

Best Practices and Smart Solutions for High Performance Research Computing: Promoting Global Collaboration and Innovation with Researchers and IT at the Center of the Higher Education Enterprise

Sponsored by IBM – March 2010

Sriini Chari, Ph.D., MBA,

<mailto:chari@cabotpartners.com>

Introduction – Research Computing is Key to the Future of Innovation and Discovery

Higher education institutions regard investments in research computing strategic to recruit outstanding faculty, students, and staff, improve brand equity of their institutions, build world-class research and innovation capability, and enhance the long-term economic and competitive positions of their stakeholders.

The high performance research computing landscape at major universities continues to evolve. The pendulum swings back-and-forth between central and local computing architectures resulting¹ in a healthy tension between researchers and their IT support organizations. Several factors like funding, technological advances, the complexity of the engineering and scientific challenges, and the availability of skilled professionals to support research computing have had a profound impact throughout this time period and have caused this flux. However, one theme that remains unchanging is that the needs of research computing continue to drive the envelope of computing in industry, government, and academic institutions.

This article draws upon several public documents, recent trends, and over ten in-depth, one-on-one interviews with pioneers and leaders of research computing deployments at various higher education institutions. Our survey covered over forty questions ranging from high-level business and scientific areas to more detailed technical, organizational, and cultural issues facing these institutions. Here, we provide a detailed perspective on the evolving landscape of research computing and the current and future needs of researchers along with best practices and solutions adopted in recent years to overcome these challenges. IBM academic initiatives and leading-edge high performance computing solutions to help research institutions address these challenges are highlighted through illustrative examples and client case studies.

What is Research Computing?

Research computing through advances in computational simulation and modeling has **dramatically extended the exploration of fundamental processes** in nature e.g., the interactions between elementary particles that give rise to matter, the interactions of proteins and other molecules that sustain life, the prediction of weather and climate, the formation of galaxies and stars – as well mankind's ability to predict the behavior of complex natural and engineered systems e.g., nanoscale devices, microbial cells, fusion reactors, performance of aerospace and automotive vehicles, location of petroleum and natural gas resources, and the efficacy of therapeutic drugs.

Research computing bridges theory and experiment in science and engineering by allowing the execution of “computational experiments” on systems, including those that do not yet exist in the physical world. Moreover, computational science and experimental science complement each other by permitting a complete picture of the system that neither one alone can provide. Often these computational experiments provide a breakthrough insight into physical phenomena that even the best modern instruments like telescopes, microscopes, imaging and spectroscopic equipment cannot easily provide or – as in many cases – the experiment is either too hazardous or expensive or even infeasible to conduct. Additionally, modern instruments and data acquisition devices frequently used by experimental scientists require modern computer platforms that need to be integrated and supported with the computational scientist's platforms.

Copyright © 2009. Cabot Partners Group, Inc. All rights reserved. The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States and/or other countries: IBM, the IBM logo, AIX, System x, System p, Blue Gene, iDataPlex, General Parallel File System, GPFS, and 1350. Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both. Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both. Other companies' product names or trademarks or service marks are used herein for identification only and belong to their respective owner. All images were obtained from IBM or from public sources. The information and product recommendations made by the Cabot Partners Group are based upon public information and sources and may also include personal opinions both of the Cabot Partners Group and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. The Cabot Partners Group, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document.

Research computing relies heavily on advances in applied mathematics and computational science. There is a cascadic flow of information between the engineers and scientists and the applied mathematicians and computer scientists. In this collaborative context, the role of the applied mathematician is to provide a rational means of passing between the physical model to the discrete computational representation and for effectively manipulating that representation to obtain results that are meaningful. The computer scientist devises algorithms and processes to perform these manipulations on leading-edge computing systems, and provides ways to leverage these algorithms and processes across other applications and propagate these concepts over a complex and constantly-evolving set of computing architectures^{2,3}.

A Historical Background of Four Decades of Research Computing in Higher Education

Over the last 40 years, there have been several trends in research computing where the pendulum has swung back-and-forth between centralized data center approaches and a more local approach to research computing support.

