



Cloud Computing

Authors: Greg Boss, Padma Malladi, Dennis Quan, Linda Legregni, Harold Hall,
Management Contact: Dennis Quan
Organization: High Performance On Demand Solutions (HiPODS)
Web address: www.ibm.com/developerworks/websphere/zones/hipods/
Date: 8 October 2007
Status: Version 1.0

Abstract: This paper describes cloud computing, a computing platform for the next generation of the Internet. The paper defines clouds, explains the business benefits of cloud computing, and outlines cloud architecture and its major components. Readers will discover how a business can use cloud computing to foster innovation and reduce IT costs. IBM's implementation of cloud computing is described.

To receive information about future workshops and seminars pertaining to this solution, send an email to hipods@us.ibm.com.

Executive summary

Innovation is necessary to ride the inevitable tide of change. Indeed, the success of the transformation of IBM to an On Demand Business depends on driving the right balance of productivity, collaboration, and innovation to achieve sustained, organic top line growth — and bottom line profitability.

Enterprises strive to reduce computing costs. Many start by consolidating their IT operations and later introducing virtualization technologies. Cloud computing takes these steps to a new level and allows an organization to further reduce costs through improved utilization, reduced administration and infrastructure costs, and faster deployment cycles. The cloud is a next generation platform that provides dynamic resource pools, virtualization, and high availability.

Cloud computing describes both a platform and a type of application. A cloud computing platform dynamically provisions, configures, reconfigures, and deprovisions servers as needed. Cloud applications are applications that are extended to be accessible through the Internet. These *cloud applications* use large data centers and powerful servers that host Web applications and Web services.

Cloud computing infrastructure accelerates and fosters the adoption of innovations

Enterprises are increasingly making innovation their highest priority. They realize they need to seek new ideas and unlock new sources of value. Driven by the pressure to cut costs and grow—simultaneously—they realize that it's not possible to succeed simply by doing the same things better. They know they have to do new things that produce better results.

Cloud computing enables innovation. It alleviates the need of innovators to find resources to develop, test, and make their innovations available to the user community. Innovators are free to focus on the innovation rather than the logistics of finding and managing resources that enable the innovation. Cloud computing helps leverage innovation as early as possible to deliver business value to IBM and its customers.

Fostering innovation requires unprecedented flexibility and responsiveness. The enterprise should provide an ecosystem where innovators are not hindered by excessive processes, rules, and resource constraints. In this context, a cloud computing service is a necessity. It comprises an automated framework that can deliver standardized services quickly and cheaply.

Cloud computing infrastructure allows enterprises to achieve more efficient use of their IT hardware and software investments

Cloud computing increases profitability by improving resource utilization. Pooling resources into large clouds drives down costs and increases utilization by delivering resources only for as long as those resources are needed. Cloud computing allows individuals, teams, and organizations to streamline procurement processes and eliminate the need to duplicate certain computer administrative skills related to setup, configuration, and support.

This paper introduces the value of implementing cloud computing. The paper defines clouds, explains the business benefits of cloud computing, and outlines cloud architecture and its major components. Readers will discover how a business can use cloud computing to foster innovation and reduce IT costs.

Contents

| | |
|---|----|
| Executive summary | 2 |
| Contents..... | 3 |
| What is a cloud?..... | 4 |
| Definition | 4 |
| Benefits | 4 |
| Usage scenarios | 5 |
| Architecture | 6 |
| Cloud provisioning and management..... | 7 |
| Automated provisioning | 7 |
| Reservation and scheduling | 8 |
| Change management | 8 |
| Monitoring | 10 |
| Open source | 12 |
| Virtualization..... | 12 |
| Storage architecture in the cloud | 13 |
| Piloting innovations on a cloud | 14 |
| Conclusion..... | 15 |
| References..... | 16 |
| Acknowledgements | 16 |
| Notices | 17 |

What is a cloud?

Cloud computing is a term used to describe both a platform and type of application. A cloud computing platform dynamically provisions, configures, reconfigures, and deprovisions servers as needed. Servers in the cloud can be physical machines or virtual machines. Advanced clouds typically include other computing resources such as storage area networks (SANs), network equipment, firewall and other security devices.

Cloud computing also describes applications that are extended to be accessible through the Internet. These *cloud applications* use large data centers and powerful servers that host Web applications and Web services. Anyone with a suitable Internet connection and a standard browser can access a cloud application.

