# Cisco CloudCenter Solution Use Case: Capacity Augmentation

## Overview

Enterprise IT organizations find it difficult, or impossible, to accurately predict hardware capacity requirements. Traditionally, IT has overallocated capacity based on a just-in-case model to reduce the risk of disrupting critical business services. The result has been underutilized infrastructure.

Many enterprise applications exhibit usage patterns that vary significantly over time. For example, application deployments in development and test environments are designed to be temporary, but they account for four to five times as many virtual machines deployed for long-running applications in production environments.

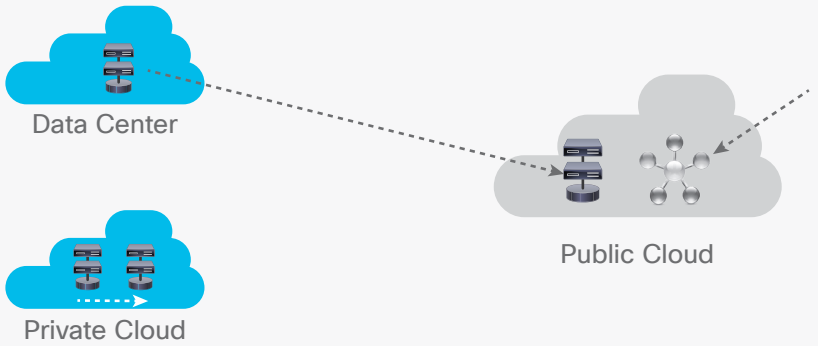
Web and mobile applications show dramatic swings in workload volume during the day. And data-intensive applications increase resource consumption for short amounts of time during periods of heavy processing like during end of month and end of quarter accounting.

Cloud computing provides a solution. The cloud is instantly on, and has a highly scalable architecture that helps enable IT to deliver capacity when and where it is needed, eliminating the risks associated with underallocation and the costs associated with overallocation.

As Figure 1 illustrates, IT teams can use cloud technology to address the capacity challenge in two ways:

- **Use in-house resources more fully.** Cloud management technology deployed in the data center allows applications to be deployed on demand and then deleted, or to be scaled out and back horizontally as the workload expands and contracts. Deallocated resources returned to a resource pool are available for other applications. The result is higher infrastructure utilization.
- **Offload demand to outside resources.** Public cloud services that provide capacity on a pay-as-you-go basis supplement in-house capacity by enabling bursting of highly scalable workloads.

Figure 1. Cloud enables capacity optimization




## The cloud capacity augmentation challenge

Cloud technology provides an effective means of augmenting data center capacity because of the ease and speed at which resources can be provisioned, scaled, and released when no longer needed. However, the complexity of modern data centers make a hybrid IT strategy difficult to implement. To augment data center capacity effectively, IT has to achieve three major objectives:

- **Application focus:** IT must tie deployment automation and scaling policies to the needs of the application. It is impractical to rewrite applications to accommodate each environment in which they must run. Instead, the needs of the application should direct the provisioning of infrastructure.
- **Policy-based automation:** Automation must be based on a high level of standardization and guided by predefined rules and policies. Performance metrics and predefined triggers should guide horizontal scaling or cross-cloud bursting. Policies should also automate resource decommissioning when jobs are complete or applications are no longer in use.
- **Cross-environment automation:** Automation that scales or bursts across multiple environments must work properly across all those environments. However, infrastructure provisioning automation and application stack deployment automation are typically performed using separate tools that require different processes and skill sets. Moreover, each automation script is typically hard-wired to a single environment. The result is a complex mix of automation artifacts that must be version controlled and separately maintained.

What is needed is a solution that takes full advantage of cloud agility and scale and supports automation based on the needs of the application and not the infrastructure. That automation must work transparently across different data center and cloud environments.

# The Answer: The Cisco CloudCenter Solution

The Cisco CloudCenter™ solution is a hybrid cloud management platform that provisions infrastructure through cloud APIs and deploys and orchestrates fully configured application stacks in more than 19 data center, private cloud, and public cloud environments.

