Charting IBM's Smart Journey from Linux to Cloud Computing: Maximizing Business Value while Mastering the Odyssey of Scale, Complexity, and Escalating Costs

Sponsored by IBM

Srini Chari, Ph.D., MBA

December, 2009

mailto:chari@cabotpartners.com

Introduction

In today's climate, companies must innovate with flexibility and speed in response to customer demand, market opportunity, regulatory changes, or a competitor's move. However, the current economic downturn and the escalating energy and people costs for IT, will force companies to reevaluate how they can maximize their return on IT investments. They will need smarter approaches to reduce costs, manage complexity, improve productivity, reduce time to market, and enable innovation. Simply put, companies must and will carefully examine the business (value and costs) case of IT investments as follows:

Value

- Business Value: e.g. customer revenues, new business models, compliance regulations, better products, increased business insight, and new breakthrough capability,
- Operational Value: e.g. faster time to results, more accurate analyses, more users supported, improved user productivity, better capacity planning,
- IT Value: e.g. improved system utilization, manageability, administration, and provisioning, scalability, reduced downtime, access to robust proven technology and expertise.

Costs

- Data Center Capital Purchases Avoided e.g. new servers, storage, networks, power distribution units, chillers, etc.
- Data Center Facilities Not Built e.g. land, buildings, containers, etc.
- Operational Costs: e.g. labor, energy, maintenance, software license, etc.
- Other Costs: e.g. deployment and training, downtime, bandwidth, etc.

Through in-depth research and interviews, this paper charts the industry and IBM's journey from Linux to Cloud Computing. In 2008, building upon prior initiatives in service-oriented architectures (SOA), Linux, and energy-efficient dynamic infrastructure, IBM announced a significant company-wide cloud computing initiative, tying its systems, software, and services businesses; driven and governed by a central cross-brand group. Since then, IBM outlined its cloud computing vision, strategy, and a detailed roadmap of specific Linux-based solution offerings deployed internally at IBM and at customer sites with the following bottom-line business benefits:

- ***$1.5 billion in savings over four years*** for IBM's internal IT transformation driven by the CIO,
- ***10x reduction in deployment time*** for new business services at a major South African bank,
- ***84 percent reduction in TCO*** using the IBM innovation cloud implemented by the office of the CIO
- ***Accurate, fast results in about 2 months instead of the typical 9*** at a petroleum services provider
- ***Improved automated service management and flexibility*** for a cloud services provider in Europe.

Taken together, we believe, IBM's portfolio of Linux-based cloud solutions and their future roadmaps will enable and empower enterprises and users - across multiple industries, geographies, and user scenarios - to maximize business value while overcoming the challenges of escalating costs, complexity, volatility, and scale. In today's environment of flat or shrinking IT budgets, IBM's cloud initiative is in many ways a natural evolutionary step – but one that could change the IT industry's current conversation and definition of what cloud computing is.

From Linux to SOA to Cloud Computing – a Business Driver View Over the Last Decade

The journey to cloud computing is, in many ways, natural and evolutionary; building upon prior overlapping trends (Linux, SOA, virtualization, consolidation, etc.) that have significantly transformed the IT industry. While, here, we only chart some key trends over the last decade, it is equally important to recognize that several key facets of cloud computing e.g. time sharing, virtualization, metering, etc. have been in play since the advent of the IBM mainframes in the 1970s.

The rise and growth of Linux: Shortly after the dot-com bust and during the brief recession of the early 2000s, companies entered a phase of delaying or reducing new IT capital expenses (CapEx) and reining in operational expenses (OpEx). Then, Linux and subsequently open source software and web solutions adoption grew rapidly as customers benefited from broad access to attractively priced solutions that worked on multi-vendor hardware spanning multiple architectures while preserving traditional rich Unix-like features of ease-of-configurability, management, reliability, and resilience. Moreover, with access to the Linux source-tree, customers were able to rapidly deploy and customize their IT environments.

Emerging from this recession in 2002, customers further benefited from the advent of enterprise class Linux distributions on enterprise grade hardware such as the IBM System z and Power systems. Today, Linux is ubiquitous and is expected to grow faster than alternative operating systems[1]. Linux supports the widest range of architectures and hardware devices and it performs, scales, is modular and energy-efficient, and supports an extensive and growing ecosystem of ISV applications[2].

The emergence and value of SOA: While Linux did an admirable job of reducing IT infrastructure costs and complexity within enterprises, it was only through the transformative power of service-oriented architecture (SOA) implementations did enterprises unlock and realize significant added business value. SOA is an approach to design software in building-blocks, components, or "services" often achieved by splitting up functionality from existing applications. These services can be used independent of the applications and platforms they were previously a part of. Companies can then quickly integrate and group these services in various ways to efficiently create new capabilities, products, or services often leveraging web technologies.

With SOA, many enterprises, often under the sponsorship and direction of business executives, have been able to meet their current and future business needs, build the flexibility to deploy new business services, drive automation and collaboration within multiple organizations in the business and with suppliers and partners, and foster innovation at an unprecedented scale and speed[3]. In addition to achieving significant additional cost reductions, these enterprises have been able to increase revenues, build new products/services, ensure compliance with changing regulations, and implement innovative business models.

Addressing the data center crisis with centralization and virtualization: The electricity costs in data centers to power and cool hardware are expected to increase by 11.2 percent while new server spend is expected to remain almost flat[4]. The issue of power and cooling has become a top priority for IT executives and a crisis at enterprise data centers. The Uptime Institute has also defined additional metrics[5] that define data center "greenness". Bringing sound environmental and management principles to bear in capacity planning and operating a data center can become a competitive advantage and a source of operational efficiency and increased reliability for many information-intensive industries.

Data centers typically account for 25 percent of total corporate IT budgets when facilities, storage devices, servers, and staffing are included. The costs of operating a data center facility is growing by as much as 20 percent a year, far outpacing overall IT spending[6]. Moreover, the investment required to launch a large-enterprise data center has risen to $500M, from $150M, over the past 5 years and larger data centers take 2

---

[1] Al Gillen, "The Opportunity for Linux in a New Economy", April, 2009, http://www.linuxfoundation.org/sites/main/files/publications/Linux_in_New_Economy.pdf
[2] Amanda McPherson, "Linux: The Operating System of the Cloud", http://www.linuxfoundation.org/sites/main/files/publications/linuxincloud.html
[3] Jay DiMare, "Service-oriented architecture: Measuring SOA's ROI in the New Economic Environment", IBM Institute of Business Value, ftp://ftp.software.ibm.com/common/ssi/pm/xb/n/gbe03202usen/GBE03202USEN.PDF
[4] Jed Scaramella, "Worldwide Server Power and Cooling Expense 2006-2010 Forecast", IDC, September 2006.
[5] John R. Stanley, Kenneth Brill, and Jonathan Koomey, "Four Metrics Define Data Center "Greenness"", White Paper, Uptime Institute.
[6] McKinsey on Business Technology, Innovations in IT Management, Number 14, Winter 2008.

years or more to design and build and are expected to last for 12-15 years. Yet, server utilization generally tops out at 5 to 15 percent, wasting both energy and employed capital. Many IT managers indicate that excess servers exist to provide for extreme situations e.g. holiday seasons. However, this assertion may not be completely true.

