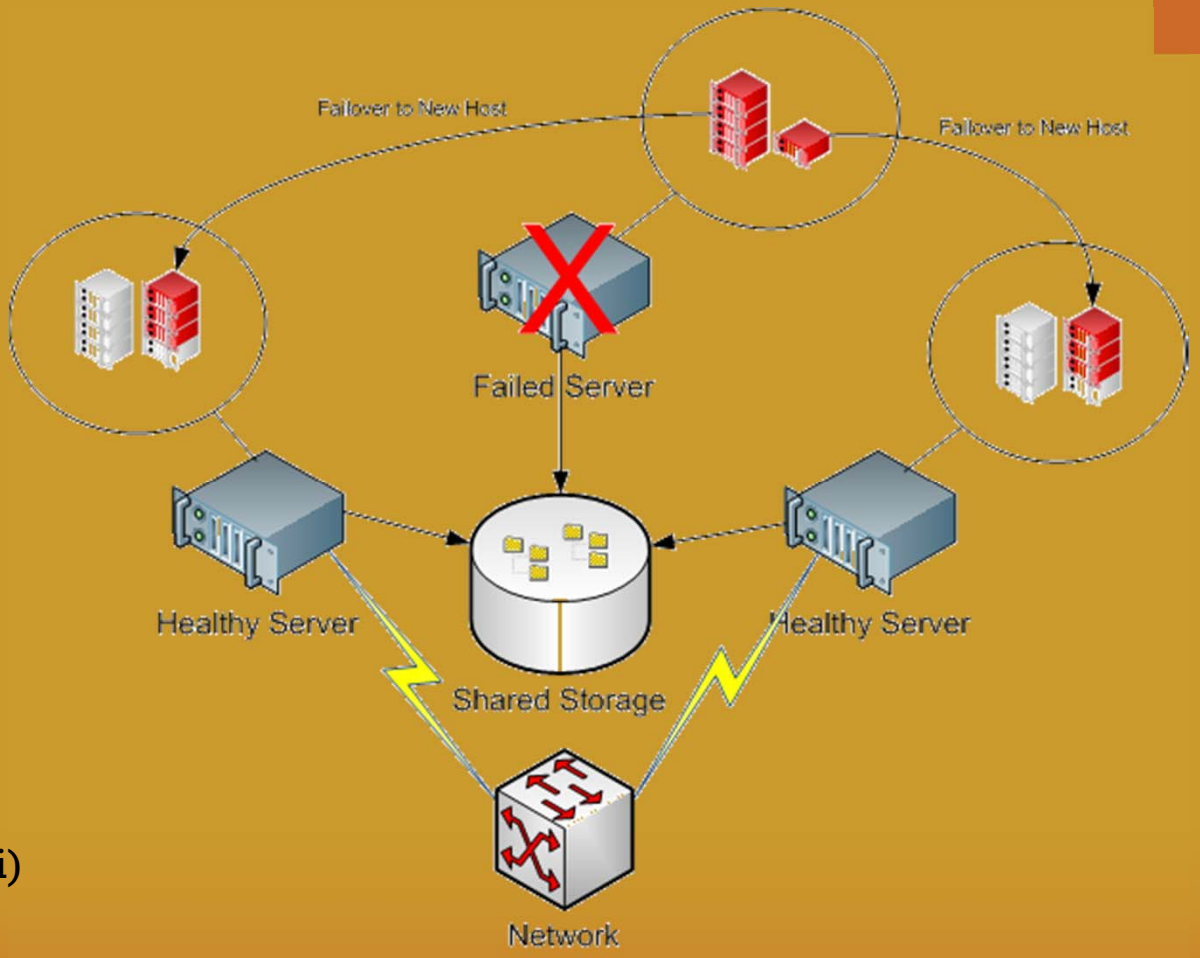
Greg Shields
Senior Partner and Principal Technologist,
Concentrated Technology, LLC
http://ConcentratedTech.com

concentrated technology

Reality Moment:
HA/DRS Solve Two Problems

Reality Moment:
HA/DRS Solve Two Problems

Problem #1:
Protection from
Unplanned Host
Downtime



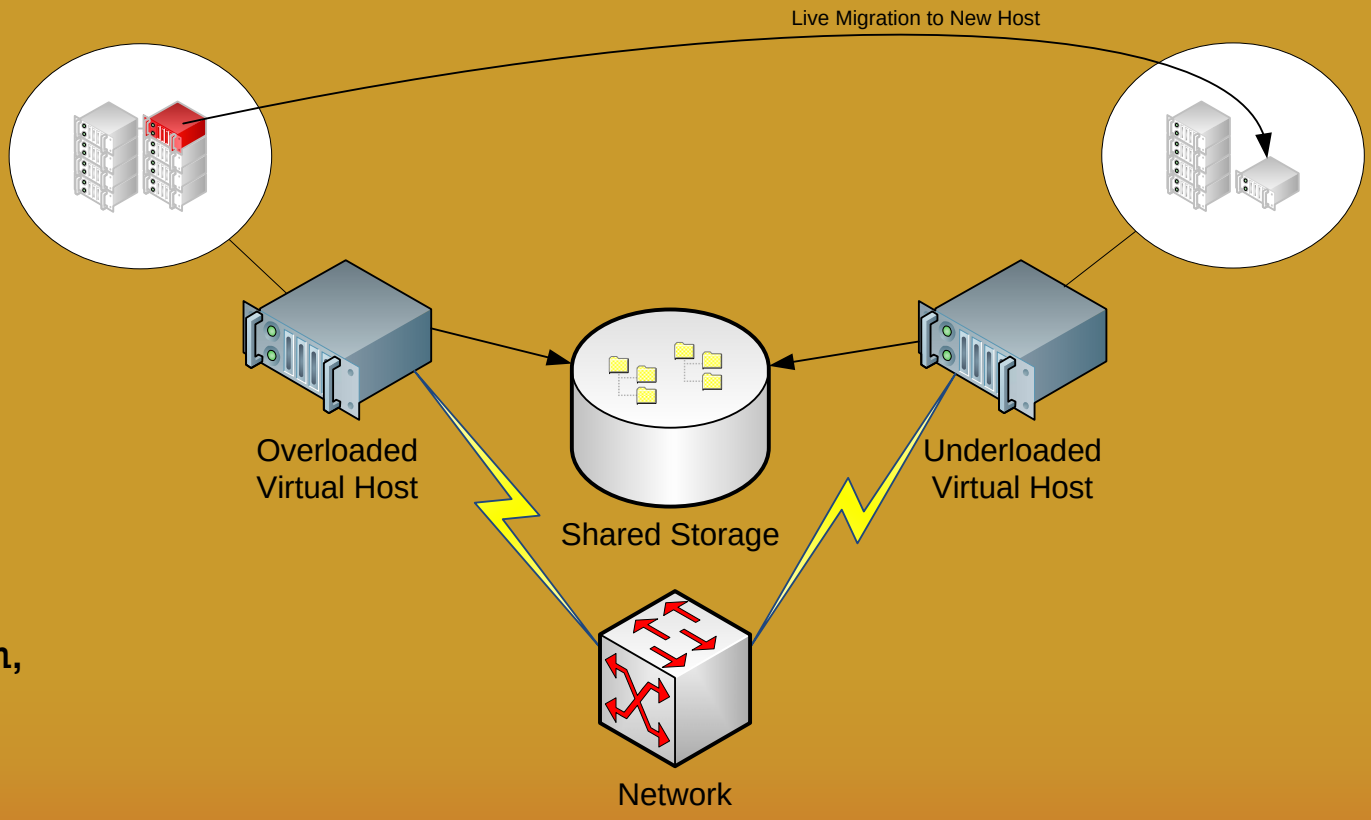
(This is relatively rare)

