

February 2009



Seeding the Clouds: Key Infrastructure Elements for Cloud Computing



Table of Contents

Executive summary	3
Introduction	4
Business value of cloud computing	4
Evolution of cloud computing	6
The dynamic data center model.....	8
Architecture framework and technology enablers	9
Virtualized environment.....	10
<i>What is virtualization?</i>	10
<i>How does server virtualization work?</i>	10
Infrastructure management	11
<i>Automation</i>	11
<i>Self-service portal</i>	12
<i>Monitoring</i>	14
<i>Capacity planning</i>	15
Business use cases	16
Innovation enablement.....	16
<i>IBM's internal innovation portal</i>	16
<i>Sogeti innovation cloud</i>	17
Software development and test environments	19
<i>China Cloud Computing Center at Wuxi</i>	19
Advanced computing model for data-intensive workloads.....	20
<i>IBM/Google Academic Initiative</i>	20
Summary	22
References	23

Executive summary

Cloud computing is an emerging computing model by which users can gain access to their applications from anywhere, through any connected device. A user-centric interface makes the cloud infrastructure supporting the applications transparent to users. The applications reside in massively scalable data centers where computational resources can be dynamically provisioned and shared to achieve significant economies of scale. Thanks to a strong service management platform, the management costs of adding more IT resources to the cloud can be significantly lower than those associated with alternate infrastructures.

What is driving the adoption of cloud computing? Many factors, including the proliferation of smart mobile devices, high-speed connectivity, higher density computing and data-intensive Web 2.0 applications.

As a result, vendors across the IT industry have announced cloud computing efforts of varying capabilities and among corporate clients there is an increasing interest in aspects of the cloud, such as infrastructure outsourcing, software as a service key processes as a service and next-generation distributed computing.

IBM is uniquely positioned to help these clients adopt cloud computing technologies and management techniques to improve the efficiency and flexibility of their data centers. With proven expertise dating back to its pioneering position in the virtualization space during the 1960s, IBM has recently introduced its vision of a data center that supports a dynamic infrastructure. This vision brings together the strengths of the Web-centric cloud computing model and today's enterprise data center. It provides request-driven, dynamic allocation of computing resources for a mix of workloads on a massively scalable, heterogeneous and virtualized infrastructure. Furthermore, it is optimized for security, data integrity, resiliency and transaction processing. Thanks to extensive experience in both enterprise data centers and cloud computing, IBM is exceptionally well prepared to provide clients with the best solutions to achieve this vision.

IBM has been working with leading-edge clients around the world, such as Google and the Government of Wuxi in China, to define best practices for running data centers with workloads ranging from Web 2.0 applications to mission-critical transaction processing systems. Specifically, IBM has been defining and enhancing a cloud computing framework for running large scale data centers that enables key functionality for hosting a wide range of applications. This framework now includes automation for the complex, time-consuming processes of provisioning servers, networks, storage, operating systems and middleware. It also provides support for extremely data-intensive workloads and supports requirements for resiliency and security.

IBM's hands-on experience setting up cloud data centers represents a major step in the evolution of data centers in support of a dynamic infrastructure. This paper describes a high-level cloud computing infrastructure services framework and the underlying technology enablers, such as virtualization, automation, self-service portal, monitoring and capacity planning. It also discusses examples of, and the value propositions for, certain data centers that have been built in this manner to date. These data centers can host a mix of workloads, from Java™ 2 Enterprise Edition (J2EE) applications to software development to test environments to data-intensive business intelligence analytics.

Introduction

Business value of cloud computing

Cloud computing is both a business delivery model and an infrastructure management methodology. The business delivery model provides a user experience by which hardware, software and network resources are optimally leveraged to provide innovative services over the Web, and servers are provisioned in accordance with the logical needs of the service using advanced, automated tools. The cloud then enables the service creators, program administrators and others to use these services via a Web-based interface that abstracts away the complexity of the underlying dynamic infrastructure.

The infrastructure management methodology enables IT organizations to manage large numbers of highly virtualized resources as a single large resource. It also allows IT organizations to massively increase their data center resources without significantly increasing the number of people traditionally required to maintain that increase.

For organizations currently using traditional infrastructures, a cloud will enable users to consume IT resources in the data center in ways that were never available before. Companies that employ traditional data center management practices know that making IT resources available to an end user can be time-intensive. It involves many steps, such as procuring hardware; finding raised floor space and sufficient power and cooling; allocating administrators to install operating systems, middleware and software; provisioning the network; and securing the environment. Most companies find that this process can take upwards of two to three months. Those IT organizations that are re-provisioning existing hardware resources find that it still takes several weeks to accomplish. A cloud dramatically alleviates this problem by implementing automation, business workflows and resource abstraction that allows a user to browse a catalog of IT services, add them to a shopping cart and submit the order. After an administrator approves the order, the cloud does the rest. This process reduces the time required to make those resources available to the customer from months to minutes.