The mismatch could be because many companies have difficulties in accurately forecasting workload demands. The long-term (12 years or more) lifecycle of a data center investment coupled with the lack of a holistic and integrated view of future workload needs make enterprise-wide capacity planning a daunting task. Many companies would have difficulty forecasting whether a 50% increase in demand would require 20% or 80% more server and data center capacity. In the extreme, companies that rely entirely on in-house computing capacity may be stuck with excess wasted space in a datacenter, or may have to undertake the large expense of building a new data center almost immediately.

What's needed to address this data center crisis is a comprehensive approach with a centralized governance model in which the CIO is empowered by the CEO and is accountable and responsible for data center management. Enterprise CIOs can reform data center operations to further increase enterprise value by:

- managing IT assets aggressively through virtualization, consolidation, and standardization,
- implementing new metrics for data center efficiencies that account for energy, utilization, and floor space,
- improving forecasting and minimizing deviations from real demand through better capacity planning and management processes,
- treating all IT resources as scarce resources or "services" that can be quickly scaled up or down, and,
- ensuring that business units implement a total cost of ownership (TCO) or return on investment (ROI) model for new systems and applications.

While these recommendations are primarily targeted at improving an enterprise's internal data center operations, they constitute the use of some key enabling technologies and policies which - in addition to Linux, SOA, energy-efficient systems and storage, better systems and power management software - form the foundation of cloud computing services.

As Cloud Computing Mainstreams, Multiple Delivery Models Emerge. Most are on Linux.

In many ways, cloud computing is the next logical evolutionary step, building upon the industry's rapid adoption of Linux, open source solutions, SOA, and more recently virtualization. While many definitions and analogues to the electric utility industry exist, we believe that most do not completely capture the cloud's broader concepts and business benefits. Here, we present a perspective that motivates this discussion with cloud business models that parallel the passenger airline transportation industry.

The value of cloud computing: Cloud computing promises to provide dynamically scalable and often virtualized IT (hardware, software, and applications) resources as a service transparently to a large set of users who may possess a broad but differing range of knowledge, expertise in, or control over the technology infrastructure. The concept incorporates software as a service (SaaS), Web 2.0 and other popular, recent, Internet computing trends such as SOA, and also builds upon recent IT infrastructure solution concepts such as grid computing, utility computing, and autonomic computing.

With a spectrum of flexible offerings and pricing models, cloud services providers are well-poised to provide secure, affordable, elastic, often automated with "self-service" access to IT resources for companies that need to quickly scale-up or scale-down their IT needs to adapt to their business demands. Cloud computing can transform companies of all sizes to become more agile and develop sustainable competitive advantage while reining in costs.

With cloud computing solutions, smaller companies – that typically face steep entry cost barriers to access IT resources - such as Internet companies, service providers, or Independent Software Vendors (ISVs) will no longer need large capital outlays in hardware or facilities to deploy their services or the labor to operate these IT facilities. On the other hand, larger companies benefit from the increased business value resulting from the added capability and flexibility to rapidly deploy standardized yet customizable "self-service" solutions that automate and scale business processes end-to-end while minimizing escalating labor and infrastructure costs.

A broad range of business and delivery models: Over the last two years, with the increasing interest in cloud computing, many excellent articles[7, 8, 9, 10] have characterized or defined cloud computing. Briefly, following typical IT architectural stacks, cloud services are delivered as: Infrastructure – servers, storage, etc. - as a Service (IaaS), Platform - a software development environment – as a Service (PaaS), Software – typically applications – as a Service (SaaS), or even Business processes as a Service. These services can be implemented and delivered as a private (within an enterprise) cloud or as a public (accessible through the internet) cloud, or as a secure hybrid (extending private) cloud. While analogues to the electric utility industry are popular, we believe, from a customer perspective, it is more illustrative (and realistic) to depict the wide spectrum of cloud business model choices with an analogy from the passenger airline industry as summarized in the following figure.

IT Industry	Transportation	Cost Model	Business Impact
In-House Data Center or Private Cloud (not-outsourced)	Own planes	CapEx, OpEx, long term	Total control and responsible for entire infrastructure including service levels
Private Cloud (outsourced)	Lease planes and service	CapEx and OpEx in medium term lease	Control of infrastructure but not fully responsible for service levels
Hybrid	Charter planes	Pay for duration, short to medium term/trip	Some control of infrastructure and depend on provider for service levels for duration
Public Cloud	Commercial airlines	Pay as you use, very short term/trip/user	Almost no control of infrastructure or service levels

While the evolution of public cloud adoption has been rapid, particularly with smaller businesses and individual developers, early adopters at larger enterprises are increasingly turning to private and hybrid clouds to address concerns (with public clouds) of security, regulatory compliance, governance, reliability, and IP protection. These larger enterprises, who already have substantial in-house IT investments, are implementing private or hybrid clouds to improve utilization levels, reduce operational expenses, and can - through "self-service" portals - dynamically provision IT services in minutes or hours instead of months. With Linux powering their cloud implementations, these enterprises will be able to rapidly develop, customize, test, and roll-out new business services and applications to their users and clients while reining in costs.

Today, most large public cloud-service providers such as Google and Amazon use Linux[2,11] in their large-scale, high-demand, extremely-elastic cloud environments – another testament to the value of Linux for cloud computing.

[7] Wikipedia
[8] Tim Jones, "Cloud Computing with Linux", http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux/l-cloud-computing/l-cloud-computing-pdf.pdf
[9] Jeffrey Rayport and Andrew Hayward, "Envisioning the Cloud: The Next Computing Paradigm", March, 2009.
[10] Ashar Baig, "A Cloud Guide for HPC", May 2009.
[11] WebHostingUnleashed, "Cloud-Computing Services Comparison Guide", http://www.webhostingunleashed.com/whitepaper/cloud-computing-comparison/

Why is Linux the Ideal Operating System for Public and Private Clouds?

The TCO advantages of Linux as an applications server operating system over other operating systems including Solaris and Windows has been well-documented[12]. Linux is versatile; even Windows desktop users can transparently use Linux clouds on the back-end. Moreover, Linux supports the widest range of architectures and hardware devices and it performs, scales, is modular and energy-efficient, and supports an extensive and growing ecosystem of ISV applications[2]. Using Linux-based clouds, customers can:

Build a flexible scalable architecture: The Linux kernel supports a degree of componentization and customization that is unrivaled among general purpose operating systems, and is remarkably adaptable to a broad range of computing environments. It's configurable and powers everything from a hand-held device to a laptop to a supercomputer. Linux today supports more hardware devices than any other operating system in the history of the world[13]. This is particularly valuable in highly customized, large-scale cloud environments, which typically run on heterogeneous ensembles of servers, networking, and storage systems. Customers will also take advantage of the ability to modify the Linux source code in order to optimize and customize the kernel to their specific needs and hardware.

Drive standardization, interconnectedness, and collaboration: As cloud offerings continue to grow, there's evidence of incompatibility between these nascent offerings[14]. The lack of standards and the resulting potential for lock-in into a specific vendor's proprietary islands of insecure and hard to manage offerings, are key concerns. To achieve the promise of cloud computing, there must be open standards around data formats, packaging, security, runtimes, and management. Fortunately, the differences between Linux instances in cloud environments and those hosted locally or at a data center are in general minor. For the time being, customers can leverage Linux and service-oriented approaches, as cloud offerings continue to evolve and the industry arrives at standards. By standardizing on Linux workloads, customers will have the unique flexibility to deploy private, public, or hybrid clouds to drive an unprecedented level of interconnectedness and collaboration.

