

POSSIBLE
BEGINS
WITH YOU

Extreme Performance Series: Performance Best Practices

Valentin Bondzio, VMware, Inc.

#vmworld

#VIN2677BE

vmware®

Disclaimer

This presentation may contain product features or functionality that are currently under development.

This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.

Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.

Technical feasibility and market demand will affect final delivery.

Pricing and packaging for any new features/functionality/technology discussed or presented, have not been determined.

Agenda

Introduction

Getting started

CPU Usage Accounting

Repeating some basics

Latency Sensitivity

What it does and when it doesn't

Active Memory

How it is measured and for what

Introduction

What this isn't

Quick Stats

Released 2018-08-03

1.32 MB

88 Pages

4 Chapters

514 Best Practices

~2-3 hour read (with note taking)

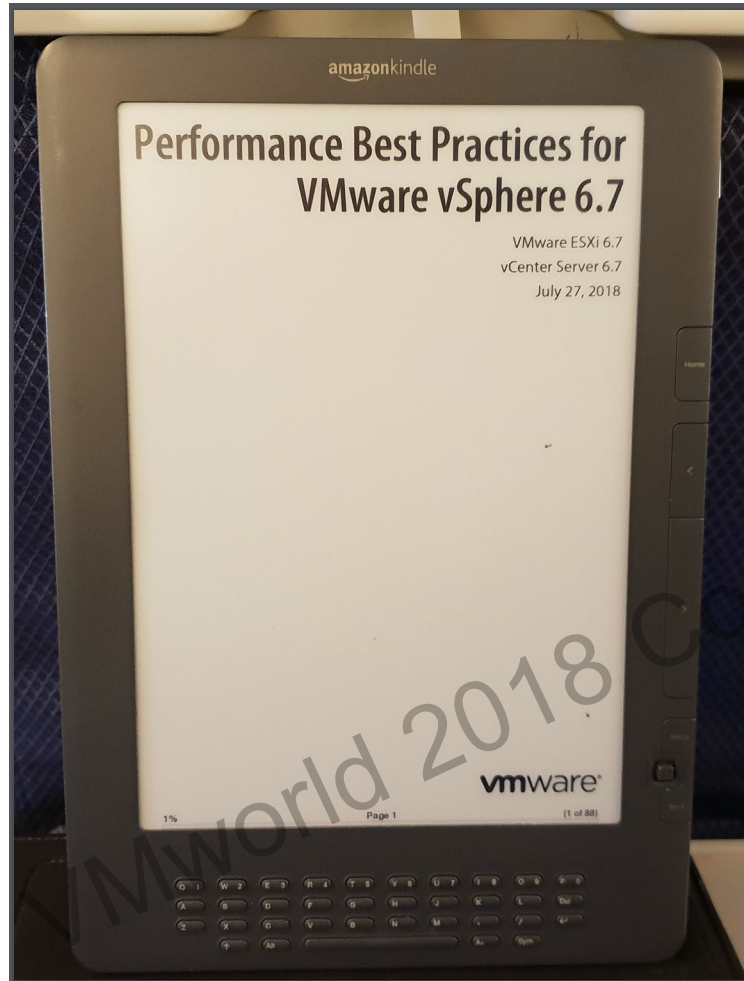
Performance Best Practices for VMware vSphere 6.7

VMware ESXi 6.7
vCenter Server 6.7
July 27, 2018

vmware®

Introduction

What this isn't



Quick Start

Release

1.32 MB

88 Pages

4 Chapters

514 Best Practices

~2-3 hour read (with time to talk)

This bad boy can fit so many best practices in it

slaps cover of document



Introduction

We could be here all day ...

Keep everything reasonably up-to-date

HW selection makes a difference

Refer to existing best practice documentation

Rightsize your workloads

Evaluate your power management policy

Use resource pools properly, or not at all

Use DRS to manage contention

Monitor oversubscription

Define and monitor application level KPIs

Use paravirtualized drivers

Understand your workload

Don't compare apples with oranges

...

Introduction

Where to go for more information

Performance Best Practices for vSphere 6.7

- <https://blogs.vmware.com/performance/2018/08/performance-best-practices-guide-for-vsphere-6-7.html>

Application Specific Best Practice Guides (SQL Server, Oracle, etc.)

- <https://www.vmware.com/solutions/business-critical-apps.html>

VROOM! Blog

- <https://blogs.vmware.com/performance/>

Performance Community

- <https://communities.vmware.com/community/vmtn/performance>

CPU usage accounting

VMworld 2018 Content: Not for publication or distribution

CPU Usage Accounting

What states are there



CPU Usage Accounting

What states are there



CPU Usage Accounting

What states are there



CPU Usage Accounting

In an ideal world



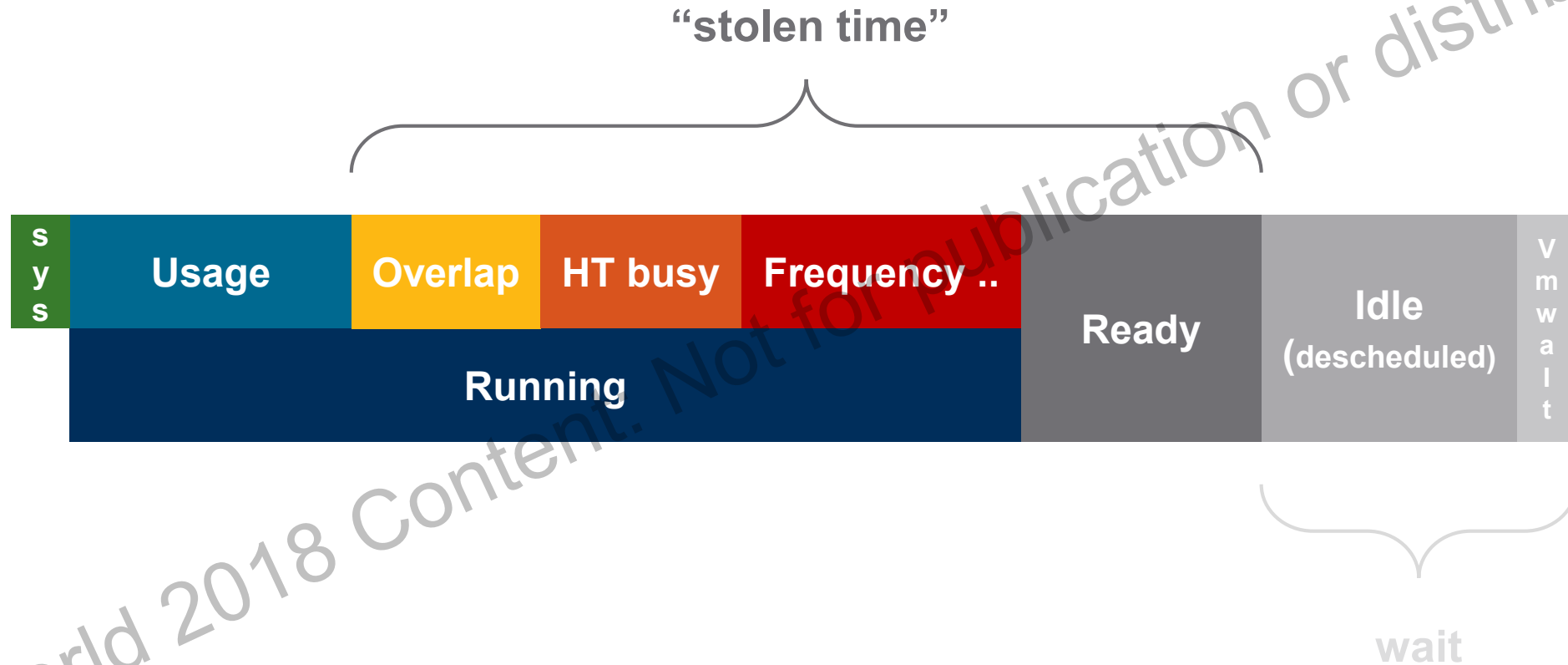
What is “CPU Usage” for a VM / vCPU world

Compared to the nominal / base frequency of a single core



CPU Usage Accounting

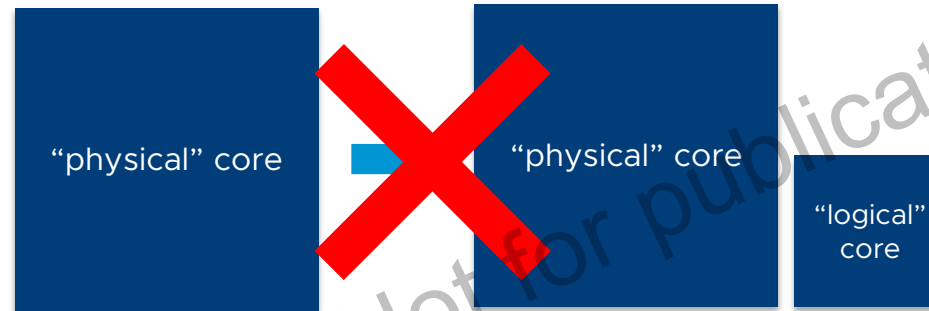
What is charged against the VM



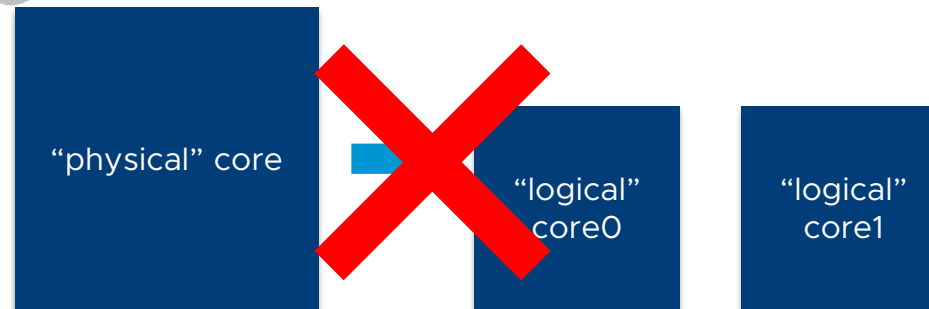
Intel® Hyper-Threading Technology

Cores and Threads

Does enabling HT “spawn” a less capable “logical core”?



Maybe two slightly less capable “logical” cores?



Intel® Hyper-Threading Technology

Individual throughput reduction, aggregated throughput increase at high load

100



100

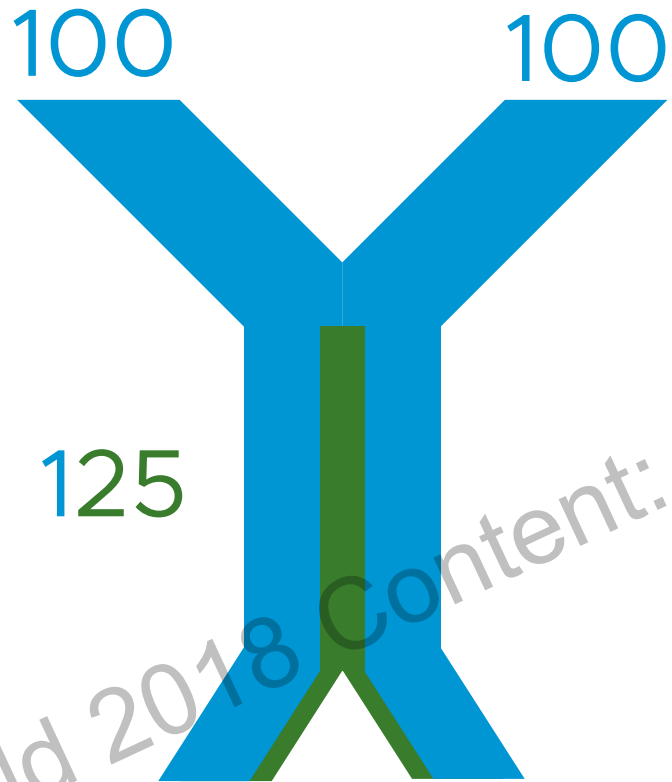


~125



Intel® Hyper-Threading Technology on ESXi

Throughput reduction is accounted for in USED



`HTEfficiencyShift` - Default: 2

HT is:

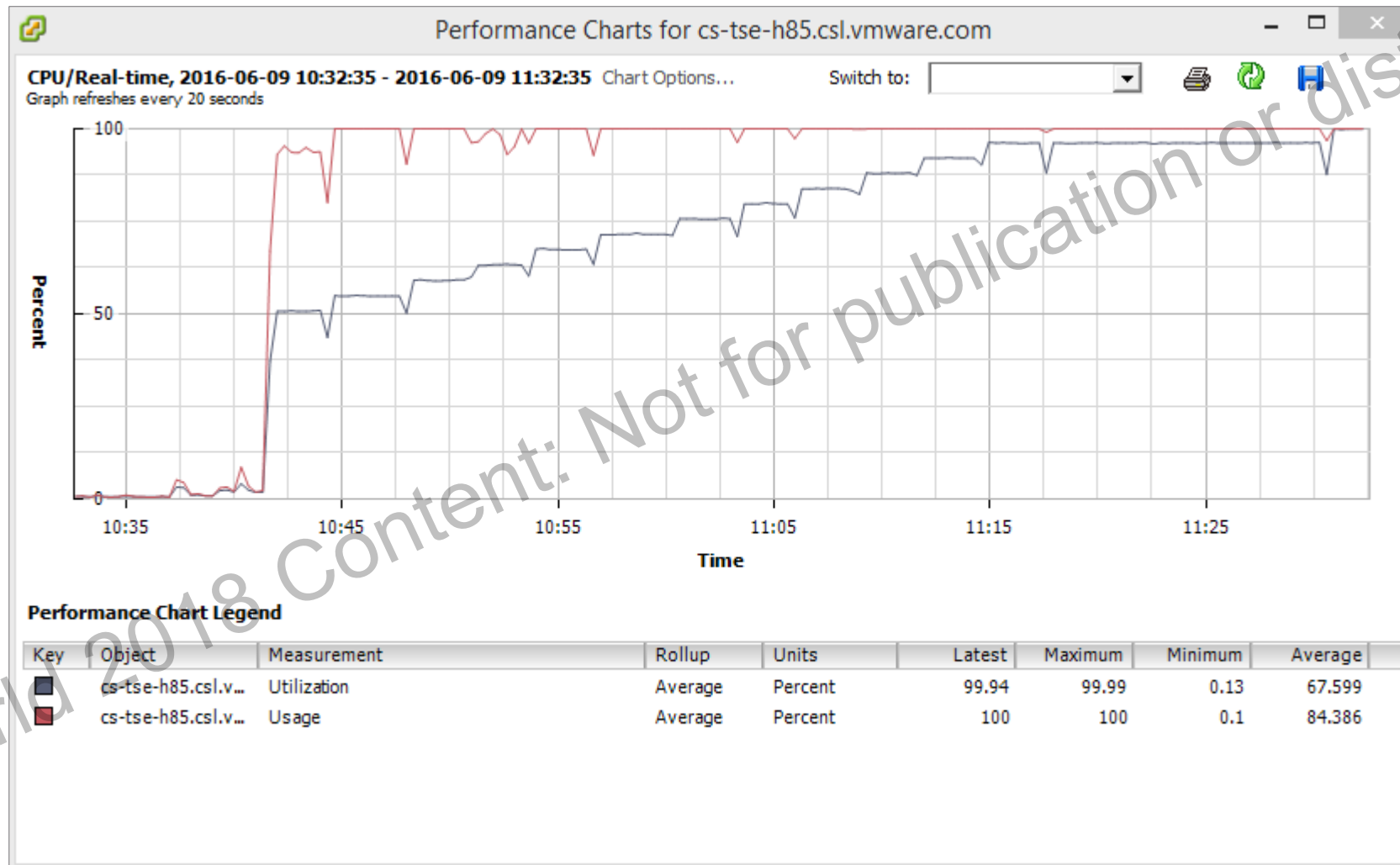
1:	50	%
2:	25	%
3:	12.5	%
4:	6.25	%
5:	3.125	%

more efficient than no-HT

$$2 \times 50 + 12.5 = 62.5$$

CPU Usage Accounting

Usage vs. Utilization



Are we talking about L1TF?

