

Keeping Enterprises' Head Above The Clouds!

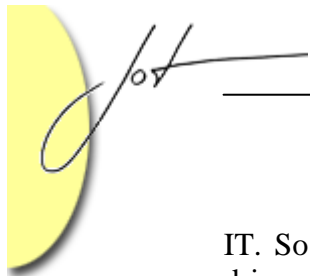
Mahesh H. Dodani, IBM, U.S.A.

1 TODAY'S DATA CENTER WEATHER REPORT: HOT, CROWDED AND CHAOTIC

“While the Internet and World Wide Web may seem invisible, residing somewhere out in the ether, in fact they reside in a network of interconnected data centers, also known as server farms. These data centers usually hold thousands of computer servers, jam-packed on racks, one on top of the other, which store and transmit the data and Web pages available on the Internet. Together, all of these servers in all of these data centers are known as "the cloud," and today more and more of what you do when you fire up your computer doesn't happen on the little hard drive under your desk, but actually happens out there in the network cloud. All the servers in America today would take about six or seven 1-gigawatt nuclear power plants to keep running 24/7, with that number going up every year.” – Thomas Friedman: Hot, Flat and Crowded

Over my last five articles, I have laid a foundation for a [Service](#) Oriented Architecture (SOA) as the enterprise architecture of the [globally integrated enterprise](#) and focused on how to define and establish the business side of the enterprise through a well defined [business architecture](#). I have continued over the last few articles to focus on the IT architecture side of the equation, and covered the application and infrastructure architectures. This article serves two purposes: as the last in the SOA series by focusing on the infrastructure architecture, as well as the transition into a new series that I want to kick off focusing on cloud computing.

A key theme for ensuring the success of the globally integrated enterprise as an agile business that is able to respond to change by seamlessly handling new customer demands, market conditions, competitive threats, Government regulation & compliance, and acquisitions is the alignment of the business to its IT. Over the past decade we have all experienced many fundamental changes that technology has enabled in supporting the agile enterprise. However, these technology advancements has placed a tremendous strain on an enterprises' data centers and IT operations. IT professionals have had to balance the challenges associated with managing data centers as they increase in cost and complexity with the need to be highly responsive to ongoing demands from the business placed on



IT. So on one side you have an accelerating pace of business model changes that can drive competitive advantage but can wreak havoc with existing IT infrastructures. On the other side – you have a set of operational challenges around cost, service delivery, business resiliency and security, and “green” initiatives that have many data centers at a breakpoint. Never before has the enterprise data center faced such a “perfect storm” of forces that drives the need for true data center transformation.

Let us examine the operational issues a little closer. The IT infrastructure at the core of the world’s most serious business – like government, healthcare and financial services – is still highly fragmented, inefficient, underutilized and unable to keep up with the pace of transactions. Underlying this IT infrastructure is a trillion dollars of investment in applications, information stores, and data centers that exist as islands of one kind or another – and have to move forward into the Internet age. Today, IT organizations are mired down in a sea of operational issues and at the same time have to deal with the added issues of compliance and security, environmental concerns, and the ever growing demand for tackling ever increasing amounts of information in “real time.” Consider: 10GB Ethernet ports projected to triple over the next five years, Apple’s iPhone 3G has a capacity of 16 GB! Costs to power and cool systems have doubled or risen eight-fold, depending on the survey that you read. Costs to manage this IT infrastructure has quadrupled in the last 5 years and projected to continue to grow at least 10% a year. The overall impact? Enterprises report that IT operational overhead is up to 70% of IT budget and growing, leaving precious few resources for new initiatives. Figure 1 summarizes the IT spending trends from several studies.

