Smart IBM Solutions for High Performance Engineering Clouds: Advanced Performance for Complex Engineering Problems Delivered in a Flexible and Secure Mode through Private and Private-Hosted Clouds.

Sponsored by IBM

Srini Chari, Ph.D., MBA

June, 2011

mailto:chari@cabotpartners.com

Why Engineering Needs High Performance Computing (HPC) and Cloud Solutions

Product and design chain complexity are pushing the limits of engineering processes and computing, leading to increased demand on computing resources and rapidly escalating costs. Further, a wide array of sophisticated skills and capabilities are required to become proficient in developing and delivering innovative products in the manufacturing industry. Today, the design and engineering function in companies is in a state of flux.

The engineering community must deliver innovative designs better, faster and cheaper; design high quality products with fewer designers or resources and across a distributed ecosystem of partners; and respond to increased CIO cost control of engineering IT.

And they must do all of this in an operational reality of siloed data centers tied to projects and locations; limited or poor operational insight; underutilized resources still not able to respond to peak demands; and designers tied to local desk side workstations and with limited collaboration.

The way to overcome these issues is to transform siloed environments into shared engineering clouds – starting from a private and private-hosted environment and maturing to public over time. To achieve this, engineering functions require interactive and batch remote access; shared and centralized engineering IT; and an integrated business and technical environment. This unlocks designer skills from a location; provides greater access to compute and storage resources; aligns resource to project priorities; and consequently companies realize improved operational efficiencies and competitive cost savings.

IBM offers a rich portfolio of HPC cloud computing solutions with substantial business benefits: This paper describes IBM's rich portfolio of HPC cloud computing offerings for engineering and other technical workloads. IBM has been the leader in clustered high performance computer systems for many years and consistently dominates the TOP500 and Green500 (www.green500.org) list of the world's most powerful and environmentally responsible supercomputers. 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 (www.ibm.com/cloud), strategy, and a detailed roadmap of specific HPC cloud solutions deployed at IBM customer sites**.** IBM HPC cloud computing offerings deliver the performance and value of supercomputing to users in an affordable, efficient and secure manner on private clouds. This paper describes in greater detail deployments of HPC cloud offerings from IBM with the following bottom-line benefits:

- ***Enhanced competitive differentiation*** for Exa, a major fluids and MCAE application software provider
- ***Built an ultra-efficient private HPC cloud called Magellan*** at NERSC and Argonne National Laboratory for the Department of Energy (DOE)
- ***24x7 access to general and HPC cloud computing for students, researchers, and faculty*** at North Carolina State University, and
- ***Reduced the IT cost per designer by half and cut first pass design time 6 months*** for the design and development of the latest IBM POWER7 processors.

HPC is Crucial for Electronics, Mechanical Engineering, and Digital Manufacturing

High Performance Computing (HPC) solutions are crucial in the semiconductor industry. A wide range of Electronic Design Automation (EDA) solutions are used to collaboratively design, test, validate, and manufacture rapidly shrinking nanometer integrated chips leveraging advanced research, process technologies, and global research and development teams. Today, Static Timing Analysis (STA) for circuit simulation and Computational Lithography for process modeling are key HPC applications. The ultimate goal for many engineering R&D enterprises is to virtualize the full semiconductor development process. Doing so will reduce cost by requiring fewer silicon experiments and

improving time to market for next generation semiconductor technologies. It will also drive the need for more HPC resources in support of near first principle predictive models for circuit design and process modeling which builds upon the current work in timing analysis and computational lithography. Likewise, many companies are discovering that Mechanical Computer Aided Engineering (MCAE) and Digital Manufacturing can slash production time, optimize designs, and prevent expensive rework. As a result, MCAE and Digital Manufacturing have grown in importance earlier in the product development cycle, and no longer consigned to the final stages of the design and manufacturing process. Smaller component suppliers with limited skills and resources are increasingly using many MCAE and Digital Manufacturing applications for more sophisticated analysis such as crash analysis, non-linear metal forming, and computational fluid dynamics in addition to traditional structural analyses. This has increased collaborative and iterative product development and design optimization throughout the manufacturing supply chain resulting in dramatic reductions in the time to market. In order to be competitive and responsive, small and medium suppliers are increasingly using MCAE and Digital Manufacturing solutions that need HPC environments. The availability of flexible and cost-effective HPC cloud computing solutions compatible with desktop Computer Aided Design (CAD) environments has helped to further energize MCAE adoption among these suppliers. This is also true in EDA environments.

Why High Performance Computing is turning to Clouds

The current economic downturn and the escalating energy and people costs for information technology are forcing engineering companies to reevaluate how they can maximize their returns on 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 cost of IT investments.

As high performance computing (HPC) mainstreams, its business innovation impact expands: High performance computing (HPC) uses supercomputers and clustered computers ranging from a 16 node cluster to the teraflops or petaflops performance range to solve computational and data intensive problems. HPC helps enterprises achieve the speed, agility, insights, and sustained competitive advantage to deliver innovative products, increase revenues, and improve operational performance. Scientists, engineers, and analysts in leading-edge enterprises rely on HPC to solve challenging problems in fields ranging from engineering, manufacturing, finance, risk analysis, revenue management, to life and earth sciences to name a few. This mainstreaming of HPC has put considerable pressure on solution providers to provide scalable, reliable, and secure solutions while reining in costs and complexity.