Not a security expert !



Where can I hear about L1TF?

Mitigating CPU Security Vulnerabilities – A look at vSphere Mitigations [SAI3770BE]

- Thu., 08 November, 09:00 - 10:00 | Hall 8.0, Room 17
- Richard Brunner, CTO, Server Platform Technologies, VMware

There is that term again ...

“Power Management”

There is that term again ...

“Power Management”

Umbrella Term

“Power Management”

Fan Speed

Memory Refresh Rate

Turbo Boost

DMI Link Frequency

Energy Performance Bias

Redundant Power Supply Mode

Memory Frequency

C1E

Uncore Frequency

Memory Patrol Scrub

QPI Link Frequency

Energy Efficient Turbo Boost

Channel Interleaving

Maximum PCI Express Speed

Collaborative Power Control

Umbrella Term

“Power Management”

Fan Speed

Memory Refresh Rate

Turbo Boost

DMI Link Frequency

Options aka: Power Regulator, CPU Power Management, EIST
Energy Performance Bias

Redundant Power Supply Mode

Memory Frequency

C1E

Uncore Frequency

Memory Patrol Scrub

QPI Link Frequency

Energy Efficient Turbo Boost

Channel Interleaving

Maximum PCI Express Speed

Collaborative Power Control

P-States

Deep C-States

Power Management _Profiles_

HP

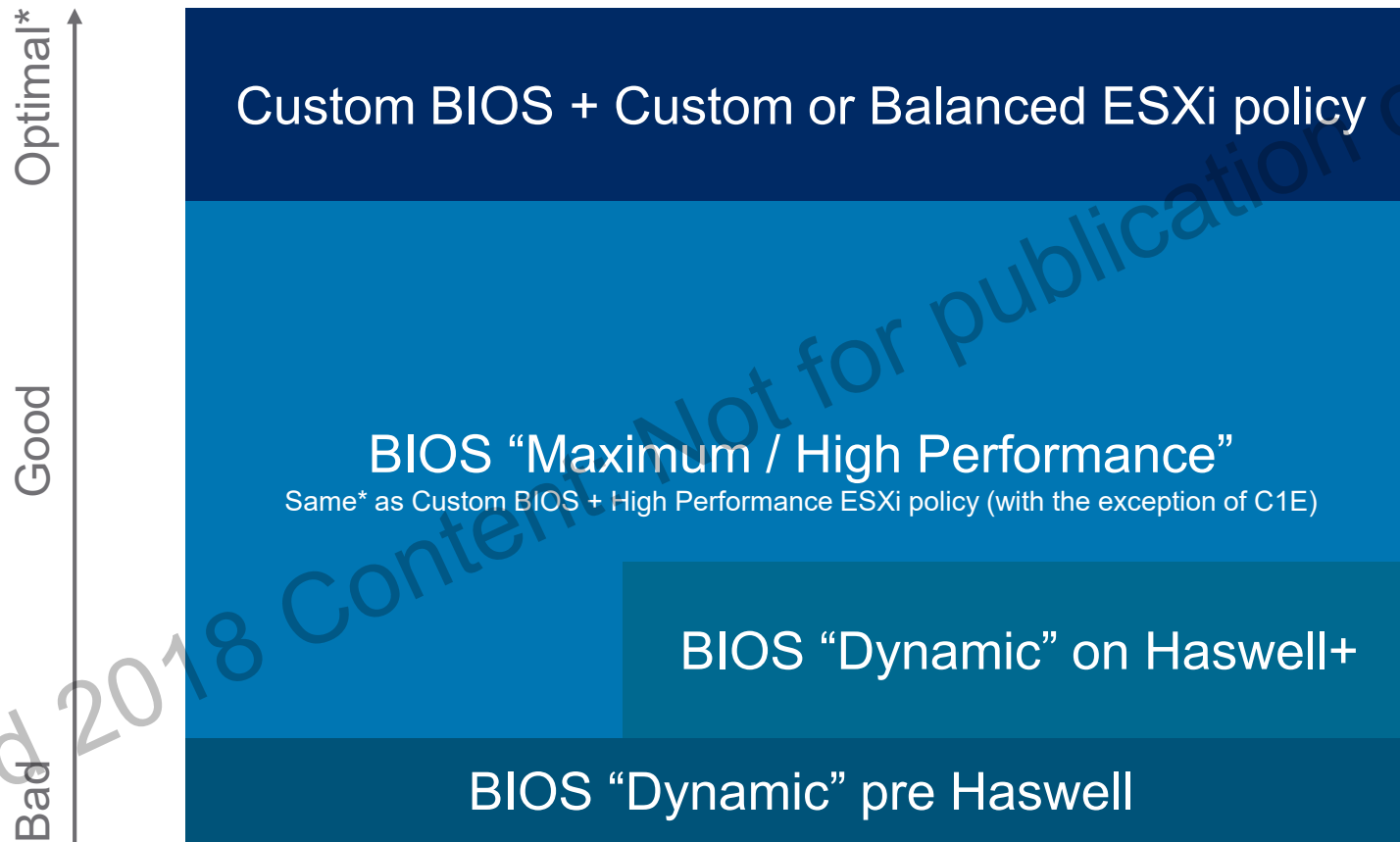
Power Management		Power Profile (Power Management Profile)			
	Power Profile	Balanced Power and Perf.	(...)	Maximum Performance	Cust. Custom Recommended
P-State	Power Regulator	Dynamic Power Savings Mode	(...)	Static High Performance Mode	any OS Control Mode or SHPM
Deep C	Min. Proc. Idle Power Core C-State	C6 State	(...)	No C-states	any C6 State
	Min. Proc. Idle Power Package Core C-State	Package C6 (retention) State	(...)	No Package State	any PC6 (retention) State
	Advanced Power Options				
	Intel QPI Link Power Management	Enabled	(...)	Disabled	any Disabled
	Intel QPI Link Frequency	Auto	(...)	Auto	any Auto
	Intel QPI Link Enablement	Auto	(...)	Auto	any Auto
	Energy/Performance Bias	Balanced Performance	(...)	Maximum Performance	any Maximum Performance
	Maximum Memory Bus Frequency	Auto	(...)	Auto	any Auto
	Channel Interleaving	Enabled	(...)	Enabled	any Enabled
	Maximum PCI Express Speed	Maximum Supported	(...)	Maximum Supported	any Maximum Supported
	Dynamic Power Savings Mode Response	Fast	(...)	Fast	any Fast
	Collaborative Power Control	Enabled	(...)	Enabled	any Disabled
	Redundant Power Supply Mode	Balanced Mode	(...)	Balanced Mode	any Balanced Mode
	Intel DMI Link Frequency	Auto	(...)	Auto	any Auto

Power Management _Profiles_

Dell

System Profile Settings		System Profile (Power Management Profile)				
	System Profile	Perf. Per Watt (DAPC)	(...)	Performance	Cust.	Custom Recommended
P-State	CPU Power Management	System DBPM (DAPC)	(...)	Maximum Performance	any	OS DBPM or Maximum Performance
	Memory Frequency	Maximum Performance	(...)	Maximum Performance	any	Maximum Performance
	Turbo Boost	Enabled	(...)	Enabled	any	Enabled
	Energy Efficient Turbo	Enabled	(...)	Disabled	any	Disabled
C1E	C1E	Enabled	(...)	Disabled	any	Enabled
Deep C	C States	Enabled	(...)	Disabled	any	Enabled
	Collab. CPU Performance Control	Disabled	(...)	Disabled	any	Disabled
	Memory Patrol Scrub	Standard	(...)	Standard	any	Standard
	Memory Refresh Rate	1x	(...)	1x	any	1x
	Uncore Frequency	Dynamic	(...)	Maximum	any	Maximum
	Energy Efficient Policy	Balanced Performance	(...)	Performance	any	Performance
	# of TB Enabled Cores for CPU 1	All	(...)	All	any	All
	# of TB Enabled Cores for CPU 2	All	(...)	All	any	All
	Monitor/Mwait	Enabled	(...)	Enabled	any	Enabled

Power Policy “playfield”



* a few workloads fare better with more deterministic performance

Umbrella ‘Brella ‘Brella

“ESXi Power Management Policy”

P-States

Deep C-States

VMworld 2018 Content: Not for publication or distribution

*Energy Performance Bias

ESXi Power Management Policy

Only affects what's presented from the BIOS

somehost-h5c-saves-paint.net | ACTIONS ▾

Summary Monitor **Configure** Permissions VMs Datastores Networks

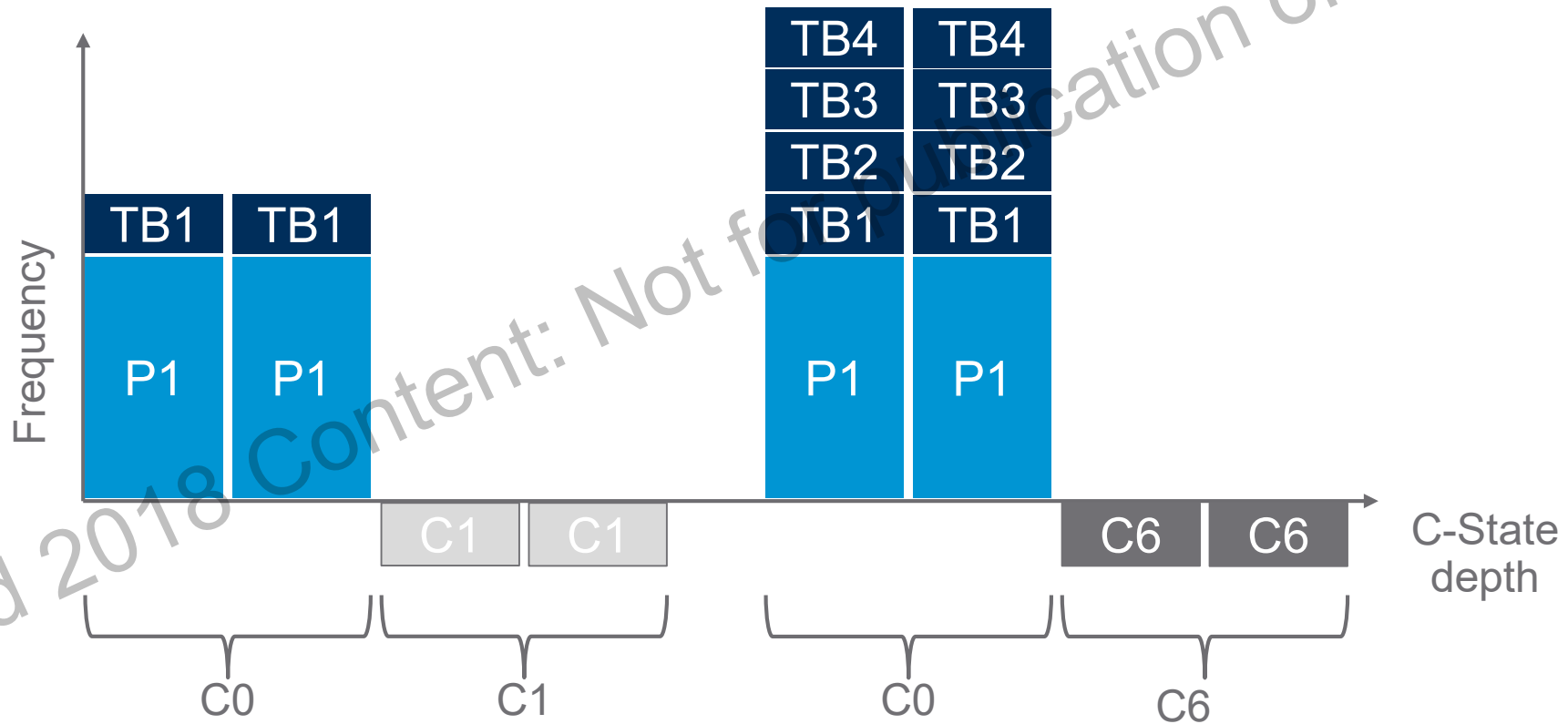
Services
Security Profile
System Swap
▼ Hardware
Processors
Memory
PCI Devices
Power Management ▼

Power Management EDIT...

