Avaya Solution & Interoperability Test Lab

Application Notes on Best Practices for Reservationless deployment of Avaya Aura® software release 10.x on VMware – Issue 1.1

Abstract

This Application Notes presents best practices for reservationless deployment of Avaya Aura® software release 10.x on VMware based infrastructure.

This Application Notes includes the Best Practices and recommendation for reservationless deployment, configurations of Alarms and Events, monitoring of CPU and Memory utilization and VMware reports/logs to be collected/analyzed in case of any issue in reservationless deployment of Avaya applications.

Table of Contents

1. Scope of the document

This Application Notes presents best practices for reservationless deployment of Avaya Aura® software release 10.x on VMware based infrastructure.

This Application Note includes the Best Practices and recommendation for reservationless deployment, configurations of Alarms and Events, monitoring of CPU and Memory utilization and VMware reports/logs to be collected/analyzed in case of any issue in reservationless deployment of Avaya applications.

Note that the “reservationless” is specific to CPU & Memory resources only. For other deployment aspects such as network requirement, storage requirement etc., guidelines provided in respective product documents must be followed.

This document shall be used in conjunction with VMWare best practices and Avaya product documents for deploying Avaya applications.

Post Avaya Aura® release 8.0, rules around deploying Avaya Aura® without reservations have been relaxed and are considered as supported if the best practices described in this document are followed.

Disclaimer
It is assumed that in reservationless deployment, extra administrative burden is going to be customer’s responsibility and that Avaya will support the solution deployed in this environment if any issue arising is not related to Virtual environment resource attrition or starvation. If such situation were to occur, the customer would be required to ensure availability of required resources prior to Avaya performing any further troubleshooting.

The images used in this document may not be identical to the latest VMware vSphere version. Please refer VMware documentation in case of discrepancies.

2. Terminologies and Acronyms

Acronym / Term	Definition
ESXi	VMware bare-metal hypervisor that installs directly onto a physical server
Host	Physical server with ESXi hypervisor installed
VM	Virtual Machine running on ESXi Host
HA	High Availability
TPS	Transparent Page Sharing
SMT	Simultaneous Multithreading
BIOS	Basic input/output system
vCPU	Virtual CPU
vCenter	Centralized Server platform from VMware
AAMS	Avaya Aura® Media Server

3. List of Supported Applications

Following Avaya applications are in scope for reservationless deployment on VMware:

- Avaya Aura® System Manager release 10.x
- Avaya Aura® Session Manager release 10.x
- Avaya Aura® Communication Manager release 10.x
- Avaya Aura® Application Enablement Services release 10.x
- Avaya Aura® Media Server releases 10.x
- Avaya Aura® Presence Services release 10.x
- Avaya Session Border Controller for Enterprise release 10.x

4. Deployment Recommendations

Oversubscribing real time application on your server infrastructure has never been and is not supported.

Recommendations mentioned in the subsequent sections needs to be followed for reservationless deployment.

4.1. Resource allocation for reservationless deployment

Even if the reservation is off, resource allocation shall be equal to the application requirement.

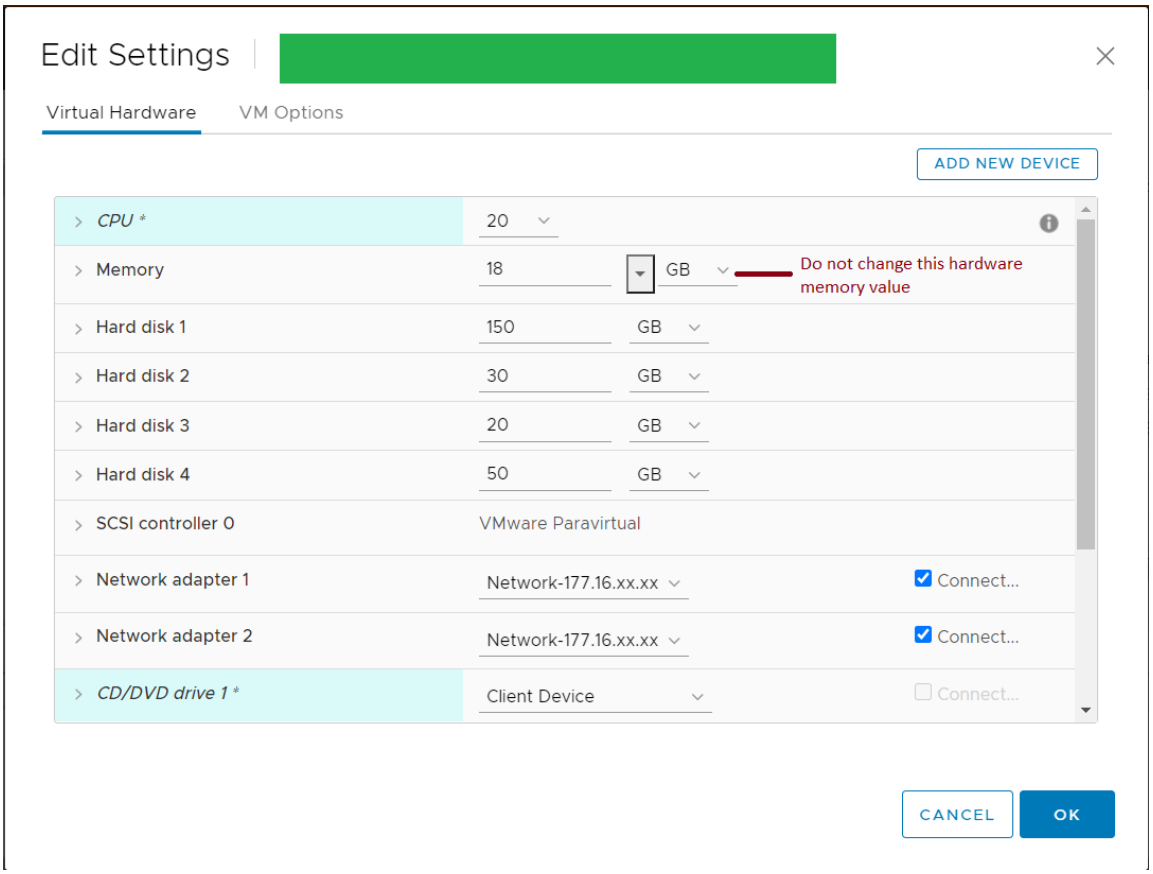
Refer the deployment guide of individual Avaya applications to get information on resource requirement.

Once the required resources have been allocated without reservation, follow the guidelines as mentioned below:

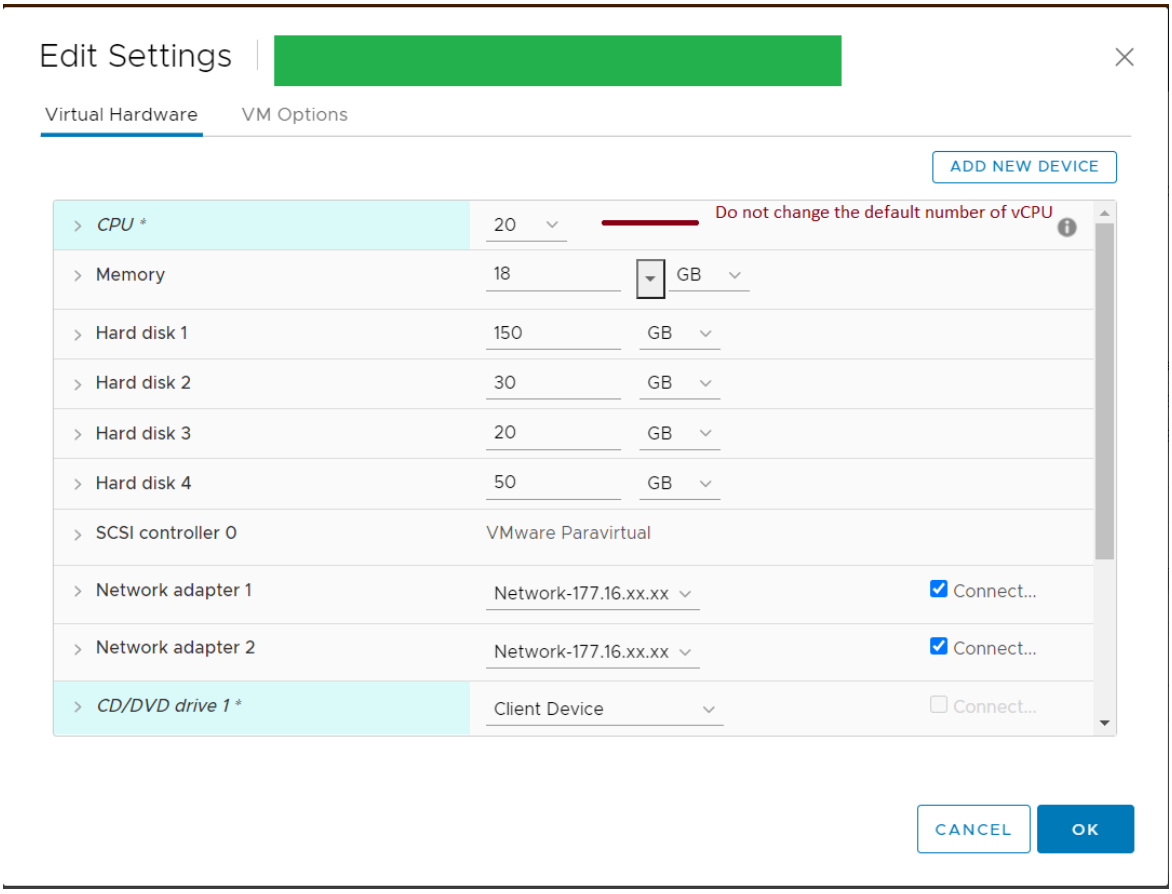
4.1.1. Default assigned memory and CPU under Virtual Hardware Tab

Once Avaya application is deployed on VMware with appropriate supported footprint, do not change the default Memory or CPU under Virtual Hardware Tab. The same is indicated in following screenshots:

Default assigned Memory status:



Default assigned CPU status:



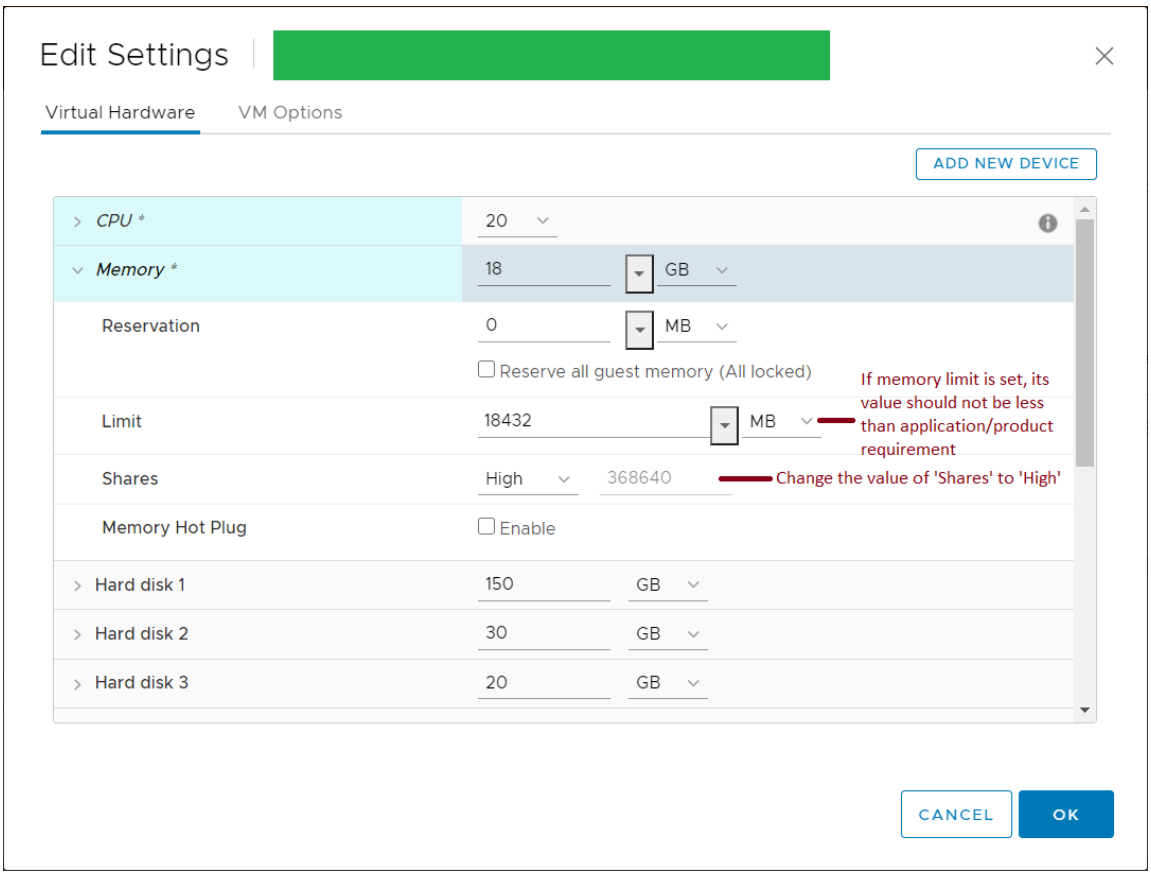
4.1.2. CPU or Memory under Virtual Hardware configuration

Once Avaya application is deployed on VMware with appropriate supported footprint, change the Resource Allocation Shares from “Normal” to “High”.