Cloud computing effectively reduces costs and complexity…but enterprise class security and reliability remain concerns: Cloud computing -- in which large amounts of data and computing resources can be accessed remotely over the Internet using a personal computer, cell phone or other device -- holds great promise in the IT market. The cloud model has the potential to cut the costs, complexity and headaches of technology deployment for companies, universities, and government agencies.

The potential benefits to enterprise clients are immense. Cloud computing turns the economics of enterprise IT on its head. Delivery of IT services (including infrastructure, platform, and applications) from the cloud has both capital expense and operation expense advantages. The ability to pool resources, virtualize them, and then dynamically provision from the resource pool yields a much higher utilization rate and thus better economics -- sometimes improving system utilization from 15 percent to 90 percent!

Already, new-generation IT companies such as Amazon.com, Google and Salesforce.com, among others, offer cloud-based Web services including e-mail, computer storage, and customer management software. But many enterprises and government agencies have been wary of cloud computing because of traditional IT concerns like data security, reliability of service, and regulatory compliance. Those concerns are now being fully addressed as trusted computer solution providers like IBM apply their fundamental strengths in deploying enterprise computing level security and bringing reliability to cloud computing.

Cloud computing will dramatically impact the way IT services are consumed and delivered in the future. Surveys suggest that cloud computing will be a $199 billion opportunity by 2015 and 40% of IT solutions will be delivered over the cloud in the next few years[1]. We believe that:

- The market opportunity is large with growth rates much faster than the overall IT industry,

[1] IBM internal assessment and Gartner 2010 CIO Survey

- Linux and open source 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.

IBM Delivers a Broad Portfolio of Workload Optimized Cloud Computing Offerings

In 2008, IBM announced a broad vision and strategy for cloud computing with an initial set of offerings spanning its major businesses – software, systems, and services. IBM then extended these initial offerings with a broader set of offerings with increased capability optimized for specific high-opportunity workloads. Branded today as the IBM SmartCloud (www.ibm.com/smartcloud), 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 secure IBM SmartCloud. In 2010, IBM successfully completed ***2000*** cloud engagements, ***50%*** of Fortune 10 and Fortune 50 are working with IBM on private clouds, and ***80%*** of Fortune 500 companies use IBM cloud capabilities. Today, IBM offers solutions across the entire cloud stack: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software/Business Process as a Service (SaaS/BPaaS) as shown in the following figure.

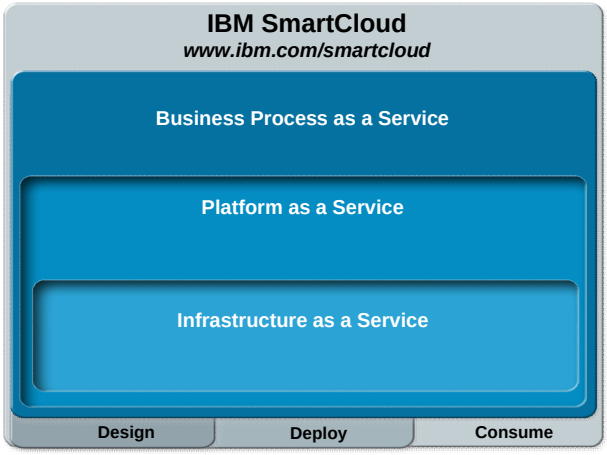


Figure 1: IBM's Broad Portfolio of Cloud Computing Offerings

According to Jeff Vance of Datamation, *"**IBM has one of the most comprehensive cloud portfolios**, with the cloud integrated throughout its many lines of business. Moreover, IBM's consulting arm has put them in touch with numerous early adopters and special use cases – all of which helps the company stay ahead of competitors."*

IBM has leveraged successes and capabilities with the SmartCloud and HPC to form the foundation of the IBM Engineering Solutions for Cloud that includes industry optimized HPC clouds for EDA and CAE workloads, in addition to IBM Rational Solution for Systems and Software Engineering, Global Collaboration Hub process integration from IBM and more. These IBM Engineering Solutions for the Cloud can leverage the IBM SmartCloud and/or the new HPC cloud offerings from IBM based on proven IBM HPC solutions as described next in greater detail.

IBM High Performance Computing Cloud Offerings

HPC cloud offerings from IBM are optimized to address the needs for speed and flexibility. These offerings are based on proven technology, leveraging IBM's leadership in scientific and technical computing over the past two decades and serve a diverse set of users across multiple industries including manufacturing, financial services, life sciences, energy, higher education, and government research. At the core of IBM's HPC cloud offerings is a light-weight, high performance job scheduler and resource manager that can schedule and provision processing nodes -often in less than four minutes, support both physical and virtual machine images, and provide users with a user-friendly internet based HPC portal in addition to standard job-queue interfaces. A highly skilled IBM Services team is available to help

clients install, configure, and tune their HPC or engineering cloud environments. A fast "Quick Start" implementation service from IBM will be introduced in 3Q11 to reduce time to deployment.

