

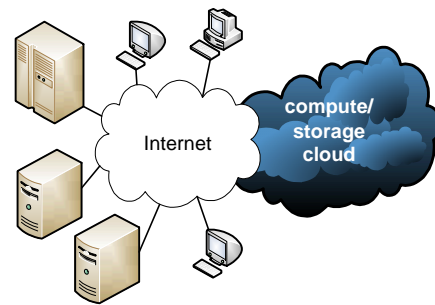
The Fragile Cloud

June 2009

Caution: Cloud computing can be hazardous to your health.

In our previous article, "The Fragile Internet,"¹ we questioned the dependability of the Internet for critical corporate functions. We pointed out several instances in which the Internet disappeared for hours or days and asked what you would do if a portion of your enterprise was suddenly left without Internet service for an extended period of time?

This quandary is magnified manyfold by a new and emerging paradigm – cloud computing. Cloud computing provides services that can be delivered and used over the Internet through an as-needed and pay-per-use business model. However, as any pilot knows, there is turbulence in a cloud; and if the cloud is a thunderstorm, that turbulence can be fatal.



In this article, we describe cloud computing and look at its recent availability experience. Based on these findings, we suggest what the proper use of the cloud is today, and what is not.

What is Cloud Computing

Cloud computing is an emerging business model by which users can gain access to their applications from anywhere, at any time, through any connected device. The applications reside in massively-scalable data centers where computational resources can be dynamically provisioned and shared to achieve significant economies of scale.²

A key requirement is that cloud services must be optimized for availability, data integrity, and security. We will focus on how well today's cloud services meet these criteria.

The Emergence of the Cloud

The scope of the term "cloud computing" varies widely in the literature. In this article, we will use it in a quite general sense, covering a wide range of efforts that started a decade ago.

¹ The Fragile Internet, *Availability Digest*; May 2009.

² Seeding the Clouds: Key Infrastructure Elements for Cloud Computing, IBM White Paper; February, 2009.

The first major form of cloud computing was the Grid. Then along came a variety of public application offerings grouped under the term Software-as-a-Service (SaaS). This has expanded into what we know today as cloud computing. The ultimate goal is the compute utility.

The Grid

Grid computing is an array of computing devices grouped to act in concert to execute very large tasks. An early use of grid computing was SETI@home, the Search for Extra-Terrestrial Intelligence project, which was started in 1999 and which harnessed over 5 million PCs around the world to parse through massive amounts of celestial data searching for any sign of intelligence.

Attempts to standardize grid computing started at about the same time when the Global Grid Forum (GGF) was created by major players in the IT industry. Through GGF's efforts, the Globus toolkit was created to help companies build their own grids. However, this effort seems to have gone the way of many other industry-supported standardization efforts – it has moved into oblivion.

Software-as-a-Service (SaaS)

In the early 2000s, service providers started offering products that performed the same function as major systems in the enterprise but which were provided via the service providers' data centers. With these services, which became known as Software-as-a-Service, or SaaS, companies could decommission their internal systems and use the SaaS services instead, thus saving significant costs associated with running their own systems.

Predominant among these are the many email services available today, many for free. There are many other examples. Salesforce.com offers Customer Relationship Management (CRM). Google offers Google Apps, which is an ad-supported free suite of useful collaboration, messaging, and office productivity services.

The Storage Cloud

As SaaS services matured, some companies with massive data-processing facilities moved to make their computing facilities available to the public on a fee basis. The first steps in this direction were to provide massive data storage in a controlled environment for companies. Thus was born the Storage Cloud.

A major example of the Storage Cloud is Amazon's S3 service (Simple Storage Service). Many other companies, both small and large, offer storage services for a fee or for free.

The Compute Cloud

During this time, many companies sprouted up that offered web-hosting services on their systems. Many of these built very large and powerful data centers. Typically, hundreds of web sites shared one large server; and these companies built data centers with thousands of servers hosting millions of small web sites. This was the first instance of the compute cloud.

In many cases, SaaS providers and hosting-service providers moved to open the computing capacity of their systems to the public, as did other companies with massive data centers. Salesforce.com introduced force.com. Hostway, a major hosting service, introduced FlexCloud. Rackspace, another major hosting service, introduced Mosso. Amazon introduced its Elastic Compute Cloud service (EC2).

Even the server and operating-system vendors are getting involved. Sun has announced its Sun Cloud Storage Service and its Sun Cloud Compute Service. Microsoft is experimenting with its Azure Services Platform.

These are all examples of today's state-of-the-art in cloud computing.

The Compute Utility

All of these cloud services fall short in one common area. They are all proprietary and cannot interoperate with each other. There are many advantages to cloud interoperability, including being able to store your data in multiple clouds and to run your applications in multiple clouds while accessing your data in some other cloud. This would be the ultimate in availability.

