CliQr®

CliQr is now
part of Cisco.

Capacity Augmentation

Capacity Augmentation

Enterprise IT organizations find it difficult if not impossible to accurately predict hardware capacity requirements. Traditionally IT has overallocated capacity based on a just-in-case model to minimize the risk of disrupting critical business services. The result has been underutilized infrastructure.

Many enterprise applications exhibit usage patterns that vary significantly over time. While application deployments in development and test are temporary in nature, they account for four to five times the number of deployments of long-running applications in production. Web and mobile applications show dramatic swings in workload volume during the day. And data-intensive applications increase resource consumption for short periods of time during heavy processing.

Cloud computing provides a solution. The cloud's instant-on, highly scalable architecture enables IT to deliver capacity when and where it is needed, eliminating the risks associated with underallocation and the cost associated with overallocation.

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

1. **Utilize in-house resources more fully.** Cloud technology deployed in the datacenter allows applications to scale out and back horizontally as workload expands and contracts. Deallocated resources returned to a resource pool are available for other applications. The result is higher infrastructure utilization.
2. **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.

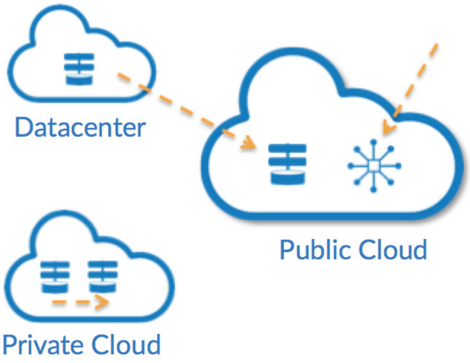


Figure 1. Cloud enables capacity optimization

The Cloud Capacity Augmentation Challenge

Cloud technology provides an effective means of augmenting datacenter 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 datacenters make a hybrid cloud strategy difficult to implement. To augment datacenter capacity effectively, IT has to achieve three major objectives:

1. **Application focus.** IT must tie deployment automation and scaling policies to the needs of the application. It's 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.

2. **Policy-driven 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 complete or applications are no longer in use.

3. **Cross-environment automation.** Automation that scales or bursts across multiple environments must work properly across all those environments. The problem is that infrastructure provisioning automation and application stack deployment automation are typically done with 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's needed is a solution that takes full advantage of cloud agility and scale, and drives automation based on the needs of the application and not the infrastructure. That automation must work seamlessly across different datacenter and cloud environments.

The Answer: CliQr CloudCenter

CloudCenter is an application-centric cloud management platform that provisions infrastructure via cloud application programming interfaces (APIs), and deploys and orchestrates fully configured application stacks in more than 19 datacenter, private, and public cloud environments.

CloudCenter automates the deployment of workloads that range in complexity from a single virtual machine (VM) or operating system (OS) image, to complex multitier, multiservice applications with 50+ components. IT can use CloudCenter to implement various capacity augmentation scenarios, including:

- **Temporary high-performance compute.** High-performance compute jobs such as Blender, and big data analytics jobs such as Hadoop are ideal candidates for the elasticity of the public cloud. With CloudCenter, IT can deploy both clusters and applications to various cloud environments based on business needs. CloudCenter directs infrastructure resources to meet the needs of the applications based on various performance metrics and policies.
- **In-place horizontal scaling.** Scaling applications by deploying additional instances as needed and then deleting those instances when no longer needed utilizes datacenter infrastructure more fully and minimizes costs. CloudCenter is an effective solution for horizontal scaling of applications, even for applications that weren't designed to scale. What's more, it functions in environments that don't offer native load-balancing services. CloudCenter 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 datacenter 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 CloudCenter's 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 resources and unnecessarily locking up infrastructure capacity in anticipation of workload peaks. The IT department can centrally establish and manage the policies that guide these capabilities, ensuring consistent implementation across the enterprise.

Note – CloudCenter can also be used to migrate applications to the cloud – See the Migrate and Manage use case.

