Cabot Partners Group, Inc. 100 Woodcrest Lane, Danbury CT 06810, www.cabotpartners.com

WHITE PAPER

Confronting the Data Center Crisis: A Cost - Benefit Analysis of the IBM Computing on Demand (CoD) Cloud Offering

Sponsored by IBM

Srini Chari, Ph.D., MBA

March, 2009

mailto:chari@cabotpartners.com

Executive Summary

Proven cloud or Infrastructure as a Service (IaaS) computing solutions such as IBM's Computing on Demand (CoD) are very attractive for many commercial enterprises especially in today's environment of flat or shrinking IT budgets. With a spectrum of flexible offerings and pricing models, IBM's CoD solution provides secure, affordable, elastic, and risk-free access to IT infrastructure resources for companies that need to quickly scale-up or scale-down their IT needs to adapt to their business demands.

Through IBM client interviews across several industries and other industry and workload analyses, this paper examines the benefits and costs of IBM's CoD offering for high performance business and technical computing. In evaluating the business (benefits and costs) case for IBM's CoD, the following criteria were used as applicable:

Benefits

- 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.

For a sample configuration using IBM CoD, TCO savings ranged from 30% to 69%. In addition, three specific IBM CoD clients discussed in detail in the case studies have benefited from:

- ***new capability*** to solve a previously intractable problem,
- ***faster time to results*** for end of day processing, and
- the ability to implement a ***new innovative business model.***

In all cases, the companies realized the business and operational benefits they were seeking, and the flexible pricing models backed up with outstanding service and support provided by IBM were an added bonus.

With CoD, IBM has demonstrated the successful delivery of a secure cloud service with robust technology, efficient processes, and skilled personnel to solve many high performance business and technical computing problems across a range of industries. This agile CoD cloud solution enables better and more efficient alignment of a client's business with IT especially in today's volatile business environment.

Cloud Computing for Commercial Enterprises

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. Dynamic infrastructure and cloud computing can transform companies to become more agile and develop sustainable competitive advantage. It is a key driver of growth and productivity. High performance business and technical computing solutions enable the accurate modeling of intricate phenomena and the processing of mammoth data into actionable knowledge and insights. It requires the use of large scalable systems[1] and a smart, dynamic IT infrastructure in the data center.

However, the current economic downturn and the escalating energy and people costs to build and operate datacenters, will force companies to reevaluate how they can maximize their return on IT investments. They will need smarter approaches to reduce costs and manage complexity. Large government laboratories may continue to invest the several hundred million dollars needed to build and operate large data centers, but many information intensive commercial enterprises will likely use a combination of in-house computing augmented with economical private off-site services such as secure cloud computing.

Cloud computing[2] promises to provide dynamically scalable and often virtualized IT (hardware, software, and applications) resources as a service transparently over the Internet to users who need not have knowledge of, 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, and also builds upon recent IT infrastructure solution concepts such as grid computing, utility computing, and autonomic computing. With a cloud computing service, even smaller companies such as Internet companies 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.

Cloud computing solutions are still evolving. There is very little published data on the quantitative demonstration of ROI for cloud solutions. TCO and ROI calculators[3] comparing in-house IT solutions to cloud solutions that consider several hidden costs and value process improvement benefits are just beginning to appear. While these calculators are extensive and useful for traditional applications such as ERP, we believe they do not capture many of the hidden costs involved in building data centers such as the energy and facilities costs or quantify the incredible business value delivered by cloud solutions such as IBM's CoD solution.

Challenges in Enterprise Data Centers and Solutions

According to IDC[4], 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. The issue of power and cooling has become a top priority for IT executives. The IT industry is defining additional metrics such as gigaflops/watt that rate systems. The Uptime Institute has also defined additional metrics[5] that define data center "greenness". The Green500 list[6] is becoming as important as the Top500[7] list of supercomputers as vendors compete for bragging rights. 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 compute-intensive industries.

According to recent research by McKinsey[8], 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

---

[1] The IBM System x iDataPlex, http://www-03.ibm.com/systems/x/hardware/idataplex/
[2] Wikipedia
[3] Cloud Apps Hosting Business Case and ROI Calculator, http://spreadsheets.google.com/pub?key=pNRbuDlSBmus_GfaoIcbICw&output=html
[4] Jed Scaramella, "Worldwide Server Power and Cooling Expense 2006-2010 Forecast", September 2006.
[5] John R. Stanley, Kenneth Brill, and Jonathan Koomey, "Four Metrics Define Data Center "Greenness"", White Paper, Uptime Institute.
[6] The Green 500 List, www.green500.org/Lists.html
[7] The Top 500 list at www.top500.org
[8] McKinsey on Business Technology, Innovations in IT Management, Number 14, Winter 2008.

center facility is growing by as much as 20 percent a year, far outpacing overall IT spending, which is increasing at a rate of 6 percent a year. 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 years or more to design and build and are expected to last for 12 years. Also consistent with other research, the McKinsey research found that server utilization generally tops out at 5 to 10 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, the McKinsey research indicates that this assertion may not be true.

McKinsey postulates that this mismatch could be because many companies have difficulties in accurately forecasting workload demands. The long-term (12 years) 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 25% or 100% 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.

