the Availability Digest

Can You Trust the Compute Cloud?

August 2008

Thousands of small to medium businesses (and some large ones as well) use Amazon.com's services to implement their online presences. Amazon significantly extended these services recently with AWS, the Amazon Web Services, which opens Amazon's massive infrastructure to its customers. Predominant among these services are S3, the Simple Storage Service, and EC2, the Elastic Compute Cloud. S3 allows customers to store their data on Amazon's massively-redundant distributed database. EC2 provides server capacity on demand for the customer by creating virtual machines that run on Amazon's server infrastructure.[1] These services are known as cloud computing.

If these services should go down, thousands of businesses go out of business during the outage. And that is exactly what has happened this year. Amazon has racked up hours of downtime in several incidents during the past six months.

Can cloud computing be trusted? What can a business do to protect itself from outages over which it has no control? Does cloud computing have a future? We explore these questions in this article.

What is a Compute Cloud?

Building a data center is serious business. Companies must worry about costs, facilities, staffing, administration, security, availability, disasters, and many other factors. Given a commitment to build its own data center, a company often finds that it is a victim of its own efforts. In today's competitive environment, there must be fast response to market pressures. New applications must be put into service quickly, even if on a prototype basis to test new approaches.

However, the company's data center is constrained by capacity and budget. New projects compete with existing projects for valuable resources. Fast-moving market factors may require a rapid response that is outside of the budgeting process. Can the data center accommodate a new request? On what servers should it run? How long will it take to get management approval and budget?

Werner Vogels, Amazon's CTO, has said that if managing a massive data center isn't a core competency of your business, maybe you should get out of this business and pass the responsibility to someone who has that competency.

Like Amazon.

[1] How Many 9s in Amazon?, *Availability Digest*; July 2008.

Vogels submits that Amazon has a massive computing infrastructure and over a decade of experience managing it. Why not host a business' applications on Amazon's systems and let Amazon do all of the work? A company can move much faster at far lower costs. Businesses can design their own databases and can dynamically allocate only the server capacity required. Best of all, it's cheap!

This capability is known as "cloud computing." Instead of a company managing its own data center, it uses the services of a data utility. Just as an electric bill is determined by usage of electricity, a data utility bill is determined by how much data is stored and by how much computing capacity is used.

Forrester Research defines cloud computing as:[2]

> *A pool of abstracted, highly scalable, and managed compute infrastructure capable of hosting end-customer applications and billed by consumption.*

This is exactly what Amazon is offering with its S3 and EC2 services, among others.

Amazon's Dismal Cloud-Computing Record

Amazon today is the leading provider of cloud computing. Cloud computing is used primarily by small companies that do not have IT investments to manage. They are attracted by the very low cost of computing compared to the equipment and personnel costs that they would incur if they were to build their own data centers. Furthermore, they have near-infinite flexibility and are able to increase or decrease their processing capacity and storage requirements virtually on-the-fly.

However, Amazon's compute cloud is still in its early stages. It is far from mature. Amazon is reluctant even to provide service level agreements, though in the last year they started to offer a rather toothless SLA for their S3 service.

The lack of stability of the Amazon cloud has been dramatically demonstrated in the last six months. In February, 2008,[1] Amazon's S3 service was overwhelmed by a flood of authorization requests and was down for over three hours in the United States and Europe.

Then in July of this year, S3 died again, this time for eight hours.[3] In this case, Amazon reported that the problem was due to communication problems between several of its components. Though EC2 was not affected directly, the S3 failure prevented new virtual machines from being registered and existing virtual machines from being launched.

The good news is that no data was lost. Amazon stores S3 data in several geographically-distributed locations to ensure that there will be no data loss, even in the event of a disastrous loss of an entire data center.

The bad news is that thousands of online stores were down for hours. Per its SLA, Amazon refunded hundreds of dollars in fees to companies to compensate them for perhaps hundreds of thousands of dollars in lost revenues.

Amazon's infrastructure failures are not limited to its cloud-computing services. In June of this year, Amazon suffered when its own online store went down for over three hours. Amazon's only explanation was that its "systems are very complex and on rare occasions … may experience problems."