Leverage an extensive application and ISV ecosystem: With thousands of Linux compatible and certified applications available, customers have many options for their specific workloads. By standardizing on Linux in their departments and data centers, these users will be able to transparently extend these capabilities to clouds – private, public, and hybrid.

For cloud-service providers wanting to transfer cost-advantages to their clients, Linux is the logical choice. Like web and service-oriented architectures, many public cloud computing platforms often leverage other open source projects and industry alliances, from databases to file systems to application and web servers to language runtimes to device drivers. By virtue of its ubiquity, quality, scalability, performance, and open source nature, Linux has become the first choice development and deployment system for these projects. As a result, cloud service providers and their customers benefit from this large application portfolio.

Benefit from attractive licensing costs and terms: While there still exists some misconception that Linux is always free, the majority of enterprise production deployments are commercially licensed and supported. Despite this, a comparative application server TCO study[12] found that Linux could reduce on-going licensing and maintenance costs especially when per-processor licensing models are used. The study participants who used Linux drove their systems to higher utilization levels and achieved greater consolidation benefits.

However, cloud service providers do have a compelling option to run non-commercial distributions. Cloud service providers such as Google and Amazon have indeed chosen this path as they do possess adequate

---

[12] Robert Francis Group, "TCO for Application Servers: Comparing Linux with Windows and Solaris, http://www-03.ibm.com/linux/whitepapers/robertFrancesGroupLinuxTCOAnalysis05.pdf

[13] "The Linux Driver Model: A Better Way to Support Devices", https://www.linuxfoundation.org/en/Linux_Device_Driver_Model#Executive_Summary

[14] Bob Sutor, "To Deliver on the Promise of Cloud Computing, Follow an Open Path", Linux + DVD, April, 2009, pp. 62-63.

internal skills and capabilities to support a non-commercial Linux version, and avoid the uneconomical licensing costs for tens of thousands of nodes with traditional per-processor licensing models. By leveraging this lower cost approach, cloud providers can pass on the savings to their customers.

Operate energy-efficient IT infrastructures: Beyond its advantages in licensing, Linux is a very cost-effective system to enhance data center energy efficiency. With increasing usage in small, power sensitive devices such as netbooks, embedded devices, and other mobile devices, Linux has benefited from a great deal of research in lowering total energy consumption. With features like the tickless kernel and PowerTop[15] that avoid or mitigate traditional high-alert processor time-keeping in somnolent or quiescent states, Linux has become more energy-efficient relative to competitors. When combined with other energy reduction strategies[16] within cloud data centers, Linux helps significantly lower the TCO for cloud customers.

Access and use a large existing skill pool and IT resources: Enterprise IT organizations are very often challenged to hire and train the best skills to run and operate their environments. They seriously consider and evaluate this large expense during the purchase and deployment of technology. Fortunately, because managing and developing for Linux are common skills, customers can easily access a large existing skill base or hire new personnel easily in the market. By deploying Linux-based clouds, these customers can avoid costly re-training for system administrators and developers and also better leverage existing IT investments that already work in Linux environments.

Improve patch management, security, and availability: Keeping the IT systems up-to-date and compliant with the corporate security requirements is complex, because some vendors release a large number of updates on a frequent basis, the tracking and evaluation of the appropriate software updates is time-consuming, the patch deployment must be managed to reduce downtime, and auditing and reporting is required. In a study[12] comparing Linux deployments with Windows and Solaris, it was found that it is easier to lock down a Linux system and deploy patches, and that Linux patch deployments reduced downtime. This also reduces security vulnerabilities and improves reliability and availability; factors crucial to combat the much publicized recent concerns with public clouds. With Linux on enterprise grade systems such as the IBM System z mainframes, enterprises could further improve reliability and security on clouds.

Enhance utilization and TCO through virtualization and consolidation: Today, virtualization is a mainstream technology in most data centers, enterprises, and cloud platforms. Simply put, virtualization enables the ability to abstract operating system or application instances from the underlying platform. With virtualization, Windows or Solaris images or applications, can be hosted and run on top of a Linux platform. At the core of a broad array of virtualization technologies are hypervisors that actually make virtualization possible. With complementary "self-service" systems management tooling, virtualized IT resources can be dynamically assembled and provisioned in a matter of minutes instead of months.

Coupled with live migration features, virtualization can consolidate workloads to improve data center utilization and energy-efficiency, and also bridge into cloud environments. Linux when coupled with virtualization technologies like VMware, KVM or Xen can be the host or guest cloud platform.

Optimize workloads to drive efficiencies and innovation across the business: Linux supports the broadest range of workloads from traditional transaction processing to high-performance computing (HPC). As cloud computing evolves, these workloads will move to the cloud at differing rates. Many web and HPC workloads have already moved to the cloud. With over 78% penetration[17], Linux dominates the Top500 list of supercomputers – yet another testament to the affinity of Linux to cloud computing.

---

[15] Stephen Shankland, "Linux Coders Tackle Power Efficiency", http://news.cnet.com/Linux-coders-tackle-power-efficiency---page-2/2100-1007_3-6192865-2.html?tag=mncol
[16] Data Center Efficiency, http://www.ibm.com/itsolutions/energy-efficiency/
[17] Sean Michael Kerner, "Linux dominates top 500 supercomputer list, November, 2009, http://blog.internetnews.com/skerner/2009/11/linux-dominates-top-500-superc.html

Workload Classification: Making the Case for Cloud Computing

Accurately forecasting application workload requirements is a major challenge for IT managers and planners. This challenge becomes even more acute as businesses depend increasingly on high performance analysis and web applications which have more variability than traditional enterprise business applications. Using a simple workload characterization and classification model developed here, we examine workload trends that are poised to fundamentally transform the delivery of IT solutions through cloud computing.

First, we examine a wide range of workloads typical in many IT applications across several dimensions. These applications range from traditional enterprise business and transactional applications to more compute intensive High Performance Computing (HPC) applications and web/business analytics. These applications are classified according to their typical workload characteristics with compute-intensive/job on the x-axis and workload variability[18], V_W, on the y-axis. The bubble size in this chart is indicative of the total server capacity deployed globally to execute these applications. The arrows give an idea of workload growth in the future across the two primary dimensions. For example, traditional transactional applications and ERP comprise a large part of the workload today but they are not very compute-intensive and exhibit low variability. On the other hand, web analytics is an emerging area that is currently small but expected to grow rapidly. HPC applications are normally very compute intensive often requiring 100s of CPUs to execute one job and since these applications are often used for complex analyses, workload variability is large and frequently difficult to predict *a-priori*. Web searching capability is becoming deeper and more complex with multi-modality capability beyond simple text searches. With more users using complex search, transactional web applications, and web analytics, we expect the web workload to become more compute-intensive with increasing variability. Business analytics is a large portion of the workload and is becoming more variable and compute-intensive. The large development and test environments have cyclical patterns with increasing variability to adapt to severe pressures to deliver new products and services at ever reducing cycle times. And finally, with the interconnected and mobile nature of today's business environment, collaboration workloads are large, will grow, and become more variable.