McKinsey suggests using a centralized governance model in which the CIO is empowered by the CEO and is accountable and responsible for data center management. CIOs at large companies can reform data center operations and could double data center energy efficiency by:

- managing IT assets aggressively through virtualization,
- providing incentives to IT personnel to improve forecasting and minimize deviations from real demand,
- treating data center resources as scarce resources and ensuring that business units implement a total cost of ownership (TCO) model for new systems and applications,
- implementing new metrics for data center efficiencies that account for energy, utilization, and floor space.

These recommendations are primarily targeted at improving existing data center efficiencies, but significant business and technical challenges remain. These challenges will become more acute in the future as companies perform more complex analyses seeking competitive advantage in a dynamic and volatile business environment. This is particularly true for web and high performance business and technical computing workloads. What's needed is a comprehensive approach that includes energy-efficient and computationally dense systems, software for better systems and power management and utilization, and flexible delivery models such as cloud computing that adapt easily to computing demands.

In recent years, IT systems providers have made significant innovations in "green" and next generation data centers[9] to reduce the Total Cost of Ownership (TCO), reflecting not just capital costs but operational and maintenance costs as well. In addition to energy-efficient and computationally dense systems[10, 11], these innovations also include software solutions to increase IT asset utilization through virtualization, workload management, and consolidation. Recently, IBM has made significant announcements in cloud computing[12] and dynamic infrastructure[13] with a comprehensive roadmap of specific systems, software, services, and solutions to help address these crucial data center challenges.

---

[9] The New Enterprise Data Center, http://www-03.ibm.com/systems/nedc/index.html

[10] Srini Chari, "IBM System x iDataPlex: The Newest Economical Workhorse in the Computing Cloud for Next Generation HPC Data Centers", Cabot Partners White Paper, April, 2008, ftp://ftp.software.ibm.com/common/ssi/sa/wh/n/xsw03016usen/XSW03016USEN.PDF

[11] Srini Chari, "A Total Cost of Ownership Study (TCO) Comparing the IBM Blue Gene/P with Other Cluster Systems for High Performance Computing", Cabot Partners White Paper, November 2008, http://www-03.ibm.com/systems/resources/tcopaper_finalfinal_2008.pdf

[12] IBM Cloud Computing, http://www.ibm.com/ibm/cloud/

[13] IBM Prepares to Take on 21st Century Infrastructure, http://www.networkworld.com/news/2009/020909-ibm-dynamic-infrastructure.html

Workload Classification: Making the Case for Cloud Computing

Accurately forecasting 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 provide a framework to examine the emerging workload trends that are poised to fundamentally transform the delivery of IT solutions through cloud computing utilities.

First, we examine a wide range of workloads typical in many IT applications across several dimensions. These applications range from traditional enterprise business applications such as Enterprise Resource Planning (ERP) to more compute intensive High Performance Computing (HPC) applications and web/business analytics. These applications are classified according to the typical workload characteristics with compute-intensive/job on the x-axis and workload variability (V_W, defined in the **Appendix**) 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 most 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.

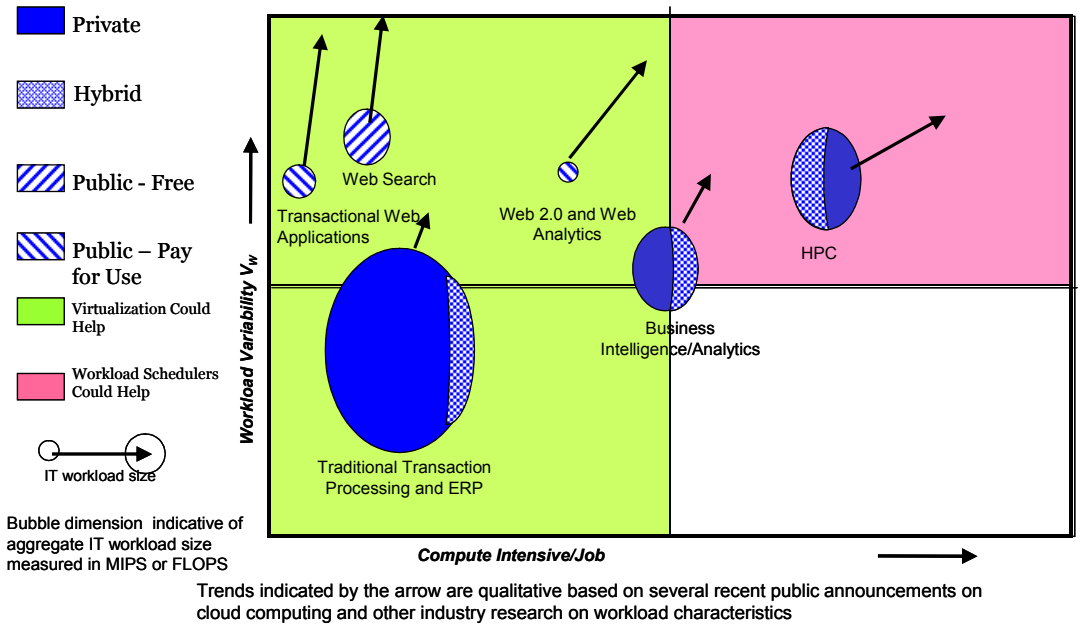

Figure 1: High Level Workload Classification by Application Domains

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). The IBM CoD solution is an excellent example of the Hybrid model. 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.