Global Annual IT Spending *Estimated US\$B 1996-2010*

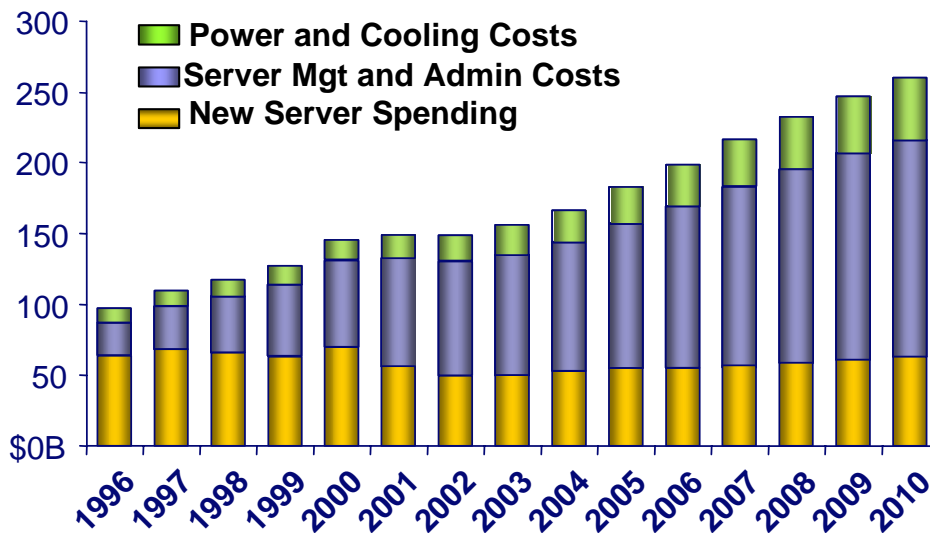


Figure 1: Enterprise IT Spending Challenge



Exacerbating this rise in IT spend, the “flattening” of the world economy has enabled vast pools of capacity and talent to be available everywhere. Advances in technology have lowered the barrier for organizations who want to make use of those resources, and allowed them to evolve to reap the benefits of global integration. During the first half of the 20th century, most large companies were the prototypical international enterprise with scores of sales offices overseas and exporting products to customers all over the world. Today, most large companies have moved to an international or multi-national model with regional offices and supporting data centers proliferating around the world. As we have discussed, to compete and grow in today’s environment, an enterprise must evolve to a globally-integrated enterprise. It is an evolution made possible by the rise of a globally networked infrastructure, allowing access to new skills, high growth markets and free-trade agreements, in new and expanding parts of the world.

In parallel to this, and in many ways supporting this, global integration trend was a tremendous proliferation of devices. In less than a decade, the Internet has connected a million businesses and a billion people – making it, in essence, the world’s largest operational infrastructure. It has changed the way we think about doing business. Along with this “connectedness” has come an explosion of transactions and information. Data volumes and bandwidth consumed are doubling every 18 months with devices accessing data over networks doubling every 2.5 years. It is projected that we will add 200 million new users each year for the foreseeable future. Soon, trillions upon trillions of things will be ubiquitously connected to the Internet. Right now, we have nearly 3 billion people subscribed to wireless technology. In many of the emerging countries, wireless is the only communication service which has facilitated the growth of these countries by leap-frogging the need for expensive and time-consuming infrastructure. So the explosion of information and the need to better secure and manage the delivery of that information becomes critical.

Business Models have evolved along the way as well. Initially, it was about finding ways to allow businesses to interact better together. Then, it was about how to offer better access to information for consumers. As we have moved into the latter part of this century – how has all of this come together? Well, along with new platforms come new business opportunities and the internet as a platform has spawned completely new ways of doing business. Companies like e-Bay have brought about a fundamentally new model of consumer to consumer businesses and Google has forever changed the way we access information. The acceptance of these emerging business models are accelerating and changing forever the dynamics of how we all interact, collaborate and deliver value.

Data centers are at a tipping point and need to change to accommodate the business needs, technological advances, and the increase in connectedness and information. We continue to see significant increases in computing demand – between 2000 and 2010 server capacity is expected to grow by six times, and storage capacity by 69 times. In fact, blades (dense server environments) will represent about one-quarter of all server shipments around the world. So this trend shows both a volume and a growth of IT

resources. However, these resources typically have 10-30% utilization rates which will need to be optimized to handle the explosion in information.