(They will get even more
rare as we migrate to ESXi)

concentrated technology

Reality Moment:
HA/DRS Solve Two Problems

Problem #2:
Load Balancing of VM &
Host Resources

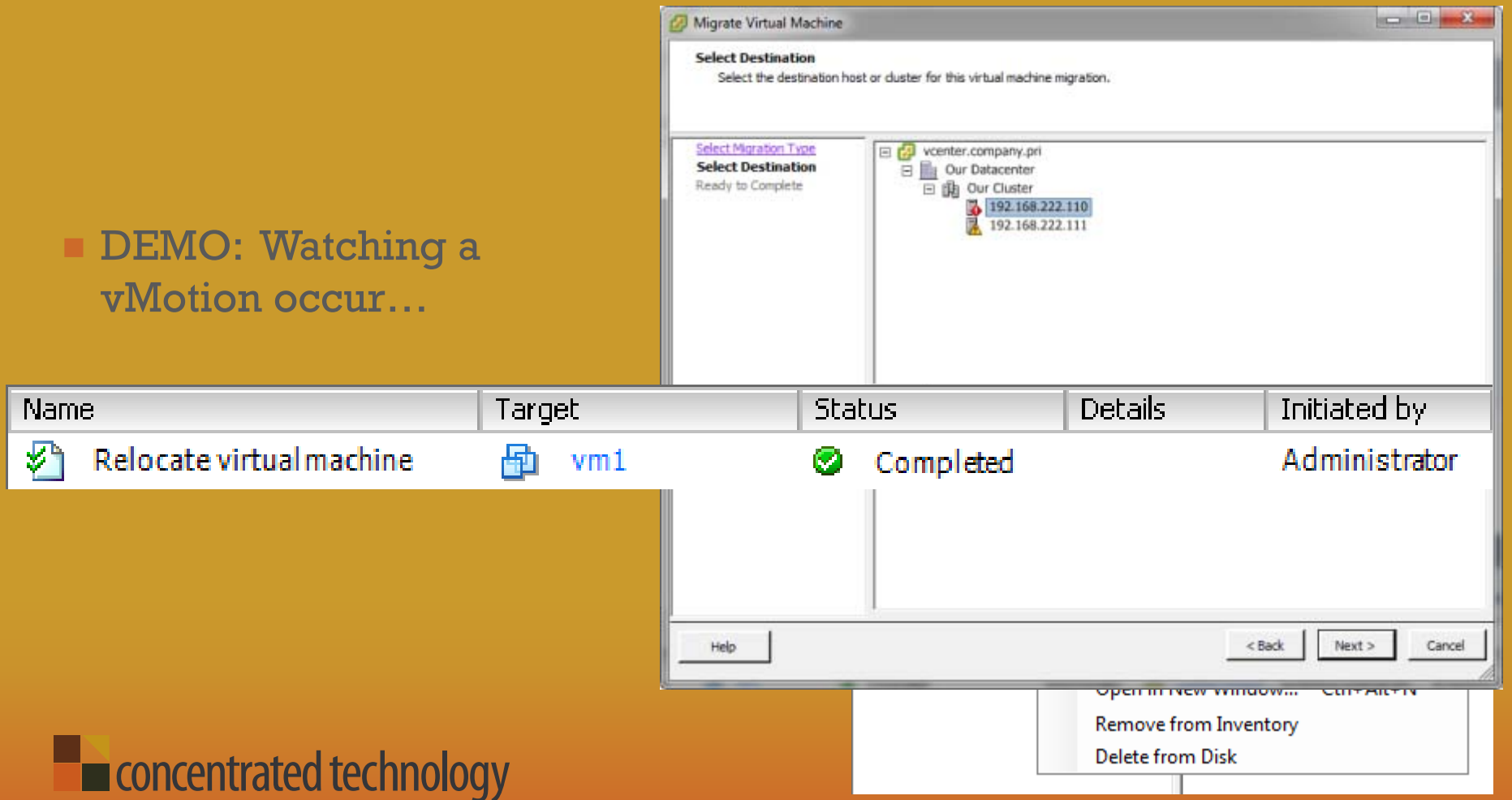


(Much more common,
where enabled)

concentrated technology

Contrary to Popular Belief…

- …seeing the actual vMotion process occur isn't all that sexy.
- DEMO: Watching a vMotion occur…



concentrated technology

Useful, However…

- …is recognizing where bad HA and DRS settings impact vMotion's ability to do its job.
  - A surprising number of environments have configured HA/DRS settings incorrectly.
  - Some do so because of hardware constraints.
  - Others have not designed architecture with HA/DRS in mind.
  - Even others have introduced problems as they scale upwards.

Useful, However…

- …is recognizing where bad HA and DRS settings impact vMotion's ability to do its job.
  - A surprising number of environments have configured HA/DRS settings incorrectly.
  - Some do so because of hardware constraints.
  - Others have not designed architecture with HA/DRS in mind.
  - Even others have introduced problems as they scale upwards.

- What follows are 16 big mistakes you'll want to avoid as you build or scale your HA/DRS cluster(s).

Big Mistake #1: Not Planning for HW Change

- Successful vMotion requires similar processors.
  - Processors must be from the same manufacturer.
    - No Intel-to-AMD or AMD-to-Intel vMotioning.
  - Processors must be of a proximate families.
  - This bites people a-few-years-down-the-road all the time!

Legend: CPU Architecture
AMD Opteron™ Generation 1 (Rev. E)
AMD Opteron™ Generation 2 (Rev. F)
AMD Opteron™ Generation 3 (Greyhound)

	EVC Cluster Baseline			
AMD Opteron™ Host CPU Model	AMD Opteron™ Generation 1 EVC Mode	AMD Opteron™ Generation 2 EVC Mode	AMD Opteron™ Generation 3 EVC Mode	AMD Opteron™ Generation 3 (no 3DNow!™) EVC Mode
1xx Series	Yes	No	No	No
2xx Series	Yes	No	No	No
8xx Series	Yes	No	No	No
12xx Series	Yes	Yes	No	No
22xx Series	Yes	Yes	No	No
82xx Series	Yes	Yes	No	No
13xx Series	Yes	Yes	Yes	Yes
23xx Series	Yes	Yes	Yes	Yes
24xx Series	Yes	Yes	Yes	Yes
83xx Series	Yes	Yes	Yes	Yes
84xx Series	Yes	Yes	Yes	Yes
61xx Series	Yes	Yes	Yes	Yes
41xx Series	Yes	Yes	Yes	Yes

concentrated technology

Big Mistake #1: Not Planning for HW Change

Legend: CPU Architecture
Intel® Core™2 (Merom)
Intel® 45nm Core™2 (Penryn)
Intel® Core™ i7 (Nehalem)
Intel® 32nm Core™ i7 (Westmere)

	EVC Cluster Baseline			
Intel® Xeon® Host CPU Model	Intel® Xeon® Core™2 EVC Mode	Intel® Xeon® 45nm Core™2 EVC Mode	Intel® Xeon® Core™ i7 EVC Mode	Intel® Xeon® 32nm Core™ i7 EVC Mode
30xx Series	Yes	No	No	No
32xx Series	Yes	No	No	No
51xx Series	Yes	No	No	No
53xx Series	Yes	No	No	No
72xx Series	Yes	No	No	No
73xx Series	Yes	No	No	No
31xx Series	Yes	Yes	No	No
33xx Series	Yes	Yes	No	No
52xx Series	Yes	Yes	No	No
54xx Series	Yes	Yes	No	No
74xx Series	Yes	Yes	No	No
35xx Series	Yes	Yes	Yes	No
55xx Series	Yes	Yes	Yes	No
34xx Lynnfield Series	Yes	Yes	Yes	No
34xx Clarkdale Series without AES and PCLMULQDQ	Yes	Yes	Yes	No
65xx Series	Yes	Yes	Yes	No
75xx Series	Yes	Yes	Yes	No
i3/i5 Clarkdale Series without AES and PCLMULQDQ	Yes	Yes	Yes	No
34xx Clarkdale Series with AES and PCLMULQDQ	Yes	Yes	Yes	Yes
56xx Series	Yes	Yes	Yes	Yes
36xx Series	Yes	Yes	Yes	Yes
i3/i5 Clarkdale Series with AES and PCLMULQDQ	Yes	Yes	Yes	Yes

concentrated technology

Big Mistake #1: Not Planning for HW Change

- As a virtual environment ages, hardware is refreshed and new hardware is added.
  - Refreshes sometimes create "islands" of vMotion capability

Big Mistake #1: Not Planning for HW Change

- As a virtual environment ages, hardware is refreshed and new hardware is added.
  - Refreshes sometimes create "islands" of vMotion capability
- How can we always vMotion between computers?

Big Mistake #1: Not Planning for HW Change

- As a virtual environment ages, hardware is refreshed and new hardware is added.
  - Refreshes sometimes create "islands" of vMotion capability
- How can we always vMotion between computers?
  - You can always refresh all hardware at the same time (Har!)