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, with HPC and other compute intensive workloads often requiring several CPUs per job, workload scheduling and management solutions from Platform, DataSynapse, 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/domain decomposition algorithms.

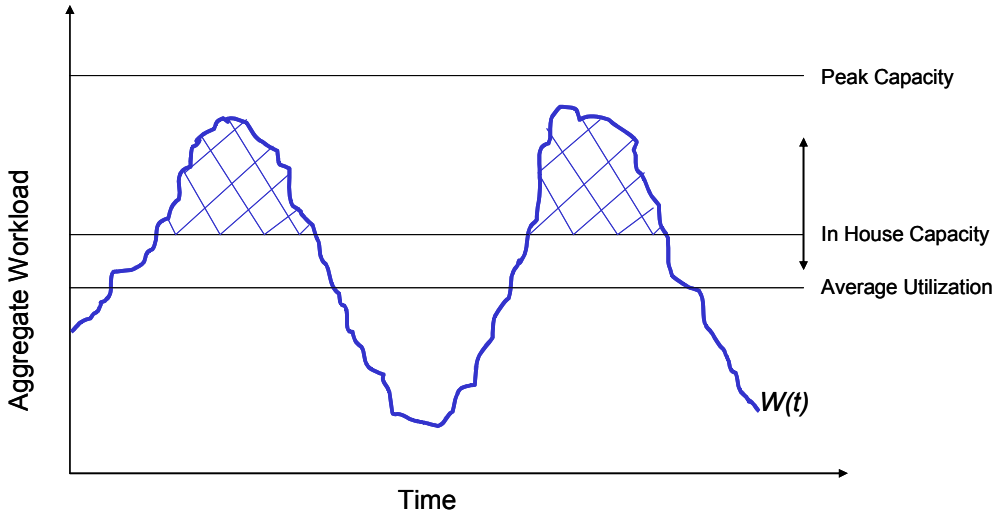

Figure 2: Typical Low Variability Workload Curve - Common in Many Enterprise Business Applications

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.[14] 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 depicted in the shaded area. This would allow them to maintain their current private in-house infrastructure and get the additional computing capacity on demand. 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 these companies would continue to add more capacity in their data centers which could result in the need to upgrade, build, buy, or lease data centers.

[14] 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.

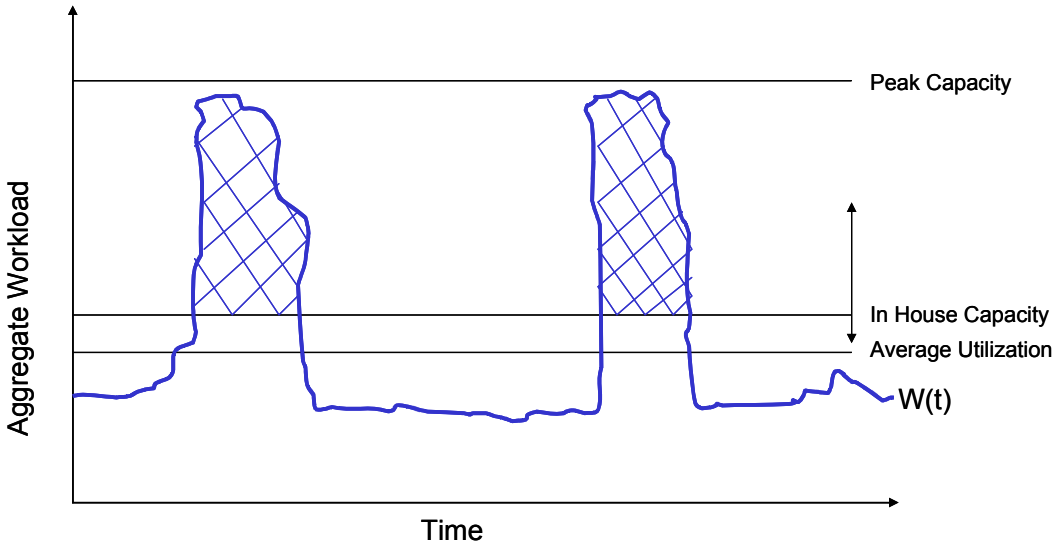

Figure 3: Pulsed High Variability Workload - Common in Departmental or Small and Medium Business HPC and Web Applications

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"[15] 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 model is the only economically viable model for these workloads and hence has attracted many early adopters with these workloads. For these businesses, the 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 economically infeasible.

