Yan Han

On the Clouds: A New Way of Computing

This article introduces cloud computing and discusses the author's experience "on the clouds." The author reviews cloud computing services and providers, then presents his experience of running multiple systems (e.g., integrated library systems, content management systems, and repository software). He evaluates costs, discusses advantages, and addresses some issues about cloud computing. Cloud computing fundamentally changes the ways institutions and companies manage their computing needs. Libraries can take advantage of cloud computing to start an IT project with low cost, to manage computing resources cost-effectively, and to explore new computing possibilities.

Scholarly communication and new ways of teaching provide an opportunity for academic institutions to collaborate on providing access to scholarly materials and research data. There is a growing need to handle large amounts of data using computer algorithms that presents challenges to libraries with limited experience in handling nontextual materials. Because of the current economic crisis, academic institutions need to find ways to acquire and manage computing resources in a cost-effective manner.

One of the hottest topics in IT is cloud computing. Cloud computing is not new to many of us because we have been using some of its services, such as Google Docs, for years. In his latest book, *The Big Switch: Rewiring the World, from Edison to Google*, Carr argues that computing will go the way of electricity: purchase when needed, which he calls "utility computing." His examples include Amazon's EC2 (Elastic Computing Cloud), and S3 (Simple Storage) services.[1] Amazon's chief technology officer proposed the following factors related to cloud computing: infinite computing resources available on demand, removing the need to plan ahead; the removal of an up-front costly investment, allowing companies to start small and increase resources when needed; and a system that is pay-for-use on a short-term basis and releases customers when needed (e.g., CPU by hour, storage by day).[2] National Institute of Standards and Technology (NIST) currently defines cloud computing as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. network, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."[3]

As there are several definitions for "utility computing" and "cloud computing," the author does not intend to suggest a better definition, but rather to list the characteristics of cloud computing. The term "cloud computing" means that

- customers do not own network resources, such as hardware, software, systems, or services;
- network resources are provided through remote data centers on a subscription basis; and
- network resources are delivered as services over the Web.

This article discusses using cloud computing on an IT-infrastructure level, including building virtual server nodes and running a library's essential computer systems in remote data centers by paying a fee instead of running them on-site. The article reviews current cloud computing services, presents the author's experience, and discusses advantages and disadvantages of using the new approach.

All kinds of clouds

Major IT companies have spent billions of dollars since the 1990s to shape cloud computing. For example, Sun's well-known slogan "the network is the computer" was established in late 1980s. Salesforce.com has been providing on-demand Software as a Service (SaaS) for customers since 1999. IBM and Microsoft started to deliver Web services in the early 2000s. Microsoft's Azure service provides an operating system and a set of developer tools and services. Google's popular Google Docs software provides Web-based word-processing, spreadsheet, and presentation applications. Google App Engine allows system developers to run their Python/Java applications on Google's infrastructure. Sun provides $1 per CPU hour. Amazon is well-known for providing Web services such as EC2 and S3. Yahoo! announced that it would use the Apache Hadoop framework to allow users to work with thousands of nodes and petabytes (1 million gigabytes) of data. These examples demonstrate that cloud computing providers are offering services on every level, from hardware (e.g., Amazon and Sun), to operating systems (e.g., Google and Microsoft), to software and service (e.g., Google, Microsoft, and Yahoo!). Cloud-computing providers target a variety of end users, from software developers to the general public. For additional information regarding cloud computing models, the University of California (UC) Berkeley's report provides a good comparison of these models by Amazon, Microsoft, and Google.[4]

As cloud computing providers lower prices and IT advancements remove technology barriers—such as virtualization and network bandwidth—cloud computing has moved into the mainstream.[5] Gartner stated, "Organizations are switching from

Yan Han (hany@u.library.arizona.edu) is Associate Librarian, University of Arizona Libraries, Tucson.

company-owner hardware and software to per-use service-based models."[6] For example, the U.S. government website (http://www.usa.gov/) will soon begin using cloud computing.[7] The *New York Times* used Amazon's EC2 and S3 services as well as a Hadoop application to provide open access to public domain articles from 1851 to 1922. The *Times* loaded 4 TB of raw TIFF images and their derivative 11 million PDFs into Amazon's S3 in twenty-four hours at very reasonable cost.[8] This project is very similar to digital library projects run by academic libraries. OCLC announced its movement of library management services to the Web.[9] It is clear that OCLC is going to deliver a Web-based integrated library system (ILS) to provide a new way of running an ILS. DuraSpace, a joint organization by Fedora Commons and DSpace Foundation, announced that they would be taking advantage of cloud storage and cloud computing.[10]

On the clouds

Computing needs in academic libraries can be placed into two categories: user computing needs and library goals.

