Going Beyond 100%*

The Little Hypervisor That Could

Valentin Bondzio
Sr. Staff TSE / Global Support

2021-07-20

“Light at the end of the tunnel” Edition

Agenda

Intro

Motivation

CPU

Usage / Accounting

Memory

Consumed / Accounting

Takeaway*

*Hopefully

Motivation

Profitable + Sustainable = Efficient

“We are trying to stay around 60-70% average and not exceed 85% peak utilization.”

Rule of Thumb

vSphere Client | Menu | Search in all environments

ACTIONS

Summary | Monitor | Configure | Permissions | Hosts | VMs | Datastores | Networks | Updates

Hosts | Resource Pools

Filter

Name	State	Status	Cluster	Consumed CPU %	Consumed Memor...	HA State
l.vmware.com	Connected	Warning		98%		N/A
l.vmware.com	Connected	Warning		98%		N/A
l.vmware.com	Connected	Warning		98		N/A
l.vmware.com	Connected	Warning				N/A
l.vmware.com	Connected	Warning				N/A
l.vmware.com	Connected	Warning				N/A
l.vmware.com	Connected	Warning				N/A
l.vmware.com	Connected	Warning				N/A
l.vmware.com	Connected	Warning				N/A
l.vmware.com	Connected	Warning				N/A

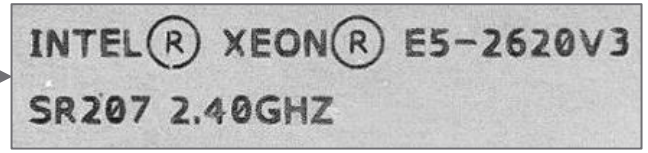
Export | 4 items

CPU Usage (host level) is accounted:
at nominal frequency and without SMT

100% Host CPU Usage =
nominal frequency * number of cores

Nominal (Base) Frequency

“What’s **on** the box”




```
[root@cs-tse-d93:~] vsish -e get /hardware/cpu/cpuModelName
Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz

[root@cs-tse-d93:~] gzip -dc /var/log/boot.gz | grep -i timer.*measured
TSC: 4945939461 cpu0:1)Timer: 737: cpu 0: early measured tsc speed is 2399997831 Hz
0:00:00:02.412 cpu0:1)Timer: 1196: cpu 0: measured cpu speed: 2399998045 Hz, TSC speed: 2399997831 Hz, bus speed: 99999904 Hz
```

Nominal (Base) Frequency and Beyond

Turbo Boost

Table 2. Intel® Xeon® Processor E5-1600, E5-2400, E5-2600 and E5-4600 v3 Product Families Turbo Bins (Sheet 2 of 3)

S-Spec No	Stepping	Model Number	TDP (W)	# Cores	Intel® Turbo Boost Technology Maximum Core Frequency (GHz)										Notes
					Core 1 -2	Core 3	Core 4	Core 5	Core 6	Core 7	Core 8	Core 9	Core 10	Core 11+	
SR1XV	M1	E5-2658 v3	105	12	2.9	2.7	2.6	2.5	2.5	2.5	2.5	2.5	2.5	2.5	1,2,3,7
SR1XW	M1	E5-2648L v3	75	12	2.5	2.3	2.2	2.1	2.1	2.1	2.1	2.1	2.1	2.1	1,2,3,7
SR1XZ	M1	E5-2628L v3	75	10	2.5	2.3	2.2	2.2	2.2	2.2	2.2	2.2	2.2	NA	1,2,3,7
SR20H	R2	E5-1680 v3	140	8	3.8	3.6	3.5	3.5	3.5	3.5	3.5	NA	NA	NA	1,2,3,6,7
SR20J	R2	E5-1650 v3	140	6	3.8	3.6	3.6	3.6	3.6	NA	NA	NA	NA	NA	1,2,3,6,7
SR207	R2	E5-2620 v3	85	6	3.2	2.9	2.8	2.7	2.6	NA	NA	NA	NA	NA	1,2,3,7
SR1YC	M1	E5-2609 v3	85	6	1.9	1.9	1.9	1.9	1.9	NA	NA	NA	NA	NA	1,2,3,4,5
SR20N	R2	E5-1660 v3	140	8	3.5	3.3	3.3	3.3	3.3	3.3	3.3	NA	NA	NA	1,2,3,6,7
SR20B	R2	E5-2608L v3	50	6	2	2	2	2	2	NA	NA	NA	NA	NA	1,2,3,7
SR20P	R2	E5-1620 v3	140	4	3.6	3.6	3.6	NA	NA	NA	NA	NA	NA	NA	1,2,3,6,7
SR21P	R2	E5-2608L v3	52	6	2	2	2	2	2	NA	NA	NA	NA	NA	1,2,3,7
SR22P	C1	E5-4660 v3	120	14	2.9	2.7	2.6	2.5	2.4	2.4	2.4	2.4	2.4	2.4	1,2,3,7,8
SR22J	C1	E5-4650 v3	105	12	2.8	2.6	2.5	2.4	2.4	2.4	2.4	2.4	2.4	2.4	1,2,3,7,8
SR22L	C1	E5-4640 v3	105	12	2.6	2.4	2.3	2.2	2.1	2.1	2.1	2.1	2.1	2.1	1,2,3,7,8

Intel® Xeon® Processor E5 v3 Product Family
Specification Update
September 2017

9

```
[root@cs-tse-d93:~] sched-stats -t ncpus
24 PCPUs
12 cores
 2 LLCs
 2 packages
 2 NUMA nodes

[root@cs-tse-d93:~] esxtop
 8:49:09pm up 52 days 11:49, 970 worlds, 0 VMs, 0 vCPUs; (...)
Power Usage:  140W, Power Cap:  N/A
PSTATE MHZ: 2401 2400 2300 (...) 1300 1200

CPU %USED %UTIL %C0 %C1 %C2 %P0 %P1 %P2 (...) %P12 %P13 %A/MPERF
  0   0.0   0.1   0 100   0 100    0    0 (...)     0     0     108.3
  1   0.0   0.0   0 100   0 100    0    0 (...)     0     0     108.4
(...)
 11   0.1   0.1   0 100   0 100    0    0 (...)     0     0     108.3
 12   0.1   0.1   0 100   0 100    0    0 (...)     0     0     108.3
(...)
 22   0.0   0.0   0 100   0 100    0    0 (...)     0     0     108.2
 23   0.0   0.0   0 100   0 100    0    0 (...)     0     0     108.3

[root@cs-tse-d93:~] awk "BEGIN{printf \"%.1f\n\", (26 / 24) * 100}"
108.3
```

https://cdrdv2.intel.com/v1/dl/getContent/330785

Nominal (Base) Frequency and Beyond

Turbo Boost

Table 2. Intel® Xeon® Processor E5-1600, E5-2400, E5-2600 and E5-4600 v3 Product Families Turbo Bins (Sheet 2 of 3)

S-Spec No	Stepping	Model Number	TDP (W)	# Cores	Intel® Turbo Boost Technology Maximum Core Frequency (GHz)										Notes
					Core 1 -2	Core 3	Core 4	Core 5	Core 6	Core 7	Core 8	Core 9	Core 10	Core 11+	
SR1XV	M1	E5-2658 v3	105	12	2.9	2.7	2.6	2.5	2.5	2.5	2.5	2.5	2.5	2.5	1,2,3,7
SR1XW	M1	E5-2648L v3	75	12	2.5	2.3	2.2	2.1	2.1	2.1	2.1	2.1	2.1	2.1	1,2,3,7
SR1XZ	M1	E5-2628L v3	75	10	2.5	2.3	2.2	2.2	2.2	2.2	2.2	2.2	2.2	NA	1,2,3,7
SR20H	R2	E5-1680 v3	140	8	3.8	3.6	3.5	3.5	3.5	3.5	3.5	NA	NA	NA	1,2,3,6,7
SR20J	R2	E5-1650 v3	140	6	3.8	3.6	3.6	3.6	3.6	NA	NA	NA	NA	NA	1,2,3,6,7
SR207	R2	E5-2620 v3	85	6	3.2	2.9	2.8	2.7	2.6	NA	NA	NA	NA	NA	1,2,3,7
SR1YC	M1	E5-2609 v3	85	6	1.9	1.9	1.9	1.9	1.9	NA	NA	NA	NA	NA	1,2,3,4,5
SR20N	R2	E5-1660 v3	140	8	3.5	3.3	3.3	3.3	3.3	3.3	3.3	NA	NA	NA	1,2,3,6,7
SR20B	R2	E5-2608L v3	50	6	2	2	2	2	2	NA	NA	NA	NA	NA	1,2,3,7
SR20P	R2	E5-1620 v3	140	4	3.6	3.6	3.6	NA	NA	NA	NA	NA	NA	NA	1,2,3,6,7
SR21P	R2	E5-2608L v3	52	6	2	2	2	2	2	NA	NA	NA	NA	NA	1,2,3,7
SR22P	C1	E5-4660 v3	120	14	2.9	2.7	2.6	2.5	2.4	2.4	2.4	2.4	2.4	2.4	1,2,3,7,8
SR22J	C1	E5-4650 v3	105	12	2.8	2.6	2.5	2.4	2.4	2.4	2.4	2.4	2.4	2.4	1,2,3,7,8
SR22L	C1	E5-4640 v3	105	12	2.6	2.4	2.3	2.2	2.1	2.1	2.1	2.1	2.1	2.1	1,2,3,7,8

Intel® Xeon® Processor E5 v3 Product Family
Specification Update
September 2017

9

```
[root@cs-tse-d93:~] sched-stats -t ncpus
24 PCPUs
12 cores
 2 LLCs
 2 packages
 2 NUMA nodes

[root@cs-tse-d93:~] esxtop
 8:49:09pm up 52 days 11:49, 970 worlds, 0 VMs, 0 vCPUs; (...)
Power Usage:  140W, Power Cap:  N/A
PSTATE MHZ: 2401 2400 2300 (...) 1300 1200

CPU %USED %UTIL %C0 %C1 %C2 %P0 %P1 %P2 (...) %P12 %P13 %A/MPERF
  0   0.0   0.1   0 100   0 100    0    0 (...)     0     0     108.3
  1   0.0   0.0   0 100   0 100    0    0 (...)     0     0     108.4
(...)
 11   0.1   0.1   0 100   0 100    0    0 (...)     0     0     108.3
 12   0.1   0.1   0 100   0 100    0    0 (...)     0     0     108.3
(...)
 22   0.0   0.0   0 100   0 100    0    0 (...)     0     0     108.2
 23   0.0   0.0   0 100   0 100    0    0 (...)     0     0     108.3

[root@cs-tse-d93:~] awk "BEGIN{printf \"%.1f\n\", (26 / 24) * 100}"
108.3
```

https://cdrdv2.intel.com/v1/dl/getContent/330785

Nominal (Base) Frequency and Beyond

Turbo Boost on modern CPUs

intel

Non-Intel® Advanced Vector Extensions (Non-Intel® AVX), Intel® Advanced Vector Extensions (Intel® AVX), and Intel® Advanced Vector Extensions 512 (Intel® AVX-512) Turbo Frequencies