IBM Private HPC Cloud: When a cloud environment is created inside an enterprise firewall, it can provide users with the same rapid access to IT as the public model, but with less exposure to Internet security. This can often make private clouds more appropriate for high value programs and systems unique to organizations with sensitive data that must be protected or mitigate challenges related to transferring large file sets or data across the internet. It also allows clients to scale within their own enterprise, safely behind their firewall, when departmental, divisional, and/or geographically-based systems are pooled into a large virtual HPC resource pool.

For many years, IBM had deployed its own private HPC grid for developing future processors. Later, with the objective to streamline the development process and reduce costs, IBM determined it was critical to move to a centralized datacenter private cloud model where the data / storage devices could be housed in a centralized location, along side the compute resources. Further, a highly optimized 2D accelerated portal was added to enable developers throughout the world to access the cloud, work efficiently and in a collaborative manner. Five years later, the benefits to implementing a private cloud model are enormous. IBM reduced the IT cost per designer by 50% and took 6 months out of their Power7 development schedule.

IBM has now packaged and hardened the components of this internal HPC cloud as a new offering – the IBM HPC Management Suite for Cloud. This suite helps deploy HPC applications in the cloud, dynamically provision bare metal HPC clusters, and tailor virtual machines by the number of cores, disk space, or the amount of real or virtual memory. Acting as a policy-based resource manager, this suite also consolidates underutilized clusters or other scattered resources onto a larger infrastructure. A self-service web portal enhances user productivity for administrators and end-users alike.

In addition to the HPC Management Suite for the Cloud, an IBM HPC private cloud typically includes an IBM Intelligent Cluster or an IBM Power Systems Cluster coupled with a choice of additional components from the IBM HPC Cloud Software stack:

Base components of the HPC Management Suite for Cloud include:

- **xCAT (Extreme Cloud Administration Toolkit – www.xcat.sourceforge.net).** An open source scalable distributed computing management and provisioning tool that provides a unified interface for hardware control, discovery, and operating system diskful/diskfree deployment. This robust toolkit can be used for the deployment and administration of AIX or Linux clusters and Microsoft Windows. Its features are based on user requirements, and take advantage of IBM System x and Power Systems hardware.
- **Tivoli Workload Scheduler LoadLeveler – www.ibm.com/systems/software/loadleveler.** Used for dynamic workload scheduling, Tivoli Workload Scheduler LoadLeveler is a distributed network-wide job management facility designed to dynamically schedule work to maximize resource utilization and minimize job completion time. Jobs are scheduled based on job priority, job requirements, resource availability and user-defined rules to match processing needs with resources. LoadLeveler provides consolidated accounting and reporting and supports IBM servers including IBM Power Systems and IBM System x environments.
- **HPC Cloud Portal –** This self-service portal enables machine requests, usage reports, workload submission, monitoring, and more. Administrators can use this web portal to setup policies such as a power state management policy with the ability to override manually or restrict a specific job type to a certain set of nodes. A command line interface is also supported.

Value added components of the IBM HPC cloud offerings include:

- **GPFS (General Parallel File System – www.ibm.com/systems/software/gpfs).** A high-performance cluster file management infrastructure for AIX, Linux, Windows and mixed clusters (x86 and Power) that provides performance, scalability and availability for file data. It is designed to optimize the use of storage, to support scale-out applications and to provide a high availability platform for data intensive applications. GPFS provides online storage management, scalable access and tightly integrated information lifecycle tools capable of managing petabytes of data and billions of files. GPFS can help clients move beyond simply adding storage to optimizing data management.

- **SONAS (Scale Out Network Attached Storage –** www.ibm.com/systems/storage/network/sonas). Based on GPFS, SONAS is an easy-to-install, turnkey, modular, scale out Network Attached Storage (NAS) solution that provides the performance, clustered scalability, high availability (HA) and functionality that are essential to meeting strategic petabyte age and cloud storage requirements.
- **Other HPC storage –** This includes the recently launched IBM System Storage DCS3700 which is designed to help organizations accommodate large and rapidly growing data volumes while conserving data center resources. This dense solution enables organizations to pack up to sixty 2 TB drives in a single 4U rack space. Implementing GPFS with the DCS3700 provides an excellent combination of performance, cost-effectiveness and efficient management for big data. Further, as disk-based storage fills, GPFS works with IBM Tivoli Storage Manager software to back data up to cost-effective tape, such as the IBM System Storage TS3500 Tape Library, so organizations can retain the information they need—for as long as they need it.
- **HPC Open Software Stack** –IBM has enhanced the Open Software stack to include several distinct software tools that have been tested and integrated by IBM. These include OpenCL Common Runtime for Linux on x86 Architecture, xCAT, Advance Toolchain for Power Systems; install scripts, a resource-management tool and a cluster-administration toolkit. The Open Software Stack makes it easy for universities and academic researchers and others– who need an open source community development environment with full access to source code.