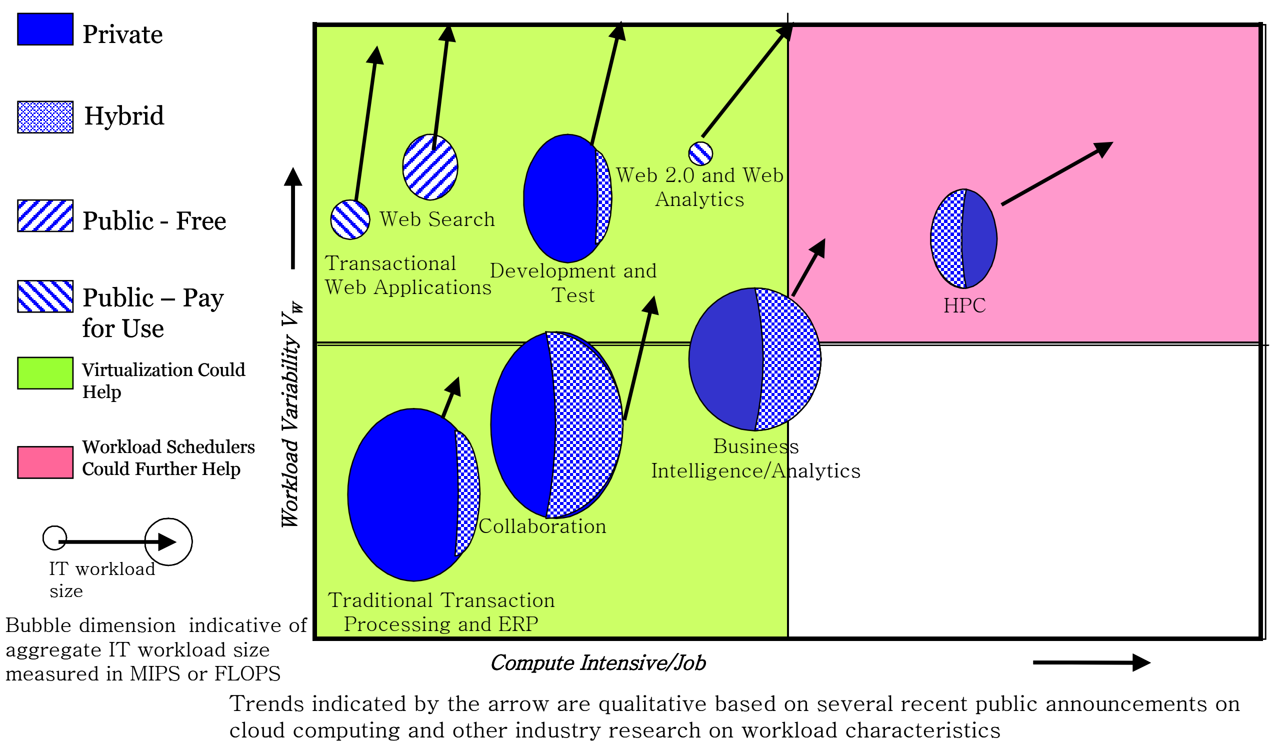


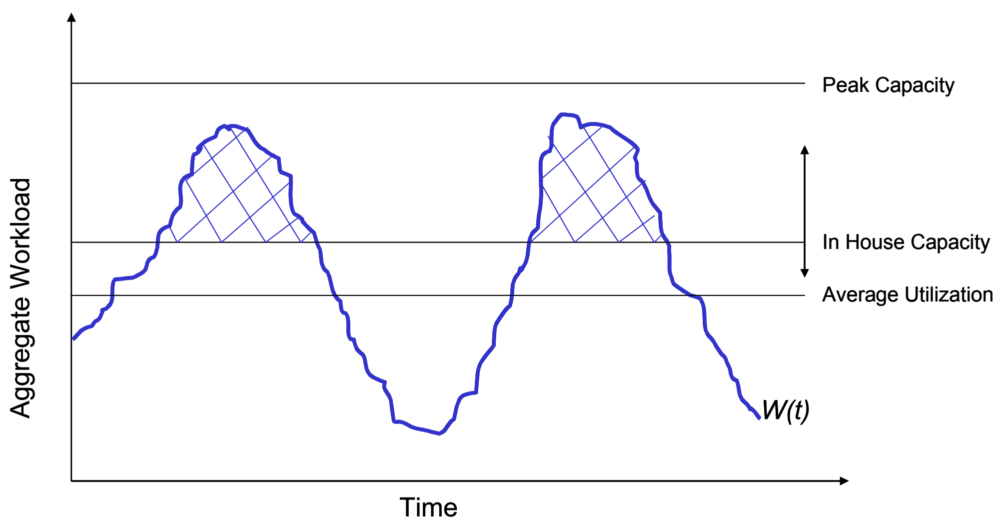
Figure 1: High Level Workload Classification by Application Domains

---

[18] Srini Chari, "Confronting the Data Center Crisis: A Cost Benefit Analysis of IBM's Computing on Demand Cloud Offering", http://www-03.ibm.com/systems/resources/tcocodpaperfinal.pdf

Another important dimension in the workload analysis is the way these applications are typically delivered and executed today and what may be expected in the future. Again, in a non-prescriptive manner, we break up the typical delivery models into Private (In-House), Public (public access through the Internet), and Hybrid (Internet access through Virtual Private Networks (VPNs) contracted and customized to the needs of the end-users). Public is further divided into Public-Free which is a free access to the end-user (e.g. Google search) and Public-Pay for Use (e.g. Amazon Web Services) which is a pay-for-service infrastructure utility model. As depicted in Figure 1 and consistent with the opinion of most analysts, we believe that most of the future cloud opportunity is in private or hybrid clouds.

Virtualization and workload scheduling and management are key software solutions that often help increase system utilization and overall data center efficiency. With virtualization and consolidation, many low-to-moderate compute intensive workloads can be mapped onto fewer physical systems without adversely impacting service levels. Hence, workloads depicted in the left quadrants in Figure 1 could benefit, substantially resulting in efficiency gains for customers. VMware is one prominent example of a virtualization solution. The IBM System z mainframe is another excellent example. However, at the other end of the spectrum, with HPC and other compute intensive workloads often requiring several CPUs per job, workload scheduling and management solutions from Platform and the IBM Tivoli LoadLeveler have been used to increase overall system utilization and throughput in cluster configurations. Virtualization solutions that usually consolidate several jobs onto one CPU may actually adversely impact the scalability and performance of HPC and analytics jobs especially parallel batch applications that use data partitioning.