Figure 1. 2nd Gen Intel® Xeon® Scalable Processors Non-Intel® AVX Turbo Frequencies

82xx, 62xx, and 52xx Processors

					# of active cores / maximum core frequency in turbo mode (GHz)																								
SKU	Cores	LLC (MB)	TDP (W)	Base non-AVX Core Freq. (GHz)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
8280	28	38.5	205	2.7	4.0	4.0	3.8	3.8	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.5	3.5	3.5	3.3	3.3	3.3	3.3
8276	28	38.5	165	2.2	4.0	4.0	3.8	3.8	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.4	3.4	3.4	3.4	3.1	3.1	3.1	3.0	3.0	3.0	3.0
8270	26	35.75	205	2.7	4.0	4.0	3.8	3.8	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.5	3.5	3.5	3.5	3.4	3.4		
8268	24	35.75	205	2.9	3.9	3.9	3.7	3.7	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.5	3.5	3.5	3.5				
8260	24	35.75	165	2.4	3.9	3.9	3.7	3.7	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.3	3.3	3.3	3.3	3.1	3.1	3.1	3.1				
8256	4	16.5	105	3.8	3.9	3.9	3.9	3.9																								
8253	16	22	125	2.2	3.0	3.0	2.8	2.8	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.5	2.5	2.5	2.5												
6254	18	24.75	200	3.1	4.0	4.0	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	3.9										
6252	24	35.75	150	2.1	3.7	3.7	3.5	3.5	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.2	3.2	3.2	3.2	3.0	3.0	3.0	3.0	2.8	2.8	2.8	2.8				
6248	20	27.5	150	2.5	3.9	3.9	3.7	3.7	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.4	3.4	3.4	3.4	3.2	3.2	3.2	3.2								
6246	12	24.75	165	3.3	4.2	4.2	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1																
6244	8	24.75	150	3.6	4.4	4.4	4.3	4.3	4.3	4.3	4.3	4.3																				
6242	16	22	150	2.8	3.9	3.9	3.7	3.7	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.5	3.5	3.5	3.5												
6240	18	24.75	150	2.6	3.9	3.9	3.7	3.7	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.4	3.4	3.4	3.4	3.3	3.3										
6238	22	30.25	140	2.1	3.7	3.7	3.5	3.5	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.1	3.1	3.1	3.1	2.9	2.9	2.9	2.9	2.8	2.8						
6234	8	24.75	130	3.3	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0																				
6230	20	27.5	125	2.1	3.9	3.9	3.7	3.7	3.6	3.6	3.6	3.6	3.4	3.4	3.4	3.4	3.0	3.0	3.0	3.0	2.8	2.8	2.8	2.8								
6226	12	19.25	125	2.7	3.7	3.7	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5																
5222	4	16.5	105	3.8	3.9	3.9	3.9	3.9																								
5220	18	24.75	125	2.2	3.9	3.9	3.7	3.7	3.6	3.6	3.6	3.6	3.1	3.1	3.1	3.1	2.8	2.8	2.8	2.8	2.7	2.7										

12

2nd Gen Intel® Xeon® Scalable Processors Specification Update

SKU	Cores	NF	MACF	ratio	↓ plus_cores
8276	28	2.2	3.0	1.36	10.2
6238T	22	1.9	2.7	1.42	9.3
6252	24	2.1	2.8	1.33	8.0
5220T	18	1.9	2.7	1.42	7.6
6262V	24	1.9	2.5	1.32	7.6
6238	22	2.1	2.8	1.33	7.3
8260	24	2.4	3.1	1.29	7.0
8270	26	2.7	3.4	1.26	6.7
6230	20	2.1	2.8	1.33	6.7
6222V	20	1.8	2.4	1.33	6.7
6230T	20	2.1	2.8	1.33	6.7
8280	28	2.7	3.3	1.22	6.2
6248	20	2.5	3.2	1.28	5.6
8268	24	2.9	3.5	1.21	5.0
6240	18	2.6	3.3	1.27	4.9
6254	18	3.1	3.9	1.26	4.7
5218T	16	2.1	2.7	1.29	4.6
5220	18	2.2	2.7	1.23	4.1
(…)					

MACF = Maximum All Core Frequency
Maximum Frequency with a “High Performance” BIOS or ESXi policy*

https://www.intel.co.jp/content/www/jp/ja/products/docs/processors/xeon/2nd-gen-xeon-scalable-spec-update.html

Identifying MACF

Not a common industry acronym!

Configure whatever “Maximum / High Performance” Power Management policy available

↵

```
[root@cs-tse-d93:~] esxtop
(…)
```

p, f, f

↵

```
 4:38:51pm up 1 min, 1276 worlds, 0 VMs, 0 vCPUs; CPU load average: 0.02, 0.00, 0.00
Power Usage:  142W, Power Cap:  N/A
PSTATE MHZ:

CPU %USED %UTIL %C0 %C1 %A/MPERF
  0   0.0   0.1   0 100    108.3
  1   0.1   0.1   0 100    108.4
  2   0.1   0.1   0 100    108.3
  3   0.0   0.1   0 100    108.4
 (…)
 20   0.0   0.1   0 100    108.4
 21   0.0   0.0   0 100    108.3
 22   0.0   0.0   0 100    108.3
 23   0.2   0.2   0 100    108.3
```

MACF is 8.3 % on top of NF for this CPU / Host

Cycle “Capacity”

Intel Xeon Platinum 8276

2.20 GHz	NF
3.00 GHz	MACF
4.00 GHz	M2CF

36% additional capacity @ MACF
82% additional speed @ M2CF*

* needs deep C-States / non “High Performance” policy

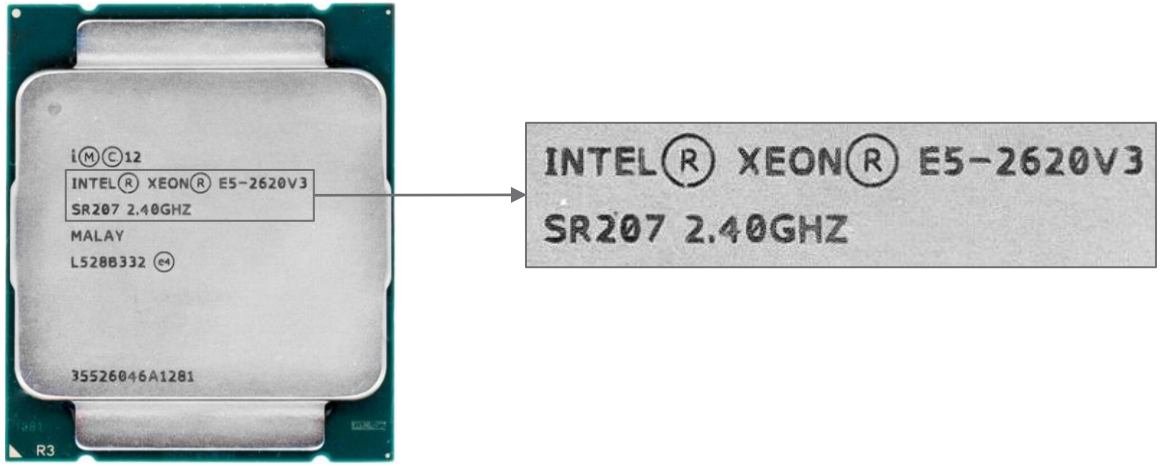
[Modify Frequency Info]

Mode	Base	Turbo Frequency/Active Cores																											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Normal	2,200MHz	4,000MHz	4,000MHz	3,800MHz	3,800MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,700MHz	3,400MHz	3,400MHz	3,400MHz	3,400MHz	3,100MHz	3,100MHz	3,100MHz	3,100MHz	3,000MHz	3,000MHz	3,000MHz	3,000MHz
AVX2	1,700MHz	3,800MHz	3,800MHz	3,600MHz	3,600MHz	3,500MHz	3,500MHz	3,500MHz	3,500MHz	3,500MHz	3,500MHz	3,500MHz	3,500MHz	3,200MHz	3,200MHz	3,200MHz	3,200MHz	2,900MHz	2,900MHz	2,900MHz	2,900MHz	2,700MHz	2,700MHz	2,700MHz	2,700MHz	2,600MHz	2,600MHz	2,600MHz	2,600MHz
AVX512	1,300MHz	3,700MHz	3,700MHz	3,500MHz	3,500MHz	3,300MHz	3,300MHz	3,300MHz	3,300MHz	2,900MHz	2,900MHz	2,900MHz	2,900MHz	2,600MHz	2,600MHz	2,600MHz	2,600MHz	2,300MHz	2,300MHz	2,300MHz	2,300MHz	2,200MHz	2,200MHz	2,200MHz	2,200MHz	2,100MHz	2,100MHz	2,100MHz	2,100MHz

https://en.wikichip.org/wiki/intel/xeon_platinum/8276

Cycle “Capacity”

Intel Xeon E5-2620 v3



2.40 GHz	NF
2.60 GHz	MACF
3.20 GHz	M2CF

8% additional capacity @ MACF
33% additional speed @ M2CF*

* needs deep C-States / non “High Performance” policy

Table 2. Intel® Xeon® Processor E5-1600, E5-2400, E5-2600 and E5-4600 v3 Product Families Turbo Bins (Sheet 2 of 3)