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

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

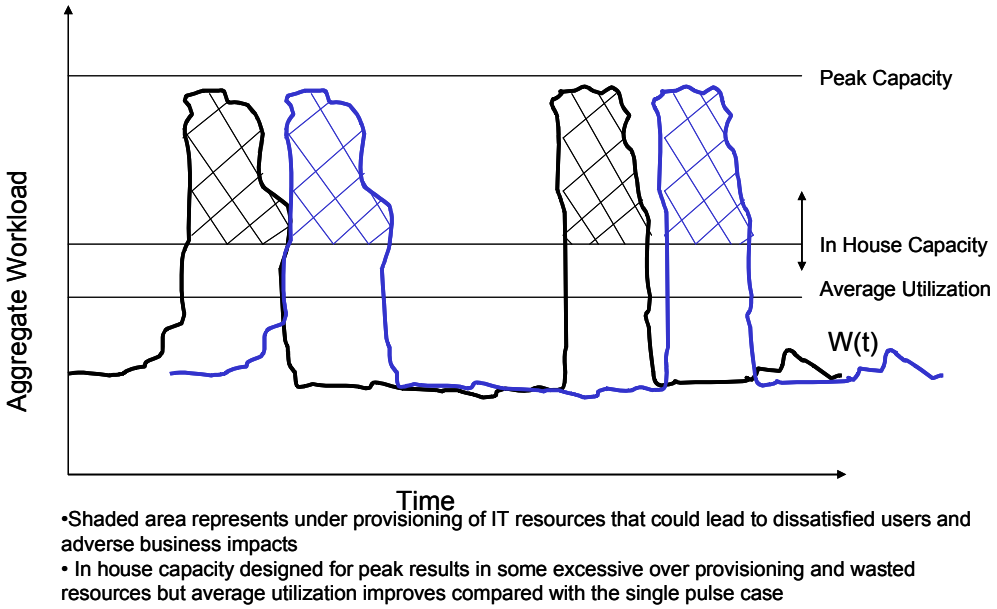

Figure 4: Many Pulsed Workload (Low to Medium Variability) - Common in Large Enterprise HPC

Mature 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 or will have to justify the need to build, buy, or lease new data centers. If new pulse workloads are substantial in intensity – say at least 3-5 times the normal pulse, they could benefit from a cloud computing service.

The IBM Computing on Demand (CoD) Solution

The IBM Computing on Demand (CoD) is an offering that provides companies with flexible access to vast computing power capable of handling workloads of all sizes. CoD users have access to security-rich supercomputing environments that can be used like on-site hardware, but without the capital commitment, management, and maintenance costs. When computing demands exceed in-house capacity clients can easily shift the excess workload to an IBM CoD cloud center and purchase the additional processing capacity necessary to help meet demand. The hardware is hosted, maintained, and supported by IBM to deliver cost-effective capacity that helps free-up companies to focus on business operations.

With access to CoD's on-tap supercomputing capability, a company could maintain business critical processor-intensive tasks in-house while delegating urgent computations to reserved CoD capacity. Instead of incurring large capital investments to buffer capacity from spikes in demand, CoD users can treat additional capacity as a value-driven operational expense. If users only have to pay for capacity when they need it, the existing capital may be recommitted to strategic business objectives.

IBM Computing on Demand provides customers a spectrum of delivery and pricing options metered by CPU Hour and storage per week. Supported CoD models include dynamic (hourly), variable (weekly), and dedicated (yearly) capacity options. IBM currently operates seven global Computing on Demand cloud centers throughout the world. Combined, these centers offer over 13,000 cluster processors, 54 Terabytes of storage, diverse interconnects and operating systems. With very high levels of utilization, IBM clients have benefited from this efficient, cost effective way to purchase compute power. According to Ms. Christina Cunningham, Global Executive, Worldwide IBM Computing on Demand, "We have consistently observed that our CoD clients have very high server utilization rates. Typically the systems are in use over 75% of the time for a given client engagement".

A Cost – Benefit Framework for the IBM CoD Solution

The escalating costs of building and operating data centers are not due to IT capital costs. They are primarily because of increasing energy, facilities, and other operational costs. Evaluating systems solely on IT acquisition costs and price/performance is seriously flawed. Moore's law continues to persist at the processor level and every 18 months or so, the computational performance delivered by new generation of systems continues to more than double at roughly the same price point. IT acquisition costs as a fraction of the Total Cost of Ownership (TCO) are expected to decline in the future. The TCO over several years, say 3, must be assessed in order to make objective cost decisions while evaluating various solution options. But the TCO alone is inadequate. What's needed is a framework that examines the total costs incurred and the benefits delivered by IT solutions.

First, we examine the many pulsed workload case for a large enterprise HPC environment depicted in Figure 4. To illustrate our analysis, we assume the ideal situation, where the workload pulses of equal intensity are perfectly sequenced with no overlap and the system is fully utilized and the workload is executed perfectly. In this ideal situation, the workload graph is flat over time as depicted in Figure 5.