The cloud also provides a user interface that allows both the user and the IT administrator to easily manage the provisioned resources through the life cycle of the service request. After a user's resources have been delivered by a cloud, the user can track the order, which typically consists of some number of servers and software, and view the health of those resources; add servers; change the installed software; remove servers; increase or decrease the allocated processing power, memory or storage; and even start, stop and restart servers. These are self-service functions that can be performed 24 hours a day and take only minutes to perform. By contrast, in a non-cloud environment, it could take hours or days for someone to have a server restarted or hardware or software configurations changed.

The business model of a cloud facilitates more efficient use of existing resources. Clouds can require users to commit to predefined start and end dates for resource requests. This helps IT organizations to more efficiently repurpose resources that often get forgotten or go unused. When users realize they can get resources within minutes of a request, they are less likely to hoard resources that are otherwise very difficult to acquire.

Clouds provide request-driven, dynamic allocation of computing resources for a mix of workloads on a massively scalable, heterogeneous and virtualized infrastructure. The value of a fully automated provisioning process that is security compliant and automatically customized to user's needs results in:

- *Significantly reduced time to introduce technologies and innovations;*
- *Cost savings in labor for designing, procuring and building hardware and software platforms;*
- *Cost savings by avoiding human error in the configuration of security, networks and the software provisioning process; and*
- *Cost elimination through greater use and reuse of existing resources, resulting in better efficiency.*

The cloud computing model reduces the need for capacity planning at an application level. The user of an application can request resources from the cloud and obtain them in less than an hour. A user who needs more resources can submit another request and obtain more resources within minutes, and in a policy-based system, no interaction is needed at all; resource changes are performed dynamically. Thus, it is far less important to correctly predict the capacity requirements for an application than it is in traditional data centers, and capacity planning is simplified because it is performed only once for the entire data center.

Today's IT realities make cloud computing a good fit for meeting the needs of both *IT providers* (who demand unprecedented flexibility and efficiency, lower costs and complexity and support for varied and huge workloads) and *Internet users* (who expect availability, function and speed).

As technology such as virtualization and corresponding management services like automation, monitoring and capacity planning services become more mature, cloud computing will become more widely used for increasingly diverse and even mission-critical workloads.

Evolution of cloud computing

This section reviews the history of cloud computing and introduces the IBM vision for cloud computing that supports dynamic infrastructures. The following section introduces an infrastructure framework for a data center and discusses the virtualized environment and infrastructure management. Subsequently, existing cloud infrastructures and their applications are described.

Cloud computing is an important topic. However, it is not a revolutionary new development, but an evolution that has taken place over several decades, as shown in Figure 1.

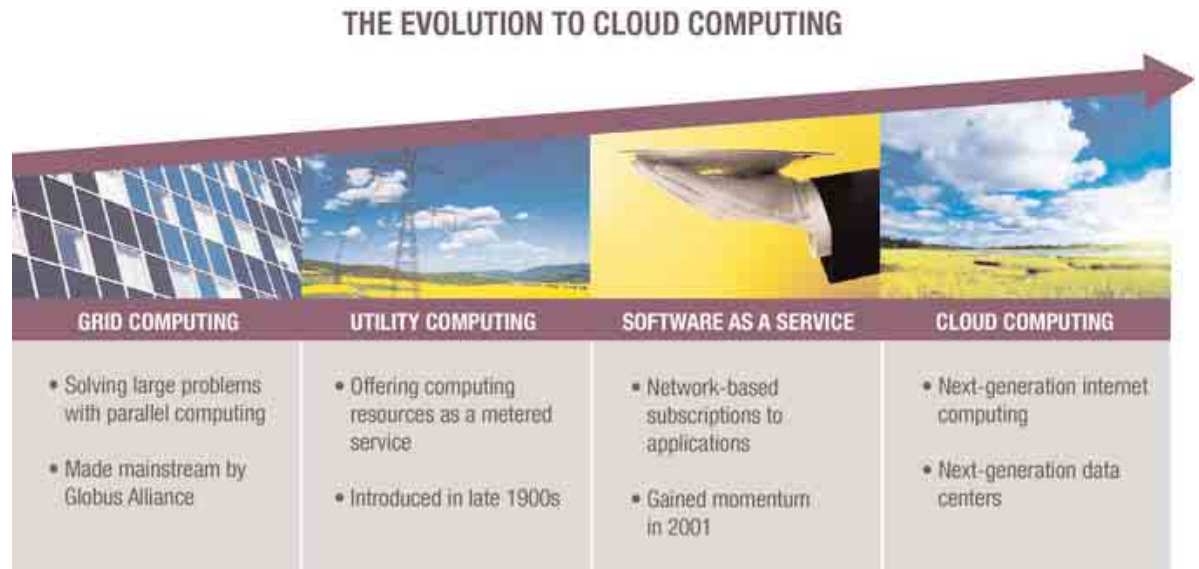


Figure 1. Evolution toward cloud computing

The trend toward cloud computing started in the late 1980s with the concept of grid computing when, for the first time, a large number of systems were applied to a single problem, usually scientific in nature and requiring exceptionally high levels of parallel computation.