IBM Intelligent Cluster for HPC Clouds: IBM Intelligent Clusters, based on IBM System x x86 servers and related storage and file management resources, offer leading edge technology, flexibility, and high performance at an attractive price point, energy efficiency, and ease of deployment for clients looking to deploy HPC or technical computing cloud solutions.

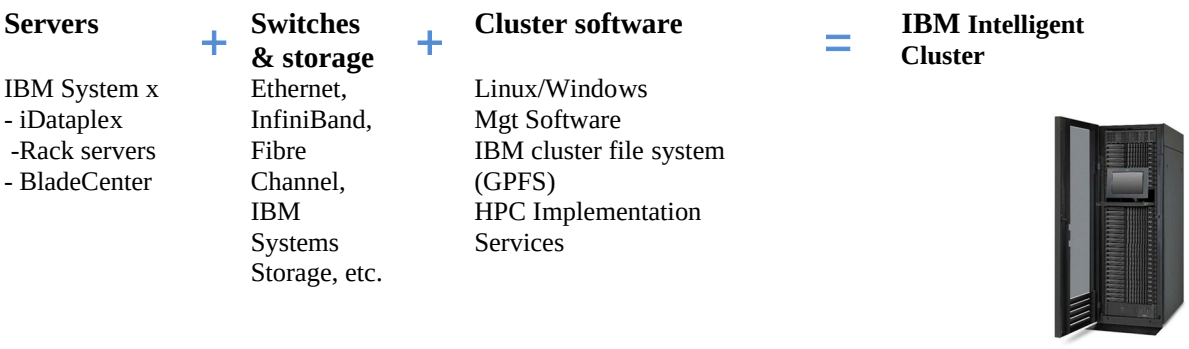


Figure 2: The IBM Intelligent Cluster for HPC Clouds

IBM HPC and Cloud Service Offerings: IBM offers services to simplify and expedite the planning and implementation of a cloud environment. The sample of offered cloud services include:

- **HPC Cloud Implementation Services from IBM.** This is a service to quickly implement an HPC cloud based on the IBM HPC Management Suite for Cloud**.** This new quick-start service helps install, configure and optimize a private or private-hosted cloud with system administrator training to help ensure ongoing success.
- **IBM Deep Computing Services.** The IBM Deep Computing Services offering is a market driven response to customers seeking highly specialized skills and expertise from the IBM research and development teams. These services can be provided as an element of the total opportunity, or the service offering can be used to facilitate a more formal collaboration. In either case, the appropriate skills will be identified and made available to meet the customer's need to design, install and optimize a private or private-hosted cloud.
- **IBM University Delivery Services – for Virtual Computing Lab support.** Cloud computing for education and healthcare programs based on virtual computing lab (open source). Includes installation, configuration, instructions, and support.

- **IBM Cloud Infrastructure Workshop.** Feasibility assessment for implementing advanced IBM technologies for a private cloud with estimates for the expected cost and operational benefits.

- **IBM Cloud Security.** Cloud consulting services from IBM to help assess a client's cloud environment and formulate a plan to improve its security posture. Specific Application Security Services for the Cloud also help determine the right balance of internal control and service provider autonomy.

- **IBM Managed Services.** IBM provides private-managed or private-hosted cloud services operated by IBM. These clouds can be hosted on dedicated infrastructure at a client's premises (private-managed) or on an IBM (private-hosted) datacenter accessible through a VPN. IBM can help clients deploy secure cloud environments, access critical applications and capabilities and help ensure the resilience and security of their systems and applications more easily, effectively and efficiently than ever before.

IBM Engineering Solutions for Cloud Combine HPC and Clouds for EDA and Mechanical

The IBM Engineering Solutions for Cloud is designed to achieve the above by delivering the deep computing power of HPC over the cloud and the benefits are substantial.

IBM Engineering Solutions for Cloud: Electronics

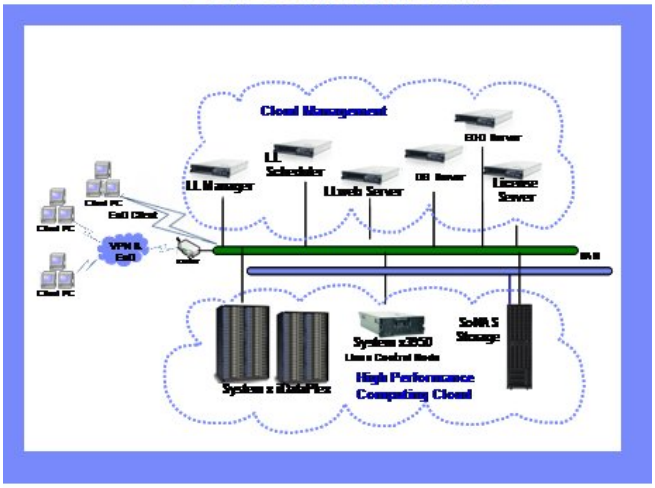
Solution Definition

- Web portal access to centralized engineering Desktops and EDA-optimized private or private hosted HPC Cloud

Key Capabilities

- Accelerated 2D remote graphics
- High Performance Computing Infrastructure
- Agile Systems and Workload management
- ISV Partner Integration