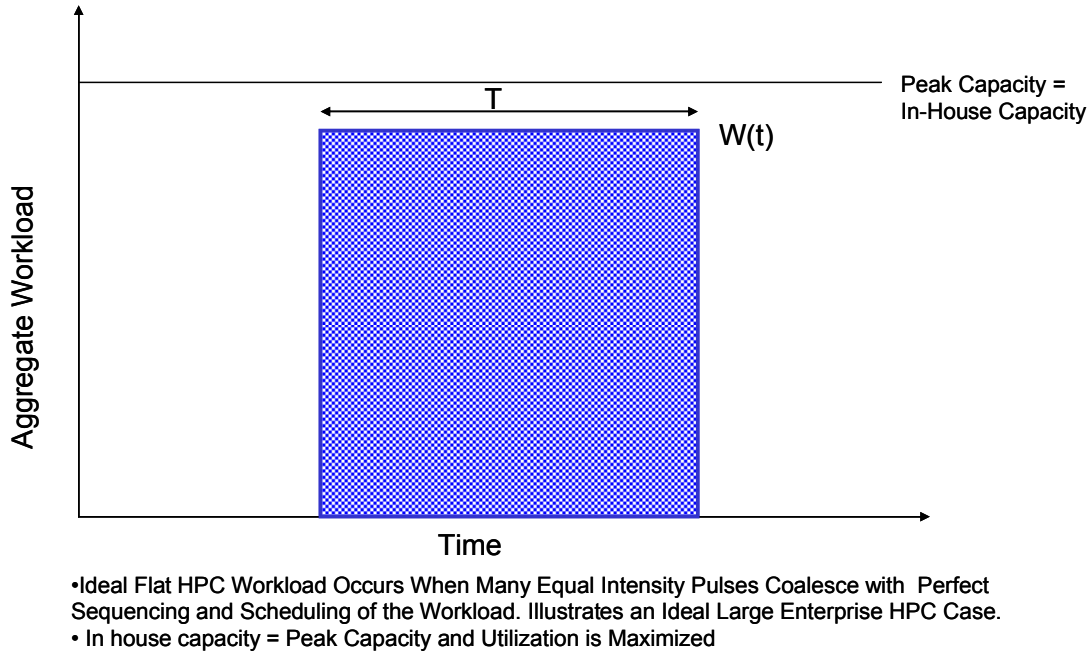

Figure 5: Ideal Flat Many Pulsed Large Enterprise HPC Workload

In this instance, a customer could build in-house capacity or lease a dedicated CoD solution from IBM. We analyze the TCO of an in-house build using the Uptime Institute framework[16] and customized by Cabot Partners[11] versus a leased CoD solution. We used a dual core x86 cluster system with 840 CPUs (cores) and 1TB of storage. We further assumed that the dedicated CoD pricing is $.25 per CPU hour and $.25/GB for storage per week. For the in-house alternative, we assumed minimal personnel for IT, site, maintenance, and security for three shifts. Figure 6 shows the results and the TCO advantages of the IBM CoD solution.

Middleware and application software costs are not considered in both cases. Moreover, the Uptime framework does not consider additional costs that will be incurred in recruiting and training staff, deployment and customization costs, and potential costs due to failures and unplanned outages. There is also a critical lead time cost not considered in the in-house alternative. As mentioned earlier, it may take many months (even years to build large data centers) to deploy an in-house solution whereas the IBM dedicated CoD solution can be deployed almost immediately for small configurations and weeks to a few months for

[16] Jonathan Koomey, "A Simple Model for Determining True Total Cost of Ownership for Data Centers", White Paper, The Uptime Institute, 2007.

much larger deployments. This "value of immediacy" of the IBM CoD solution can be immense for many customers. Finally, the costs of being unable to predict demand for compute capacity could be substantial. If in-house capacity turns out to be excessive then utilization will be very low. If in-house capacity becomes inadequate then expansion costs could be large. The "flexibility value" of the IBM CoD solution can be huge and significantly alleviates the customer's burden of having to accurately predict demand for computing capacity as McKinsey recommends for in-house data center efficiencies.

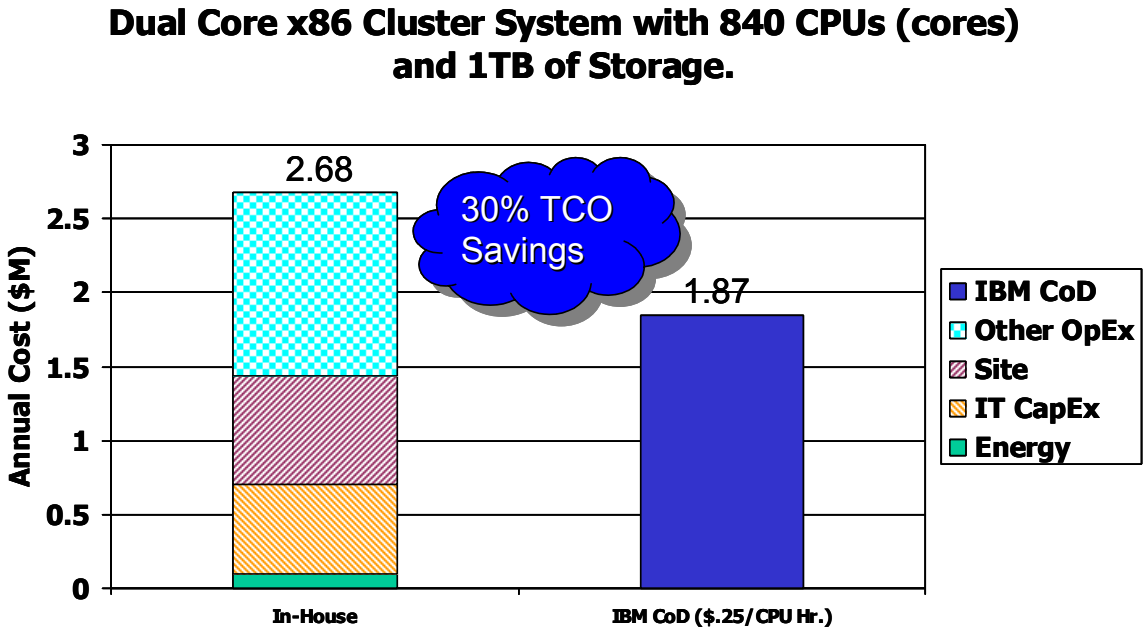

Figure 6: TCO Analysis of IBM CoD Dedicated vs. In-House

Next we examine, the single pulse workload case depicted in Figure 3 typical in departmental or small and medium business analytics or web workloads. To illustrate our analysis, we assume the ideal situation, where the workload pulse is periodic over T and flat for a short duration T_P as shown in Figure 7.

