Cloud Computing Pilots for Enterprise Data Center Applications

Casimer DeCusatis
Distinguished Engineer, IBM Corporation, Poughkeepsie, NY
decusat@us.ibm.com

Abstract

Enterprise class architectures for cloud computing have begun to emerge, driven by recent advances in virtualization of servers, storage, and access networks. We present a hierarchy classification scheme for cloud services, based on whether the intrastructure, platform, or software is being provisioned and virtualized. We then discuss early adoption of public and private clouds based on the client workload affinity, and present several recent case studies of cloud implementations. We also describe a virtual storage cloud pilot currently operating in the New York City area, which relies on virtual private network extensions over extended distances. Finally, we summarize a proposal from the 2010 Internet 2 workshop to apply the design principles of the virtual storage cloud to the construction of dynamically provisioned cloud data centers utilizing next generation Internet technology.

Introduction

Cloud computing is an emerging method of delivering information technology (IT), in which applications, data, and resources within the data center are virtualized, scaled, and rapidly provisioned on demand. This provides the benefits of a traditional enterprise data center without requiring the end user to purchase, manage, or maintain IT resources. Applications are provided as standardized offerings to end users from a centralized, dynamically reconfigurable data center. This allows for a more flexible cost and pricing business model, and enables the enterprise to provision new resources much faster than current methods. As shown in figure 1, capital equipment expenditures for enterprise-class data centers have remained relatively flat over the past decade, while operating expenses (management, administration, and energy consumption) continue to grow. Cloud computing helps address the growing total cost of ownership for enterprise data centers. In many cases, it's possible to reduce labor cost by up to 50% (configuration management, maintenance, monitoring, and operations) and to reduce capital expenditures by up to 75% (significantly reducing software licensing costs). The cycle time for provisioning new features can be reduced from weeks to minutes, and quality is improved by eliminating up to 30% of software defects.

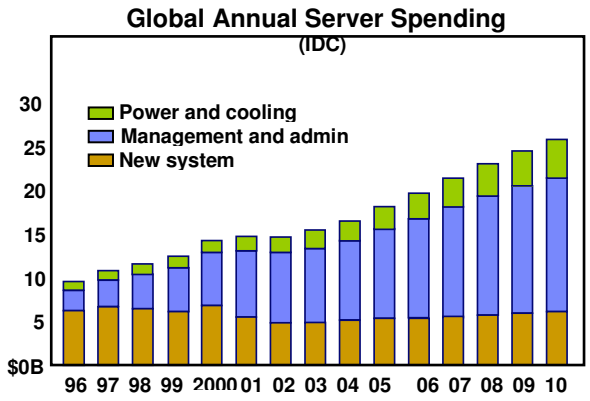

Figure 1 – Global Annual Server Spending from 1996 – 2010 (source: IDC 2010 report)

There has been a great deal of publicity around content delivery clouds for commercial applications, such as Facebook and Flickr, or desktop applications such as Google Docs. However, this represents only a subset of the possible cloud deployments and service delivery models [1-4]. So-called private clouds will exist within a corporate data center to address applications which are sensitive to security or performance issues, as well as for test and development environments. Public clouds will extend beyond the confines of a single enterprise. Applications related to web hosting, software as a service, and others will benefit from this approach.

A confluence of technology trends, including faster processors (multi-core designs with higher clock frequencies), solid state memory for storage, and highly scalable networking technologies, have contributed to dramatic cost reductions in key data center resources. These, in turn, have helped make cloud computing deployments feasible. Consider, for example, the recently announced Intel Core processor family, which promises a remarkable cost/performance ratio (under $300 for a quad core, 2.66 GHz processor) [5]. Within the network, technology improvements have driven down network bandwidth costs to the point where communications technology usually reserved for telecommunication central offices has become practical to install in corporate data centers. As processing power, storage, and bandwidth become available to anyone who needs them, businesses are able to more easily afford the computing resources they need to remain competitive. Moreover, they no longer have to purchase, manage, and support these computing resources themselves; nearly free IT resources have made cloud computing possible by allowing providers to deploy massive amounts of computational power and make it available on demand to end users.