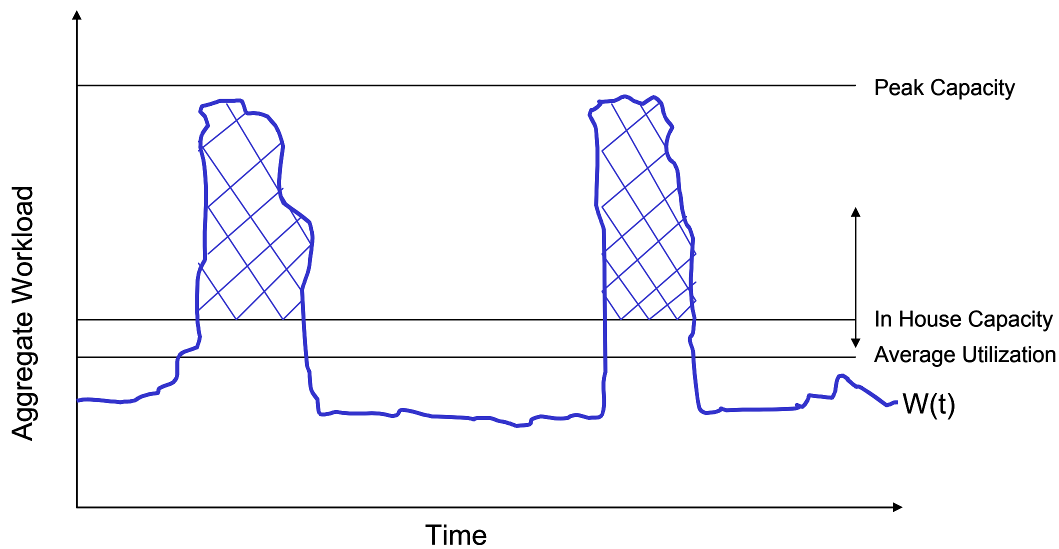
•Shaded area represents under provisioning of IT resources that could lead to dissatisfied users and adverse business impacts
•In house capacity designed for peak results in some over provisioning and wasted resources

Figure 2: Typical Low Variability Workload Curve - Common in Many Enterprise Business Applications

Figure 2 depicts a typical low variability workload for enterprise business applications such as transaction processing or ERP. The peak capacity of the data center must be over-provisioned by about 20%-30% to ensure there is enough usable capacity to prevent any compromises in service levels[19]. Thus the data center peak capacity is higher than the workload peak. If the in-house capacity is set to the peak capacity with some over-provisioning and wasted resources, the entire workload can be executed and the average utilization level would be as indicated in the figure. If however, the in-house capacity is less than the peak capacity, then the workload indicated in the shaded area would be compromised and this may result in frustrated users, loss of revenue, delay in time to results, or even no results as the jobs may be ejected from the queue. As the in-house capacity is moved up or down, the associated total cost of computing (TCO) will also go up or down. If other internal business/technical obstacles are not an issue, an enterprise can access computing resources through a cloud computing service to execute the workload over in-house capacity as

[19] Abramson, D., Buyya, R., and Giddy, J., "A computational economy for grid computing and its implementation in the Nimrod-G resource broker", Future Generation Computer Systems, 18, 8, (2002), 1061-1074.

depicted in the shaded area. This would allow them to maintain their current private in-house infrastructure and get the additional computing capacity as needed. While this could be a very economical model, many enterprises are expected to adopt this utility model gradually over the next 3-5 years especially for mission critical business applications. So we expect that many of these companies would continue to add more capacity in their data centers and implement private clouds.



•Shaded area represents under provisioning of IT resources that could lead to dissatisfied users and adverse business impacts

• In house capacity designed for peak results in excessive over provisioning and wasted resources

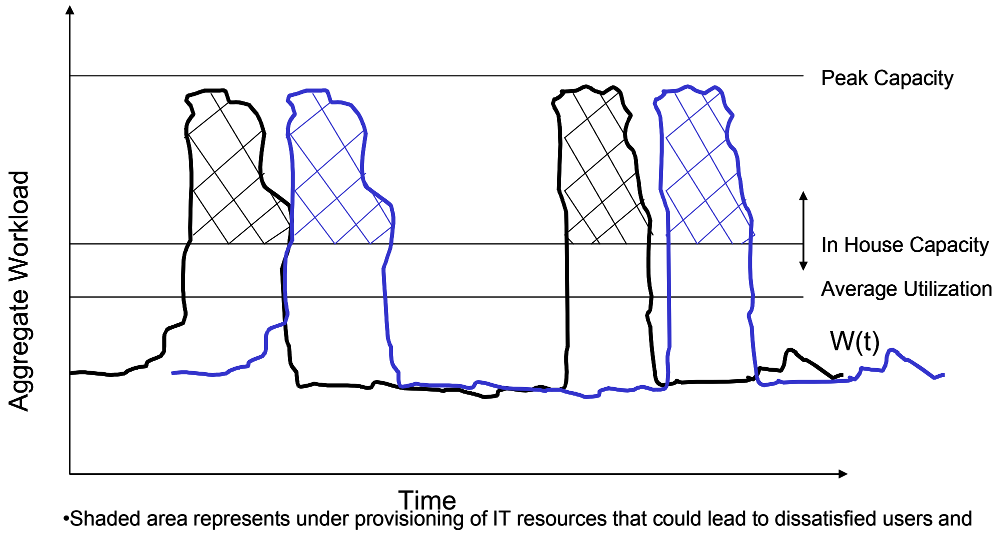
Figure 3: Pulsed High Variability Workload - Common in Departmental or Small and Medium Business HPC and Web Applications

Figure 3 depicts a pulsed high variability workload typical in many departmental or small and medium business HPC applications and web workloads such as search, analytics, and transactional web applications. HPC workloads are often highly parallel and can benefit from "cost associativity"[20] of cloud computing. So if 1 CPU takes 1000 hours to execute this workload and if this workload is perfectly scalable, then 1000 CPUs would execute this in 1 hour. Furthermore, either model on a cloud would cost roughly the same but the business benefits of this substantial reduction in the time to results could be immense. Likewise web workloads exhibit this extreme pulsed behavior during very attractive product sales promotions or rapidly breaking news events when 1000s of interested users access the system concurrently. If the in-house capacity is set to the peak capacity with excessive over-provisioning and substantial wasted resources, the entire workload can be executed but the average utilization level would be very low as indicated in the figure. If however, the in-house capacity is quite less than the peak capacity, then the workload indicated in the shaded area would be severely compromised and this will almost always result in frustrated users, substantial loss of revenues, and the inability to execute critical workload during that short duration. Adding some additional in-house capacity will provide minimal incremental business benefits. The cloud computing pay-as-you-go model is the only economically viable model for these workloads and hence has attracted many early adopters with these workloads. For these businesses, this type of cloud computing model is a business necessity – they would either go out of business or will be unable to start a new business. The IT alternatives are just economically infeasible.

Figure 4 depicts a many-pulsed, low-medium variability workload typical in large enterprises developing or testing business applications, or using multiple analytics applications. Often in these environments, several departments or locations execute workloads that would be series of phased pulses. As these pulses get added, the aggregate workload becomes smooth and could even become almost flat if sequenced (or

[20] Armbrust, M., et. al., "Above the Clouds: A Berkeley View of Cloud Computing", Technical Report No. UCB/EECS-2009-28.

scheduled) optimally for the available in-house resources. This increases overall system utilization and throughput. If the in-house capacity is set to the peak capacity, the entire workload can be executed and the average utilization level would increase as more pulses (new workloads) are added and scheduled optimally.



•Shaded area represents under provisioning of IT resources that could lead to dissatisfied users and adverse business impacts
• In house capacity designed for peak results in some excessive over provisioning and wasted resources but average utilization improves compared with the single pulse case

Figure 4: Many Pulsed Workload (Low to Medium Variability) - Common in Large Enterprise Development and Test and Analytics Environments

Virtualization, consolidation, and sometimes workload scheduling solutions are widely used in these environments and help IT managers extract the most value from their in-house resources. IT managers will be able to add capacity to existing data centers more economically and implement private clouds. If new pulse workloads are substantial in intensity – say at least 3-5 times the normal pulse, they could benefit from an additional cloud computing service which could be pay-as-you-go.

