# The Economics of the AWS Cloud vs. Owned IT Infrastructure

*Published: December 7, 2009*

Amazon Web Services (AWS) offers companies of all sizes an elastic, reliable, flexible, low-cost infrastructure web services platform in the cloud. Many companies have already launched applications in the cloud while others are currently evaluating the costs and benefits of moving some or all of their IT infrastructure to the cloud. This document presents a qualitative discussion of the costs of Amazon Web Services vs. traditional IT infrastructure alternatives. The discussion below offers a comparative analysis of several direct costs of ownership such as hardware costs and asset utilization, data redundancy and security, supply chain management, power and cooling efficiency, and personnel costs. Also included in this document is a brief discussion on the indirect costs of running your own data centers. Finally, AWS has published an *Amazon EC2 Cost Comparison Calculator*, a Microsoft Excel-based quantitative tool, to help financial decision makers quantify the direct economic benefits of cloud computing compared to traditional IT infrastructure alternatives. A current version of the Amazon EC2 Cost Comparison Calculator is available for free download at http://aws.amazon.com/economics.

# Direct Costs

## Asset Utilization

Utilization of hardware assets is one of the key areas where enterprises can benefit from deploying to the cloud. In traditional enterprise-owned data centers, server utilization commonly averages 5%-20% when measured annually.[1] While investments in virtualization and related technologies can improve server utilization, the CIOs with whom AWS is in regular contact believe that post-virtualization utilization rates of 20%-25% are still the highest they can achieve. In contrast, the AWS pay-for-use pricing model only charges customers for resources they actually use, so customers can effectively achieve close to 100% utilization. AWS is able to achieve greater overall utilization of its hardware assets because of its large and heterogeneous customer population. Within that population exists thousands of workloads, with non-correlating peaks and valleys. As an example, a financial services firm with peaks at the beginning and end of each trading day will have their utilization offset by an ecommerce firm with a shopping peak in the middle of the day and by a pharmaceutical company data analysis job running overnight. In addition, this large customer base allows AWS to make larger investments in efficiency innovations than individual enterprises, leading to continuous maximization of its infrastructure efficiency, ultimately benefitting AWS customers.

Moreover, Amazon Elastic Compute Cloud (Amazon EC2) features such as Auto Scaling and Elastic Load Balancing enable businesses to automatically grow or shrink their usage of AWS based on the actual performance of their application. In so doing, they can minimize their waste of AWS resources and achieve a utilization rate that truly does approach 100%.

## Hardware Costs

Thinking about long-term value of assets is critical for enterprises making multi-million dollar investments in IT infrastructure. In typical enterprise data centers, large initial capital outlays make ongoing upgrades in technology (i.e., the newest servers, routers, or load balancers) prohibitively expensive. Over time costs remain fixed, but so does performance. The economies of scale available with the cloud allow AWS to purchase large volumes of hardware at very low costs. Consequently, AWS customers reap the benefits of decreasing costs, increasing performance, and enhanced functionality over time. The expectation of improving performance at lower costs is illustrated by the cost of Reserved Instances for Amazon EC2, which offers over 50% savings from On-Demand (hourly) prices. Reserved instances can also be turned off at any time they're not being used to avoid usage charges (e.g., to cover costs of cooling, power, etc.). These are costs that enterprises can't avoid if they are running data centers themselves.

## Power Efficiency

According to most industry reports, the average data center Power Usage Effectiveness (PUE) is 2.5. This means that for every 1 Watt of power that is delivered to the servers, 1.5 Watts are wasted in overhead. Serious energy-efficiency efforts require dedicating IT and Engineering resources, using the most efficient equipment, and adhering to industry best practices, which often are not feasible expenses for enterprises. However, heavily-invested data centers, such as the ones that make up the AWS cloud, are far more efficient than average. Businesses looking to run their own data centers would need to invest heavily in ongoing efficiency efforts to decrease the PUE ratio of their facilities. However, to justify the needed investments, businesses must be operating on a large scale, with a large number of servers across

[1] Source: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf

multiple data centers. Managing multiple data centers, each with a large number of servers, is more efficient than managing an enterprise-sized data center.