Secondly, we are seeing real changes in cost dynamics between capital expenditures vs. operating expenditures. Over a 20-year time period, an enterprise will spend three to five times in operating costs what it originally spent in capital costs to build that data center. Using the industry average of \$1500-\$2500 per square foot for a new data center, a 20,000 square foot data center costs \$30M-\$50 million to build, and \$90M-\$150M to operate. Over half that cost is in energy, and commercial energy costs in the U.S. and around the world will continue to increase.

There is also a significant mismatch between the facilities that enterprises are running and the capabilities that they want. As you look at technologies they want to install between 2000 and 2010 the density of that technology is expected to grow by 20 times over the period. Data centers have clearly not increased their power and cooling capacity by 20 times in that same time period. In fact, most data centers have not changed at all from the day they were built. So we really need a significantly different model for data center design and the management of those data centers, including the software elements, that really addresses the key needs. There is a clear mismatch between data centers today and the underlying technology. Over 78 percent of the datacenters around the world are more than seven years old. These datacenters are expected to last a long time – about one-third of the enterprises are expecting their data centers to continue to last 20 or even 30 years.

How can an enterprise scale their data centers to meet their business growth requirements, and to take advantage of technologies that may be available over a 10-year, 20-year, 30-year time period. Enterprises need to focus on data center designs and solutions that can meet their business and IT growth requirements and evolve effectively to leverage everchanging technologies. These data center designs must be able to reduce their capital and operational costs by focusing on improved energy efficiency and improve the utilization of IT resources. Finally, there needs to be a focus on reducing the risk by providing more available, predictable datacenter operations and more reliable and predictable datacenter designs. The main approaches to achieve these goals is through consolidation and virtualization. Virtualization is a logical representation of resources not constrained by physical limitations. Virtualization allows a single physical server to be partitioned into multiple logical servers to support server consolidation—to reduce the number of servers required to support IT services. Through consolidation and virtualization, enterprises can achieve a simpler, more scalable, more cost-efficient IT infrastructure that aligns more flexibly with emerging business goals. The rest of this article introduces cloud computing as the model to facilitate the “flattening” of the IT infrastructure, the ability to flexibly scale the resources “on demand” based on needs, the delivery of the services to users as a “utility”, and the effective management of the qualities of services based on established “agreements.”



2 CLOUD COMPUTING TO THE RESCUE: FLATTENING THE INFRASTRUCTURE

“The old one-server-per-application model has created a dire situation in the data centers of large enterprises. Their infrastructures are becoming too complex and expensive to maintain, and smaller organizations want new ways to grow and expand their businesses without falling victim to these same issues. Cloud computing, or network-delivered services and software, can save customers up to 80 percent on floor space and 60 percent on power and cooling costs, and deliver triple asset utilization(1). While the economics are compelling to businesses of all sizes, concerns over security, data portability and reliability are causing reluctance among enterprise customers.” – [IBM Cloud Computing Press Release](#)

We need a new computing model to handle the maturing role of the Internet, the rise of social networking, globalization, the availability of global resources, and the need to access information in real time. These are all becoming interconnected phenomena – and the advancements in technology are driving their growth at breakneck speed. Cloud computing has emerged as the model that will help address these issues and that will greatly change the way people acquire, deploy and manage IT services. We have seen this type of major shift before when (a decade ago) traditional businesses were faced with decisions on how to embrace the web for innovation. Some thought it would be a web model overtaking the traditional IT model. However, what emerged was “e-business”, a model that merged the best of both styles of computing.