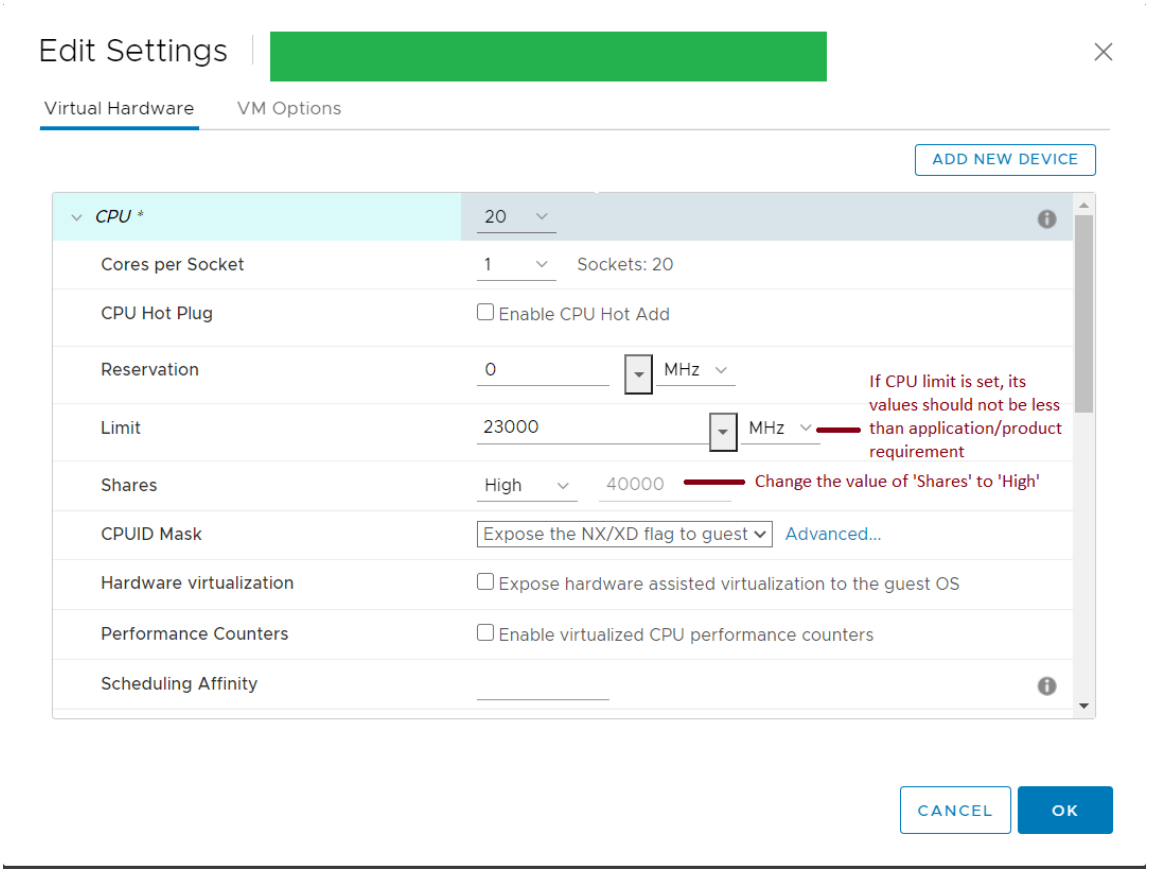
Do not configure the limit for the CPU or Memory values which are less than the product/application requirement.

The same is indicated in following screenshots:

Memory Limit Screen:



CPU Limit Screen:



4.1.3. Resource Over-commitment and Resource Starvation

Reservationless deployment has no impacts to the Avaya application when host resources are not over-committed as resource starvation will not occur (Except Avaya Aura® Media Server, see guidelines mentioned in below paragraph for Avaya Aura® Media Server). But once reservations are removed, it is possible to over-commit the resources of the Host by adding more Virtual Machines and this may result in resource starvation which is not supported by Avaya.

Avaya Aura® Media Server is handling the real-time media traffic and it is very sensitive with respect to any processing delay. Hence if Avaya Aura® Media Server is used in reservationless deployment, following recommendation shall be followed:
"Number of vCPUs assigned across all VMs must not exceed total physical CPUs available in an ESXi host which contains Avaya Aura® Media Server. Physical CPU refers to physical CPU core and not hyper-threaded core"

Avaya does not support any issues of Avaya Aura® Media Server, if number of vCPUs assigned across all virtual machines exceeds total physical CPUs available in an ESXi host.

If CPU and/or Memory resource is over-committed, ensure that the actual resource usage of the Host is not crossing the recommended thresholds, which are 75% for CPU and 70% for Memory. It is highly recommended to take corrective actions if resource usage crosses above threshold for CPU or Memory to avoid resource starvation.

4.1.4. Server size for Reservationless Deployment

It is recommended to use larger servers for reservationless deployment instead of smaller servers.

4.1.5. With and without reservation deployment

Do not mix non-Avaya applications having reservation with reservationless Avaya applications.

Ensure that third party reserved virtual machines are not deployed in the same host where Avaya reservationless applications are deployed. If third party virtual machines are deployed in the same host where Avaya reservationless applications are deployed, then they must not be reserved. This is required for correct monitoring of resource availability for reservationless applications.

4.1.6. Caution

Except not reserving CPU & Memory, all the other guideline provided in respective product documents such as CPU & memory allocation, network requirement, storage requirement etc. must be followed.

Resource starvation is not allowed in reservationless deployment. Impacts of resource starvation can be varied and severe such as poor voice quality, talk path issues, all the users out of service, application reboots etc.

Following are the examples of some of the resource sensitive activities. During such activities, spike of resource starvation can lead to major impacts on Avaya application and/or entire Avaya solution if full allocated resources are not available.

- Power-on and reboot of any machine
- Fresh installation of any Virtual machine
- Upgrade, Patching, backup and restore activity of any application
- Planned failover & recovery between the data centers

Incase VMware HA is configured, any host failure event results in more over-subscription on the remaining hosts due to VM reallocation in those hosts. If resource reservation is removed, host failure event may result in resource starvation.

4.2. Memory Reclamation Features

Memory reclamation features minimize the impact of memory starvation on the host. However, **Avaya application may have unpredictable behavior in case of resource starvation.**

ESXi uses following techniques to reclaim virtual machine memory when host is starving for Memory resources:

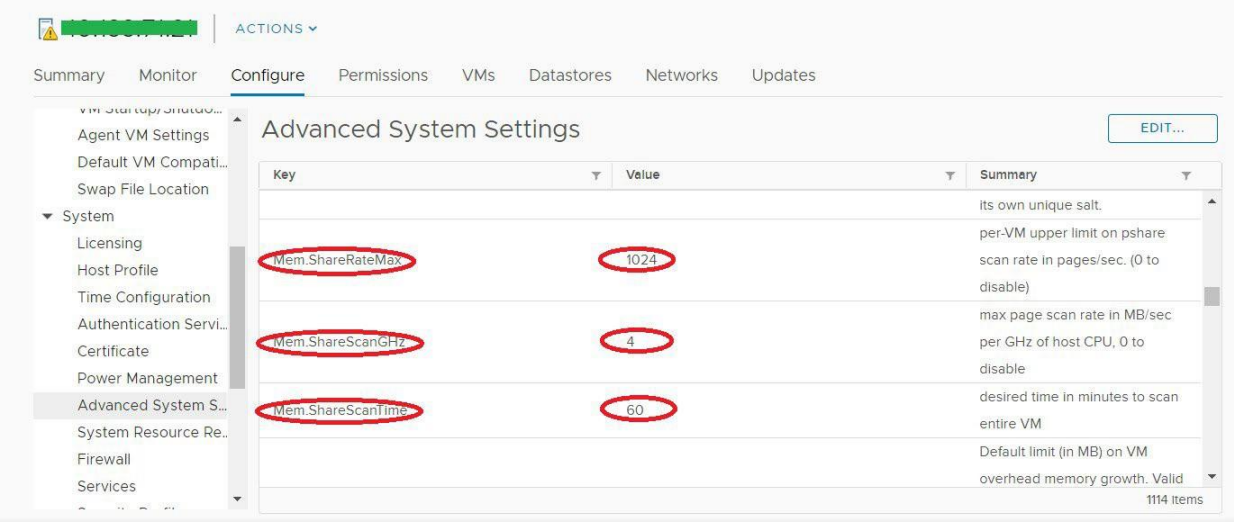
1. Transparent page sharing (TPS): - Reclaims memory by removing redundant pages with identical content
2. Ballooning: - Reclaims memory by artificially increasing the memory pressure inside the guest
3. Memory compression: - Reclaims memory by compressing the pages that need to be swapped out
4. Hypervisor swapping: - Reclaims memory by having ESXi directly swap out the virtual machine's memory

If ESXi is starting to use Ballooning, Memory compression or Hypervisor swapping, it indicates that there is a memory starvation and Avaya applications may be impacted. This should be avoided at all cost as Avaya applications do not support any Memory starvation. Details on how to monitor the memory starvation is described under "Monitoring" section of this document.

4.2.1. Transparent Page Sharing (TPS)

When multiple virtual machines are running, some of them may have identical sets of memory content. This presents opportunities for sharing memory across virtual machines (as well as sharing within a single virtual machine). For example, several virtual machines may be running the same guest operating system, have the same applications, or contain the same user data. With page sharing, the hypervisor can reclaim the redundant copies and keep only one copy, which is shared by multiple virtual machines in the host physical memory. As a result, the total virtual machine host memory consumption is reduced and memory over commitment is possible.

In VMware ESXi, the hypervisor scans the guest physical pages randomly with a base scan rate specified by Mem.ShareScanTime, which specifies the desired time to scan the virtual machine's entire guest memory. The maximum number of scanned pages per second in the host and the maximum number of per-virtual machine scanned pages, (that is, Mem.ShareScanGHz and Mem.ShareRateMax respectively) can also be specified in ESXi advanced settings. An example is shown in below Figure.



The default values of these three parameters are chosen to provide sufficient sharing opportunities while keeping the CPU overhead negligible. ESXi adjusts the page scan rate based on the amount of current shared pages. If the virtual machine's page sharing opportunity seems to be low, the page scan rate will be reduced accordingly and vice versa. This optimization further mitigates the overhead of page sharing.

4.2.2. Ballooning

Ballooning is a completely different memory reclamation technique compared to transparent page sharing. Due to the virtual machine's isolation, the guest operating system is not aware that it is running inside a virtual machine and is not aware of the states of other virtual machines on the same host. When the hypervisor runs multiple virtual machines and the total amount of the free host memory becomes low, none of the virtual machines will free guest physical memory because the guest operating system cannot detect the host's memory shortage. Ballooning makes the guest operating system aware of the low memory status of the host.

In ESXi, a balloon driver is loaded into the guest operating system as a pseudo-device driver. It has no external interfaces to the guest operating system and communicates with the hypervisor through a private channel. The balloon driver polls the hypervisor to obtain a target balloon size. If the hypervisor needs to reclaim virtual machine memory, it sets a proper target balloon size for the balloon driver, making it "inflate" by allocating guest physical pages within the virtual machine.

VMware Tools should be running for ballooning.

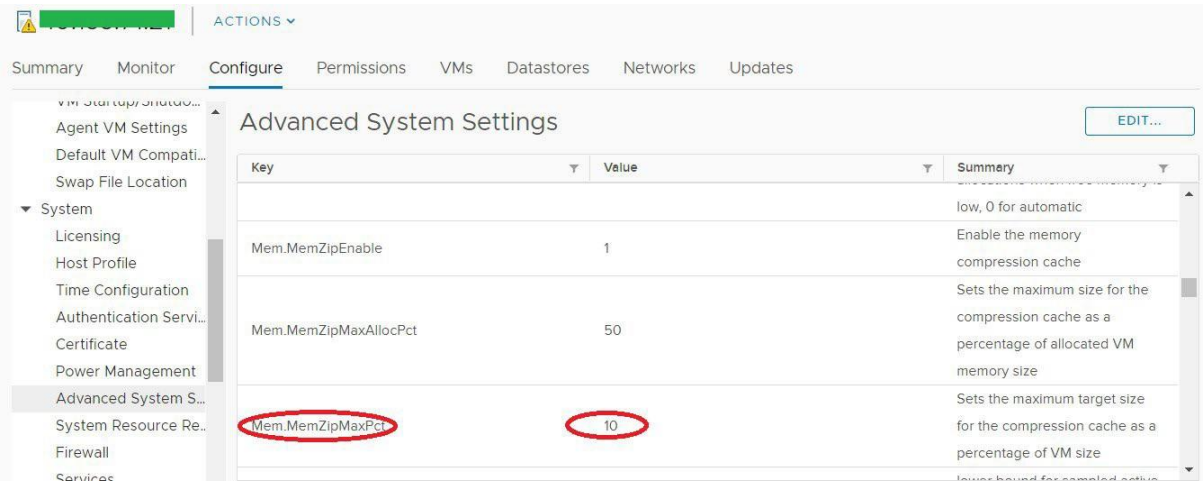
Ballooning might not reclaim memory quickly enough to satisfy host memory demands. Ballooning may impact applications.