## Enabling Redundancy

A highly reliable and available IT infrastructure requires enterprises to not only maintain reliable storage and backup devices, but also operate a reliable network with redundant networking devices, transit connections, and physical connections between data centers. This goes beyond RAID, given the average failure rates inherent in any single storage array or device. In addition to backup and reliable networking, enterprises must also have a tested, working solution for disaster recovery. This includes deploying data and applications across multiple data centers – either with failure resilient software or in a more traditional hot/cold standby approach. To achieve realistic disaster recovery, all of the data centers and servers involved have to be constantly utilized; if they sit idle, it's almost certain they won't function as desired when activated from a cold start. So an enterprise needs to account for both the cost and the complexity of this redundancy when evaluating their deployment. In contrast, AWS includes all this in its simple usage charges, and lets customers easily do things like deploy servers in multiple Availability Zones, which will not fail due to the same physical causes (e.g. power failures, cooling failures, fire, lightning, etc.).

## Security

Another direct cost for enterprises running their own data center is ensuring the confidentiality, integrity, and availability of business critical data. Examples of security costs for enterprises include capital expenditures for network security devices, security software licenses, staffing of an information security organization, costs associated with information security regulatory compliance, physical security requirements, smart cards for access control, and so on. To provide end-to-end security and end-to-end privacy in the cloud, AWS builds services in accordance with security best practices and features, and clearly documents how developers can effectively use those features. AWS customers thus take advantage of Amazon's reliable and secure global computing infrastructure, which has been the backbone of Amazon.com's multi-billion dollar retail business for more than 15 years, at no additional cost to the customer. For more information on AWS security, consult the *Amazon Web Services: Overview of Security Processes* whitepaper at aws.amazon.com/security.

## Supply Chain Management

In traditional enterprise data centers, it is fairly common to experience capacity constraints caused by the time that passes from when hardware is ordered to when it is brought online – often running many months. Such long lead times necessitate having excess capacity that spreads throughout the pipeline and increases costs. Dedicated service providers like AWS minimize this excess capacity by devoting significant resources to effectively managing its supply chain and amortizing these investments over a large customer and hardware base. It is difficult for enterprises to justify spending as much time and money when amortizing these investments over even a large enterprise data center, as it would typically serves fewer customers and contains far less hardware than the AWS cloud.

## Personnel

Personnel costs include the cost of the sizable IT infrastructure teams that are needed to handle the "heavy lifting" – managing heterogeneous hardware and the related supply chain, staying up-to-date on data center design, negotiating contracts, dealing with legacy software, operating data centers, moving facilities, scaling and managing physical growth, etc. – all the things that an enterprise needs to do well if it wants to achieve low infrastructure costs in the areas discussed above. For example:

- Hardware procurement teams are needed, who have to spend a lot of time evaluating hardware, negotiating, holding hardware vendor meetings, managing delivery and installation, etc. It's expensive to have a staff with sufficient knowledge to do this well.
- Data center design and build teams are needed to create and maintain reliable and cost-effective facilities.
- Operations staff is needed 24/7/365 in each facility.
- Networking teams are needed for running a highly available network. Expertise is needed to design, debug, scale, and operate the network and deal with the external relationships necessary to have cost-effective internet transit.
- Security personnel are needed at all phases of the design, build, and operations process.

# Indirect Costs

As important as direct cost savings are, there are many indirect costs that attract customers of all sizes to the AWS cloud. Foremost among them is the opportunity cost of owning, operating, and maintaining traditional IT infrastructure. Running large scale, high availability infrastructure requires the efforts of many talented staff members and the dedicated attention of upper level management. This represents lost opportunity for enterprises to focus on and innovate in their core businesses.

Many customers have stated that AWS is more reliable than what they achieve themselves. For example, consider how often corporate email is inoperative or unreliable for many companies. AWS offers in-the-cloud services that many thousands of external customers use for mission-critical applications, and this requires AWS to prioritize operational excellence and spend significant resources monitoring its systems 24 hours a day, 7 days a week. Indeed, operational excellence has always been the lifeblood of Amazon.com. AWS also publishes a service health dashboard to provide continuous, real-time visibility into operational performance.

Moreover, many enterprises simply don't have the capital budgets required to build, extend, or replace IT infrastructure in a capital-constrained environment. As a result, many are simply foregoing important projects due to lack of capital. Enterprises also find that the flexibility offered by the AWS platform enhances the agility of their business, improving their ability to requisition compute resources, experiment quickly, or manage unanticipated demand. Using Amazon Web Services, an e-commerce web site can weather unforeseen demand with ease; a pharmaceutical company can "rent" computing power to execute large-scale simulations without having to go through a laborious requisition process; a media company can, within minutes, serve unlimited videos, music, and more; and an enterprise can deploy bandwidth-consuming services and training to its mobile workforce. The cloud is not simply a way of saving money, but it also makes it possible to be more productive, more agile, and more responsive to opportunity than is possible by simply provisioning physical hardware in an enterprise's own data center.