Sun is attacking this problem. Sun's vision is a world of clouds that are both open and that interoperate. They have announced the Sun Open Cloud Platform, which is a cooperative effort in the open-source community to develop an open and standard cloud.

Private Clouds

The public clouds described above are made available to users via the Internet and are free or inexpensive to use. Private clouds offer many of the same benefits as public clouds, most importantly the dynamic allocation of compute resources as they are needed. However, private clouds are managed within the enterprise. Private clouds can overcome some of the concerns that a company may have with public clouds, such as availability, control, and security.

Though there are no unified products available today to build a private cloud, many large vendors – particularly IBM and HP with its Cloud Assure services – offer consulting help to companies who may wish to build their own private clouds.

How Reliable is the Cloud?

Now we come to the meat of this article – is the cloud right for you? For your noncritical applications, it may be. However, for your critical applications, tread carefully. Let us look at some recent history.

Software as a Service

- December 20, 2005 – Salesforce.com upgraded to Oracle 10g. Unfortunately, the upgrade went horribly wrong; and Salesforce.com was down for most of the day. Another outage on January 5, 2006, interrupted service for almost three hours, followed by a several-hour outage on January 30th. February 9th and February 15th saw further outages like aftershocks from a large earthquake. Salesforce.com has since installed additional data centers that back each other up.³
- August 6, 2008 – An outage locked out Gmail users and Google Apps customers for 15 hours. On August 15, 2008, another outage locked out some Google users for a day. Two months later, on October 16, 2008, users went without Gmail access for thirty hours.

February 24, 2009, saw a two and a half hour outage affecting almost all Gmail and Apps Premier customers. During the maintenance of one of Google's European data centers, its traffic was routed to another nearby data center. This inadvertently overloaded that data center, which caused a cascading effect from one data center to another.⁴

³ On-Demand Software Utility Hits Availability Bump, *Availability Digest*, October 2007.

⁴ Has Gmail Become Gfail? *Availability Digest*, March 2009.

On May 14, 2009, Google Apps went down for almost two hours. A routing error redirected traffic through Asia, overloading data centers there. 14% of Google's customers (millions of users) were affected. A typical story was that of a California bank that lost its online banking services because they depended upon Google Analytics.

- August 7, 2008 - Citrix's GoToMeeting and GoToWebinar services were temporarily unavailable, resulting in meetings and webinars that could not be held. Citrix blamed a surge in demand
- January 7, 2009 – Salesforce.com went down for an hour. Even its status page, trust.salesforce.com, went down; so no one knew what was going on.

Compute Clouds

- July 27, 2007 – Hostway, a major web-site hosting service, planned a move of a newly-acquired data center from Miami, Florida, to Tampa, Florida. It informed its affected customers that they would be offline twelve to fifteen hours over the weekend. However, server problems caused by the move and network problems in the Tampa data center extended this to days and in some cases to over a week.⁵
- November 12, 2007 – Rackspace, another major hosting service, lost power when a truck hit a transformer outside of its data center. Emergency personnel would not let Rackspace use its emergency generator or switch to backup primary power because of danger to the rescue personnel. Thousands of web sites were down for a day while servers were restored to service. Many of these web sites were run by SaaS providers serving millions of end users.⁶
- May 31, 2008 – Operating six data centers, The Planet is the largest privately-held dedicated server hosting company and the fourth largest in the world. On May 31st, a battery-room explosion in one of its data centers blew out three interior walls and destroyed the power-transfer switch to its backup generator. The fire department evacuated the building. When personnel were allowed back in, they found that 9,000 servers leased by 7,500 customers were taken down by lack of power. Four days after the explosion, some customers still remained offline.⁷

On May 12 and May 13, 2009, The Planet had short half-hour shortages when an operator error caused IP addresses to be improperly advertised. Millions of web sites were affected.

- March 3, 2009 – Media Temple, yet another major web-site host, had a storage-system crash that corrupted its files. It was down for 38 hours while the database was restored. 3,000 customers were out of business for almost two days.

On March 6th, Media Temple again went down for two days for a similar reason. This outage took out 15,000 web sites. Media Temple offered all affected customers a one-year credit – an offer that cost it potentially millions of dollars.

⁵ Hostway's Web Hosting Service Down for Days, *Availability Digest*, September 2007.

⁶ Rackspace – Another Hosting Service Bites the Dust, *Availability Digest*, December 2007.

⁷ The Planet Blows Up, *Availability Digest*, September 2008.

