Managing the Real Cost of On-Demand Enterprise Cloud Services with Chargeback Models


Introduction

Enterprises today are looking to cloud computing to help them better deliver existing as well as new, innovative services on demand across network, computing, and storage resources—at reduced cost.

Cloud computing—IT resources and services that are abstracted from the underlying infrastructure, pooled, and then provided on demand and at scale in a shared, multitenant and elastic environment—is often described as a transformational model for the enterprise.

Cloud computing can be deployed within an enterprise, as a private cloud, since it is restricted in access to the enterprise’s private network. In contrast, a service provider can provide cloud services to its customers in a public cloud. An environment that transparently combines multiple clouds, private or public, is known as a hybrid cloud.

A cloud can provide IT infrastructure (for example, computer servers and storage), or infrastructure as a service (IaaS); an application runtime platform, or platform as a service (PaaS); or subscription-based software, or software as a service (SaaS). IT services that are delivered as cloud services typically have the following characteristics:

- A pay-as-you-go model with minimal or no upfront costs
- Usage-based pricing, so that customer costs are based on actual usage
- Elasticity, so that customers can dynamically consume more or less resources as and when they need it

From its ability to break down IT silos to its flexibility, efficiency, and democratization around resource allocation, resulting in agile IT service delivery, cloud computing changes the way we do business and engage with each other.
Most importantly, the way in which information and services are provided to and consumed by enterprise users in the cloud—shared, self-service, scale on demand, automated recovery, provisioning on demand, and pay per use—enables new economies of scale, along with new mechanisms for cost-effective services.

Pay for use—that is, charging customers based on their usage and consumption of a service—has long been a cornerstone of certain businesses, such as utilities, wireless phone companies, and even web collaboration services. Pay for use makes users keenly aware of the cost of doing business and consuming a resource, since the cost comes out their pockets, or, in the enterprise world, their own budgets. And with awareness of the costs comes more efficient and selective usage, thus resulting in less waste and lower costs.

In enterprise computing, pay for use has increasingly gained acceptance, as IT strives to lower costs across infrastructures, applications, and services and pushes back the IT costs to the consumer. And now, with cloud computing, pay for use has become both necessary in a shared environment and easier to implement. And, on top of it, pay for use delivers a range of benefits beyond lower costs and cost management.

**Chargeback Models for Computing**

In the financial world, chargeback refers to a reversal of a charge where the financial institution (for example, a credit card company) credits a consumer’s account. In the cloud computing world, chargeback is an IT term for recovering the cost of providing cloud computing services from the service consumers: that is, making the consumer pay for the usage.

Enterprises that are deploying private clouds will benefit from developing a chargeback model with billing based on actual resource usage, instead of resource allocation or reservation, as a fundamental component of the cloud architecture, even if intradepartmental billing is not required. A chargeback model in the cloud delivers many benefits, including the most obvious:

- Correlating utilization back to cloud consumers or corporate departments, so that usage can be charged if desired
- Providing visibility into resource utilization
- Facilitating capacity planning, forecasting, and budgeting
- Providing a method for managing IT demand
- Enabling cloud users to know their compute footprint, thus encouraging good citizenship and a perhaps a greener approach
- Encouraging the use of emerging technologies, which might be priced lower than other services as an incentive (For example, virtual machines will cost less than physical servers)
- Bringing transparency to enterprise IT, which is a pivotal step in the transformation of enterprise IT from a cost center to a business enabler
- Providing a mechanism for the enterprise IT function to justify and allocate their costs to their stakeholder business units

When designing or evaluating a chargeback service, you should create a chargeback model with the following characteristics:

- **Accurate**: assess charges for actual resource usage accurately and fully
- **Auditable**: store and retrieve detailed records on all charges to handle billing inquiries and disputes
- **Flexible**: modify easily to handle pricing variations, for example, for promotions and specials that might vary over time or by region
- **Scalable**: scale components easily to handle cloud-sized workloads

The methodology for defining and deploying a chargeback service can be applied to private or public clouds, as well as hybrid clouds. The chargeback service discussed here is for an IaaS cloud, although the same methodology can be applied to other cloud delivery models as well.
The Virtualization Conundrum

For enterprises, server virtualization, which is a crucial enabling technology for a cloud environment, has some advantages as well as challenges that affect the chargeback model.