Definition

A cloud is a pool of virtualized computer resources. A cloud can:

- Host a variety of different workloads, including batch-style back-end jobs and interactive, user-facing applications
- Allow workloads to be deployed and scaled-out quickly through the rapid provisioning of virtual machines or physical machines
- Support redundant, self-recovering, highly scalable programming models that allow workloads to recover from many unavoidable hardware/software failures
- Monitor resource use in real time to enable rebalancing of allocations when needed

Cloud computing environments support grid computing by quickly providing physical and virtual servers on which the grid applications can run. Cloud computing should not be confused with grid computing. Grid computing involves dividing a large task into many smaller tasks that run in parallel on separate servers. Grids require many computers, typically in the thousands, and commonly use servers, desktops, and laptops.

Clouds also support nongrid environments, such as a three-tier Web architecture running standard or Web 2.0 applications. A cloud is more than a collection of computer resources because a cloud provides a mechanism to manage those resources. Management includes provisioning, change requests, reimaging, workload rebalancing, deprovisioning, and monitoring.

Benefits

Cloud computing infrastructures can allow enterprises to achieve more efficient use of their IT hardware and software investments. They do this by breaking down the physical barriers inherent in isolated systems, and automating the management of the group of systems as a single entity. Cloud computing is an example of an ultimately virtualized system, and a natural evolution for data centers that employ automated systems management, workload balancing, and virtualization technologies.

A cloud infrastructure can be a cost efficient model for delivering information services, reducing IT management complexity, promoting innovation, and increasing responsiveness through real-time workload balancing.

The Cloud makes it possible to launch Web 2.0 applications quickly and to scale up applications as much as needed when needed. The platform supports traditional Java™ and Linux, Apache, MySQL, PHP (LAMP) stack-based applications as well as new architectures such as MapReduce

and the Google File System, which provide a means to scale applications across thousands of servers instantly.

Large amounts of computer resource, in the form of Xen virtual machines, can be provisioned and made available for new applications within minutes instead of days or weeks. Developers can gain access to these resources through a portal and put them to use immediately. Several products are available that provide virtual machine capabilities, including proprietary ones such as VMware, and open source alternatives, such as XEN. This paper describes the use of XEN virtualization.

Many customers are interested in cloud infrastructures to serve as platforms for innovation, particularly in countries that want to foster the development of a highly skilled, high-tech work force. They want to provide startups and research organizations with an environment for idea exchange, and the ability to rapidly develop and deploy new product prototypes.

In fact, HiPODS has been hosting IBM's innovation portal on a virtualized cloud infrastructure in our Silicon Valley Lab for nearly two years. We have over seventy active innovations at a time, with each innovation lasting on average six months. 50% of those innovations are Web 2.0 projects (search, collaboration, and social networking) and 27% turn into products or solutions. Our success with the innovation portal is documented in the August 20 *Business Week* cover story on global collaboration.

Usage scenarios

Cloud computing can play a significant role in a variety of areas including internal pilots, innovations, virtual worlds, e-business, social networks, and search. Here we summarize several basic but important usage scenarios that highlight the breadth and depth of impact that cloud computing can have on an enterprise.

Internal innovation

Innovators request resources online through a simple Web interface. They specify a desired start and end dates for their pilot. A cloud resource administrator approves or rejects the request. Upon approval, the cloud provisions the servers. The innovator has the resources available for use within a few minutes or an hour depending on what type of resource was requested.

Virtual worlds

Virtual worlds require significant amounts of computing power, especially as those virtual spaces become large or as more and more users log in. Massively multiplayer online games (MMPOG) are a good example of significantly large virtual worlds. Several commercial virtual worlds have as many as nine million registered users and hundreds and thousands of servers supporting these environments.

A company that hosts a virtual world could have real time monitors showing the utilization level of the current infrastructure or the average response time of the clients in any given 'realm' of the virtual world. Realms are arbitrary areas within a virtual world that support a specific subset of people or subset of the world. The company discovers that realm A has an significant increase in use and the response times are declining, whereas realms S and Z have decreased in use. The company initiates a cloud rebalance request to deprovision five servers each from realms S and Z and provision ten servers to Realm A. After a couple of minutes the ten servers are relocated without interruption to any users in any of the realms and the response time for realm A has returned to acceptable levels. The company has achieved significant cost savings by reusing

underutilized equipment, maintained high customer satisfaction, avoided help desk calls from users and completed in minutes what would previously have taken days or weeks to accomplish.

e-business

In e-business, scalability can be achieved by making new servers available as needed. For example, during a peak shopping season, more virtual servers can be made available that can cater to high shopper demand. In another example a company may experience high workloads on weekends or evenings as opposed to early mornings and weekdays. If a company has a significantly large cloud, they could schedule computer resources to be provisioned each evening, weekend, or during a peak season. There are more opportunities to achieve efficiencies as the cloud grows. Another aspect of this scenario involves employing business policies to decide what applications receive higher priorities and thus more computing resources. Revenue generating applications may be rated higher than research and development or innovation pilots. For several months IBM has been running a cloud infrastructure that adjusts computer resources appropriately and automatically according to business policies.