User computing needs

Academic libraries usually run hundreds of PCs for students and staff to fulfill their individual needs (e.g., Microsoft Office, browsers, and image-, audio-, and video-processing applications).

Library goals

A variety of library systems are used to achieve libraries' goals to support research, learning, and teaching. These systems include the following:

- *Library website*: The website may be built on simple HTML webpages or a content management system such as Drupal, Joomla, or any home-grown PHP, Perl, ASP, or JSP system.
- *ILS*: This system provides traditional core library work such as cataloging, acquisition, reporting, accounting, and user management. Typical systems include Innovative Interfaces, SirsiDynix, Voyager, and open-source software such as Koha.
- *Repository system*: This system provides submission and access to the institution's digital collections and scholarship. Typical systems include DSpace, Fedora, EPrints, ContentDm, and Greenstone.
- *Other systems*: for example, federated search systems, learning object management systems, interlibrary loan (ILL) systems, and reference tracking systems.
- *Public and private storage*: staff file-sharing, digitization, and backup.

Due to differences in end users and functionality, most systems do not use computing resources equally. For example, the ILS is input and output intensive and database query intensive, while repository systems require storage ranging from a few gigabytes to dozens of terabytes and substantial network bandwidth.

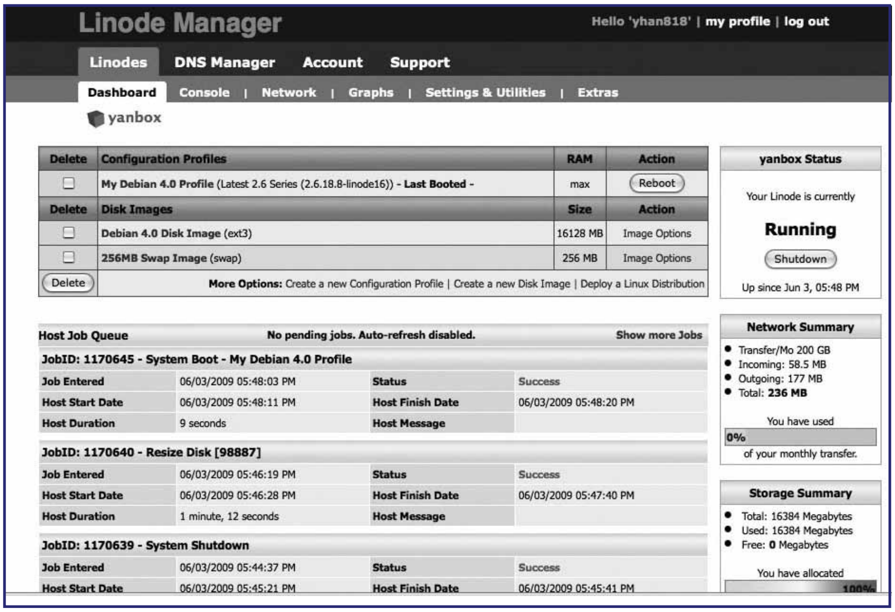
Cloud computing brings a fundamental shift in computing. It changes the way organizations acquire, configure, manage, and maintain computing resources to achieve their business goals. The availability of cloud computing providers allows organizations to focus on their business and leave general computing maintenance to the major IT companies. In the fall of 2008, the author started to research cloud computing providers and how he could implement cloud computing for some library systems to save staff and equipment costs. In January 2009, the author started his plan to build library systems "on the clouds."

