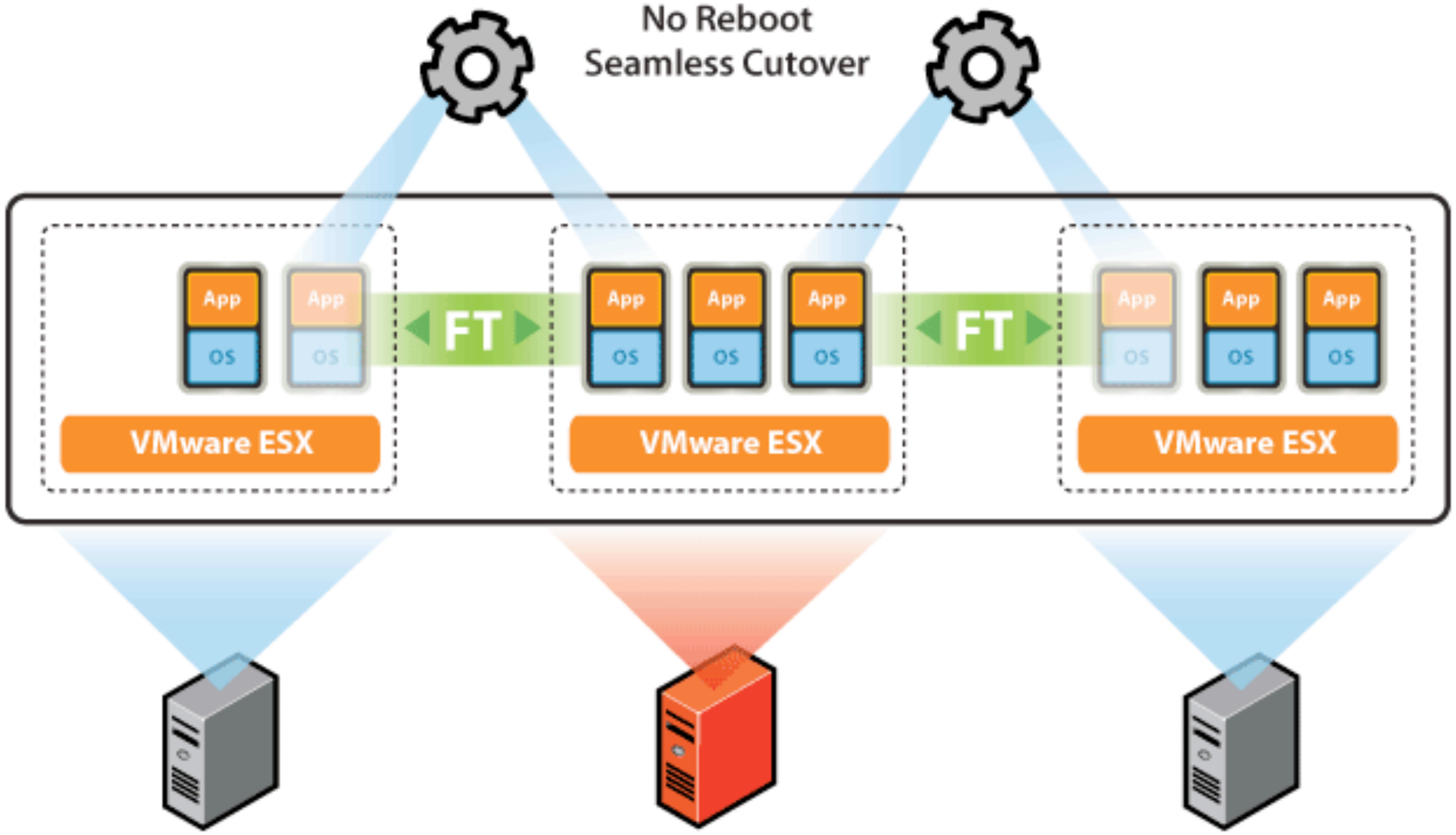
VMware Fault Tolerance Deep Dive

rod@hiperlogic.com

Why VMware Fault Tolerance?

Eliminates any workload disruption due to server hardware failure.



High Level Overview

- FT Is Enabled on a Virtual Machine (This is the “Primary” VM)
- “Secondary” VM is created on another host using Distributed Resource Scheduler.
- Primary and Secondary run in lockstep.
- If Primary dies, secondary takes over with no data loss or interruption, new secondary started automatically.

Use Cases

- FT “On-Demand” to protect VM’s during critical periods, like Accounting VM at end of quarter running long running reports.
- Protect workloads/OS’s that don’t have clustered solutions available. (BES)
- Protect workloads that were previously to expensive/difficult to provide FT.

Performance Impact

- Little impact to throughput when primary and secondary have enough CPU headroom.
- FT Logging Traffic really is dependent on application behavior, incoming I/O is key factor. If congested latency-bound apps may be impacted.
- Negligible I/O latency increase (few hundred microseconds)
- See VMware whitepaper for performance numbers on Exchange, SQL Server etc.

General FT Requirements

Requirements

- Additional dedicated GbE FT Logging NIC.
- Shared Storage.
- Thick-Eager Zero Disks.
- HA Cluster.
- All ESX hosts running same build number.
- Hardware Virtualization Enabled in BIOS.