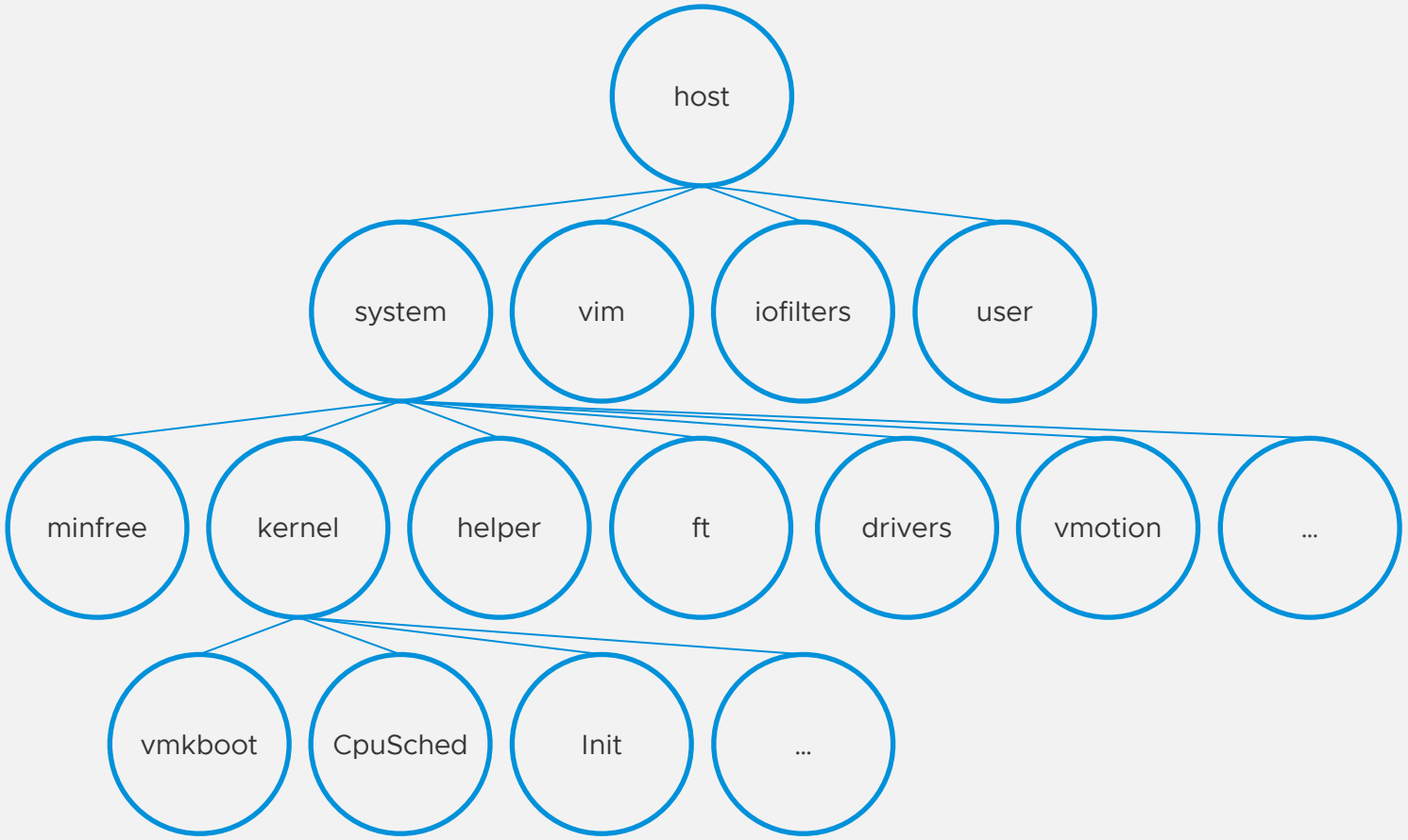
S-Spec No	Stepping	Model Number	TDP (W)	# Cores	Intel® Turbo Boost Technology Maximum Core Frequency (GHz)										Notes
					Core 1 -2	Core 3	Core 4	Core 5	Core 6	Core 7	Core 8	Core 9	Core 10	Core 11+	
SR207	R2	E5-2620 v3	85	6	3.2	2.9	2.8	2.7	2.6	NA	NA	NA	NA	NA	1,2,3,7

https://cdrdv2.intel.com/v1/dl/getContent/330785

Reservable == Capacity?

Hierarchical Resource Groups

From an ESXi perspective



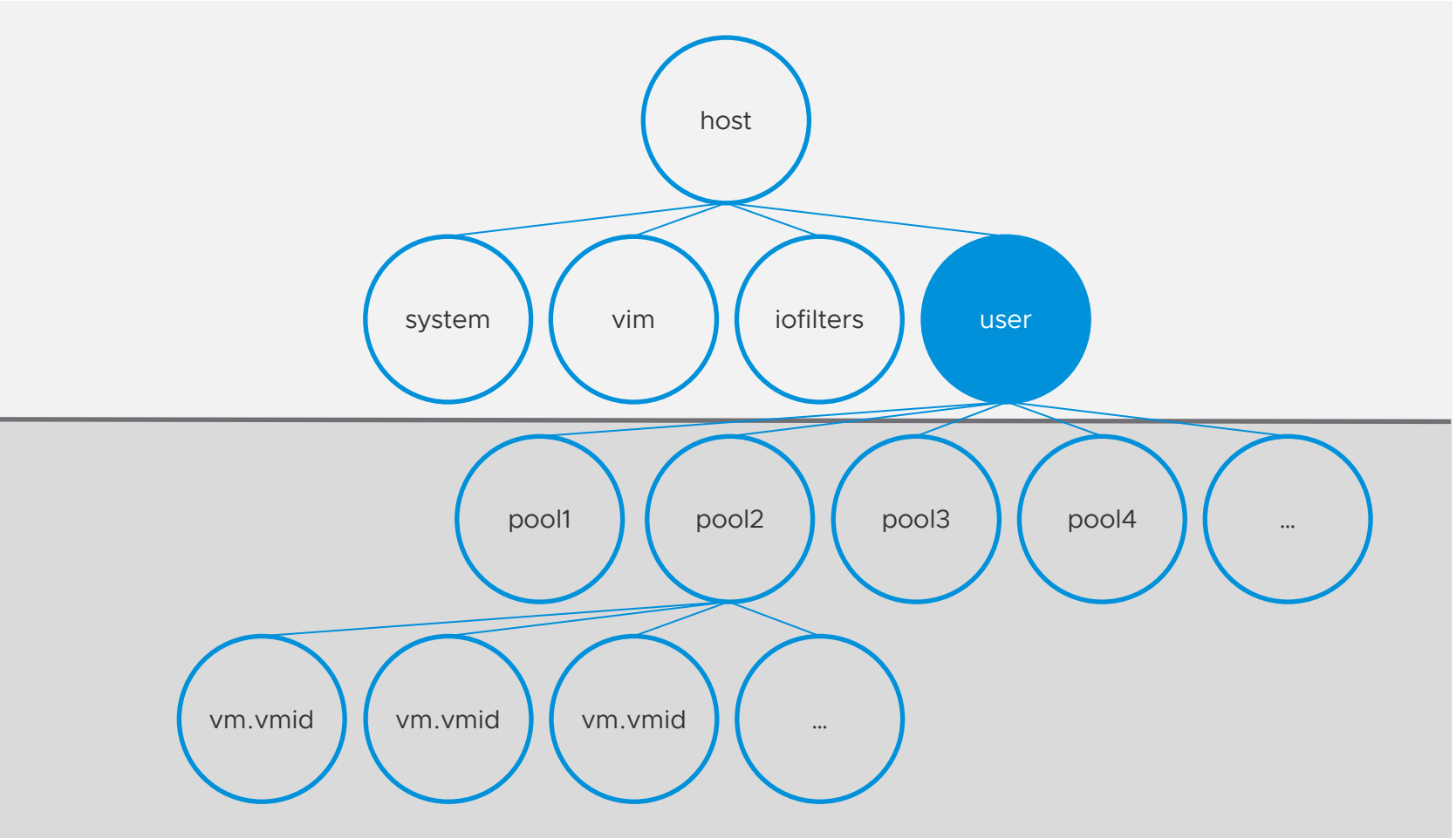
The host owns all resources

Those are distributed by hierarchical resource groups

Consumers can demand (request) resources

Hierarchical Resource Groups

From an ESXi perspective



vCenter shows the sum of all user resources as:

Total Reservation Capacity

Global Resource Pools are then distributed back to hosts into Local RPs

- Based on VMs demand

Cycle “Capacity”

Intel Xeon E5-2620 v3

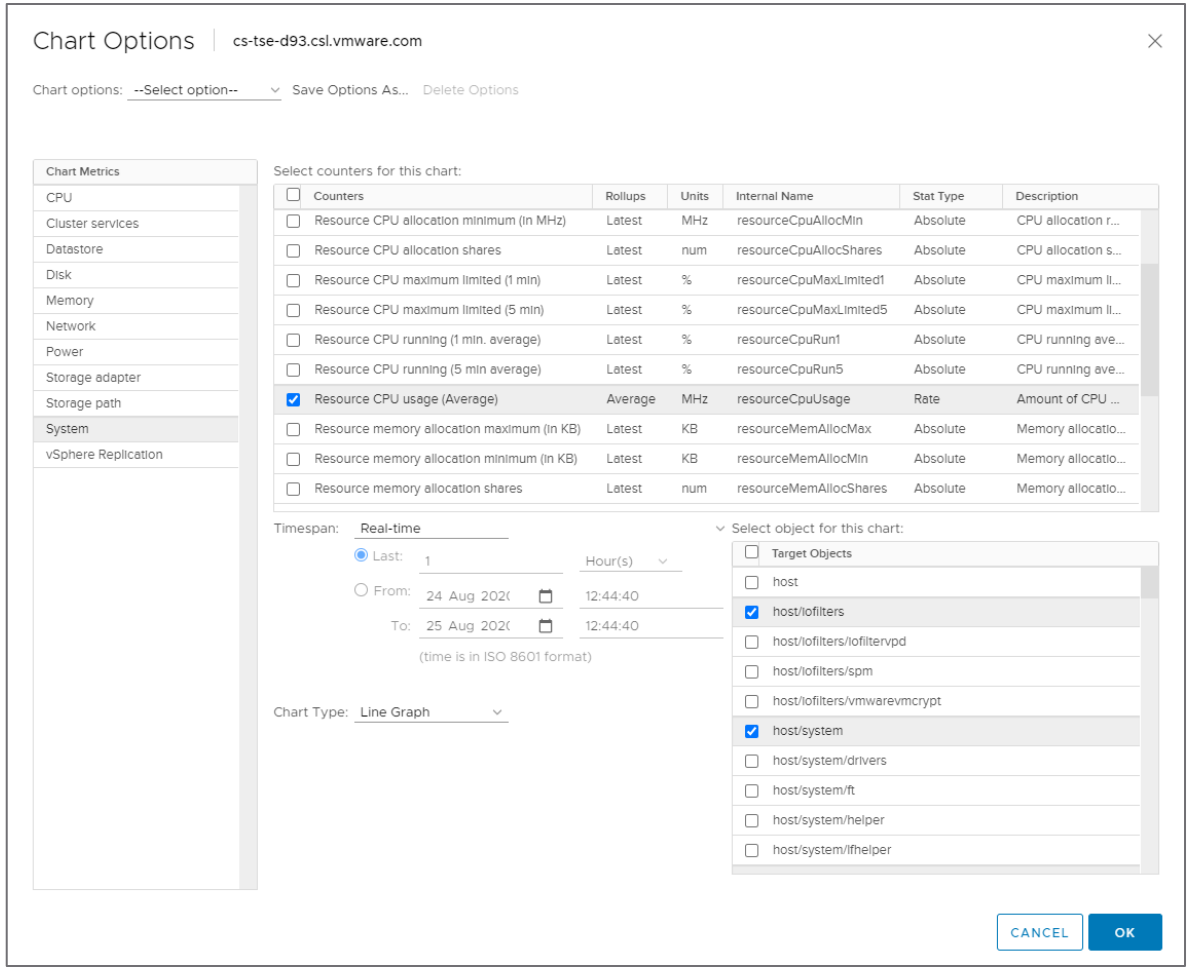
```
[root@cs-tse-d93:~] vsish -e get /sched/groups/0/stats/capacity
group-capacity {
   cpu-reserved:4296 MHz
   cpu-unreserved:24504 MHz
   mem-reserved:7005904 KB
   mem-unreserved:59674248 KB
}
[root@cs-tse-d93:~] awk "BEGIN{printf \"%.0i\n\", (4296 + 24504) / 12 }"
2400

[root@cs-tse-d93:~] vsish -e get /sched/groups/4/stats/capacity
group-capacity {
   cpu-reserved:0 MHz
   cpu-unreserved:24504 MHz
   mem-reserved:33340 KB
   mem-unreserved:59674248 KB
}
[root@cs-tse-d93:~] sched-stats -t groups -z vmgid:name:pgid:pname:size:vsmps:amin:amax:units:resvMHz:availMHz
| awk 'NR == 1 || $2 ~ /^(vm\.|pool)[0-9]+/ || /^[0-4] / {print $0}'

vmgid   name              pgid    pname             size vsmps amin   amax   units    resvMHz   availMHz
0       host              0       none              4    969   1200   1200   pct      4296      24504
1       system            0       host              690  695   10     -1     pct      288       24504
2       vim               0       host              4    271   4008   -1     mhz      3648      24864
3       iofilters         0       host              3    3     0      -1     pct      0         24504
4       user              0       host              0    0     0      -1     pct      0         24504
```