The University of Arizona Libraries (UAL) has been a key player in the process of rebuilding higher education in Afghanistan since 2001. UAL Librarian Atifa Rawan and the author have received multiple grant contracts to build technical infrastructures for Afghanistan's academic libraries. The technical infrastructure includes the following:

- Afghanistan ILS: a bilingual ILS based on the open-source system Koha.[11]
- Afghanistan Digital Libraries website (http://www.afghandigitallibraries.org/): originally built on simple HTML pages, later rebuilt in 2008 using the content management system Joomla.
- A digitization management system.

The author has also developed a Japanese ILL system (http://gifproject.libraryfinder.org) for the North American Coordinating Council on Japanese Library Resources. These systems had been running on UAL's internal technical infrastructure. These systems run in a complex computing environment, require different modules, and do not use computing resources equally. For example, the Afghan ILS runs on Linux, Apache, MySQL, and Perl. Its OPAC and staff interface run on two different ports. The Afghanistan Digital Libraries website requires Linux, Apache, MySQL, and PHP. The Japanese ILL system was written in Java and runs on Tomcat. There are several reasons why the author moved these systems to the new cloud computing infrastructure:

- These systems need to be accessed in a system mode by people who are not UAL employees.
- System rebooting time can be substantial in this infrastructure because of server setup and IT policy.
- The current on-site server has



Linode Manager
Hello 'yhan818' | my profile | log out
Linodes
DNS Manager
Account
Support
Dashboard
Console | Network | Graphs | Settings & Utilities | Extras
yanbox

Delete	Configuration Profiles	RAM	Action
☐	**My Debian 4.0 Profile** (Latest 2.6 Series (2.6.18.8-linode16)) - **Last Booted -**	max	Reboot
Delete	**Disk Images**	**Size**	**Action**
☐	**Debian 4.0 Disk Image** (ext3)	16128 MB	Image Options
☐	**256MB Swap Image** (swap)	256 MB	Image Options
Delete	**More Options:** Create a new Configuration Profile | Create a new Disk Image | Deploy a Linux Distribution		

Host Job Queue **No pending jobs. Auto-refresh disabled.** **Show more Jobs**

JobID: 1170645 - System Boot - My Debian 4.0 Profile

Job Entered	06/03/2009 05:48:03 PM	**Status**	Success
Host Start Date	06/03/2009 05:48:11 PM	**Host Finish Date**	06/03/2009 05:48:20 PM
Host Duration	9 seconds	**Host Message**	

JobID: 1170640 - Resize Disk [98887]

Job Entered	06/03/2009 05:46:19 PM	**Status**	Success
Host Start Date	06/03/2009 05:46:28 PM	**Host Finish Date**	06/03/2009 05:47:40 PM
Host Duration	1 minute, 12 seconds	**Host Message**	

JobID: 1170639 - System Shutdown

Job Entered	06/03/2009 05:44:37 PM	**Status**	Success
Host Start Date	06/03/2009 05:45:21 PM	**Host Finish Date**	06/03/2009 05:45:41 PM

yanbox Status

Your Linode is currently

Running

Shutdown

Up since Jun 3, 05:48 PM

Network Summary

- Transfer/Mo 200 GB
- Incoming: 58.5 MB
- Outgoing: 177 MB
- Total: **236 MB**

You have used

0%

of your monthly transfer.

Storage Summary

- Total: 16384 Megabytes
- Used: 16384 Megabytes
- Free: **0** Megabytes

You have allocated

100%

Figure 1. Linux Node Administration Web interface

reached its life expectancy and requires a replacement.

By analyzing the complex needs of different systems and considering how to use resources more effectively, the author decided to run all the systems through one cloud computing provider. By comparing the features and the costs, Linode (http://www.linode.com/) was chosen because it provides full SSH and root access using virtualization, four data centers in geographically diverse areas, high availability and clustering support, and an option for month-to-month contracts. In addition, other customers have provided positive reviews. In January 2009, the author purchased one node located in Fremont, California, for $19.95 per month. An implementation plan (see appendix) was drafted to complete the project in phases. The author owns a virtual server and has access to everything that a physical server provides. In addition, the provider and the user community provided timely help and technical support.