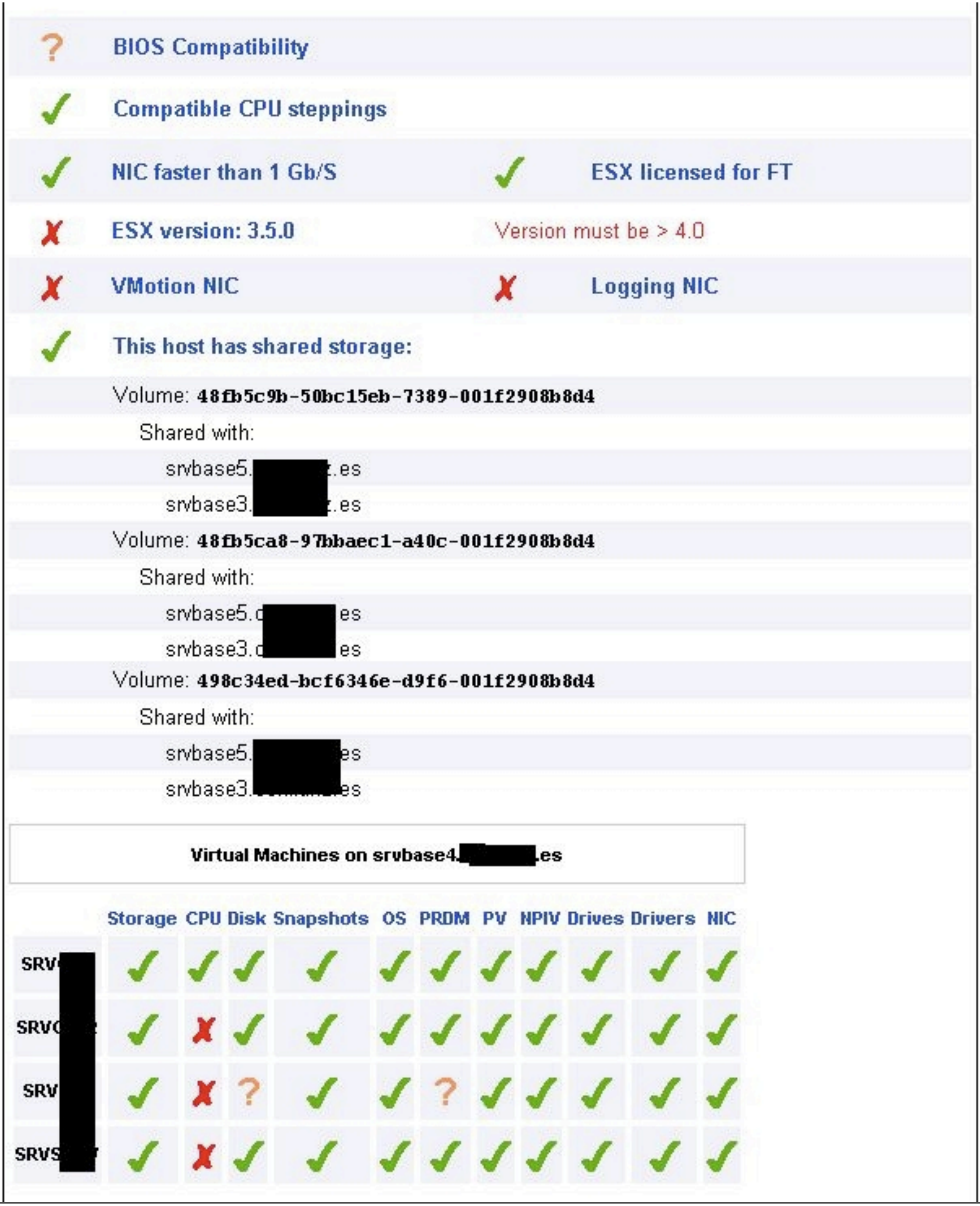
VMware vSphere Advanced or Higher

Use Hardware Site Survey

Download and Install the Fault Tolerance Site Survey:

http://www.vmware.com/download/shared_utilities.html

FT Capable CPU's with hardware virtualization (AMD-V, Intel VT) are REQUIRED.



? BIOS Compatibility

✓ Compatible CPU steppings

✓ NIC faster than 1 Gb/S ✓ ESX licensed for FT

✗ ESX version: 3.5.0 Version must be > 4.0

✗ VMotion NIC ✗ Logging NIC

✓ This host has shared storage:

Volume: `48fb5c9b-50bc15eb-7389-001f2908b8d4`

Shared with:

srvbase5.[redacted].es
srvbase3.[redacted].es

Volume: `48fb5ca8-97bbaec1-a40c-001f2908b8d4`

Shared with:

srvbase5.[redacted]es
srvbase3.[redacted]es

Volume: `498c34ed-bcf6346e-d9f6-001f2908b8d4`

Shared with:

srvbase5.[redacted]es
srvbase3.[redacted]es

Virtual Machines on srvbase4.[redacted].es

	Storage	CPU	Disk	Snapshots	OS	PRDM	PV	NPIV	Drives	Drivers	NIC
SRV[redacted]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SRV[redacted]	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
SRV[redacted]	✓	✗	?	✓	✓	?	✓	✓	✓	✓	✓
SRV[redacted]	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓

Site Survey Results

Hardware/Guest Requirements

	Intel 45	Intel i7	AMD
2008 64-bit	■	□	□
Vista	■	□	□
2003 64-bit	■	□	□
2003 32-bit	■	□	□
XP 64-bit	■	□	□
XP 32-bit	■	□	🟥
2000	□	□	🟥
NT 4	□	□	🟥
Linux	■	□	□

■ YES

□ YES/OFF

🟥 NO

Intel Xeon based on 45nm Core 2 Microarchitecture Category:

- 3100 Series Wolfdale
- 3300 Series Yorkfield
- 5200 Series Wolfdale (DP)
- 5400 Series Harpertown
- 7400 Series Dunnington

Intel Xeon based on Core i7 Microarchitecture Category:

- 5500 Series (Nehalem)