The mainframe era – it was centralized in the beginning! During the 1970s and into the early 1980s, bulk of research computing was done on traditional mainframe environments that used a centralized time-sharing model for computing. While this model worked well for a large constituent of users (about 80%), 20% of the research users who had an increasing appetite for computing capability were not adequately served with this model. This caused them to look for alternative computing paradigms to address their unique needs and they built up local (housed in departments) infrastructures to address these needs. This gave rise to the client-server model of computing at higher education institutions.

The client-server era – moving to distributed architectures. From the mid-1980s to the early 1990s, this client-server model was prevalent in research computing; as end-users, professors, research scientists and engineers began to deploy these Unix based client-server architectures in their own departments. The growth of the power of the microprocessors (as per Moore's law), coupled with significant equipment funding from NSF, NIH, DARPA, and other federal/state funding agencies enabled research computing users to tackle significant computing problems of their day and thus could push the envelope of science and engineering.

The exotic parallel supercomputing era – tackling grand challenges with centralized scalable parallel. From the mid-1990s onward, research computing problems became increasingly complex in size, scale, and the resolution needs of the underlying physical phenomena yielded more fine grain structures. This required the researchers to map their computational problems onto massively parallel supercomputers and other large centralized supercomputers. The acquisition costs of these computers coupled with the diminishing funding from the federal agencies such as the NSF, NIH, etc. for individual researchers to procure computing infrastructure, enabled the creation of centralized national supercomputer centers at several higher education institutions. Most of these supercomputer centers received substantial grants from federal agencies and large equipment manufacturers to create these centralized facilities for research computing. Individual researchers were no longer adequately funded for procuring equipment that satisfied all their needs. The agencies that sponsored their research largely focused on funding costs that were directly related to the actual research and encouraged individual researchers to use these centralized supercomputing facilities. Again, due to the lack of flexibility and the often cumbersome (to the individual researchers) parallel application enablement, centralized policies of system administration, and lower prioritization caused these researchers with very voracious computing appetites to look for alternatives.

The commodity cluster era – providing economical capacity and performance local to the researcher. The continuing growth in the performance of microprocessors, high-speed interconnects, coupled with innovations in open source software efforts (e.g. Linux, Internet, the w3, etc.) enabled research users to invest in building computing capability by clustering local resources that were largely in their own control. Even the hardware manufacturers began shipping low-priced, affordable “cluster in a box” and blade systems that were very attractive to solve the research problems of the day. Very soon the utilization levels at the centralized supercomputing facilities began to fall as workload migrated back to the individual departments in the late 1990s to the early 2000s. However, over time, many institutions realized the problems with server sprawl and the associated increase in the total cost of computing, and once again the

pendulum began to swing towards centralization of research computing to manage this escalating TCO. Furthermore, grand challenge problems still required centralized ultrascale systems.

Today's cyberinfrastructure (CI)-harnessing campus and national resources. Since the 2000s, there has been a concerted effort in the United States – largely sponsored by funding agencies such as the NSF and the Office of Cyberinfrastructure (OCI) – to coordinate and leverage research computing activities and interests at a national (even international) level with diverse stakeholders⁴. The following figure depicts the existing research computing landscape as viewed through the traditional lens of the Branscomb Pyramid⁵.

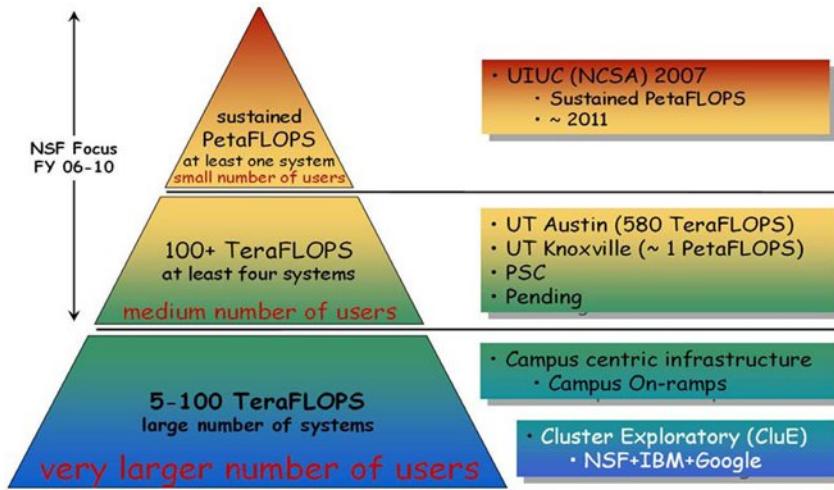


Figure 1: The Research Computing Branscomb Pyramid

At the top of the pyramid are a few national centers characterized by ultrascale supercomputing centers with petaflop performance supporting relatively few research computing users seeking solutions to grand challenge problems. At the bottom of the pyramid is the campus infrastructure, ranging from a single research laboratory to regional/state resources, supporting a very large number of research computing users.