Server virtualization allows a physical machine to be used as multiple virtual machines: a very cost-effective advantage in terms of reducing overall capital and operating expenses by deploying fewer physical servers. Each virtual machine uses a portion of the host machine’s available CPU, memory, storage, and network resources on the physical server. In type 1 (or bare metal) virtualization, a software module called the hypervisor sits in between the hardware and the virtual machines and allows each virtual machine’s operating system to think that it has its own dedicated resources. The hypervisor also manages the host’s physical resources and coordinates usage across virtual machines.

A virtual machine might be assigned resources, but the hypervisor can allow other virtual machines to utilize these resources based on the combined workload and configuration policies. For example, a virtual machine with 2 GB of RAM configured might not be actually utilizing 2 GB of RAM on the host system. Based on policies, the hypervisor can use the unused physical RAM for other virtual machines.

Although server virtualization can consolidate servers and reduce costs, it makes billing models based solely on allocation inadequate.

Server virtualization systems do allow reservation of resources, in which case the hypervisor will not allocate the reserved resources to other virtual machines. This makes the system more predictable but erodes some of the benefits of virtualization because resources can no longer be dynamically pooled and allocated based on demand.

If resources are reserved, the system can no longer use variations across workloads to dynamically balance and optimize resource usage, achieve better server consolidation, and reduce power and energy costs. Hence, a virtualization best practice is to consider reserving resources only when absolutely needed and to otherwise rely on virtualization and resource management software to monitor and optimize resource allocation and usage.

Initial reservation of resources can also lead to noncompetitive costs for cloud computing services.

To gain maximum benefits from server virtualization and for competitive service pricing, it is necessary to consider a chargeback model with billing based on actual resource usage instead of resource allocation or reservations.

Chargeback Methodology

The methodology for creating a chargeback model is straightforward. To develop a chargeback model, enterprises should:

1. Analyze and document all relevant costs.
2. Identify the billable items and, for each one, identify the smallest unit that will be available as a service to customers. A unit can be defined as a unit of compute power, a portion of the data center, or a combination of compute, network, and storage resources. This small unit becomes the “atomic unit” that drives data collection, billing, and reporting.
3. Define a pricing strategy by choosing pricing options for each billable item.
4. Identify, integrate, and deploy the tools necessary to collect billing data and to mediate the data into a billing record for an upstream system.

These steps are illustrated in Figure 1.
Costs

Chargeback involves collecting and correlating billing data records and applying pricing models to the records to generate customer bills. To be able to determine a pricing model that provides business value, it is necessary to know the direct and indirect costs of providing cloud services. For an IaaS deployment, the cost can be modeled as a fully loaded cost per server or per virtual machine.

Costs can be initial or ongoing. Initial costs, also known as capital expenditures, or CapEx, include the costs to acquire assets such as hardware and facilities. The following costs are examples of items that should be considered when determining the CapEx for the deployment:

- Facility construction or acquisition
- Power and cooling infrastructure
- Server, network, and storage hardware
- Software licenses, including operating system and application software
- Racks, cables, and installation

Ongoing costs, also known as operational expenditures, or OpEx, include all costs for keeping the business or facility running. The following costs can be considered to calculate the OpEx for the deployment:

- Payroll
- Facilities maintenance
- Hardware maintenance
- Software maintenance
- Other fees such as insurance, legal, and accounting fees
Cost of the Cloud

The capital and operational expenditures can be used to calculate the monthly costs for the cloud computing deployment.

For capital expense cost items, the cost of each item needs to be amortized over the life of the item. Typically servers have a lifespan of 3 to 5 years, while data center facilities have a life of 10 to 15 years.

Here is the formula to calculate a monthly cost, using a simple straight-line depreciation model:

\[
\text{Monthly Cost} = \frac{\text{Total Cost}}{1 - \left(1 + \frac{1}{N \times R}\right)}
\]

Where:  
- \( R \) is the monthly cost of money or inflation rate (for example, 3%/12)
- \( N \) is the life of the item in months

The monthly costs for the operational expense items can then be combined with the calculated monthly costs for the capital expense items, to determine the total monthly costs of the cloud deployment.