The migration of systems was straightforward: A Linux kernel (Debian 4.0) was installed within an hour, domain registration was complete and the domains went active in twenty-four hours, the Afghanistan Digital Libraries' website (based on Joomla) migration was complete within a week, and all supporting tools and libraries (e.g., MySQL, Tomcat, and Java SDK) were installed and configured within a few days. A month later, the Afghanistan ILS (based on Koha) migration was completed. The ILL system was also migrated without problem. Tests have been performed in all these systems to verify their usability. In summary, the migration of systems was very successful and did not encounter any barriers. It addresses the issues facing us: After the migration, SSH log-ins for users who are not university employees were set up quickly; systems maintenance is managed by the author's team, and rebooting now only takes about one minute; and there is no need to buy a new server and put it in a temperature and security controlled environment. The hardware is maintained by the provider.

The administrative GUI for the Linux Nodes is shown in figure 1.

Since migration, no downtime because of hardware or other failures caused by the provider has been observed. After migrating all the systems successfully and running them in a reliable mode for a few months, the second phase was implemented (see appendix). Another Linux node (located in Atalanta, Georgia) was purchased for backup and monitoring (see figure 2). Nagios, an open-source monitoring system, was tested and configured to identify and report problems for the above library systems. Nagios provides the following functions: (1) monitoring critical computing components, such as the network, systems, services, and servers; (2) timely alerts delivered via e-mail or cell phone; and (3) report and record logs of outages, events, and alerts. A backup script is also run as a prescheduled job to back up the systems on a regular basis.

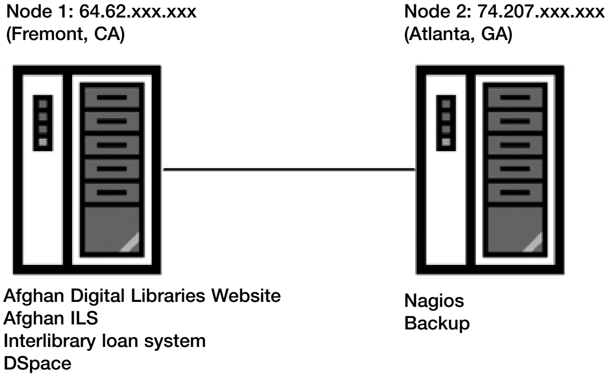


Figure 2. Two Linux Nodes located in two remote data centers

Findings and discussions

Since January 2009, all the systems have been migrated and have been running without any issues caused by the provider. The author is very satisfied with the outcomes and cost. The annual cost of running two nodes is $480 per year, compared to at least $4,000 dollars if the hardware had been run in the library.[12]

From the author's experience, cloud computing provides the following advantages over the traditional way of computing in academic institutions:

- *Cost-effectiveness*: From the above example and literature review, it is obvious that using cloud computing to run applications, systems, and IT infrastructure saves staff and financial resources. UC Berkeley's report and Zawodny's blog provide a detailed analysis of costs for CPU hours and disk storage.[13]
- *Flexibility*: Cloud computing allows organizations to start a project quickly without worrying about up-front costs. Computing resources such as disk storage, CPU, and RAM can be added when needed. In this case, the author started on a small scale by purchasing one node and added additional resources later.
- *Data safety*: Organizations are able to purchase storage in data centers located thousands of miles away, increasing data safety in case of natural disasters or other factors. This strategy is very difficult to achieve in a traditional off-site backup.
- *High availability*: Cloud computing providers such as Microsoft, Google, and Amazon have better resources to provide more up-time than almost any other organizations and companies do.
- *The ability to handle large amounts of data*: Cloud computing has a pay-for-use business model that allows academic institutions to analyze terabytes of data using distributed computing over hundreds of computers for a short-time cost.