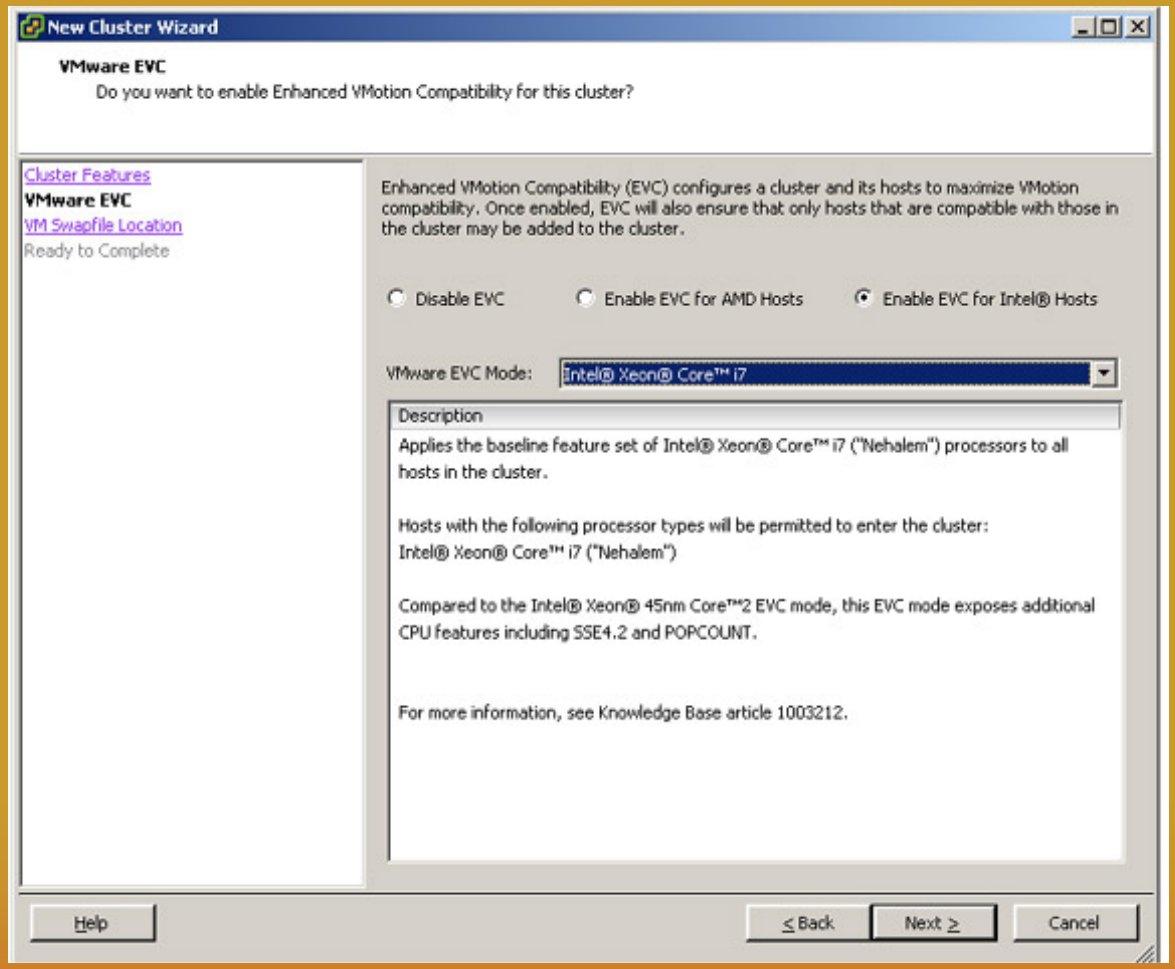
Big Mistake #1: Not Planning for HW Change

- As a virtual environment ages, hardware is refreshed and new hardware is added.
  - Refreshes sometimes create "islands" of vMotion capability
- How can we always vMotion between computers?
  - You can always refresh all hardware at the same time (Har!)
  - You can cold migrate, with the machine powered down. This always works, but ain't all that friendly.

Big Mistake #1: Not Planning for HW Change

- As a virtual environment ages, hardware is refreshed and new hardware is added.
  - Refreshes sometimes create "islands" of vMotion capability
- How can we always vMotion between computers?
  - You can always refresh all hardware at the same time (Har!)
  - You can cold migrate, with the machine powered down. This always works, but ain't all that friendly.
  - You can use vMotion Enhanced Compatibility Mode to manage your vMotion-ability. Create islands as individual clusters.
- SOLUTION: vMotion EVC

Big Mistake #1:
Not Planning for HW Change



concentrated technology

Big Mistake #2: Not Planning for svMotion

- Storage vMotion has some special requirements.
  - Virtual machines with snapshots cannot be svMotioned.
  - Virtual machine disks must be persistent mode or RDMs.
  - The host must have sufficient resources to support two instances of the VM running concurrently for a brief time.
  - The host must have a vMotion license, and be correctly configured for vMotion.
  - The host must have access to both the source and target datastores.

Big Mistake #3: Not Enough Cluster Hosts

- You cannot change the laws of physics.
  - For HA to failover a VM, there must be resources available elsewhere in the cluster.
  - These resources must be set aside. Reserved. "Wasted".

Big Mistake #3: Not Enough Cluster Hosts

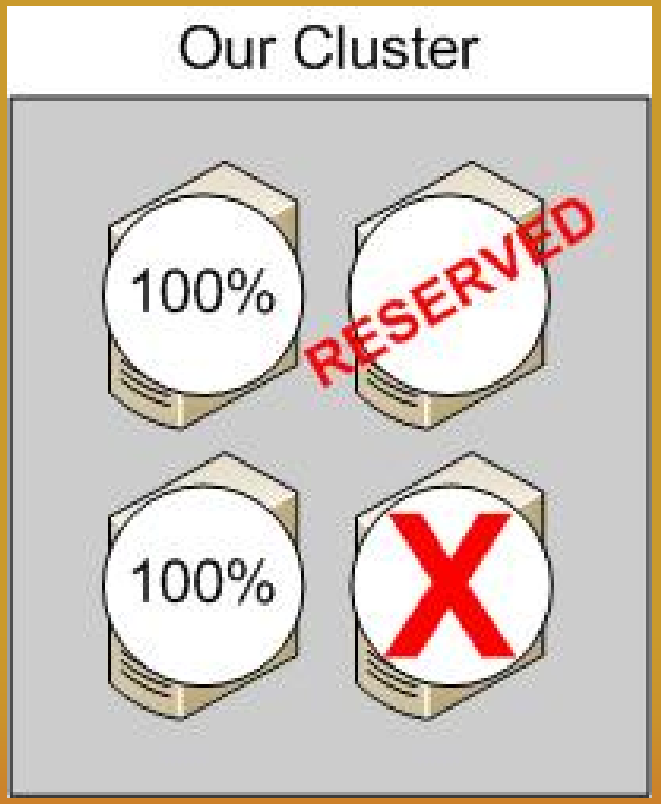
- You cannot change the laws of physics.
  - For HA to failover a VM, there must be resources available elsewhere in the cluster.
  - These resources must be set aside. Reserved. "Wasted".

- Many environments don't plan for cluster reserve when designing their clusters.
  - Nowhere for VMs to go…

concentrated technology

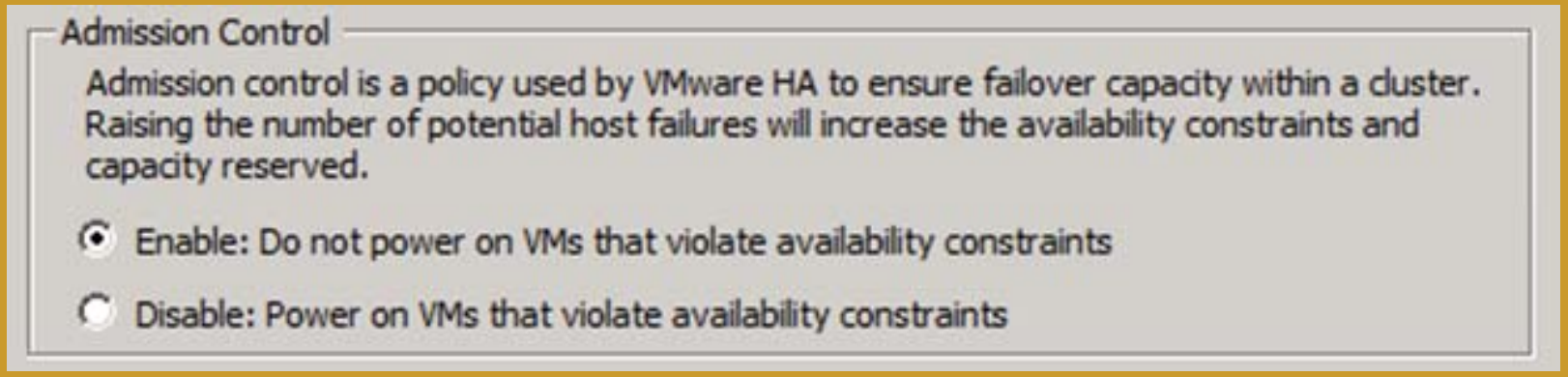
Big Mistake #3: Not Enough Cluster Hosts

- A fully-prepared cluster must set aside one full server's worth of resources in preparation for HA.