Storage Clouds

- February 15, 2008 – An unanticipated increase in database authentication traffic caused Amazon's S3 servers to overload and be inaccessible. The infallible Amazon S3 and EC2 services were down for up to eight hours. A notable web site that was taken of the air by this outage was Twitter.⁸
- October, 2008 - Digital Railroad abruptly shut down. The online storage service posted a note to its web site stating that it had ran out of money and would have to close. Digital Railroad gave customers 24 hours to remove their images before the files would be destroyed. Subsequent access to Digital Railroad was severely limited as a crush of customers rushed to save images hosted on the company's servers. Many could not recover their files and lost most or all of their data.
- December, 2008 – JournalSpace, a popular blog-hosting service, lost its entire database when a disgruntled employee – the IT manager, of all people – wiped out the database. To add insult to injury, he had never established a backup procedure, a failure that had gone unnoticed by management. Within days, the company closed its doors; and many bloggers lost all of their work.⁹
- March 27, 2009 – Computerworld reported that online storage service Carbonite is suing two equipment manufacturers for faulty equipment that caused Carbonite to lose the data of 7,500 customers two years ago.
- March, 2009 - The Linkup, an online storage service, permanently closed up after losing access to unspecified amounts of customer data

Several online storage services have announced closings in the recent past, often because the storage cloud business model was found to be unprofitable. However, unlike the stories of Digital Railroad, JournalSpace, and The Linkup, described above, these service providers gave sufficient advance notice so that customers could retrieve their data. Such services have included AOL (Xdrive and AOL Pictures), Hewlett-Packard (Upline), Sony (Image Station), and Yahoo (Briefcase).

Kodak is another current provider of online storage services. On its web site, it urges customers to keep a copy of each image they upload to the site in a separate and secure place.

How Secure is the Cloud?

Is it safe to keep your data in the cloud, whether it is data that you are storing in a storage cloud or data that you are generating in a compute cloud? It all depends.

From a security viewpoint, you must assume that the privacy of your data may be violated. This could happen in several ways:

- Data stored in a cloud is ripe pickings for hackers. Why waste your time on a small web site when you can have access to millions of rows of data in a cloud infrastructure?
- Data stored in a cloud is ripe picking for government authorities, who can subpoena data without a court warrant. A statute called the Stored Communications Act allows the government to require an ISP to hand over the contents of your e-mails without a

⁸ How Many 9s in Amazon?, *Availability Digest*, July 2008.

⁹ Why Back Up?, *Availability Digest*, April 2009.

warrant.¹⁰ At least if your data is on your private site, a warrant would be required in order to force you to hand it over.

True, you can protect your data from eavesdropping by encrypting it during transmission. The use of a VPN (virtual private network) to connect to the cloud would also help. Does your cloud provider support these features?

In terms of providers, who can you trust? It has been suggested that free services are probably less trustworthy than fee-based services. The reason is that free services are often paid for by reduced privacy and the right of the service provider to retain your data for an indefinite length of time. Google's Gmail service, for instance, permits automated review of the contents of e-mail for advertising purposes. A fee-based provider stands to lose a lot more than a free services provider if it is careless with its customers' data.

Even given that, do you trust your data to be stored at unknown locations managed by unknown administrators employed by a company that you know only through the web?

If you must use online storage, a good rule to consider is to not put personally-identifiable data in the cloud. Data like CRM data (à la Salesforce.com) is probably safe because it has little value to anyone else. However, data that contains personal information, credit-card data, bank-account numbers, and the like are best relegated to a popular payment service such as PayPal that guarantees privacy.¹¹

Of course, the best solution to the security problem is for a company to maintain its own private cloud.¹²

The Good and Bad of Clouds

Cloud Computing Offers Many Advantages

Running a data center is a complex and expensive business. The cloud abstracts away the complexity of the underlying infrastructure, freeing the company to focus on its applications.¹³ The "use what you need, pay for what you use" business model of the cloud has many advantages:

- It significantly reduces the time required to introduce new applications and innovations.
- It eliminates labor and capital costs for designing, procuring, building, and managing hardware and software platforms.
- It eliminates human error in the configuration of security, networks, and software.
- It provides more efficient use of computing resources, eliminating peak load performance problems.

Cloud Computing Has Major Problems

Improving Availability

History aside, the cloud can probably achieve a better uptime than you can. But be prepared – it can and will fail. What do you do to protect yourself from potentially being out of business for days due to a failed upgrade, a corrupted database, an overloaded data center, or a data-center disaster as so many other companies have experienced?

¹⁰ Mark Rasch, [Get Off My Cloud](#), *Security Focus*; August 19, 2008.

¹¹ John Benson, *Codeasaurus Rex* blog; June 24, 2007.

¹² Ron LaPedis, [Virtualization, Cloud Computing, and Business Continuity ... Oh My!](#), *Disaster Recovery Journal*; April 2009.

¹³ [Seeding the Clouds: Key Infrastructure Elements for Cloud Computing](#), *IBM White Paper*, February 2009.