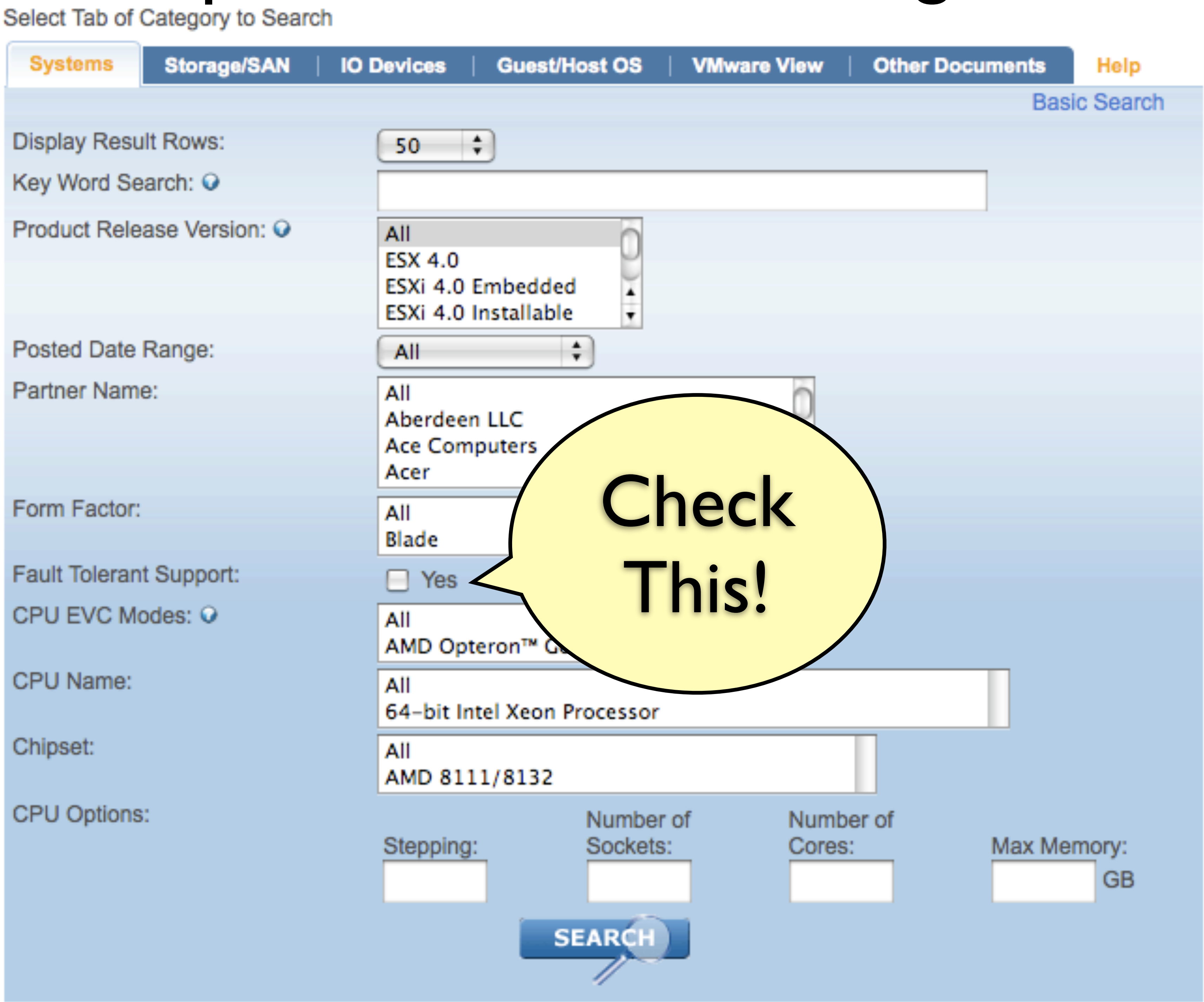
AMD 3rd Generation Opteron Category:

- 1300 Series Budapest
- 2300 Series Barcelona (DP)
- 8300 Series Barcelona (MP)

http://kb.vmware.com/kb/1008027

HCL Requirements

http://www.vmware.com/go/hcl



Select Tab of Category to Search

Display Result Rows: 50

Key Word Search:

Product Release Version: All
ESX 4.0
ESXi 4.0 Embedded
ESXi 4.0 Installable

Posted Date Range: All

Partner Name: All
Aberdeen LLC
Ace Computers
Acer

Form Factor: All
Blade

Fault Tolerant Support: □ Yes

CPU EVC Modes: All
AMD Opteron™ G

CPU Name: All
64-bit Intel Xeon Processor

Chipset: All
AMD 8111/8132

CPU Options:

Stepping: Number of Sockets: Number of Cores: Max Memory: GB

Check This!

VM Requirements

NO

vSMP

Thin Provisioned Disks

Snapshots

Storage VMotion

Physical Raw Disk Mappings

Physical CD/Floppy

Nested/Extended Page Table

NPIV

VM hardware < v7

Paravirtualized Drivers

vmxnet3, sound, USB

Hot Plug

Local Storage

MSCS Clustering

Turn On/Enable Detail

Unsupported Devices Removed (Sound, USB, Phys CD/Floppy)

Memory Reservation on VM set to Prevent Swapping/Ballooning

Hardware MMU (AMD RV/Intel EPT) Turned Off on VM (OFF)

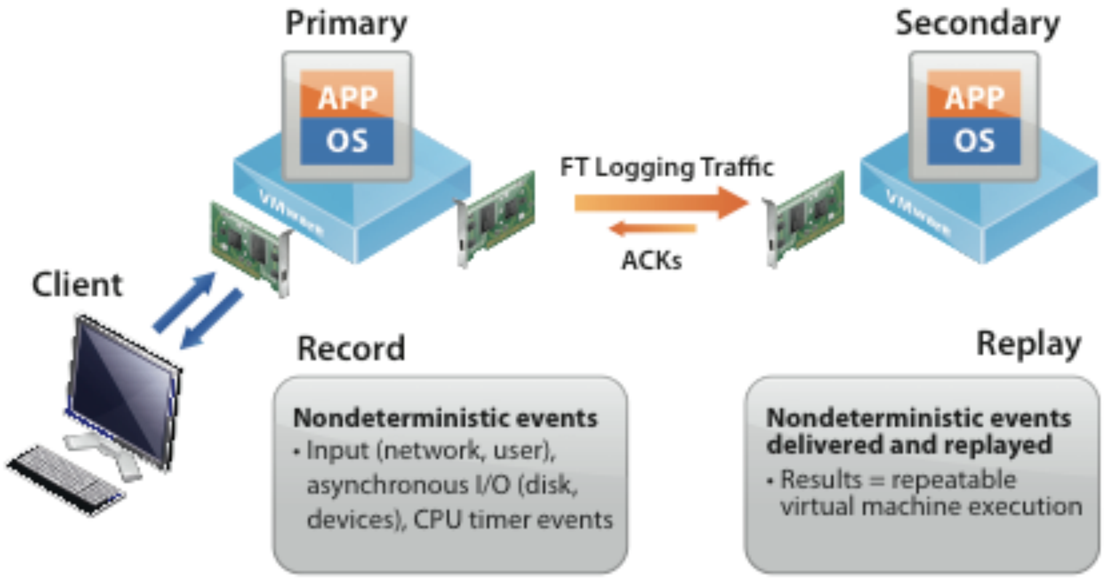
DRS For VM Turned Off

Thin Disk Converted to Thick-Eager Zeroed (Off)