4.2.3. Memory Compression

With memory compression, ESXi stores pages, which would otherwise be swapped out to disk through host swapping, in a compression cache located in the main memory. Memory compression outperforms host swapping because the next access to the compressed page only causes a page decompression, which can be an order of magnitude faster than the disk access.

ESXi determines if a page can be compressed by checking the compression ratio for the page. Memory compression occurs when the page's compression ratio is greater than 50%. Otherwise, the page is swapped out. Only pages that would otherwise be swapped out to disk are chosen as candidates for memory compression. This means ESXi will not proactively compress guest pages when host swapping is not necessary. In other words, memory compression does not affect workload performance when host memory is under committed.

The default maximum compression cache size is set to 10% of configured VM memory size. This value can be verified through the vSphere Client in Advanced Settings by changing the value for Mem.MemZipMaxPct, which is shown in below screenshot.



4.2.4. Hypervisor Swapping

In the cases where ballooning, transparent page sharing, and memory compression are not sufficient to reclaim memory, ESXi employs hypervisor swapping to reclaim memory. At virtual machine startup, the hypervisor creates a separate swap file for the virtual machine. Then, if necessary, the hypervisor can directly swap out guest physical memory to the swap file, which frees host physical memory for other virtual machines.

Both page sharing, and ballooning take time to reclaim memory. In contrast, hypervisor swapping is a guaranteed technique to reclaim a specific amount of memory within a specific amount of time. However, hypervisor swapping has a huge impact on the Virtual Machines.

4.3. CPU Hyper threading

Hyper-threading technology (sometimes also called simultaneous multithreading, or SMT) allows single physical processor core to behave like two logical processors, essentially allowing two independent threads to run simultaneously. Unlike having twice as many processor cores, that can roughly double performance, hyper-threading can provide anywhere from a slight to a significant increase in system performance by keeping the processor pipeline busier.

If the hardware and BIOS support hyper-threading, ESXi automatically makes use of it. For the better Performance, it is recommended to enable hyper-threading, which can be accomplished as follows:

a) Ensure that your system supports hyper-threading technology. It is not enough that the processors support hyper-threading. The BIOS must support it as well. Consult your system documentation to see if the BIOS includes the support for hyper-threading.
b) Enable hyper-threading in the system BIOS. Some manufacturers label this option Logical Processor while others label it Enable Hyper-threading.

Note: While deployment planning, total processing capacity of Host should be considered with actual hardware cores and not with the hyper-threaded cores. For example, host with CPU cores of 2.4 GHz (2400 MHz) and 16 real cores, will show 32 vCPUs with Hyper-threading. However, total processing capacity of a host should be 38,400 MHz (16X2400) and not 76,800 MHz (32X2400). Example mentioned here is theoretical for understanding purpose only and actual processing capacity may be different due to hyperthreading and other overheads.

5. Configuration for monitoring of CPU and Memory resources

For reservationless deployment, it is very critical to configure proper monitoring mechanisms for VMware resource utilization. It helps in taking timely corrective action to avoid possible resource starvation on the applications.

It is strongly recommended to configure proper CPU and Memory Alarms and associated actions for each ESXi Host.

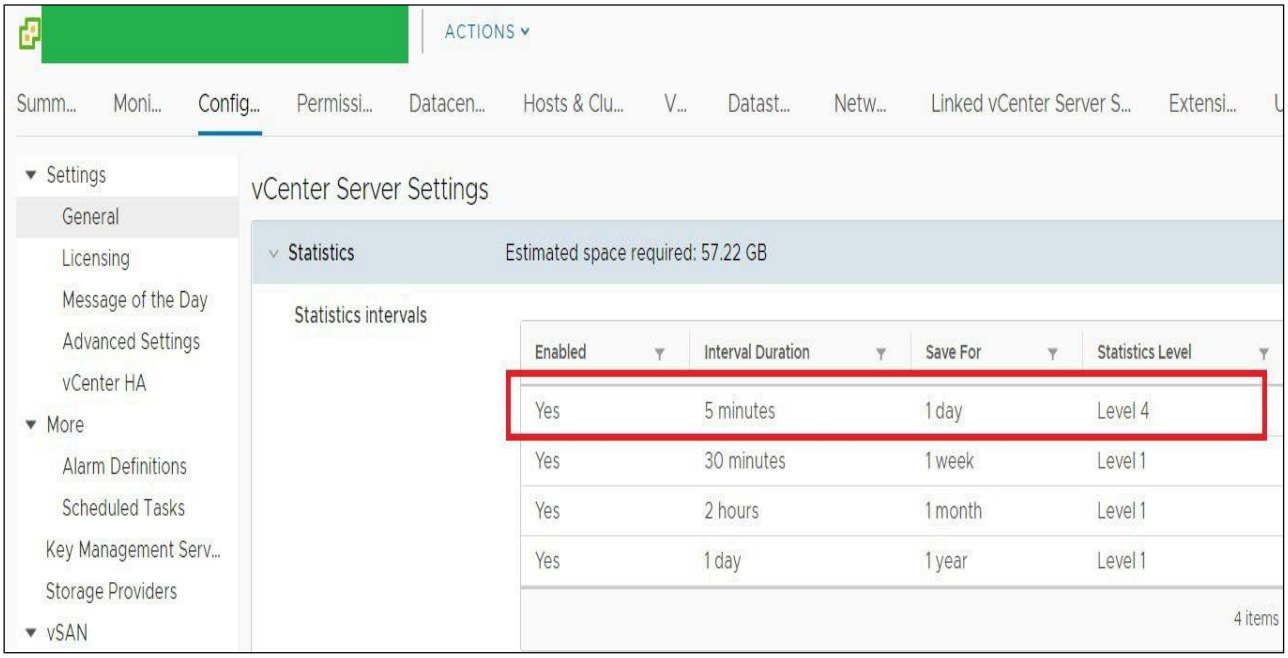
VMware allows to configure alarms and to specify the actions the system should take when they are triggered.

It is possible to monitor inventory objects such as Cluster, Host or Virtual machines by setting alarms on them. Setting an alarm involves selecting the type of inventory object to monitor, defining when and for how long the alarm will trigger, and defining actions that will be performed because of the alarm being triggered.

5.1. Configure VMware Statistics from vCenter

Ensure that statistics are properly configured and enabled in vCenter. This will enable to capture the VMware historical reports up to one day with frequent intervals.

Login to vCenter, Go to “Host and Clusters”, click on vCenter name, now click on “Configuration” tab from right hand side pane. This will open up vCenter Settings, click on option “General”. Expand “Statistics” option in right hand pane and verify that the “interval duration” of 5 Minutes is enabled; “Save For” is at least “1 Day” and “Statistic Level” is set to “4”. Following screenshot indicates the same.



5.2. Configure Events and Alarms for CPU and Memory Thresholds

Ensure that the below mentioned thresholds for alarm triggers are configured for CPU and Memory utilization. It is recommended to configure "Email Notification Action" when any of these Alarms triggers.

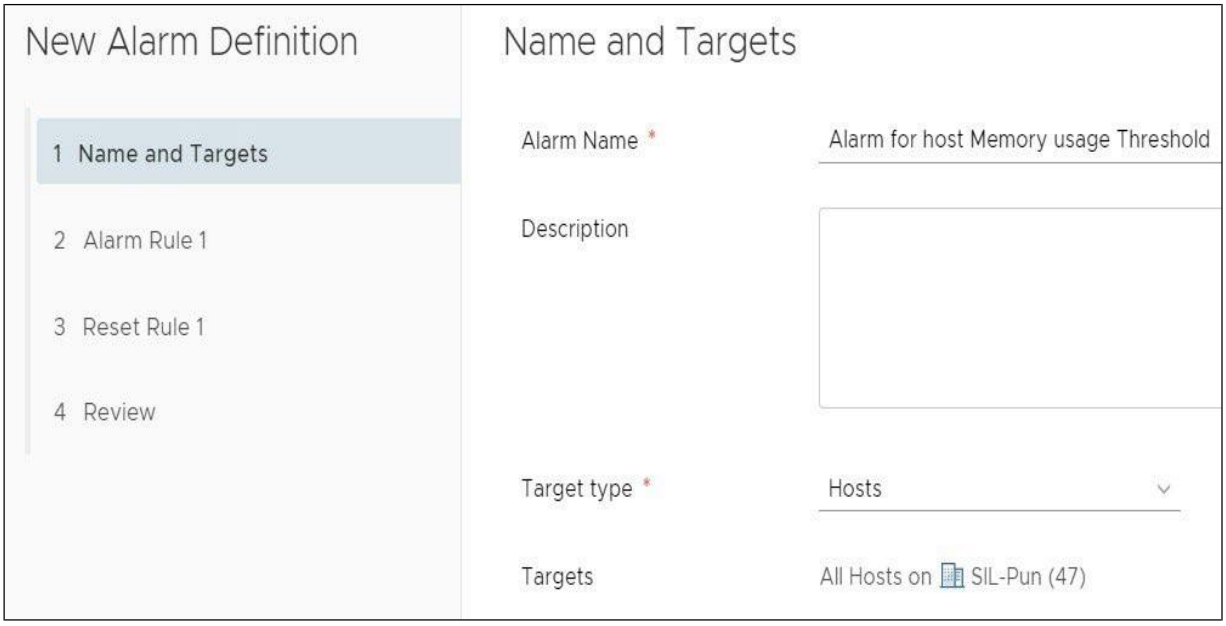
Modify default alarm monitoring configurations for CPU and Memory as per the guidelines below:

5.2.1. Configure Memory Alarms

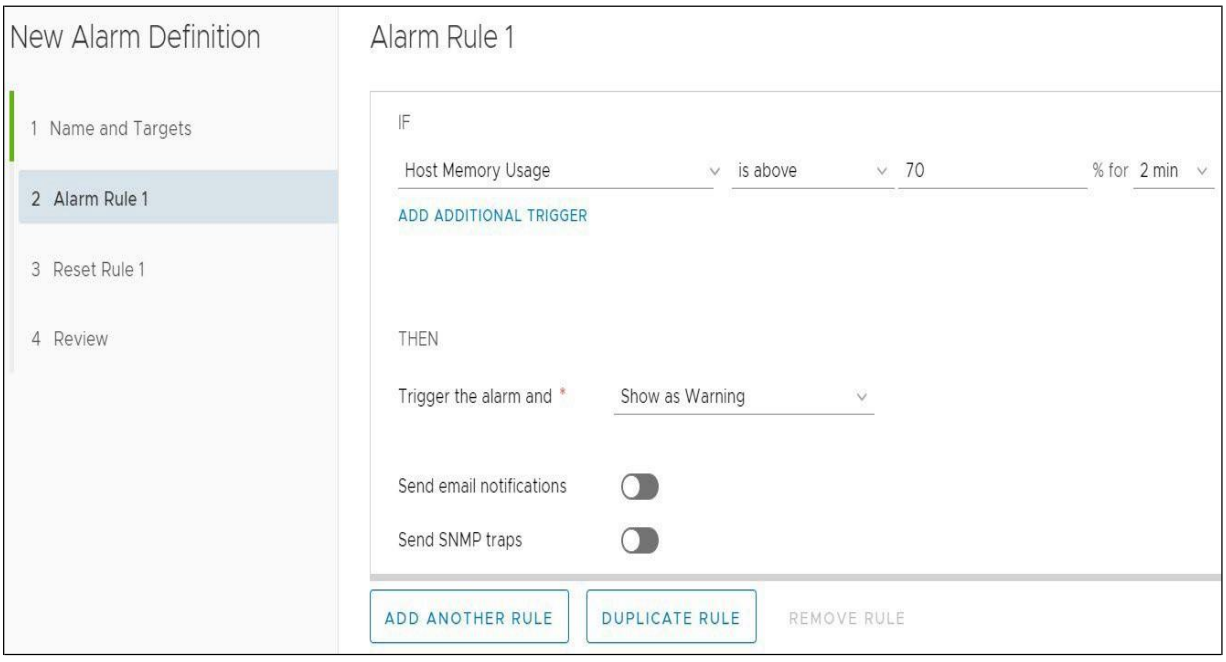
It is recommended to configure following Memory Alarms:

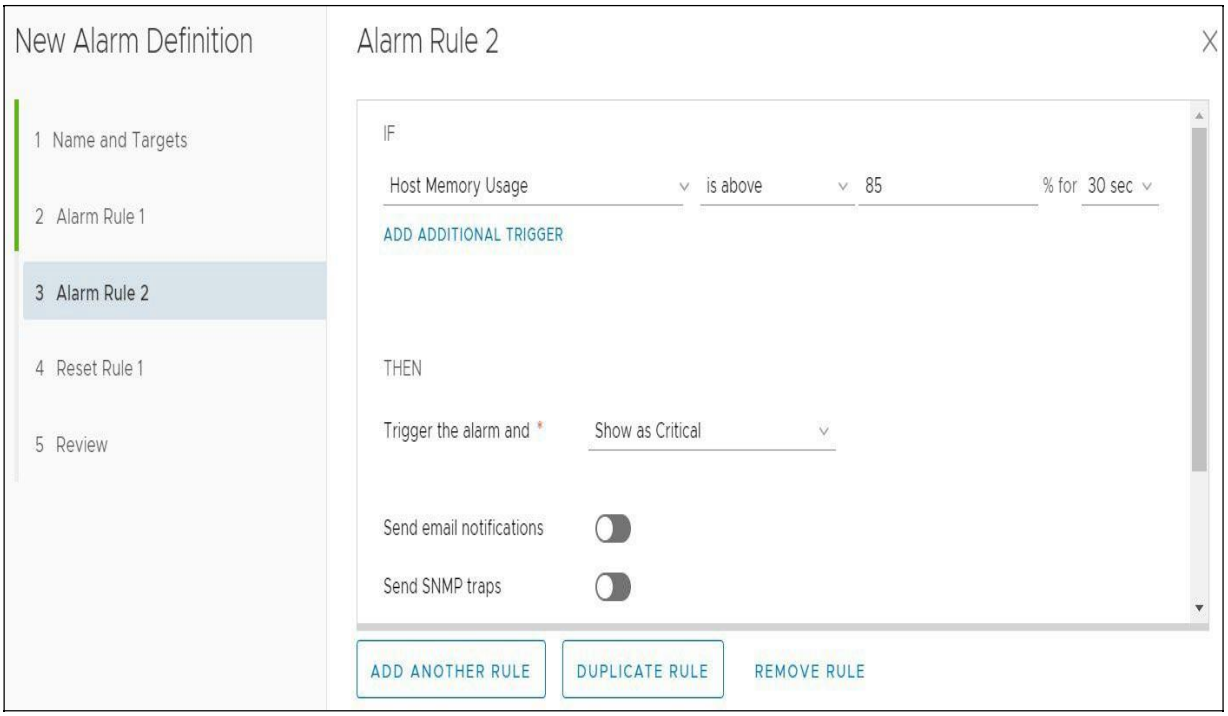
- Warning: 70% for 2 minutes
- Critical: 85% for 30 Seconds

To configure memory alarm, Login to vCenter, go to "Hosts and Clusters", click on the desired Data Center, on right hand side pane click on "Configure" Tab, under this tab select "Alarm Definition". On right hand side pane click on "ADD" button. It will open a window called "Name and Targets", put a descriptive alarm name e.g. Alarm for host Memory usage Threshold. Select the target as Hosts, click on "Next"

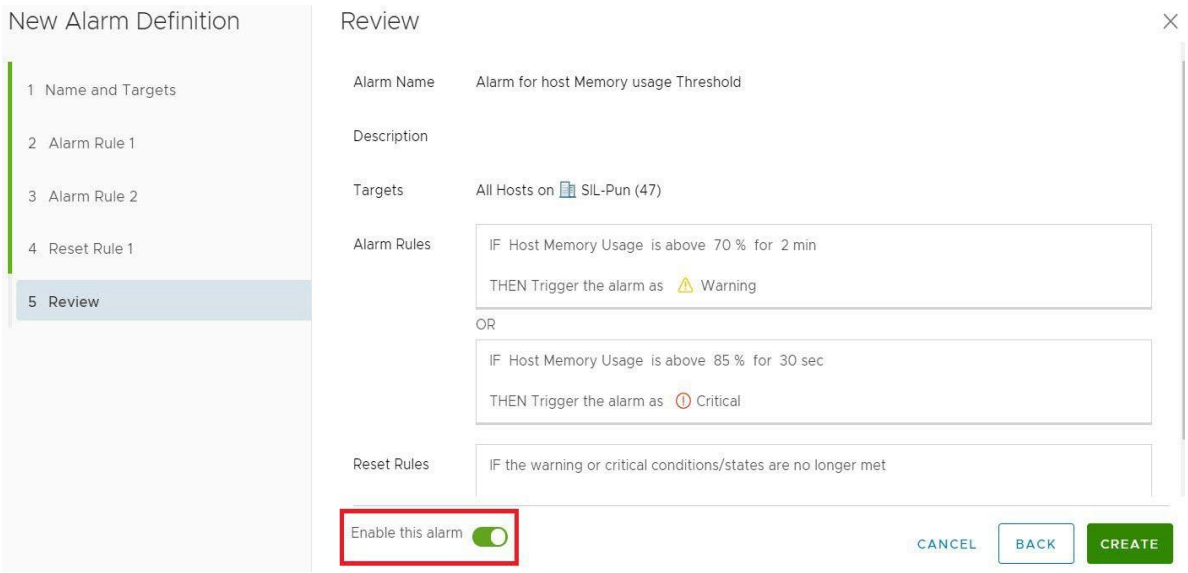


On Alarm Rule 1 window, under "IF" statement click on "Select Triger" dropdown arrow, now expand "Capacity and Usage" option, select Host Memory Usage. Now click on "Select an Operator" dropdown and select "is above", type 70 in the next field for %, in drop down "For" Select "2 Min". Now Under "THEN" statement for "Select Severity" dropdown, select "Show as Warning" option. To add one more rule, click on "Add another Rule" and configure it as displayed in second screenshot for Alarm Rule2.





After configuration of Alarm Rule 2 , click on “Next” button , again click “Next” button on Reset Rule1 window. In next window review the rules created are reflecting fine, make sure that “Enable this Alarm” option is enabled as shown in the screenshot and click on “Create” Button.

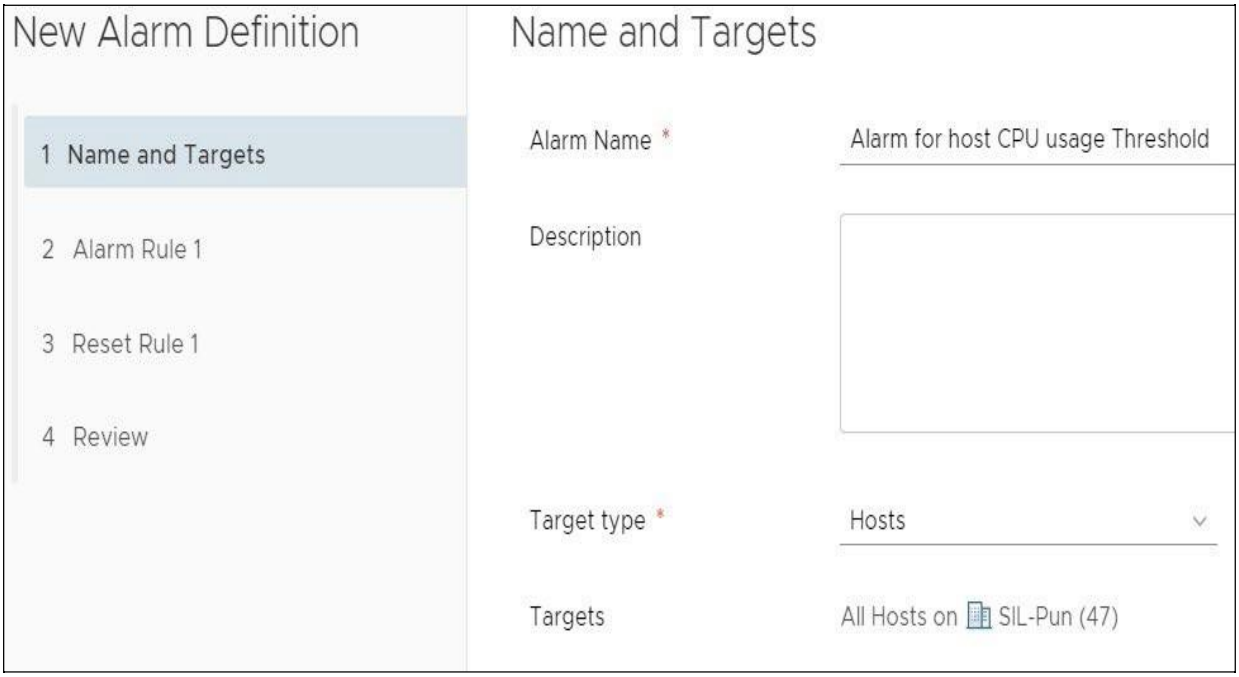


5.2.2. Configure CPU Alarms

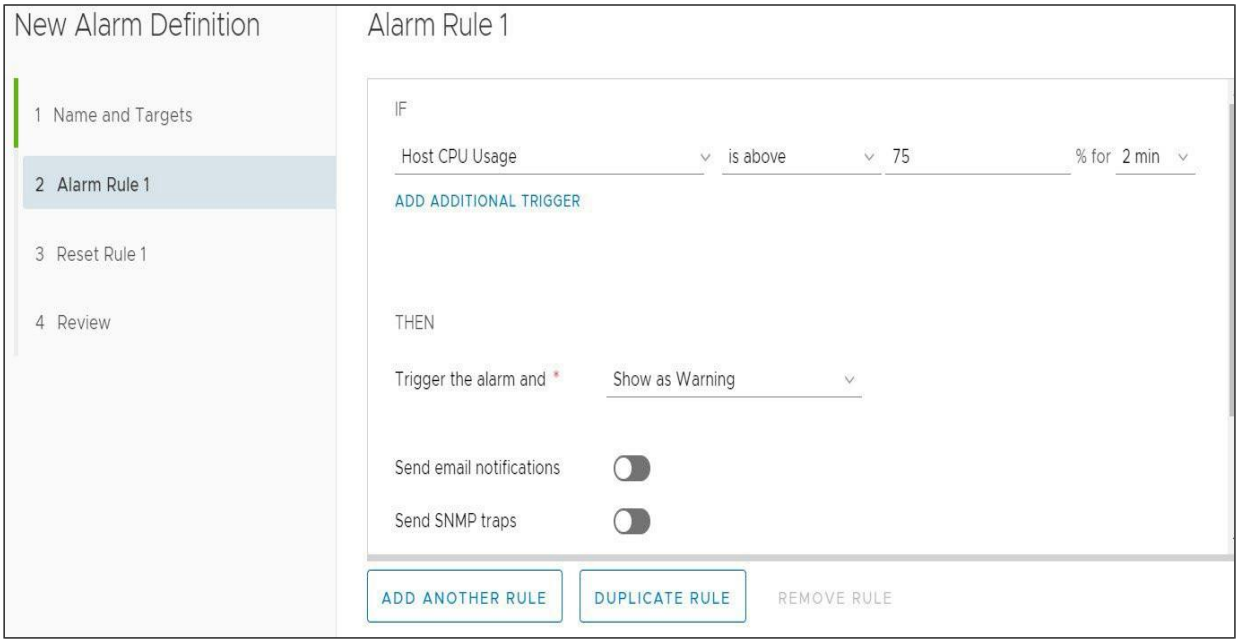
It is recommended to configure following Alarms for CPU.

