Scalability and sizing

This section discusses MediaSense scalability and sizing.

Performance

The supported capacity for MediaSense is a function of the hardware profile that the system selects at startup time. The hardware profile depends on which VM template the node is deployed on, and the VM template depends partially on what type of hardware you are deploying. (See "Virtual machine configuration" for a full description of each template.) The "Hardware profiles" section below shows the actual capacity when using each type of VM template.

For example, for each "7 vCPU" template node (the standard for large production deployments) the MediaSense server supports up to 400 media streams simultaneously (200 calls) at a sustained busy hour call arrival rate of two calls per second on up to 12 terabytes of disk space. The 400 represents all streams used for recording, live monitoring, playback, .mp4 or .wav conversion, and HTTP download; all of which may occur in any combination. Conversion and download are not strictly speaking streaming activities, but they do use system resources in a similar way and are considered to have equal weight. Playback of a video track takes 9 times more resources than playback of an audio track. As a result, each uploaded video playback (one video track + one audio track) has the weight of 10 audio tracks, leading to a maximum capacity of 40 simultaneous video playbacks per node.

In determining how many streams are in use at any given time, you need to predict the number of onsets for each activity per unit time as well as their durations. Recording, live monitoring, and playback have a duration that is equal to the length of the recording. Video playbacks, if configured to play once only, have a duration equal to the length of the video. Video playbacks for hold purposes must be estimated to last as long as each video caller typically remains on hold. The .mp4 conversions, .wav conversions, and HTTP download durations are estimated at about 5 seconds per minute of recording.

To determine the number of servers required, evaluate

- the number simultaneous audio streams needed plus 10 times the number of videos being played, divided by the number of audio-weight media streams supported by each node
- the number of busy hour call arrivals divided by the maximum call arrival rate for each node
- the space required for retained recording sessions divided by the maximum media storage for each node.

The number of servers required is equal to the largest of the above three evaluations (rounded up).

Video playback for VoH, ViQ, and Video Messaging is further limited on 2\- and 4-vCPU virtual hardware and depends on the type of physical hardware being used. See the Hardware profiles section for details.

Another factor that significantly impacts performance is the number of MediaSense API requests in progress. This is limited to 15 at a time for 7-vCPU systems, with the capability to queue up to 10 more (the figures are reduced for smaller systems). These numbers are per node, but they can be doubled for MediaSense clusters that contain both a primary and a secondary node. For more information, see "System resiliency and overload throttling".

The media output and conversion operations (monitoring, playback, convert to MP4 or WAV, and HTTP download) are entirely under client control. The client enforces its own limits in these areas. The remaining operations (call recording and uploaded media file playback) are not under client control. The deployment can be sized so that the overall recording and video playback load will not exceed a desired maximum number cluster-wide (leaving room for an enforceable number of monitoring, playback, and HTTP download operations). The recording and video playback load is balanced across all servers. (Perfect balance will not always be achieved, but each server has enough room to accommodate most disparities.)

Hardware profiles

When MediaSense nodes are installed, they adjust their capacity expectations according to the hardware resources they discover from the underlying virtual machine. When the server is installed using one of the Cisco-provided OVA templates, the correct amount of CPU and memory are automatically provisioned and a matching hardware profile will be selected. The hardware profile determines:

- the number of audio-equivalent calls supported,
- the number of concurrent API requests supported,
- the maximum call arrival rate supported,
- the maximum number of nodes supported in the cluster,
- the maximum amount of media storage available,
- the cap on number of video playbacks supported, and
- a number of other internal parameters

as a function of the number of vCPUs, CPU speed, and amount of memory provisioned.

If an incorrect OVA template is used, or if the virtual machine's configuration is changed after the OVA template is applied such that the virtual machine does not exactly match one of the existing hardware profiles, the server is considered to be unsupported and the capacities in the "Unsupported" category are used.

For more information, see the Hardware Profile table at http://docwiki.cisco.com/wiki/Virtualization_for_Cisco_MediaSense.

Maximum session duration

MediaSense can record calls that are up to eight hours in duration. Beyond that duration, some sessions may end up being closed with an error status and HTTP download and .mp4 or .wav conversion functions may not succeed.

Storage

The amount of storage space required depends on a number of factors, such as the mix of codecs in use, the number of calls, the call arrival rate, duration, and duty cycle; and the retention period desired. Since most of these parameters are very difficult to estimate, the focus is on only the number of recording session hours and the retention period so that the amount of space required to retain h hours of recordings for d days can be accurately calculated.

Here is the formula:

Write Rate (W) = B * P * U, in hours of storage per hour of elapsed time

where

- B is the codec bit rate (in MB/hour for two streams)
- P is the number of phones
- U is the average usage ratio of each phone (in hours per day)

Retention (R) in hours = S * 1024 / W

where S is the total storage available across all servers (in GB).