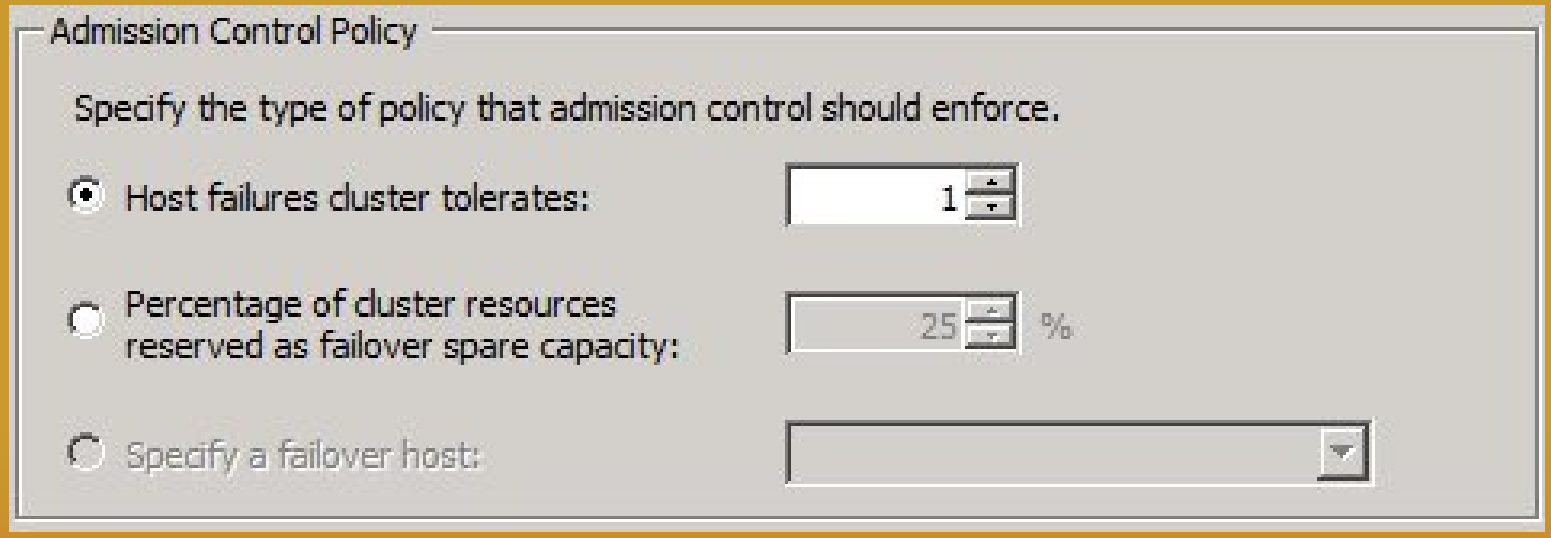
Big Mistake #3: Not Enough Cluster Hosts

- A fully-prepared cluster must set aside one full server's worth of resources in preparation for HA.
  - This is done in your Admission Control Policy.
  - First, Enable Admission Control.



Big Mistake #3: Not Enough Cluster Hosts

- A fully-prepared cluster must set aside one full server's worth of resources in preparation for HA.
  - This is done in your Admission Control Policy.
  - Then, set Host failures cluster tolerates to 1 (or more).



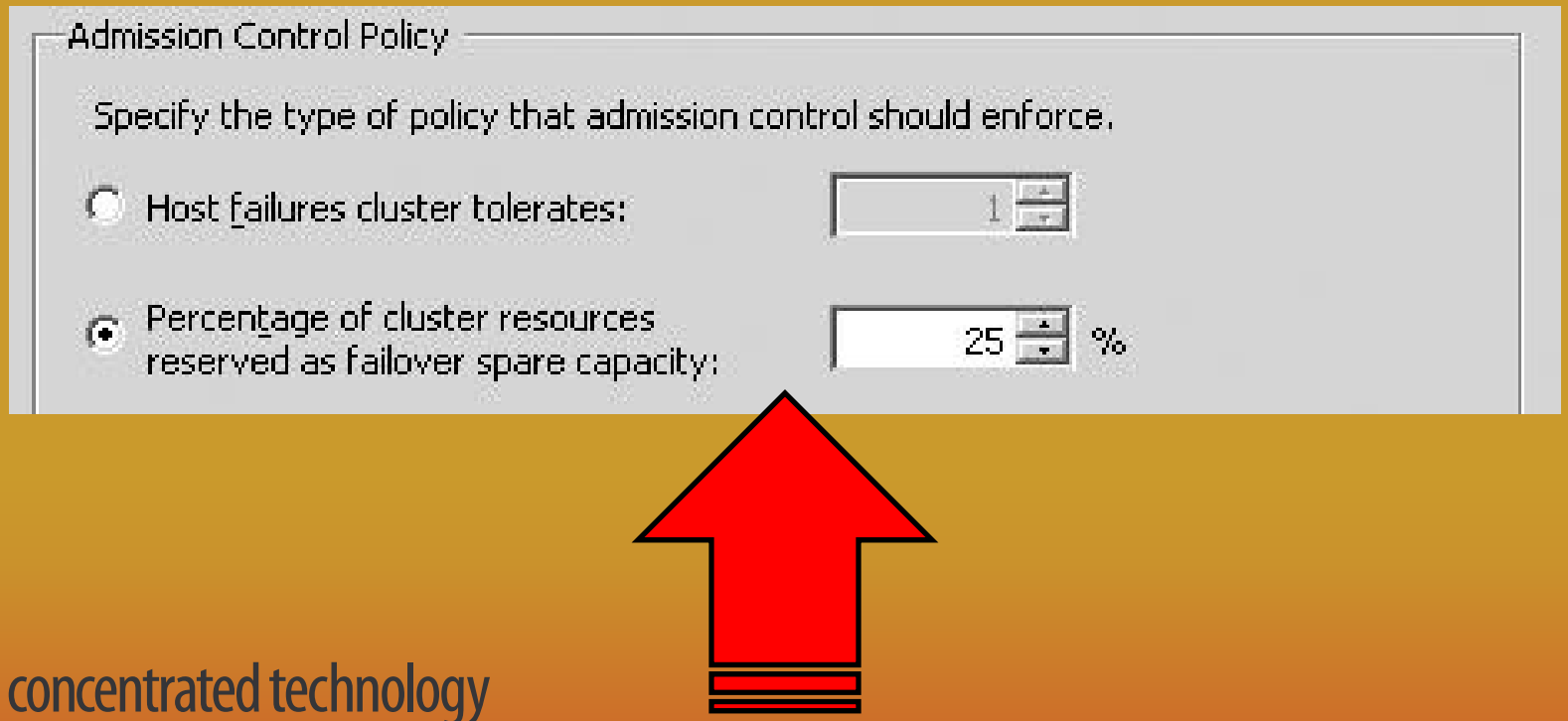
concentrated technology

Big Mistake #4: Setting Host Failures Cluster Tolerates to 1.

- Setting Host failures cluster tolerates to 1 may unnecessarily set aside too many resources.
  - Not all your VMs are Priority One.
  - Some VMs can stay down if a host dies.
  - Setting aside a full host is wasteful, particularly when your number of hosts is small.

Big Mistake #4: Setting Host Failures Cluster Tolerates to 1.