Technology	ACPI P-states, ACPI C-states
Active policy	Balanced

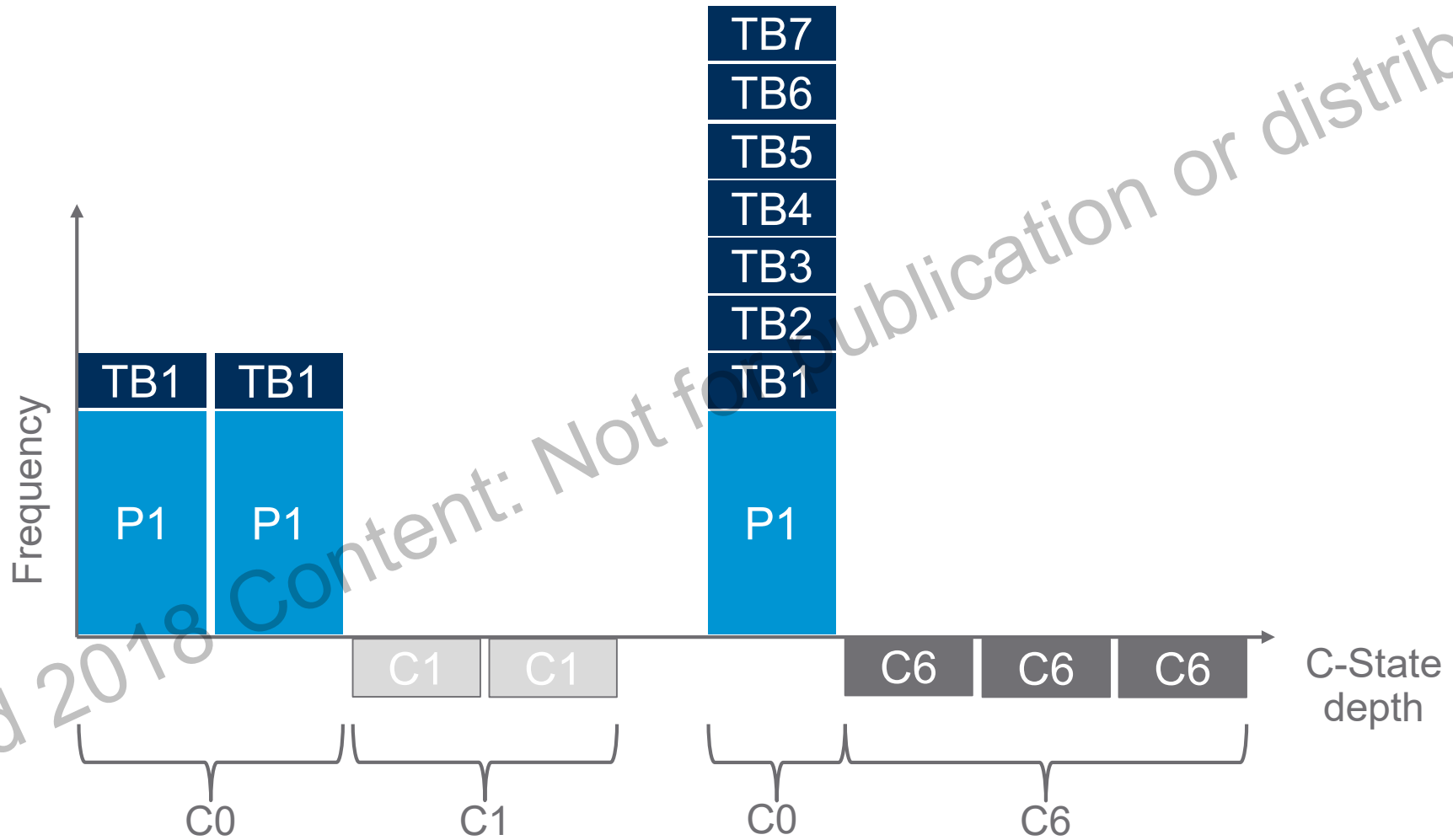
The magic of Turbo Boost

Dynamic, supported overclocking



The magic of Turbo Boost

Dynamic, supported overclocking



CPU Usage Accounting

Usage > Run thanks to Turbo Boost (or I/O)

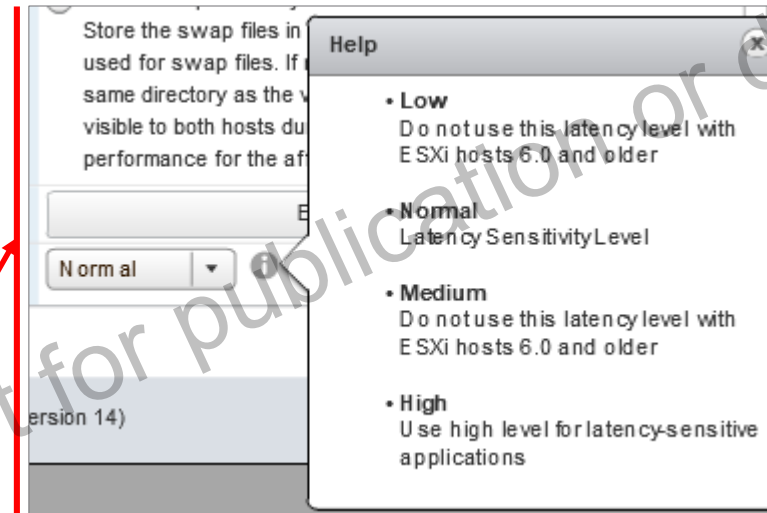
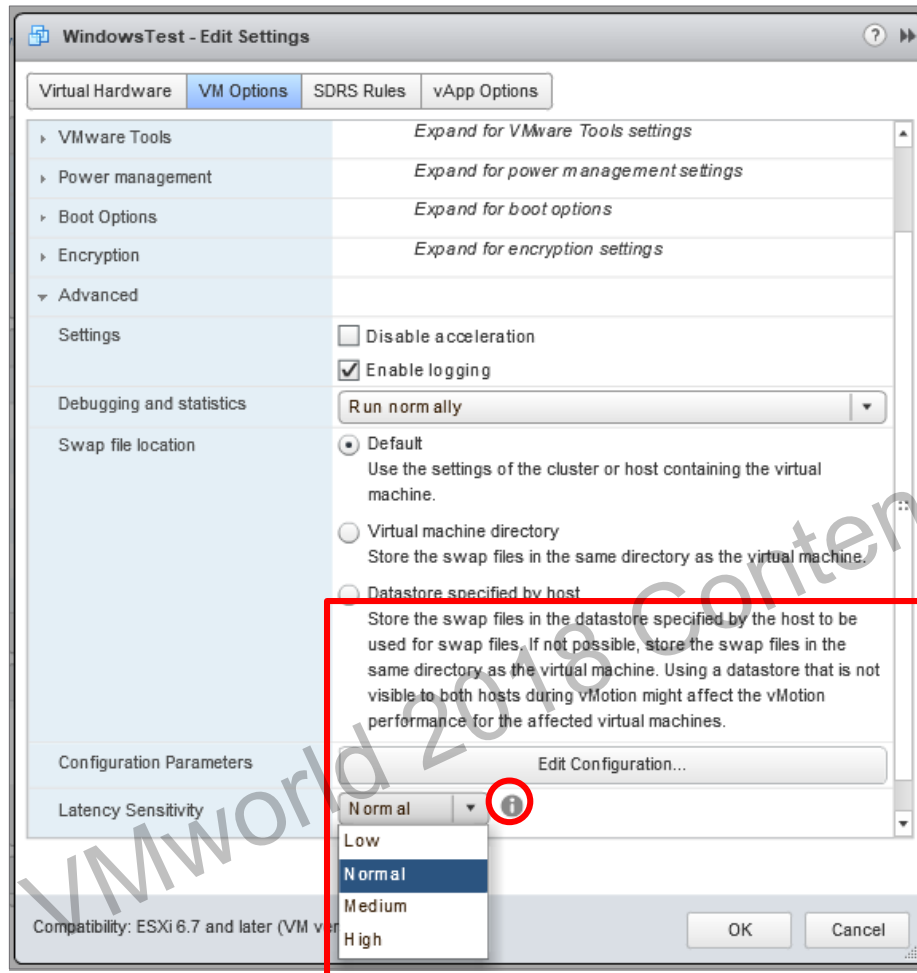


Latency Sensitivity

VMworld 2018 Content: Not for publication or distribution

Latency Sensitivity Options

What and Where



- Low: Network coalescing changes*
- Normal: Default
- Medium: Network coalescing changes*
- High: Affects CPU, Memory, Network

Latency Sensitivity Options

Since 6.7 also via the HTML 5 client

The screenshot shows the 'Edit Settings' window for a virtual machine named 'WindowsTest'. The 'VM Options' tab is selected. Under the 'Swap file location' section, three radio buttons are visible: 'Default' (selected), 'Virtual machine directory', and 'Datastore specified by host'. Below this is a 'Configuration Parameters' section with an 'EDIT CONFIGURATION...' button. The 'Latency Sensitivity' dropdown menu is open, showing 'Normal' as the selected option, with 'Normal' and 'High' as visible choices. A 'Fibre Channel NPIV' section is partially visible at the bottom.

Latency Sensitivity Options

Since 6.7 also via the HTML 5 client

The screenshot shows the 'Edit Settings' dialog for a virtual machine named 'WindowsTest'. The 'VM Options' tab is selected. Under the 'Swap file location' section, three radio button options are visible: 'Default' (selected), 'Virtual machine directory', and 'Datastore specified by host'. Below this is a 'Configuration Parameters' section with an 'EDIT CONFIGURATION...' button. The 'Latency Sensitivity' dropdown menu is open, showing 'Normal' as the selected option, with 'Normal' and 'High' also visible in the list. A 'Fibre Channel NPIV' section is partially visible at the bottom.

Clarity Dark Theme

Fully Supported in 6.7 U1

New in 6.7 U1

The screenshot displays the VMware vSphere Client interface in a dark theme. The top navigation bar includes the 'vm vSphere Client' logo, a 'Menu' dropdown, a search bar, and the user 'Administrator@VSPHERE.LOCAL'. The left sidebar shows a tree view of the environment, with 'WindowsTest' selected under 'VMworld2018EU'. The main content area shows the 'WindowsTest' VM summary, including its status 'Powered Off', guest OS 'Microsoft Windows Server 2016 (64-bit)', and hardware details. The 'ACTIONS' menu is open, and the 'Switch Theme' option is highlighted with a red box. Other options in the menu include 'Change Password', 'Change Time Format', 'Change VM Console Preference', and 'Logout'. The 'VM Hardware' section shows 8 CPU(s), 4 GB memory (0 GB active), and a 40 GB hard disk. The 'Notes' and 'Custom Attributes' sections are also visible.

Clarity Dark Theme

Fully Supported in 6.7 U1

New in 6.7 U1

The screenshot displays the vSphere Client interface in a dark theme. The top navigation bar includes the VMware logo, 'vSphere Client', a menu, a search bar, and the user 'Administrator@VSPHERE.LOCAL'. The left sidebar shows a tree view of the environment, with 'WindowsTest' selected under 'VMworld2018EU'. The main content area shows the 'WindowsTest' VM details, including its status 'Powered Off' and various configuration tabs. A red box highlights the 'ACTIONS' menu, which is open, showing options like 'Change Password', 'Change Time Format', 'Change VM Console Preference', and 'Switch Theme'. The 'Switch Theme' option is also highlighted with a red box. Below the VM details, the 'VM Hardware' section shows CPU (8 CPU(s)), Memory (4 GB, 0 GB memory active), Hard disk 1 (40 GB), and Network adapter 1 (VM Network (disconnected)).

vm vSphere Client Menu Search in all environments Administrator@VSPHERE.LOCAL

ee-vc-1.csl.vmware.com
Datacenter
vb-DC
vb-CL
VMworld2018EU
cs-tse-d95.csl.vmware.com
foo
h5ngcVA
WindowsTest

WindowsTest
ACTIONS

Summary Monitor Configure Permissions Datastores Networks U
Powered Off
Guest OS: Microsoft Windows Server 2016 (64-bit)
Compatibility: ESXi 6.7 and later (VM version 14)
VMware Tools: Not running, version:10305 (Current)
More info
DNS Name: WIN-PBAJ3K1F1ID
IP Addresses:
Host: cs-tse-d95.csl.vmware.com

Launch Web Console
Launch Remote Console

VM Hardware
CPU 8 CPU(s)
Memory 4 GB, 0 GB memory active
Hard disk 1 40 GB
Network adapter 1 VM Network (disconnected)

Change Password
Change Time Format
Change VM Console Preference
Switch Theme
Logout

0 B
STORAGE USAGE
40 GB

Notes
Edit Notes...

Custom Attributes

Latency Sensitivity Options

Since 6.7 also via the HTML 5 client

The screenshot shows the 'Edit Settings' window for a virtual machine named 'WindowsTest'. The 'VM Options' tab is selected. Under the 'Swap file location' section, three radio buttons are visible: 'Default' (selected), 'Virtual machine directory', and 'Datastore specified by host'. Below this is the 'Configuration Parameters' section, which contains a table of settings. A red box highlights the 'Turbo!! No downsides!!' setting, which has a dropdown menu open showing three options: 'No!!!!' (selected), 'Of course!', and 'Of course!'. The 'Fibre Channel NPIV' setting is also visible below it.

Configuration Parameter	Value
Turbo!! No downsides!!	No!!!!
> Fibre Channel NPIV	Fibre Channel NPIV settings

Latency Sensitivity

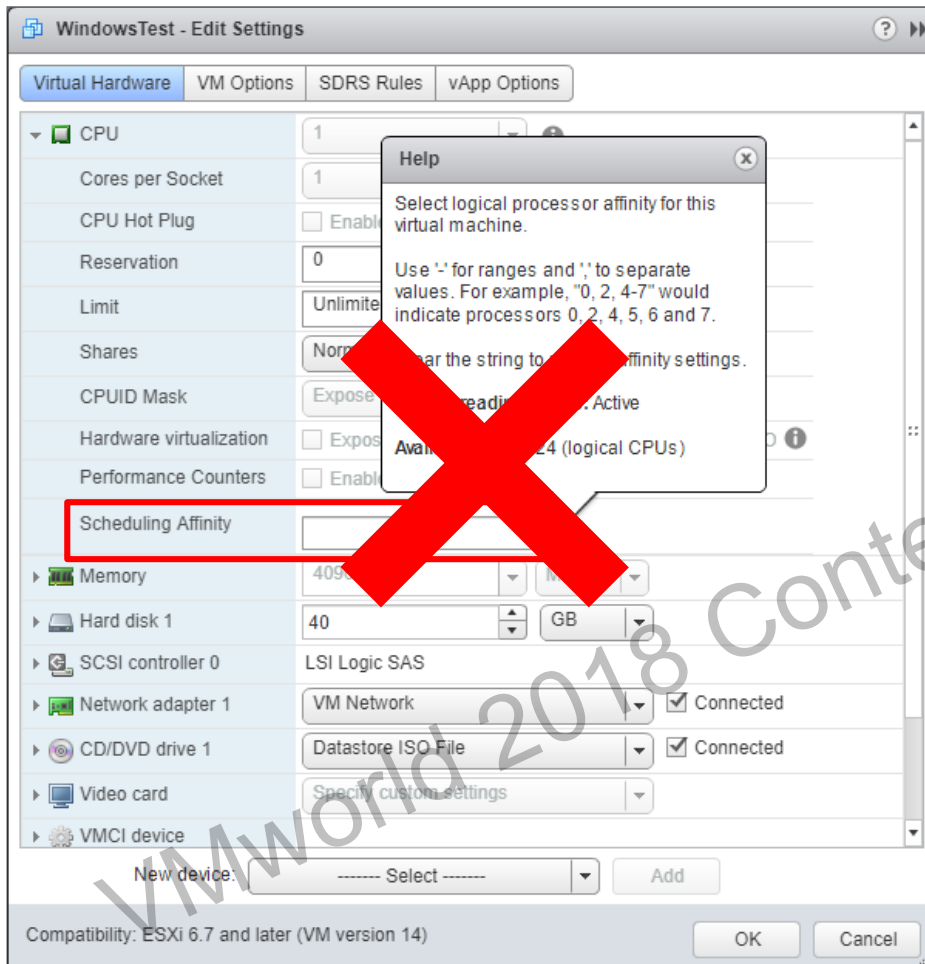
What happens when setting it to High

`sched.cpu.latencySensitivity = high`



Latency Sensitivity = High

Direct and exclusive vCPU to PCPU affinity



It's not the VM vCPU affinity that is selectable in the UI!