Figure 2 shows an example of the major monthly percentage costs for a cloud deployment.

**Figure 2. Sample Monthly Percentage Costs for a 50,000-Server Cloud**

Costs per Server

The next step is to determine the monthly or hourly cost per physical and virtual machine.

The cost per physical server can be calculated by dividing the total monthly costs by the number of servers. The cost per virtual machine can be obtained by dividing the costs per server by the target consolidation ratio of virtual machines per server. Since the actual consolidation ratio will vary based on workload and configuration policies, it is best to use a conservative ratio.
Billable Items

Billable resources are items for which customers will be charged. In a cloud deployment, these items will be part of the IT service catalog, and customers will be able to purchase these items using the cloud service portal.

Table 1 is an example of the types of billing resources for an IaaS cloud.

<table>
<thead>
<tr>
<th>Billable Resource</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual machine</td>
<td>• CPU</td>
</tr>
<tr>
<td></td>
<td>• Memory</td>
</tr>
<tr>
<td></td>
<td>• Storage capacity</td>
</tr>
<tr>
<td></td>
<td>• Disk and network I/O</td>
</tr>
<tr>
<td>Server blade</td>
<td>Options will vary by type and size of the hardware</td>
</tr>
<tr>
<td>Network services</td>
<td>• Load balancer</td>
</tr>
<tr>
<td></td>
<td>• Firewall</td>
</tr>
<tr>
<td></td>
<td>• Virtual router</td>
</tr>
<tr>
<td>Security services</td>
<td>• Isolation level</td>
</tr>
<tr>
<td></td>
<td>• Compliance level</td>
</tr>
<tr>
<td>Service-level agreements</td>
<td>• Best effort (Bronze)</td>
</tr>
<tr>
<td>(SLAs)</td>
<td>• High availability (Silver)</td>
</tr>
<tr>
<td></td>
<td>• Fault tolerant (Gold)</td>
</tr>
<tr>
<td>Data services</td>
<td>• Data encryption</td>
</tr>
<tr>
<td></td>
<td>• Data compression</td>
</tr>
<tr>
<td></td>
<td>• Backups</td>
</tr>
<tr>
<td></td>
<td>• Data availability and redundancy</td>
</tr>
<tr>
<td>WAN services</td>
<td>• VPN connectivity</td>
</tr>
<tr>
<td></td>
<td>• WAN optimization</td>
</tr>
</tbody>
</table>
Atomic Units
An atomic unit is the smallest possible unit of measurement and collection, for a billable item, that will be used for billing purposes. The consumer bill will typically contain information on how many atomic units of a resource were used.

The goal is to find a reasonable unit of measurement, collection, and billing that makes it easy to aggregate and store the billing data and also provides an appealing price point to the consumer.

Pricing Plans
After the cost model has been built, the billable Items have been identified, and the atomic units have been defined, it is possible to build one or more pricing plans.

Building a pricing plan requires associating costs to pricing for billable items and then combining these into different plans that are offered to cloud consumers. Each billable item can have different pricing options that can be either fixed or variable pricing. In fact, the same billable item might be offered with different pricing options as part of different plans or sales promotions. There are several typical pricing options.

Fixed Recurring Pricing
Fixed recurring pricing is the simplest pricing option, where the billable item has a fixed periodic cost. For example, a virtual machine can be offered at a fixed cost per month. The consumer is billed the same amount every month without consideration for actual usage. Fixed recurring pricing can also be used as a base price for a billable item.

Variable Pricing by Resource Consumption
Variable pricing by resource consumption involves billing the consumer for the actual amount of atomic units of the billable item that were used during the billing period. For example, an IaaS deployment might measure the CPU usage of virtual machines and calculate pricing based on the consumed amount of processing power measured in 1 GHz units.