If you are using a storage cloud to host your data, the answer is simple – maintain your own backup copy of the data. Using the storage cloud to host your data has the advantage that the data is available anywhere, anytime - until it isn't. Most storage cloud services allow you to download your data for backup purposes. Do so to your internal servers or perhaps to an independent storage-cloud service.

Protecting service availability is a little more difficult. One way is to run your applications in more than one compute cloud. If the data is stored in an independent storage cloud, then it is available to applications running in any other compute cloud.¹⁴ If your storage cloud data is backed up on another storage cloud, and if you can run your applications in two or more clouds, then you can achieve centuries of uptime. This approach multiplies the expense of cloud computing but is probably a fraction of the cost of building your own active/active system.

Another approach is to expect the cloud provider to offer continuous availability services, probably at an additional cost. Amazon now provides such a service via its *Availability Zones*.¹⁵ Amazon divides the world into geographic regions, each containing several Availability Zones. 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. The database in the backup instance is kept synchronized with the primary data database via data replication. Following a primary failure, the backup application instance will assume the IP addresses used by customers; and the application will be back up and running.

To ease the impact of an outage, some cloud applications such as Google's Gmail and Salesforce.com's CRM have introduced an offline capability so that users can at least perform some functions against downloaded data if the service goes down.

Improving Security

The other Achilles heel of the cloud is security. As we have discussed above, it may be wise to avoid using the cloud to store sensitive information, whether it be personal information, corporate banking-account details, competitive information, or any other kind of company-confidential information.

True, technologies such as communication encryption and virtual private networks can thwart eavesdropping of data sent between you and the cloud. But once the data is in the cloud, its privacy cannot be guaranteed. You have lost control of it. Your best bet if you must store data in the cloud is to store it in encrypted form.

Transparency

An early problem with cloud providers was that when their clouds went down, their staffs were so busy trying to recover that there was no capacity to respond to their customers as to what the situation was and when recovery could be expected. Blogs and Twitter are full of the furious messages that erupted during these periods.

Both Amazon and Salesforce.com were guilty of this in their early cloud days. Along with much improved redundancy for availability, both now maintain a status board that customers can readily access over the web to see the status of the clouds and to get updates during outages. Amazon maintains its status on its Amazon Web Services Service Health Dashboard (<http://status.aws.amazon.com/>). Salesforce.com posts its status on its Service Performance History at www.trust.salesforce.com.

¹⁴ For instance, Amazon's S3 storage cloud service allows data to be accessed by URLs.

¹⁵ Can You Trust the Compute Cloud?, *Availability Digest*, August 2008.

Service Level Agreements

What little protection you have in the cloud should certainly be the subject of a service level agreement (SLA). At the very least, the SLA should cover the following topics:¹⁶

- The amount of uptime that is guaranteed.
- The provider's response time to an outage.
- The backup policies for data – frequency, technology, and storage location.
- The provider's data security policies.
- The posting of status information concerning the health of the cloud.
- The maximum time that a response will be given to an email or a telephone query during an outage.
- The credits or remuneration that will be paid in the event of an SLA violation.

Even with a strong SLA, you should realize that it may not be of much benefit to you. Amazon's EC2 SLA (<http://aws.amazon.com/ec2-sla>) guarantees an uptime of 99.95% per year. If it violates its SLA, it will reimburse each customer 10% of the last month's payment. Look at what this really means:

- Amazon can be down four hours per year without violating the SLA.
- If you are a small user paying \$20 per month for EC2 services, and if Amazon should have a two-day outage, you will be reimbursed \$2. Does this reimburse you for two days of lost business?

When to Use the Cloud and When Not To

Russ Daniels, the VP and CTO of HP's cloud-services business, has summarized all of the above succinctly:¹⁷

"You need to be thoughtful about how you use cloud resources so that the things you do have lower risk. If it takes an extra day to run, you don't really care. Be thoughtful about where this stuff sits rather than imagining that your existing systems will be replaced by stuff in the cloud and that it will all be OK."

In short, if it's not critical, use the cloud. If it is critical, the cloud isn't for you – yet.

Summary

Because of the economies provided by the cloud, small to medium-sized businesses have been the pioneers in its use, with web hosting being the first major widespread acceptance of this technology. Larger companies are moving more slowly as they evaluate the issues associated with availability, security, and regulatory issues.

However, the business model of the cloud is so compelling that major investments are being made by service providers, equipment vendors, and consulting organizations to bring this technology to maturity. The cloud could well become the next important paradigm shift in corporate computing, eliminating the data center just as the power grid eliminated the need for each company to have its own generator.

¹⁶ Marcia Gulesian, When the Internet Fails: Application Availability, SLAs, and Disaster Recovery Planning, *Enterprise IT Planet*; September 24, 2008.

¹⁷ Outages Force Cloud Computing Users To Rethink Tactics, *Information Week*; August 16, 2008.