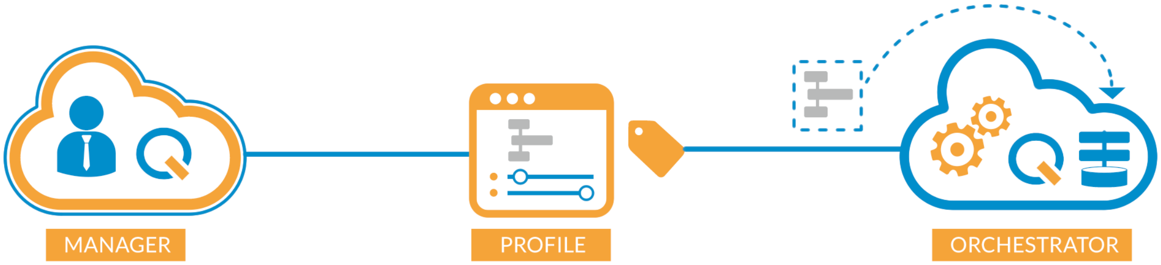
Advanced Features

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

Application Profile and Orchestrator Combination

CloudCenter's patented technology is key to enabling agile and efficient capacity augmentation. As Figure 2 illustrates, CloudCenter combines its cloud-agnostic Application Profile with its cloud-specific Orchestrator.

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



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 datacenter, private cloud, or public cloud environment.

The Application Profile can be deployed to any supported environment to scale out or burst. The Orchestrator can deploy additional instances in any environment, 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.

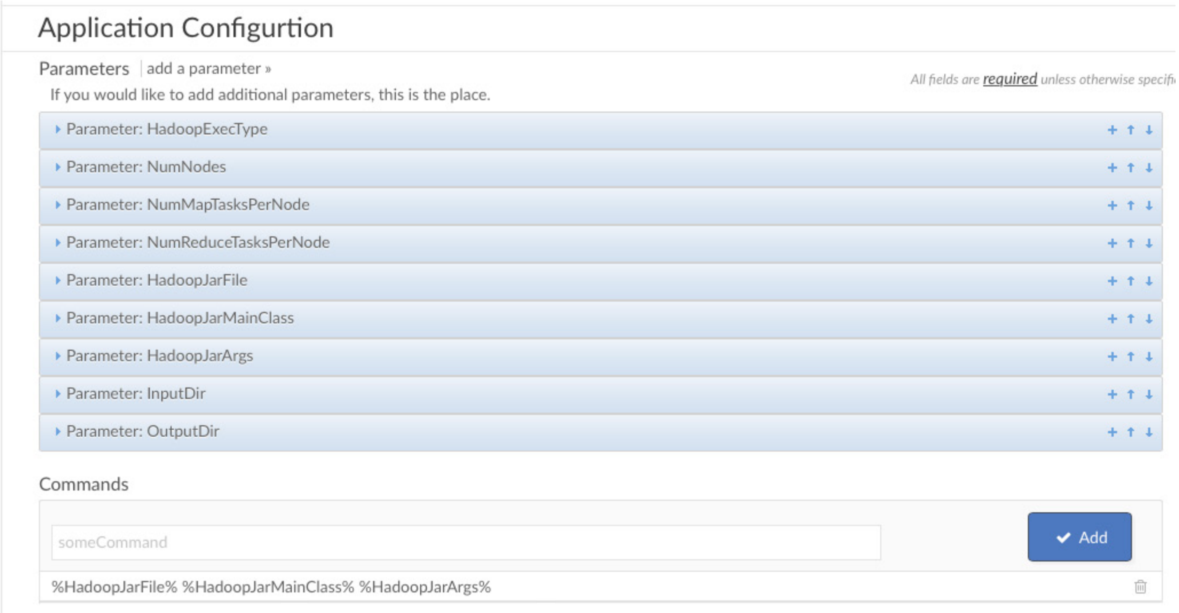
Onboard High-performance Compute Application

With CloudCenter, IT can easily onboard and run high-performance compute applications such as Blender. The process is easy and straightforward:

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

CloudCenter can also be used to automate deployment of a Hadoop cluster, and, as Figure 3 shows, a Hadoop Mapreduce application.

Figure 3. Configuring a Hadoop Mapreduce application profile



Application Configurtion

Parameters | add a parameter »

If you would like to add additional parameters, this is the place.