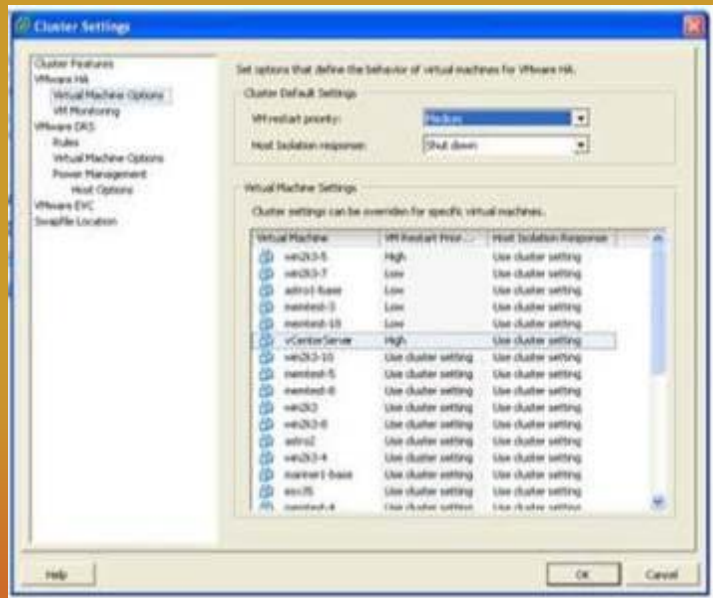
- Tune your level of waste with Percentage of cluster resources reserved as failover capacity.
  - Set this to a lower value than one server's contribution.



concentrated technology

Big Mistake #5: Not Prioritizing VM Restart.

- VM Restart Priority is one of those oft-forgotten settings.
  - A default setting is configured when you enable HA.
  - Per-VM settings must be configured for each VM.
- These settings are most important during an HA event.
  - Come into play when Percentage policy is enabled.



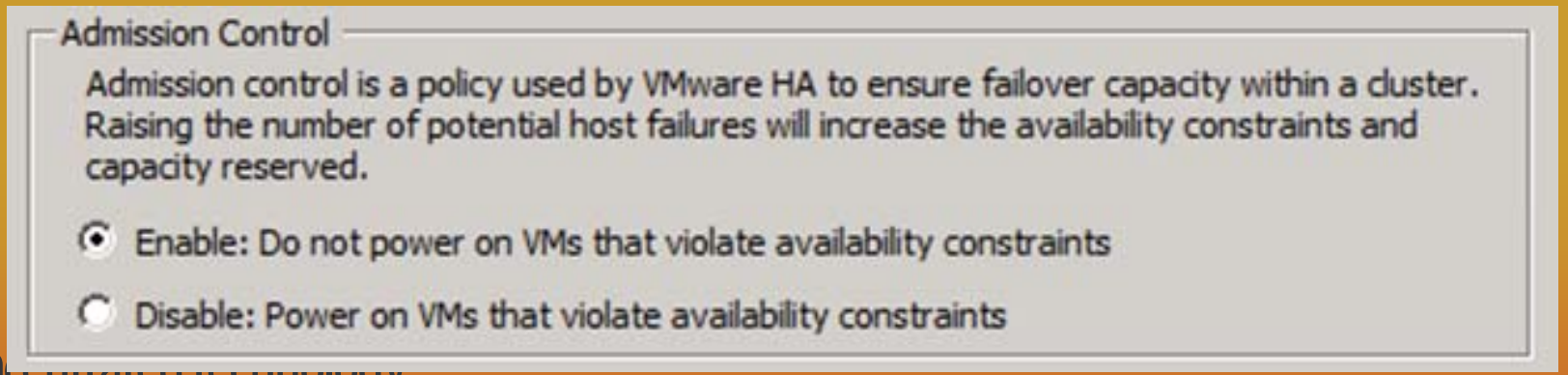
concentrated technology

Big Mistake #6: Disabling Admission Control

- Every cluster with HA enabled will have "waste".
  - Some enterprising young admins might enable HA but disable Admission Control.
  - "A-ha," they might say, "This gives me all the benefits of HA but without the waste!"

Big Mistake #6: Disabling Admission Control

- Every cluster with HA enabled will have "waste".
  - Some enterprising young admins might enable HA but disable Admission Control.
  - "A-ha," they might say, "This gives me all the benefits of HA but without the waste!"
  - They're wrong. Squeezing VMs during an HA event can cause downstream performance effects as hosts begin swapping.
  - Never disable Admission Control.



Big Mistake #7: Not Updating Percentage Policy

- The Percentage policy may need to be adjusted as your cluster size changes.
  - Adding servers can change the percentage of resources that must be set aside.
  - Take a look at adjusting percentage every time you add servers.

Big Mistake #7: Not Updating Percentage Policy

- The Percentage policy may need to be adjusted as your cluster size changes.
  - Adding servers can change the percentage of resources that must be set aside.
  - Take a look at adjusting percentage every time you add servers.
- Host failures cluster tolerates does not require adjusting.
  - No matter how many hosts you have, this policy setting will always set aside one server's worth of resources.
  - Yet here danger lies…

Big Mistake #8: Buying Dissimilar Servers

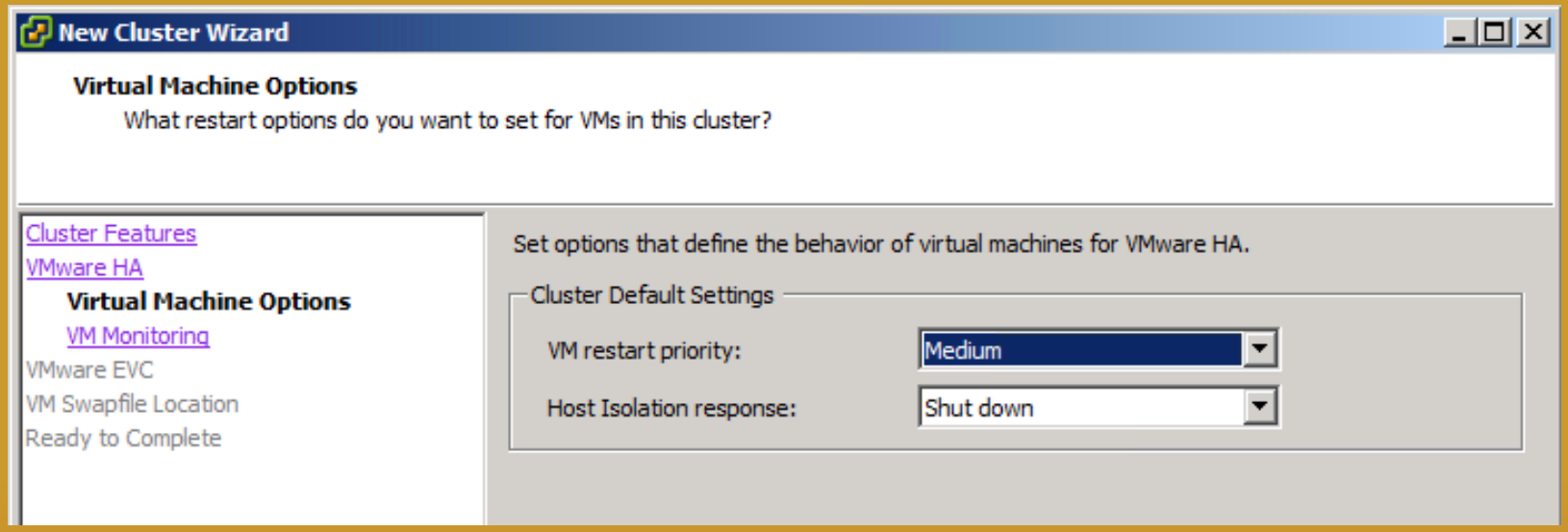
- Host failures cluster tolerates sets aside an amount of resources that are needed to protect every server.
  - This means that any fully-loaded server will be HA-protected.
  - Including, your biggest server!

Big Mistake #8: Buying Dissimilar Servers

- Host failures cluster tolerates sets aside an amount of resources that are needed to protect every server.
  - This means that any fully-loaded server will be HA-protected.
  - Including, your biggest server!
- Thus, Host failures cluster tolerates must set aside resources equal to your biggest server!
  - If you buy three small servers and one big server, you're wasting even more resources!
  - This is necessary to protect all resources, but wasteful if your procurement buys imbalanced servers.

Big Mistake #9: Host Isolation Response

- Host isolation response instructs the cluster what to do when a host loses connectivity, but hasn't failed.
  - That host is isolated from the cluster, but its VMs still run.