While NSF's TeraGrid initiative has greatly advanced coordination of access to national resources, these national resources have not been coherently integrated with campus resources or other non-TeraGrid resources. Broader integration challenges are both socio-cultural and technical in nature. What's required is a seamless and efficient way for campus researchers to integrate their computational workflows across both local and national resources while reining the overall costs of research computing. This should alleviate the existing tension between researchers and centralized research computing organizations. Both parties must have a stake in building their institution's long-term brand for pioneering innovation and discovery.

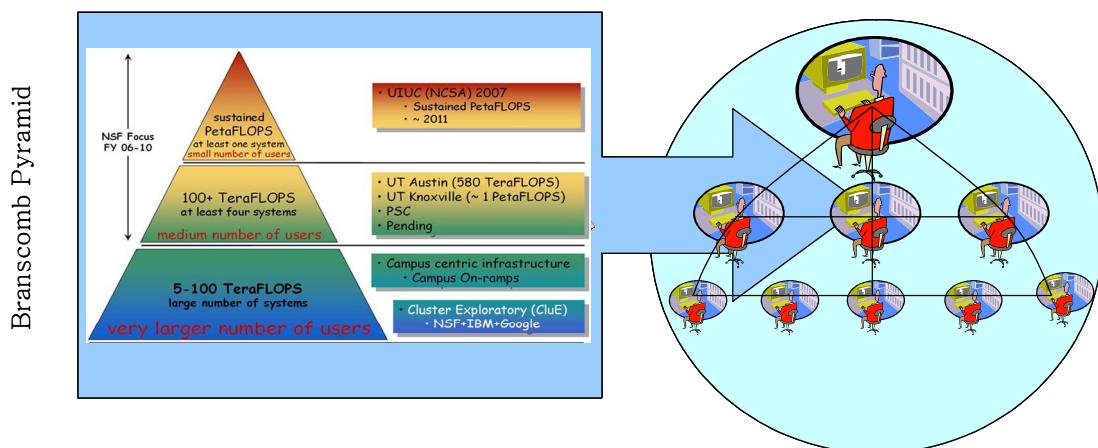


Figure 2: Collaboration/Seamless Access to Computing Resources to Foster Discovery and Innovation

Current Challenges and Trends in Research Computing at Higher Education Institutions

We interviewed several research computing IT directors at large and small computing centers at various phases of their life cycle. Some centers have been operating for over a decade while others are earlier in their life cycle; operating for just a few years. While some challenges are unique, many business, scientific, and technical challenges are common across many of these institutions. These challenges along with some current research computing tends are summarized here.

Ongoing funding challenges require new approaches. Today, over 50% of the funding at smaller state or campus research computing centers comes from their institutions/state. Research computing IT directors are pressured to move a larger fraction of their center's funding jointly with researchers who must have appropriate incentives to share in this funding. IT directors must invest in specific efforts to enhance:

- **Growth** – need outreach and business development to grow user base beyond the traditional user base while sustaining current base. Avenues for growth must include expanding the applications of research computing to emerging areas of life sciences, energy, social sciences, and humanities, and also with businesses and industries in the state.
- **Advocacy** – articulate the value of HPC to senior administrators and strategic decision makers while educating researchers on the TCO advantages of centralization, and
- **Skills and environment** - primarily people, software, policies, and standardized processes to support needs of researchers especially the non-traditional HPC users in the social sciences and humanities, and other industrial users. Beyond traditional HPC systems administration skills, investments in computational application scientists who have a combination of domain specific and IT skills are required.