Live-Migrate Primary to Create Secondary

The FT Secret Sauce

vLockStep



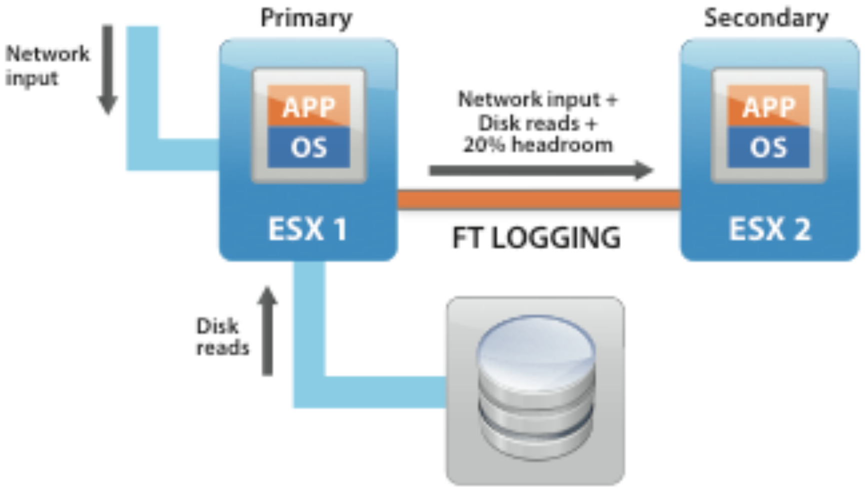
Nondeterministic events
- Input (network, user), asynchronous I/O (disk, devices), CPU timer events

Nondeterministic events delivered and replayed
- Results = repeatable virtual machine execution

Only Primary Performs
“Writes” to Network/Disk.

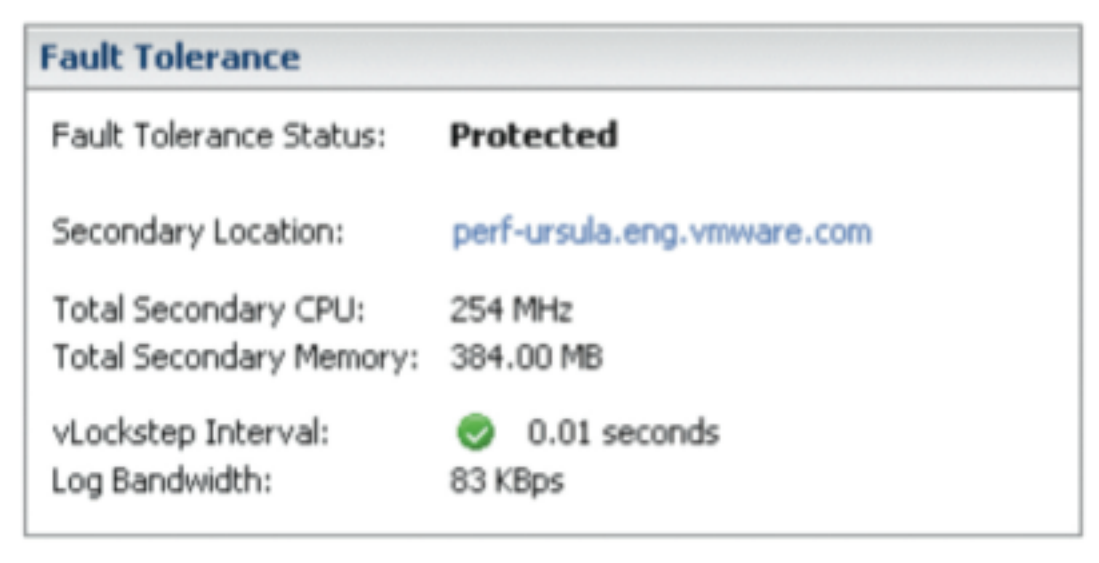
How Much FT Logging Bandwidth?

FT Logging Bandwidth ~= (Avg disk reads (MB/s) x 8 + Avg network input (Mbps)) x 1.2



What is vLockstep Interval?

- vLockstep Interval: Average time delay.
- Secondary has info to catch up, even if primary host dies.
- VMware will slow primary down if needed.
- Shown in vCenter.



Fault Tolerance

Fault Tolerance Status:	**Protected**
Secondary Location:	perf-ursula.eng.vmware.com
Total Secondary CPU:	254 MHz
Total Secondary Memory:	384.00 MB
vLockstep Interval:	✓ 0.01 seconds
Log Bandwidth:	83 KBps

How is the Secondary VM Placement Chosen?

- On Initial Creation DRS chooses if enabled.
- On Failure VMware HA chooses.
- In either case, automatic placement of secondary is NOT under user control.