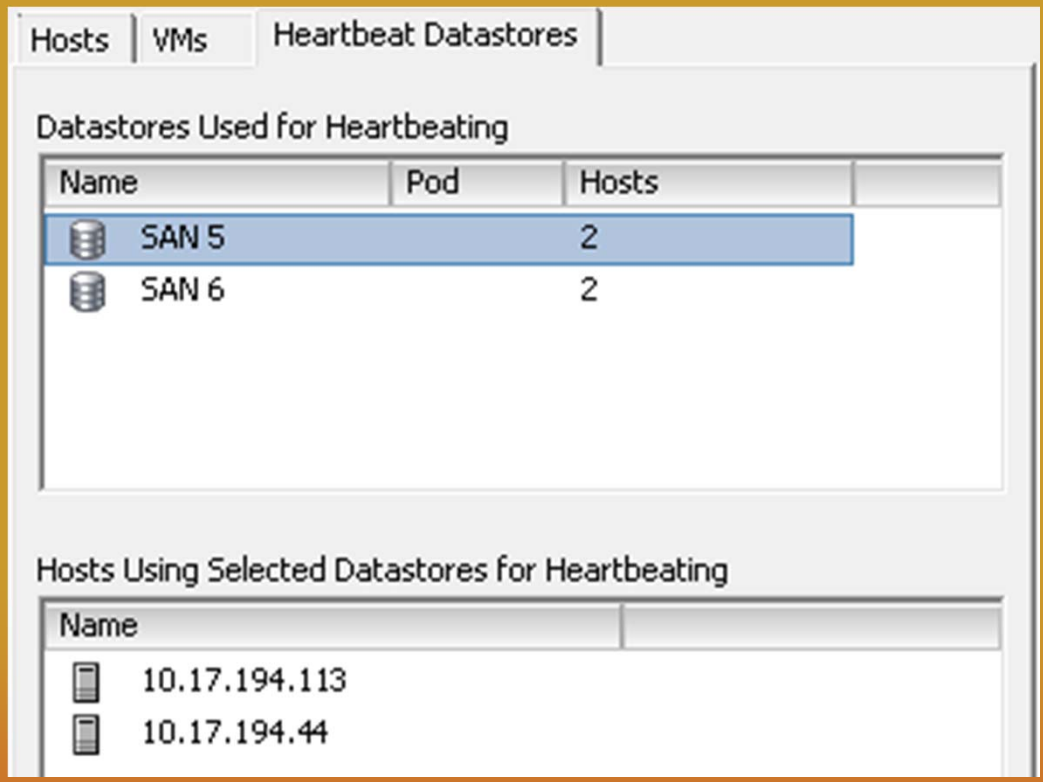
concentrated technology

Big Mistake #9: Host Isolation Response

- Host isolation response instructs the cluster what to do when a host loses connectivity, but hasn't failed.
  - That host is isolated from the cluster, but its VMs still run.
- Three options available:
  Leave powered on / Power off / Shut down.
  - VMs that remain powered on cannot be managed by surviving cluster hosts. Egad, its like split brain in reverse!
  - VMFS locks prevent them from being evacuated to "good" host.
  - Suggestion: Shut Down will gracefully down a VM, but will release VMFS locks so VM can be again managed correctly.
  - Adjust per-VM settings for important VMs.

Ta-Da!, v5.0!: Host Isolation Response

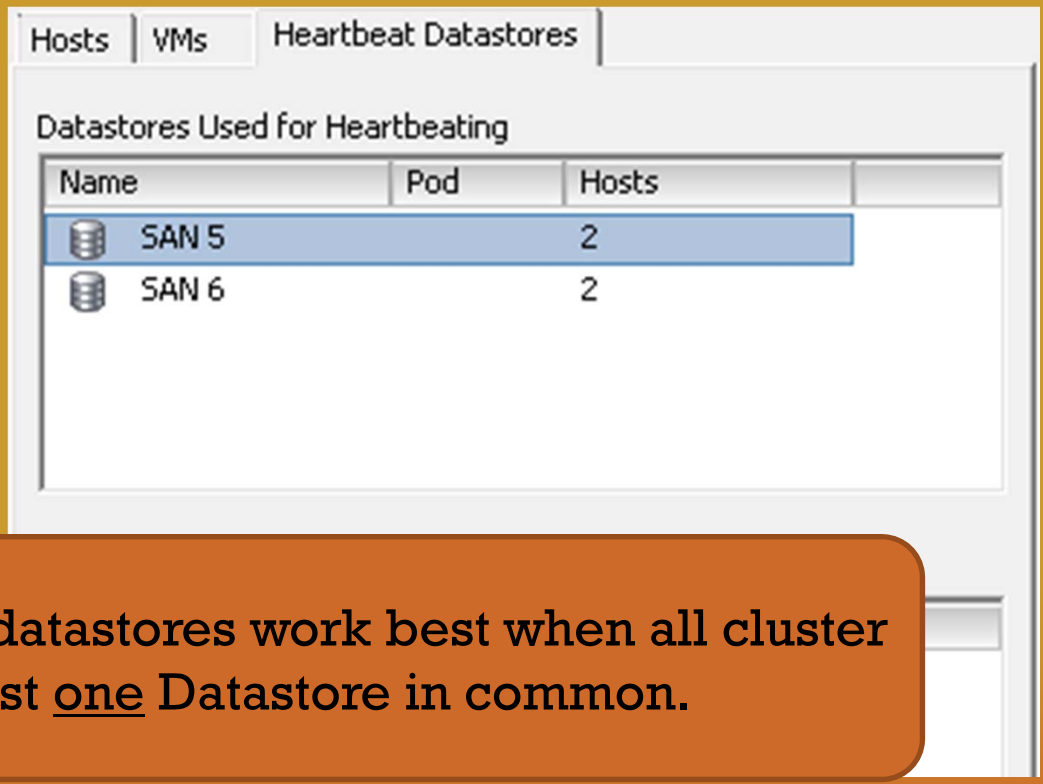
- Heartbeat Datastores, New in v5.0.
  - vSphere HA in v5.0 can now use the storage subsystem for communication.
  - Adds redundancy.
  - Used as communication channel only when the management network is lost
  - Such as in the case of isolation or network partitioning.



concentrated technology

Ta-Da!, v5.0!: Host Isolation Response

- Heartbeat Datastores, New in v5.0.
  - vSphere HA in v5.0 can now use the storage subsystem for communication.
  - Adds redundancy.
  - Used as communication channel only when the management network is lost
  - Such as in the case of isolation or network partitioning.



IMPORTANT: Heartbeat datastores work best when all cluster hosts share at least one Datastore in common.

Big Mistake #10: Overdoing the Reservations, Limits, and Affinities

- HA may not consider these "soft affinities" at failover.
- However, they will be invoked after HA has taken its corrective action.
  - Reservations and limits can constrain resulting calculations.
  - Affinities add more constraints, particularly in smaller clusters.

Big Mistake #10: Overdoing the Reservations, Limits, and Affinities

- HA may not consider these "soft affinities" at failover.
- However, they will be invoked after HA has taken its corrective action.
  - Reservations and limits can constrain resulting calculations.
  - Affinities add more constraints, particularly in smaller clusters.
- Use shares over reservations and limits where possible.
  - Shares balance VM resource demands rather than setting hard thresholds. Less of an impact on DRS, and thus HA.
  - Limit the use of affinities.

Big Mistake #11: Doing Memory Limits at All!

- Don't assign Memory Limits. Ever.
  - Let's say you assign a VM 4G of memory.
  - Then, you set a 1G memory limit on that VM.
  - That VM can now never use more than 1G of physical RAM. All other memory needs above 1G must come from swap or ballooning.

Big Mistake #11: Doing Memory Limits at All!