Research IT needs to support large-scale global collaboration. Scientific and engineering research is increasingly moving towards a model where researchers and analysts are less concentrated at a single location^{6,7,8}. A lot of physical sciences research is headed in the direction of the Large Hadron Collider (LHC)⁹, Panoramic Survey Telescope and Rapid Response (Pan-STARRS)¹⁰, Sloan Digital Sky Survey (SDSS)¹¹ type of large collaborations across multiple institutions spread across different continents. This frequently involves the use of remote equipment, access to remote data and applications, and distributed analysis. A purely central data-center type solution targeted towards a single institution fails to address these needs for global collaboration and pushes the costs of such collaborations back onto the individual researchers who will respond by moving back to more local architectures for computing support.

Grand challenge and smaller production simulations must be supported simultaneously. Effective computational science and engineering requires a fusion of interdisciplinary advances in scientific models, mathematical algorithms, information technology architectures, and disciplined software engineering. Moreover, the interdisciplinary nature of today's large scale scientific and engineering problems (e.g. reacting turbulent flows, fusion energy sciences, pharmacogenomics, environmental and space sciences, etc.), requires a deeper level of sustained global collaborations and the balanced use of computing capability for heroic simulations (grand challenge simulations) coupled with capacity computing for production simulations or early testing and debugging of grand challenge simulations.

Escalating research computing costs are driving IT centralization. In recent years and with the current economic downturn, computing support for research in large universities and other institutions is increasingly being centralized¹² driven by equipment procurement costs, funding, and administrative issues prevalent in totally local models. It is vital to look at the total cost of ownership (TCO) for research computing along with the total value of ownership¹³. Direct software, support, maintenance costs, and other escalating operational costs such as energy, power and cooling, people, and facilities are often transferred back to the researcher or their departments. These costs could dominate the costs in the research computing life cycle¹⁴. The following figure qualitatively depicts the changing costs of research computing over the last four decades. While the IT hardware costs as measured by \$/performance have come down significantly with the adoption of new HPC hardware technologies, the associated software and operational costs continue to rise, and in fact, today, these costs dominate the TCO.

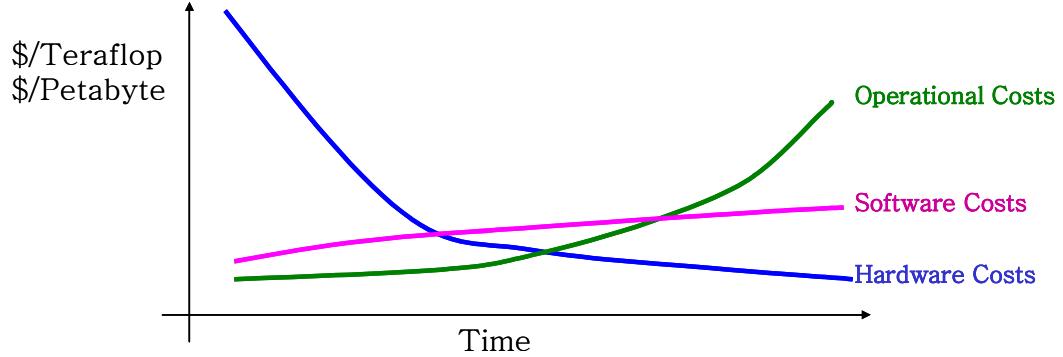


Figure 3: The Shifting Costs of Research Computing Motivate IT Centralization

An older 2005 TCO¹⁵ study for campus research computing conducted by Indiana University illustrates these TCO drivers for one teraflops (a medium scale system at the time) and hundred teraflops of peak performance (a very large scale system at the time). Three cases for each instance were investigated: worst case distributed, best case distributed, and completely centralized. Hardware costs are the same for each case as should be expected. The power costs for facilities and IT equipment are lowest for centralized. Staff costs for centralized are higher; probably because in distributed scenarios, graduate students may have managed many of these systems, and their costs were not explicitly considered. More importantly, the downtime and under-use penalties for centralized are lowest; with this differential being greater on the larger system. Total centralized costs as measured by \$/CPU-hour is lower for the large system – further illustrating the economies of scale of centralization: improved energy-efficiency, utilization, and reliability.