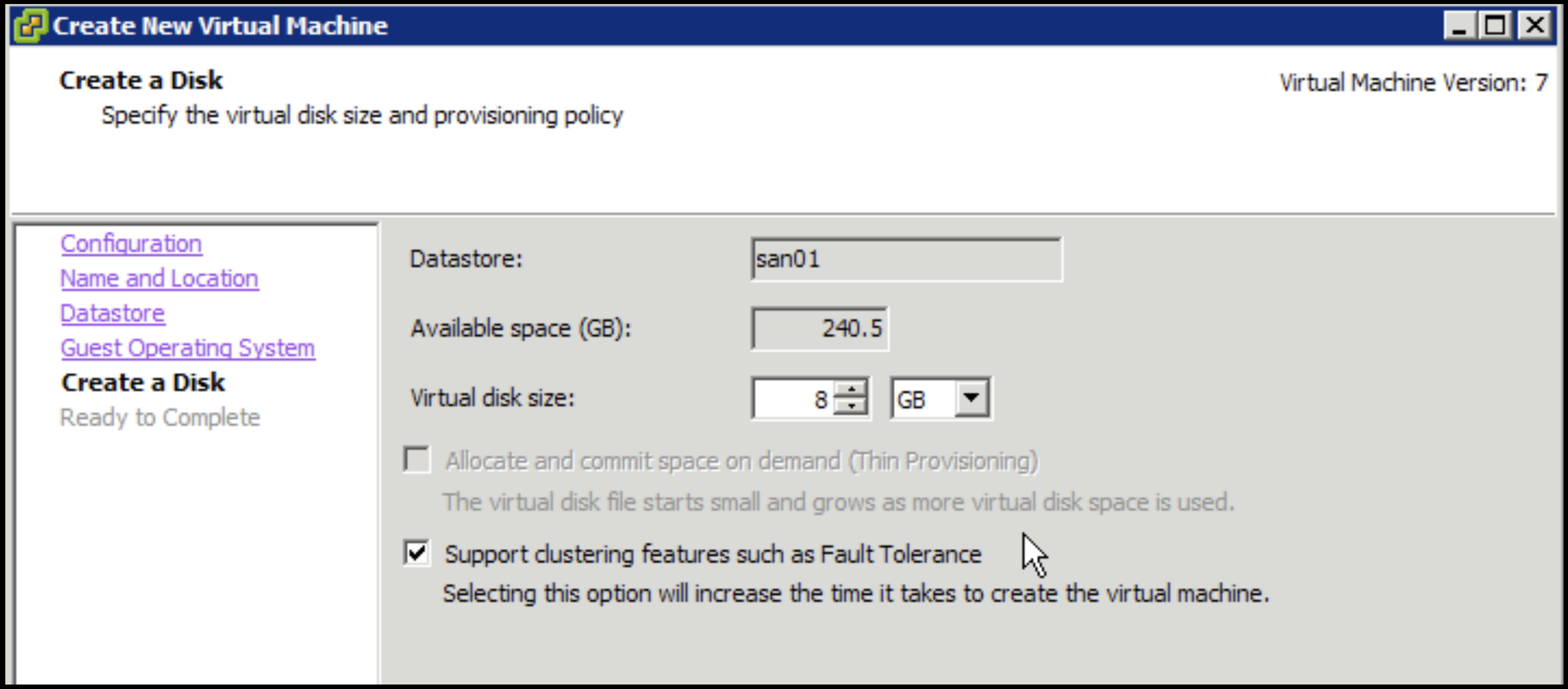
Secondary VM A Full Citizen?

- Named “VMName (secondary) in VC.
- Shows up in VM list, but not in inventory.
- CAN move secondary to another host, open console, etc.

Storage Considerations

Thick and Eager Zeroed

- Thick and Eager Zeroed (All blocks pre-zeroed) VM Required.



Create a Disk

Specify the virtual disk size and provisioning policy

Virtual Machine Version: 7

Datastore: san01

Available space (GB): 240.5

Virtual disk size: 8 GB

□ Allocate and commit space on demand (Thin Provisioning)

The virtual disk file starts small and grows as more virtual disk space is used.

☑ Support clustering features such as Fault Tolerance

Selecting this option will increase the time it takes to create the virtual machine.

Thick and Eager Conversion

- Let FT do the conversion when you enable FT.
- Use vmkfstools --diskformat eagerzeroedthick
- Set "cbtmotion.ForceEagerZeroedThick="true" in .vmx and Storage VMotion to do conversion.

Disk Read Intensive Workload Optimization

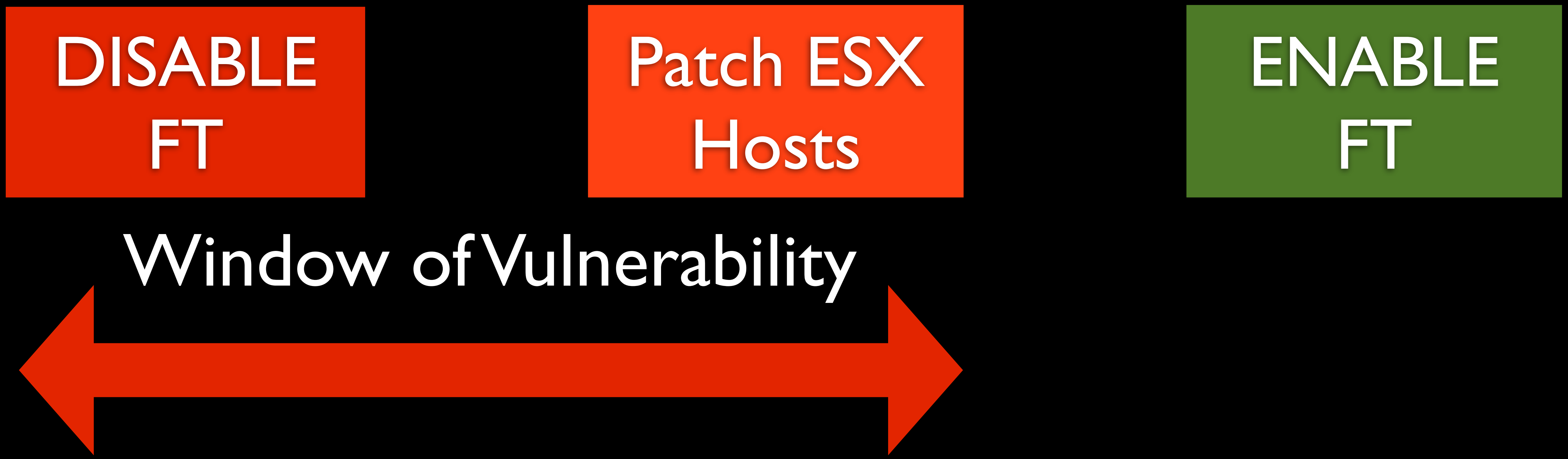
Have secondary read disk I/O from disk instead of primary sending over the FT network.