Personal hobbies

Innovation is no longer a concept developed and owned by companies and businesses. It is becoming popular at the individual level, and more individuals are coming up with innovations. These individuals could be requesting servers from a cloud to work on their innovations.

Architecture

Figure 1 illustrates the high level architecture of the cloud computing platform. It's comprised of a data center, IBM® Tivoli® Provisioning Manager, IBM® Tivoli® Monitoring, IBM® Websphere® Application Server, IBM® DB2®, and virtualization components. This architecture diagram focuses on the core back end of the cloud computing platform; it does not address the user interface.

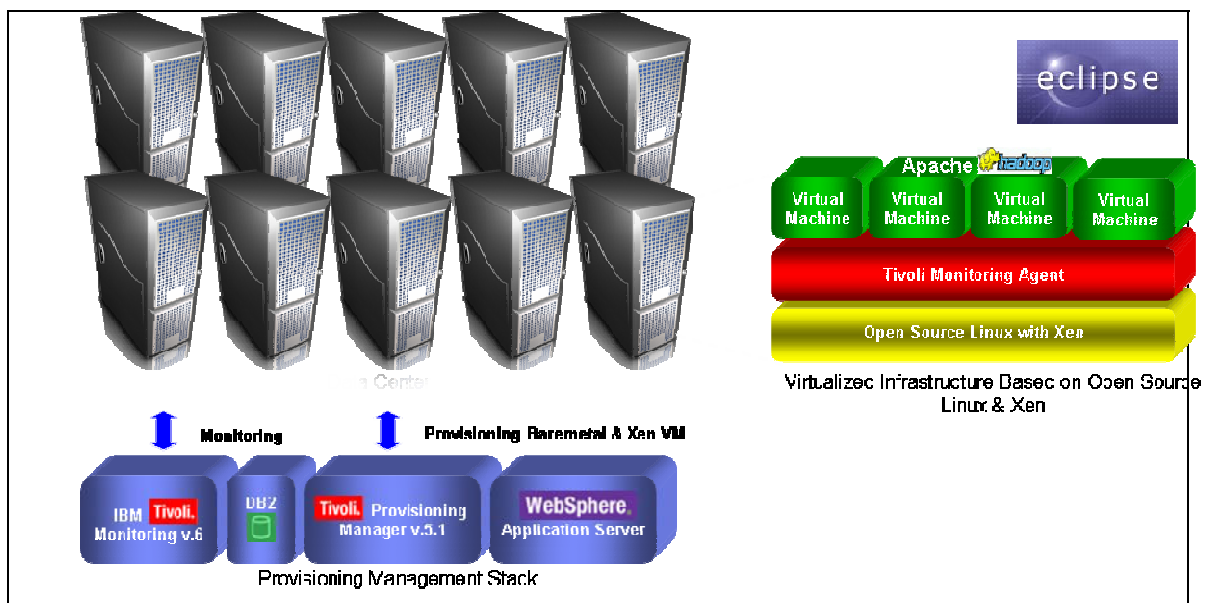


Figure 1. High level cloud architecture

Tivoli Provisioning Manager automates imaging, deployment, installation, and configuration of the Microsoft Windows and Linux operating systems, along with the installation / configuration of any software stack that the user requests.

Tivoli Provisioning Manager uses WebSphere Application Server to communicate the provisioning status and availability of resources in the data center, to schedule the provisioning and deprovisioning of resources, and to reserve resources for future use.

As a result of the provisioning, virtual machines are created using the XEN hypervisor or physical machines are created using Network Installation Manager, Remote Deployment Manager, or Cluster Systems Manager, depending upon the operating system and platform.

IBM Tivoli Monitoring Server monitors the health (CPU, disk, and memory) of the servers provisioned by Tivoli Provisioning Manager.

DB2 is the database server that Tivoli Provisioning Manager uses to store the resource data.

IBM Tivoli Monitoring agents that are installed on the virtual and physical machines communicate with the Tivoli Monitoring server to get the health of the virtual machines and provide the same to the user.

