Capacity Planning Preparation

Getting Started

The first thing you must do to get started with a capacity management plan is to establish a baseline – answer the question: "What is my capacity utilization today?" To answer this question, you must first determine the busiest, recurring period within a reasonable timeframe. For many customers, there is usually a 1-hour period of each day that is typically the busiest. Moreover, there can be busier days of the week (for example Monday vs. Wednesday); busier days of the month (last business day of the month) or busier weeks of the year (for example, the first week in January for insurance companies, or for the United States Internal Revenue Service (IRS), the first two week of April). These traditionally busy hours, days, or weeks represent the most taxing period on the deployment; these are the periods during which a capacity utilization calculation is best because you always want to ensure that your deployment can handle the worst.

The steps to getting started are:

1. Set up basic sampling (daily).

    Sample the performance counter values: CPU, Memory, Disk, Network, Call and Unified Communications (UC) Application Traffic

2. Determine the busy period.

    Identify the recurring busy period – worst case scenario – by:

    - Per Component
    - Solution Wide

3. Establish a baseline of utilization for the target period.

    - Determine hardware capacity utilization
    - Identify components with high capacity utilization

4. Generate a recurring collection plan.

    Devise a plan that is repeatable – such as automated – that can be done on a weekly basis whereby samples are obtained during the busiest hour of the week.

After you establish a baseline and identify a busy hour, daily sampling is no longer necessary; you must sample only during the busy hour on a weekly basis. However, if regular reporting shows that the busy hour may have changed, then you must complete daily sampling again so that you can identify the new busy hour. After you identify the new busy hour, weekly sampling during the busy hour can resume.

Busy Hour Determination

The busy hour is defined as the peak one-hour period during the day when the largest volume of subscriber activity is handled by a platform in terms of CPU, memory utilization, or storage input/output operations per second (IOPS). The busy hour can be broken down into various components provided below. You should collect at least one week of field data to determine trends in subscriber behavior. To assist in the analysis, remove outlying data such as weekends when the utilization is very low, or holidays when utilization is atypically high.

To determine the busy hour, start continuous sampling to cover at minimum a full week, 24 hours a day:

- ESXi Hosts— Look for the peak CPU utilization and peak memory utilization within a 24-hour period. Correlate the CPU and memory utilization busy hour with the applications that run on the ESXi Host. Create a graph to visualize the trend and mark the CPU and memory utilization on the chart.

- UC Applications — Look for the peak CPU utilization and peak memory utilization within a 24-hour period. Correlate the CPU and memory utilization busy hour with the traffic profile (most importantly the call attempts, Messaging, Voicemail). Create a graph to visualize the trend and mark the CPU and memory utilization on the chart.

- Management Applications (e.g. Hosted Collaboration Mediation-Fulfillment, Cisco Unified Communications Domain Manager, Cisco Contact Center Domain Manager, Prime Collaboration Assurance)— Look for the peak CPU utilization and peak memory utilization within a 24-hour period. Correlate the CPU and memory utilization busy hour with the traffic profile (most importantly the provisioning profile). Create a graph to visualize the trend and mark the CPU and memory utilization on the chart.

- Shared Storage (SAN or NAS)—Look for the peak IOPS within a 24-hour period based on delay criteria. Correlate the CPU utilization busy hour with the traffic profile (most importantly the call attempts). Create a graph and set the IOPS utilization on the chart. This point determines the node busy hour for IOPS. Consider maintenance, backup, and upgrade activities in the busy hour determination (in lightly loaded data centers, the maintenance period might show a perceived busy hour due to the amount of IOPS required.) You should also consider utilization spikes during database optimization, which is unrelated to traffic.

You should also monitor CPU usage for the Shared Storage to make sure that the system is not overloaded. This may require tools from your Shared Storage equipment provider, or a 3rd party monitoring tool.

System Usage Patterns

In a typical Cisco HCS deployment, the Virtual Machines (VMs), if correctly engineered, will not encounter issues with the CPU, memory, or input/output operations per second (IOPS) under normal daytime operations. However, tasks such as nightly backup add an extra load to the IOPS at the chassis level (independent of the type of Shared Storage deployed).

You must schedule these tasks to optimize the backup time and avoid IOPS overruns. You can simultaneously back up a maximum of two VMs per blade. You must determine the specific scheduling order by mapping the customer logical unit number (LUN) to the physical blade. After you complete the scheduling order, you can enter the order into the Platform Manager tool to automate the scheduling.

Capacity Planning Process

Change an existing Unified Communications (UC) deployment in small steps. Then analyze the impact of each step with a well-established, repeatable process. This process includes the following phases (steps):

1. Sample Phase: Start data sampling at the same time for the same interval for each change made.
2. Collect and Categorize Phase: Collect the samples and distribute to appropriate buckets.
3. Analysis Phase: Check application resource boundaries – has any component exceeded utilization limits? Determine best fit for new deployment requirements. Estimate solution level capacity utilization for new requirements.
4. Change Phase: Implement changes to solution based on analysis and estimate of impact.

We recommend that you repeat the process more than once to ensure that results are consistent under the same conditions.