Add to VMX file:

```
replay.logReadData = checksum
```

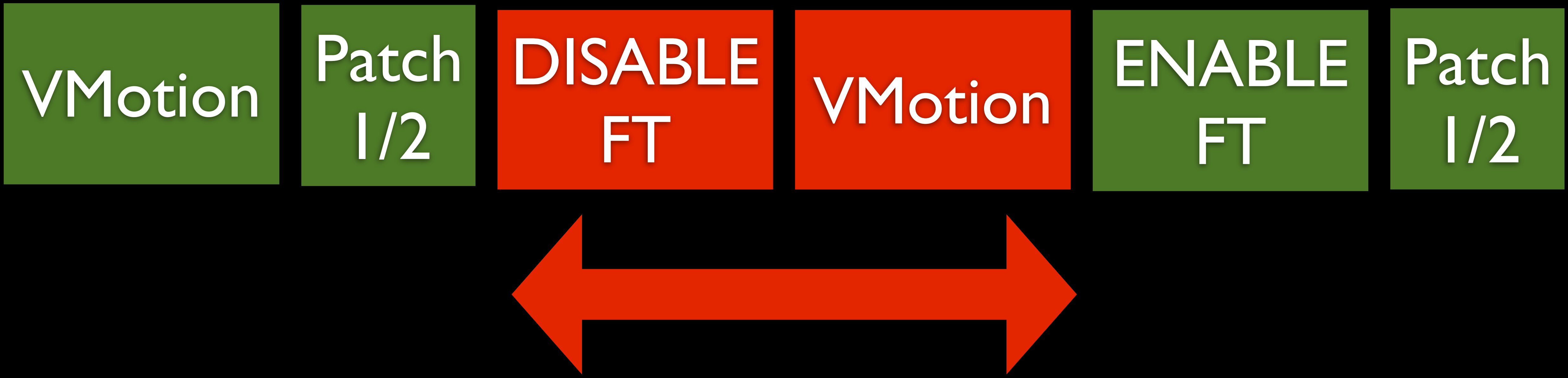
Patching Best Practices with Fault Tolerance

Patching Option 1



FT Requires all ESX hosts be at same patch level.

Patching Option 2



VMotion

Patch 1/2

DISABLE FT

VMotion

ENABLE FT

Patch 1/2

SHORTER Window of Vulnerability

Use With Four or More ESX Hosts With Enough Resources

Other Things to Consider

- Secondary consumes resources.
- Additional resources consumed on primary.
- The secondary can slow the primary down due to different clock, power management, or VM contention.
- Turning ON (not just enabling) can have negative performance impact on VM.

Performance Recommendations

- Disable BIOS CPU Power Management.
- Distributes Primaries among hosts.
- Use dedicated GbE links for FT/VMotion.
- No more than 4 FT VM’s per host.
- Use CPU reservations as needed.
- All ESX hosts have identical CPU frequency.
- Use FT On/Off Sparingly, stage operations.

The Ideal VM for FT

- Runs well on uniprocessor VM.
- Tolerate a small increase in latency.
- Medium network bandwidth requirements (< 600 Mbps).
- Doesn't require heavy disk reads.
- Expensive or not possible to protect otherwise.
- Can tolerate windows of vulnerability.

FT Network Performance

- Latency on FT network less than 1ms. Check with vmkping
- Use Jumbo Frames
- 1 GbE minimum, 10 GbE better, Infiniband best.
- Minimum recommendation: one NIC for FT, one NIC for VMotion, and one NIC as a shared failover for both.

FT Logging NIC Teaming Best Practice

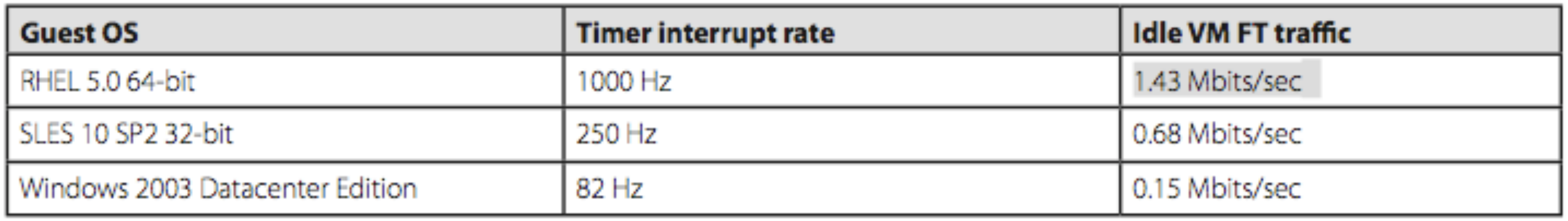
Note all FT logging from a host has same source port and MAC, if using multiple uplinks you must use IP hash policy which uses source AND destination to determine uplink used for load balancing.

Note ports on the physical switch must be in etherchannel mode (See KB Article 1004048)

KB Article: **1011966, 1004048**

FT Logging NIC Guest Optimization

For Linux Guests, reduce the default timer interrupt to reduce amount of unnecessary traffic over FT NIC. See kb 1005802



Guest OS	Timer interrupt rate	Idle VM FT traffic
RHEL 5.0 64-bit	1000 Hz	1.43 Mbits/sec
SLES 10 SP2 32-bit	250 Hz	0.68 Mbits/sec
Windows 2003 Datacenter Edition	82 Hz	0.15 Mbits/sec