The Cisco CloudCenter solution automates the deployment of workloads that range in complexity from a single virtual machine or operating system image to complex multitier and multiservice applications with 50 or more components. IT can use the solution to implement various capacity augmentation scenarios, including:

- **Temporary high-performance computing:** High-Performance Computing (HPC) jobs such as Blender workloads, and big data analytics jobs such as Hadoop workloads are well suited for the elasticity of the public cloud. With the Cisco CloudCenter solution, IT can deploy both clusters and applications to various cloud environments based on business needs. The solution directs infrastructure resources to meet the needs of the applications based on various performance metrics and policies.
- I**n-place horizontal scaling:** Scaling applications by deploying additional instances as needed and then deleting those instances when they are no longer needed uses data center infrastructure more fully and reduces costs. The Cisco CloudCenter solution provides an effective way to horizontally scale applications, even for applications that weren't designed to scale. In addition, it functions in environments that don't offer native load-balancing services. The solution also allows the use of a wide range of metrics to activate predetermined triggers that guide scale-out and scale-back actions.
- **Cross-cloud bursting:** IT can respond to workload spikes that deplete data center resources by offloading excess demand to the public cloud. Administrators can easily deploy any application to any supported cloud with a single click. In addition, IT can schedule deployments based on expected usage spikes, or can take advantage of policy-based automation that detects when thresholds are exceeded and deploys additional application instances in another cloud.

With Cisco CloudCenter autoscaling and cross-cloud bursting capabilities, applications can expand temporarily in place or move to another cloud, all based on predefined policies. Autoscaling and bursting help IT organizations avoid overprovisioning of resources and unnecessarily locking up infrastructure capacity in anticipation of workload peaks. The IT organization can centrally establish and manage the policies that guide these capabilities, helping ensure consistent implementation across the enterprise.

# Advanced features

In addition to scaling and bursting, the unique Cisco CloudCenter solution includes a wide range of capabilities that meet the complex needs of enterprise IT organizations.

## Application profile and orchestrator combination

The patented Cisco CloudCenter technology plays a crucial role in enabling agile and efficient capacity augmentation. As Figure 2 illustrates, the Cisco CloudCenter solution combines a cloud-independent application profile with a cloud-specific orchestrator.

The application profile is a deployable blueprint that combines infrastructure and application stack automation instructions. The orchestrator abstracts the infrastructure API and application services that are unique to the data center, private cloud, or public cloud environment and interprets the needs of the application for that environment.

The application profile can be deployed in any supported environment for scale-out or bursting. The orchestrator can deploy additional instances in any environment and then remove them automatically when they are no longer needed. Users can add instances to meet peak workload requirements or use policies to automate deployment in scale-out or bursting scenarios.

## High-performance computing applications

With the Cisco CloudCenter solution, IT can easily configure and run HPC applications such as Blender. The process is easy and straightforward:

- Load the application binary files and the job's data files in an artifact repository.
- Model an application profile in the graphical topology modeler by dragging application services from the service library.
- Configure the application by pointing to the appropriate repository of binary files and data files, and set the desired scaling properties.

You can also use the Cisco CloudCenter solution to automate deployment of a Hadoop cluster, and, as Figure 3 shows, a Hadoop MapReduce application.

Any authorized user can deploy an application profile to any supported data center, private cloud, or public cloud. The cloud-resident orchestrator calls APIs to provision the necessary resources and then deploys the profile and related data. Deployment can be scheduled based on expected workload or resource availability, or policies can guide automated scaling based on properties set in the application profile.

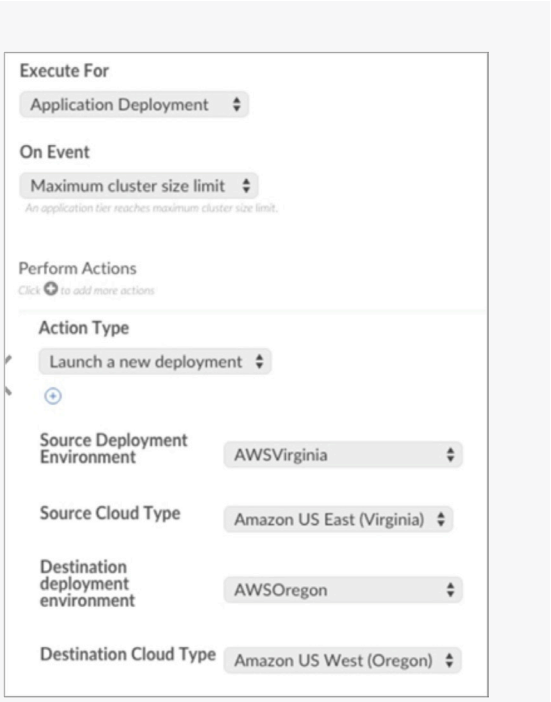
Figure 2. Cisco CloudCenter technology provides a strong foundation for capacity augmentation


Figure 3. Configuring a hadoop MapReduce application profile

## Runtime policies

Runtime policies guide scaling, bursting, and aging based on prespecified rules.