In this paper, we will describe different types of cloud service offerings and applications which are well suited for early cloud adoption. We describe storage, compute, and development clouds which have realized significant cost reductions and performance

efficiencies. Finally, we review a recent pilot for a virtual storage cloud using shared network infrastructure, which may be extensible to Internet 2.

Information Technology As A Service

The first steps towards cloud deployment involve reducing infrastructure complexity and staffing requirements, thereby realizing reduced operational costs. This value proposition has led nearly one-third of companies currently using cloud computing to increase their investment in the technology, despite the weak global economy. The next steps towards cloud computing transformation involve removing physical resource boundaries on key applications, sharing storage and computing resources, and implementing granular service metering and billing models. According to recent reports [6-8], the leading obstacles which hinder the widespread adoption of cloud computing include security, data transfer bottlenecks, unpredictable performance, energy consumption, and ease of use. For example, when compared against widely used alternatives, the data center network needs to exhibit up to two orders of magnitude better cost/performance in order to fully realize the potential of cloud computing environments. In an effort to address these obstacles, it may be preferable to take a phased approach to cloud implementation.

We can segment the cloud computing market into three hierarchical tiers, depending on whether infrastructure, platforms, or software is being provisioned, as shown in figure 2. At the lowest level, infrastructure as a service (IaaS) encompasses the shared, virtualized, dynamic provisioning of servers, storage, and networking. Some early examples of this include cloud service deployments from Amazon and other content providers; however, most of these offering suffer from a lack of adequate service and performance guarantees, since they are based on the existing public Internet. There are many emerging network technologies which are expected to enable IaaS deployments in the future. Examples include data center networks which can be managed through a common hypervisor with server resources, enabling migration of server virtual partitions from one physical server to another without service disruptions.

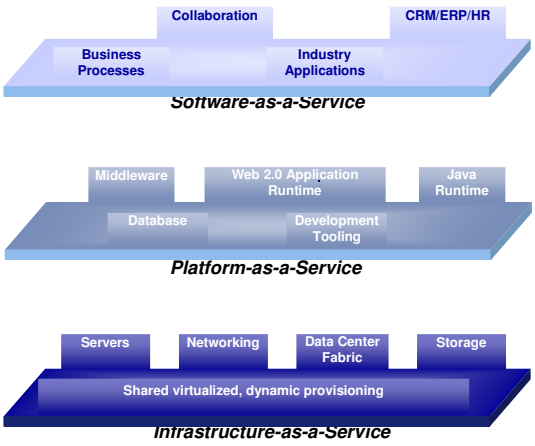

Figure 2 – Segmentation of IT as a service

Platform as a service (PaaS) includes database and development tools, middleware, and various Web 2.0 applications, enabled to provide virtual application services (such as Google Docs) or collaborative software development environments. Finally, software as a service (SaaS) enables entire business processes (such as SalesForce.com) or applications such as video conferencing (Lotus Live).

These three tiers of cloud service are being adopted by a range of applications, as shown in figure 3. We can view the cloud as a service delivery model; it is important to select the right workloads and optimizing them for cloud delivery. Early applications are based on workload affinity, including storage clouds, desktop clouds, collaborative services, web serving, and developer clouds. Many successful cloud data centers have been deployed worldwide, as described in the following sections.