- VMX option: “`sched.cpu.affinity`”
- This would sets an affinity for all worlds of the VM
 - Meaning all of the worlds could run on all of the PCPUs

Latency sensitivity sets a per vCPU affinity:

- VMX option: “`sched.vcpuX.affinity = Y`”
- Those vCPUs can only run on the specific PCPU
- Only one vCPU per HyperTwin

+ Exclusive Affinity:

- VMX option: “`sched.cpu.affinity.exclusive`”
- Prevents anything else from executing on the same PCPU

Latency Sensitivity = High

Idle the HyperTwin / partner PCPU and don't deschedule

```
8:57:23pm up 2 days 9:17, 751 worlds, 1 VMs, 4 vCPUs; CPU load average: 0.35, 0.21, 0.08
PCPU USED(%): 0.0 0.2 0.3 0.0 0.2 0.0 0.0 0.0 0.2 0.0 0.0 0.2 0.0 0.1 0.2 0.1 13 0.0 0.0 0.0 0.2 0.1 0.0 0.0 AVG: 0.6
PCPU UTIL(%): 0.1 0.7 100 0.0 100 0.0 0.1 0.1 100 0.0 0.0 100 0.1 0.1 0.4 0.2 10 0.1 0.0 0.1 0.3 0.2 0.2 0.0 AVG: 17
CORE UTIL(%): 0.7 100 100 0.3 100 100 0.2 0.6 10 0.0 0.4 0.2 AVG: 34
```

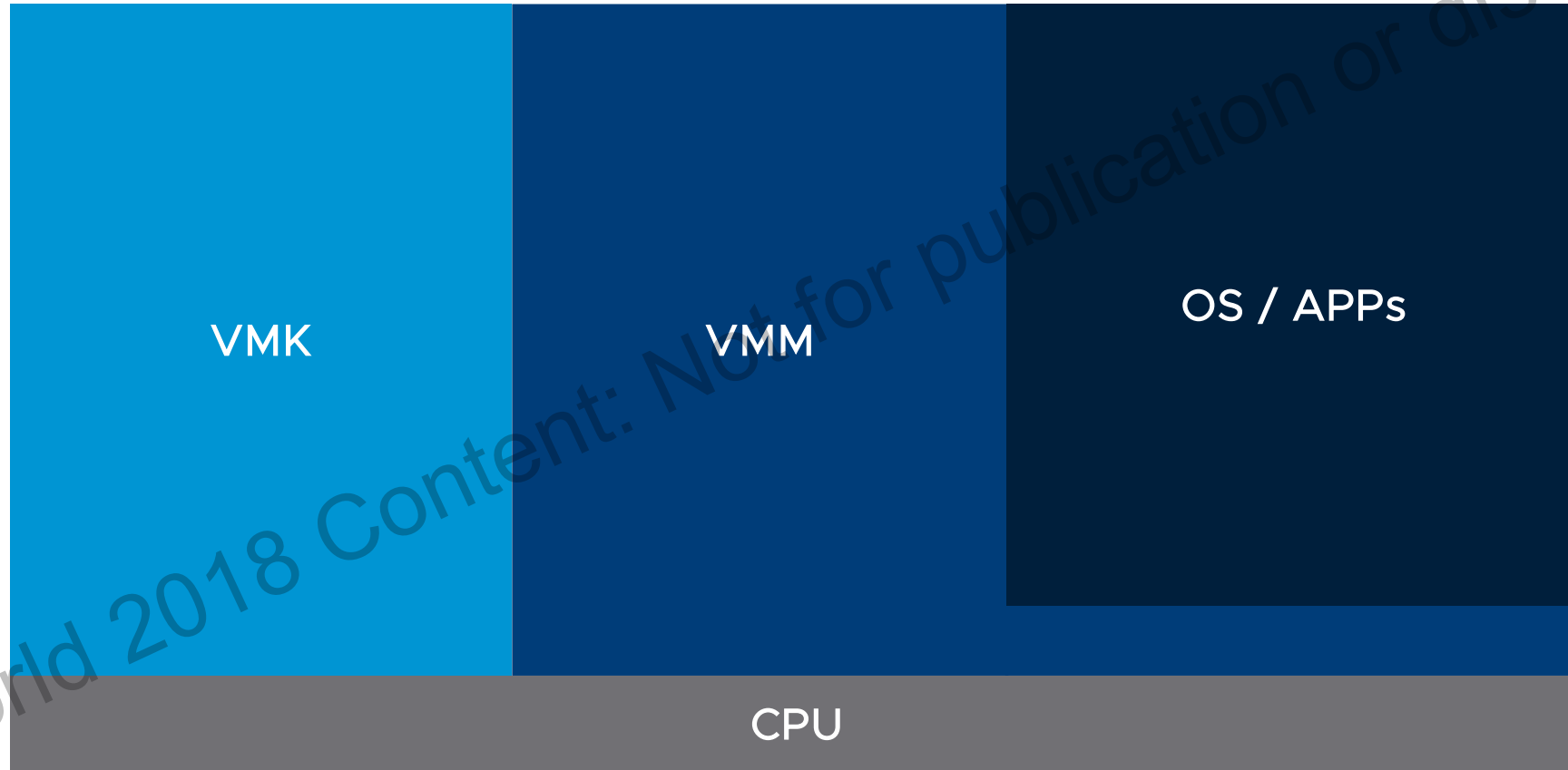
ID	GID	NAME	NWLD	%USED	%RUN	%SYS	%WAIT	%VMWAIT	%RDY	%IDLE	(...)
2127519	153670	vmx	1	0.03	0.03	0.00	100.00	-	0.00	0.00	(...)
2127521	153670	NetWorld-VM-2127520	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)
2127522	153670	NUMASchedRemapEpochInitialize	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)
2127523	153670	vmast.2127520	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)
2127527	153670	vmx-vthread-212	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)
2127528	153670	vmx-filtPoll:WindowsTest	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)
2127529	153670	vmx-mks:WindowsTest	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)
2127530	153670	vmx-svga:WindowsTest	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)
2127531	153670	vmx-vcpu-0:WindowsTest	1	0.31	100.21	0.00	0.00	0.00	0.00	0.00	(...)
2127533	153670	vmx-vcpu-1:WindowsTest	1	0.16	100.21	0.00	0.00	0.00	0.00	0.00	(...)
2127534	153670	vmx-vcpu-2:WindowsTest	1	0.15	100.21	0.00	0.00	0.00	0.00	0.00	(...)
2127535	153670	vmx-vcpu-3:WindowsTest	1	0.15	100.21	0.00	0.00	0.00	0.00	0.00	(...)
2127532	153670	LSI-2127520:0	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)
2127536	153670	vmx-vthread-212:WindowsTest	1	0.00	0.00	0.00	100.00	-	0.00	0.00	(...)

RUN is ~100% when exclusive affinity is active

- VMX option: "monitor_control.halt_desched = false"
- VMX option: "monitor_control.halt_in_monitor = true"