Figure 4. Setting bursting policy




- **Scaling policies:** IT can set scaling policies with various polling intervals that deploy additional instances in a cluster and are triggered based on performance metrics such as CPU and memory utilization. Scaling automation includes triggers that scale back and delete unused instances, also based on performance metrics.
- **Bursting policies:** IT can create action policies that cause a fresh deployment to burst to a new environment when the current environment reaches the maximum cluster size as specified by the cluster policy. The administrator sets the Action Type to "Launch a new deployment" and specifies the source and destination deployment environments (Figure 4).
- **Aging policies:** Administrators can apply aging policies to delete instances based on time criteria.

# Benchmark

Workload footprints vary widely, and the configuration of cloud resources dramatically affects price and performance. In some cases, cost is the most important factor. In other cases, speed and performance are more important.

With the Cisco CloudCenter benchmark capability, IT can deploy variations of a single application to compare the resulting price-to-performance metrics and determine the optimum configuration.

Figure 5 shows actual Cisco CloudCenter benchmark results for the Blender application that reveal the price-to-performance characteristics for different cloud machine instance sizes. As can be seen, increasing the size from XL to 2XL nearly doubles the cost and reduces job time by 13 minutes. But doubling the cost again saves only an additional 2 minutes. Doubling the cost again saves 2 more minutes. With this real price-to-performance data, IT can optimize the price-to-performance trade-off.

Figure 5. Cost and time trade-offs for a rendering job


| Instance | Size | Cost (US$) | Time (min) |
|---|---|---|---|
| ◆ | 8XL | $2.4 | 22 |
| ◆ | 4XL | $1.2 | 24 |
| ◆ | 2XL | $0.6 | 26 |
| ◆ | XL | $0.335 | 33 |

# Real-world examples

Cisco customers have used the power of the Cisco CloudCenter platform in a range of capacity augmentation scenarios.

## Automobile racing engineering company

Race engineers have as little as a week to optimize race-day vehicle configurations. During that time, they have to process dozens of gigabytes of data collected during the previous race and optimize the upcoming race-day scenario with respect to tires, brakes, suspension, engine, racetrack, and expected weather conditions. They needed to dramatically reduce the simulation time to enable analysis of multiple what-if scenarios.

Prior to the Cisco CloudCenter solution, the engineers ran simulations on a single eight-core workstation. With the Cisco CloudCenter solution, they parallelized and ran simulations across multiple machines in the public cloud, deploying and scaling to 500 virtual CPU (vCPU) instances. In doing so, they increased processing power 300 times without changing application code and reduced simulation time from 14 days to just 5 hours at a cost of only US$62 per run.

## Electronics design software provider

The semiconductor design process is complex and computation intensive. This electronics design software provider realized that offering a Software-as-a-Service (SaaS) version of its software would make the software more affordable and more easily accessible, presenting a new revenue opportunity for the company. The company could create instances of the software just for the duration of a customer's project. To help ensure a viable SaaS delivery model, the provider had to be able to quickly deploy the software in a variety of data center and cloud environments to meet the needs of its customers.

With Cisco CloudCenter automated provisioning, the company was able to reduce the provisioning time of each cloud-based design environment from weeks to only 30 minutes. In addition, the software can use IBM Platform Load-Sharing Facility (LSF) cluster technology to increase performance. It can also use hardware security module technology to help ensure protection of customers' sensitive design data.

## Media streaming content provider

Media streaming is characterized by extreme fluctuations in streaming traffic. Fortunately, the traffic variations are often predictable. For example, the content provider knows ahead of time how many subscribers have paid to stream a popular sporting event. Consequently, the provider can schedule deployment of enough servers to accommodate event traffic.

With a fully automated Cisco CloudCenter deployment, the company can test various server configurations to determine the optimum balance of service quality and cost for the event. Then, before the event, the company can quickly deploy hundreds of optimized streaming servers. After the event, the servers can be decommissioned and returned to the cloud's resource pool.



C22-737218-01 07/17