The cloud computing platform has two user interfaces to provision servers.

- One interface is feature rich -- fully loaded with the WebSphere suite of products -- and relatively more involved from a process perspective. For more information on this interface, see *Cloud provisioning and management*.
- One interface provides basic screens for making provisioning requests.

All requests are handled by Web2.0 components deployed on the WebSphere Application Server. Requests are forwarded to Tivoli Provisioning Manager for provisioning/deprovisioning servers.

Cloud provisioning and management

Automated provisioning

The core functionality of a cloud is its ability to automatically provision servers for innovators and to enable innovators, administrators, and others to use that function with a Web-based interface. The role-based interface abstracts out the complexity of IBM Tivoli Provisioning Manager, Remote Deployment Manager, Network Installation Manager, business process execution language (BPEL), and Web services.

Typically, a pilot team needs four to twelve weeks to identify, procure, and build a pilot infrastructure and additional time to build a security compliant software stack so that developers can begin building or deploying applications and code. The cloud provides a framework and offering that reduces that boarding process to approximately one hour.

We accomplish this through a role-based Web portal that allows innovators to fill out a form defining their hardware platform, CPU, memory, storage, operating system, middleware, and team members and associated roles. This process takes about five minutes. After submitting the request through the portal, a cloud administrator is notified and logs in to approve, modify, and/or

reject the request. If approved, the system begins a process involving Web services, Tivoli Provisioning Manager, and optionally IBM Tivoli Security Compliance Manager, BPEL, IBM® Enterprise Workload Manager, and Remote Deployment Manager/Cluster Systems Management/Network Installation Manager to build the server(s). This process is fully automatic and completes in about an hour.

The value of having a fully automated provisioning process that is security compliant and automatically customized to innovators' needs is manifested in reduced time to introduce technologies and innovations, cost savings in labor for designing, procuring, and building hardware and software platforms, and cost avoidance in higher use and reuse of existing resources.

Reservation and scheduling

Critical to an environment like a cloud is the ability to understand what your current and future capacity is to accommodate customers. Without that understanding you cannot accurately forecast how many customers you can support, nor can you ensure that you maintain a steady pipeline of innovation. For this reason, projects cannot board the cloud without an agreed upon end date. This date which is part of the contract (an approved request for resources) offers an incentive for the project team to work aggressively to meet their target or risk the removal of the resource assigned. Contractual end dates also allow cloud administrators to accurately schedule resources for future dates.

For this purpose the cloud also requires a contractual start date so that it is possible to reserve resources for a future time. Contract start dates give innovators an accurate expectation of when they will receive their approved resources. The reservation system in the cloud provides a system of checks and balances so that new reservations cannot be approved for resources that don't exist or that are already committed for the specified time frame.

Change management

Extending contracts

As with most innovations, projects, pilots, and prototypes often have unknown elements related to development delays, new requirements, and funding. The unknowns sometimes make it difficult to adhere to a contractual end date, especially if the end date was agreed to several months before the project delays. For this reason, the cloud allows innovators to request extensions to their original contract end date.

Authorized members of the project can log in to the cloud portal and request a contract end date extension (see Figure 2). This request is evaluated by the cloud administrator for both resource capacity and business justification. Availability of resources is revealed to the administrator through the cloud portal. Business justification is a human task that corresponds with the same BPEL approval step.

When an extend contract request is made, the administrator can log in to the Web interface and view the request. Upon approval of the new date, the appropriate BPEL task is executed and the contract is updated to reflect the new date. A grace period is implemented so that any implementation of the Cloud can define a several day or week buffer at the end of a project. This allows for flexibility in granting approvals of minor extend contract requests while still maintaining a structured environment that facilitates long term resource management.


| | |
|----------------------|---|
| Project Name: | XL Software |
| State: | Active |
| Status: | Service Provision Completed |
| Request Date: | 09-10-2007 |
| Start Date: | 06-13-2007 |
| End Date: | 02-05-2008 extend to: <input type="text"/>  Extend contract date |

Figure 2. Changing end date

Changing contracts

The cloud has many types of innovative pilots and prototypes and is specifically designed to be as flexible and accommodating as possible. It is not uncommon for a risky pilot using emerging or untested technologies to need to wipe the server clean and start fresh, or in some cases to discover midprocess that they need additional middleware or an alternative operating system. The cloud can accommodate this need with the *change contract* function.