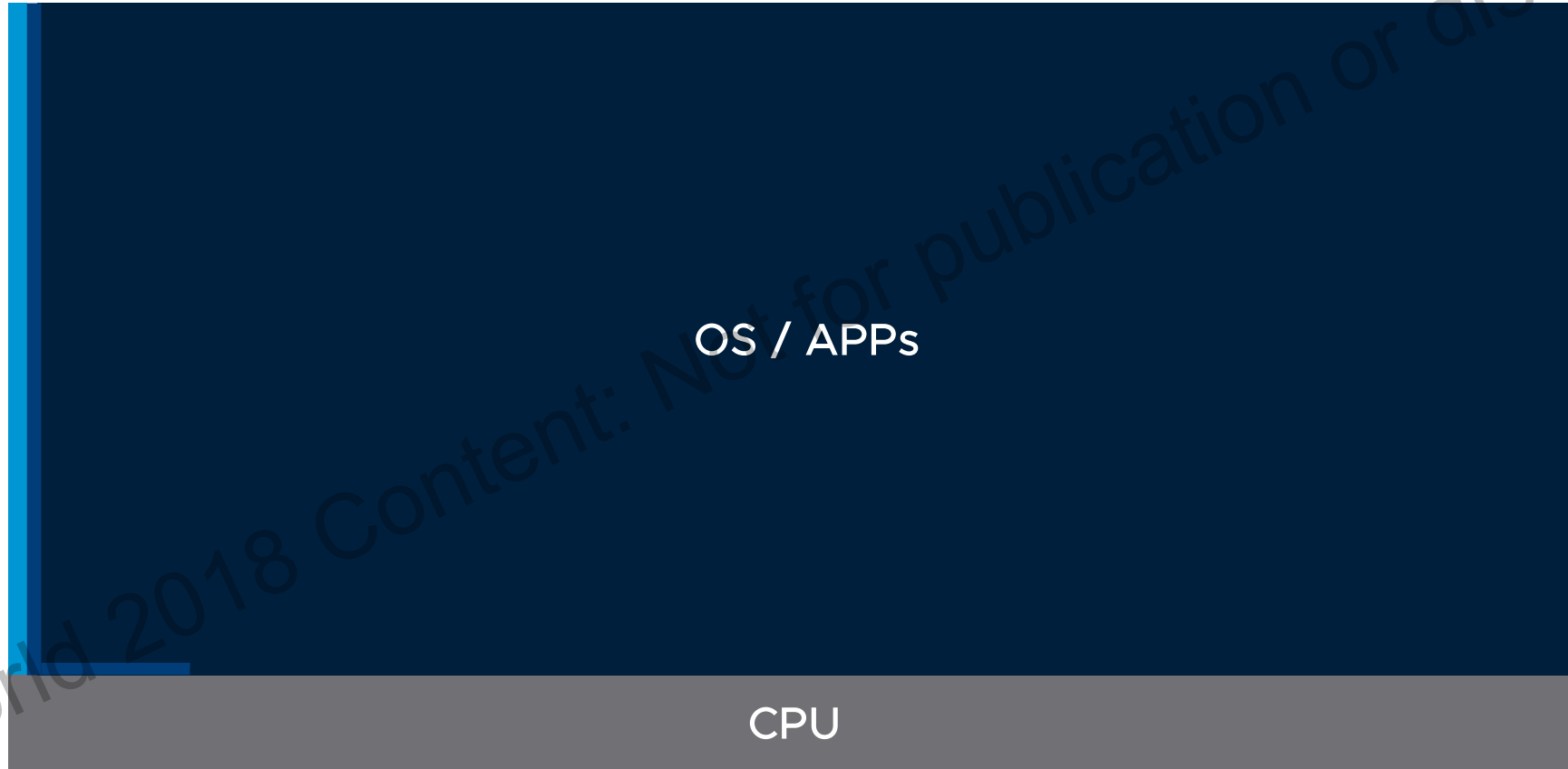
What runs where and when

The high level picture



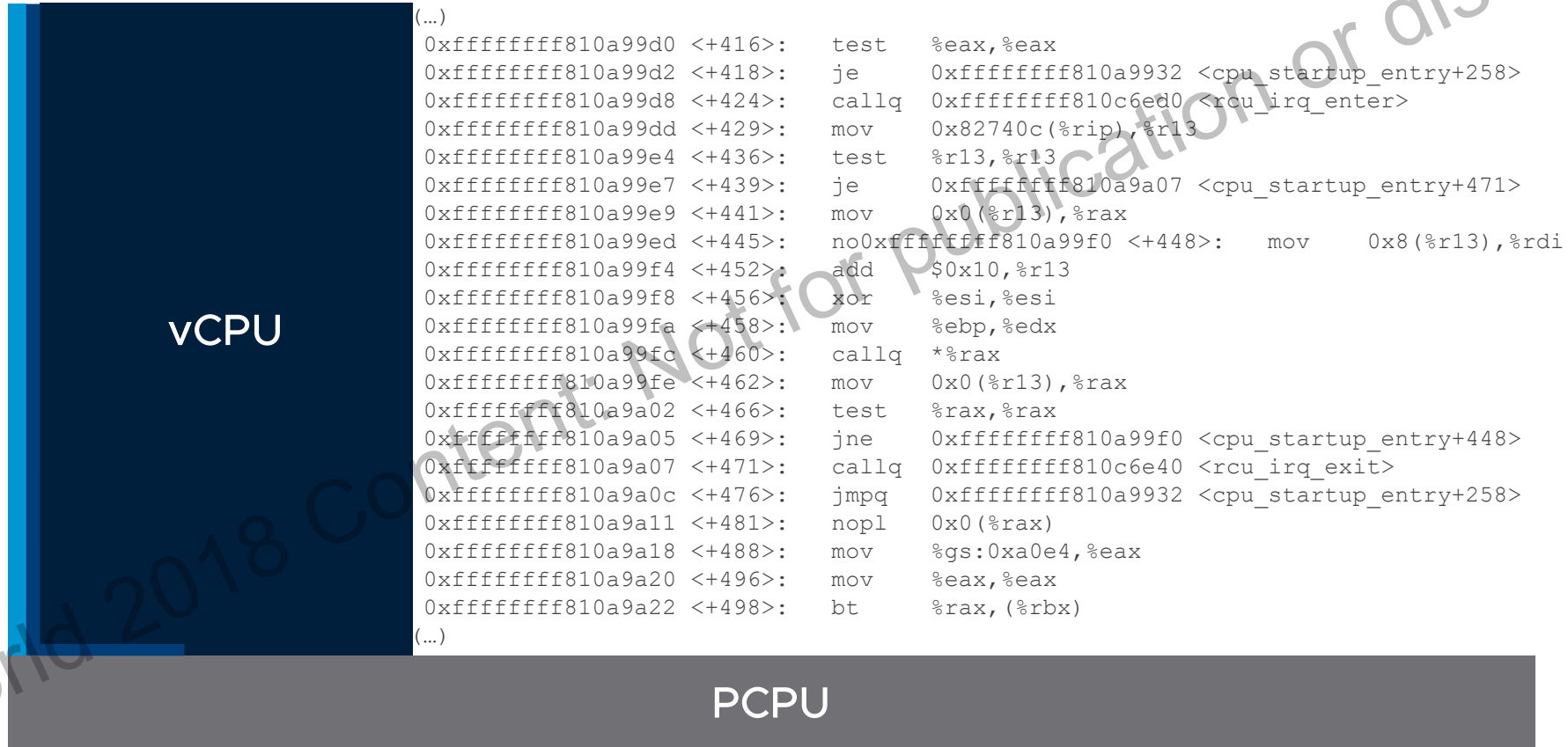
What runs where and when

Mostly Direct Exec



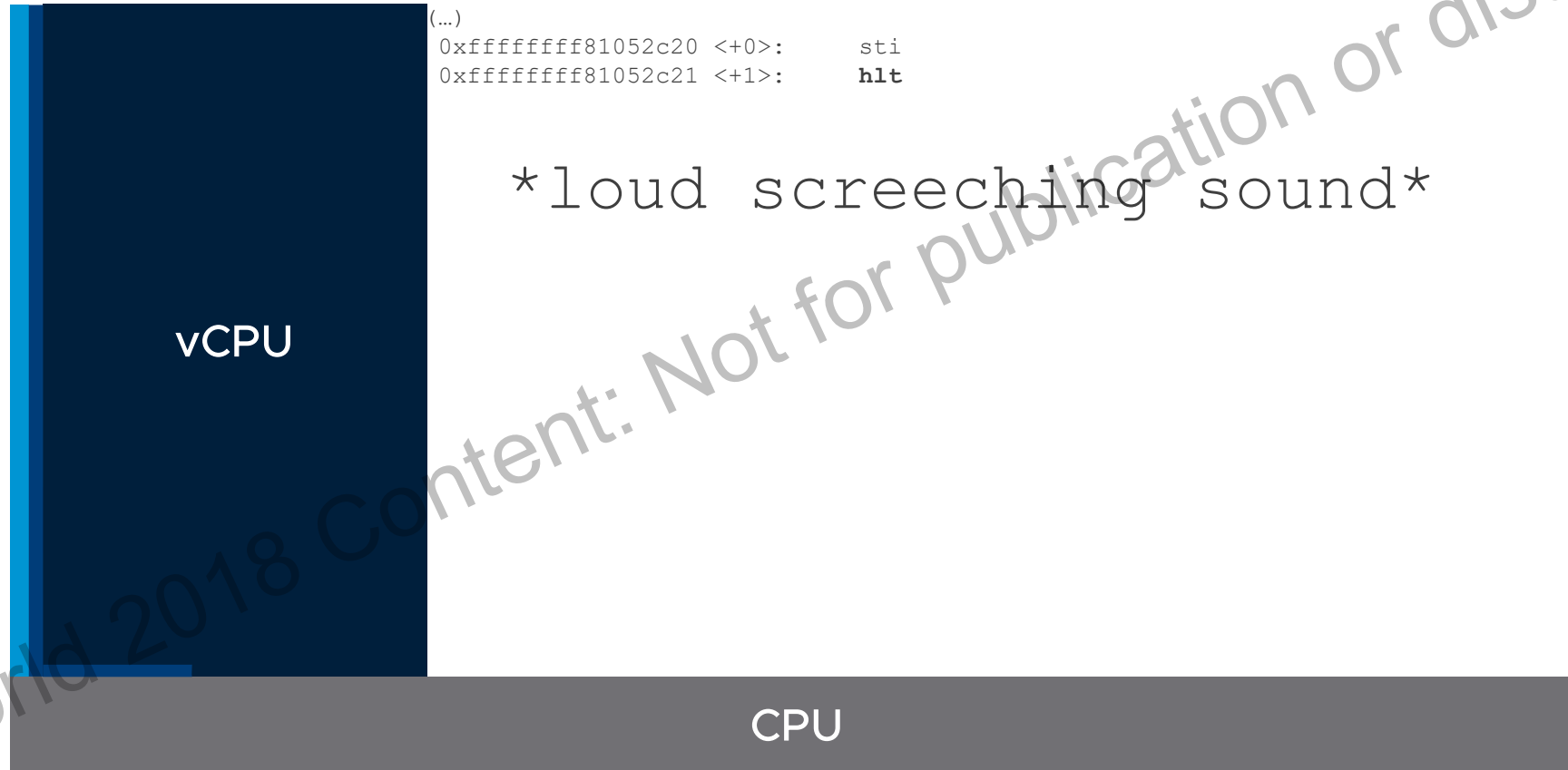
What runs where and when

Mostly Direct Exec



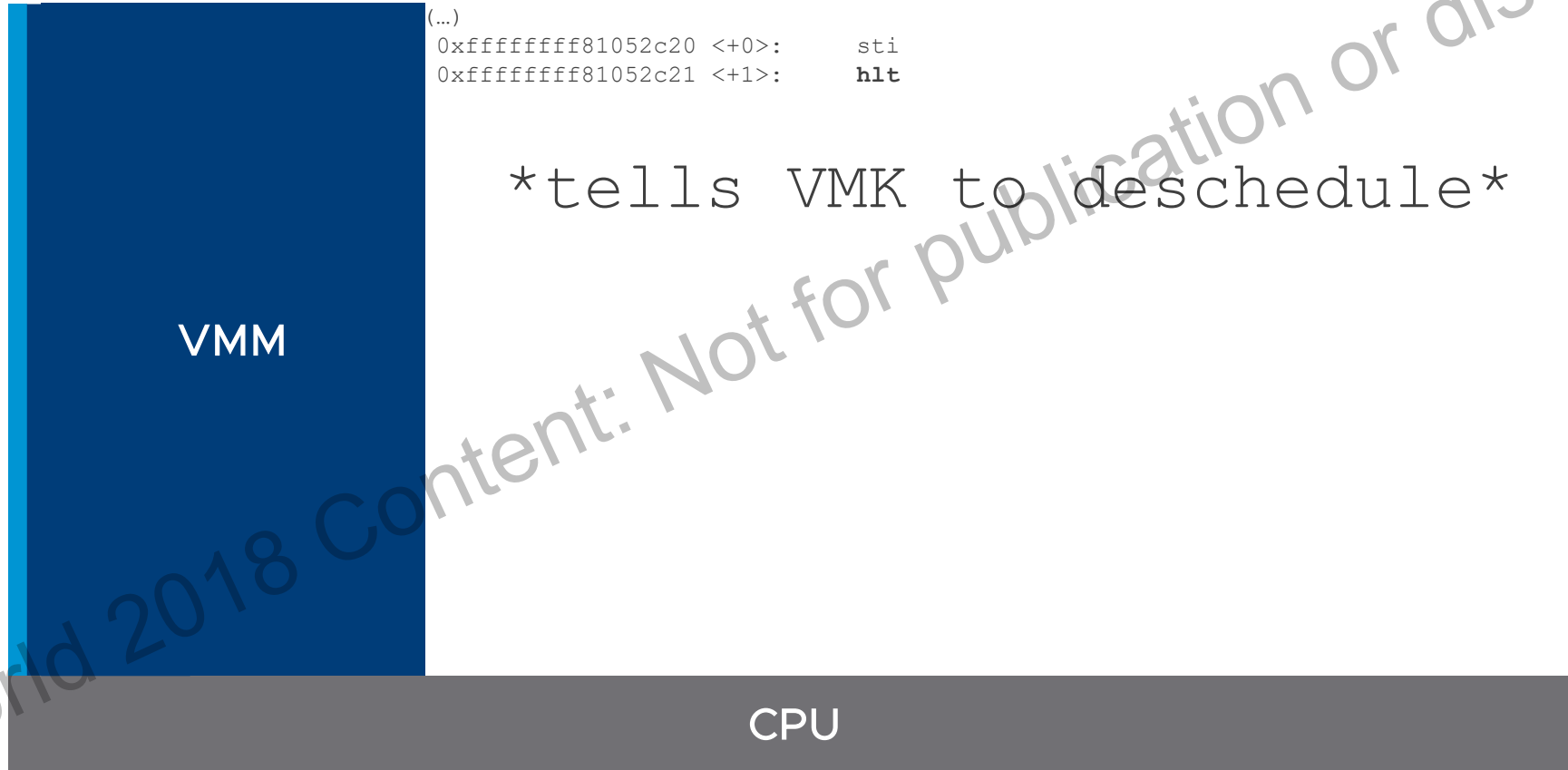
What runs where and when

What about Idle?