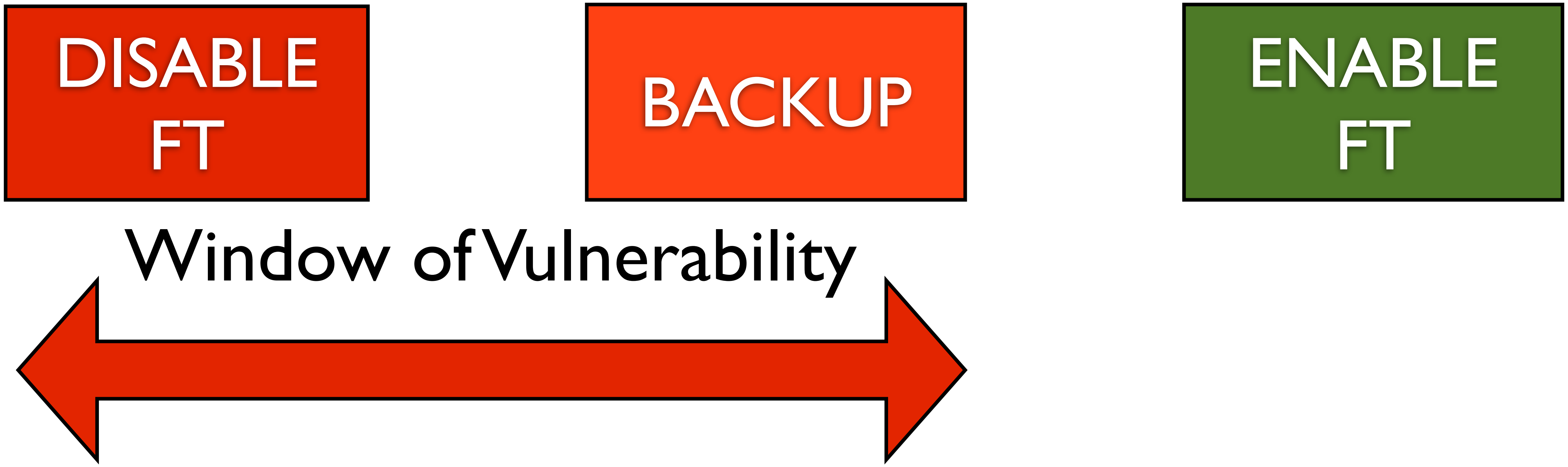
FT Logging Network Placement Best Practice

- Most traffic is from primary to secondary, secondary only sends back ACK's
- DON'T put all primary's on one host, match primaries with other secondaries to balance FT Logging NIC traffic. VMotion secondaries as needed.

FT ONLY PROTECT AGAINST HOST FAILURES

Best Practice: USE Storage Multi Pathing and Fully Redundant NIC Teaming to protect against component failure.

FT Backup Strategy



DISABLE FT

BACKUP

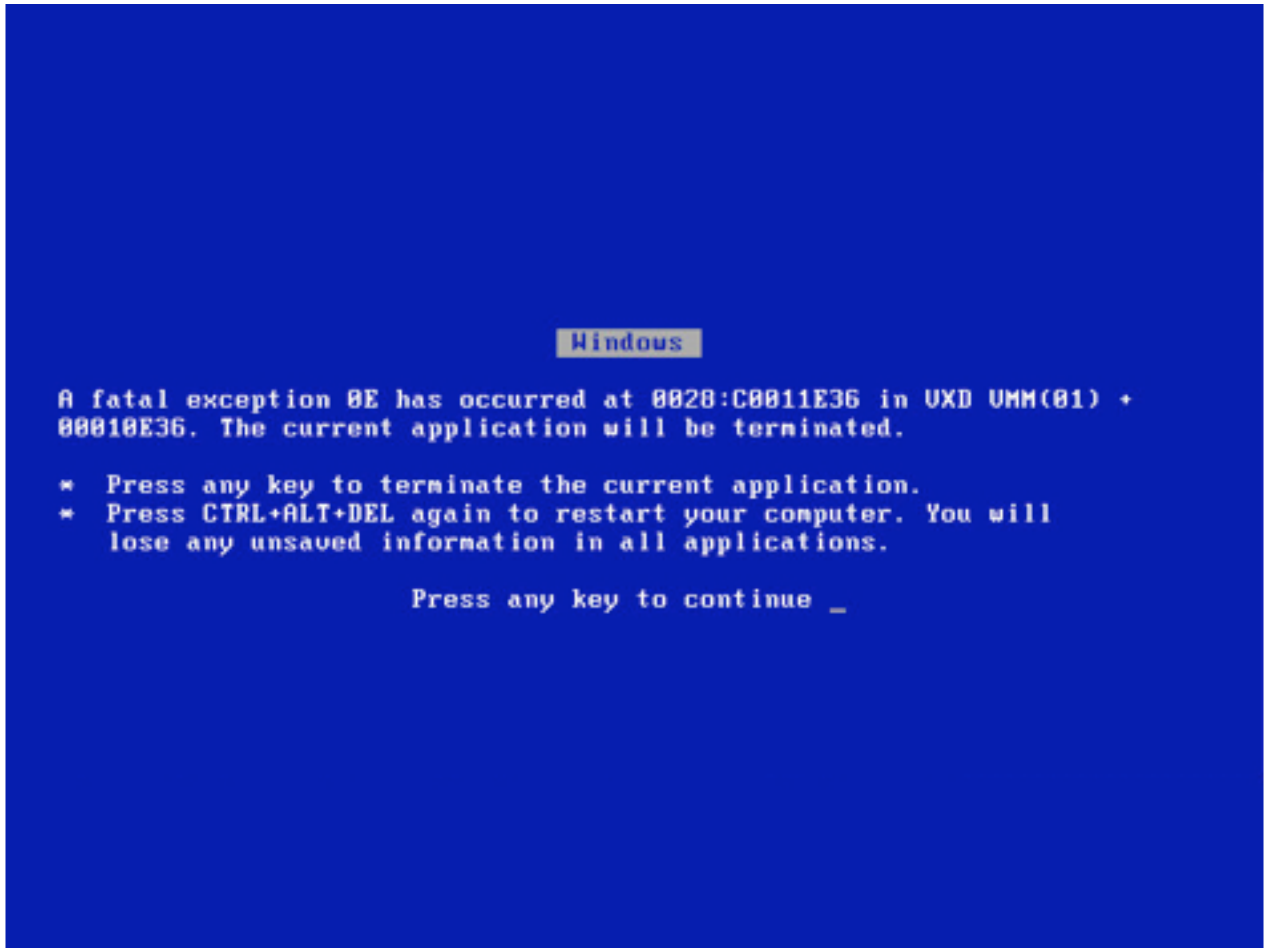
ENABLE FT

Window of Vulnerability

Can I do Host Level Backup in FT VM?

- Possible, but could overload the FT logging network with all the disk and network I/O. NOT recommended.

