Capacity and Performance Monitoring

Benefits of Capacity and Performance Monitoring

Capacity and performance monitoring ensures that the current system is running within safe engineering limits on the customer's network. Understanding the current capacity of the system and monitoring this capacity profile on a regular basis allows both the customer and Cisco to understand the subscriber growth and behavior for future trending and capacity planning activities. Regular capacity and performance monitoring also ensures the health and stability of the nodes, which helps to prevent capacity and performance exposures.

Monitoring Strategy

Along with specific vendor and industry Best Practices, capacity and performance monitoring within Cisco HCS can be divided between three major areas. The following are not meant to be an all-inclusive but simply a high-level checklist of critical areas:

1. UC provisioning Domain Manager tool reporting tools
2. Cisco Prime Collaboration Assurance (PCA) for UC Video and Voice monitoring and reporting
3. Infrastructure:
   - Compute
     - Hardware, Traffic and NetFlow monitoring. Cisco UCS Manager supports the entire Cisco UCS server and Cisco HyperFlex Series hyperconverged infrastructure.
   - Network
     - Physical and Virtual network resources, devices, connections, and performance
     - Firewall monitoring using Cisco Adaptive Security Device Manager (ADSM)
     - Monitoring of all NX-OS-enabled deployments using Cisco Data Center Network Manager (DCNM)
   - Storage
     - IOPS to check for spikes, especially during operations like provisioning or backups

  - Latency which is a good indicator of congestion in the system
  - Disk space within the SAN
  - Space utilization within VMs themselves. Use Cisco Unified Real-Time Monitoring Tool (RTMT) for UC apps.

- VMware vCenter
  - Virtualization CPU monitoring
    - Monitor both pCPU and vCPU
  - Virtualization Memory monitoring
    - ESXi Host RAM utilization
    - VM RAM utilization

Note RAM oversubscription of Cisco UC and Cisco Management apps is not supported

- WAN
  - Safeguard that remote links are operating properly
  - Ensure capacity between data centers and customer sites

Capacity Monitoring Metrics

The two main objectives of Cisco HCS capacity reporting are:

1. Identify immediate performance concerns.
2. Provide trending information based on usage patterns to ensure timely upgrades.

Cisco HCS collects metrics in the following areas on a per-VM basis, regardless of the applications on the VM:

- CPU (Average, Peak, Ready Time)
- Memory (Average, Peak, Swap, Overhead)
- Disk Usage (Latency and IOPS)

Network infrastructure metrics are also important during network integration through a fresh install, or when growth occurs in the network that requires software and hardware upgrades. Depending on the topology of the network deployed, overall system capacity in terms of supported subscribers and customers can vary. You must track network infrastructure metrics regularly to manage potential growth as the Cisco HCS system expands.

Monitoring capacity metrics will help you identify bottlenecks in the system will become evident, and perform capacity planning to ensure that bottlenecks do not inhibit system resources or end-user experience.

The following tables provide a list of metrics that are important to monitor and trend on a regular basis.

Table 1: System metrics

System metrics	Source
Subscribers	Cisco Prime Collaboration

System metrics	Source
Customers /Tenants	Cisco Prime Collaboration
Endpoints	HCM-F
Busy Hour Call Attempts (BHCA)	Cisco Prime Collaboration

Table 2: VM metrics

VM metrics	Source
CPU Utilization	vCenter
Memory Utilization	vCenter
IOPS	vCenter
CPU ready	vCenter
IOPS read instructions	vCenter
IOPS write instructions	vCenter
VM Name	vCenter
vCPUs	vCenter
CPU Reservation	vCenter
Memory Reservation	vCenter
Total NICs	vCenter
VMState	vCenter

Table 3: CPU Metrics

CPU Metrics	Source
Average	vCenter
Peak	vCenter
Ready Time	vCenter
CPU Idle	vCenter
CPU Used	vCenter
CPU Wait	vCenter