Other non-application specific IT workloads could also benefit from Linux-based private or hybrid clouds built on virtualization technologies. These include storage clouds, desktop clouds, and information archival clouds for back-up and recovery, and clouds for automated image management.

Cloud computing will dramatically impact the way IT services are consumed and delivered in the future. Our broad workload-driven approach strongly enforces the view that cloud computing is a rapidly growing IT opportunity throughout the value chain; estimates range from the tens of billions of dollars to even over hundred billion dollars in the next 3-5 years[21]. Regardless of the precise estimate, we believe:

- The market opportunity is large with growth rates much faster than the overall IT industry,
- Linux will remain pervasive throughout the cloud ecosystem; becoming even more dominant with the emergence of cloud standards to drive interoperability between clouds,
- Private and hybrid clouds will become the dominant cloud delivery models as enterprise workloads begin to leverage the promise of clouds and security concerns persist with public clouds,
- Before making substantial new cloud investments, businesses will carefully examine the business case that will be primarily driven by their current and future workload needs, and lastly,
- Customers will rely on cloud providers who have the deepest insights into their workloads and can deliver a broad portfolio of cloud services and systems optimized to these workloads.

[21] Markus Klems, "Merrill Lynch Estimates Cloud Computing to be $100 Billion Market", http://cloudcomputing.sys-con.com/node/604936

IBM Delivers a Broad Portfolio of Workload Optimized Cloud Computing Offerings

Since announcing a major corporate cloud initiative in 2008, IBM outlined, in the summer of 2009, a broad vision and strategy for cloud computing with an initial set of offerings spanning its major businesses – software, systems, and services. Again, in the fall of 2009, IBM extended these initial offerings with a broader set of offerings with increased capability optimized for specific high-opportunity workloads. Branded as IBM Smart Business services and IBM Smart Business systems (www.ibm.com/cloud), these offerings are tailored for development and test, collaboration, business analytics, and information-intensive (compute and storage) workloads behind a company's firewall or on the IBM Cloud. Of special mention are IBM Smart Business Development & Test Cloud which leverages Linux; and the IBM Smart Analytics Cloud and the IBM CloudBurst which leverage Linux on the IBM System z mainframe and on the IBM System x BladeCenter respectively. The following figure (courtesy IBM) summarizes the current – as of November 2009 - IBM cloud computing offering portfolio.

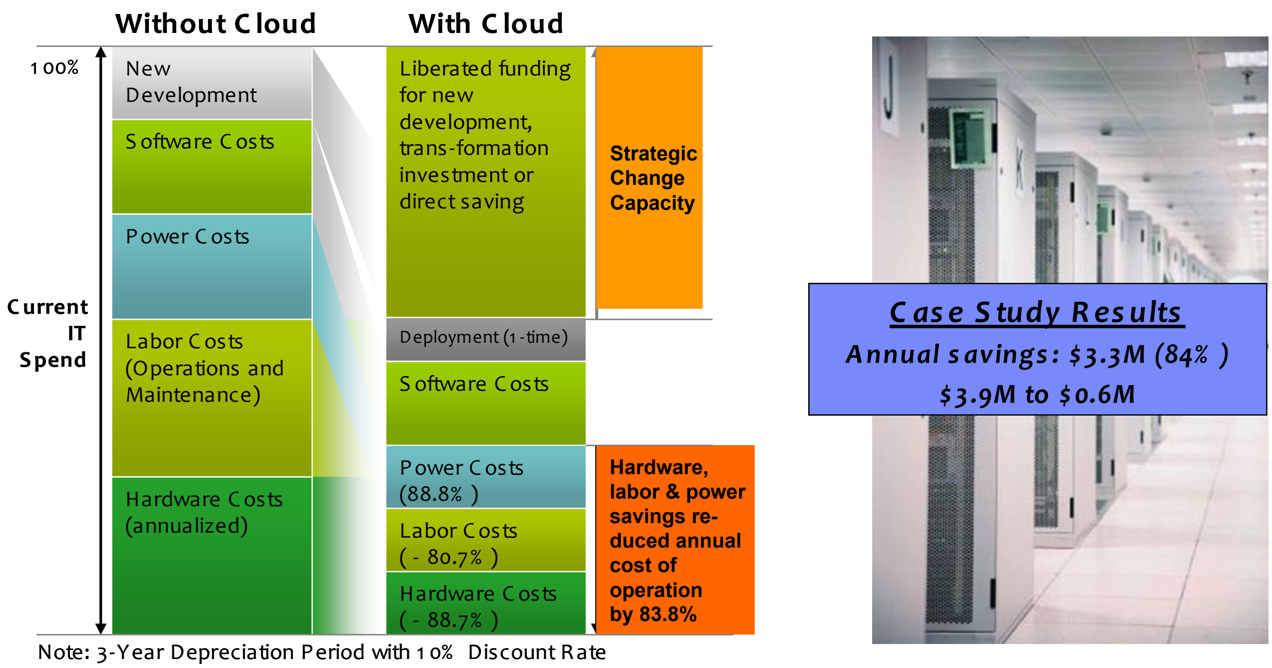
While many of IBM's cloud offerings are recent, it is important to note that IBM has been working on cloud computing for over three years for internal transformation and with early adopter customers for "in-market" experiments through (now twenty) global cloud laboratories. Lessons learned from these efforts have formed the foundation of IBM's current cloud initiative. Here, we summarize some of these efforts to demonstrate the bottom-line business benefits already achieved on Linux-based clouds.

Internal projects sponsored by the IBM CIO[22]: A global IBM internal business transformation cloud project was deployed within IBM's Technology Adoption Program (TAP)[23] designed to support innovators and early adopters of emerging technologies. This "TAP cloud" has been rolled-out for 110,000 employees worldwide. Results achieved in this transformation effort include a consolidation of 155 data centers to five,

[22] Jay Subrahmonia, "Cloud Computing Customer Examples", August, 2009, https://www-950.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/Jay%20and%20Matt%20Cloud%20Customer%20Examples/$file/Jay%20and%20Matt%20Cloud%20Customer%20Examples.pdf
[23] Avi Alkalay, *et. al.*, "Supporting Innovators and Early Adopters: A Technology Adoption Program Cookbook", IBM Redpaper, http://www.redbooks.ibm.com/redpapers/pdfs/redp4374.pdf

16,000 applications to 4,500, and $1.5 Billion in savings. Linux on IBM System z powered many of these applications and data centers.

Next, an IBM innovation cloud was implemented by the CIO's office; based on the internal implementation of the TAP cloud which has now gone into full-scale production. This cloud consists of an Innovation Portal, Cloud Catalog with many self-service system images consisting of Linux and Windows bundled with IBM software group products to enable the rapid customization of cloud environments to deliver capacity provisioning of resources efficiently. Results include consolidation and virtualization through Xen of 400 servers to 55 servers for 120 projects. As shown from the following IBM provided figure, the TCO (costs for hardware, labor, facilities, energy, etc.) was reduced from $3.9M to $.6M annually – an 84 percent reduction in TCO.