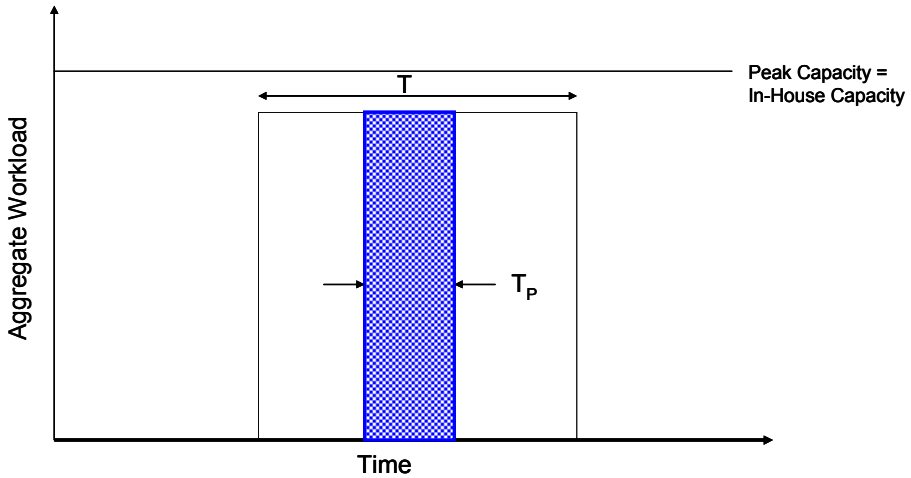

Figure 7: Ideal Single Short Pulse Workload

In these instances, a cloud computing solution such as the IBM CoD variable or dynamic solution will be the only economical alternative. In-house alternatives with significant cost barriers will be financially infeasible. Nevertheless, using the same cluster system as the base case, we compare the TCO between an in-house build and the IBM CoD variable and dynamic solution. We assume that variable solution pricing to be $.60/CPU Hour with a one week commitment and used 12 times a year. The dynamic pricing is assumed to be $.80/CPU Hour with a 10 hour commitment for 5 days and used 24 times in a year. Figure 8 shows the results and the significant TCO advantages of the IBM CoD solution.

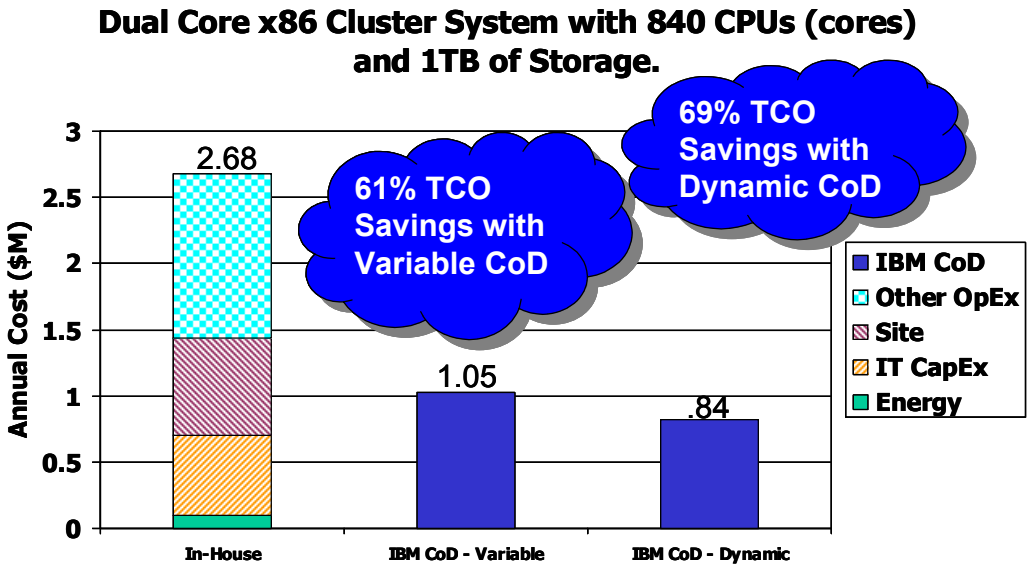

Figure 8: TCO Comparison of IBM CoD Variable or Dynamic vs. In-House

While the TCO advantages of the IBM CoD cloud solution are substantial, companies must also assess the business benefits of such innovative cloud computing solutions. We use several customer case studies to illustrate the business benefits and value of the IBM CoD solution in addition to the TCO advantages.