Table 4: Disk Usage Metrics

Disk Usage Metrics	Source
deviceLatency	vCenter
deviceReadLatency	vCenter
deviceWriteLatency	vCenter
kernelLatency	vCenter
kernelReadLatency	vCenter
kernelWriteLatency	vCenter
maxTotalLatency	vCenter
numberRead	vCenter
numberWrite	vCenter
Total IOPS	vCenter

Table 5: Memory Metrics

Memory	Source
Average	vCenter
Peak	vCenter
Swap	vCenter
Overhead	vCenter

Table 6: Provisioning metrics

Provisioning metrics	Source
Cisco Unified Communications Managers	Service Inventory / HCM-F
Cisco Unified Communications IM and Presence	Service Inventory / HCM-F
Cisco Unity Connection	Service Inventory / HCM-F
Cisco HCM-F	Service Inventory / HCM-F
Cisco Prime Collaboration Assurance	Service Inventory / HCM-F
Cisco Unified Contact Center	Service Inventory / HCM-F

Table 7: Network infrastructure metrics

Network infrastructure metrics	Source
VLANs	CLI
VLAN Port Instances	UCS Manager
VRFs	CLI
UCS Chassis	CLI
BGP Peers	CLI
Static Routes	CLI
HSRP Instances	CLI
OSPF Adjacencies (if applicable)	CLI
VMs	vCenter
Server Blades/Hosts	vCenter

Table 8: CUCM Stats

CUCM Stats	Source
CPU Utilization	Cisco Prime Collaboration
Memory Utilization	Cisco Prime Collaboration
IOPS	vCenter
Disk Utilization	vCenter
MTP Resources	Cisco Prime Collaboration
MOH Resources	Cisco Prime Collaboration
Conferencing Resources (HW & SW)	Cisco Prime Collaboration
Location Bandwidth	Cisco Prime Collaboration
Calls Attempted	Cisco Prime Collaboration
Calls Completed	Cisco Prime Collaboration
Calls in Progress	Cisco Prime Collaboration
Number of Registered Phones	Cisco Prime Collaboration
Number of Registered Gateways	Cisco Prime Collaboration
Number of CTI Ports	Cisco Prime Collaboration

Table 9: Cisco Unified IM and Presence Stats

Cisco Unified IM and Presence Stats	Source
CPU Utilization	Cisco Prime Collaboration
Memory Utilization	Cisco Prime Collaboration
IOPS	vCenter
Disk Utilization	Cisco Prime Collaboration

Table 10: Unity Stats

Unity Stats	Source
CPU Utilization	Cisco Prime Collaboration
Memory Utilization	Cisco Prime Collaboration
IOPS	vCenter
Disk Utilization	Cisco Prime Collaboration
Percentage Active Inbound CUCxn Ports	Cisco Prime Collaboration
Percentage Active Outbound CUCxn Ports	Cisco Prime Collaboration

Table 11: Blade Metrics

Blade Metrics	Source
Server Model #	vCenter
# of Nics	vCenter
# of Cores	vCenter
# of Threads	vCenter

Table 12: Performance metrics

Performance metrics	Source
Call Success Rate	Cisco Prime Collaboration

Table 13: ASA Metrics

ASA Metrics	Source
ASA Security Contexts	ASA CLI
# of pps	ASA CLI
# of Mbps	ASA CLI

Capacity Exhaustion

Capacity exhaustion refers to system failure as a result of one or more components reaching or exceeding capacity, usually during a heavy load. Cisco HCS is designed to avoid capacity exhaustion, but you must monitor performance to ensure that components are operating within healthy parameters.

Capacity Exhaustion Example

This is an example of a system that is experiencing sufficient load to put it at risk of capacity exhaustion.

Table 14: Field Resource Utilization (Example)

CPU: 40%	EngPoint: 80%	Utilization: 40/80 (50%)
Memory: 30%	EngPoint: 80%	Utilization: 30/80 (37.5%)
IOPS: 25%	EngPoint: 75%	Utilization: 25/75 (33.3%)

"EngPoint" refers to the Engineering Point, or maximum capacity of the system for the subscriber profile. The Engineering Point is calculated based on reasonable worst-case assumptions with regard to resource usage and traffic spikes.