Variable Pricing by Time
Variable pricing by time involves billing the consumer based on how long the billable item was used. The units here are modeled as time increments. For example, the price can be modeled per hour of usage. For a virtual machine priced on an hourly basis, the consumer can be billed monthly for the total number of hours the virtual machine was in use (powered on) in that month.

Cost Multipliers
Some billable resources might be modeled as cost multipliers, rather than giving a standalone price. For example, data services such as compression and encryption can be modeled as multipliers on the cost of storage. SLAs can be priced in a similar manner.

Figure 4 shows a sample pricing model for an IaaS deployment.

In this plan, virtual machines are offered both on a monthly rate and at an hourly rate. Physical machines and other services can be purchased at a fixed monthly rate. Bandwidth is priced by the gigabytes used. Some services, such as data encryption and data backups, are priced based on the data size. SLAs are priced as multipliers of the base virtual machine rates.
Figure 4. Sample Pricing Plan

<table>
<thead>
<tr>
<th>Item#</th>
<th>Type</th>
<th>Name</th>
<th>Pricing Type</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>Virtual Machine</td>
<td>Compute Profile 1</td>
<td>Monthly Rate</td>
<td>$20.00</td>
</tr>
<tr>
<td>1001</td>
<td>Virtual Machine</td>
<td>Compute Profile 1</td>
<td>Hourly Rate</td>
<td>$0.30</td>
</tr>
<tr>
<td>1002</td>
<td>Virtual Machine</td>
<td>Compute Profile 2</td>
<td>Monthly Rate</td>
<td>$21.00</td>
</tr>
<tr>
<td>1003</td>
<td>Virtual Machine</td>
<td>Compute Profile 2</td>
<td>Hourly Rate</td>
<td>$0.70</td>
</tr>
<tr>
<td>2000</td>
<td>Physical Machine</td>
<td>Blade1</td>
<td>Monthly Rate</td>
<td>$120.00</td>
</tr>
<tr>
<td>2001</td>
<td>Physical Machine</td>
<td>Blade2</td>
<td>Monthly Rate</td>
<td>$210.00</td>
</tr>
<tr>
<td>3000</td>
<td>Network Services</td>
<td>Bandwidth</td>
<td>Usage / GB</td>
<td>$0.02</td>
</tr>
<tr>
<td>3001</td>
<td>Network Services</td>
<td>Load Balancer</td>
<td>Monthly Rate</td>
<td>$130.00</td>
</tr>
<tr>
<td>3002</td>
<td>Network Services</td>
<td>Firewall</td>
<td>Monthly Rate</td>
<td>$80.00</td>
</tr>
<tr>
<td>3003</td>
<td>Network Services</td>
<td>VLAN</td>
<td>Monthly Rate</td>
<td>$2.00</td>
</tr>
<tr>
<td>3001</td>
<td>WAN</td>
<td>VPN</td>
<td>Monthly Rate</td>
<td>$299.00</td>
</tr>
<tr>
<td>4000</td>
<td>Storage Data Services</td>
<td>Encryption</td>
<td>Monthly Rate Per GB</td>
<td>$0.01</td>
</tr>
<tr>
<td>4001</td>
<td>Storage Data Services</td>
<td>Daily Backup</td>
<td>Monthly Rate Per GB</td>
<td>$0.10</td>
</tr>
<tr>
<td>5000</td>
<td>SLA</td>
<td>Bronze</td>
<td>Multiplier</td>
<td>1.00</td>
</tr>
<tr>
<td>5001</td>
<td>SLA</td>
<td>Silver</td>
<td>Multiplier</td>
<td>0.10</td>
</tr>
<tr>
<td>5002</td>
<td>SLA</td>
<td>Gold</td>
<td>Multiplier</td>
<td>0.15</td>
</tr>
<tr>
<td>5002</td>
<td>SLA</td>
<td>Platinum</td>
<td>Multiplier</td>
<td>0.25</td>
</tr>
<tr>
<td>6000</td>
<td>Security</td>
<td>Compliance - PCI</td>
<td>Multiplier</td>
<td>0.10</td>
</tr>
<tr>
<td>6001</td>
<td>Security</td>
<td>Isolation - Physical</td>
<td>Multiplier</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Figure 5 shows a sample monthly bill for a customer that has purchased an hourly rate virtual machine and a monthly rate virtual machine, with different SLAs for each. The consumer also used the backup service, and the bill includes the cost for the bandwidth used.