Customer Case Studies Highlighting Business Benefits

The cost – benefit analysis presented here is based on feedback obtained from interviewing three current CoD customers who represent a range of industries and application uses. We discuss both the value and cost benefits realized by these customers. In all cases, the companies realized the business and operational benefits they were seeking, and the flexible pricing models backed up with outstanding service and support provided by IBM were an added bonus.

Case Study #1: Top International Wealth Management Savings Company – New Capability

We interviewed the Vice President of the US subsidiary of Corporate Actuary. This company provides life insurance products, annuities, and other financial products.

Applications Used: Currently this company uses actuarial software which helps insurers and financial services companies gain competitive advantage and satisfy forthcoming regulatory requirements. An actuary uses statistics and historical data in an attempt to measure the risk of a particular investment. This application is used for a variety of tasks, including product development, pricing, valuation, cash flow testing, liability and asset/liability modeling. These applications are fully transparent, allowing insurers to see the actual code used to make the calculations and, in turn, to trace the calculations and ease audit ability. This platform also supports distributed processing and grid computing on Microsoft Windows to provide scalability for very large projection tasks.

Challenges: There is a significant change in the regulations for financial reporting for insurance firms that will go into effect after December 31, 2009. These European regulatory requirements – the Market-

Consistent Embedded Value (MCEV) Principles include clear guidance on calculating and require insurance companies to provide more detailed financial information than ever before.

This subsidiary had to increase by at least a factor of five the range and number of statistical analyses performed on a routine basis. This required an immediate solution in order to be ready to comply with new regulations. The existing compute server grid was inadequate to handle this immediate need for additional capacity. Also, physical space to host an increased number of servers was an issue. Even if these additional statistical analyses could be done in-house by extending the existing grid, additional expense and inconvenience would be incurred to switch back-and-forth between the existing IT environment and the one needed for the new regulations.

Alternatives Investigated: Several alternatives were investigated to address these challenges in addition to the IBM CoD solution. These included:

1. Upgrading/adding existing servers in their in-house compute grid. This alternative entailed expensive capital acquisition and the timing was infeasible. Physical space was also a limiting factor.
2. Using the services and infrastructure provided by the ISV. The ISV's HPC infrastructure was not robust and scalable to address the extreme computing needs for the added statistical analyses.
3. A "container" type solution was also evaluated. This was infeasible. There was no available land in the vicinity to deploy a container solution.

Why IBM CoD was Chosen: The IBM CoD solution provided flexible and variable access to over 100 IBM System x servers for 3 weeks in a month. The pricing and terms were attractive and VPN access options made this environment secure, stable, and isolated from the existing compute grid. Furthermore, additional savings of about 1PY were possible as IBM provided the support and service necessary to maintain this compute capacity. IBM was able to collaboratively solve minor migration issues with application deployment on an IBM Windows HPC cluster at the CoD center.

The Ongoing IBM CoD Value: The IBM CoD solution with variable pricing terms has been consistently used over the last 6 months with over 75% utilization. This has made feasible a previously "intractable" problem as alternative approaches did not satisfy the immediate business need of compliance with upcoming regulations. The company implemented an IBM CoD solution in one subsidiary in the United States. They're now able to get 5-7 times more work done in the same timeframe or the same amount of work done 5 times faster. It is expected that this solution would also be considered in the corporate parent organization in Europe and at other subsidiaries worldwide. With the current global economic crisis and the need for more disclosures in financial statements, even more statistical analyses will be needed, further driving up the compute requirements. The access to almost "infinite" resources and the pricing elasticity offered by IBM CoD make this solution well poised for future growth needs.

Case Study #2: Major New York Based Financial Conglomerate - Faster Time to Results

We interviewed the Director Solutions Delivery, Corporate IT responsible for providing internal IT services and support for several business line groups including actuary analysts, legal, compliance, corporate finance, and other HQ functions. This company provides life insurance products, annuities, and other financial products.

Applications Used: Currently this firm uses the MG-ALFA application from Milliman, Inc; a global software and services firm providing a range of application solutions for the insurance and financial services industry. MG-ALFA uses DataSynapse grid software for scaling up actuarial analyses. In-house actuary analysts build stochastic analysis models on their Windows desktops using MG-ALFA and other tools and execute these models on a grid/cluster infrastructure using DataSynapse as the parallel scheduler and enabler.

Challenges: The internal IT group supports a wide range of internal clients and has to adhere to very stringent service level expectations for risk analysis. End-users demand simple, reliable, timely, secure, and "transparent" IT services. End-of-day analyses must be turned around in 10 hours or less. The current financial crisis, increased regulations, and market volatility have spiked the demand for more complex

stochastic risk analyses. Future workload is expected to grow even more with newer market regulations for reporting and risk assumption.

Alternatives Investigated: The firm investigated adding more servers in the compute grid. This IT capital expansion was expensive and would be grossly under-utilized when time critical risk analyses workload was not running.

Why IBM CoD was Chosen: The IBM CoD solution provided a flexible and dynamic access to between 200 and 350 CPUs. There have been 4-5 times in the last year that an urgent increase in capacity of up to 500 CPUs was needed. IBM was able to respond in a matter of hours to these business critical requests. In every case, IBM exceeded this firm's expectations.