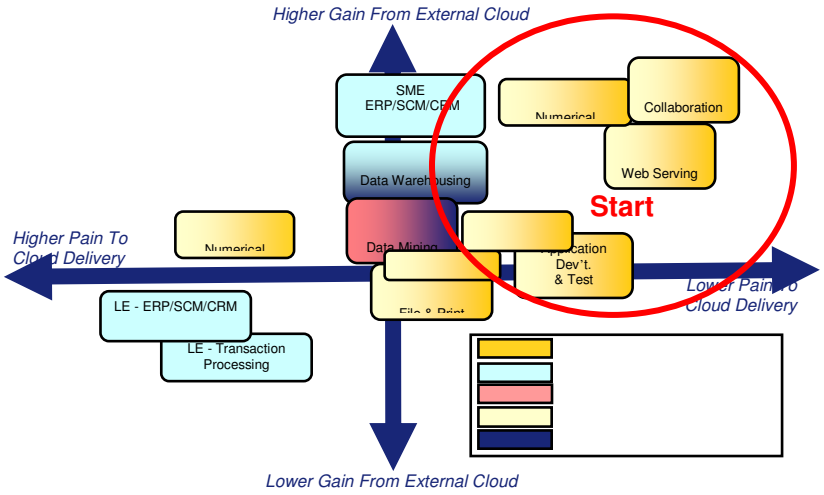

Figure 3 – Cloud adoption based on workload affinity; gain can be measured in terms of ROI, time to value, or similar metrics

China Software Developer Cloud

IBM has collaborated with the municipal government of Wuxi, near Shaghai, to create the first commercial cloud computing center in China. Eleven locations are linked together in a multi-tenant public software development cloud, supporting a number of start-up companies. Resources are access through either the Internet or secure, virtualized, private network connections. Each of the companies has access to a private development environment with Rational software development and test tools, WebSphere security, and Tivoli autonomous provisioning. Using IBM Blue Cloud technology, this offering includes IBM System X BladeCenter (HS21), Power systems (System P560Q), and IBM TotalStorage products and switches including the SAN16B-2 and DS4700. Backup and restore capabilities for the cloud are provided by Tivoli Storage Manager. Software development clouds such as this accelerate development and test cycles, and allow quick on-boarding of new resources as required. It is hoped that efforts such as this will help drive the region's transformation to a service-led economy.

Video/Animation Compute Cloud

Kantana Animation Studios is one of the leading entertainment companies in Thailand. Their work includes character rendering, wire frame modeling for 3D structures, and video files with both live action and animated content. This work is placing increasing demands on their data storage capacity, which needs to scale in a cost effective way, as well as their ability to store and retrieve extremely large files at high speed. They have implemented an IBM Smart Business Storage Cloud based on System X and IBM TotalStorage technology, managed under a service offering from IBM Global Services. This allows the storage of large amounts of data in a single location, accessible to all their animators, and to increase storage capacity as dictated by business requirements. This solution has enabled the exponential growth of storage capacity while reducing management costs. They have also documented an increase in productivity derived from the use of centralized file storage, since animators are now free to take on new projects instead of concerning themselves with infrastructure requirements.

IBM/Google/National Science Foundation Training Cloud

One of the largest production clouds in existence, this joint initiative employs over 1,000 servers, 6 TBytes of RAM, and over 800 TBytes of storage. This multi-tenant public cloud supports nearly 40 universities and over 850 students and researchers. By allowing users to tap into levels of compute power not previously available, this cloud promotes open development standards and the Hadoop massively parallel computing model. Students perform research and are also trained in next generation computing skills. During its first 6 months of operation, over 50,000 service requests were processed by this cloud. The software includes Tivoli information management and Websphere tools, with a combination of hypervisors using Linux and Windows platforms. IBM Power systems, System X, and BladeCenter technology are all being leveraged to provide cloud services which could serve as a model for transforming education in much the same way that the Internet changed our approach to technical education.

Virtual Storage Cloud for Internet 2

While the storage clouds presented in our earlier examples offer good business value, it is desirable to share the resources of a single storage cloud among multiple customers. A pilot for such a virtual storage cloud is currently in progress in the New York City metropolitan area, tying three of the world's major video broadcasting companies into a common virtual storage cloud. Video serving demands have been placing an increasing burden on each company's data storage infrastructure; they share a common need for cost effective, highly scalable storage solutions. As illustrated in figure 4, IBM has implemented a Smart Business Storage Cloud on Madison Avenue in New York City, with a virtualized network infrastructure interconnecting multiple users within a 20-50 km radius. A single common repository for data storage is based on XIV gen2 with both scale-out file servers (SOFS) and scale-out network attached storage (SONAS). This allows for economies of scale in management, and enables elastic scaling of storage capacity with short provisioning time. This is made possible by networking technology

including IBM and Juniper network switches within the cloud data center, which interconnect to a private optical network managed by Level(3) corporation. Dual redundant optical wavelength division multiplexing equipment from Adva Optical Networking interconnects all locations, and wavelengths are provisioned using Tivoli connected to the wavelength multiplexer control plane. Optical wavelengths are reserved for each end user, providing secure, private data transfer. This virtualized network infrastructure is connected directly into the cloud data center, which enables new functionality such as the ability to boot devices off the network or migrate server virtual partitions between different physical locations. Furthermore, the optical network has been tested and enabled for future protocols such as Fibre Chanel over Converged Enhanced Ethernet (FCoCEE), which is expected to further reduce infrastructure costs, as well as InfiniBand protocols which will enable low, consistent latencies.

