the Availability Digest

How Many 9s in Amazon?

July 2008

Founded in 1994 and headquartered in Seattle, Washington, Amazon.com is arguably the largest online retailer in the world. Supporting not only its own retail operations but also those of thousands of small retailers (and some large ones as well), Amazon welcomes about 60 million shoppers to its stores each month.

Amazon started out as an online bookstore. It soon diversified and now carries a wide range of products, including video tapes, DVDs, CDs, computers, video games, electronics, toys, sports equipment, apparel, and tools. It has web sites in the U.S., Canada, the U.K., France, Germany, China, and Japan. Its annual revenues exceed $10 billion, and Amazon is now one of the stocks in the S&P 500.

Amazon Web Services

In 2002, Amazon launched Amazon Web Services (AWS). AWS is a collection of online services for other web sites or client-side applications using the immense Amazon infrastructure. These services offer functionality that other developers can use. Amazon claims over 300,000 developers using AWS.

Among the dozen services provided by AWS are the Amazon Simple Storage Service (S3) and the Amazon Elastic Compute Cloud (EC2). These two services are central to our *Never Again* story.

Amazon Simple Storage Service (S3)

In early 2006, Amazon launched Amazon S3. S3 is essentially what its name implies – an online storage service that a customer can use to store an unlimited number of data objects through a simple web services interface. Data objects may be from one byte to five gigabytes in size, and they are distributed by HTTP.

Amazon charges for storage and for bandwidth used. S3 uses the same scalable storage infrastructure that Amazon.com uses to run its global e-commerce network. It is reported that Amazon S3 currently stores more than fourteen billion objects.

S3's design goals include scalability, availability, and low latency at commodity costs. Objects are organized into *buckets* owned by an AWS customer. Within a bucket, objects are identified by unique keys. Bucket names and keys are structured so that objects are addressable via standard URLs. For instance, one URL format is http://s3.amazonaws.com/bucket/key.

Access to a bucket can be controlled by an access control list, and requests for data can be authenticated through the AWS Authentication service. As we shall see, this authentication service turned out to be an Achilles' heel for S3 customers.

Because objects are accessible by unmodified HTTP clients, S3 can be used to provide a standard web-hosting infrastructure. As a consequence, many small enterprises use Amazon S3 for web hosting, for image hosting, for database backup, and for many other purposes. S3 by any measurement has been hugely successful for Amazon.

Amazon Elastic Compute Cloud (EC2)

Later in 2006, Amazon introduced its EC2 service. EC2 provides scalable virtual private servers using the Xen open source virtualization hypervisor.[1] The virtualized server farm uses the Amazon server infrastructure with its high availability to run diverse customer applications, including web hosting.

Customers can buy the amount of virtualized computing capacity that they require. EC2 provides scalable deployment of applications by providing a web services interface via which customers can request any number of virtual machines. Each of Amazon's servers can run many such virtual machines. Customers can run any applications that they desire on any of their virtual machines. They can create, launch, and terminate virtual machine instances on demand; thus, the term "elastic."

When creating a virtual machine, the customer can specify that it be equivalent to a small, large, or extra-large server:

- A small virtual server is the equivalent of a one gigahertz Opteron or Xeon single-core 32-bit processor with 1.7 gigabytes of memory and 160 gigabytes of disk storage.

- A large virtual server is the equivalent of a 64-bit quad-core Opteron or Xeon processor with 7.5 gigabytes of memory and 850 gigabytes of storage.

- An extra-large virtual server is the equivalent of an eight-core processor with 15 gigabytes of memory and 1.7 terabytes of storage.

The combination of Amazon's S3 and EC2 web services is ideal for many companies that want to build large applications without the headache of managing their own server and storage farms. These AWS services now host thousands of small and large web sites.

Service Level Agreements

Initially, Amazon would not agree to provide service level agreements (SLAs). They argued that since S3 and EC2 services used Amazon's own infrastructure, of course they would be reliable. If any problem occurred, it would be Amazon's top priority to correct the situation.