Om Malik, a well-known technology author, wrote the following after the latest Amazon debacle:

[2] Is Cloud Computing Ready for the Enterprise?, *Forrester Research*; March 7, 2008.
[3] More outages hit Amazon's S3 storage service, *Networkworld.com*; July 21, 2008.

"But even as cloud computing starts to gain traction with companies … and most of our business and communication activities are shifting online, web services are still fragile, in part because we are still using technologies built for a much less strenuous web."

In short, he submits that the compute cloud is fragile. Amazon's experiences support that conjecture.

What's a Small Business To Do?

The cost, administration, and flexibility arguments for cloud computing use by small companies are compelling. On the other hand, the lack of high availability is likely to be a great concern for many.

Some companies are willing to accept reduced availability in return for cloud computing's advantages. SmugMug, a major photo and video storage site, said the following after Amazon's latest outage:[4]

"Every component SmugMug has ever used, whether it's networking providers, datacenter providers, software, servers, storage, or even people, has let us down at one point or another. It's the nature of the game, and our job is to handle these problems and outages as best we can."

It appears that SmugMug feels that it can be down occasionally because the primary impact is the unavailability of material that its customers have stored on its site.

However, those who lose significant revenue when their site is down or who feel that their services are more critical will probably disagree with this. What is needed for critical applications are:

- Redundancy so that backup facilities can be employed following a cloud failure.
- Good communication from the cloud provider concerning the status of the compute cloud during an outage.
- Customer monitoring of the compute cloud so that problems can be isolated to the cloud or to the customer's systems.

Redundancy

If high availability is to be achieved with cloud computing, the preferred solution is the same as it is with any computing infrastructure – redundancy. If the cloud goes down, you must have a backup plan. Typically, this means that you should have a backup site to which you can switch during the outage to provide at least the minimum services that your business requires.

An example of a company with a backup plan is Mediafed.[5] Mediafed provides traffic analysis of RSS (Really Simple Syndication) feeds to European media companies such as BBC and LeMonde. The company uses two cloud providers side-by-side – Amazon and UK-based Flexiscale. Both clouds run at the same time, backing each other up. Customers are split between the two clouds, and impacted customers can be quickly switched over to the surviving cloud in the event of a cloud outage. Mediafed survived the Amazon S3 failure with no problem.

[4] Amazon's S3 experiences outage, *News.cnet.com*; July 20, 2008.
[5] S3 outage: time to double up, Phil Wainewright, *blogs.zdnet.com/SAAS/;* July 21, 2008.

Amazon's New Availability Zones

Recognizing the need for redundancy for critical applications, Amazon introduced this past March a powerful redundancy option – Availability Zones and Elastic IP Addresses.

Amazon divides the world into geographic regions. Each region contains several *Availability Zones.* Each Availability Zone is a distinct location within a region and is insulated from faults in other Availability Zones. Communication connections are provided between the Availability Zones, which are close enough together so that communication latency should not be a significant factor in application performance (round-trip communication latency between two Availability Zones 500 miles apart is about twenty milliseconds).

A customer can select an Availability Zone to launch an instance of his application. He can also launch a backup instance in another Availability Zone in the same region. One of these instances is the primary instance. The database in the backup instance is kept synchronized with the primary data database via data replication.

Amazon's *Elastic IP Addresses* allow the customer to dynamically associate an IP address with an instance of his application. In normal operation, the IP address points to the primary instance of his application. However, should the primary instance fail, user requests can be rapidly switched to the backup instance.

Following a primary failure, the backup database will be brought into a state of consistency. The backup applications will be started and will connect to the backup database. The backup application instance will assume the IP addresses used by customers, and the application will be back up and running. This process should take minutes compared to the hours of downtime experienced during Amazon's recent outages.

If desired, the new primary instance can at this time start another backup instance in a surviving Availability Zone.

Communication

A common problem with cloud computing and other Software as a Service offerings is communication of status during outages. Following Amazon's first outage in February, there was a great outcry from its customer base due to the lack of information forthcoming from Amazon.