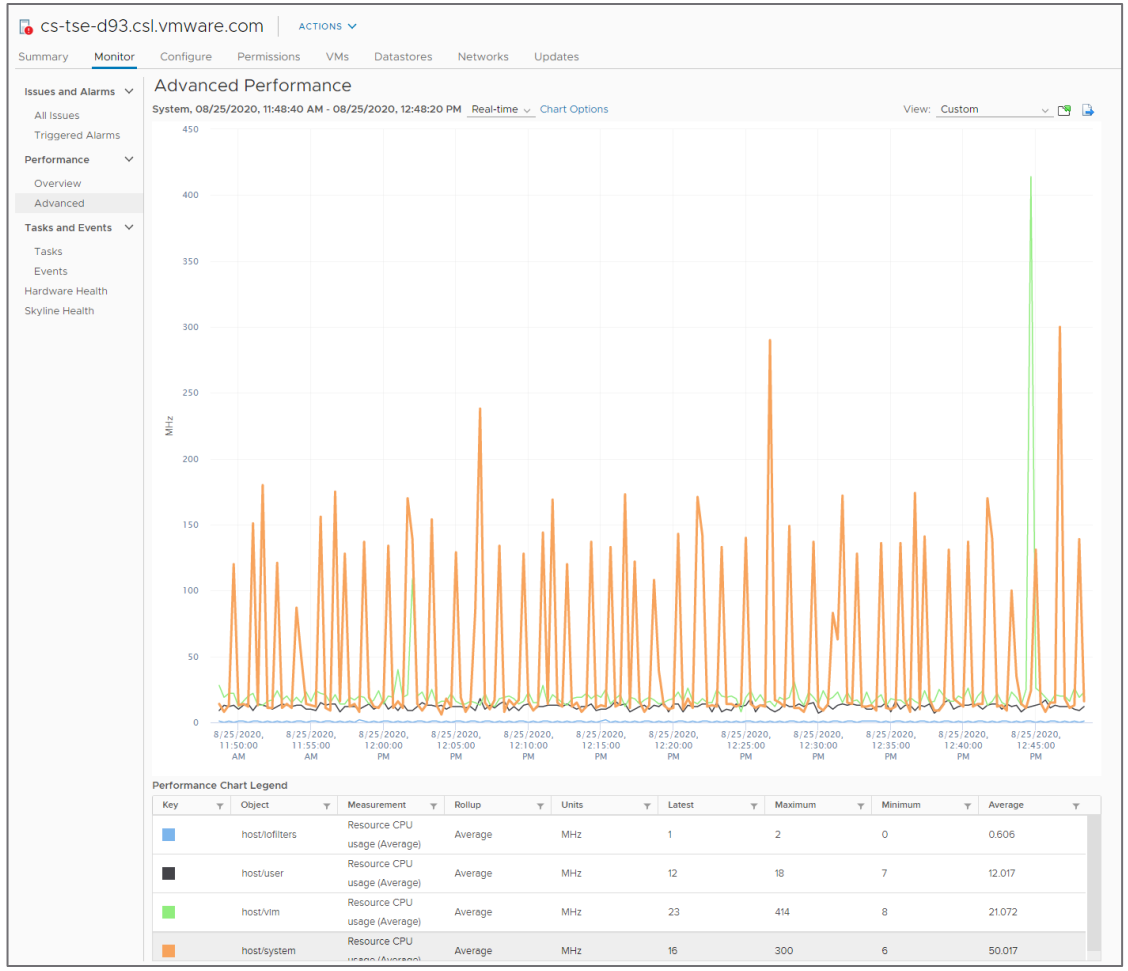
NF

-z in 7.0.2

System CPU Usage

2nd level resource pools





Test Setup

Scripts and Helpers

esxi_one-host-sized-vm_settings-echo.sh 769 Bytes

```
schedNcpus=$(sched-stats -t ncpus)
numPcpus=$(echo "${schedNcpus}" | sed -n 's/\([0-9]\+\) PCPUs$/\1/p')
numCores=$(echo "${schedNcpus}" | sed -n 's/\([0-9]\+\) cores$/\1/p')
numNumaNodes=$(echo "${schedNcpus}" | sed -n 's/\([0-9]\+\) NUMA nodes$/\1/p')
# don't care for package size / LLC etc.
schedDomainSize=$((${numPcpus} / ${numNumaNodes}))

echo "cpuid.coresPerSocket = ${schedDomainSize}"
echo "numa.vcpu.maxPerVirtualNode = ${schedDomainSize}"
# theoretically only necessary with > 2 NUMA nodes but lets be safe
# [ "$((${numPcpus} / ${numCores}))" -gt "1" -a "${numNumaNodes}" -gt "2" ]
if [ "$((${numPcpus} / ${numCores}))" -gt "1" ]; then
        echo "numa.vcpu.preferHT =  true"
fi
for i in $(seq 0 1 ${numPcpus})
        do echo -e "sched.vcpu${i}.affinity = ${i}"
done
```

```
# ./esxi_one-host-sized-vm_settings-echo.sh

cpuid.coresPerSocket = 12
numa.vcpu.maxPerVirtualNode = 12
numa.vcpu.preferHT =  true
sched.vcpu0.affinity = 0
sched.vcpu1.affinity = 1
sched.vcpu2.affinity = 2
sched.vcpu3.affinity = 3
sched.vcpu4.affinity = 4
sched.vcpu5.affinity = 5
sched.vcpu6.affinity = 6
sched.vcpu7.affinity = 7
(…)
sched.vcpu16.affinity = 16
sched.vcpu17.affinity = 17
sched.vcpu18.affinity = 18
sched.vcpu19.affinity = 19
sched.vcpu20.affinity = 20
sched.vcpu21.affinity = 21
sched.vcpu22.affinity = 22
sched.vcpu23.affinity = 23
sched.vcpu24.affinity = 24
```

Test Setup

Scripts and Helpers

esxi_pre-bench_cosched-control.sh 496 Bytes

```
if [ -z "$1" ]; then
    echo "call with either \"disable\", \"max\" or \"reset\" (to default)"; exit 0
fi

case $1 in

        "disable")
                start=0
                stop=0
                echo "disabled cosched"
                ;;

        "max")
                start=99000
                stop=100000
                echo "maxed cosched"
                ;;

        "reset")
                start=2000
                stop=3000
                echo "reset cosched to system defaults"
                ;;

esac

vsish -e set /config/Cpu/intOpts/CoschedCostartThreshold ${start} &>/dev/null
vsish -e set /config/Cpu/intOpts/CoschedCostopThreshold ${stop} &>/dev/null
```

```
# ./esxi_pre-bench_cosched-control.sh disable
disabled cosched
```

Test Setup

Scripts and Helpers

esxi_pre-bench_power-stabilizer.sh 992 Bytes Edit Web IDE

```
if [ -z "$1" ]; then
    echo "call with either \"set\" or \"reset\" (to default)"; exit 0
fi

hwSupport=$(vsish -e get /power/hardwareSupport | sed -n 's/   CPU power management:\(ACPI P-states, ACPI C-states\)$/\1/p')

if [ -z "$hwSupport" ]; then
    echo "Check that BIOS is properly configured"; exit 0
fi

cpuSupport=$(vsish -e get /hardware/cpu/cpuList/0 | sed -n 's/   Name:\(GenuineIntel\)/\1/p')

if [ -z "$cpuSupport" ]; then
    echo "Only known to work on Intel"; exit 0
fi

case $1 in

        "set")
                policy=4
                max=99
                min=99
                dc=0
                echo "disabled Turbo Boost and deep C-States, vote for P1/NF"
                ;;

        "reset")
                policy=2
                max=100
                min=0
                dc=1
                echo "reset custom power options to system default and policy to balanced"
                ;;

esac

vsish -e set /power/currentPolicy ${policy}
vsish -e set /config/Power/intOpts/MaxFreqPct ${max} &>/dev/null
vsish -e set /config/Power/intOpts/MinFreqPct ${min} &>/dev/null
vsish -e set /config/Power/intOpts/UseCStates ${dc} &>/dev/null
```

```
# ./esxi_pre-bench_power-stabilizer.sh set
disabled Turbo Boost and deep C-States, vote for P1/NF
```