All fields are ***required*** *unless otherwise specifi*

- Parameter: HadoopExecType
- Parameter: NumNodes
- Parameter: NumMapTasksPerNode
- Parameter: NumReduceTasksPerNode
- Parameter: HadoopJarFile
- Parameter: HadoopJarMainClass
- Parameter: HadoopJarArgs
- Parameter: InputDir
- Parameter: OutputDir

Commands

someCommand

%HadoopJarFile% %HadoopJarMainClass% %HadoopJarArgs%

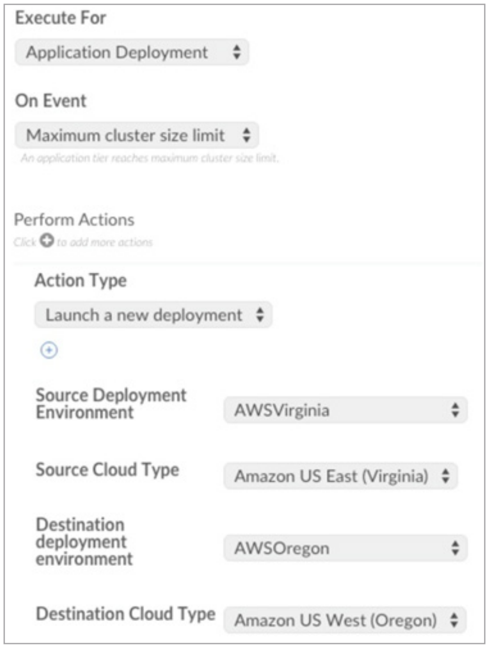
Any authorized user can deploy an application profile to any supported datacenter, private cloud, or public cloud. The cloud-resident orchestrator calls APIs to provision the necessary resources, then deploys the profile and related data. Deployment can be scheduled based on expected workload or resource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

Runtime policies

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

- **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, again 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 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. (See Figure 4.)
- **Aging policies.** Admins can apply aging policies to delete instances based on time criteria.

Figure 4. Set bursting policy



Execute For

Application Deployment

On Event

Maximum cluster size limit

An application tier reaches maximum cluster size limit.

Perform Actions

Click ⊕ to add more actions

Action Type

Launch a new deployment

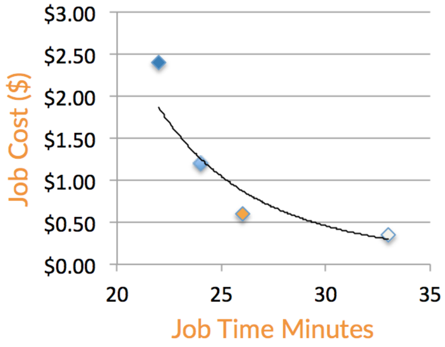
Source Deployment Environment	AWSVirginia
Source Cloud Type	Amazon US East (Virginia)
Destination deployment environment	AWSOregon
Destination Cloud Type	Amazon US West (Oregon)

Benchmark

Workload footprints vary widely and the configuration of cloud resources dramatically impacts price and performance. In some cases, cost is the most important factor. In other cases, speed and performance are more important. With the CloudCenter benchmark capability, IT can deploy variations of a single application to compare the resulting price/performance metrics and determine the optimum configuration.

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

Figure 5. Cost and time tradeoffs for a rendering job



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

Real-world Examples

CliQr customers have leveraged the power of the CloudCenter platform in a range of capacity augmentation scenarios.

Automobile racing engineering firm

Race engineers have as little as a week to optimize race-day vehicle configuration. 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. That means slashing the simulation time to enable analysis of multiple what-if scenarios.

Prior to CloudCenter, the engineers ran simulations on a single eight-core workstation. With CloudCenter, they parallelized and ran simulations across multiple machines in the public cloud, deploying and scaling to 500 1 vCPU instances. In doing so, they increased processing power 300 times without changing application code and reduced simulation time from 14 days to just five hours at a cost of only $62 per run.

Electronics design software provider

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

With CloudCenter's automated provisioning, the company was able to reduce the provisioning time of each cloud-based design environment from weeks to only 30 minutes. What's more, the software can leverage IBM Platform LSF cluster technology to maximize performance. It can also leverage hardware security module technology to 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 spinning up enough servers to accommodate event traffic.

With CloudCenter's fully automated deployment, the company is able to 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 spin up hundreds of optimized streaming servers. After the event, the servers can be decommissioned and returned to the cloud's resource pool.

CliQr Technologies
1732 North First St., Suite 100, San Jose, CA 95112
888.837.2739 • info@cliqr.com • www.cliqr.com