That said, it's important to distinguish between grid computing and cloud computing. Grid computing specifically refers to leveraging several computers in parallel to solve a particular, individual problem, or to run a specific application. Cloud computing, on the other hand, refers to leveraging multiple resources, including computing resources, to deliver a "service" to the end user.

- *In grid computing, the focus is on moving a workload to the location of the needed computing resources, which are mostly remote and are readily available for use. Usually a grid is a cluster of servers on which a large task could be divided into smaller tasks to run in parallel. From this point of view, a grid could actually be viewed as just one virtual server. Grids also require applications to conform to the grid software interfaces.*
- *In a cloud environment, computing and extended IT and business resources, such as servers, storage, network, applications and processes, can be dynamically shaped or carved out from the underlying hardware infrastructure and made available to a workload. In addition, while a cloud can provision and support a grid, a cloud can also support nongrid environments, such as a three-tier Web architecture running traditional or Web 2.0 applications.*

In the 1990s, the concept of virtualization was expanded beyond virtual servers to higher levels of abstraction—first the virtual platform, including storage and network resources, and subsequently the virtual application, which has no specific underlying infrastructure. Utility computing offered clusters as virtual platforms for computing with a metered business model. More recently software as a service (SaaS) has raised the level of virtualization to the application, with a business model of charging not by the resources consumed but by the value of the application to subscribers.

The concept of cloud computing has evolved from the concepts of grid, utility and SaaS. It is an emerging model through which users can gain access to their applications from anywhere, at any time, through their connected devices. These applications reside in massively scalable data centers where compute resources can be dynamically provisioned and shared to achieve significant economies of scale. Companies can choose to share these resources using public or private clouds, depending on their specific needs. Public clouds expose services to customers, businesses and consumers on the Internet. Private clouds are generally restricted to use within a company behind a firewall and have fewer security exposures as a result.

The strength of a cloud is its infrastructure management, enabled by the maturity and progress of virtualization technology to manage and better utilize the underlying resources through automatic provisioning, re-imaging, workload rebalancing, monitoring, systematic change request handling and a dynamic and automated security and resiliency platform.

The dynamic data center model

As more and more players across the IT industry announce cloud computing initiatives, many CIOs are asking IBM how they can adopt cloud computing technologies and management techniques to improve the efficiency and flexibility of their own data centers and other computing environments. In response, IBM has recently introduced its vision of data centers which support a dynamic infrastructure that unifies the strengths of the Web-centric cloud computing model and the conventional enterprise data center (as shown in Figure 2).

These data centers will be virtualized, efficiently managed centers, which will employ some of the tools and techniques adopted by Web-centric clouds, generalized for adoption by a broader range of customers and enhanced to support secure transactional workloads. With this highly efficient and shared infrastructure, it becomes possible for companies to respond rapidly to new business needs, to interpret large amounts of information in real time and to make sound business decisions based on moment-in-time insights. The data center that supports a dynamic infrastructure is an evolutionary new model that provides an innovative, efficient and flexible approach in helping to align IT with business goals.

The remaining sections of this paper provide a high-level description of the data center in support of a dynamic infrastructure with underlying technology enablers, such as virtualization, automation, provisioning, monitoring and capacity planning. Finally, examples of actual data center implementations are discussed to reveal which characteristics of dynamic infrastructure models can offer the most value to customers of any size, across a variety of usage scenarios.