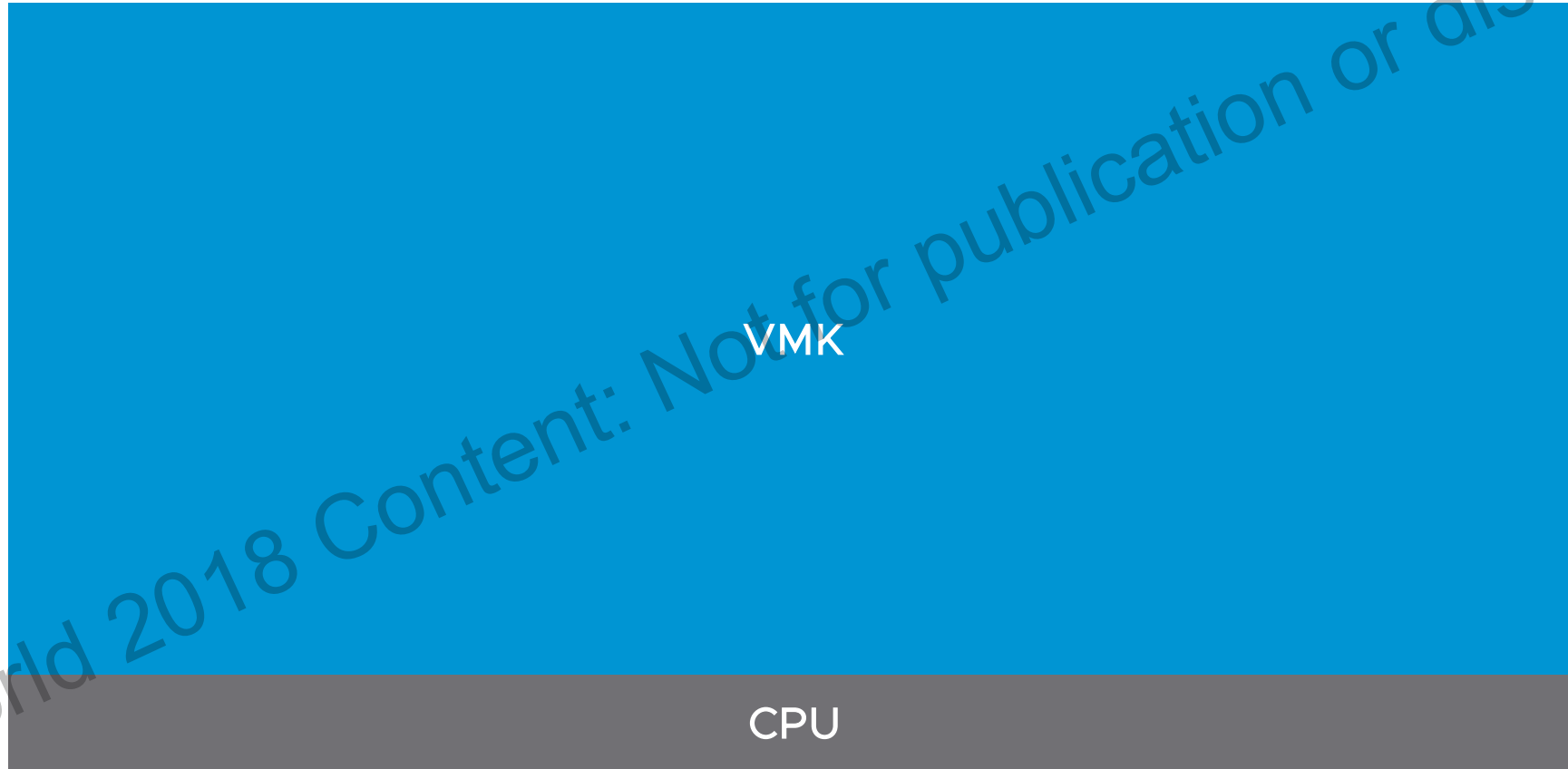
What runs where and when

VMM traps on the privileged instruction and puts (with VMK) the vCPU to “sleep



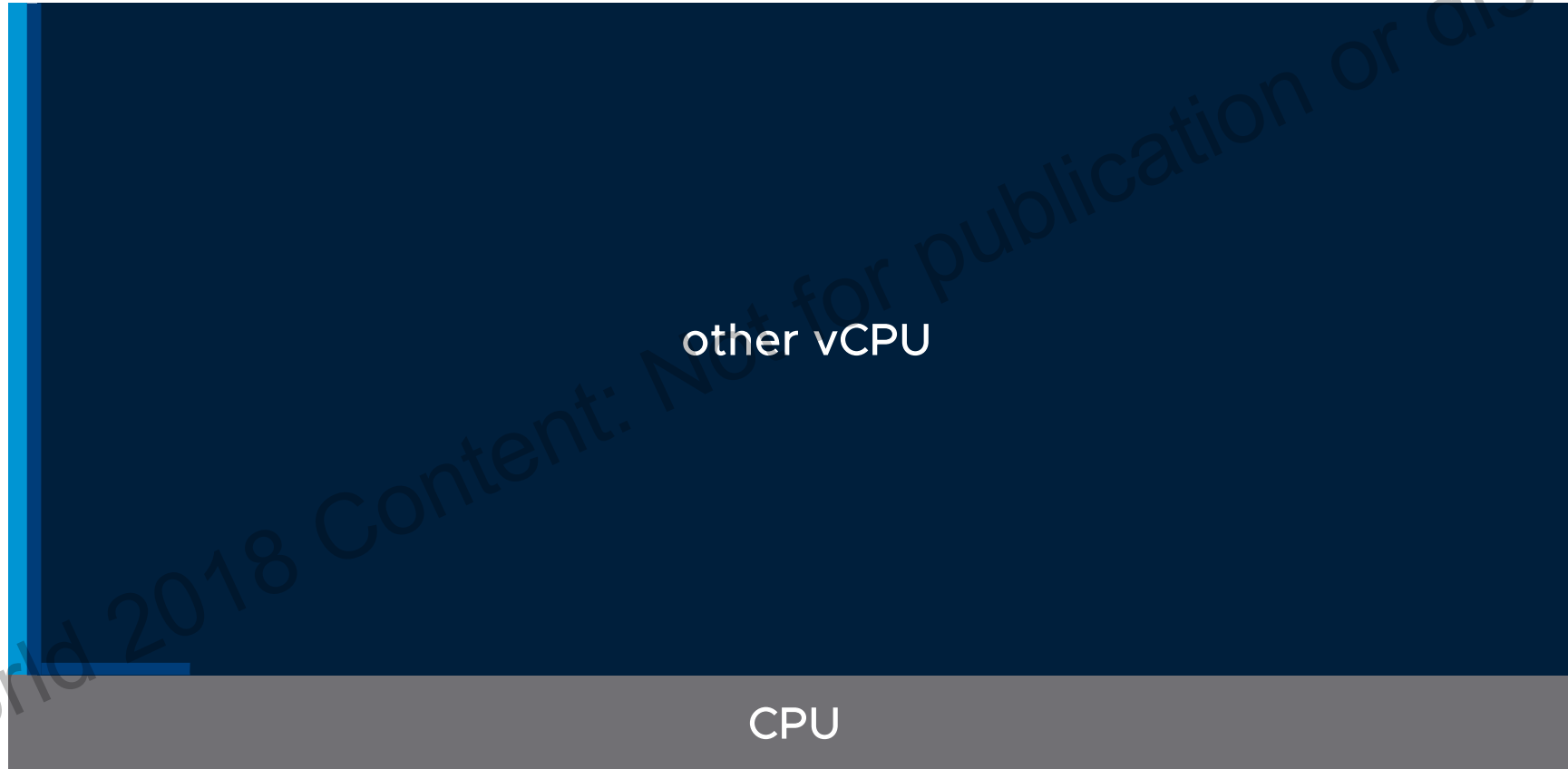
What runs where and when

The scheduler decides what next to run



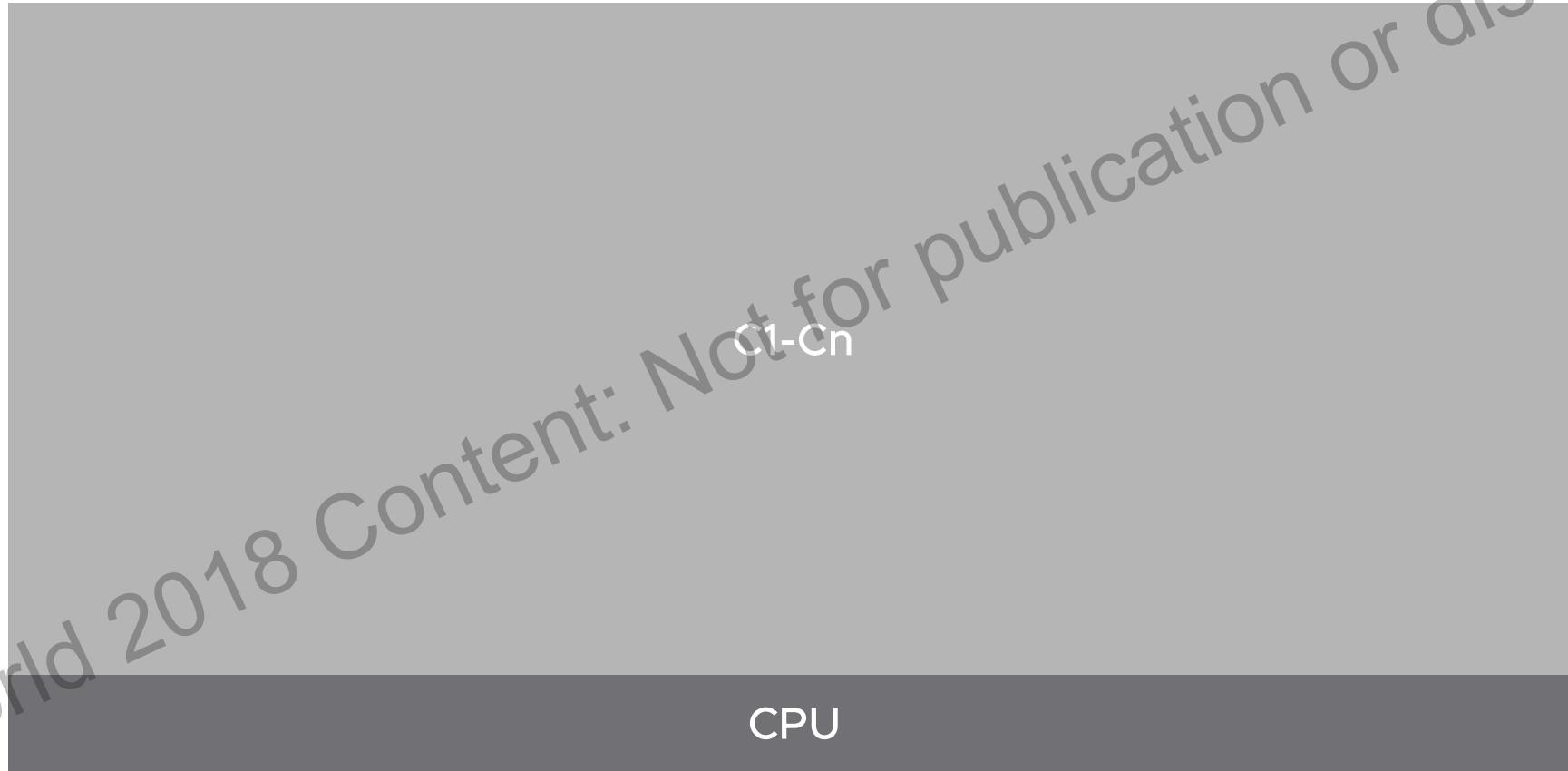
What runs where and when

E.g. a vCPU / world that is ready to run



What runs where and when

ESXi's `_own_` idle thread



Latency Sensitivity = High

Idle the HyperTwin / partner PCPU and don't deschedule

9:01:35pm up 2 days 9:21, 751 worlds, 1 VMs, 4 vCPUs; CPU load average: 0.35, 0.35, 0.17

Power Usage: 136W, Power Cap: N/A

PSTATE MHZ: 2401 2400 2300 2200 2100 2000 1900 1800 1700 1600 1500 1400 1300 1200

CPU	%USED	%UTIL	%C0	%C1	%C2	%P0	%P1	%P2	%P3	%P4	%P5	%P6	%P7	%P8	%P9	%P10	%P11	%P12	%P13	%A/MPERF	
0	0.1	0.2	0	3	97	2	0	0	0	0	0	0	0	0	0	0	0	0	0	98	53.2
1	0.1	0.3	0	4	95	97	0	0	0	0	0	0	0	0	0	0	0	0	0	3	61.6
2	0.3	100.0	100	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63.2
3	0.0	0.0	0	10	90	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58.6
4	0.2	100.0	100	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57.5
5	0.0	0.0	0	11	89	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57.2
6	0.1	0.1	0	10	90	15	0	0	0	0	0	0	0	0	0	0	0	0	0	85	73.2
7	0.2	0.6	1	1	98	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54.7
8	0.2	100.0	100	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	76.0
9	0.0	0.0	0	11	89	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57.0
10	0.0	0.0	0	11	89	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57.0
11	0.2	100.0	100	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63.6
12	3.4	3.6	4	9	87	4	0	1	0	0	0	0	0	0	0	0	0	0	0	95	94.4
13	0.0	0.1	0	14	86	32	0	0	0	0	0	0	0	0	0	0	0	0	0	68	60.2
(...)																					

HyperTwin / Partner PCPU is idled

- Best effort, can still be “broken” by overcommitting via affinity
 - e.g. "numa.nodeAffinity"

Latency Sensitivity = High

How else to verify exclusive affinity?

```
9:49:36pm up 2 days 10:09, 751 worlds, 1 VMs, 4 vCPUs; CPU load average: 0.35, 0.35, 0.35
PCPU USED(%): 0.2 0.0 13 0.0 4.1 0.0 0.1 0.0 103 0.0 0.0 3.7 0.1 0.0 0.4 0.1 0.1 0.0 0.0 0.0 33 0.1 0.1 0.1 AVG: 6.7
PCPU UTIL(%): 0.4 0.0 100 0.0 100 0.0 0.2 0.0 100 0.0 0.0 100 0.2 0.0 0.4 0.1 0.2 0.0 0.0 0.1 25 0.3 0.2 0.2 AVG: 17
CORE UTIL(%): 0.2 100 100 4.3 100 100 0.3 0.0 5.1 4.9 26 0.3 AVG: 36
```

ID	GID	NAME	NWLD	%LAT_C	%LAT_M	%DMD	EMIN	TIMER/s	AFFINITY_BIT_MASK	CPU	EXC_AF
2127519	153670	vmx	1	0.0	0.0	0	1557	-	0-23	17	-
2127521	153670	NetWorld-VM-2127520	1	0.0	0.0	0	2400	-	0-23	0	-
2127522	153670	NUMASchedRemapEpochInitialize	1	0.0	0.0	0	1557	-	0-23	12	-
2127523	153670	vmast.2127520	1	0.0	0.0	0	157	-	0-23	17	-
2127527	153670	vmx-vthread-212	1	0.0	0.0	0	1557	-	0-23	0	-
2127528	153670	vmx-filtPoll:WindowsTest	1	0.0	0.0	0	1557	-	0-23	12	-
2127529	153670	vmx-mks:WindowsTest	1	0.0	0.0	0	1557	-	0-23	16	-
2127530	153670	vmx-svga:WindowsTest	1	0.0	0.0	0	1557	-	0-23	14	-
2127531	153670	vmx-vcpu-0:WindowsTest	1	0.0	0.0	99	6224	-	0-23	2	Y
2127533	153670	vmx-vcpu-1:WindowsTest	1	0.0	0.0	99	6224	-	0-23	8	Y
2127534	153670	vmx-vcpu-2:WindowsTest	1	0.0	0.0	99	6224	-	0-23	4	Y
2127535	153670	vmx-vcpu-3:WindowsTest	1	0.0	0.0	99	6224	-	0-23	11	Y
2127532	153670	LSI-2127520:0	1	0.0	0.0	0	1557	-	0-23	1	-
2127536	153670	vmx-vthread-212:WindowsTest	1	0.0	0.0	0	1557	-	0-23	23	-

Exclusive affinity flag, added to esxtop in ESXi 6.7

- Part of the field “SUMMARY STATS“ in the CPU view
 - “AFFINITY_BIT_MASK“ not set for Latency Sensitive = High VMs

Latency Sensitivity = High

How else to verify exclusive affinity?

```
[root@cs-tse-d95:~] sched-stats -t cpu 2> /dev/null | (line break)
awk 'NR == 1 || $(NF-5) ~ /^vm\.|pool|[0-9]+/ {printf ("%10s %-25s %-15s %+5s %+5s %+8s\n", $1, $4, $17, $19, $21, $22)}'
```

vcpu	name	group	cpu	mode	affinity
2127519	vmx	vm.2127519	13	0	0-23
2127522	NUMASchedRemapEpochI	vm.2127519	14	0	0-23
2127523	vmast.2127520	vm.2127519	17	0	0-23
2127527	vmx-vthread-212	vm.2127519	0	0	0-23
2127528	vmx-filtPoll:Windows	vm.2127519	12	0	0-23
2127529	vmx-mks:WindowsTest	vm.2127519	23	0	0-23
2127530	vmx-svga:WindowsTest	vm.2127519	21	0	0-23
2127531	vmx-vcpu-0:WindowsTe	vm.2127519	2	2	2
2127533	vmx-vcpu-1:WindowsTe	vm.2127519	8	2	8
2127534	vmx-vcpu-2:WindowsTe	vm.2127519	4	2	4
2127535	vmx-vcpu-3:WindowsTe	vm.2127519	11	2	11
2127532	LSI-2127520:0	vm.2127519	7	0	0-23
2127521	NetWorld-VM-2127520	vm.2127519	1	0	0-23
2127536	vmx-vthread-212:Wind	vm.2127519	23	0	0-23

Latency Sensitivity = High

Full CPU Reservation is a requirement since 6.7

Edit Settings | WindowsTest

Virtual Hardware | VM Options

CPU: 4

Cores per Socket: 1 Sockets: 4

CPU Hot Plug: Enable CPU Hot Add

Reservation: 0 MHz

Limit: 24,504 MHz

Shares: 0

CPUID Mask: Expose the VMX ID tag to guest Advanced...

Hardware virtualization: Expose hardware assisted virtualization to the guest OS

Performance Counters: Enable virtualized CPU performance counters

I/O MMU: Enabled

Invalid CPU reservation for the latency-sensitive VM. (sched.cpu.min) should be at least 9596 MHz.

CANCEL OK

“Maximum” is unreserved in user pool pre ESXi 6.5 U2 / 6.7 U1

For earlier releases, error on power on will show the required CPU reservation

Can be disabled via VMX option:

- "latency.enforceCpuMin = false"

Can doesn't mean you should

Latency Sensitivity = High

Ensure all pages are mapped before the guest starts

```
[root@cs-tse-d95:~] memstats -r vm-stats -s name:memSize:touched:consumed:allocTgt:mapped -u mb 2> /dev/null | (linebreak)
sed -n '/ \+name/,/ \+Total/p'
```

name	memSize	allocTgt	consumed	mapped	touched
vm.2127519	4096	4096	4096	4096	4096
Total	4096	4096	4096	4096	4096

All guest and VM memory has to be available immediately:

- VMX option: “`sched.mem.prealloc.pinnedMainMem = true`”
 - Disables: remap, defrag (i.e. memory can’t be “moved” around) and sampling (touched / active metric)
 - Also required for 1GB pages, Fault Tolerance, Passthrough and vGPU
- VMX option: “`sched.mem.prealloc.guestMem = true`”
 - Disables: TPS, balloon, zip, swap
 - Enforces full reservation

Latency Sensitivity = High

Active / touched constantly at 100% is the most visible “symptom”

Virtual machine memory usage Acknowledge Reset To Green

VM Hardware

- > CPU 4 CPU(s)
- > Memory ■ 4 GB, 4 GB memory active
- > Hard disk 1 40 GB
- > Network adapter 1 VM Network (connected)
- > CD/DVD drive 1 Connected

Notes

Custom Attributes

Attribute	Value
-----------	-------

Acknowledge Reset To Green

Issue	Type	Trigger Time	Status
Virtual machine memory usage	Triggered Alarm	08/26/2018, 9:05:06 PM	! Alert

Latency Sensitivity = High

Always requires a full memory reservation

Invalid memory setting: memory reservation (sched.mem.min) should be equal to memsize(4096).