EDA Cloud Architecture



Private or Private Hosted Cloud

IBM Engineering Solutions for Cloud: MCAE including Automotive, Aerospace, and Digital Manufacturing

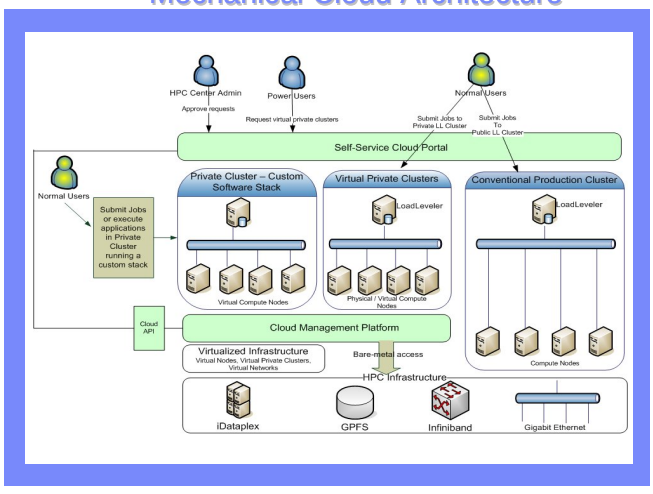
Solution Definition

- Engineering portal and HPC Cloud Infrastructure optimized to support 2D/3D workloads in batch or interactive mode for design, simulation and analytics.

Key Capabilities

- High Performance Computing Infrastructure
- Agile Systems and Workload Management
- Accelerated 3D remote graphics (via Partner)
- ISV Partner Integration

Mechanical Cloud Architecture



Private or Private Hosted Cloud

IBM Engineering Solutions for Cloud have been designed to address technical and business issues within and across engineering domains (see following figure). Both EDA and Mechanical clouds are a key part of this overall architecture.

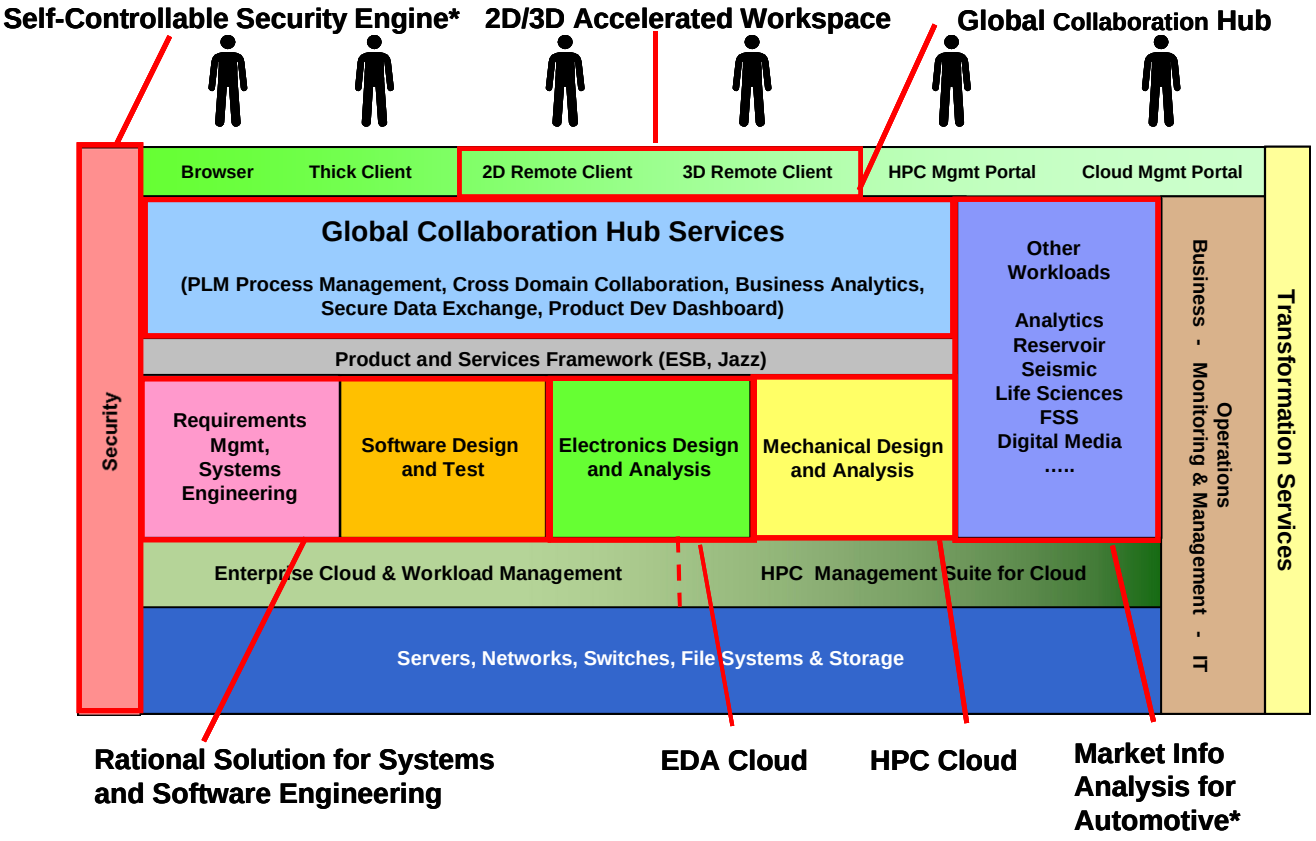


Figure 3: Conceptual Architecture of IBM Engineering Solutions for Cloud (*Source: IBM, June 2011)*