IBM-Google-NSF academic partnership: IBM and Google have partnered to promote and nurture cloud skills at over 35 universities world-wide. Today, this Linux and Xen infrastructure leverages IBM's software stack (WebSphere, Information Management, and Tivoli) in addition to open source software (Hadoop and Eclipse) with over 2000 cores, 6 TB Ram, PB Storage Space and supports:

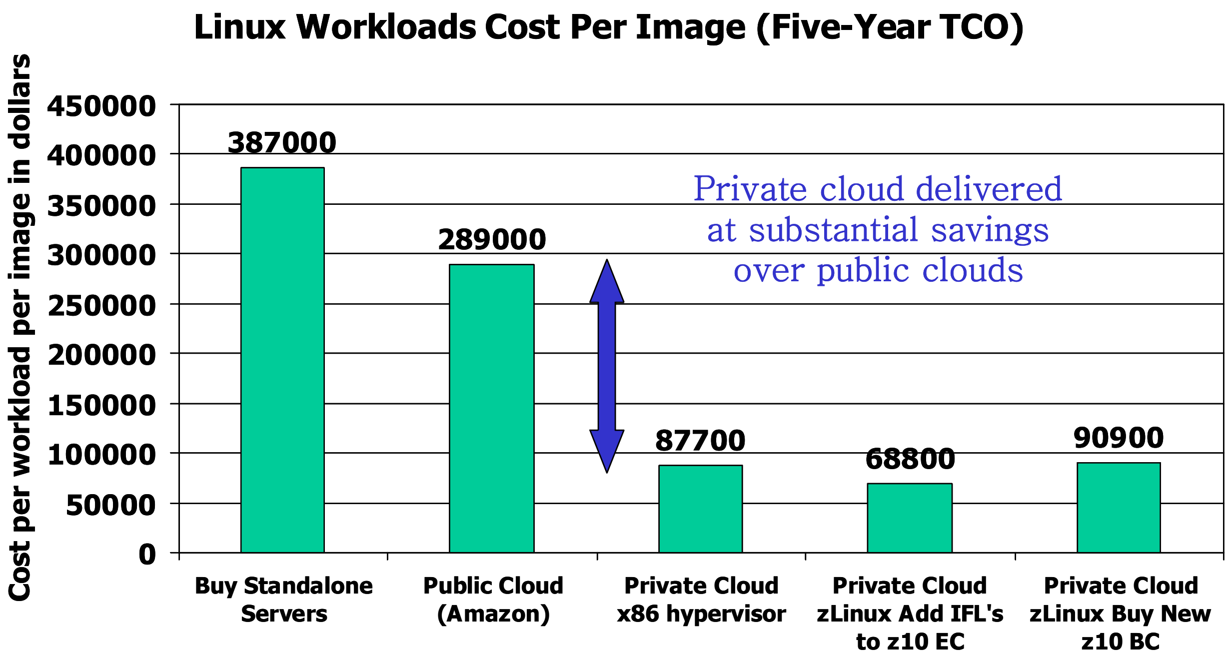
- Over 850 projects, with 50k requests satisfied
- New undergraduate programming courses
- Graduate research projects

Universities include Yale, Johns Hopkins, University of California, Berkeley, Technische Universitat, Dresden, and several other universities in Asia, North America, and Europe.

IBM benchmark on a Linux virtualized private cloud on the System z10 mainframe[24]: A recent IBM study revealed that the System z10 mainframe with Linux and z/VM as a hypervisor provided a cheaper, easier-to-manage cloud solution. A sample banking application was built using IBM Websphere Application Server and DB2 running on Linux. Several systems were benchmarked including the Amazon public cloud, and IBM x86 and z10 systems. Results depicted in the following figure show that private cloud

[24] Scott Bain, "IBM study reveals the System z10 provides a cheaper, easier-manage cloud solution", IBM Systems Magazine, September/October, 2009, pp. 37-41, www.ibmsystemsmag.com/mainframe

implementations with virtualized x86 and System z servers can be up to 75 percent less expensive than public cloud options and around 80 percent less than a distributed stand-alone server.



Source: Scott Bain, "IBM study reveals the System z10 provides a cheaper, easier-manage cloud solution", IBM Systems Magazine, September/October, 2009, pp. 37-41, www.ibmsystemsmag.com/mainframe

Customer Examples Highlighting the *Business Benefits* of IBM Cloud Implementations

Through in-market experimentation and pilot engagements over the last three yeas, IBM has been driving customer-based innovation. These early cross-brand cloud solutions, developed by IBM and partners, have initially focused on delivering rapid scalable and flexible provisioning, automation, and increased efficiencies through the combined use of technologies such as virtualization, system management, image management, and high-volume web infrastructures. More details on some of these client success stories[22] can also be found on http://www.youtube.com/ibmcloud. Another example discussed here is how a small petroleum services company has used the IBM Computing on Demand[18] solution to implement a new business model.

iTricity, a European cloud services provider – *automated service management*

Goals were to transform from a traditional hosting model to a cloud model and provide "Compliant Infrastructures as a Service" for enterprise customers.

Challenges included the need for iTricity needed to adopt subscription and usage based billing – "pay by the hour" and host multiple customers and workloads previously done through manual operations which were very cumbersome.

Solution implemented over several phases consisted of TAP components on Linux and Windows with Xen on IBM System x in the form of IBM CloudBurst enabling iTricity with rapid provisioning of server and network resources, automation, and efficiencies without manual scripting through self-service portals and virtual systems images.

Benefits included lower costs - automated service management, flexibility to extend enterprise data centers to the iTricity cloud using secure MPLS/VPLS technologies using cloud computing compliant with regulations (Basel-II, SOX, SAS-70, Healthcare, ISO 27002). Significant labor cost reductions were realized end-to-end.

A leading South African bank – ***reduced deployment time for new business services***

Goals were to automate SOA business process deployments, reduce cost, accelerate core banking transformation, and time to market.

Challenges included skills shortage, long time to deploy new business services - took ten weeks to setup and deploy test environments, multiple teams needed for each build of the software stack: Tivoli, WebSphere, DB2 support, application, and AIX.

Solution implemented consisted of IBM CloudBurst bundled with WebSphere Process Server 6.0 (Cluster), WebSphere MQ 6.0.2.2 (HACMP), ITCAM for SOA 6.1, and DB2 7.2 (HACMP), and Power Systems p570 and p595 bundled with WebSphere Application Server 6.0 (Cluster), IBM HTTP Server 6.0 (Load Balanced), WebSphere Edge Server 6.0, and WebSphere Message Broker 6.1 (HACMP).

Benefits realized with the IBM cloud include reduced deployment time for new business services by 10x to less than one week, optimized resource utilization – equipment and people, and accelerated business transformation. The IBM cloud solution also provided consistency of processes and consolidation of resources. This enabled this bank to deliver differentiated new customer services faster.

According to a Software Service Executive at the bank, *"Previously, we have to spend a lot of time supporting projects, and every time we have to provision a new environment, we take the time away from the project. Using cloud, we can provision the environment overnight (2 hours)"*.

Services Cloud, a developer of cloud services in a growth market – ***stimulated innovation***

Goals were to stimulate software start-up company growth, launch new business models, and accelerate transformation to a services-led economy.

Challenges included the need to efficiently provide software development and test tools to software startup companies for competitive services to internet users.

Solution implemented consisted of IBM CloudBurst bundled with Rational development and test tools and backup/restore capabilities with Tivoli Storage Management to protect customer assets.

Benefits realized include accelerated development and test cycles through quick resource on-boarding and a cloud with access through the Internet or a secure connection to virtualized, secure, and network isolated environments.