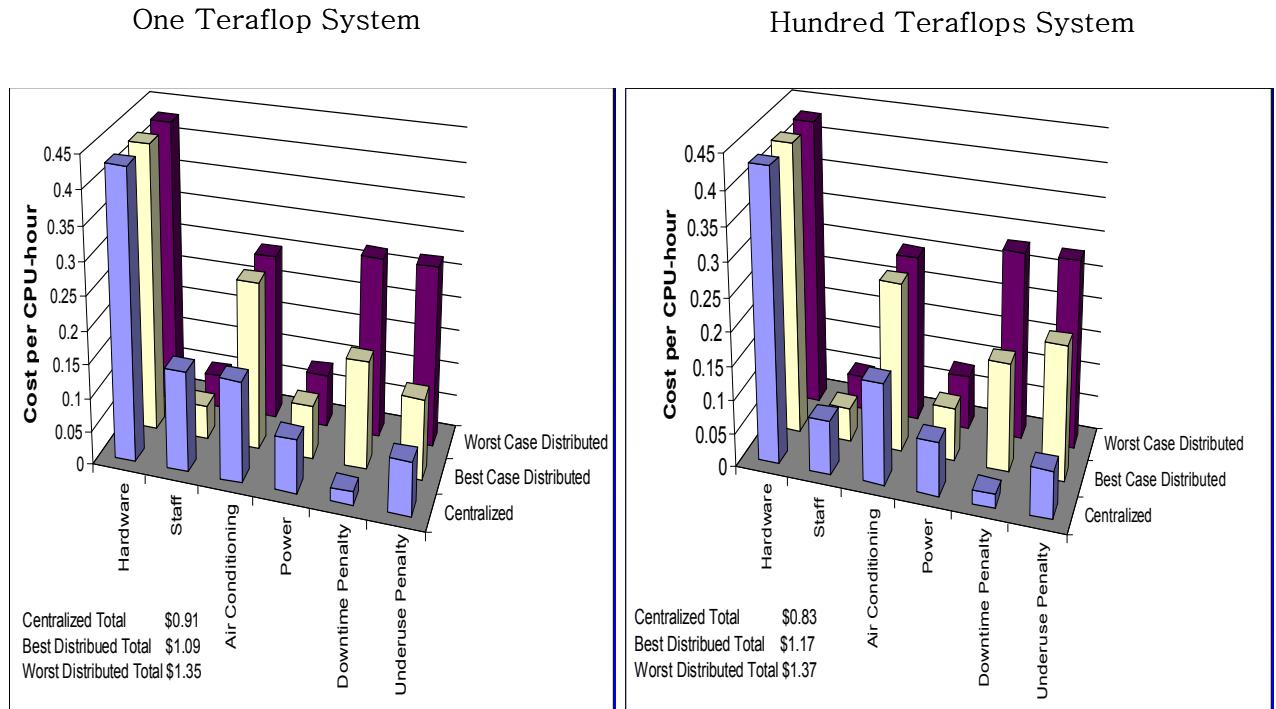


Figure 4: 2005 Indiana University TCO Analyses for Distributed vs. Centralized Research Computing

Reliability concerns magnify with scale. In addition to application performance, users demand systems that deliver excellent reliability, availability, and serviceability (RAS) – just as these systems are increasingly built with thousands of components (processor, memory, disk, switch, power supplies, and so on) to deliver the needed performance and functionality. This is a challenging engineering design and

deployment problem as failures grow with the number of components. It's crucial for today's petascale and tomorrow's exascale (1000 petaflops) systems to possess the RAS characteristics approaching today's mission-critical business systems. Merely scaling up systems designs with "commodity" components will just not cut it.

Most parallel HPC applications have multiple processes that run concurrently on many nodes with communication over high-speed interconnects. A single failure in one of these processes can cause an outage in the entire application, requiring the user to restart the application from the previously checkpointed state. This hampers a user's productivity, diminishes innovation, and increases the annual TCO. In fact, a 1% reduction in system availability could result in a loss of several million dollars of bottom-line benefits at institutions that rely on large-scale research computing environments¹⁶.

The road to exascale will require significant innovations. Today, in 2010, a ten - hundred teraflops system would be considered a medium scale system and the larger ones are approaching or at the petaflops range. The electricity costs for today's systems continue to escalate even more rapidly driven by increased unit costs and denser systems packaging. In fact, annual energy cost estimates for exascale systems are in excess of \$100M¹⁷; probably more than the expected annual capital costs for these systems. Clearly, significant technology and process innovations in energy-efficiency and reliability would be required to support the performance needs of future research computing users, particularly at national centers.