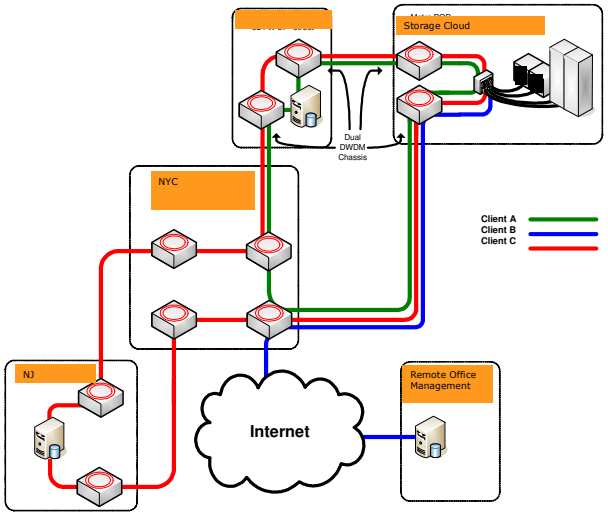

Figure 4 – Virtual Storage Cloud Pilot

We have proposed that a virtual network similar to this pilot, which overcomes traditional one-for-one mapping of network resources to data center applications or operating systems, could be enacted on a much larger scale, using high speed Internet connectivity. As shown in Figure 5, the Internet 2 backbone is a national scale, 10 Gbit/s network based on optical fiber provided by Level(3), which also feeds many smaller capillary networks. This was the first wavelength routed network deployed on a large scale; the infrastructure can dynamically provision additional capacity as required and monitor performance at various levels. Future versions will implement GMPLS wavelength routing. Cloud service providers are using similar tools for their applications, and it has recently been proposed that the features of a virtual storage cloud are well suited to implementation over the Internet 2 network [9].