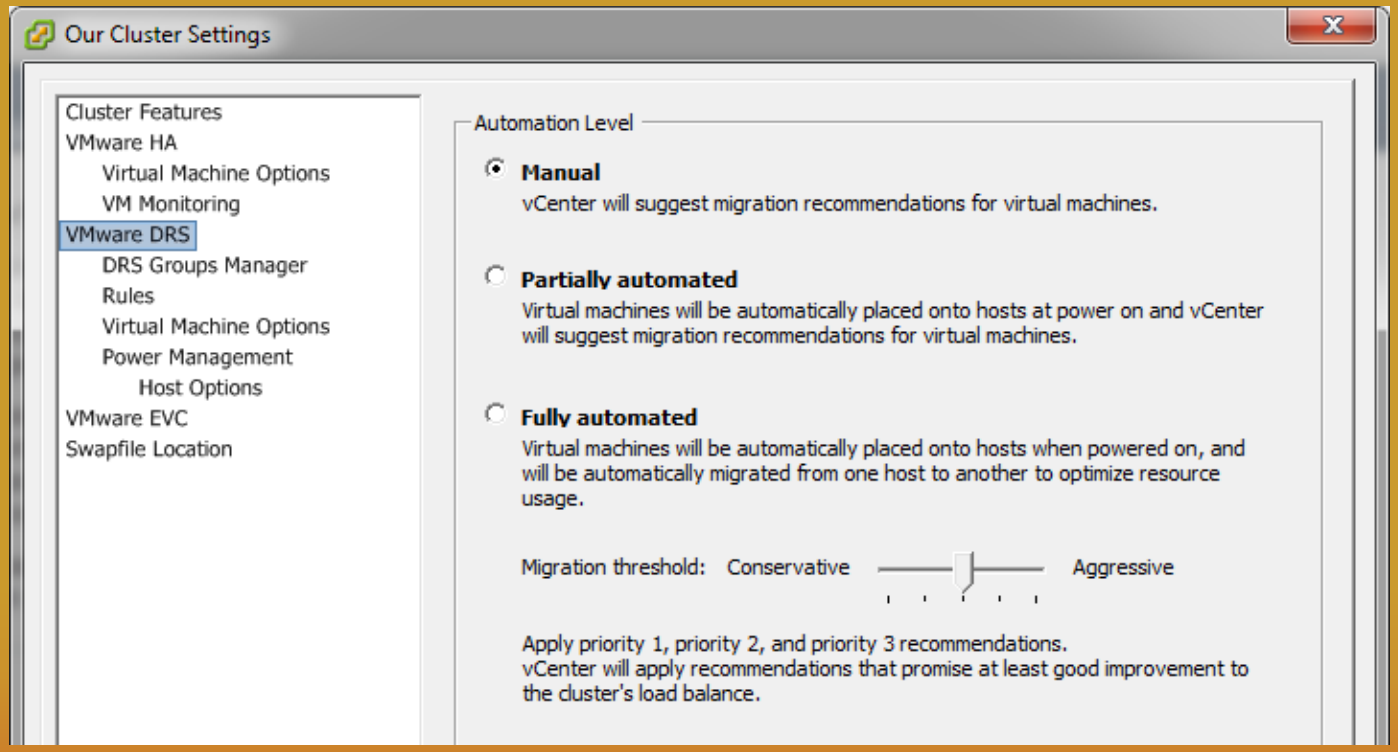
- Don't assign Memory Limits. Ever.
  - Let's say you assign a VM 4G of memory.
  - Then, you set a 1G memory limit on that VM.
  - That VM can now never use more than 1G of physical RAM. All other memory needs above 1G must come from swap or ballooning.
- Always best to limit memory as close to the affected application as possible.
  - Limit VM memory, but even better to throttle application memory.
  - Example:
    Limiting SQL > Limiting Windows > Limiting the VM > Limiting the Hypervisor.

concentrated technology

Big Mistake #12: Thinking You're Smarter than DRS

('cuz you're not!)

- No human alive can watch every VM counter as well as a monitor and a mathematical formula.



concentrated technology

Big Mistake #13: Not Understanding DRS' Equations.

- DRS is like a table that sits atop a single leg at its center.
  - Each side of that table represents a host in your cluster.
  - That leg can only support the table when all sides are balanced.
  - DRS' job is to relocate VMs to ensure the table stays balanced.

Big Mistake #13: Not Understanding DRS' Equations.

- DRS is like a table that sits atop a single leg at its center.
  - Each side of that table represents a host in your cluster.
  - That leg can only support the table when all sides are balanced.
  - DRS' job is to relocate VMs to ensure the table stays balanced.
- Every five minutes a DRS interval is invoked.
  - During that interval DRS analyses resource utilization counters on every host.
  - It plugs those counters into this equation:

$$\frac{\sum(VM\ Entitlements)}{Host\ Capacity}$$

Big Mistake #13: Not Understanding DRS' Equations.

- VM entitlements
  - CPU resource demand and memory working set.
  - CPU and memory reservations or limits.
- Host Capacity
  - Summation of CPU and memory resources, minus…
    - VMKernel and Service Console overhead
    - Reservations for HA Admission Control
    - 6% "extra" reservation

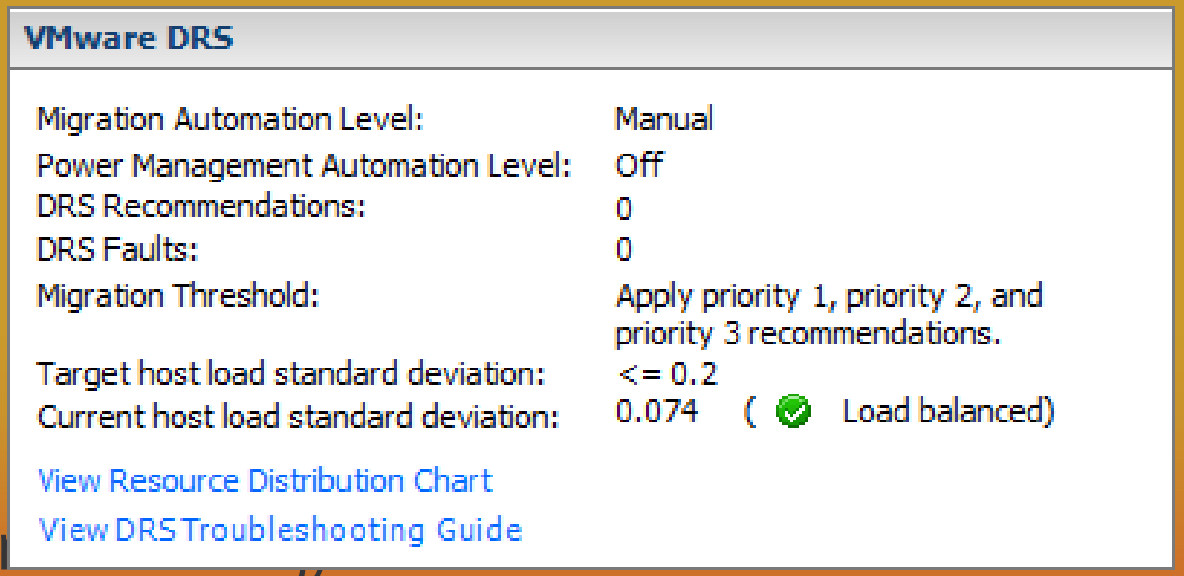
$$\frac{\sum(VM\ Entitlements)}{Host\ Capacity}$$

Big Mistake #13: Not Understanding DRS' Equations.

- A statistical mean and standard deviation can then be calculated.
  - Mean = Average load
  - Standard deviation = Average deviation from that load

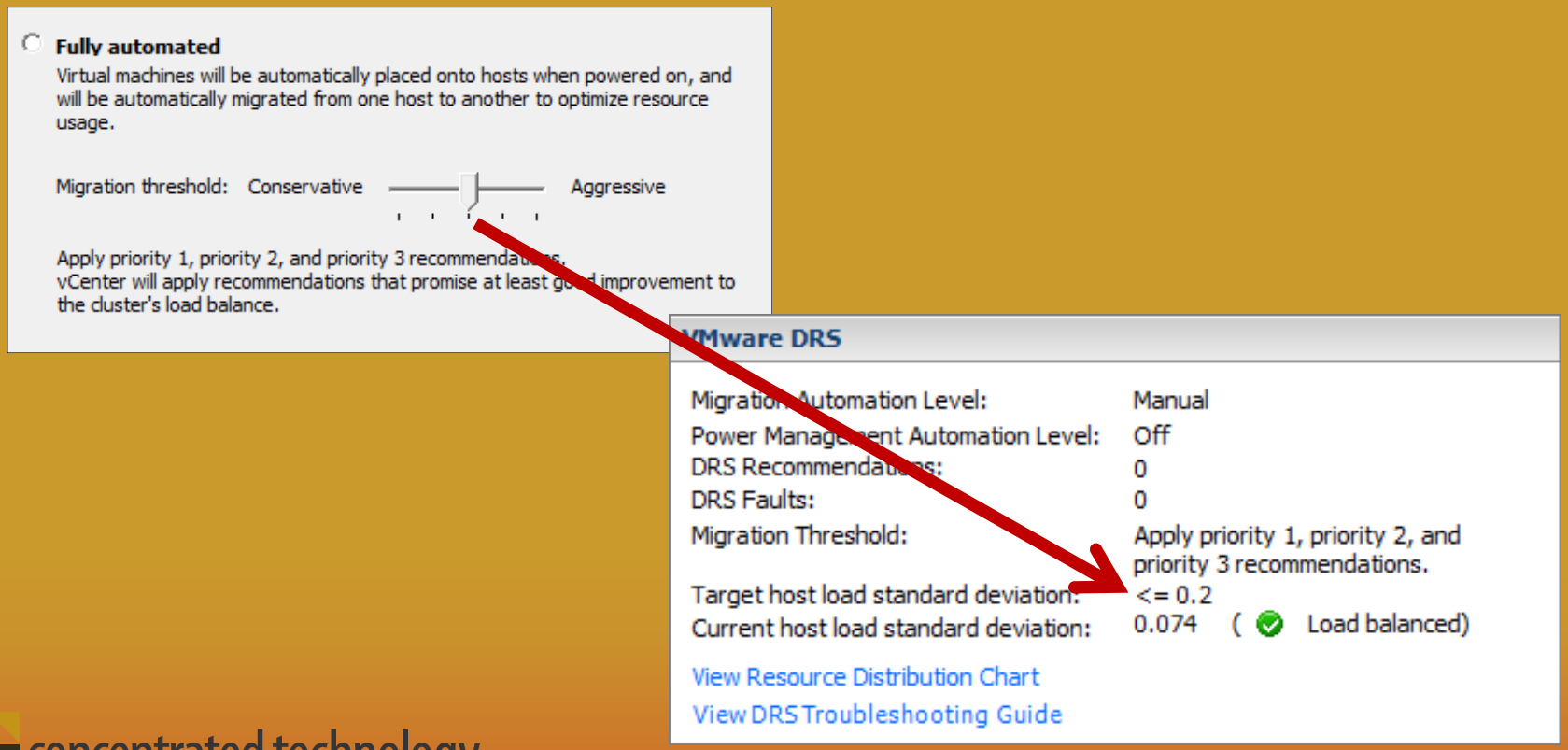
Big Mistake #13: Not Understanding DRS' Equations.