Setting	Value
CPU	4
Memory	4 GB
Reservation	4096 MB
Limit	4 GB
Shares	Default
Memory Hot Plug	Disable
Hard disk 1	40 GB
SCSI controller 0	LSI Logic SAS
Network adapter 1	VM Network

Latency Sensitivity = High

Network latency optimization on the VM level

Disable LRO (Large Receive Offload)

- VM equivalent of “`Net.Vmxnet3SwLRO = false`” (host wide setting)
- Small packets are no longer concatenated into larger ones

Disable (vNIC) coalescing

- VMX option: “`ethernetX.coalescingScheme = disabled`”
- Issue TX immediately and immediately interrupt on RX

Disable Dynamic queueing

- NetQueue feature, load balances and combines less used queues
- Disabling guarantees a single queue for the VM

Network

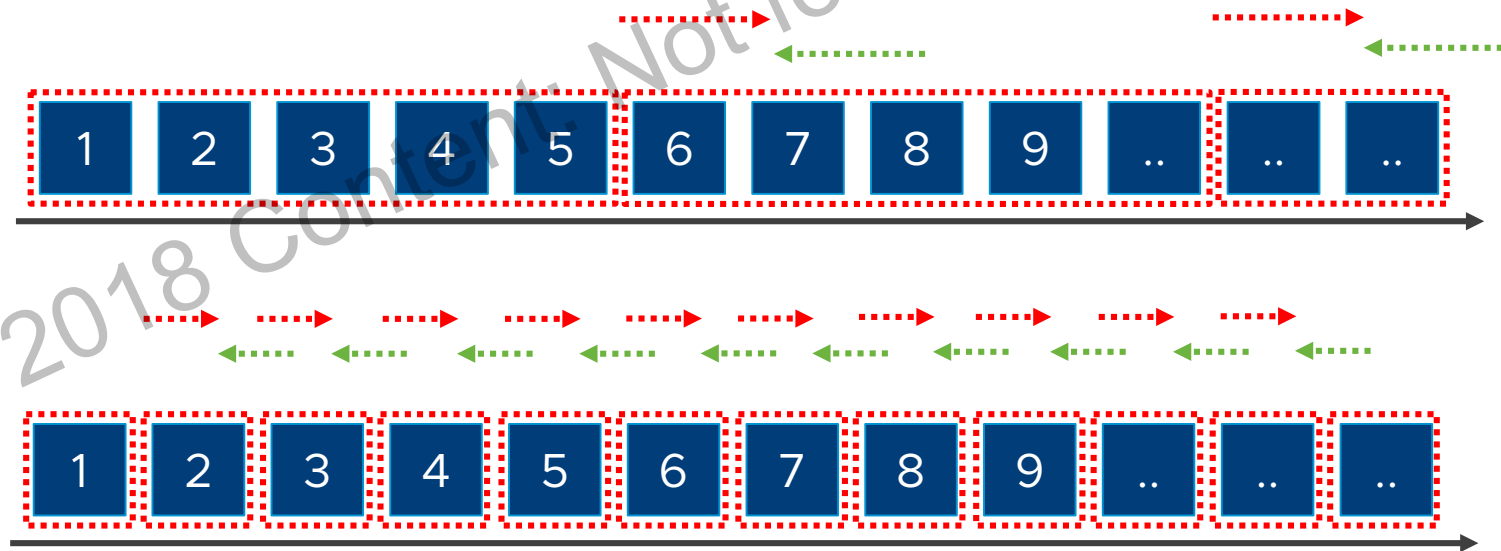


Virtual NIC coalescing - recap

Trading CPU Cycles for Lower Latency

By default, vSphere tunes for lower CPU usage by batching I/O operations

- By default, that is also the case for the RX and TX path on vNICs (here vmxnet3)
- When disabled:
 - Every packet received interrupt immediately
 - Every packet will be issued immediately



Latency Sensitivity = High

Checking Coalescing settings

```
[root@cs-tse-d95:~] net-stats -l
PortNum          Type SubType SwitchName      MACAddress      ClientName
33554434         4     0 vSwitch0      ec:f4:bb:ec:9d:a0 vmnic0
33554436         3     0 vSwitch0      ec:f4:bb:ec:9d:a0 vmk0
50331650         4     0 vSwitch1      ec:f4:bb:ec:9d:a2 vmnic1
50331652         3     0 vSwitch1      00:50:56:6d:67:1c vmk1
50331658         5     9 vSwitch1      00:50:56:98:23:14 WindowsTest
50331659         5     7 vSwitch1      00:50:56:98:67:52 WindowsTest
```

```
[root@cs-tse-d95:~] net-stats -i 1 -t c -p 50331658 | grep coalesce
"coalesce": { "scheme": "disabled", "params": ""},
```

```
[root@cs-tse-d95:~] vsish -e get /net/portsets/vSwitch1/ports/50331658/vmxnet3/rxqueues/0/rxCoalesce
VSISHCmdGetInt():Get failed: Not supported
```

```
[root@cs-tse-d95:~] vsish -e get /net/portsets/vSwitch1/ports/50331658/vmxnet3/txqueues/0/txCoalesce
VSISHCmdGetInt():Get failed: Not supported
```

It's possible to only use the CPU / Memory optimizations of LS=High

- In case you need network throughput at lower CPU cost
- Host wide(!) option: “Net.NetLatencyAwareness = false“

Latency Sensitivity = High

Common Issues

Exclusive affinity disables silently when VM drops below full core entitlement

- Can only happen when VM isn't fully CPU reserved
- Network and Memory related changes stay active
- Check 100% RUN / Exclusive Affinity Bit in esxtop
 - entitled-cores can't be smaller than # vCPUs

```
[root@cs-tse-d95:~] vsish -e get /sched/Vcpus/2132151/stats/debugStats
vcpu-debug-stats {
  wait-event:0
  vtime-aged:0 vtime
  bound-lag-behind:0
  bound-lag-ahead:0
  bound-lag-total:0 vtime
  local-wakeup-sample-rate:1
  remote-wakeup-sample-rate:2
  entitled-cores:8
  extended-cores:0
  progressSema:0
  guestProgress:79803335365329
}
```

Latency Sensitivity = High

Common Issues

Full CPU ($\# \text{ vCPUs} \times \text{Nominal Frequency}$) can't be reserved due to vCenter / ESXi issue

- Fixed in vSphere 6.5 GA / 6.0 U3 / 5.5 U3e – P08

vCD by default limits the VM to exactly $\# \text{ vCPUs} \times \text{Nominal Frequency}$

- Any IO load charged to the VM would cause MLMTD on the vCPU
 - Again, causing Exclusive Affinity to fail
- Set: “Allow CPU resources to grow beyond reserved value”
 - under Organization VDC Properties, Allocation (vCD 9.1 +)

The difference between USED and RUN is wrongly accounted for in LAT_C

- Fixed in vSphere 6.7 GA

Latency Sensitivity = High

Common Issues

Latency Sensitive = High VMs might have 100% NUMA remote memory on ESXi 6.5

- Fixed in vSphere 6.7 GA / not yet in 6.5
- Workaround, VMX option: “`numa.nodeAffinity = <node>`”
 - Only guaranteed to work for 6.5 U1 +

```
[root@cs-tse-d89:~] sched-stats -t numa-clients | awk 'NR == 1 || $2 ~ /8979272/ {print $0}'
```

groupName	groupID	clientID	homeNode	affinity	nWorlds	vmmWorlds	localMem	remoteMem	currLocal%	cummLocal%
vm.1002410911	8979272	0	1	0x3	1	1	8192	516096	1	1

```
[root@cs-tse-d89:~] sched-stats -t numa-migration | awk 'NR == 1 || $2 ~ /8979272/ {print $0}'
```

groupName	groupID	clientID	balanceMig	loadMig	localityMig	longTermMig	monitorMig	pageMigRate
vm.1002410911	8979272	0	0	0	0	0	0	0

Latency Sensitivity = High

Latency Sensitivity Band Classification

Band Classification

- Band 1: <10 μ s (microseconds)
 - Not a good candidate for virtualization today
- Band 2: 10s of μ s
 - Good support since vSphere 5.1, much improved since vSphere 5.5
- Band 3: 100s of μ s to ms (milliseconds)
 - Surprisingly, *might* not be a good candidate because of high cost of Latency Sensitivity = High

Active Memory

VMworld 2018 Content: Not for publication or distribution

Active Memory

Not the same as guest stats!

The screenshot shows the 'VM Hardware' configuration for a virtual machine named 'WindowsTest'. The 'Memory' section is highlighted with a red box, showing '16 GB, 0 GB memory active'. Other hardware components include 8 CPU(s), a 40 GB Hard disk 1, and a VM Network (connected) network adapter.

Component	Configuration
CPU	8 CPU(s)
Memory	16 GB, 0 GB memory active
Hard disk 1	40 GB
Network adapter 1	VM Network (connected)

The screenshot shows the Windows Task Manager Performance tab. The Memory section is highlighted with a red box, showing '7.2 GB (1.5 GB) 8.5 GB' available. The total memory is 16.0 GB, with 15.9 GB used. The 'In use (Compressed)' and 'Available' values are highlighted in red.

Category	Value
Memory	7.4 / 15.9 GB (47%)
Disk 0 (C:)	3%
Ethernet	S: 0 R: 0 Kbps
Ethernet	Not connected
Ethernet	S: 0 R: 0 Kbps
Ethernet	S: 0 R: 0 Kbps

Memory Usage	Memory Composition
In use (Compressed): 7.2 GB (1.5 GB)	Committed: 13.5 / 23.6 GB
Available: 8.5 GB	Cached: 5.3 GB
	Paged pool: 480 MB
	Non-paged pool: 461 MB

Active Memory

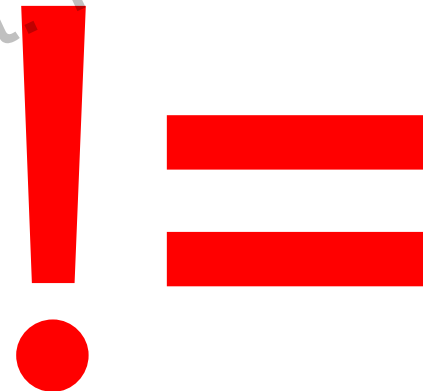
Not the same as guest stats!

The screenshot shows the vSphere VM Summary page for a VM named 'WindowsTest'. The 'VM Hardware' section is expanded to show the 'Memory' configuration, which is highlighted with a red box. The configuration shows '16 GB, 0 GB memory active'. Other hardware details include 8 CPU(s) and a 40 GB hard disk. The guest OS is Microsoft Windows Server 2012 R2, and the VMware Tools are running.

VM Hardware	Configuration
CPU	8 CPU(s)
Memory	16 GB, 0 GB memory active
Hard disk 1	40 GB
Network adapter 1	VM Network (connected)

The screenshot shows the Windows Task Manager Performance tab for Memory. It displays a total of 16.0 GB of memory. The 'Memory usage' bar shows 15.9 GB used. The 'Memory composition' bar shows 7.2 GB in use (1.5 GB compressed) and 8.5 GB available. The 'Committed' memory is 13.5/23.6 GB, and the 'Cached' memory is 5.3 GB. The 'Paged pool' is 480 MB and the 'Non-paged pool' is 461 MB.

Memory	16.0 GB		
Memory usage	15.9 GB		
60 seconds	0		
Memory composition			
In use (Compressed)	Available	Speed:	186...
7.2 GB (1.5 GB)	8.5 GB	Slots used:	2 of 2
Committed	Cached	Form factor:	Chip
13.5/23.6 GB	5.3 GB	Hardware reserved:	153 MB
Paged pool	Non-paged pool		
480 MB	461 MB		

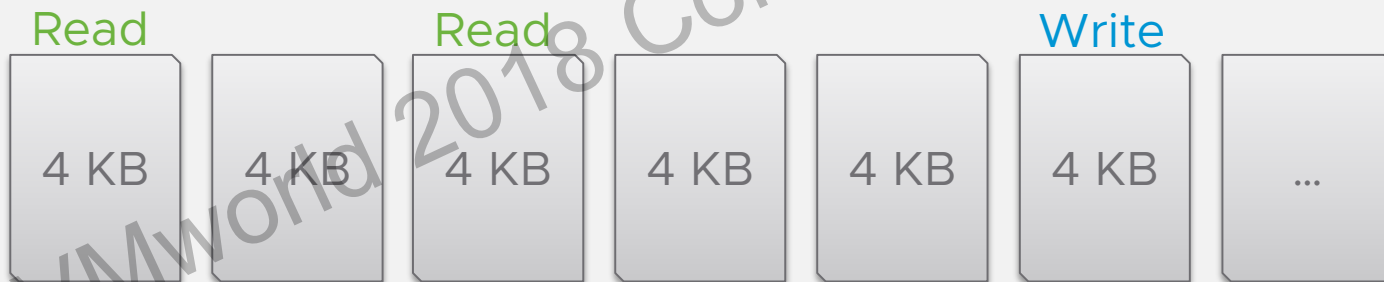


Active Memory

aka Touched

VM mapped memory

100 x 4 KB / min



ESXi VM level heuristic

- Weighted, moving average
- OS / VMTTools independent
- “Memory Sampling”

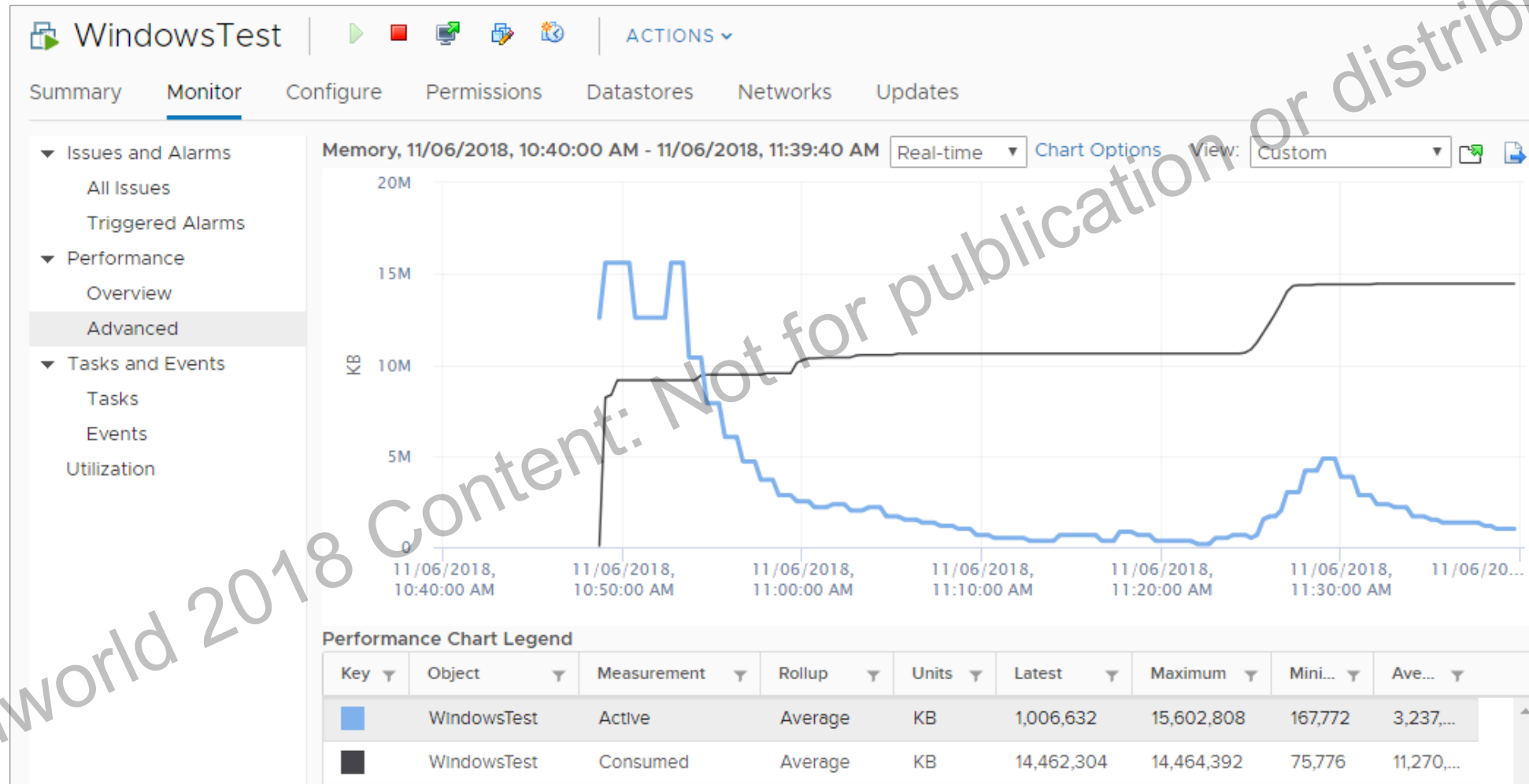
Un-maps 100 random pages over the entire VMs mapped address space

Monitors R/W for a minute (access traps to the VMM)

After one minute, re-maps all remaining pages, starts again

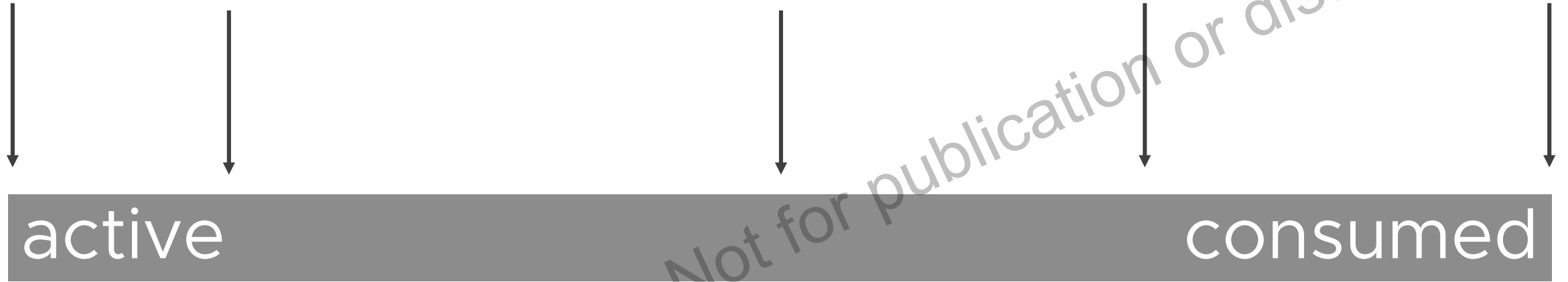
Active Memory

vs. Consumed



Active Memory

What to trust?