This example assumes that:

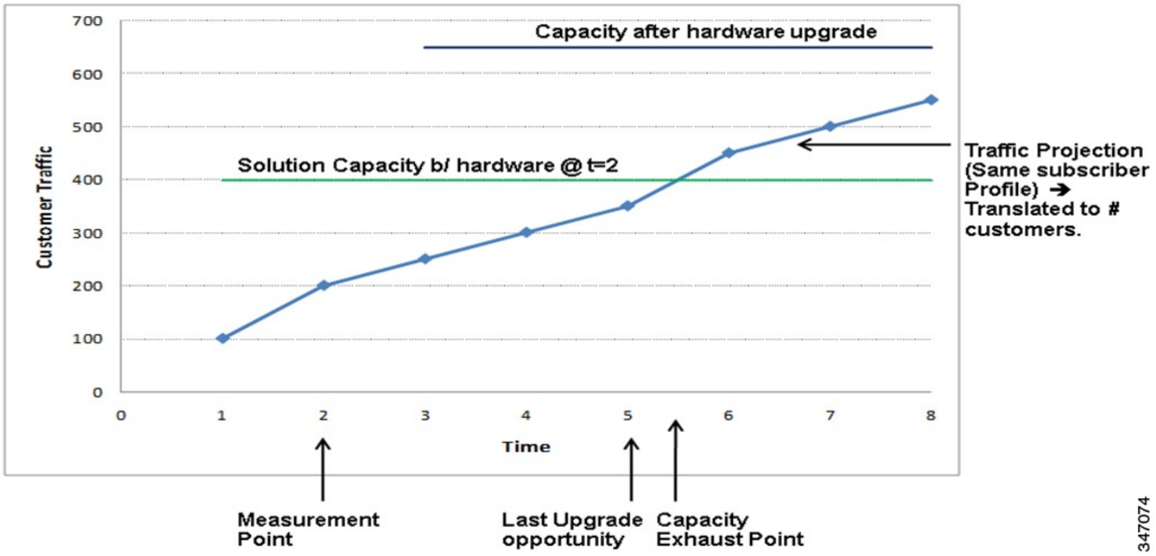
- No-load usage is negligible and the usage pattern is linear.
- All latency parameters are within tolerance.
- The Engineering Point (EngPoint) accounts for traffic spikes (peaks).

Based on these assumptions, we can conclude that the example system described above is CPU-limited. It may be possible to increase the capacity of the example system by upgrading hardware.

Capacity Exhaustion Planning

The following chart shows an example of planning to prevent capacity exhaustion. Based on traffic calculations made at time point 2, system exhaustion is projected between time points 5 and 6 (assuming linear growth). An upgrade should be performed no later than time point 5 (but preferably earlier) in order to avoid system failure.

Figure 1: Capacity planning and forecasting



Capacity Alarms

The Session Controller will raise alarms when any of the scalability metrics reach certain percentages of the licensed limit. There are three alarm thresholds: minor, major, and critical. The critical alarm is always raised when the licensed limit is breached. The other thresholds (and the level at which the critical alarm is cleared) are configurable, with the following default values:

- The minor alarm is raised if a metric reaches 60% of the licensed capacity and cleared if it drops below 50%.
- The major alarm is raised if a metric reaches 80% of the licensed capacity and cleared if it drops below 70%.
- The critical alarm is cleared if the metric for which it was raised drops below 90% of the licensed capacity.

Backup and Restore

A service provider must take great care with backup and restore activities because there are engineering rules associated with these activities. For full details on backup and restore activities for Cisco HCS, refer to the *Cisco Hosted Collaboration Solution Maintain and Operate Guide*: http://www.cisco.com/en/US/partner/products/ps11363/prod_maintenance_guides_list.html.

Regarding backup and restore activities for capacity planning, we recommend that you map Virtual Machines (VMs) to physical LUNs in the storage system.

A service provider should map the applications provisioned on the Cisco HCS system with the corresponding SAN RAID group and LUN location. This mapping should help identify high IOPS risk areas and allow for IOPS load balancing across the system during backup and restore activities.

You can use the Platform Manager to assist with these activities by scripting a sequence of actions.