According to the CEO of this Cloud Computing Service Provider, *"IBM has its vision to help the Chinese government, as well as the company, to use the resource in a smarter way. Cloud computing is one of the most important platforms that enables us to do so"*.

Ingrain Rocks, a small solution provider for the petroleum industry – ***enabled a new business model***

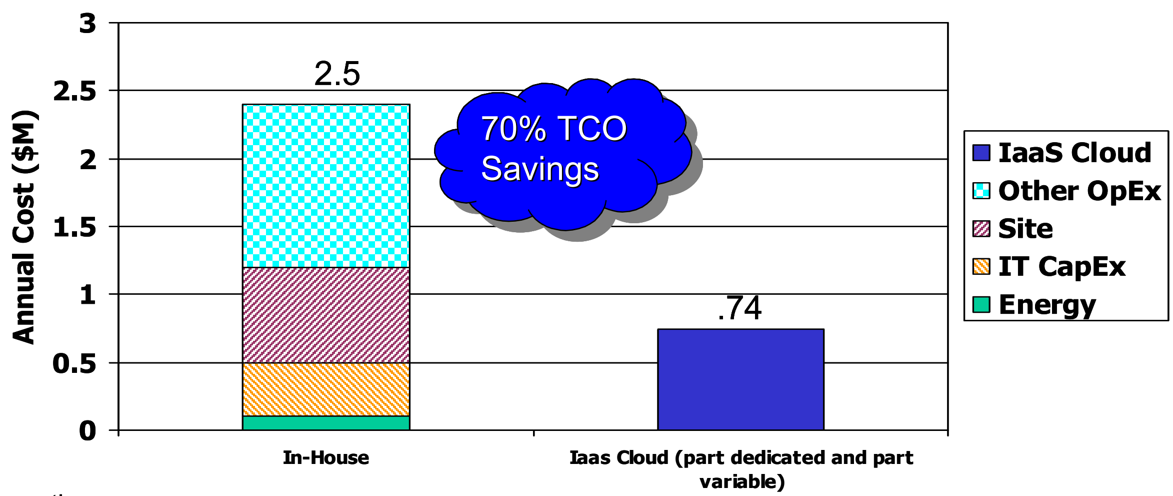
Goals were to provide radical competitive differentiation and reduce service delivery time from months to days with more accurate analysis of rock properties through HPC.

Challenges include the inability to invest in large in-house IT resources, yet needing global access to secure, flexible, and "infinite" computing resources while ensuring that IT costs track client workload. A key computational analysis step is expected to become more critical as client-delivery time shrinks; further increasing IT elasticity needs.

<u>Solution</u> implemented consisted of optimizing and deploying the Linux based parallel rock property analysis application on the IBM Computing on Demand (CoD) IaaS cloud which provided a range of systems, flexible configurations, almost "infinite" capacity; at attractive price points.

<u>Benefits</u> include the ability to tackle more challenging projects that even some of the largest petroleum companies struggle with and a 70 percent savings in TCO as shown in the following figure. In 12-18 months, Ingrain expects to have a very efficient end-to-end process that will enable them to deliver results to their clients in days instead of months.

Dual Core x86 Cluster System with 384 CPUs (sockets) and 1184 GB of Storage.



Assumptions:

1. 224 CPUs dedicated for 52 weeks at $.32/CPU hour
2. 140 CPUs variable for 12 weeks at $.49/CPU hour
3. Includes annual storage and fixed annual contract cost
4. Minimal personnel assumed for 3 shifts for In-House system

* Jonathan Koomey, "A Simple Model for Determining True Total Cost of Ownership for Data Centers", White Paper, The Uptime Institute, 2007.
* TCO model adapted by Cabot Partners

According to Mr. Barry Stewart, CFO, Ingrain Rocks, *"The IBM CoD solution is an absolute business necessity for Ingrain. It is a must have solution. IBM has provided outstanding service and support to help Ingrain migrate and optimize their applications on the CoD clusters. Ingrain has completed testing and quality assurance of their application in a very short time. Ingrain has observed a 15%-20% improvement in performance and can now scale the workload to much larger configurations".*

Why is IBM Well Positioned to Establish Leadership in Cloud Computing?

While cloud computing is still early in the lifecycle, it is a reality. Many enterprises have launched several executive sponsored cloud projects. Over the past three years, building upon prior successes with Linux and SOA, IBM has partnered with several of these early adopters through "in-market" experiments and delivered bottom-line business benefits through a three-prong strategy of:

Delivering broad heterogeneous offerings: Today, IBM delivers enterprise-class systems ranging from the IBM System x, z, and Power, plus storage; a choice of operating systems including Linux, AIX, and Windows; a range of cloud models that include private, public, and hybrid clouds, and support of a robust software stack including virtualization: VMWare, Xen, and KVM.

Starting small with secure private clouds and then scaling up: This helps enterprises maintain security, compliance, governance, and ease of deployment. With IBM Global Technology Services for strategy and deployment, early adopters have substantially reduced the risks associated with deploying emerging technologies.

Pursuing a top-down and bottoms-up approach: With a focus on the customers' strategic business objectives, and with gradual project-based deployments delivering ROI, IBM has succeeded in implementing key business and system use cases.

Now, with its broader cloud initiative and a roadmap of workload-optimized offerings, IBM is well positioned to build on its three-prong strategy, and evolve to a leadership position in cloud computing for the following additional reasons:

1. IBM will have the broadest portfolio of cloud solutions with the flexibility to offer the best mix of functionality, scale, reliability, security, and a range of consumption models delivered and supported through its Global Technology Services (GTS) organization,
2. While cloud users will be empowered through "self-service" flexibility, the technologies underpinning clouds are primarily "back-end" and targeted at data centers – an IBM strength,
3. Since the cloud ecosystem is large, diverse, and is evolutionary (rather than disruptive) with no dominant incumbent, customers will demand solution providers to provide the flexibility that open systems typically provide. IBM's continuing investments in Linux, open-source, and KVM virtualization will serve it well,
4. Customers will continue to demand bottom-line business benefits unique to their industry and workloads. Over time customers will no longer ask "What is cloud computing?" but rather "What can the cloud do for me?" IBM's traditional vertical industry-oriented focus in its Global Business Services and Sales and Distribution organizations will become a source of differentiation as was the case with SOA,
5. And finally, and perhaps most strategic, IBM recently announced a business and predictive analytics initiative[25] with several billion dollars of investment to build integrated (hardware, software, and services) solutions. The best systems and software designs are driven by and optimized for workloads. Many of IBM's future investments in its software and systems portfolio for business analytics should also help with cloud computing workloads.

IBM has the unique opportunity to influence the current industry conversation on what cloud computing is. Unlike a conversation on data centers which is very relevant for IT-based enterprises, or a conversation around calendaring and email services which is typically very relevant for individuals and users, the conversation around cloud computing must include both constituents. IBM has historically done a fine job with enterprises, while Amazon and Google have leveraged their mind share with consumers to influence the cloud computing dialogue and build cloud mind share. IBM must seize this opportunity to also reach out to consumers and individual developers to build and capture cloud computing mind share. After all, mind share does indeed influence market share.

[25] IBM to Acquire SPSS Inc. to Provide Clients Predictive Analytics Capabilities, http://www-03.ibm.com/press/us/en/pressrelease/27936.wss