CLOUD COMPUTING AND THE DYNAMIC ENTERPRISE DATA CENTER

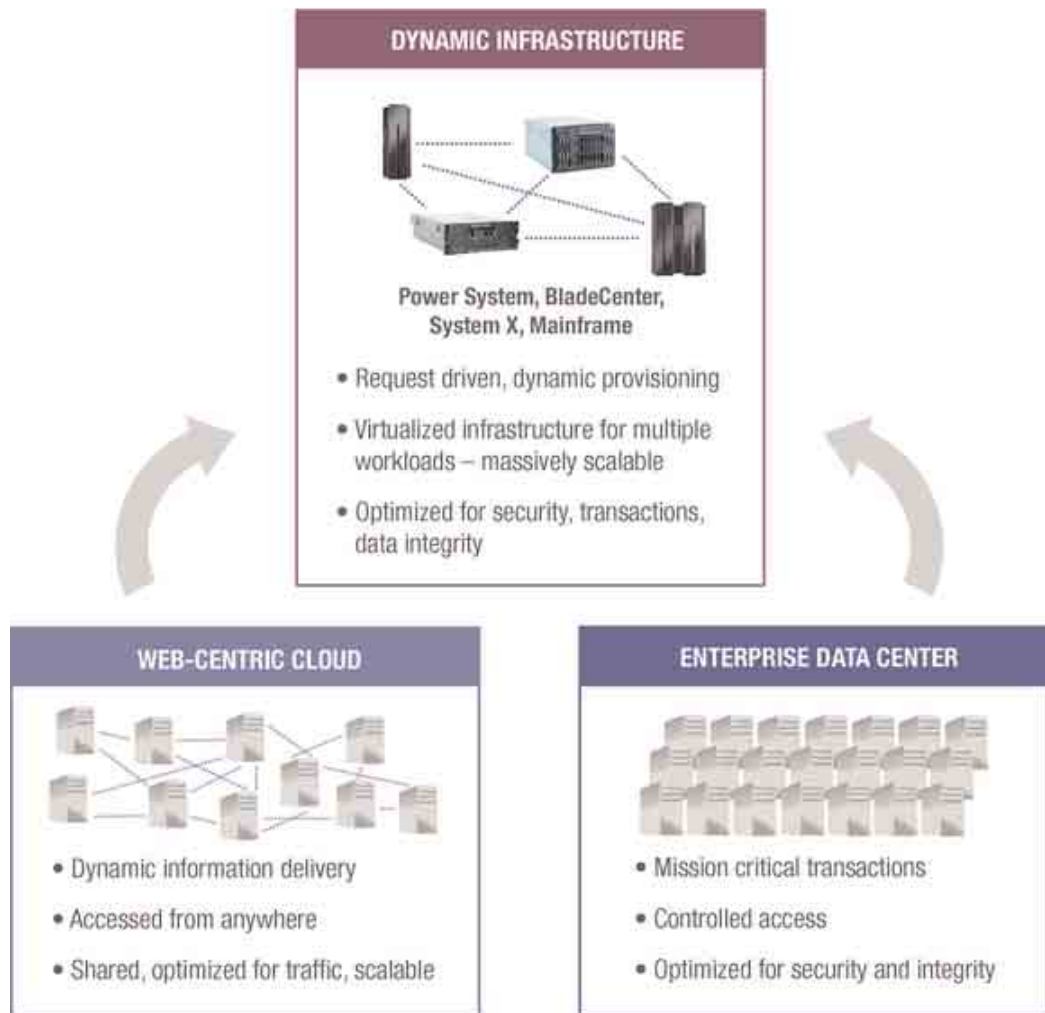


Figure 2. Cloud computing and the data center that supports a dynamic infrastructure

One solution to such problems lies in the MapReduce distributed parallel programming model, which is increasingly used to write these types of business analytics programs. MapReduce allows the processing to be distributed across hundreds to thousands of nodes, all of them working in parallel on a subset of the data. The intermediate results from these nodes are combined, sorted and filtered to remove duplicates before arriving at the final answer.

The majority of processing at Google is based on this MapReduce model. Google's success has prompted other companies to follow the same model, and the increasing popularity of MapReduce-style programming inspired the Apache Hadoop project, which is an open-source implementation of the MapReduce programming framework. Many companies use Apache Hadoop for large-scale business analytics, ranging from understanding users' navigation patterns and trends on their Websites to building targeted advertising campaigns.

The characteristics of the MapReduce programming model require an underlying compute infrastructure that is highly scalable. A data center platform that supports dynamic infrastructure provides an ideal foundation for such workloads.

To promote this surge of interest in MapReduce-style programming, IBM and Google announced a partnership in October 2007 to provide a number of data centers for use by the worldwide academic community. These centers are powered by an IBM solution based on dynamic infrastructure architecture for data center management, which allows users to quickly provision large Hadoop clusters for students who might otherwise be short of required IT resources to complete their lab assignments or run their research programs. Some of the leading universities using these centers and teaching courses on the MapReduce methods are University of Washington, University of Maryland, Massachusetts Institute of Technology, Carnegie Mellon University, University of California Berkeley and Colorado State University.

The heart of this solution is to automatically provision a large cluster of virtual machines for students to access through the Internet to test their parallel programming projects. As a result, physical machines, or virtual machines created using the Xen hypervisor, can be provisioned rapidly and automatically using