Test Setup

Scripts and Helpers

linux_stress-ng_usage-util_loadramp.sh 1007 Bytes

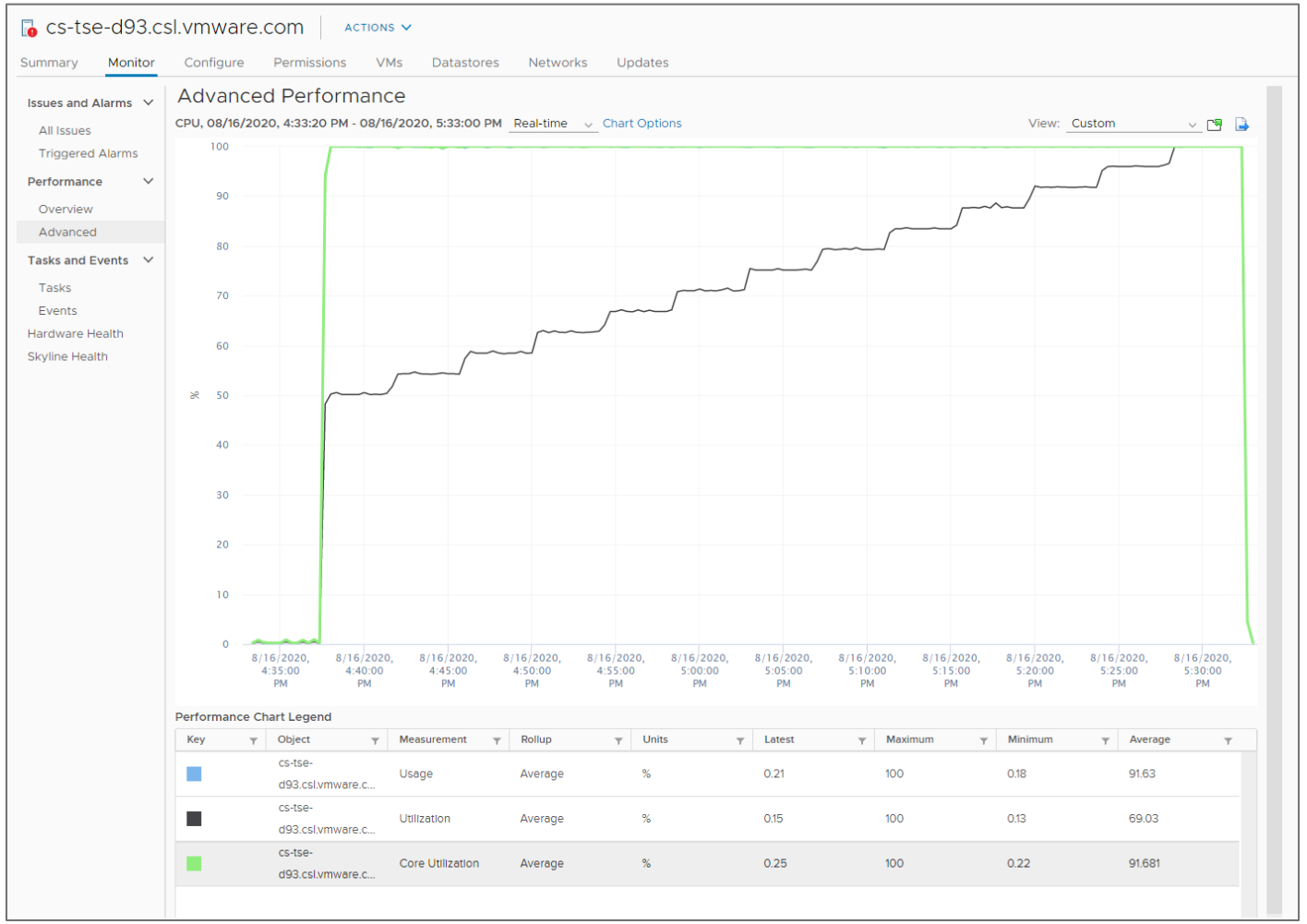
Edit | Web IDE | Lock | Replace | D

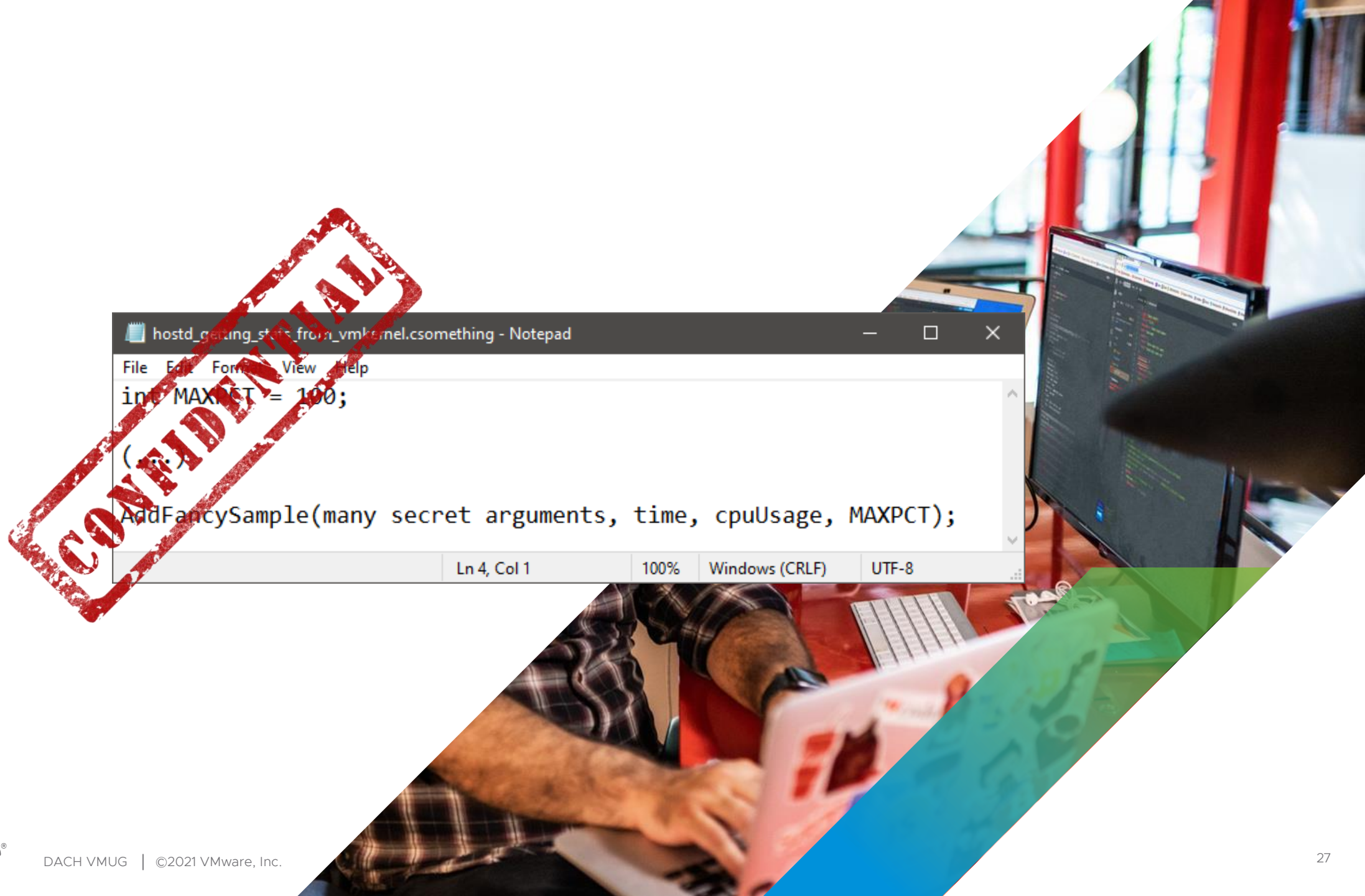
```
if [ -z "$1" ]; then
    echo "call with runtime in minutes (default: 5)"
        # I usually select a duration that fits into realtime charts, so max. 60
        testDurationMinutes=5
else
        testDurationMinutes=$1
fi

# maybe include --cpu-load and change --cpu-method from loop to something else
cpuLoadMethod="loop"

cpus=$(nproc --all)
testDurationSeconds=$(($testDurationMinutes*60))
testIntervalSeconds=$(($testDurationSeconds / (($cpus / 2) + 1)))
initialTaskset=$(seq 0 2 $(($cpus - 1)) | paste -s -d ,)

# run stress-ng in background and quiet
# baseline -> half of the vCPUs, all of the cores
stress-ng --taskset $initialTaskset --cpu $(($cpus / 2)) --cpu-method $cpuLoadMethod -t $testDurationSeconds -q &

loop=0
for iterateTaskset in $(seq 1 2 $(($cpus - 1)))
do
    loop=$(($loop + 1))
    sleep $testIntervalSeconds
    # add another worker to each hypertwin
    stress-ng --taskset $iterateTaskset --cpu 1 --cpu-method $cpuLoadMethod -t $(($testDurationSeconds - ($loop * $testIntervalSeconds))) -q &
done
```

```
# ./ linux_stress-ng_usage-util_loadramp.sh
stress-ng: info:  [2293] dispatching hogs: 12 cpu
stress-ng: info:  [2311] dispatching hogs: 1 cpu
stress-ng: info:  [2317] dispatching hogs: 1 cpu
stress-ng: info:  [2323] dispatching hogs: 1 cpu
stress-ng: info:  [2329] dispatching hogs: 1 cpu
stress-ng: info:  [2335] dispatching hogs: 1 cpu
stress-ng: info:  [2341] dispatching hogs: 1 cpu
stress-ng: info:  [2347] dispatching hogs: 1 cpu
stress-ng: info:  [2366] dispatching hogs: 1 cpu
stress-ng: info:  [2372] dispatching hogs: 1 cpu
stress-ng: info:  [2387] dispatching hogs: 1 cpu
stress-ng: info:  [2399] dispatching hogs: 1 cpu
stress-ng: info:  [2448] dispatching hogs: 1 cpu
```