So what does cloud computing bring to the equation? The banking industry is a good example of how the business model has changed and how cloud computing can help. Thirty five years ago, people used human tellers at a branch location to make cash transactions. Banking hours were highly limited—what didn’t take place by 3 p.m. on a Friday had to wait to 10 a.m. the following Monday. In the late 1970s, banks introduced a new innovation that revolutionized the industry: the automated teller machine (ATM.) For the first time ever, money was made accessible 24 hours, seven days a week at a physical location. Enter the 1990s and the Internet. Suddenly, instantaneous access to a host of financial services is made available from a home computer. In a few simple clicks, users today can transfer money from a savings to a checking account, apply for a credit card and even take out a loan. This model is changing, too. By embracing emerging technologies, like Web 2.0 and cloud computing, the financial services industry will have the capacity to dynamically deliver even more banking services to end users in the form they choose. And, these technologies also allow banks to more effectively address the back-office operational overhead that cuts into profit. How is this possible?

Cloud computing un-tethers the applications from the underlying physical infrastructure and delivers them to the end user over the internet or intranet, meaning that access to computing can be done without a direct connection to the computer. In the 21st century, the intelligence resides in the infrastructure.

To summarize, the demand for people to be connected globally, to do business transparently, and to take computing to new frontiers, combined with the proliferation of smart devices and connected objects, will generate massive amounts of data. Turning that data into insight creates the opportunity to make organizations, industries and our world more intelligent and dynamic. In this new environment, business need to address several issues to remain competitive and “in the game”, including rapidly implementing new ways to interact and collaborate, accessing and sharing information on demand, accessing scalable computing power and more intelligent networks transparently, while reducing manpower needs, physical space requirements, and energy consumption. Cloud computing enables this business requirement by making new applications and services available through highly efficient virtualized compute resources that can be rapidly scaled up and down in a flexible yet secure way to deliver a high quality of service. In essence, cloud computing “flattens” the infrastructure by un-tethering the applications from the underlying physical infrastructure and delivers them to the end user over the internet or intranet. Figure 2 summarizes the main characteristics of cloud: **advanced virtualization** of IT resources that can **scale elastically** to handle demand and growth, which are **rapidly provisioned** for users based on their needs and delivered through **flexible pricing** models.