Figure 5. Sample Monthly Bill

Chargeback Implementation

To implement chargeback, you need the following types of tools to be built or purchased and integrated:

1. Data collection tools
2. Chargeback mediation tool
3. Billing system
4. Capacity analysis tool

Figure 6 shows the relationships and interaction between each of these tools.
Data Collection Tools

Data collection tools are required to collect and store data, based on the identified atomic units, for each type of billable item. The collection tools will vary based on the type of resource, and it is likely that multiple tools will be needed, one for each major system that is used. For example, the element management system (EMS) for a storage vendor might report detailed information on storage usage, and another EMS or tool might provide similar information on network usage.

In some cases, customized data collection—for example, use of the Cisco Unified Computing System XML API—may be required to collect appropriate, real-time, usage information.

If usage-based pricing is required, data collection for virtual machines cannot be performed at the guest OS level, because that data is not an accurate measure of the actual resource usage. Instead, the per virtual machine data must be collected from the hypervisor or virtualization management system.

Chargeback Mediation System

A mediation system is used to correlate, aggregate, and store the data collected from the various system components. The data is correlated into a billing record. The mediation tool can also map the billing record to a customer or account identifier. This requires the mediation tools to be integrated with a system or database that can provide the mapping of resource identifiers to a customer identifier.

Billing System

The billing system applies the pricing model to the collected data to generate a periodic billing report per customer or account. The billing system, often purchased as an off-the-shelf component, should be flexible in allowing the input of various pricing models and associating these to a billing entity. In some cases, the same tool might provide both the billing and the mediation functions.

Capacity Analysis

Cloud computing deployments require dynamic capacity management. A capacity analysis tool periodically collects system usage data and can provide trending analysis features to understand usage patterns and variations. An important aspect of this analysis is to understand periods of peak utilization and low utilization.

Capacity analysis can also be used to plan growth. It can lead to adjustments in the pricing model and strategy. For example, incentive pricing can be offered for periods of low utilization.
Designing the Solution—The Case for Professional Services

Designing and integrating chargeback solutions for cloud computing deployments is a demanding task. It requires not only detailed chargeback process design experience and IT services pricing policy design, it also requires detailed knowledge of the instrumentation present in the underlying network and compute infrastructure. It also requires experience of integrating IT systems management and billing tools, and—being a significant project in its own right—a structured approach to program management.

Cisco Services has significant expertise and experience in each of these areas, backed up by multiple years of helping customers transform their data centers. Cisco Services offers a range of Cloud Enablement Services, from strategy to planning and design to implementation. Chargeback is an integral component in each of these services, and detailed chargeback methodologies have already been developed by the cloud computing experts in Cisco Services. Cisco Services, therefore, are ideally placed to help you design and develop a state of the art chargeback mechanism for your cloud computing deployment.

Summary

A pay-for-use approach, or chargeback model, is an essential component of a cloud computing architecture. Even when billing and payment are not required, chargeback provides a way to map resource usage within a cloud to the service users, which is essential for proper capacity management and growth planning.

Implementing chargeback requires an understanding of direct and indirect deployment costs. The costs are used to build pricing plans, which contain the billable items that the consumer can purchase as a service.

Because of its wide scope, implementing chargeback requires integration across multiple tools and systems, which is performed by a mediation platform that collects, aggregates, and correlates the billing data. A well-implemented chargeback system helps lower costs and provides support for flexible pricing options while enabling financial modeling to maximize profitability and helps make sure of accountability for a cloud deployment. Implementing a chargeback strategy is often a pivot point in the enterprise journey to cloud computing, as it brings a new level of transparency and efficiency to enterprise IT. Finally, Cisco Services are ideally placed to help you devise and develop an advanced chargeback approach that will help justify your investment in cloud computing.

For more information, please visit [www.cisco.com/go/cloudenablement](http://www.cisco.com/go/cloudenablement) or consult your local Cisco representative.