However, in October of 2007, Amazon began offering an SLA for its S3 service. The SLA guaranteed a 99.9% uptime on a monthly basis. S3 is the only AWS service for which Amazon currently offers an SLA.

The SLA did not have a lot of teeth for the customer. If the service did not achieve 99.9% uptime but did achieve at least 99% uptime, then 10% of the monthly fee charged to the customer would be applied against the next month's charges. This adjustment would apply if S3 was down for more than 45 minutes per month, and it covered downtime up to seven and a half hours per month. For a typical customer, this might amount to a few hundred dollars a month in compensation for hours of lost business.

[1] Fault Tolerance for Virtual Environments – Part 2, *Availability Digest*; April, 2008.

If S3 achieved less than 99% uptime, then 25% of the customer's monthly fee would be applied to the following month's charges.

Friday, February 15, 2008 – The Dreaded S3 Outage.

Remember the S3 authentication services provided by AWS? These authenticated data requests are encrypted and consume significantly more computing resources than normal data requests. Amazon provides the server capacity that is required to handle this authentication load based on two years of experience.

However, early in the morning of Friday, February 15, 2008, around 3:30 AM, system operators started seeing unusually elevated levels of authentication requests for multiple users at one of their three locations that provide AWS services. Shortly before 4:00 AM, several other users significantly increased their volume of authenticated requests. This overloaded the S3 infrastructure, and Amazon was unable to process any requests at that location. The infallible S3 services were down!

It took until 6:48 that morning to reconfigure enough capacity to handle the increased authorization load and to return S3 services to Amazon's customers.

For almost three hours, thousands of web sites that depended upon S3 and its companion EC2 service were down. This included such sites as social networking site Twitter, browser technologist AdaptiveBlue, photo-sharing service SmugMug, and web-application developer 37Signals.

An outcry arose from the customer community about Amazon's lack of communication concerning the problems during the outage. Customers complained that they had no way to know whether the problem was due to their equipment or to Amazon's services. There were neither email notifications nor updates on Amazon's AWS blog. Customers felt that there should at least have been a notice posted on the front page of Amazon's web services site so that they would have known that the problem was not theirs.

Following restoration of services, Amazon acknowledged the severity of the problem and listed a set of actions that it intended to take to prevent such an occurrence in the future. These actions included:

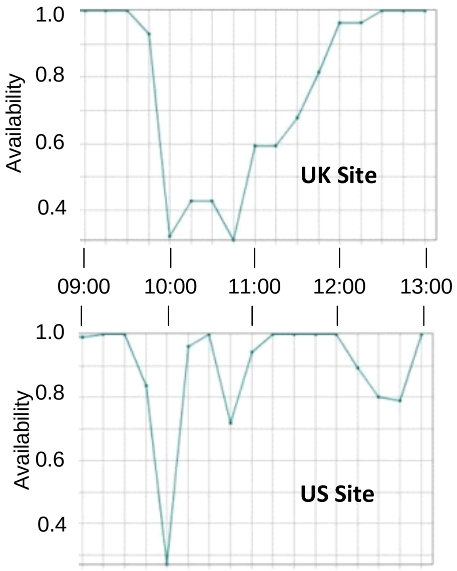
- improving monitoring of authenticated call requests.
- increasing authentication service capacity.
- adding additional defensive measures around authenticated calls.
- developing a service health dashboard (an uptime dashboard).

This episode wasn't the end of EC2 problems. Early the morning of April 7, EC2 reportedly again went down for about an hour. However, this time it affected only some U.S. customers. This, of course, is small consolation to those affected.

Friday, June 6, 2008 – An Encore

During the morning of Friday, June 6, Amazon suffered another embarrassment. This time it was in its core business - the Amazon.com online retail store. Amazon.com went down for over three hours, affecting U.S and U.K. customers. All these customers received was an error message that said "Http/1.1 Service Unavailable." Other international customers were not affected, nor were AWS services. However, U.S. and U.K. retail customers who could not connect to Amazon.com were not redirected to other operational sites, as they should have been.

U.S. and U.K. sites went down rather suddenly at about 9:30 AM. U.S. services returned about an hour later but were then sporadic until about 1:00 PM. U.K. services were restored around noontime.