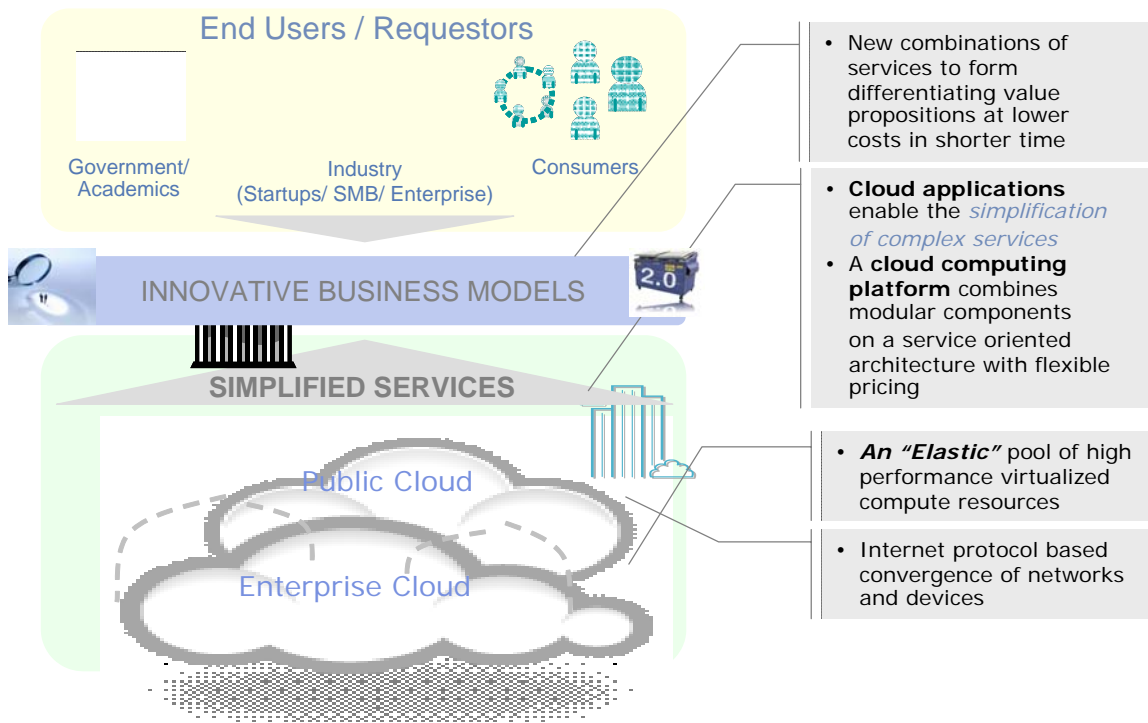


Figure 2: Cloud Computing Characteristics

Cloud Computing can be viewed from different perspectives:



From a user perspective, cloud computing provides a means of acquiring computing services via the internet while making the technology beyond the user device almost invisible.

From an organization perspective, cloud computing delivers services for consumer and business needs in a simplified way, providing unbounded scale and differentiated quality of service to foster rapid innovation and decision making.

Cloud computing provides anytime, anywhere access to IT resources delivered dynamically as a service. As an acquisition and delivery model of IT services, if properly used within an overall IT strategy, it can help improve business performance and control the costs of delivering IT resources to the organization. Cloud computing can be seen as an evolution from grid computing that focused on solving large problems with parallel computing, to utility computing that focused on offering computing resources as a metered service, and to software-as-a-service that focused on network-based subscriptions to applications.

The cloud services can range from infrastructure-as-a-service that facilitates access to IT resources (including servers, storage and networks), to platforms-as-a-service that facilitates access to platforms ranging from O/S through middleware, to applications-as-a-service that facilitates access to applications, processes and information. Each of these cloud services can be delivered through a public cloud, a private cloud or a hybrid of the two.

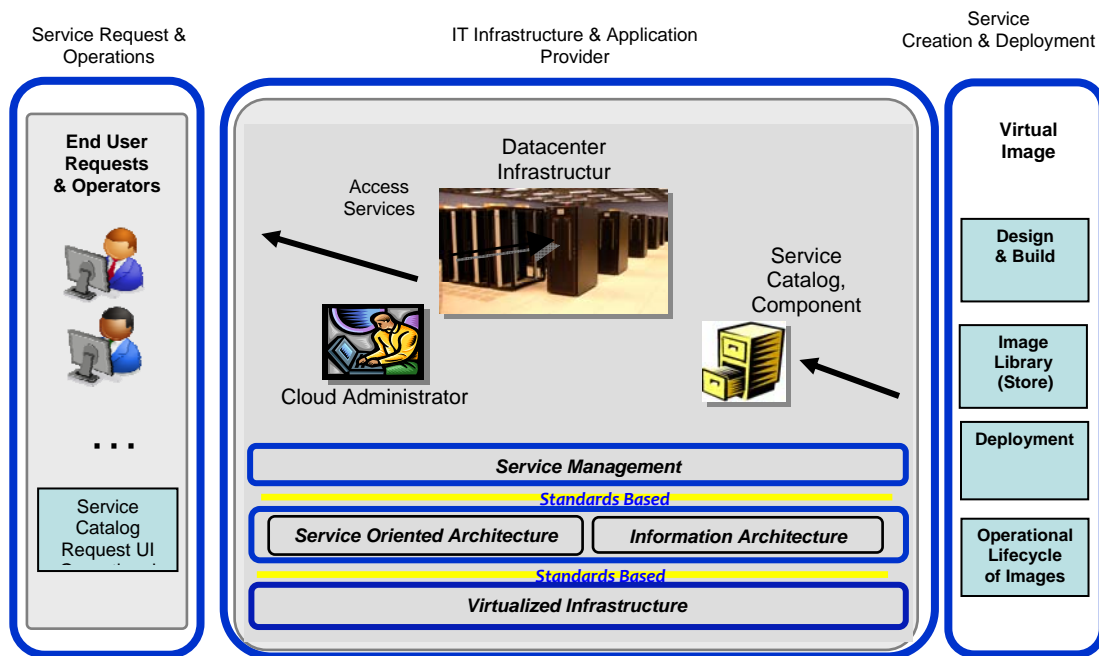


Figure 3: Cloud Computing Reference Architecture

Figure 3 shows the major components of IBM's proposed Cloud Computing Reference Architecture, and provides a guide to understanding the capabilities required for an enterprise to derive value from implementing cloud. The architecture follows the separation of concerns principle, and subdivides the capabilities of cloud into the three main "parties" – the consumer of the cloud service, the provider of the cloud service, and the creator of the crowd service.