Active Memory

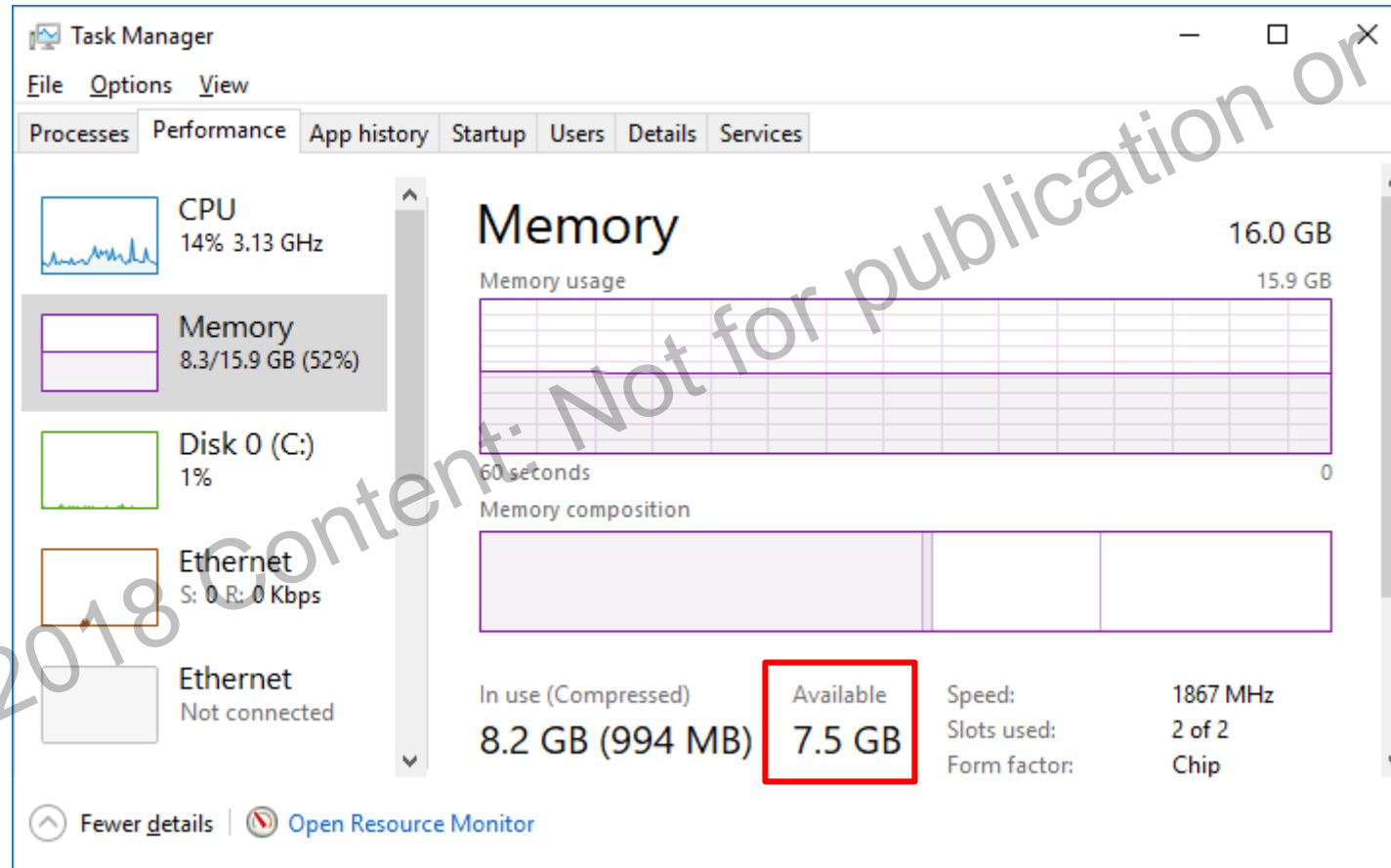
What to trust?

active

consumed

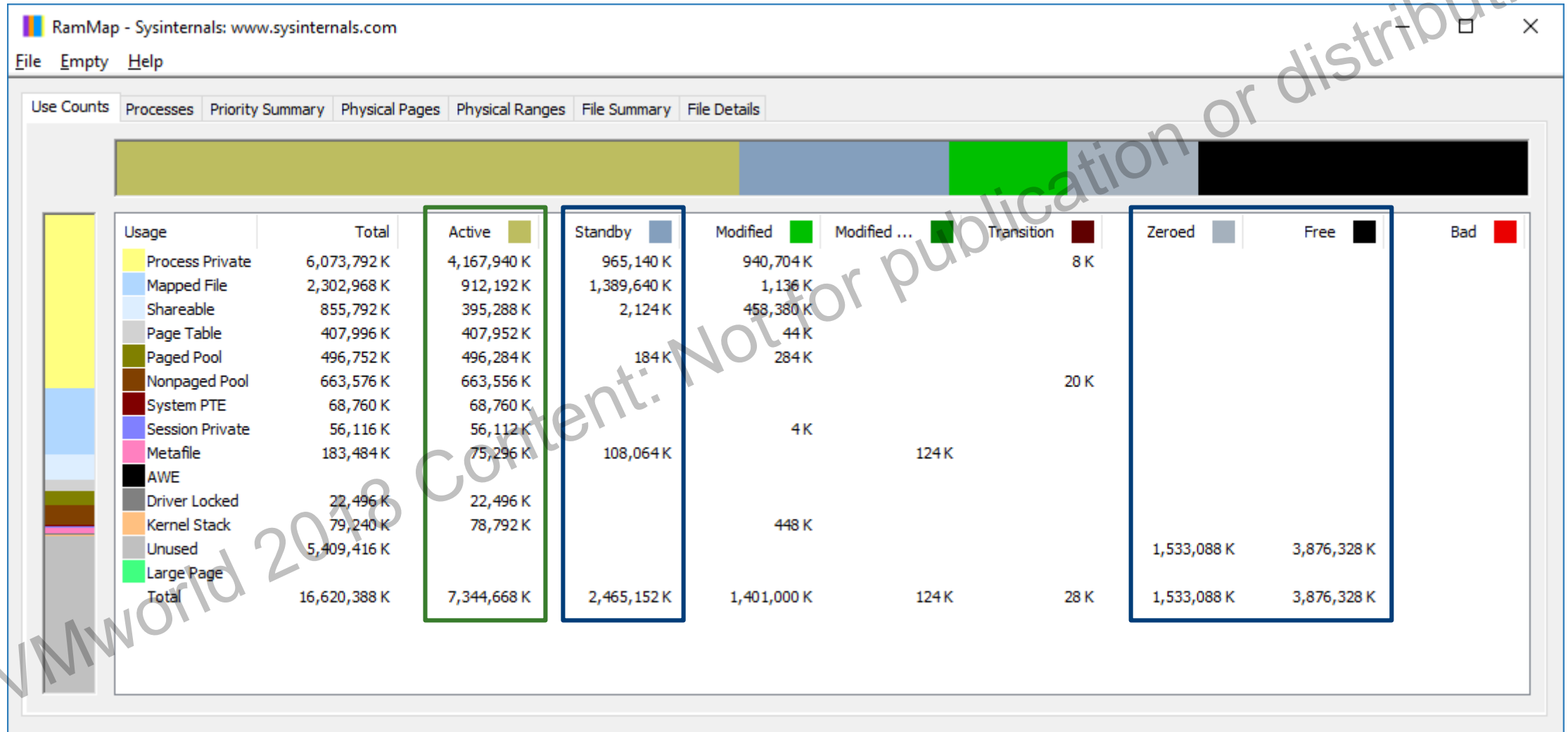
Guest Memory Metrics

In a nutshell



Guest Memory Metrics

In a nutshell



Active Memory

Guests working set tends to be between active and consumed



active guest WS consumed

Active Memory

Guest WS might over report (greedy app)

active

guest WS

Active Memory

But guest WS will not underreport

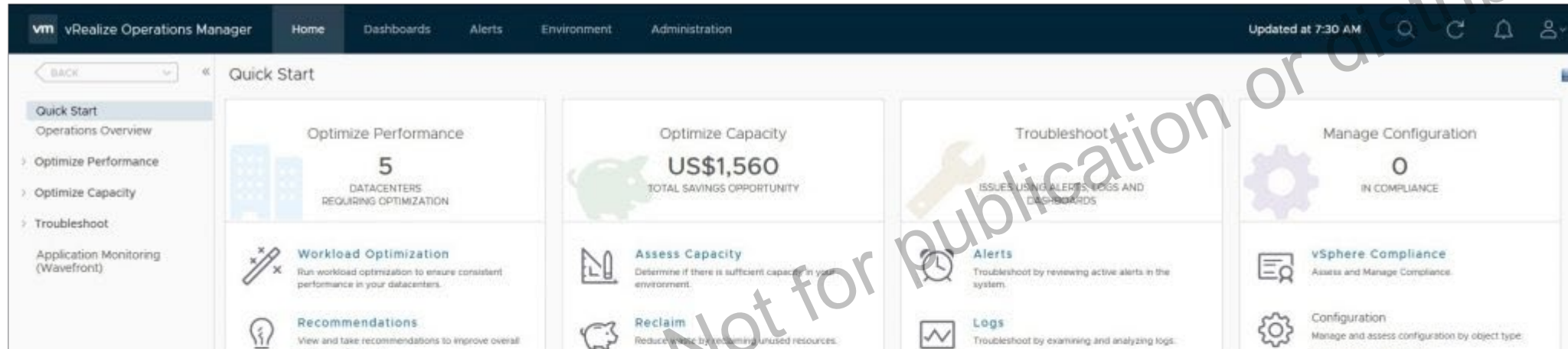
~~guest WS~~

active

consumed

Active Memory

Not then end all of guest workload estimation



Memory Usage % Metric in vRealize Operations Manager 6.7 always shows much higher than the guest OS utilization (55675)

Last Updated: 22.8.2018 Categories: Troubleshooting

✓ Symptoms


- Memory Usage % always shows much higher than the guest OS utilization on some or all applicable objects in vRealize Operations Manager 6.7.

Active Memory

Why do some VMs have a memory usage alarm?

Virtual machine memory usage [Acknowledge](#) [Reset To Green](#)

VM Hardware

- > CPU 4 CPU(s)
- > Memory  4 GB, 4 GB memory active
- > Hard disk 1 40 GB
- > Network adapter 1 VM Network (connected)
- > CD/DVD drive 1 Connected


Notes

[Edit Notes...](#)

Custom Attributes

Attribute	Value
-----------	-------

[Acknowledge](#) [Reset To Green](#)

Issue	Type	Trigger Time	Status
Virtual machine memory usage	Triggered Alarm	08/26/2018, 9:05:06 PM	 Alert

Active Memory

Why do some VMs have a memory usage alarm?

VM memory usage heuristic over-reporting on ESXi 6.5 (2149496)

Last Updated: 25.7.2017 Categories: Informational

Memory usage alarm triggers for certain types of Virtual Machines in ESXi 6.x (2149787)

Last Updated: 28.4.2017 Categories: Informational

✓ Symptoms

- Virtual machine memory usage alarm is displayed for the virtual machine.
- The Virtual Machine's Memory Usage/Active performance metric is continually reported as 100%.
- On vSphere 6.0, the Virtual Machine is configured with Latency Sensitivity set to High.
- On vSphere 6.5 and later the Virtual Machine is configured with at least one of these options:
 - PCI passthrough devices
 - Fault Tolerance (FT) enabled
 - Latency Sensitivity set to High

Extreme Performance Series - Sessions

VIN2677BE Performance Best Practices

VAP1492BE Performance of SQL Server, Oracle, and SAP workloads in VMware Cloud on AWS

VAP1900BE High Performance Big Data and Machine Learning on VMware Cloud on AWS

VIN2183BE vSphere PMEM = Storage at Memory Speed

VAP1620BE Improve App Performance with Micro-Segmentation and Distributed Routing

NFV2917BE Breaking the Virtual Speed Limit: Data Plane Performance Tuning

VIN1759BE vCenter Performance Deep Dive

VIN1782BE vSphere Compute & Memory Schedulers

Extreme Performance Series – Hand On Labs

SPL-1904-01-SDC vSphere 6.7 Performance Diagnostics and Benchmarking

- Each module dives deep into vSphere performance best practices, diagnostics, and optimizations using various interfaces and benchmarking tools.

SPL-1904-02-CHG VMware vSphere 6.7 – Challenge Lab

- Each module places you in a different fictional scenario to fix common vSphere operational and performance problems.

SPL-1947-01-EMT Machine Learning Workloads in vSphere Using GPUs – Getting Started

- Explore how to accelerate Machine Learning Workloads on vSphere using GPUs. Learn about Passthrough and vGPU NVIDIA GRID mechanisms to access GPU from a VM, how to run machine learning workloads using TensorFlow and GPUs on vSphere.

POSSIBLE
BEGINS
WITH YOU

DON'T FORGET
TO FILL OUT
YOUR SURVEY.

VMworld 2018 Content: Not for publication or distribution

#vmworld

#VIN2677BE

POSSIBLE
BEGINS
WITH YOU

THANK YOU!

VMworld 2018 Content: Not for publication or distribution

#vmworld

#VIN2677BE