Sizing and Scaling Considerations

Sizing Considerations

Some Cisco UCS Central customers have scaled their UCS environments from dozens to hundreds of domains and from hundreds to thousands of servers. The scaling metric to determine your needs is not the number of registered UCS domains, but rather the number of blades and servers. Cisco UCS Central has been tested to support up to 10,000 servers. There is no software limit.

If you have a large environment, analyze the performance of the Cisco UCS Central virtual machines (VMs) to determine if you need to increase the CPU and memory resources.

The *Cisco UCS Central Installation Guide* defines the minimum requirements for running Cisco UCS Central. The deployment of the OVA file:

- Creates 2 x 40-GB disks
- Allocates 4 x vCPUs
- Allocates 12-GB RAM

For small to medium-size environments, those settings provides sufficient performance. However, for large environments we recommend increasing those resources by a factor of two. (Cisco IT runs at least 4 vCPUs and 24-GB RAM for their instance of Cisco UCS Central, which exceeds 200 UCS domains and 6,000 servers.)

The best practice is to:

1. Monitor the performance metrics of your Cisco UCS Central VM in your new or growing infrastructure.
2. Adjust CPU and RAM resources accordingly.
3. Verify vCPU and memory usage if performance is noticeably inadequate.
4. Make needed changes with your virtualization administrator.

Scaling Global Service Profiles and Global Service Profile Templates

There is no limit to the number of global service profiles that you can attach to a global service template.

Leveraging templates to make widespread changes is beneficial for simplified management, but it is not best for all environments. Consider the ramifications of a single disruptive change to the environment. Determine how many resources you can safely attach to an updating global service profile template. One mistake could cause a significant outage. Consider reducing the potential for damage. For example, you could decide that a ratio of 100 global service profiles attached to each service profile template is the maximum limit.

Another important consideration is the number of policies. For example, if you have ten global service profile templates, with 100 global service profiles attached, that represents 1,000 blades or servers. If you had all ten templates accessing the same disruptive policies, you could have a single policy that could potentially disrupt 1,000 blades or servers. To reduce the possibilities for failure, it is best to scale across multiple service profile templates, and multiple policies. It is acceptable to have more than one policy that contains the same settings, but is used by different templates.

If you have too many global service profiles attached to too few templates, you can clone the global service profile template. Then unbind some of the profiles from the original template and bind them to the clone. The system is flexible in design and operations. You can balance and scale accordingly.

Note We recommend testing this scenario before deploying it in a production environment.

Although this guide discusses reducing the number of global service profile templates for ease of management, do not attach too many global service profiles to a template. Also, do not use a single policy for too many global service profiletemplates and profiles. This sounds contradictory; however, consider balancing reduction of templates for improved OPEX, with the widespread damage that could be caused by having too many service profiles accessing a template, or policy.