- A statistical mean and standard deviation can then be calculated.
  - Mean = Average load
  - Standard deviation = Average deviation from that load
- This value is the Current host load standard deviation.



Big Mistake #13: Not Understanding DRS' Equations.

- Your migration threshold slider value determines the Target host load standard deviation.



concentrated technology

Big Mistake #13: Not Understanding DRS' Equations.

- DRS then runs a series of migration simulations to see which VM moves will have the greatest impact on balancing.
  - For each simulated move, it calculates the resulting Current host load standard deviation.
  - Then, it plugs that value into this equation

$$6 - \left\lceil \frac{Current\ Host\ Load\ Standard\ Deviation}{.1} * \sqrt{\#\ Hosts\ in\ Cluster} \right\rceil$$

Big Mistake #13: Not Understanding DRS' Equations.

- The result is a priority number from 1 to 5.
  - Migrations that have a greater impact on rebalancing have a higher priority.
  - Your migration threshold determines which migrations are automatically done.

Migration Threshold:	Apply priority 1, priority 2, and priority 3 recommendations.

Big Mistake #14:
Being too Liberal.

Big Mistake #14: Being too Liberal.

- …with your Migration Threshold, of course.
  - Migrations with lower priorities have less of an impact on balancing our proverbial table.
  - But every migration takes time, resources, and effort to complete.
  - There is a tradeoff between perfect balance and the resource cost associated with getting to that perfect balance.

Big Mistake #14: Being too Liberal.

- …with your Migration Threshold, of course.
  - Migrations with lower priorities have less of an impact on balancing our proverbial table.
  - But every migration takes time, resources, and effort to complete.
  - There is a tradeoff between perfect balance and the resource cost associated with getting to that perfect balance.
- Remember: Priority 1 recommendations are mandatory.
  - These are all special cases:
    - Hosts entering maintenance mode or standby mode.
    - Affinity rules being violated.
    - Summation of VM reservations exceed host capacity.

Big Mistake #15: Too Many Cluster Hosts

- vSphere 4.1 clusters can handle up to 32 hosts and 3000 VMs.
  - However, each additional host/VM adds another simulation that's required during each DRS pass.
  - More hosts/VMs mean more processing for each pass.

Big Mistake #15: Too Many Cluster Hosts

- vSphere 4.1 clusters can handle up to 32 hosts and 3000 VMs.
  - However, each additional host/VM adds another simulation that's required during each DRS pass.
  - More hosts/VMs mean more processing for each pass.
- Some experts suggest DRS' "Sweet spot" is between 16 and 24 hosts per cluster. (Epping/Denneman, 2010)
  - Not too few ("waste"), and not too many ("simulation effort").
  - Rebalance hosts per cluster as you scale upwards!
  - Mind HA's needs when considering your "sweet spot".

Big Mistake #16: Creating Big VMs

- Back during the "hypervisor wars" one of VMware's big sales points was memory overcommit.
  - "ESX can overcommit memory! Hyper-V can't!"
  - So, many of us used it.

Big Mistake #16: Creating Big VMs

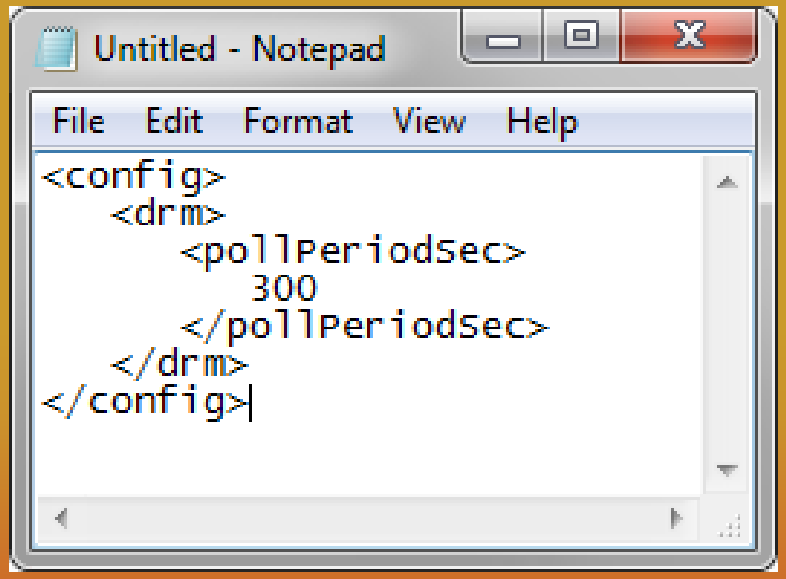
- Back during the "hypervisor wars" one of VMware's big sales points was memory overcommit.
  - "ESX can overcommit memory! Hyper-V can't!"
  - So, many of us used it.
- Overcommittment creates extra work for the hypervisor.
  - Ballooning, host memory swapping, page table sharing, etc.
  - That work is unnecessary when memory is correctly assigned.
- Assign the right amount of memory (and as few processors as possible) to your VMs.
  - Creating "big VMs" also impacts DRS' load balancing abilities.
  - Fewer options for balancing bigger VMs.

Easter Egg: Change DRS Invocation Frequency

- You can customize how often DRS will automatically take its own advice.
  - I wish my wife had this setting…

Easter Egg: Change DRS Invocation Frequency

- You can customize how often DRS will automatically take its own advice.
  - I wish my wife had this setting…
- On your vCenter Server, locate C:\Users\All Users\Application Data\VMware\VMware VirtualCenter\vpxd.cfg
- Add in the following lines (appropriately!):



concentrated technology

Things to Remember…
after the Beers…

concentrated technology

Things to Remember… after the Beers…

- For the love of <your preferred deity>, Turn on HA/DRS!
  - But only if you have enough hardware!
  - You've already paid for it.
  - It is smarter than you.

Things to Remember…
after the Beers…

- For the love of <your preferred deity>,
  Turn on HA/DRS!
  - But only if you have enough hardware!
  - You've already paid for it.
  - It is smarter than you.
- Understand why your VMs move around.
  - Make sure that you've got the connected resources they need on every host!

Things to Remember…
after the Beers…

- For the love of <your preferred deity>,
  Turn on HA/DRS!
  - But only if you have enough hardware!
  - You've already paid for it.
  - It is smarter than you.
- Understand why your VMs move around.
  - Make sure that you've got the connected resources they need on every host!
- Save some cluster resources in reserve.
  - Waste is good.
  - You'll thank me for it!

concentrated technology

Avoiding the 16 Biggest DA & DRS Configuration Mistakes

Greg Shields
Senior Partner and Principal Technologist,
Concentrated Technology, LLC
http://ConcentratedTech.com

concentrated technology