Change contract is flexible enough to allow the innovator to add a software component or to start over. It also allows an innovator to add or remove servers to an existing project or contract and on capable hardware systems like IBM System p™ or Xen it allows the innovator to increase or decrease the amount of RAM or storage assigned to one or more LPARs or virtual machines. Change contract is automated and carries all the advantages of cloud provisioning: it requires no manual administrative support, is carried out in an hour or less, can be requested at anytime, and has its interface through the cloud portal.

Figure 3 shows the interface that an innovator uses to begin the change contract request.

Project Name: XL Software

| General info | Server info | Pricing info | Contacts | History | | | | | |
|--|---------------|---------------|-------------|-------------------|-------------|---|------------|-------------|----------|
| Server Type pSeries P595 LPAR | Serial Number | Host Name | Port Number | IP Address | No. of CPUs | CPU Speed(MHz) | Memory(GB) | Storage(GB) | Priority |
| | | tdlpl31 | | XXXXXXXXXX | 4.0 | 1130 | 2 | 100 | Gold |
| | Software Name | Software Type | Comments | Monitoring Status | | | | | |
| | AIX 5.3 | OS | | | | <div>CPU Memory Disk</div> <div><div></div><div></div><div></div></div> | | | |
| Server Type pSeries P595 LPAR | Serial Number | Host Name | Port Number | IP Address | No. of CPUs | CPU Speed(MHz) | Memory(GB) | Storage(GB) | Priority |
| | | tdlpl43 | | XXXXXXXXXX | 4.0 | 1130 | 2 | 25 | Gold |
| | Software Name | Software Type | Comments | Monitoring Status | | | | | |
| | AIX 5.3 | OS | | | | <div>CPU Memory Disk</div> <div><div></div><div></div><div></div></div> | | | |
| Server Type pSeries P595 LPAR | Serial Number | Host Name | Port Number | IP Address | No. of CPUs | CPU Speed(MHz) | Memory(GB) | Storage(GB) | Priority |
| | | tdlpl19 | | XXXXXXXXXX | 4.0 | 1130 | 2 | 100 | Gold |
| | Software Name | Software Type | Comments | Monitoring Status | | | | | |
| | AIX 5.3 | OS | | | | <div>CPU Memory Disk</div> <div><div></div><div></div><div></div></div> | | | |
| <div>Manage Servers</div> <div>Set Monitoring Thresholds</div> | | | | | | | | | |

Figure 3. Interface to start the change contract process

An innovator can choose to:

- Delete a server
- Add software to a server
- Reimage a server with a new operating system
- Change the memory or storage allocation of a server
- Add a server

The cloud interface is dynamic and changes itself to offer the function available according to the task chosen. See Figure 4. All changes can follow a BPEL process, which requires an administrator's approval. When the reservation is verified (if new servers are added) and approval granted the change contract request is executed through Tivoli Provisioning Manager and Web services calls are sent to the hardware management console (HMC) (if memory or storage changes are required). A lightweight implementation of the cloud can forgo BPEL implementation and operate without the approval step. In this situation typically only administrators are allowed to access the Web interface.

Figure 4. Interface to change a contract

Monitoring

Clouds typically have a significant number of servers, As the number of cloud resources increase monitoring becomes a critical requirement. The cloud includes capabilities for monitoring both individual servers and collections of servers.

Monitoring is performed using IBM® Tivoli® Monitoring. This involves installing an IBM Tivoli Monitoring agent on each cloud server and configuring the IBM Tivoli Monitoring server. The agents collect information from the cloud resource and periodically transfer that data to the monitoring data warehouse, which is an IBM® DB/2® database. The monitoring server contains three components; IBM® Tivoli® Enterprise Monitoring, IBM® Tivoli® Enterprise Portal , and the data warehouse.

As shown in Figure 5, detailed information on each monitored resource can be viewed with Tivoli Enterprise portals and can be fully integrated with the cloud portal.