Host view

VM starts with 12 vCPUs / all cores utilized



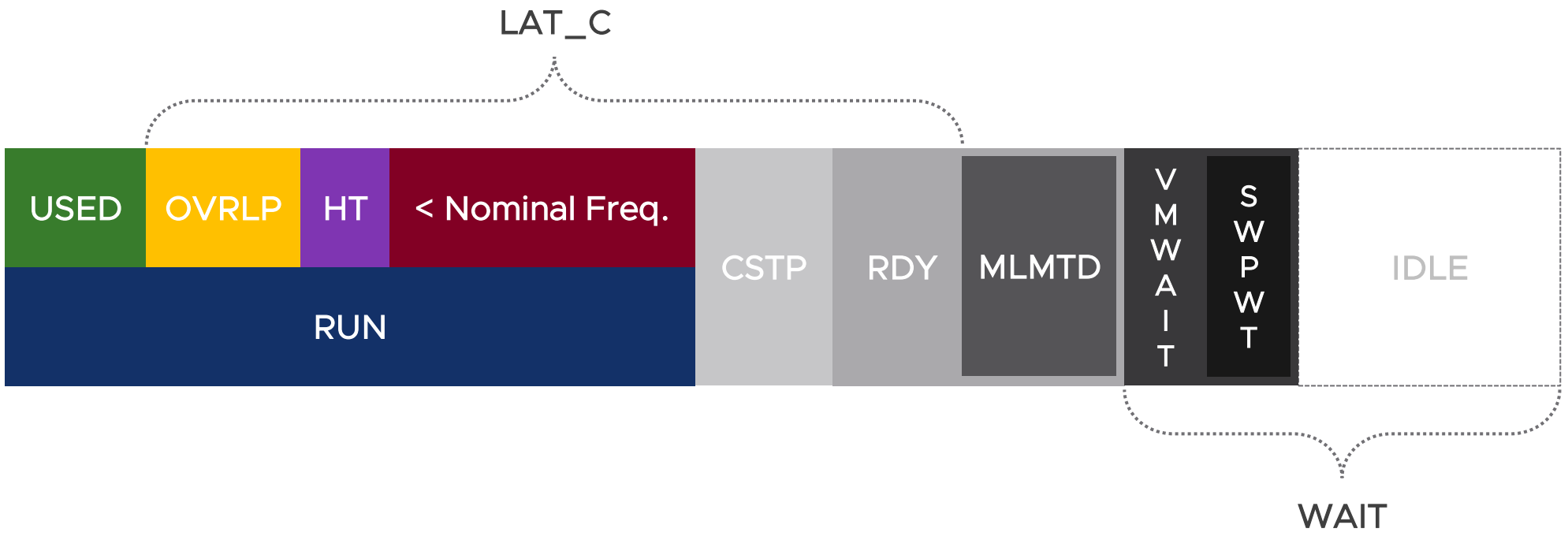

CONFIDENTIAL
hostd_getting_stats_from_vmkernel.csomething - Notepad
File Edit Format View Help
int MAXPCT = 100;
(...)
AddFancySample(many secret arguments, time, cpuUsage, MAXPCT);
Ln 4, Col 1
100%
Windows (CRLF)
UTF-8

vCPU accounting

States and Metrics – esxtop nomenclature

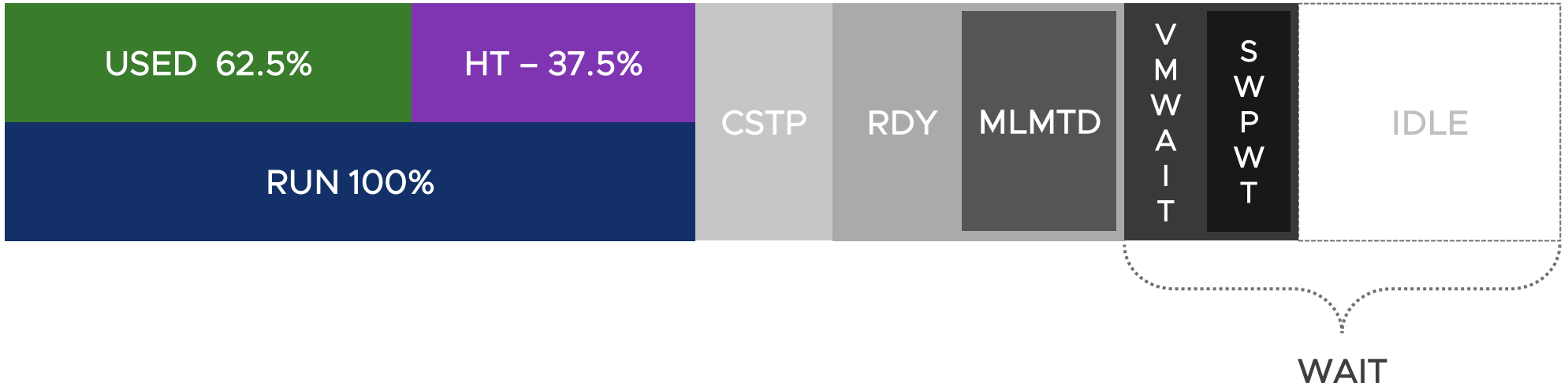
vCPU accounting

States and Metrics – esxtop nomenclature



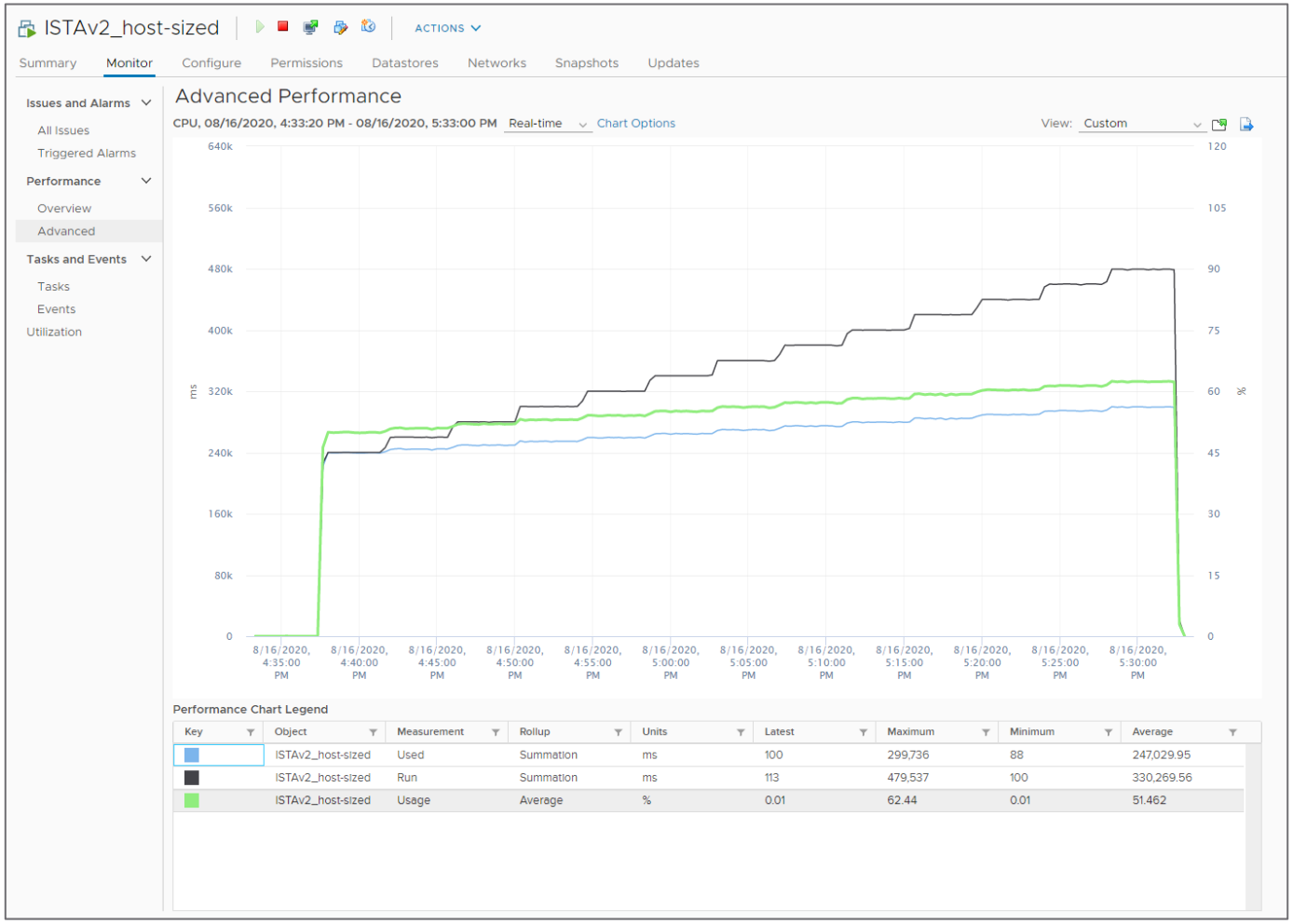
vCPU accounting

States and Metrics – esxtop nomenclature



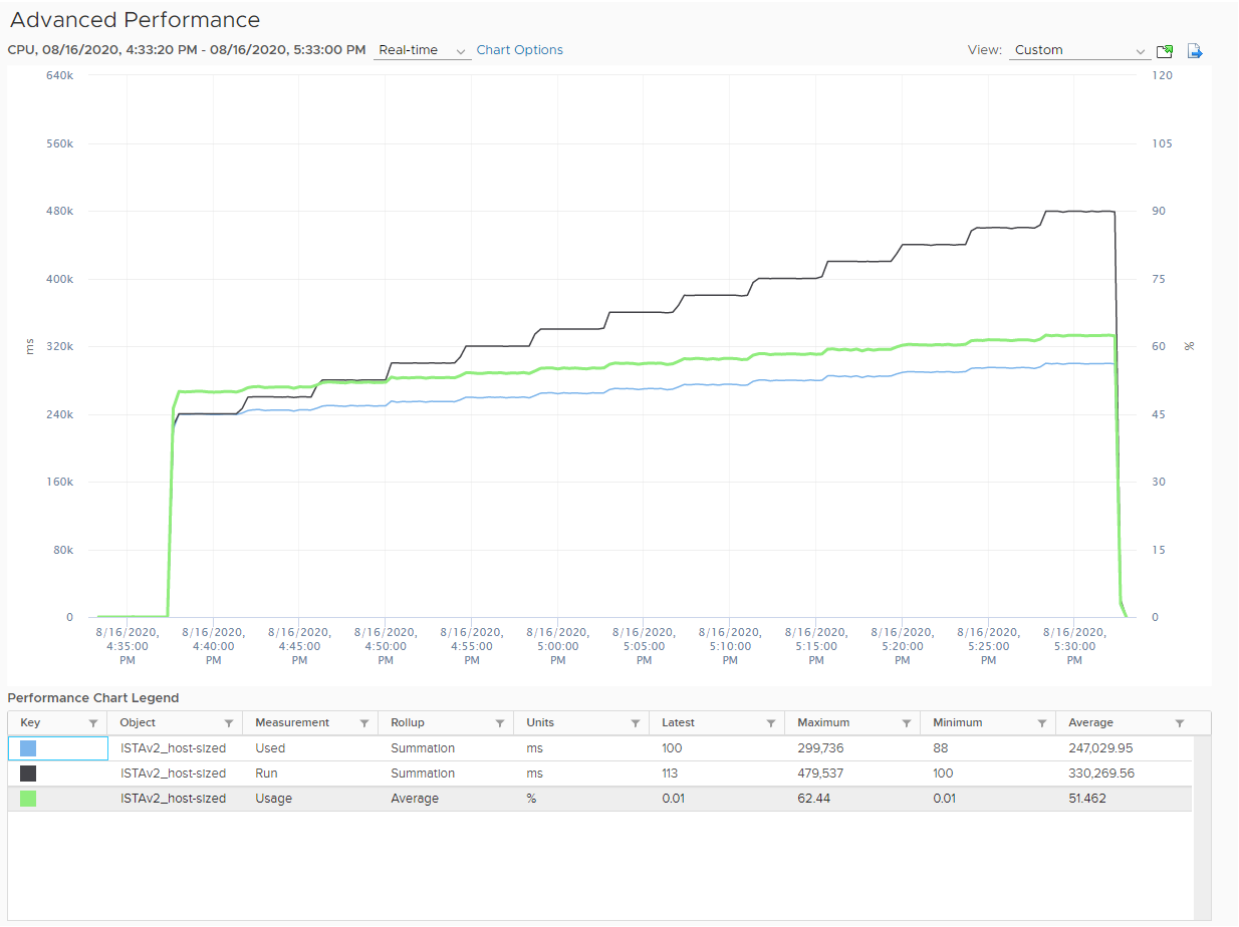
VM view

Usage only increases at a “25-degree angle”