Examples Highlighting the *Benefits* of IBM HPC Cloud Implementations

IBM and its business partners serve a variety of customers who use cloud computing environments to access HPC capabilities. These HPC cloud cases represent a range of engineering industries and application uses. In all cases, IBM's technology, expertise, resources, and support are critical to the customer.

IBM EDA Design Cloud Delivery Management System – *optimized for engineering productivity – managing computationally demanding batch workloads with time- critical interactive user workloads*

Challenges: To meet stringent design schedules and quality goals on a budget for complex batch workloads in Logic Simulation, Design for Manufacturability (DFM), Layout versus Schematic (LVS), Design Rule Checking (DRC), Optical Proximity Correction (OPC), Timing, Synthesis & Build, etc. Further, over 3000 interactive Design Cloud users demand for rapid response jobs related to Schematic and Logic Entry, Layout Entry, Design Auditing, etc.

Solution: IBM has implemented and refined a solution over the last twenty years. This solution is optimized for high-utilization for a continuous stream of demanding batch jobs and enables remote high resolution graphics on laptops. One site holds design projects servers and data accessible globally by design engineers. Interactive jobs have highest priority to minimize designer wait time. The infrastructure consists of a very large high performance cluster for OPC while Logic Simulation uses a massive FPGA system. This architecture can support many architectures and operating systems including IBM Power Systems and x86 clusters.

Benefits: Achieves 80-90% CPU utilization and 60% memory utilization with minimal overhead and no virtualization. Highly valuable designer time is also optimized for personal productivity. Stringent product design and development schedules and deadlines have been met consistently.

Exa –*enhanced competitive differentiators with a combination of private and private-hosted cloud*

Customer: Exa Corporation, based in Burlington, MA, develops, markets and supports simulation software for the fluids engineering marketplace along with a full suite of engineering consulting services.

Challenge: Exa was looking for an engineering cloud partnership that would deliver seamless integration and support of their flagship product, Exa PowerFLOW, a CFD solution for simulating complex fluid flow problems, especially for global auto majors and heavy trucks.

Solution: IBM designed and implemented a solution comprising IBM Intelligent Cluster x3650, x3550 M3 with Mellanox QDR InfiniBand interconnects; large memory and GPFS; 72 TB of DS3512 Storage with a IBM private hosted data center running Exa® PowerFLOW® Scalable, Digital Wind Tunnel™ technology and services.

Benefits: Exa now has the ability to offer PowerFLOW on demand as a turn-key service providing client's product insight much earlier in the design cycle than previously through a cost effective solution.

"Based on our experience in providing Exa PowerFLOW OnDemand to our clients hosted in IBM's Cloud datacenters, the new IBM Engineering Solutions for Cloud will be an integral part of our business strategy" - Ed Furlong, Exa CFO/COO, June 2011.

NERSC –*built Magellan - an ultra-efficient private HPC cloud for the Department of Energy (DOE)*

Customer: The National Energy Research Scientific Computing Center (NERSC) is the flagship high-performance scientific computing facility for research sponsored by the U.S. Department of Energy Office of Science. NERSC, a national facility located at Lawrence Berkeley National Laboratory, is a world leader in providing resources and services that accelerate scientific discovery through computation.

Challenges: To support Magellan, an HPC cloud project, run jointly with the Argonne National Laboratories, NERSC wished to replace two existing supercomputing clusters, to create a scalable and modular cluster that would enable expansion with a unifying architecture at known price-points. Further, to enable an orderly switch-over to the new solution, NERSC required the ability to start up the new cluster before the old clusters were decommissioned. This created extreme constraints on physical floor space and cooling capacity, and NERSC wanted the highest possible computing density for the new solution. This greatly increased the cooling challenge by packing more hot components into a smaller area. NERSC also has a strategic architecture policy of keeping shared storage separate from the computational facilities, to make it easy and non-disruptive to upgrade the computing environment.

Solution: IBM designed and implemented a new IBM Intelligent Cluster solution comprising 400 IBM System x iDataPlex dx360 M3 compute nodes with IBM Rear Door Heat eXchanger water cooling and Voltaire Grid Director 4700 QDR InfiniBand switches in a HyperScale configuration. NERSC also installed a second iDataPlex cluster, with 720 iDataPlex nodes. These computational nodes run as a traditional cluster to a fully virtualized cloud model and IBM Global Parallel File System (GPFS) offers high performance and stability for data and storage management.