The cloud service creator component has capabilities targeting all aspects of the lifecycle of the "image" that is used to bundle the service that is accessible by the consumer. This "image" can include the IT resources (e.g. server, storage, network), the operating system, the middleware, and the applications. The capabilities support the need to design and build the image, store it in the library of images that can be accessed by the users, deploy of the images into the cloud, and manage of the images through the entire operational lifecycle.

The cloud service consumer serves both the end users and the operators that manage the infrastructure. The capabilities in this component ensure that the images that can be accessed are defined in a catalog, and have appropriate role-based user interfaces to access and manipulate the images.

The key capabilities of the reference architecture are defined in the service provider component. The lowest layer of the architecture defines the capabilities of the virtualized infrastructure. These capabilities facilitate virtualization of all IT resources: server, storage and network. These virtualization capabilities can handle all types of IT resources, e.g. both mainframe and distributed servers. The next layer provides an optimized middleware with capabilities for image deployment, integrated security, workload management, and high-availability. The optimized middleware is used as the way to deliver services and information built according to well defined SOA and Information architectures. The central piece of the component is service management which provides the capabilities to manage a cloud service. These services include capabilities to handle user requests: managing the self service requests made by users, the lifecycle of images, and the provisioning of images based on the request. The capabilities also handle many of the qualities of service associated with delivering images, including availability, backup and restore, security and compliance, and performance management. To facilitate delivery through flexible models, the capabilities also support usage accounting and license management.

The reference architecture provides a comprehensive set of capabilities to ensure that cloud services can be built, deployed, accessed, delivered and managed. Each of these capabilities are supported by the appropriate standards, technologies and tools – all integrated to work together to deliver cloud computing.

3 JOURNEY TO THE CLOUDS



The journey to the cloud is no different than the SOA journey that we have been discussing over the last couple of years:

- **Transition Planning and Project Prioritization:** A key component of the journey is to establish the business goals, and then establish a roadmap for the journey. The transition planning is done by determining the as-is and to-be states, and defining a roadmap to establish the capabilities needed. Cloud Computing will never be established in one big-bang project. Instead it will be implemented incrementally and the business will demand return at each incremental project step. Consequently, the roadmap must have individual project plans to meet the most immediate goals of the business and yet created in a way that is consistent with and helps the enterprise move toward the goals articulated in the strategic vision.
- **Cloud Computing Methods:** Just as the object-oriented approach and SOA required analysis and design methods, we need similar methods to support solution modeling that is consistent with implementing cloud computing. These methods must help us address all aspects of cloud – from virtualization of IT (resources, operating systems, middleware, applications) to delivering and managing services based on user requests.
- **Governance:** Any organization transitioning to cloud computing and wanting to be successful will have to deal with big challenges including changes in behavior, ensuring rules and policies are followed, making the right decisions, and facilitating communication and collaboration. Governance establishes decision-making rights along with the associated policies and mechanisms to control and measure how these decisions are carried out.

In summary, the Infrastructure Architecture is key to aligning IT to business which is in turn an integral part of ensuring a successful globally integrated enterprise. We have introduced cloud computing as an emerging model that will facilitate the dynamic delivery of IT services. Fasten your seat belts and join me over the next few articles for a journey into the clouds!

About the author



Mahesh Dodani is a software architect at IBM focusing on Cloud Computing. His primary interests are in enabling communities of practitioners to design and build solutions that address complex business needs and deliver value. He can be reached at dodani@us.ibm.com.