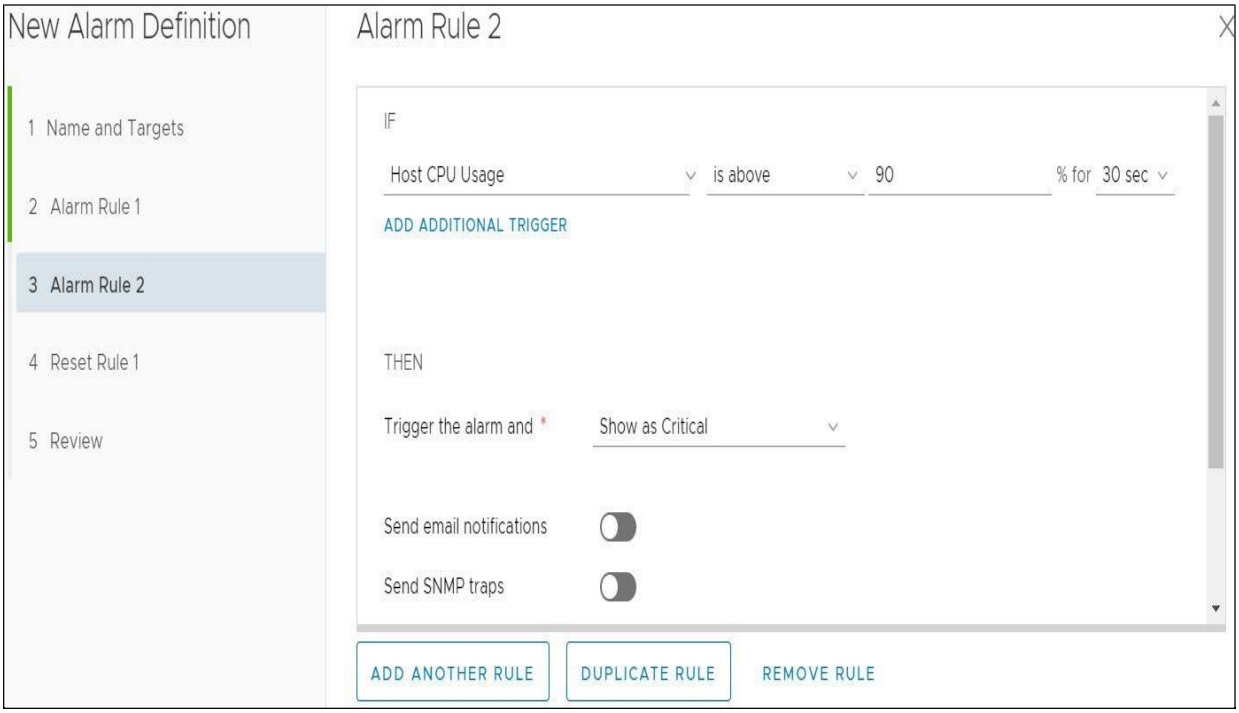
- Warning: 75% for 2 minutes
- Critical: 90% for 30 Seconds

To configure CPU alarms, Login to vCenter, go to Hosts and Clusters, click on the desired Data Center, on right hand side pane click on “Configure” Tab, under this tab select “Alarm Definition”. On right hand side pane click on “ADD” button. It will open a window called “Name and Targets”, put a descriptive alarm name e.g. Alarm for host CPU usage Threshold. Select the target as Hosts, click on “Next”

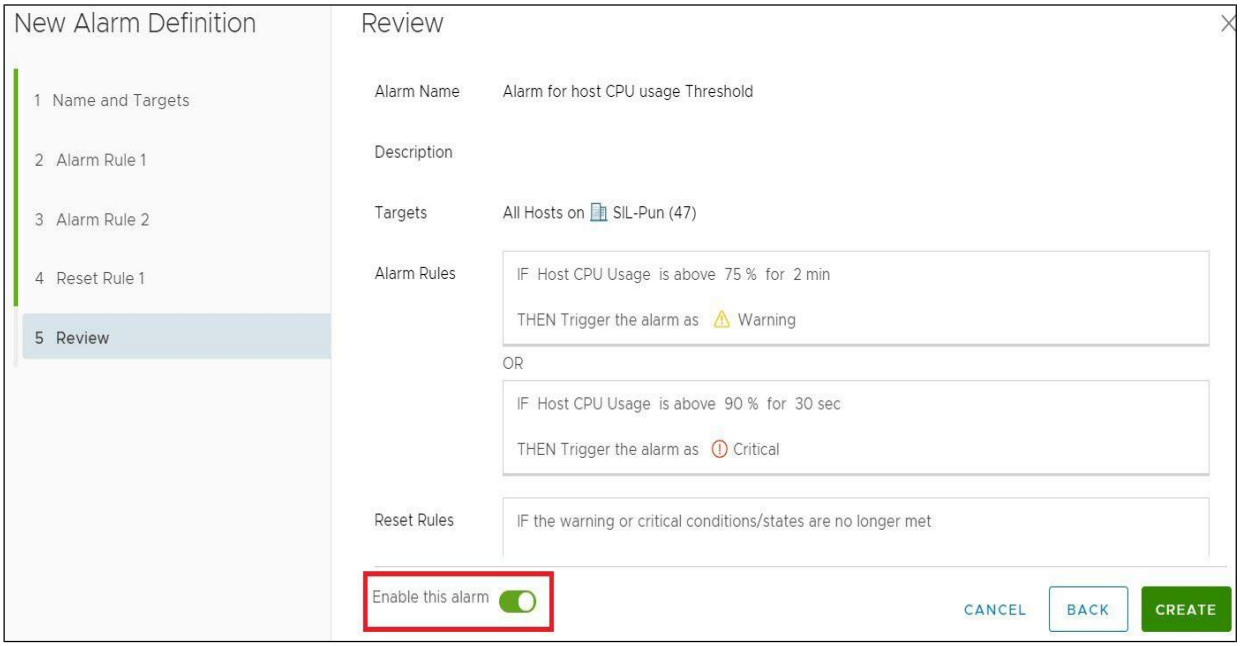


On Alarm Rule 1 window, under “IF” statement click on “Select Triger” dropdown arrow, now expand “Capacity and Usage” option, select Host CPU Usage. Now click on “Select an Operator” dropdown and select “is above”, type 75 in the next field for %, in dropdown “For” Select “2 min”. Now Under “THEN” statement for “Select Severity” dropdown, select “Show as Warning” option. To add one more rule, click on “Add another Rule” and configure it as displayed in second screenshot for Alarm Rule2.





After configuration of Alarm Rule 2 , click on “Next” button , again click “Next” button on Reset Rule1 window. In next window review the rules created are reflecting fine, make sure that “Enable this Alarm” option is enables as shown in the screenshot and click on “Create” Button.



6. Monitoring CPU and Memory resources

Following subsections explain the different monitoring mechanisms for CPU and Memory utilization. vCenter provides real time as well as historical monitoring capabilities.

It is recommended to proactively monitor the CPU and Memory utilization to avoid any possible impacts due to resource starvation.

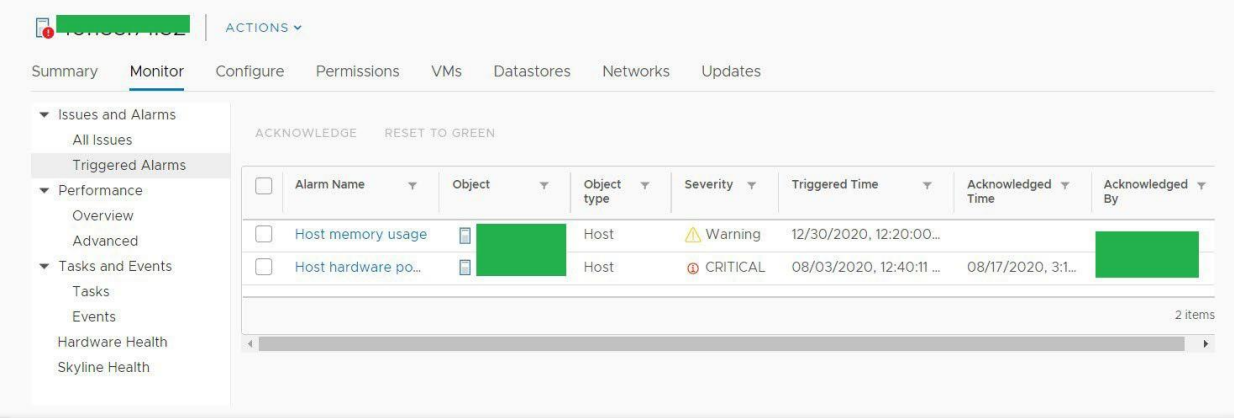
Note: Configuration mentioned in above section of this document is required to enable monitoring. For more details on VMware monitoring, it is recommended to refer VMware documentation.

6.1. Monitoring Alarms and Events

It is recommended to continuously monitor Alarms for CPU and Memory triggers.

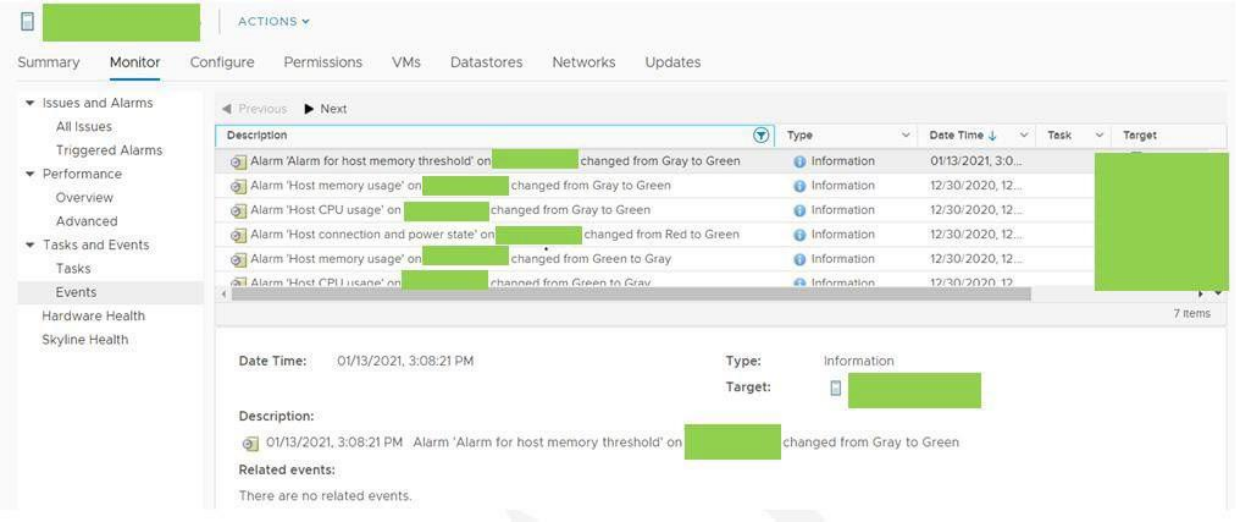
- For checking the Alarm, select the cluster or Host, click Monitor Tab. Under “Triggered Alarms” Tab, verify that there is not any Alarm triggered for CPU and Memory.
- Periodically check the Emails to verify that there is not any Alarm triggered for CPU or Memory (Considering that “Email Notification Action” is correctly configured for Alarms as recommended in above section).

Following screenshot shows the “Triggered Alarms” on the ESXi Host using vCenter



For checking Events related to Alarms, select the cluster or host; click “Monitor Tab ”. Under that view the Events by clicking the “Events” sub-tab. Look for the Events related to CPU and Memory Alarms. It shows the usage status Events like “Yellow” and “Red” which are corresponding to the Alarms thresholds “Warning” and “Alert” respectively.

Following screenshot shows the Alarm related “Events” on the ESXi Host using vCenter



6.2. Host level monitoring

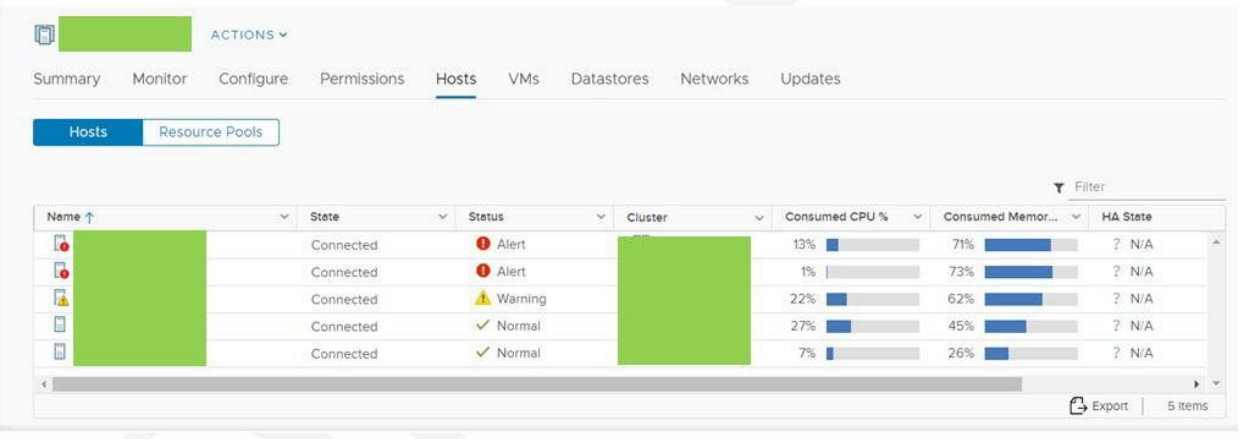
It is recommended to periodically monitor the CPU and Memory Utilization of each host for reservationless deployment.