(Credit: Keynote Systems)

This time, there was no explanation for the outage (so far as we know). Amazon's Director of Strategic Communications simply said, "Amazon systems are very complex and on rare occasions, despite our best efforts, they may experience problems." Not much help here.

Theories on the outage abound in the blogs. One common theory was that, regardless of site problems, Amazon's network is configured at several levels to redirect requests that cannot be delivered. This capability evidently did not work.

Another theory was that the sites were taken down by "bots" that created a load so great that the servers could not handle it, and they shut down. This theory was supported by the fact that this failure coincided with a known attack on Amazon's Internet Movie Database (IMDB). A single attacker flooded this site with requests for images.

The theory was further supported by many users who said that their requests were blocked because the requests were suspected of coming from a bot or an automated script. Others conjectured that the outage was caused by bots trying to scoop up Metal Gear Solid 4, an 80-gigabyte pack for PlayStation 3 that had just been released. This had happened once before on Thanksgiving Day, 2006, when Microsoft released its Xbox video games.

In any event, Ad Age estimated that the outage cost Amazon about $1.8 million dollars for a two- to three-hour outage.

Some Outside Help for EC2 Users

Hyperic, Inc. (www.hyperic.com), a young company providing open-source monitoring and management software for web infrastructure, has announced its product Hyperic CloudStatus that will provide an independent view of the health and performance of some of Amazon's web services. Included in the monitored services are S3 and EC2. This free service gives AWS customers the perspective they need to determine the cause of performance problems in their cloud-based web applications.

According to Hyperic, CloudStatus "provides a comprehensive measure of service availability, latency and throughput for cloud-based infrastructure and application services. Users can drill down for detailed metrics on any of the monitored offerings. These metrics are specific to each individual service and are designed to answer the questions most often asked by the developers and administrators that rely on the services for their business."

Lessons Learned

Customer Communications

Like so many other service providers about which we have reported, Amazon did not get kudos for its customer communications, either during its EC2 failure or during its retail-store failure. Customers can be far more forgiving and understanding if they simply know what is going on.

Cloud-Computing 9s

Cloud computing can bring tremendous advantages to small organizations that do not want to manage their own data centers. Scalable storage and computing capacity on demand make it easy to host, manage, and operate their own applications, whether it be their web sites or other corporate applications.

However, it is clear that if uptime is critical, an organization must accept that cloud-computing systems do fail and that the business will suffer downtime over which the organization has no control. Even a three 9s SLA means that there may be over eight hours of downtime per year – and that is if the SLA is met.

Therefore, some sort of backup must be provided to a cloud-computing environment to carry business operations through these outages. Even Amazon is not immune to the failure of its own highly-sophisticated infrastructure, as evidenced by the June 9th failure of its U.S. and U.K. retail operations.

Certainly, as big-technology companies such as Amazon, Google, IBM, and others start to compete in the web-based services arena through cloud computing, reliability will be one of the main features distinguishing their offerings.[2]

[2] Material for this article was taken from the following sources:
EC2, S3, AWS, *Wikipedia*.
Amazon Web Services Face Outage, *WHIR News*; February 15, 2008.
Amazon explains its S3 outage, *blogs.zdnet.com*; February 16, 2008.
Amazon S3 Outage: Do SLAs Lead to Trust?, *infoq.com*; February 28, 2008.
Amazon's Web Services Gets Another Hiccup, *techcrunch.com*; April 7, 2008.
Why Amazon Went Down, and Why It Matters, *gigiom.com*; June 6, 2008.
Amazon loses $1.8 million due to outage, *network.nationalpost.com*; June 6, 2008.
Amazon.com suffers another outage, *news.cnet*; June 9, 2008.
Bots to blame for Amazon.com outages?, *The Register*; June 9, 2008.
Amazon.com Suffers More Outages, *Information Week*; June 10, 2008.
Further Disruptions on Amazon.com, *redorbit.com*; June 10, 2008.
Hyperic To Monitor Amazon.com's Cloud Computing Availability, *Information Week*; June 24, 2008.