Planning for Coresidency

Sizing for Coresidency

When planning a coresident deployment, consider four areas: CPU, RAM, storage, and network.

For details on virtual to physical sizing rules in a coresidency context, see http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-sizing.html.

The sizing rules refer to the "Tested Reference Configuration" hardware support approach, described at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-hardware.html.

CPU

A BE6000 deployment must have a one-to-one allocation of vCPUs to physical cores. For example, if you have a host with 16 physical cores, you can deploy any combination of virtual machines with a combined requirement of no more than 16 vCPUs. There is a special case for ESXi 5.0 and 5.1 if one or more of the virtual machines are running Cisco Unity Connection. In this case, the total vCPUs must be one less than the total physical cores. For servers using ESXi5.5, you may configure Cisco Unity Connection virtual machines to use High Latency Sensitivity. In this case, and if at least one other virtual machine is configured with Normal Latency Sensitivity, the total vCPU can be the same as total physical cores.

Because the number of vCPUs must not exceed the number of physical cores, you do not have to configure CPU reservations or limits. You can never overcommit physical cores.

Note Some processors support Hyperthreading, which allows a Hypervisor to see a physical core as two logical processors. Logical processors must not be used for coresidency planning.

For more details, see the "No Hardware Oversubscription" section at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-sizing.html.

RAM

To ensure that there is no over subscription of memory, set up all coresident virtual machines with a physical memory reservation equivalent to the vRAM setting. For example, if a virtual machine is configured with 4 GB of vRAM, assign a physical memory reservation of 4-GB.

The virtualization software on a BE6000 or BE7000 appliance also requires additional physical memory to host and run the virtual machines. For the amount of memory required, see http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-sizing.html#generalrulememory.

> **Note** The overhead reservation by ESXi hosts is not applicable to BE6000S release 11.0 and older. As BE6000S is a special configuration with deployment model restrictions, in release 11.0 and older, it does not ship with, or require extra memory for ESXi as described for other Business Edition models.

For example, if the BE6000 host running ESXi 5.1has 32 GB of physical RAM, 2 GB is available for ESXi and 30 GB is available to virtual machines. For more information, see "Understanding Memory Overhead" at http://pubs.vmware.com/vsphere-51/index.jsp?topic=%2Fcom.vmware.vsphere.resmgmt.doc%2FGUID-98BD5A8A-260A-494F-BAAE-74781F5C4B87.html.

Disk Storage and Performance

The server Direct Attached Storage (DAS) must supply the combined disk space and IOPS (Input Output Operations per Second) capacity for the virtual machines that run on the host, while maintaining a minimum performance level.

It is unlikely that the latency requirements of the Business Edition applications will limit third-party applications. However, you must understand the latency and load requirements of the non-Business Edition applications before installation.

The DAS and RAID are configured during manufacturing, and no field changes are allowed to the BE6000 or BE7000 Unified Computing System Tested Reference Configuration (TRC). The BE6000 TRC is designed to meet the storage requirements of all BE6000 collaboration applications. The BE7000 TRC is designed to meet the storage requirements of higher capacity points of these applications as described at www.cisco.com go virtualized-collaboration.

To ensure the reliable operation of Business Edition applications, disk latency must not exceed 20 ms. We recommend that deployments that include non-Business Edition applications be verified such that the kernel command latency does not exceed 3 ms and the physical device command latency does not exceed 20 ms under any conditions. For more details, see "Sizing Shared Storage" at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-hardware.html#storage.

For example, when you test a TRC with DAS, all UC applications that are designed to run on the TRC are loaded with full traffic and software upgrades are run simultaneously on all virtual machines. This generates the highest IOPS load possible on the host, and if this test passes, it is likely the DAS array can handle the I/O load of the specific set of virtual machines.

To summarize the requirements, the disk subsystem must supply the disk space that is required for all the applications and support the aggregate IOPS load that the virtual machines generate, while not impacting the latency requirements of the UC applications. Otherwise, some virtual machines must be removed.

Network

The aggregate networking load of the coresident virtual machines must not exceed the capacity of the physical networking interfaces. Generally, the I/O capacity of the physical network resources on any modern server is more than adequate to meet the needs of the virtual machines that are being hosted. For UC applications, see the "QoS Design Considerations" at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-qos-designs-considerations.html.

Understand the networking requirements of the virtual machines that are deployed on the host and how to set up the host networking hardware to meet those needs. If we determine that application performance problems are due to networking congestion within the host, then some VMs must be moved off the host.

Summary of Coresidency

Under certain conditions, we allow coresident deployments with a mix of Business Edition and non-Business Edition applications. The requirements that are listed in this document provide the basis for a successful coresident deployment, but due to the unpredictable behavior of non-Business Edition applications and the impracticality of testing every possible combination of applications that could be deployed, we cannot guarantee that following these guidelines alone are sufficient.

A key principal of virtualization is based on sharing hardware resources among multiple virtual machines. When deploying only Business Edition applications on a host server, we can guarantee levels of performance, because the applications are thoroughly tested and their behavior understood. Deploying coresident third-party applications introduces a degree of unpredictability that can be mitigated by following general principals of virtualization and the specific requirements in this document.

Ultimately, we cannot guarantee that by following these guidelines alone, the virtual machines on a host will never be starved of resources. However, if this does occur, the only recourse is to remove some of the virtual machines from the host so that the load is reduced. This can be done by moving some of the virtual machines to another host, or by moving all the virtual machines to a more powerful host.