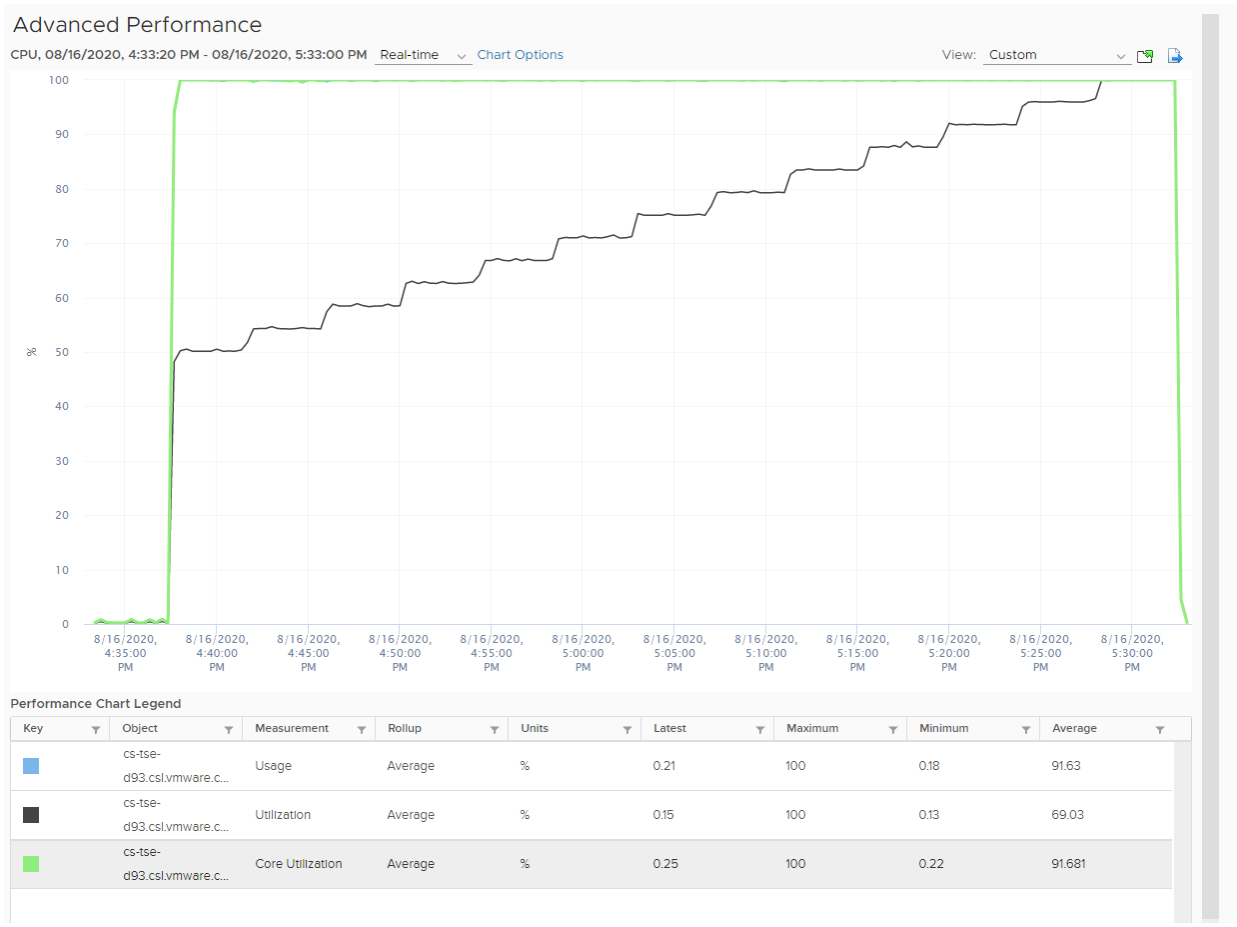
VM and Host view

Utilization / Run is the same



Performance Chart Legend

Key	Object	Measurement	Rollup	Units	Latest	Maximum	Minimum	Average
	ISTAv2_host-sized	Used	Summation	ms	100	299,736	88	247,029.95
	ISTAv2_host-sized	Run	Summation	ms	113	479,537	100	330,269.56
	ISTAv2_host-sized	Usage	Average	%	0.01	62.44	0.01	51.462

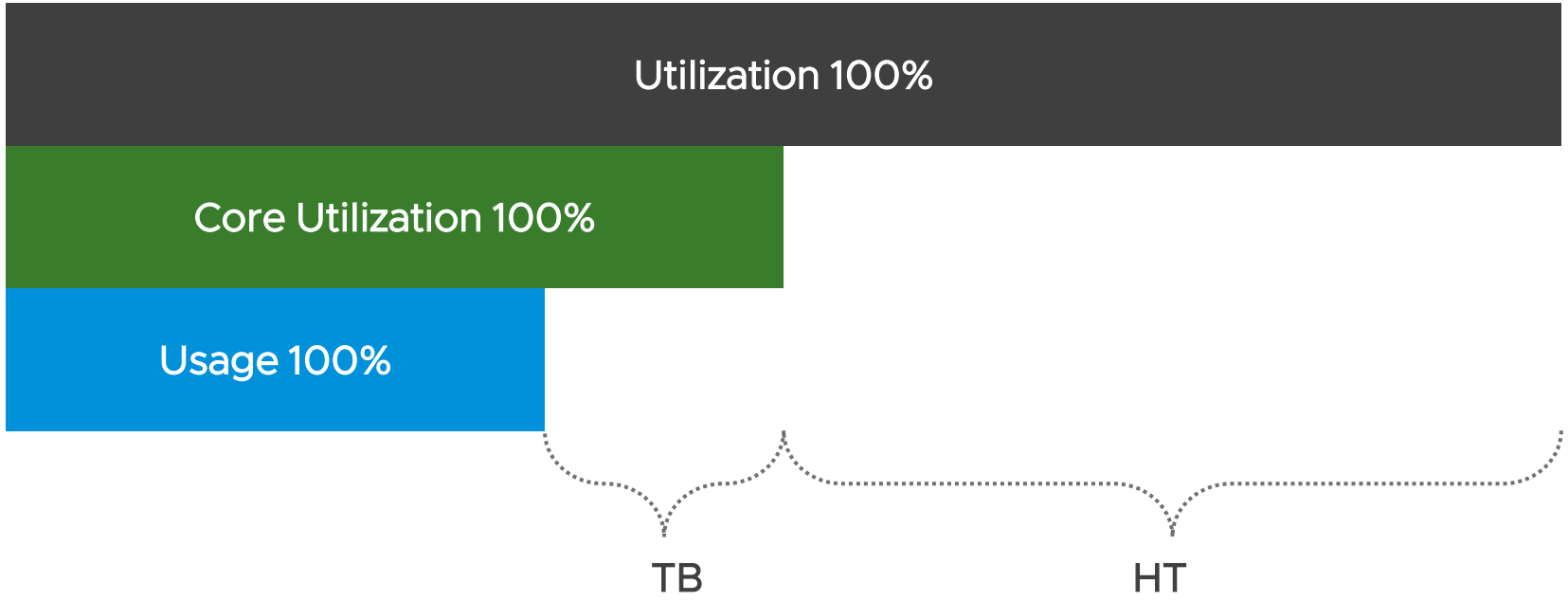


Performance Chart Legend

Key	Object	Measurement	Rollup	Units	Latest	Maximum	Minimum	Average
	cs-tse-d93.csl.vmware.c...	Usage	Average	%	0.21	100	0.18	91.63
	cs-tse-d93.csl.vmware.c...	Utilization	Average	%	0.15	100	0.13	69.03
	cs-tse-d93.csl.vmware.c...	Core Utilization	Average	%	0.25	100	0.22	91.681