6.2.1. Monitoring CPU/Memory utilization of ESXi Hosts

CPU and Memory utilization of each ESXi Hosts in a cluster can be monitored using vCenter

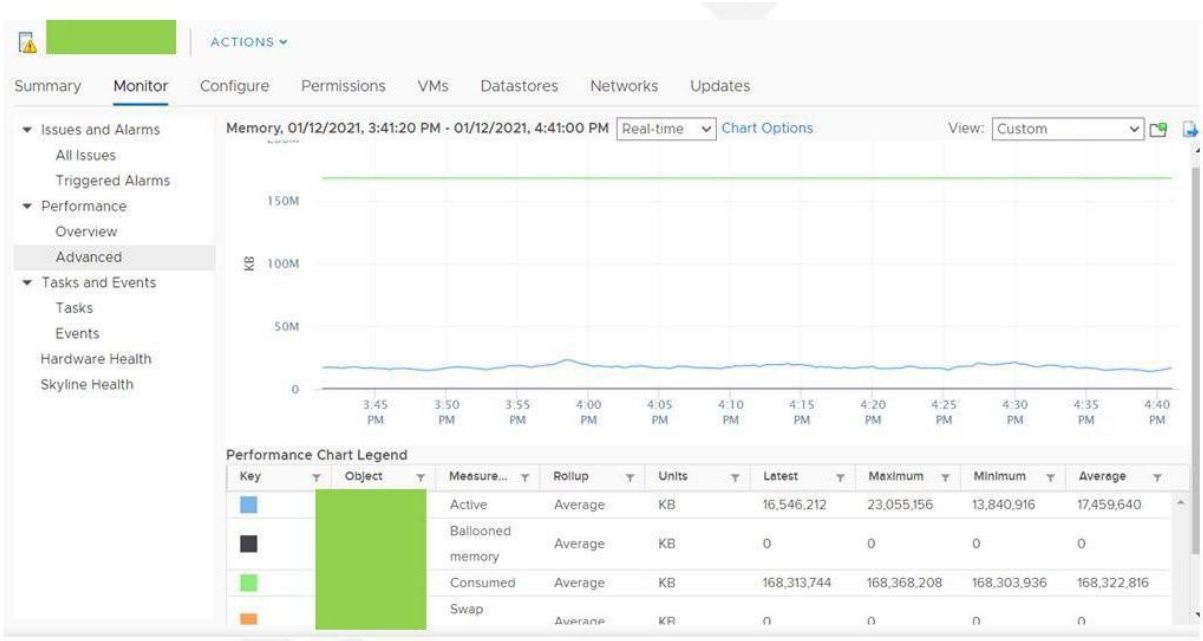
For monitoring CPU and Memory utilization of each ESXi Hosts in a cluster, select the Cluster and then click the Hosts Tab. This will display the percentage CPU and Memory usage for each host in a cluster. Ensure that the CPU and Memory utilization are less than 75% for each Host.

Following screenshot indicates the CPU and Memory utilization for different Hosts in a cluster



6.2.2. Monitoring Host using Performance Charts

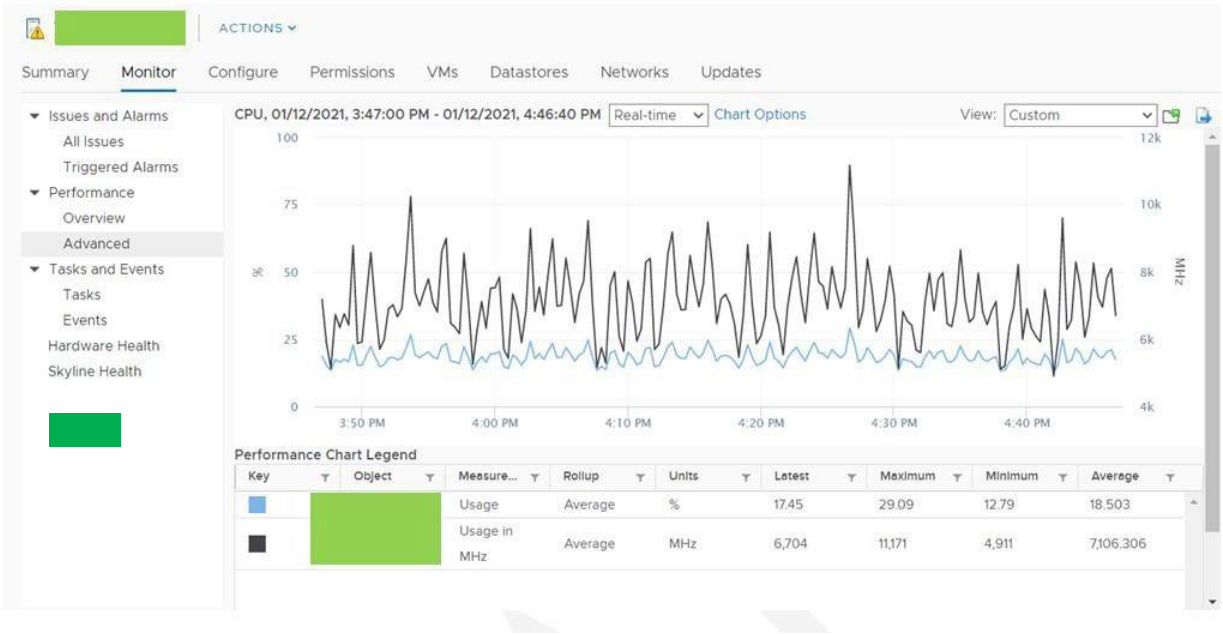
Host level Performance charts allows viewing performance data of a Host including CPU, Memory.
To view the Performance charts, select the Host and go to the Monitor Tab > Performance > Advanced. Click on the "Chart Options". Now you can select the options to view the real time or historical Memory or CPU. For Memory, select the different Memory options like Active, Consumed, Balloon, Swapped, Compressed, Usage etc. (Some Memory options like Active Memory may not be available for historical charts but available for real time charts. For Historical report, select "rollup" value as maximum for each CPU and memory options. Don't select "rollup" value of minimum or average). This lists the performance charts and data for selected items as shown below:



Following things can be monitored from the above Memory Performance charts:

- Percentage Memory Usage (Latest, Maximum, Minimum, Average).
- Consumed Memory usage of the host should be always less than 75% of the overall host Memory.
- Active Memory should be less than the Consumed Memory. If Active Memory is more than the Consumed Memory, then the Host Memory may not be sufficient to fulfill the Virtual Machine requirements.
- Ballooning and/or Compression indicates Host is starving for Memory resources. This may impact the performance of the Virtual Machines.
- Swap Memory indicates that Host is heavily starving for Memory Resources and this may have severe impact on the Virtual Machines.

Following screenshot shows the sample performance chats and data for real time CPU usage:



Following things can be monitored from the above CPU Performance charts:

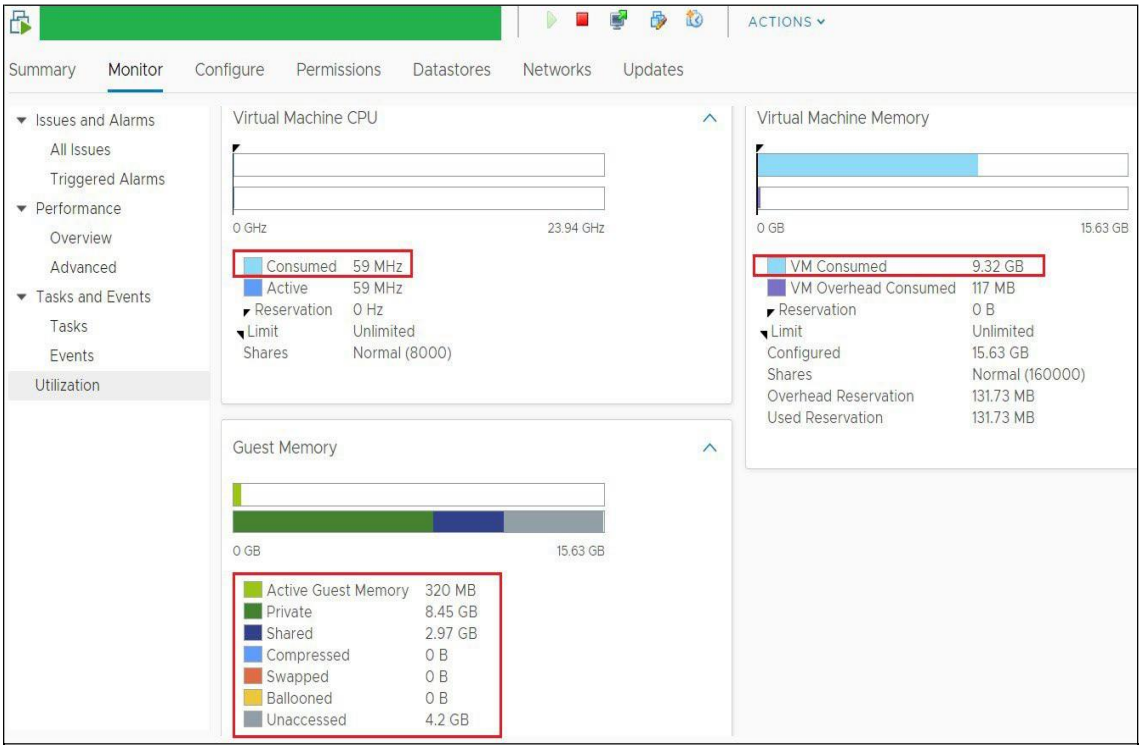
- Percentage CPU Usage (Latest, Maximum, Minimum, Average).
- Host CPU usage in MHz (Latest, Maximum, Minimum, Average).

6.3. VM (Virtual Machine) levelmonitoring

It is possible to do real time monitoring of CPU and Memory for each VM. It is also possible to extract the historical CPU and memory reports of a Virtual Machine. Monitoring real time and periodical CPU and Memory reports of Virtual Machine are very useful which may identify potential resource starvation for Virtual Machines.

6.3.1. VM real-time monitoring and Resource utilization

For monitoring current resource utilization for a Virtual Machine, select a Virtual Machine and in right hand side go to “Monitor” tab and click on “Utilization” option. There are statistics displayed for CPU and Memory as shown in below screenshot:



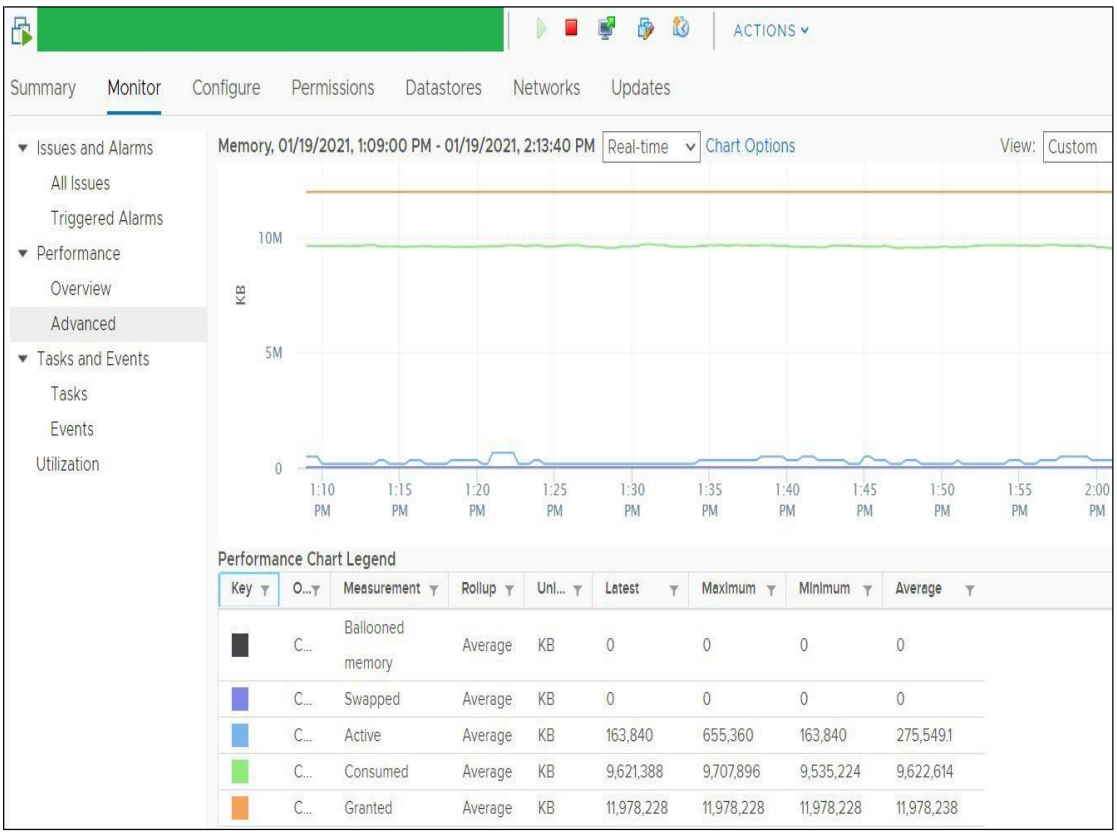
- Check that the Active CPU is not too high than Consumed CPU. If Active CPU value is much higher than the Consumed CPU, then the Virtual Machine is starving for the CPU.
- Ensure that the Consumed Memory is not less than the Active Memory. If Consumed Memory is less than the Active Memory, then the Virtual Machine may be starving for the Memory.

- Also verify that there is zero memory under Ballooned, Compressed and Swapped Memory. If there is Ballooned, Compressed or Swapped memory than this indicates the Host in which Virtual Machine resides is starving for the Memory.

6.3.2. Monitoring VM using Performance Charts

It is also possible to monitor the performance charts of Memory and CPU for a Virtual Machine like the Host Level performance charts.