On-demand data storage, high availability and data safety are critical features for academic libraries.[14] However, readers should be aware of some technical and business issues:

- *Availability of a service*: In several widely reported cases, Amazon's S3 and Google Gmail were inaccessible for a duration of several hours in 2008. The author believes that the commercial providers have better technical and financial resources to keep more up-time than most academic institutions. For those wanting no single point of failure (e.g., a provider goes out of business), the author suggests storing duplicate data with a different provider or locally.
- *Data confidentiality*: Most academic libraries have open-access data. This issue can be solved by encrypting data before moving to the clouds. In addition, licensing terms can be negotiated with providers regarding data safety and confidentiality.
- *Data transfer bottlenecks*: Accessing the digital collections requires considerable network bandwidth, and digital collections are usually optimized for customer access. Moving huge amounts of data (e.g., preservation digital images, audios, videos, and data sets) to data centers can be scheduled during off hours (e.g., 1–5 a.m.), or data can be shipped on hard disks to the data centers.
- *Legal jurisdiction*: Legal jurisdiction creates complex issues for both providers and end users. For example, Canadian privacy laws regulate data privacy in public and private sectors. In 2008, the Office of the Privacy Commissioner of Canada released a finding that "outsourcing of canada.com email services to U.S.-based firm raises questions for subscribers," and expressed concerns about public sector privacy protection.[15] This brings concerns to both providers and end users, and it was suggested that privacy issues will be very challenging.[16]

Summary

The author introduces cloud computing services and providers, presents his experience of running multiple systems such as ILS, content management systems, repository software, and the other system "on the clouds" since January 2009. Using cloud computing brings significant cost savings and flexibility. However, readers should be aware of technical and business issues. The author is very satisfied with his experience of moving library systems to cloud computing. His experience demonstrates a new way of managing critical computing resources in an academic library setting. The next steps include using cloud computing to meet digital collections' storage needs.

Cloud computing brings fundamental changes to organizations managing their computing needs. As major organizations in library fields, such as OCLC, started to take advantage of cloud computing, the author believes that cloud computing will play an important role in library IT.

Acknowledgments

The author thanks USAID and Washington State University for providing financial support. The author thanks Matthew Cleveland's excellent work "on the clouds."

References

1. Nicholars Carr, *The Big Switch: Rewiring the World, from Edison to Google*

(London: Norton, 2008).

2. Werner Vogels, "A Head in the Clouds—The Power of Infrastructure as a Service" (paper presented at the Cloud Computing and in Applications conference (CCA '08), Chicago, Oct. 22–23, 2008).

3. Peter Mell and Tim Grance, "Draft NIST Working Definition of Cloud Computing," National Institute of Standards and Technology (May 11, 2009), http://csrc.nist.gov/groups/SNS/cloud-computing/index.html (accessed July 22, 2009).

4. Michael Armbust et al., "Above the Clouds: A Berkeley View of Cloud Computing," technical report, University of California, Berkeley, EECS Department, Feb. 10, 2009, http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html (accessed July 1, 2009).

5. Eric Hand, "Head in the Clouds: 'Cloud Computing' Is Being Pitched as a New Nirvana for Scientists Drowning in Data. But Can It Deliver?" *Nature* 449, no. 7165 (2007): 963; Geoffery Fowler and Ben Worthen, "The Internet Industry Is On a Cloud—Whatever That May Mean," *Wall Street Journal*, Mar. 26, 2009, http://online.wsj.com/article/SB123802623665542725.html (accessed July 14, 2009); Stephen Baker, "Google and the Wisdom of the Clouds," *Business Week* (Dec. 14, 2007), http://www.msnbc.msn.com/id/22261846/ (accessed July 8, 2009).

6. Gartner, "Gartner Says Worldwide IT Spending on Pace to Supass $3.4 Trillion in 2008," press release, Aug. 18, 2008, http://www.gartner.com/it/page.jsp?id=742913 (accessed July 7, 2009).

7. Wyatt Kash, "USA.gov, Gobierno USA.gov move into the Internet cloud," *Government Computer News*, Feb. 23, 2009, http://gcn.com/articles/2009/02/23/gsa-sites-to-move-to-the-cloud.aspx?s=gcndaily_240209 (accessed July 14, 2009).

8. Derek Gottfrid, "Self-Service, Prorated Super Computing Fun!" online posting, New York Times Open, Nov. 1, 2007, http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/?scp=1&sq=self%20service%20prorated&st=cse (accessed July 8, 2009).