| Name | State | Status | Start Date | End Date | Submitter | Customer | Health C M D |
|----------------------------|-----------|-------------------------------|------------|------------|--------------------|---|--|
| TDIL Support Server | Active | Deploy Successful | 09-10-2007 | 12-31-2008 | Greg Boss The Man | HiPODS | <div><div></div><div></div><div></div></div> |
| SOAR | Completed | Service Deprovision Completed | 01-11-2006 | 09-14-2007 | Greg Boss | Bruce Besch | |
| ADAPT | Active | Deploy Successful | 11-06-2006 | 03-07-2008 | Chris Wyble | ADAPT | <div><div></div><div></div><div></div></div> |
| Avatar Service Framework | Active | Deploy Successful | 09-06-2007 | 03-04-2008 | Milton Bonilla | IBM Internal, Virtual World, Avatar Service | <div><div></div><div></div><div></div></div> |
| windows-2003-sp2-test1 | Completed | Service Deprovision Completed | 09-04-2007 | 09-11-2007 | hdil hdil | rpatnani@us.ibm.com | |
| GTO - VBO | Active | Deploy Successful | 07-02-2007 | 12-14-2007 | Gregory Vilshansky | IBM Sales Learning | <div><div></div><div></div><div></div></div> |
| test | Completed | Service Deprovision Completed | 08-29-2007 | 09-05-2007 | Milton Bonilla | test | |
| PassItOn | Active | Deploy Successful | 10-05-2006 | 01-18-2008 | Chris Wyble | PassItOn | <div><div></div><div></div><div></div></div> |
| Client Information Project | Active | Deploy Successful | 03-30-2006 | 01-25-2008 | Pandya Aroop | CIO | <div><div></div><div></div><div></div></div> |

Figure 5. Projects portlet showing all contracts

Summary information denoting server health can be viewed directly from the cloud portal. Figure 6 shows CPU, memory, and disk summary information that is consolidated at a project or pilot level where a project can contain more than one server or resource.

| | | | | | | | | |
|---|--|---------------|-------------|-------------------|-------------|-----------------|-------------|-----------------|
| Project Name: Avatar Service Framework | | | | | | | | |
| General info | Server info | Pricing info | Contacts | History | | | | |
| Server Type | Serial Number | Host Name | Port Number | IP Address | No. of CPUs | CPU Speed (MHz) | Memory (GB) | |
| pSeries P595 LPAR | | tdilpl45 | | 9.30.14.219 | 4.0 | 1130 | 2 | |
| | Software Name | Software Type | Comments | Monitoring Status | | | | |
| | DB2 Server V8.2 For Linux-RedHat | middleware | | | | | | |
| | WebSphere Application Server V6.1 For Linux-RedHat | middleware | | | | | | |
| | RedHat Linux 4 - pSeries | OS | | | | | | |
| | | | | | | | | CPU Memory Disk |
| | | | | | | | | |

Figure 6. Server information component of a contract

Additional server details on a per project or pilot basis is also integrated into the portal interface and is shown in Figure 6. Here innovators can log in to see summary health information for each of their servers.

Innovators and administrators gain significant benefit by having this summary and detailed monitoring information available through the cloud's Web interface. Network issues, performance problems, and capacity concerns can be quickly verified and corrected by using the monitoring functions of the cloud. Detailed problem analysis and resolution can also be added by viewing historic graphs and charts made available to the innovators and administrators.

Open source

Open source solutions played an important role in the development of the cloud. In particular, a couple of projects have been foundations for common cloud services such as virtualization and parallel processing. Xen is an open-source virtual machine implementation that allows physical machines to host multiple copies of operating systems. Xen is used in the cloud to represent machines as virtual images that can be easily and repeatedly provisioned and deprovisioned.

Hadoop, now under the Apache license, is an open-source framework for running large data processing applications on a cluster. It allows the creation and execution of applications using Google's MapReduce programming paradigm, which divides the application into small fragments of work that can be executed on any node in the cluster. It also transparently supports reliability and data migration through the use of a distributed file system. Using Hadoop, the cloud can execute parallel applications on a massive data set in a reasonable amount of time, enabling computationally-intensive services such as retrieving information efficiently, customizing user sessions based on past history, or generating results based on Monte Carlo (probabilistic) algorithms.

Virtualization

Virtualization in a cloud can be implemented on two levels. The first is at the hardware layer. Using hardware like the IBM System p™ enables innovators to request virtualized, dynamic LPARs with IBM® AIX® or Linux operating systems. The LPAR's CPU resource is ideally managed by IBM® Enterprise Workload Manager. Enterprise Workload Manager monitors CPU demand and use and employs business policies to determine how much CPU resource is assigned to each LPAR. The System p has micropartitioning capability, which allows the system to assign partial CPUs to LPARs. A partial CPU can be as granular as 1/10 of a physical CPU.

Micropartitioning combined with the dynamic load balancing capabilities of Enterprise Workload Manager make a powerful virtualized infrastructure available for innovators. In this environment pilots and prototypes are generally lightly used at the beginning of the life cycle. During the startup stage, CPU use is generally lower because there is typically more development work and fewer early adopters or pilot users. At the same time, other more mature pilots and prototypes may have hundreds or thousands of early adopters who are accessing the servers. Accordingly, those servers can take heavy loads at certain times of the day, or days of the week, and this is when Enterprise Workload Manager dynamically allocates CPU resources to the LPARs that need them.