Network limitations worsen even as computing and storage capacity barely scale with data deluge. The amount of information is soaring, estimated to double every 9-12 months¹⁸. Fortunately, storage capacity per dollar also continues to double every 12 months, so merely storing and keeping up with this information flood is somewhat practical. Analyzing it to spot patterns and extract valuable insights is hard even with computing capacity doubling every 18 months. But it's even harder to effectively share this data over collaborating research institutions, as university-to-university network performance doubles only every 4 years. From the 1970s to today, the ratio of computing performance to wide area circuit bandwidth has grown from 20 flops/bit to 1000 flops/bit; hurting the balance between computing performance to network bandwidth⁴. The ratio of data set size to network bandwidth is far worse!

Many research computing applications with very large data sets, e.g. high-resolution regional weather forecasting, require significant "computational steering" through human-computer interaction and visualization. These applications work best if network latency effects are minimized, further motivating the need to maintain significant computing capability local to these data sets at individual campuses or even at specific campus sites while accessing national resources for other data sets e.g. national weather data.

How Higher Education Institutions are Addressing Current Research Computing Challenges

It's not surprising that the higher education research computing ecosystem is a continually learning system! The newer research computing centers have learned and implemented many of the effective practices followed at more mature centers. This learning is further stimulated by government agencies such as the NSF, higher-education organizations such as EDUCAUSE, and active participation by HPC solution providers such as IBM through academic initiatives. Here we summarize and provide a framework to articulate some of the effective practices followed by these centers to continually address the challenges in research computing largely driven by the ever changing forces of funding, scientific quests, and technology.

Find innovative ways to increase funding. Research computing centers, particularly at a campus or regional level, have increased collaboration with individual researchers to jointly write research grant proposals from sponsoring funding agencies with a focus on specific application domains crucial to other state/local initiatives. Many of these centers are also actively partnering with the HPC ecosystem solution providers such as IBM to proactively seek-out targeted agency and state funding in emerging areas of national strategic importance such as life sciences, energy, and nanotechnology.

Grow user base. Many research computing centers have established targeted programs to recruit new users. They have invested in business development and outreach to recruit industrial users and users from other collaborating institutions. In addition, as research computing on campus is increasingly being used in non-

traditional disciplines such as the humanities, social sciences, and economics, many centers provide proactive support, HPC skills-training workshops, and even one-on-one mentoring to encourage computing use among these new and non-traditional users.

Standardize the computing services environment. While innovation often requires computing environments with significant flexibility; in order to manage users' expectations on resource priorities and minimize potential scheduling conflicts, research computing centers are increasingly standardizing the processes and policies of their computing services environment. Most contracts and administrative policies are standardized and transparent. This promotes increased collaboration in the user base while reducing the tension between researchers and the centers. Some mature research computing centers also track user satisfaction through regular surveys to build effective loyalty programs. This is also being implemented by newer centers.

Improve the center's advocacy at all levels. Research center directors – many of whom are faculty members - continue to spend significant efforts to educate key university decision makers such as the Provost and President, other senior university and state stakeholders on the strategic value of research computing. To promote the center's capabilities and research and operational value, most centers hold seminars, workshops, and educational events to expand reach with non-traditional users while improving loyalty with the current user base. The more mature centers have hired computational application scientists and domain experts to interface and deepen collaboration with researchers. They also have on staff seasoned marketing and business development professionals for increasing awareness of their centers and managing the complex and somewhat tenuous relationships with all key stakeholders in their ecosystem. All centers have a significant investment in providing traditional IT support and services, often 24 by 7, staffed by a pool of systems administrators who ensure that the IT value provided is maximized.

Develop a cost-value framework for research computing investments. As in the case of Indiana University, increasingly more research computing centers are developing business cases for their investment decisions in many ways similar to enterprise CIO organizations. This helps justify decisions, improves the center's advocacy, and deepens collaboration between researchers, university administrators, and the IT organizations, as they all work toward common goals. 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 of inter-related drivers and associated metrics that examine the total costs incurred and the value delivered by research computing solutions.