Benefits: NERSC was able to run their IBM HPC clusters during the decommissioning phase. The power usage effectiveness (PUE) rating of the data center was reduced from 1.35 to just 1.15.[2] These energy savings make the new iDataPlex cluster a lower-cost, higher-performance option than continuing to run the two older clusters. NERSC can continue to add more computing capability when needed in a non-disruptive mode. A commercial public cloud provider would have cost two to three times more and applications would have run up to fifty times slower than using the center's internal private cloud.[3] The center serves the needs of scientists and researchers who transmit data across the cloud. And now supports 3000 users with 300 applications that run over a cluster of 9,000 shared processors and has 1.5 petabytes of data.

North Carolina State University – *24x7 access to general and HPC cloud computing for students and faculty*

IBM and North Carolina State University have collaborated to establish the Virtual Computing Lab (VCL)[4], a cloud computing-based technology, to provide students around the state and the University of North Carolina system

[2] http://public.dhe.ibm.com/common/ssi/ecm/en/xsc03076usen/XSC03076USEN.PDF
[3] http://gcn.com/articles/2011/03/28/high-performance-clouds-in-house.aspx
[4] http://vcl.ncsu.edu/

campuses access to advanced educational materials, select software applications and computing and storage resources on demand. Today, VCL is open to 30,000+ NCSU students and faculty. At any given time, 1300 to 1800 IBM BladeCenter blades are in use of which 500 to 700 work in non-HPC mode and the rest in HPC mode. VCL delivers over 460,000 CPU hours annually to general workloads, over 7,000,000 HPC CPU hours annually, distributed over four data centers.

How it works: The VCL infrastructure consists of three tiers: a web server, a database server, and one or more management nodes. At the heart of VCL is a web-based service for scheduling and provisioning remote access to high-end computational resources. These resources consist of blade computers located in multiple data centers and other specialized University lab computers. The IBM BladeCenter compute resources are dynamically loaded on demand with a choice of operating system images and predefined application sets in either bare metal or hypervised environments. The blade servers also provide flexibility between High Performance computing and academic computing by easily repurposing during low use times between clusters used for batch processing or single seat use for less compute intensive work. IBM University Delivery cloud services are also available to other institutions and their students/faculty to replicate the VCL cloud. This service facilitates remote collaboration and access to servers and comes with additional support for installation and configuration, instruction and training, and systems/image management.

Challenges: VCL began in 2004 with a simple idea of providing dedicated remote access to a range of computing environments for students and researchers to access from any networked location either on or off campus. In a shared computing resource environment, where many students run high-end applications or experiment with computer science coding assignments on the same computer, the level of service degrades very quickly. Additionally, software media distribution for distance education students was prohibitive. Depending on the application and the license agreement, it was limited to university owned computers and was not allowed to be installed on the student's personal computer.

Why IBM: VCL exemplifies the goals of the Virtual Computing Initiative (VCI), launched by IBM and NC State University in 2006, to improve the quality of education through the application of technologies that include virtualization, cloud computing, hosted client-server models, and robust, energy efficient IBM Systems.

Conclusion

High performance computing helps academic and enterprise users achieve the speed, agility, and insights to lead the market. But handling, analyzing, and visualizing the exploding datasets resulting from HPC is straining data center capacity. Pressure on IT budgets, the need to improve utilization and return on compute dollars, and escalating energy and operational costs of building and maintaining data centers are forcing enterprises to adopt cloud computing models. HPC users are thus naturally drawn to cloud computing with its promise of pay-as-you-go supercomputing.

But apart from the traditional corporate concerns of data security of computing over a cloud, the main problem with running HPC or engineering tasks on conventional clouds is that conventional clouds are geared toward supporting general-purpose applications and services -- short transactional workloads such as Web applications and database tasks. HPC tasks, on the other hand, are mostly complex, long-running algorithms processed in parallel, with the result of one task not dependent on the outcome of another. Processing threads are brought together at the end of the activity.

Today, most HPC environments today run on bare metal (i.e. without a virtualization layer), often with high-performance interconnects to maximize absolute performance. However, in the future, we envision HPC clouds to be powered by computational clusters with software optimized for performance, flexibility, reliability, availability, and manageability.

IBM, with its long history of enterprise computing, addresses these security concerns and the specific computing needs of the HPC community through server (x86 and Power), storage, and software (IBM and partner) that is tailored for HPC requirements. These solutions are delivered over private and private-hosted clouds and implemented at many clients worldwide.

Engineering customers, in particular, will benefit from HPC over the cloud through the IBM Engineering Solutions for Cloud for EDA, MCAE, Systems and Software Engineering applications. IBM Engineering Solutions for Cloud has been optimized to address technical and business processes within and across engineering domains with the ability to extend to the engineering ecosystem. The engineering community will now be able to deliver designs better, faster and cheaper; design higher quality products in less time and across a distributed supplier network; and respond to increased CIO cost control of engineering IT.

Appendix: Cloud Computing Models and Workload Classification

As Cloud Computing Mainstreams, Multiple Delivery Models Emerge

In many ways, cloud computing is the next logical evolutionary step, building upon the industry's rapid adoption of Linux, open source solutions, high performance cluster computing, SOA, and more recently virtualization. Here we capture the cloud's broader concepts and business benefits.