The second implementation of virtualization occurs at the software layer. Here technologies such as Xen can provide tremendous advantages to a cloud environment. Our current implementations of the cloud support Xen specifically but the framework also allows for other software virtualization technologies such as VMWare's ESX product.

Software virtualization entails installing a hypervisor on an IBM System x or IBM System p physical server. The hypervisor supports multiple "guest" operating systems and provides a layer of virtualization so that each guest operating system resides on the same physical hardware without knowledge of the other guest operating systems. Each guest operating system is physically protected from the other operating systems and will not be affected by instability or configuration issues of the other operating systems.

Software virtualization allows underutilized servers to become fully utilized, saving the company significant costs in hardware and maintenance. A Xen virtualization model provides significant benefits:

- Virtual relocation: allows the cloud management system to dynamically relocate virtual machines (guest operating systems) in a matter of seconds with zero downtime.
- Instant archiving: allows the cloud to take a unused server offline with no ill affect. Later that same virtual machine can be restored and brought online in a matter of seconds.
- Instant rebalancing: allows the cloud to move over utilized virtual machines to physical machines that have unused resources (memory, CPU, disk).
- Instant deployment: allows the cloud to bring a virtual server online in a matter of seconds. Additional configurations or middleware and application provisioning may require additional time depending on the implementation.

A SAN based storage architecture must be used for some of these software virtualization benefits to be realized.

This dynamic allocation of resource and the large number of active pilots enable cloud resources to be extremely efficient. A nonvirtualized environment may well be able to handle less than half the number of projects of a virtualized cloud.

Storage architecture in the cloud

The storage architecture of the cloud includes the capabilities of the Google file system along with the benefits of a storage area network (SAN). Either technique can be used by itself, or both can be used together as needed.

Computing without data is as rare as data without computing. The combination of data and computer power is important. Computer power often is measured in the cycle speed of a processor. Computer speed also needs to account for the number of processors. The number of processors within an SMP and the number within a cluster may both be important.

When looking at disk storage, the amount of space is often the primary measure. The number of gigabytes or terrabytes of data needed is important. But access rates are often more important. Being able to only read sixty megabytes per second may limit your processing capabilities below your computer capabilities. Individual disks have limits on the rate at which they can process data. A single computer may have multiple disks, or with SAN file system be able to access data over the network. So data placement can be an important factor in achieving high data access rates. Spreading the data over multiple computer nodes may be desired, or having all the data reside on a single node may be required for optimal performance.

The Google file structure can be used in the cloud environment. When used, it uses the disks inside the machines, along with the network to provide a shared file system that is redundant. This can increase the total data processing speed when the data and processing power is spread out efficiently.

The Google file system is a part of a storage architecture but it is not considered to be a SAN architecture. A SAN architecture relies on an adapter other than an Ethernet in the computer nodes, and has a network similar to an Ethernet network that can then host various SAN devices.

Typically a single machine has both computer power and disks. The ratio of disk capability to computer capability is fairly static. With the Google file system, the single node's computer power can be used against very large data by accessing the data through the network and staging it on the local disk. Alternatively, if the problem lends itself to distribution, then many computer nodes can be used allowing their disks to also be involved.

With the SAN we can fundamentally alter the ratio between computer power and disk capability. A single SAN client can be connected to, and access at high speeds, an enormous amount of data. When more computer power is needed, more machines can be added. When more I/O capability is needed, more SAN devices can be added. Either capability is independent of the other.

Fast write is a capability available on many SAN devices. Normal disk writes do not complete until the data has been written to disk, which involves spinning the disk, and potentially moving the heads. With fast write, the write completes when the data reaches memory in the SAN device, long before it gets written to disk. Certain applications will achieve significant performance boosts through fast write if the SAN implements it.

Flash copy is an instantaneous copy capability available with some SAN devices. Actually copying the data may take time, but the SAN device can complete the physical copying after the logical copying. Being able to make copies is essential to any storage architecture. Often copies are used for purposes such as backup, or to allow parallel processing without contention. With flash copy capabilities from the SAN, the performance of copies can be greatly improved.

Shared file systems are not part of the SAN architecture, but can be implemented on top of the SAN. Some recovery techniques such as HACMP rely on SAN technology to enable failover. While the Google file system provides similar capabilities, it is not currently integrated into most failover techniques.