Value

- Strategic Value: e.g. ability to attract and retain top faculty and staff, access to increased funding, brand equity, long-term economic development, better curricula,
- Research Value: e.g. breakthrough research and innovation, broader and deeper research collaboration, greater scientific insights, more publications,
- 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 e.g. new servers, storage, networks, power distribution units, chillers, etc.
- Data Center Facilities 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.

Universities must continually evaluate the costs and value of research computing within this broad cost-benefit framework. They must maintain a focus on maximizing value not just for one application but a

collection of workloads typical in many research environments. This should further minimize tension between researchers and IT.

Grid/cloud architectures are attractive to ameliorate ongoing tensions between researchers and IT.

Pure centralization may not adequately address all the critical current and future needs of research computing especially among “big science” researchers or with other researchers with very unique high priority local needs. Architectures based on grid/cloud computing promise to be well suited to address the challenges of research computing in a collaborative, scalable manner that put researchers at the center of the computing enterprise yet help rein in the escalating costs of research computing through standardization, self-service portals, and pay-as-you-use business models.

The grid computing era – fostering global research collaboration. Research computing problems continued to grow in size, scale, and complexity. The need to collaborate between various research teams throughout the globe and across multiple disciplines became more crucial. Again, the researchers needed computing infrastructures that could be available on demand and could dynamically scale to address these grand challenge interdisciplinary problems. Since the 2000s to the present a lot of software efforts to develop grid computing solutions helped tie-up the central underutilized supercomputing resources with local clusters to enable the researchers to access vast computing resources on demand to solve their complex problems¹⁹. This approach provides the flexibility to make discovery and development simulations locally while harnessing the collective (local-central) resources for production full scale simulations. It also enables research collaboration across the globe and across several disciplines and maximizes the use of computing cycles throughout the extended research enterprise.

The cloud computing era – putting the researcher at the center and optimizing IT. 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/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 research institutions of all sizes to become more agile and innovate while reining in costs.

Over the last two years, with the increasing interest in cloud computing, many excellent articles^{20, 21, 22, 23} 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 a research enterprise) cloud or as a public (accessible through the internet) cloud, or as a secure hybrid (extending private) cloud.

With cloud computing solutions, even smaller regional universities – that typically face steep entry cost barriers to access IT resources will no longer need large capital outlays in hardware or facilities to deploy their research computing services or the labor to operate these IT facilities. On the other hand, larger organizations benefit from the increased business and scientific value resulting from the added capability and flexibility to rapidly deploy standardized yet customizable “self-service” solutions that automate and scale research and academic computing environments end-to-end while minimizing escalating labor and infrastructure costs.

Develop and enhance long-term partnerships with other stakeholders in the research computing ecosystem. Research computing institutions have developed long-term strategic partnerships with HPC

solution providers such as IBM. These collaborations are crucial to manage the rapid changes in technology and keep up with the needs of innovation and discovery at every level in the research computing ecosystem.

IBM Delivers a Broad Optimized HPC Solutions Portfolio for Research Computing

Throughout the Branscomb Pyramid, IBM offers a wide array of HPC solutions through its multi-core processor systems, large storage systems, support for a broad range of operating systems, visualization, innovative applications, middleware and partner ISVs with proven expertise and deep industry presence. IBM has the leading portfolio²⁴ of HPC architectures, systems, and software ranging from the System x® Cluster 1350™, Blades, iDataPlex®, Power Systems®, and Blue Gene® with support for a range of operating systems including Linux®, AIX®, and Windows® together with cluster management software, a high-performance shared-disk clustered file system - General Parallel File System (GPFS™), and optimized scientific and engineering libraries. In addition, IBM has a worldwide research and technical staff of domain experts to collaborate with researchers to migrate and optimize their applications on the IBM HPC portfolio to solve their largest and most challenging problems.



IBM continues a strong tradition of peer-to-peer research collaboration with academia. Researchers at IBM laboratories work with peers in universities around the world. These collaborative relationships are fostered through fellowships, grants, and funding for programs of shared interest including supercomputing and the exploitation of multi-core technologies. For instance, IBM Research is collaborating with seven universities on various software aspects of multi-core computing²⁵.