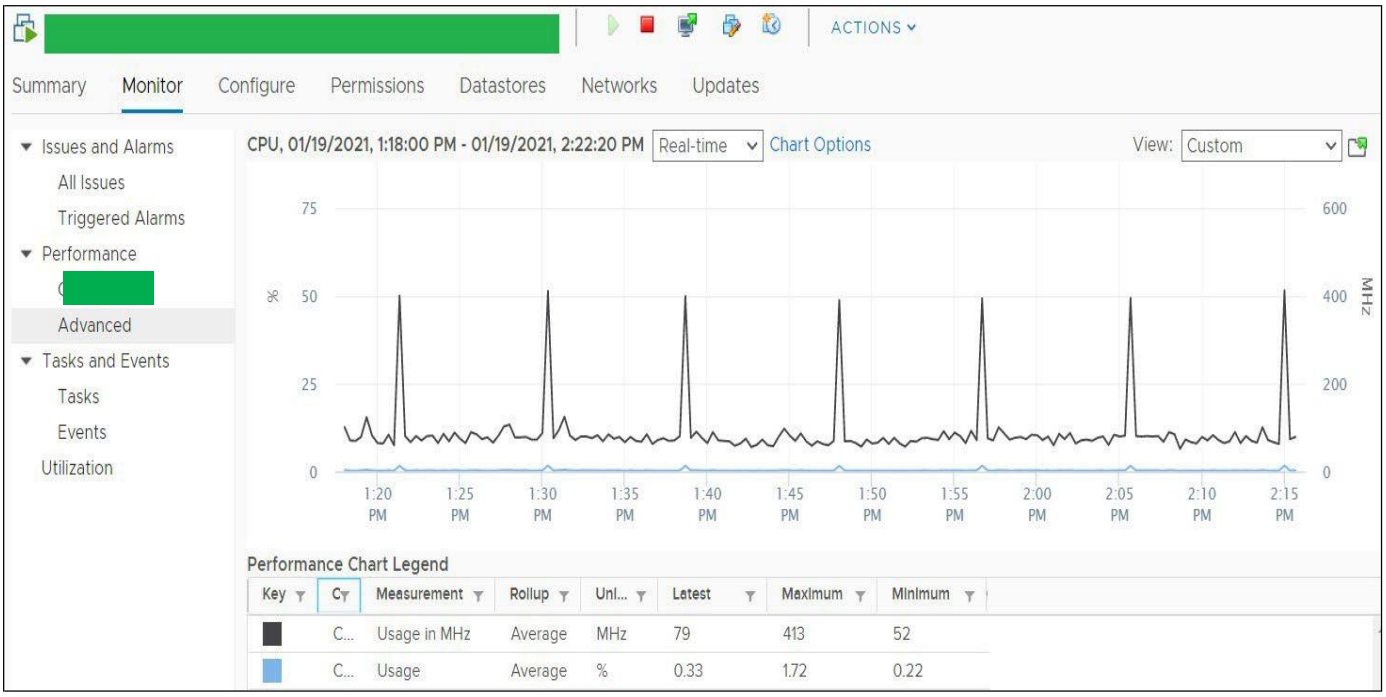
To view the Performance charts, select the Virtual Machine and on right hand side pane select "Monitor" tab, now click "Performance" and select "Advanced" Option. Click on the "Chart Options". Now from "Chart Metrics" select the options to view Memory or CPU. For Memory, select the different Memory options like Active, Consumed, Balloon, Swapped, Compressed, Usage (Some Memory options like Active Memory are not available for historical charts but they are available for real time charts). This will list the performance charts and data for selected items as shown below.



Following things can be monitored from the above Memory Performance charts:

- Percentage Memory Usage (Latest, Maximum, Minimum, Average).
- Active Memory should be less than the Consumed Memory. If Active Memory is more than the Consumed Memory, than the Virtual Machine may not be getting sufficient Memory from the Host, and this may impact the Virtual Machine Performance.
- Ballooning and/or Compression indicates Host in which Virtual Machine resides is starving for Memory resources. This may impact the performance of the Virtual Machine.
- Swap Memory indicates that Host in which Virtual machine resides is heavily starving for Memory Resources and this may have severe impact on the performance of the Virtual Machine.

Following screenshot shows the sample performance chats and data for real time CPU usage:



6.4. ESXi performance monitoring using command-line utilities

It is possible to monitor detailed information on system performance through the command line.

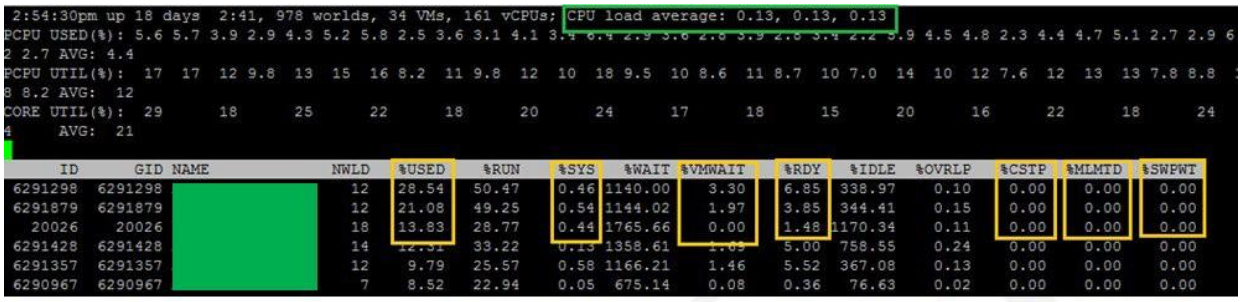
The resxtop and esxtop command-line utilities provide a detailed look at how ESXi uses resources in real time. The fundamental difference between resxtop and esxtop is that, resxtop can be used remotely, whereas esxtop can only be started through the ESXi Shell of a local ESXi host.

Virtual Machine CPU and Memory utilization as well as server-wide CPU and Memory statistics are displayed using this command line utility.

Note: By default, esxtop will provide ESXi server statistics of every 5 seconds. For monitoring resource statistics with accuracy of 2 seconds use option “d” with esxtop command to specify interval delay. (For Example: esxtop –d 2)

6.4.1. CPU Monitoring

To monitor CPU related performance parameters, ssh to ESXi server, login via root credentials, run command esxtop and press letter c from the keyboard.



Following is the brief overview:

CPU load average: CPU load average for last 1, 5 and 15 minutes.

%USED: CPU core cycles used via VM.

%SYS: Percentage of time spent by system to process interrupts and to perform other system activities on behalf of the world.

%VMWAIT: Percentage of time a VM was waiting for some VMkernel activity to complete (such as I/O) before it can continue. Includes % SWPWT and "blocked", but not IDLE time (as % WAIT does).

%RDY: Percentage of time a VM was waiting to be scheduled. If you observer values more than 10 percent, then it is a matter of concern here.
Possible Cause: CPU limit settings (Check % MLMTD) or Host CPU Availability

%CSTP: This value depicts the percentage of time a ready to run VM has spent in co-deschedule state.

%MLMTD: The percentage of time the vCPU was ready to run but deliberately wasn't scheduled because that would violate the "CPU limit" settings. Validate the CPU limits assigned for the VM.

6.4.2. Memory Monitoring

To monitor memory related performance parameters, run esxtop and press ‘m’ from the keyboard. Now to add memory compression related statistics, press ‘f” and then press ‘Q’ and hit “Enter”.

```
Current Field order: aBcDefgHijKLmnOpQ

  A:  ID = Id
* B:  GID = Group Id
  C:  LWID = Leader World Id (World Group Id)
* D:  NAME = Name
  E:  NWLD = Num Members
  F:  MEM ALLOC = MEM Allocations
  G:  NUMA STATS = Numa Statistics
* H:  SIZE = MEM Size (MB)
  I:  ACTV = MEM Active (MB)
  J:  MCTL = MEM Ctl (MB)
* K:  SWAP STATS = Swap Statistics (MB)
* L:  LLSWAP STATS = Llswap Statistics (MB)
  M:  CPT = MEM Checkpoint (MB)
  N:  COW = MEM Cow (MB)
* O:  OVHD = MEM Overhead (MB)
  P:  CMT = MEM Committed (MB)
* Q:  ZIP = MEM Compression (MB)

Toggle fields with a-q, any other key to return:
```

```
 9:32:21am up 18 days 19:05, 854 worlds, 19 VMs, 117 vCPUs; MEM overcommit avg: 0.00, 0.00, 0.00
PMEM /MB: 393181   total:  4037     vmk,195952 other, 193190 free
VMKMEM/MB: 392795 managed:  4542 minfree, 49592 rsvd, 343203 ursvd, high state
NUMA /MB: 196571 (110851), 196607 (81955)
PSHARE/MB: 46725  shared,   112  common: 46613 saving
SWAP /MB:     0    curr,     0 rclmtgt:              0.00 r/s,    0.00 w/s
ZIP  /MB:     0  zipped,     0   saved
MEMCTL/MB:    0    curr,     0  target, 84867 max

    GID NAME             MCTL?  MCTLSZ  MCTLTGT  MCTLMAX  SWCUR  SWTGT  SWR/s  SWW/s  LLSWR/s  LLSWW/s  CACHESZ  CACHEUSD  ZIP/s  UNZIP/s
5168055                    Y      0.00     0.00  1058.36   0.00   0.00   0.00   0.00     0.00     0.00     0.00      0.00   0.00     0.00
3685769                    Y      0.00     0.00 15703.49   0.00   0.00   0.00   0.00     0.00     0.00     0.00      0.00   0.00     0.00
   4321                    N      0.00     0.00     0.00   0.00   0.00   0.00   0.00     0.00     0.00     0.00      0.00   0.00     0.00
   4364                    N      0.00     0.00     0.00   0.00   0.00   0.00   0.00     0.00     0.00     0.00      0.00   0.00     0.00
   4411                    N      0.00     0.00     0.00   0.00   0.00   0.00   0.00     0.00     0.00     0.00      0.00   0.00     0.00
7024967                    N      0.00     0.00     0.00   0.00   0.00   0.00   0.00     0.00     0.00     0.00      0.00   0.00     0.00
```

Following is the brief overview:

MEM overcommit avg: Average memory overcommitment for last 1, 5 and 10 minutes. Zero indicates there is no Memory overcommitment. Non-zero value like 0.08 indicates 8% Memory is overcommitment in the ESXi host.

VMKMEM State: VMKMEM state shall remain in “high state” till free memory does not fall below 6%. When free memory falls below threshold of 6%, it will go to different states like “High”, “Soft”, “Hard” and “Low” respectively. States like “High”, “Soft”, “Hard” and “Low” indicates heavy memory starvation.

MCTLSZ: Amount of guest physical memory (MB) the ESXi Host is reclaiming by balloon driver. If larger than 0 host is forcing VMs to inflate balloon driver to reclaim memory as host is overcommitted.

SWCUR: Memory (MB) that has been swapped by VMKernel.
Possible Cause: Memory overcommitment
If larger than 0, host has swapped memory pages in the past.

SWR/s, SWW/s: Rate at which the ESXi host is writing to or reading from swapped memory. If value is larger than 0 Host is actively writing and reading from swap.
Possible cause: Memory overcommitment.

CACHEUSD: Memory (MB) compressed by ESXi Host.
Possible Cause: Memory Overcommitment
Any value greater than 1 is a matter of concern

ZIP/s: Values larger than 0 indicate that the host is actively compressing memory.
Possible Cause: Memory Overcommitment

UNZIP/s: Value larger than 0 indicate that the host is accessing compressed memory.
Possible Cause: Memory Overcommitment

7. Analysis of issues in reservationless deployment

In case on any issues observed with Avaya application having reservationless deployment, make sure that there is no starvation of CPU or Memory.

Guidelines mentioned in this section will be useful to identify possibility of issue being occurred due to resource starvation.

7.1. ESXi host analysis for resource starvation

On occurrence of any issue in reservationless deployments, check following things for ESXi host:

1. Make sure that CPU and Memory utilization of ESXi hosts are continuously less than the recommended value of 75%. This can be verified from historical and real-time usage monitoring.

2. Verify that, there is no Alarm or Events triggered for CPU or Memory thresholds.
3. Verify that, the real-time and historical (Last 24 Hours) performance charts of ESXi host for Memory and CPU usage.
    a) Look at the value of Memory Balloon (Average) in real-time and historical charts. Ballooning suggests that ESXi is under memory pressure and thus memory over commitment may be affecting the performance of the virtual machine.
    b) Check the values of Compressed, Swapped memory in real-time charts. If memory is compressed or swapped, this indicates memory starvation on the ESXi host.
    c) Compare the values of Consumed Memory and Active Memory. If consumed is lower than active, this suggests that host may not have sufficient memory resources available for virtual machines.

7.2. VM (Virtual machine) analysis for resource starvation

Following are the high-level verification steps which may identify the possibility of application being starving for CPU or Memory resources. If any of these steps indicate about CPU or Memory resource starvation, then it should be addressed to avoid any impact on the application behavior.

1. Performance charts:
   Verify following using performance charts of a Virtual machine.
    a) Select the Virtual Machine in question from vCenter, select Monitor Tab > Performance > Advanced , and then compare the values of Consumed Memory and Active Memory in real-time memory chart. If consumed is lower than active, this suggests that the Virtual machine may not be getting the required memory resources.
    b) Select the virtual machine in question from the vCenter, select the Monitor Tab > Performance > Advanced , and then look at the value of Memory Balloon (Average) in real-time and historical memory charts. Ballooning suggests that ESXi is under memory pressure and thus memory overcommitment is affecting the performance of that virtual machine.
    c) Select the Virtual Machine in question from vCenter, select Monitor Tab > Performance > Advanced , and then look at the values of Swap-In and Compressed in real-time memory chart. Swapping in and decompressing at the host level indicate more significant memory pressure.