NOT FOR APP LEVEL PROTECTION



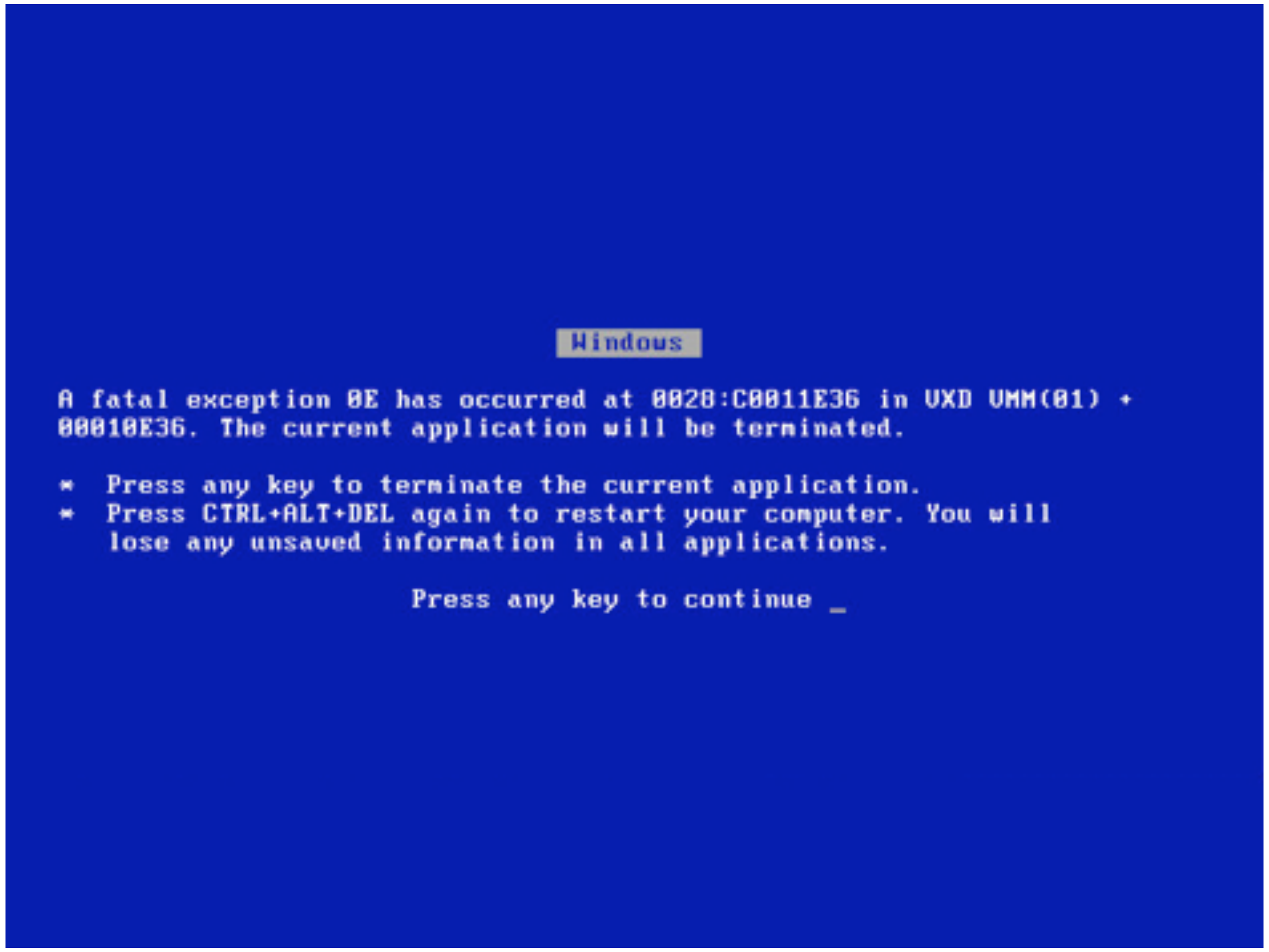
```
                        Windows

A fatal exception 0E has occurred at 0028:C0011E36 in VXD VMM(01) +
00010E36. The current application will be terminated.

*  Press any key to terminate the current application.
*  Press CTRL+ALT+DEL again to restart your computer. You will
   lose any unsaved information in all applications.

                  Press any key to continue _
```

Primary



```
                        Windows

A fatal exception 0E has occurred at 0028:C0011E36 in VXD VMM(01) +
00010E36. The current application will be terminated.

*  Press any key to terminate the current application.
*  Press CTRL+ALT+DEL again to restart your computer. You will
   lose any unsaved information in all applications.

                  Press any key to continue _
```

Secondary

VMware HA will automatically restart the failed Primary VM and re-spawn a new Secondary

Scripting FT With PowerShell

To enable FT for a VM:

```
Get-VM X | Get-View | % { $_.CreateSecondaryVM($null) }
```

To disable, run:

```
Get-VM X | Select -First 1 | Get-View | %
{ $_.TurnOffFaultToleranceForVM() }
```

Note that in PowerCLI 4.0 Get-VM will return a fault tolerant VM twice, so we select the first one.

Advanced Debugging

Errors can be cryptic, these documents are a big help:

- kb article 1010634
- vSphere Availibility Guide for Error Messages
- http://bit.ly/114K3E

Future Directions

Allow mixed builds for ESX hosts
Enable vSMP for FT VM’s
Allow Storage VMotion on FT VM’s
Allow DRS on FT VM’s
Eliminate shared storage requirement
Allow FT VM’s to span clusters
Enable WAN/Metro Support for FT VM’s
Enable VCB/Veeam Backups by allowing 1 snapshot on FT VM.

Reading List

- vSphere Migration Prerequisites Checklist
- VMware Fault Tolerance Recommendations and Considerations on VMware vSphere 4
- Performance Best Practices for vSphere 4
- Site Survey Help Guide
- vSphere Availability Guide
- Protecting Mission Critical Workloads with VMware Fault Tolerance
- VMware Fault Tolerance Architecture and Performance
- kb 1008027, 1010601, 1013428, 1011966, 1005802

What About "The Other Guys?"

- Marathon's everRUN Protection Level 3 (Xen Based) is very similar for Xen shops.
- Buy a FT server as a hardware only solution.

Rodney Mach

rod@hiperlogic.com

twitter:chamdor

rodmach.com/blog