The value of cloud computing: Cloud computing promises to provide dynamically scalable and often virtualized IT resources (hardware, software, and applications)- 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), and builds upon recent IT infrastructure solution concepts such as grid/cluster computing, utility computing, and autonomic computing.

With a spectrum of flexible offerings and pricing models, cloud service 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. These smaller companies could now use private-hosted or public clouds. On the other hand, larger organizations can use private or private-hosted clouds to 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[5, 6, 7, 8] 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 private-hosted (extending private) cloud.

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 open systems management and workload scheduling solutions 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.

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, $V_{W,}$ 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 indicate the size of workload growth in the future across the two primary dimensions. For

[5] NIST, "The Definition of Cloud Computing", http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf
[6] Tim Jones, "Cloud Computing with Linux", http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux/l-cloud-computing/l-cloud-computing-pdf.pdf
[7] Jeffrey Rayport and Andrew Hayward, "Envisioning the Cloud: The Next Computing Paradigm", March, 2009.
[8] Ashar Baig, "A Cloud Guide for HPC", May 2009.

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 10's or 100's 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.

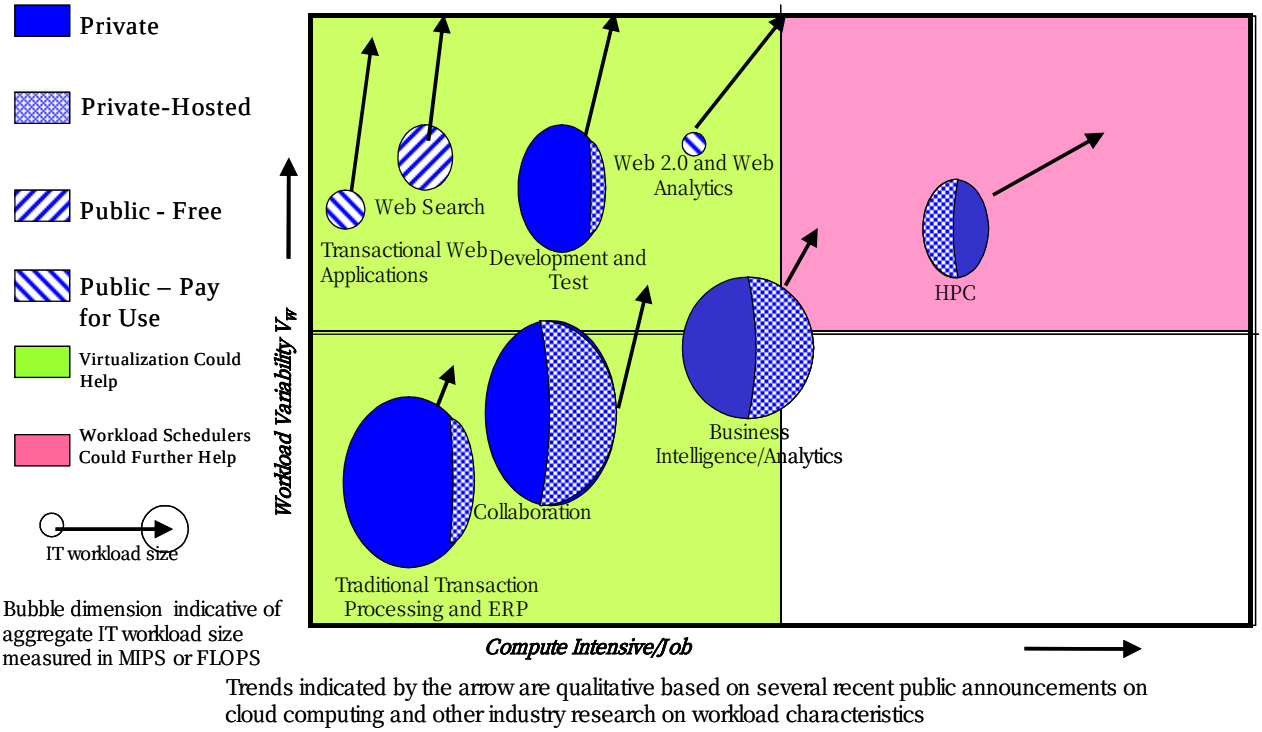


Figure 4: High Level Workload Classification by Application Domains

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), Private – Hosted (dedicated to one client, e.g. IBM hosted), Public (standardized offering, public access through the Internet), and Shared Hosted (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 and KVM are two prominent examples of virtualization solutions for x86 environments. The IBM

System z mainframe or System p servers are also excellent examples. 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 Computing, Adaptive Computing, Gridcore, 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. However recently, holistic systems designs and reliability-aware software solutions based on virtualization and scheduling solutions have been used to enhance reliability and availability and to maximize HPC application uptime[9] while delivering the quantum increase in performance and scale needed for tomorrow's HPC applications. IBM led Extreme Cloud Administration Toolkit (xCAT) solutions have been used effectively in this regard for very large scale-out cloud environments.

[9] Raju Gottumukula, *et. al*, "Reliability-Aware Resource Allocation in HPC Systems", IEEE International Conference on Cluster Computing, 2007, pp 312-321.