9. OCLC Online Computing Library Center, "OCLC announces strategy to move library management services to Web scale," press release, Apr. 23, 2009, http://www.oclc.org/us/en/news/releases/200927.htm (accessed July 5, 2009).

10. DuraSpace, "Fedora Commons and DSpace Foundation Join Together to Create DuraSpace Organization," press release, May 12, 2009, http://duraspace.org/documents/pressrelease.pdf (accessed July 8, 2009).

11. Yan Han and Atifa Rawan, "Afghanistan Digital Library Initiative: Revitalizing an Integrated Library System," *Information Technology & Libraries* 26, no. 4 (2007): 44–46.

12. Fowler and Worthen, "The Internet Industry Is on a Cloud."

13. Jeremy Zawodney, "Replacing My Home Backup Server with Amazon's S3," online posting, Jeremy Zawodny's Blog, Oct. 3, 2006, http://jeremy.zawodny.com/blog/archives/007624.html (accessed June 19, 2009).

14. Yan Han, "An Integrated High Availability Computing Platform," *The Electronic Library* 23, no. 6 (2005): 632–40.

15. Office of the Privacy Commissioner of Canada, "Tabling of Privacy Commissioner of Canada's 2005–06 Annual Report on the Privacy Act: Commissioner Expresses Concerns about Public Sector Privacy Protection," press release, June 20, 2006, http://www.priv.gc.ca/media/nr-c/2006/nr-c_060620_e.cfm (accessed July 14, 2009); Office of the Privacy Commissioner of Canada, "Findings under the Personal Information Protection and Electronic Documents Act (PIPEDA)," (Sept. 19, 2008), http://www.priv.gc.ca/cf-dc/2008/394_20080807_e.cfm (accessed July 14, 2009).

16. Stephen Baker, "Google and the Wisdom of the Clouds," *Business Week* (Dec. 14, 2007), http://www.msnbc.msn.com/id/22261846/ (accessed July 8, 2009).

Appendix. Project Plan: Building HA Linux Platform Using Cloud Computing

Project Manager:

Project Members:

Object Statement: To build a High Availability (HA) Linux platform to support multiple systems using cloud computing in six months.

Scope: The project members should identify cloud computing providers, evaluate the costs, and build a Linux platform for computer systems, including Afghan ILS, Afghanistan Digital Libraries website, repository system, Japanese interlibrary loan website, and digitization management system.

Resources:

Project Deliverable: January 1, 2009—July 1, 2009

Appendix. Project Plan: Building HA Linux Platform Using Cloud Computing (cont.)

Phase I

- To build a stable and reliable Linux Platform to support multiple Web applications. The platform needs to consider reliability and high availability in a cost-effective manner
- To install needed libraries for the environment
- To migrate ILS (Koha) to this Linux platform
- To migrate Afghan Digital Libraries' website (Joomla) to this platform
- To migrate Japanese interlibrary loan website
- To migrate Digitization Management system

Phase II

- To research and implement a monitoring tool to monitor all Web applications as well as OS level tools (e.g. Tomcat, MySQL)
- To configure a cron job to run routine things (e.g., backup)
- To research and implement storage (TB) for digitization and access

Phase III

- To research and build Linux clustering

Steps:

1. OS installation: Debian 4
2. Platform environment: Register DNS
3. Install Java 6, Tomcat 6, MySQL 5, etc.
4. Install source control env Git
5. Install statistics analysis tool (Google Analytics)
6. Install monitoring tool: Ganglia or Nagios
7. Web Applications
8. Joomla
9. Koha
10. Monitoring tool
11. Digitization management system
12. Repository system: Dspace, Fedora, etc.
13. HA tools/applications

Note

Calculation based on the following:

- leasing two nodes $20/month: $20 x 2 nodes x 12 months = $480/year
- A medium-priced server with backup with a life expectancy of 5 years ($5,000): $1,000/year
- 5 percent of system administrator time for managing the server ($60,000 annual salary): $3,000/year
- Ignore telecommunication cost, utility cost, and space cost.
- Ignore software developer's time because it is equal for both options.