Host CPU Usage and Utilization

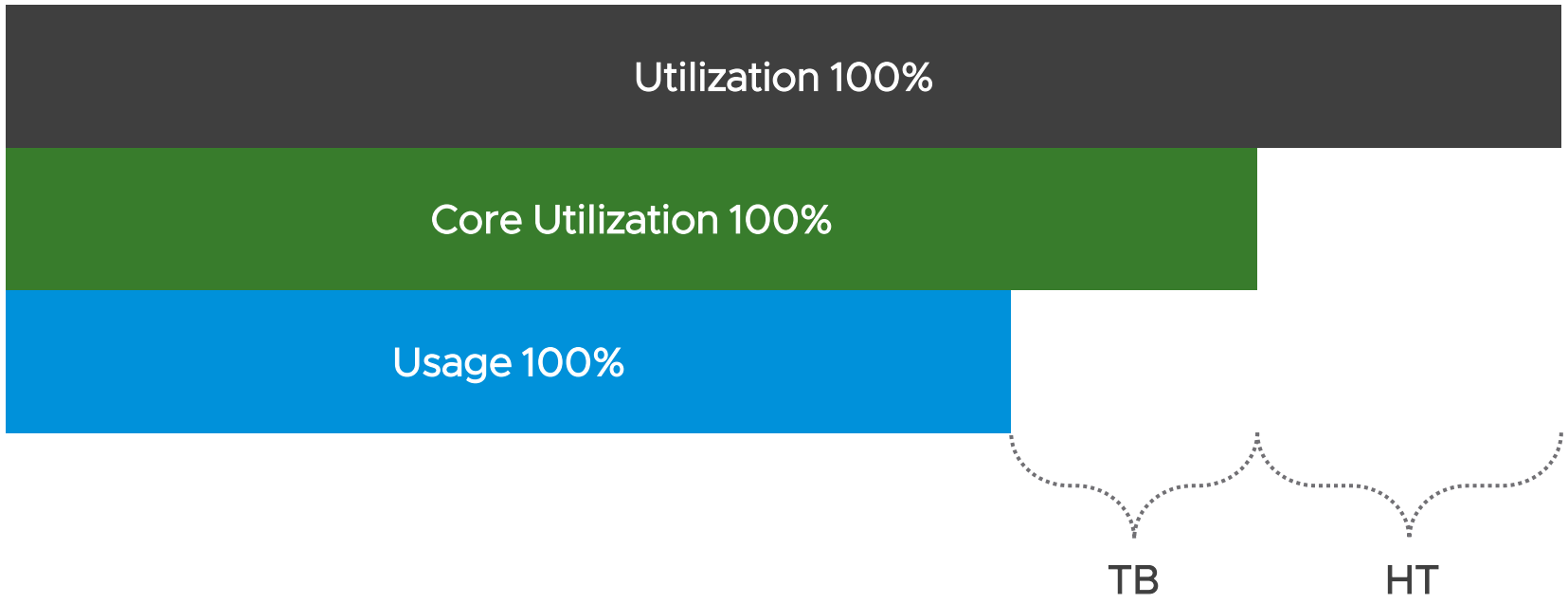
Looking at metrics



 |

Host CPU Usage and Utilization

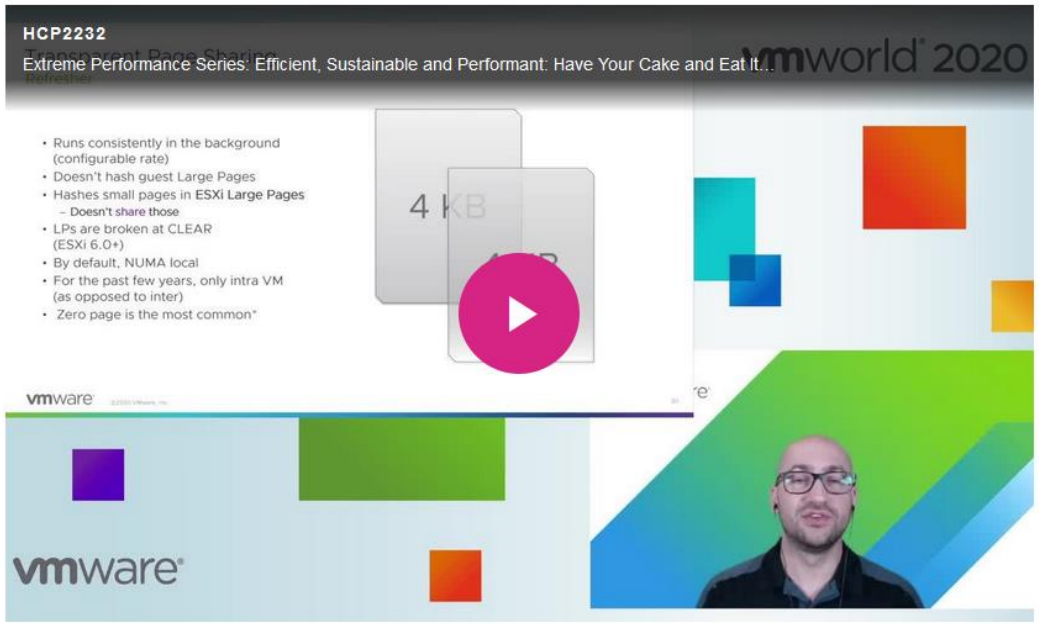
Looking at throughput



 |

Host Large Pages and Transparent Page Sharing

Need to know more?



Presentation

Download PDF

Speakers

Valentin Bondzio
Senior Staff Technical Support Engineer
VMware
Show Bio

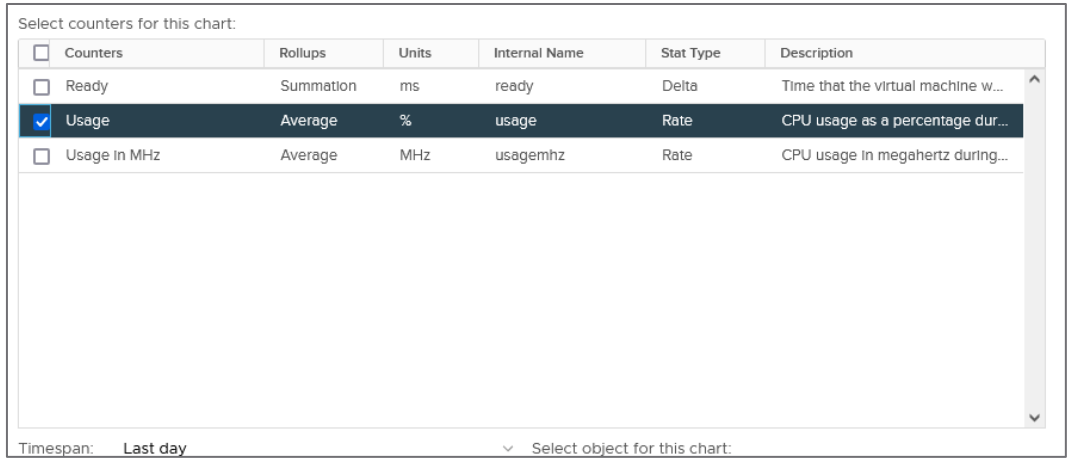
HCP2206 - Extreme Performance Series: Efficient, Sustainable and Performant: Have Your Cake and Eat It, Too
https://www.vmworld.com/en/video-library/video-landing.html?sessionid=1589493658174001Skyo

Nothing's ever easy …

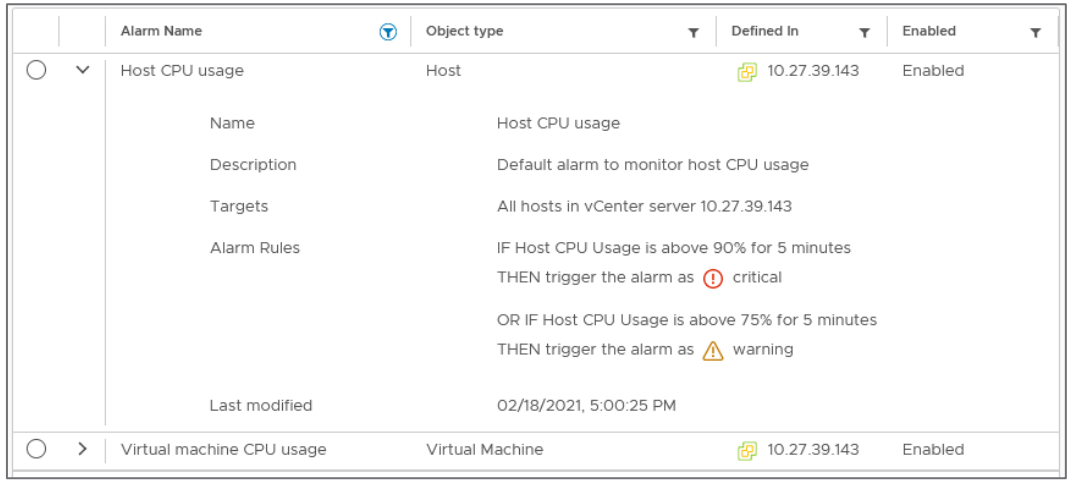
Current Problems for CPU Usage / Utilization Capacity

Host Core Utilization & Host Utilization metrics

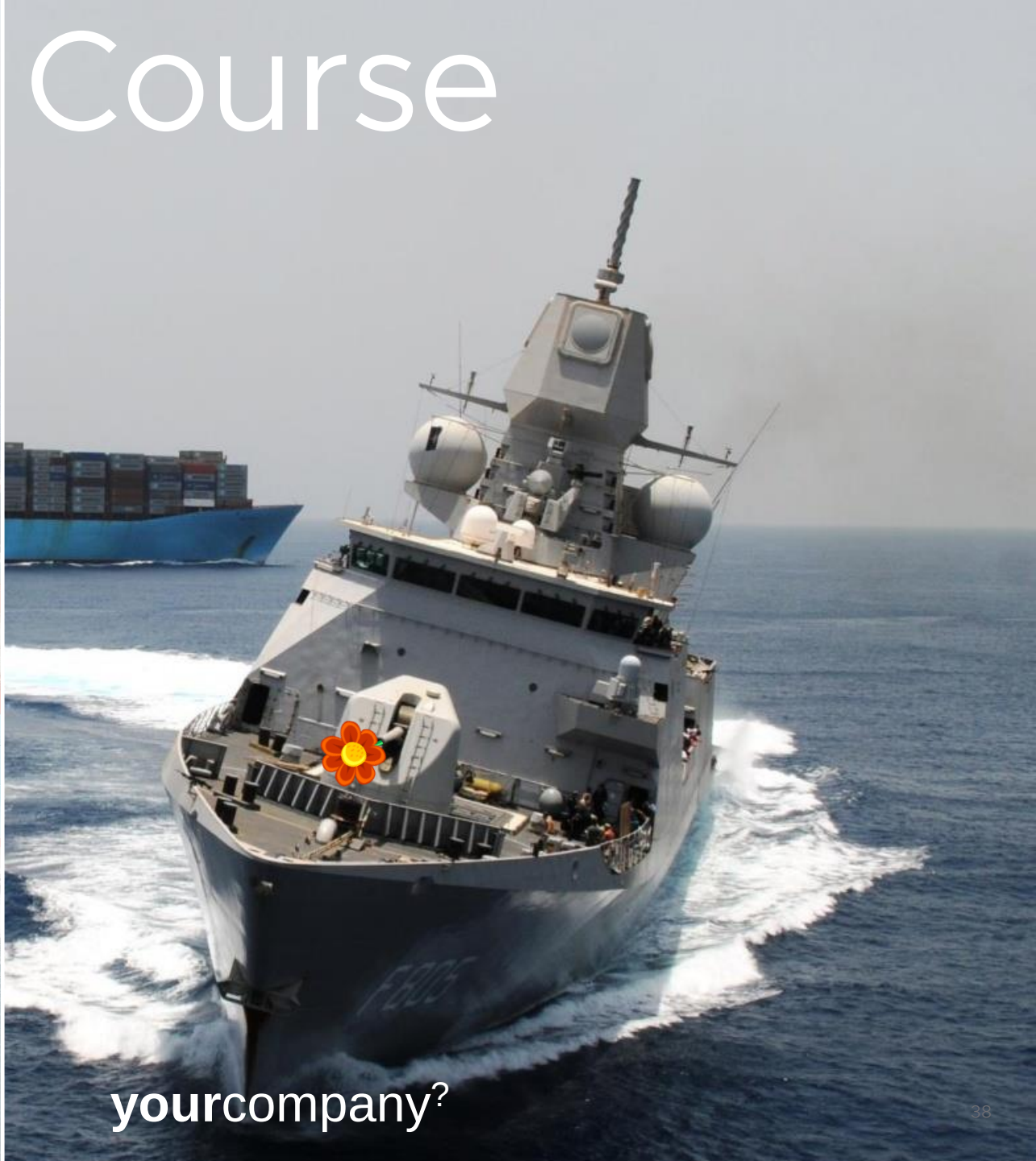
- only in Real-Time and not rolled over
  - needs to be configured manually
- not enabled by default in vROps either



vCenter Host and VM CPU Usage Alarms …



Adjusting Course



yourcompany?

Thank You

Please type your questions in the chat!