Piloting innovations on a cloud

Many companies are creating innovation initiatives and funding programs to develop innovation processes. Because innovation is an evolving topic, the team leaders often don't know where to start. More often than not, they look at traditional or existing collaboration tools to try to meet the requirements for collaborative innovation. Through numerous engagements with clients, IBM has discovered that collaboration tools by themselves will not yield the desired results as effectively as having a structured innovation platform and program in place.

IBM addressed this problem by developing a comprehensive innovation platform called Innovation Factory. The Innovation Factory removes most of the barriers that innovators experience by combining collaboration tools, search and tagging technologies, as well as site creation tools in a single unified portal.

This type of innovation platform enables innovation by putting a structure around the innovation process and providing tools for innovators and early adopters to publish, experiment, provide feedback, and enhance innovations. The Innovation Factory is a perfect complement to cloud computing because the innovators making new pilots and technologies available usually need servers or other computing resources in which to develop, test, and provide those services and applications to the early adopters.

By combining cloud computing and Innovation Factory, or any other innovation platform already in use, a company can benefit from a complete solution that provides both physical computer

resources and an innovation process combined with collaboration tools. Adding cloud computing to a company's existing innovation process reduces the time needed to develop and deliver a product, reduces the barrier to entry, and reduces costs associated with procurement, setup, management, and reuse of physical assets.

Cloud computing should be part of every innovation process when physical or virtual computer resources are needed for innovation pilots.

An overview of the IBM Innovation Factory solution is available in the HiPODS white paper *IBM Innovation Factory* (see www.ibm.com/developerworks/websphere/zones/hipods/). It describes the key components of the Innovation Factory.

Conclusion

In today's global competitive market, companies must innovate and get the most from its resources to succeed. This requires enabling its employees, business partners, and users with the platforms and collaboration tools that promote innovation. Cloud computing infrastructures are next generation platforms that can provide tremendous value to companies of any size. They can help companies achieve more efficient use of their IT hardware and software investments and provide a means to accelerate the adoption of innovations. Cloud computing increases profitability by improving resource utilization. Costs are driven down by delivering appropriate resources only for the time those resources are needed. Cloud computing has enabled teams and organizations to streamline lengthy procurement processes.

Cloud computing enables innovation by alleviating the need of innovators to find resources to develop, test, and make their innovations available to the user community. Innovators are free to focus on the innovation rather than the logistics of finding and managing resources that enable the innovation. Combining cloud computing with IBM Innovation Factory provides an end-to-end collaboration environment that could transform organizations into innovation power houses.

IBM is a leader in cloud computing and innovation technologies. IBM has been using these technologies internally to promote innovations through its own innovation portal, the Technology Adoption Program (TAP). Through the TAP program IBM employees have been able to quickly obtain computing resources. This has enable hundreds of innovation ideas to flourish within IBM. IBM can help its customers and partners do the same either as a hosted ecosystem and as a locally installed solution.

References

See all the HiPODS white papers at

www.ibm.com/developerworks/websphere/zones/hipods/library.html

Of particular interest are papers related to innovation and collaboration:

- *Innovation Factory: An integrated solution for accelerating innovation* (October 2007)
- *Introducing HiGIG: The HiPODS Global Innovation Grid* (August 2006)

Acknowledgements

We acknowledge this paper's major supporters and contributors:

- Executive sponsor: Willy Chiu
- The HiPODS Architecture Board led by Dennis Quan
- The Incubation Solutions Team that owns the Cloud strategy led by Jose Vargas
- The Innovation Factory team led by Jeff Coveyduc
- Contributors to the white paper: Greg Boss, Catherine Cuong Diep, Harold Hall, Susan Holic, Eugene Hung, Linda Legregni, Padma Malladi, Dennis Quan, John Reif, and Jose Vargas

Notices

Trademarks

The following are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both:

IBM, IBM logo, AIX, DB2, pSeries, System p, System x, Tivoli, WebSphere, xSeries

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.

Special Notice

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While IBM may have reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Anyone attempting to adapt these techniques to their own environments do so at their own risk.

While IBM may have reviewed each item for accuracy in a specific situation, IBM offers no guarantee or warranty to any user that the same or similar results will be obtained elsewhere. Any person attempting to adapt the techniques contained in this document to their own environment(s) does so at their own risk. Any performance data contained in this document were determined in various controlled laboratory environments and are for reference purposes only. Customers should not adapt these performance numbers to their own environments as system performance standards. The results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.