2. Virtual Machine Resource Utilization:
   Select the Virtual machine in question and on right hand side go to “Monitor” tab and click on “Utilization” option

    a) If consumed memory is lower than active memory, this suggests that the Virtual machine may not be getting the required memory resources.

b) Check Ballooned, Swapped and Compressed memory values. If any of these values is non-zero, it indicates that ESXi is starving for memory.
c) If consumed CPU is much lower than active CPU, this suggests the possibility of virtual machine being not getting the required CPU resources.

8. Reports and logs to be captured in case of any issues

For reservation-less deployment, it is highly recommended to capture Real-Time performance data of CPU and Memory as soon as an issue observed. Real-time CPU and Memory reports capture data for last one hour.

The data needs to be captured for all Avaya applications having reservationless deployment whenever any issue gets observed with any of the Avaya application.

Note: VMware historical reports must be captured within 24 Hours from the time of actual occurrence of the issue. Issues for Avaya applications are not supported, if VMware reports are collected within 24 Hours from the time of actual occurrence of the issue.

Avaya support may ask for additional reports like vm-support bundle. It is the responsibility of the customer to provide all the reports requested by Avaya support representative.

8.1. VMware reports and logs to be collected for Avaya support

Following VMware reports and logs should be collected while raising request with Avaya support team, if any Avaya application is running without reservation.

1. Real-Time reports for Memory and CPU of Host/s
2. Historical reports for Memory and CPU of Host/s (Last 24 Hours)
3. Real-Time reports for Memory and CPU of VM/s
4. Historical reports for Memory and CPU of VM/s (Last 24 Hours)
5. Screenshots of resource assignments for VM/s

These reports can be captured by following the below mentioned steps on vCenter:

Host level Real-Time reports for Memory and CPU:

Collect the following real-Time CPU and Memory reports as soon as an issue observed:

- Select the Host, click Monitor Tab > Performance > Advanced
- Keep the Real-time selection
- Click on “Chart Options”
- Keep Line graph in Chart Type
- On CPU Tab select following counters “Usage”, “Usage in MHz” and “Total Capacity” and then Click OK.

- On Memory Tab select following counters “Balloon memory”, “Active”, “Consumed”, “Swap consumed”, “Swap in”, and “Compressed”. Click Ok.

Host level Historical reports for Memory and CPU (Last 24 Hours):

Collect the following historical CPU and Memory reports as soon as an issue observed:

- Select the Host, click Monitor Tab > Performance > Advanced.
- Select the time as Last day.
- Click on “Chart Options”
- Keep Line graph in Chart Type On CPU Tab select following counters “Usage” (with maximum rollup) and “Usage in MHz” (with maximum rollup), Click Ok.
- On Memory Tab select following counters “Ballooned memory” and “Consumed”, click Ok.

VM level Real-Time reports for Memory and CPU:

Collect the following real-Time CPU and Memory reports as soon as an issue observed:

- Select the VM, click Monitor Tab > Performance > Advanced
- Keep the Real-time selection
- Click on “Chart Options”
- Keep Line graph in Chart Type
- On CPU Tab select following counters “Usage” and “Usage in MHz”. Click OK.
- On Memory Tab select following counters “Ballooned memory”, “Active”, “Swap out”, “Swap in”, “Consumed” and “Compressed”. Click Ok.

VM level Historical reports for Memory and CPU (Last 24 Hours):

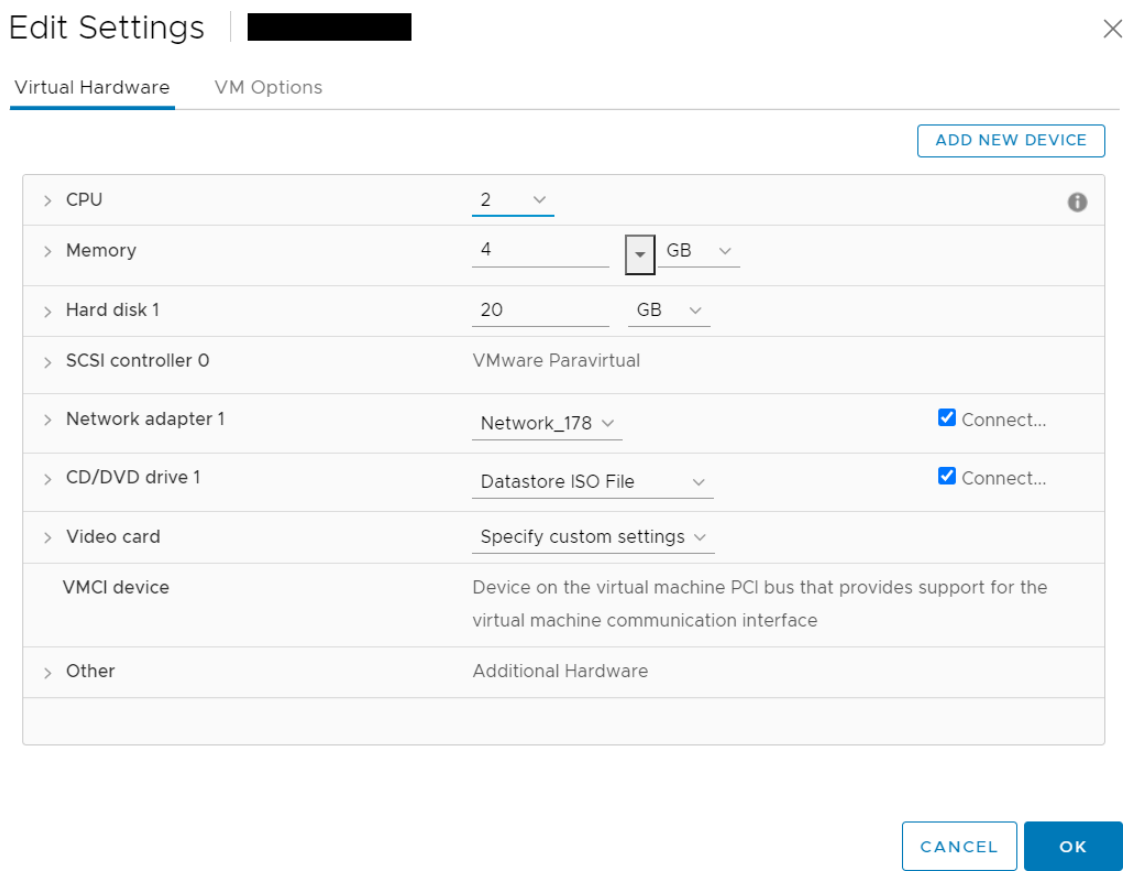
Collect the following historical CPU and Memory reports as soon as an issue observed:

- Select the VM, click Monitor Tab > Performance > Advanced.
- Select the time as Last day.
- Click on “Chart Options”
- Keep Line graph in Chart Type
- On CPU Tab select following counters “Usage” (with maximum rollup) and “Usage in MHz” (with maximum rollup), Click Ok.
- On Memory Tab select following counters “Balloon” and “Consumed”, click Ok.

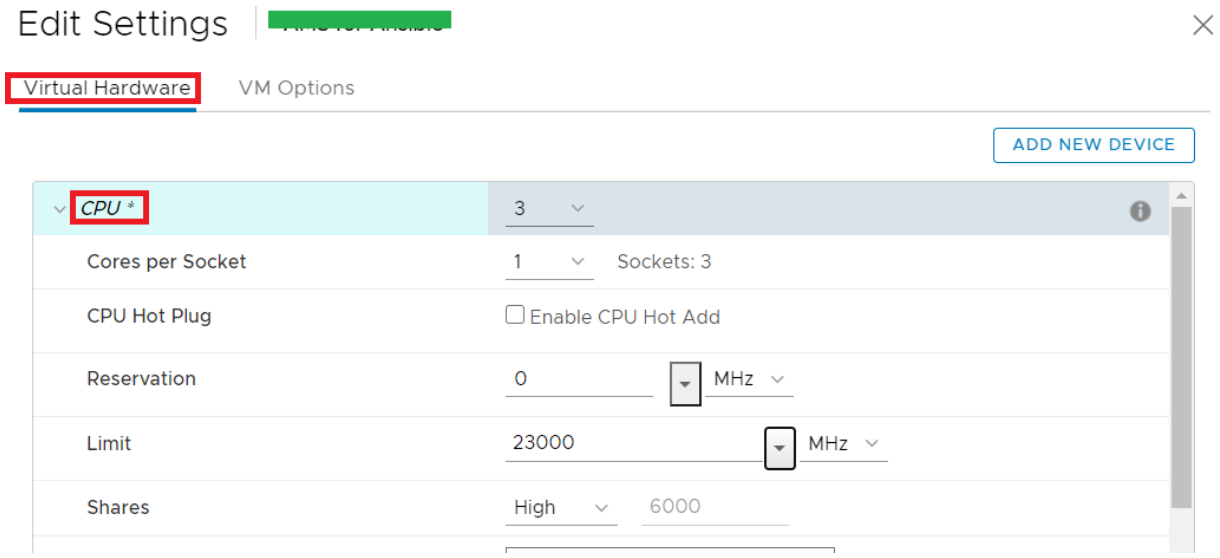
Screenshots of resource assignments for VM/s:

Attach the following screenshots of Virtual Machine resource illustrated with following example screenshots of a VM:

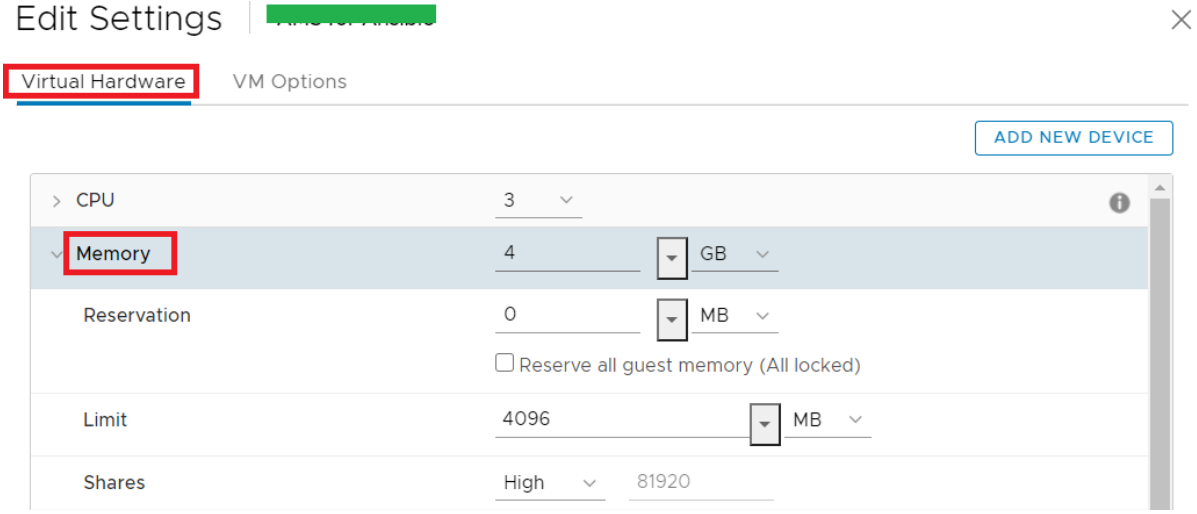
1. Select VM, right click, Edit Settings and select Virtual Hardware Tab. Take a snap.



2. Select VM, right click, Edit Setting and select Virtual Hardware Tab. Click on CPU and take a snap of details.



3. Select VM, right click, Edit Setting and select Virtual Hardware Tab. Click on Memory and take a snap of details.



9. References

Following are the references from VMware documentation:

- Performance Best Practices for VMware vSphere 7.x and 8.x
  - https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/vsphere-esxi-vcenter-server-70-performance-best-practices.pdf
  - https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/vsphere-esxi-vcenter-server-80-performance-best-practices.pdf

- vSphere Resource Management
  - https://docs.vmware.com/en/VMware-vSphere/7.0/vsphere-esxi-vcenter-server-703-resource-management-guide.pdf
  - https://docs.vmware.com/en/VMware-vSphere/8.0/vsphere-esxi-vcenter-802-resource-management-guide.pdf

- vSphere Monitoring and Performance
  - https://docs.vmware.com/en/VMware-vSphere/7.0/vsphere-esxi-vcenter-server-703-monitoring-performance-guide.pdf
  - https://docs.vmware.com/en/VMware-vSphere/8.0/vsphere-esxi-vcenter-802-monitoring-performance-guide.pdf

Please e-mail any questions or comments pertaining to these Application Notes along with the full title name and filename, located in the lower right corner, directly to the Avaya DevConnect Program at devconnect@avaya.com.