Recognizing the need for up-to-date status information concerning its cloud, Amazon implemented a web-based cloud-health dashboard. This facility played a major factor in helping customers following its July outage.

Cloud Monitoring

Another problem that has surfaced as a result of Amazon's outages is the need for a company to determine the root cause of an outage. Many companies use S3 and EC2 as an adjunct to their systems. When Amazon's cloud failed, they had no idea where the problem was. Was it in their systems, in Amazon's systems, or in some other systems upon which they may have been dependent?

This led to a realization that facilities must be provided by the company to monitor the health of the cloud so that the nature of a problem can be determined and that restoration efforts can be focused on the true problem. Some companies have built their own monitoring tools to fill this need.

In addition, 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 Amazon's cloud.

Maybe a Rainbow After the Cloudburst

The realization today is that cloud computing does not yet meet enterprise standards for information technology. It does not provide for much in the way of administration and monitoring by the user, and its availability is certainly less than desirable. In Amazon's defense, it should be noted that it is a pioneer in cloud computing. Though it has made tremendous strides in this area, it is always the pioneer that gets the arrows in his back side.

In a companion blog to Om Malik's statement referenced earlier, Larry Dignan said:[6]

> *"And like any company wrestling with legacy systems, cloud-computing vendors will dust off a tired playbook. The solutions will be the usual: Relegate legacy systems to plumbing and create more services and applications to keep infrastructure current. In other words, the cloud will likely become more of a rat's nest. What's scary about that prognosis is the cloud is already too complicated since it's built on creaky infrastructure."*

What is needed is a major effort to reinvent Web services that can reliably and adequately support cloud computing. This is spawning major research efforts to do just that.

For instance, HP, Intel, and Yahoo are joining other government and academic institutions to initiate a large-scale research project to develop a more robust cloud-computing infrastructure.[7] Joined by the Infocomm Development Authority (IDA) of Singapore, the University of Illinois at Urbana-Champaign in conjunction with the National Science Foundation, and Karlsruhe Institute of Technology of Germany, these six participants will host a multidata-center compute cloud spanning three continents for experimentation and research into cloud computing. Based largely on Intel processors and HP computers, each of the six data centers in the cloud will contain 1,000 to 4,000 processing cores.

The project will revolve around open-source distributed technology developed by the Apache Hadoop project,[8] to which Yahoo is the major contributor. It will also use Pig, a parallel programming language developed by Yahoo.

IBM and Google are also working on a major cloud-computing research program with several major universities. They plan to roll out their network over the next year. Their compute cloud will also be based on Hadoop running on Linux and will use Xen for server virtualization. IBM says that cloud computing will allow it to reach small and medium businesses, which represent $500 billion in revenues and which IBM has trouble serving profitably through its usual sales channels.

It is also reported that Microsoft and AT&T are pursuing their own research projects in cloud computing.

Where Do We Go From Here?

Many experts believe that cloud computing will be the dominant IT delivery model of the future. However, based on current experience, this technology has a long way to go.

The cloud computing provider that is furthest along is clearly Amazon with its Amazon Web Services. Amazon's efforts have to date been the major factor in the maturing of this technology.

---

[6] Amazon's S3 outage: Is the cloud too complicated? Larry Dignan, *blogs.zdnet.com*; July 21, 2008.
[7] HP, Intel, Yahoo Join Government, Academia In Cloud Computing Research, *informationweek.com*; July 29, 2008.
[8] Where did the name "Hadoop" come from? It is the name of the develpor's child's stuffed elephant.

The commercial success of its offerings has spurred several major international research projects to bring cloud computing up to enterprise standards. With the massive efforts being put into this initiative, it is quite possible that cloud computing will provide a computing utility that will transform the IT industry.

Until then, users should introduce facilities to monitor the health of their cloud provider so that they can properly identify the source of problems. In addition, businesses that want to take advantage of the low cost, simple administration, and flexibility of cloud computing must consider their availability needs. In many cases, the cloud must be backed up with a redundant cloud or equivalent. Amazon now provides Availability Zones to meet this need, though they are at this point in time too new to judge their effectiveness.