Calls using the g.711 or g.722 codecs require about 1MB per minute of dual-stream recording. The space requirements for calls using the g.729 codec (that uses a variable rate compression) are not predictable, but generally speaking require about one eighth the space needed by g.711 calls (about 128 kilobytes per minute of dual-stream recording). H.264 video is even less predictable.

Therefore, assuming the use of g.711 or g.722 codecs, we have a rate of 1MB per minute or 60MB per hour. With maximum direct-attached storage of 4TB of disk space, we can store about 70,000 hours of dual-stream recordings. With 100 phones that are actively used 80% of the time (24 hours a day), then recording occurs at a rate of 80 hours per hour of elapsed time. This uses 70,000 hours in 875 hours of elapsed time (or a little over 36 days), after which the oldest calls need to be pruned. Therefore, your retention period is 36 days.

If all of the same parameters applied but phones are only active 12 hours per day, the retention period would then be 72 days.

The above examples are based on having 4TB of storage space available. If you deploy five MediaSense servers, each with its maximum SAN storage allocation, then there are 60TB of storage space available.

The number of files converted into .mp4 format using the deprecated convertSession API is another factor to consider when calculating storage space. If you expect to be converting a significant number of recorded sessions to .mp4 and leaving them on the server, then you must increase the Write Rate (W) to account for it.

In lab trials, files in .mp4 format averaged about 18 MB/hour for dual-channel audio, and about 180 MB/hour for audio+video. (Note that the .mp4 files use AAC variable rate encoding, so the actual space used may vary considerably.) If you convert and retain an average of 50% of your recorded sessions for example, then you must increase the Write Rate by 50% times the .mp4 bit rate in MB/hour, which reduces the retention period.

Therefore the formula to calculate the Write Rate becomes:

Write Rate (W) = (B * P * U) * (1 + K * M)

where

- K is the retained .mp4 average bit rate (in MB/hour)
- M is the proportion of recorded session hours which were converted to .mp4

Clients are now encouraged to fetch .mp4 and .wav files using the wavUrl and mp4url links directly, rather than use the convertSession API. Using these links, MediaSense performs the conversions on demand, resulting in a single-step download procedure. The converted files are retained for a period of time in case a second request is received, but then they are cleaned up automatically. Therefore, you do not need to consider these files in your disk space calculations.

There is also an absolute maximum number of recordings that MediaSense can retain, no matter how much disk space is provisioned or the length of the recordings. That maximum depends on the number of tags, tracks, participants, and other metadata elements per recording; but it is generally about 16 million recording sessions.

CUBE capacity

A Cisco 3945E ISR G2 router, when running as a border element and supporting simple call flows, has a capacity of about 1000 simultaneous calls (if equipped with at least 2 GB—preferably 4 GB of memory). In many circumstances, with multiple call movements, the capacity will be lower—in the range of 800 calls (due to the additional signaling overhead). In addition, the capacity will further be reduced when other ISR G2 functions (such as QoS, SNMP polling, or T1 based routing) are enabled.

Some customers will need to deploy multiple ISR G2 routers in order to handle the required call capacity. A single MediaSense cluster can handle recordings from any number of ISR G2 routers.

Network bandwidth provisioning

For Call Recording

If Call Admission Control (CAC) is enabled, Unified Communications Manager automatically estimates whether there is enough available bandwidth between the forking device and the recording server so that media quality for either the current recording or for any other media channel along that path is not impacted. If sufficient bandwidth does not appear to be available, then Unified Communications Manager does not record the call; however, the call itself does not get dropped. There is also no alarm raised in this scenario. The only way to determine why a call did not get recorded in this situation is to examine its logs and CDR records.

It is important to provision enough bandwidth so that this does not happen. In calculating the requirements, the Unified Communications Manager administrator must include enough bandwidth for 2 two-way media streams, even though the reverse direction of each stream is not actually being used.

Bandwidth requirements also depend on the codecs in use and, in the case of video, on the frame rate, resolution, and dimensions of the image.

For Video Playback

Media connection negotiation is still bi-directional for video playback (even though MediaSense only sends data and does not receive it). This is an important consideration since the use of bi-directional media implies that you must provision double the bandwidth than what you might have otherwise expected.

Impact on Unified Communications Manager sizing

MediaSense does not connect to any CTI engines, so the CTI scalability of Unified Communications Manager is not impacted. However, when MediaSense uses Cisco IP phone built-in-bridge recording, the Unified Communications Manager BHCA increases by 2 additional calls for each concurrent recording session.

For example, if the device busy hour call rate is six (6) without recording, then the BHCA with automatic recording enabled would be 18. To determine device BHCA with recording enabled, use this calculation:

(Normal BHCA rate + (2 * Normal BHCA rate)

For more information, see "Cisco Unified CM Silent Monitoring & Recording Overview.ppt" under SIP Trunk documents at http://developer.cisco.com/web/sip/docs.