The Ongoing IBM CoD Value: The IBM CoD solution has been consistently used for time critical analyses over the last 2-3 years often with 75%-100% utilization. The firm expects increased scale and capacity use of CoD over the near future. The Director of Solutions Delivery is very pleased with IBM's exceptional service and support, rapid response to urgent requests, and deep expertise in HPC and grid computing. IBM even had a grid computing expert with in-depth knowledge of the financial services industry work with this firm's teams to help define the solution strategy and architecture. IBM's deep relationships and expertise with software partners – Milliman and DataSynapse- were also invaluable. The Director indicated that the IBM CoD solution was invaluable.

Case Study #3: Ingrain Rocks-Solution Provider for Petroleum E&P - New Business Model

We interviewed the Mr. Barry Stewart, CFO, Ingrain Rocks (www.ingrainrocks.com). Ingrain Rocks is a small startup services company that provides solutions for the Petroleum Exploration & Production (E&P) companies globally for fast accurate rock property analysis. These solutions lead the industry in measuring shales, carbonates, tight gas sands and oil sands.

Applications Used: Ingrain delivers accurate and fast results and claims to be way ahead of competition by routinely delivering results in about 2 months instead of the typical 9 months. Using proprietary high resolution imaging algorithms, Ingrain computes physical properties of reservoir rocks such as porosity, permeability, and other electric and elastic properties. Ingrain's processes work equally well with samples taken from core plugs, oil sands samples or drill cuttings. Ingrain also computes multiphase flow at the pore scale in an accurate digital representation of the pore space. Ingrain's parallel applications scale well into the 100s of processing nodes.

Challenges: As a small company, in 12-18 months, Ingrain desires to reduce end-to-end cycle time from months to days or less. A key computational analysis step is expected to become more critical as client-delivery time shrinks. Ingrain's business model is to avoid investing in IT in-house. They need global access to secure, flexible, and "infinite" computing resources and must ensure that IT costs track client workload.

Alternatives Investigated: Ingrain has cluster access through a local service to develop their applications. However, this model is inadequate for anticipated future growth needs. Ingrain examined several options including other service providers but none could meet their needs of access to a range of equipment, configurations, capacity, and price points. Flexibility in machine configurations, allows Ingrain to optimize their application workload and tune the underlying algorithms. To maintain customer intimacy and foster growth, Ingrain plans to open multiple locations globally e.g. the middle-east, and Latin America. Their computing service partner needs to have similar global capability in terms of service and support and a flexible pricing model that allows Ingrain to suitably price their services to clients. Finally, Ingrain needs a stable and trusted partner for the long haul.

Why IBM CoD was Chosen: IBM was the only provider that was global and satisfied all of Ingrain's needs by providing access to a range of systems, flexible configurations, almost "infinite" capacity, and attractive price points. Mr. Stewart says *"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".

The Ongoing IBM CoD Value: Ingrain plans to continue to use this CoD solution for all their future client engagements and expect their use of the CoD solution to grow in the future. This will allow them to tackle more challenging projects that even some of the largest E&P companies struggle with. 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 or hours instead of a couple of months.

Conclusions

High performance business and technical computing helps companies achieve the speed, agility, and insights to lead the market and together with new web workloads will drive data centers to add more capacity. But the escalating energy and operational costs of building and maintaining data centers will compel companies to adopt secure cloud computing models. The cost-benefit analysis in this paper, demonstrates the advantages of the IBM Computing on Demand (CoD) cloud solution. While the TCO savings with the CoD solution are substantial ranging from 30% for dedicated CoD to 69% for dynamic CoD, the three IBM customer case studies exemplify the business benefits of new capability, faster time to results, and new innovative business models – an IBM CoD triple play.

Appendix – Workload Variability

For simplicity and clarity, we define the following nomenclature for a homogeneous computational cluster with N CPUs each with an individual peak capacity of P_{CPU}. In the case of a multi-core node, each CPU is a core. So if the cluster is a dual-core cluster, it will have $N/2$ nodes or sockets with total peak capacity of $P = P_{CPU} * N$. Furthermore, $w_i(t)$ is the individual workload for job i at time t measured as a fraction of the individual CPU peak capacity in flops or mips. $W(t)$ is aggregate workload function in the cluster at t; so $W(t) = \Sigma w_i(t),\ i = 1, J$, where J is the number of jobs at time t. T is the total time of interest in hours.

So, $\int_0^T W(t)\,dt = A_W$; where A_W is the area under the aggregate workload curve measured in CPU Hours. And $U_W = A_W / (P*T)$ is the average utilization level for the workload over the total time of interest T. Also, $\int_0^T (W(t) - U_W)^2\,dt / T = \sigma_W$ is the workload variance. We define a new workload variability metric $V_W = \sigma_W / U_W$ that we use extensively in our non-prescriptive analysis to examine the tradeoff decisions between expanding in-house computing capacity versus getting additional computing capacity through a cloud computing utility.

More Information

To learn more about the IBM Computing on Demand (CoD) solution, contact your IBM representative or visit **http://ibm.com/deepcomputing/cod**.

To test drive IBM Computing on Demand, please visit **http://ibm.com/systems/deepcomputing/cod/testdrive.html**.

To learn more about IBM Cloud Computing Solutions, please visit **http://ibm.com/cloud**.

To learn more about IBM solutions for High Performance Computing, please visit **http://ibm.com/deepcomputing**.