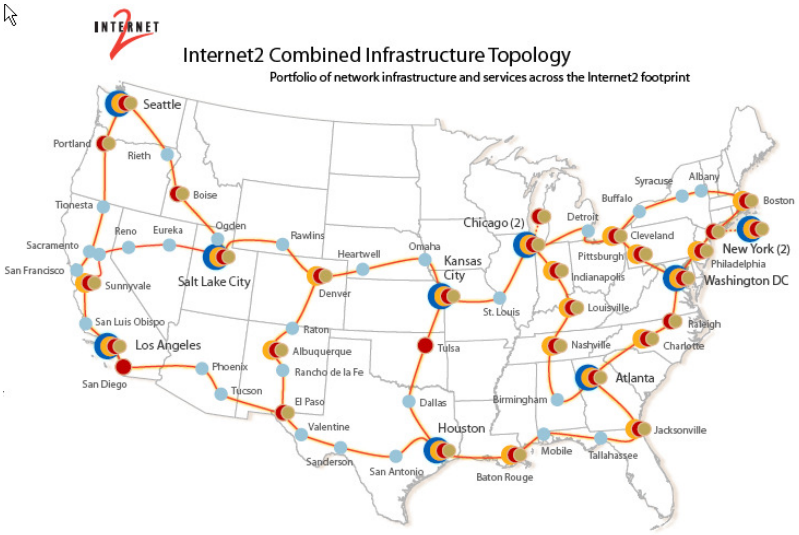

Figure 5 – Internet 2 Backbone

While most cloud computing discussions focus on processing power or software, the network architecture is playing an increasingly important role. Similar to grid computing models, some types of compute clouds are disaggregated across multiple data centers and may contain a mixture of mainframe, midrange, and other types of servers and storage. Just as optical fiber networks form the basis of telecommunication clouds, they will also be the underlying technology behind cloud computing. The cloud will be able to adapt optical routing, labeling, multiplexing, and switching techniques from related telecom practices. Applications such as virtual server migration are enabled by the joint management of network and server resources. Server partition mobility is facilitated by large layer 2 domains in the network, so high layer 2 port density becomes an important consideration for virtualized data centers. Conventional architectures statically map web services to Ethernet VLANs, where each VLAN is often constrained to a few hundred servers (due to overhead in the control plane). Spanning traffic across VLANs is a nontrivial problem, which is addressed in some cases by expensive, proprietary hardware and software. In this manner, data center resources are fragmented, and limitations on the network infrastructure (such as IP address hierarchies) make it difficult to dynamically reassign servers and applications. Current networking solutions are unable to cost effectively scale to a large Layer 2 fabric, which would help address this problem. Distributed computation is severely constrained by the low bisection bandwidth of the network, which can become a performance bottleneck over longer distances. Furthermore, conventional network designs do not scale well, since they concentrate traffic in a few large data center switches which must be disruptively upgraded to keep pace with traffic demand. This suggests the need for future switches which can be stacked through a dedicated, high bandwidth interconnect, enabling dynamic upgrades or failover without service interruptions. Another part of the scaling problem is the need for configuring large amounts of stand-alone appliances dedicated to security, load balancing, and other functions. Future designs are expected to automate and virtualize many common security functions, leveraging the existing switch fabric rather than an unwieldy deployment of point appliances. Finally, as workloads change frequently, intermittent

congestion of the network occurs, which can reduce the aggregate computational power of the data center. Existing congestion detection and management schemes are likely not adequate for future network designs; alternate means of performing flow control on virtual queues and preventing head-of-line blocking will be required.

Conclusions

Cloud computing data centers have demonstrated rapid, low cost deployment of information technology (IT) resources. This, in turn, is driving significant changes in the infrastructure supporting modern enterprise data centers. The traditional data center compute model, especially in the case of rack-mounted x86-based servers, has consisted of lightly utilized servers running a minimal operating system or a hypervisor with a small number of virtual machines. In this environment, lower bandwidth links and a fairly static infrastructure were sufficient for attachment to server resources. The industry is moving towards a higher bandwidth, dynamic infrastructure model that involves highly utilized servers running many virtual machines per server, using higher bandwidth links to communicate with virtual storage. Cloud computing offers lower capital expenses through higher utilization of servers, storage, and networks, as well as lower operational expenses through automated and integrated end-to end management. We have recently proposed a model for virtualized, shared private network infrastructure which will allow multiple customers to share a common storage data repository, and which may be extensible to networks used by the Internet 2 consortium.

References

1) Gartner Group report, "Next Generation Enterprise Data Centers", September 2007

2) Forrester report, "There are two types of compute clouds", available from www.forrester.com (November 2008)

3) Gartner Group report, "You can't do cloud computing without the right cloud (network)", available from www.gartner.com (June 2008)

4) C. Zaragoza, "Global study: cloud computing provides real business benefits…", Kelton Research study, available from www.avanade.com (February 2009)

5) J. Eijmberts, "Understanding Intel Itanium architecture values", proc. HP Dutchworld 2008 / Interex User Group Conference (November 20, 2008)

6) M. Armbrust et al., "Above the Clouds: A Berkeley View of Cloud Computing," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28, Feb 2009. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

7) C. DeCusatis, *Handbook of Fiber Optic Data Communication*, 3rd edition, Academic Press, NY (2008)

8) C. DeCusatis, "Converged networking for next generation enterprise data centers", Proc. National Science Foundation conference for the Enterprise Computing Community (ECC), June 21-23, 2009, Marist College, Poughkeepsie, NY (2009)

9) M. Haley, C. DeCusatis, T. Bundy, and G. Montanti,, "Cloud computing pilots with IBM, Level 3, and Adva Optical Networking", Proc. Internet2 Spring Member Conference, Arlington, Va. (April 26-29, 2010)