The IBM Blue Gene provides the best combination of energy-efficiency and lowest TCO for ultrascaling research computing²⁶. The platform differentiators for the Blue Gene/P are its higher energy efficiency, unparalleled scalability and smaller footprint (high package density), coupled with a standardized parallel programming model and software tools that permit the migration of a broad number of HPC applications with minimal parallel algorithmic invention.

IBM Power-based supercomputers have the best mix of performance, RAS, and utilization. We believe that over the past decade, IBM, particularly with Power-based supercomputers, has been able to deliver supercomputers, associated HPC solutions, and other complementary IT infrastructure solutions with the best mix of sustained performance, and reliability/availability, and utilization¹⁶.

The IBM iDataPlex is a scalable and economical x86 workhorse for research computing: The iDataPlex solution from IBM is a proven x86 clustered system extending the IBM blade and cluster product portfolio for high performance research computing. This “green” IBM system can help lower power consumtion by up to 40%, and can greatly reduce the air conditioning need in data centers. This architecture can economically increase the compute density by a factor of five while retaining many of the attractive attributes of current integrated blade and rack cluster solutions from IBM and is part of the IBM System x Cluster 1350 portfolio.

IBM cloud computing initiatives for research computing. IBM academic initiatives in cloud computing and high performance computing are key elements of the overall IBM cloud computing strategy (www.ibm.com/cloud). IBM delivers solutions in all the three different cloud architectures – public, private and hybrid clouds, and services to help clients with deployment and address security and privacy concerns.

Examples Highlighting the Benefits of IBM Research Computing Initiatives

IBM's collaborations with higher education institutions are broad, global, and very deep. Here we provide just a few illustrative examples covering the spectrum of IBM HPC solutions.

Rensselaer Polytechnic Institute – collaboratively pushing the frontiers of nanotechnology through supercomputing. The Computational Center for Nanotechnology Innovations (CCNI) was set up in 2007 as a collaborative venture between Rensselaer Polytechnic Institute, IBM, and New York State. Located in the Rensselaer campus and at the Rensselaer Technology Park in Troy, NY, CCNI is designed to be among the world's most powerful university-based supercomputing centers and a top supercomputing center of any kind in the world. CCNI collaborates with researchers from both universities and industry. The center is designed to both continue the impressive advances in shrinking device dimensions seen by electronics manufacturers, and to extend this model to a wide array of industries that could benefit from nanotechnology. At CCNI, researchers are pushing the edge in modeling and simulation in bio and nano technology, interoperable technologies in petascale simulations, multiscale modeling of material systems, and nano-electronics modeling and simulation involving material structure and behavior, complex flows, computational biology, and biomechanical systems.

How it works: The CCNI industry program enables multiple levels of membership to engage a broad range of industries, organizations and institutions, from start-ups to major corporations, government laboratories, and universities. Benefits include access to supercomputing facilities, collaboration with Rensselaer researchers and graduate students, a seat on the CCNI Advisory Board, confidentiality of proprietary data and intellectual property protection, and licensing arrangements for use of software. Supercomputing facilities include the IBM Blue Gene/L System, IBM Blade Server Cluster, IBM General Parallel File System, and IBM Power SMP Servers.

This active industry collaboration with over 120 ongoing research projects has produced over 150 scientific papers in last two years. The CCNI is housed in the Rensselaer technology park with 4300 sq. ft. of machine Room, business offices, systems and operations support, and domain specific scientific support. Users can also leverage the resources at the Rensselaer main campus Research Computing Center with a 2000 sq. ft. machine room.

Why IBM: In addition to providing most of the HPC systems, IBM has also invested in providing marketing and business development skills to help with evangelization and outreach. To further IBM's own R&D efforts in the area of electronic design automation, nanotechnology, and materials science, IBM has deep technical collaborations with CCNI.

IBM and CCNI researchers are using high performance computing in Electronic Design Automation (EDA) and lithography to dramatically reduce time-to-solution in chip design using highly compute intensive techniques such as Static Timing Analysis (STA) and its variations and Optical Proximity Correction (OPC). For instance, a virtual patterning flow problem that took 8 months on a serial simulator now runs in under 12 hours on the IBM Blue Gene using parallel algorithms. Another ongoing joint project IBM/CCNI is investigation of quantum-scale challenges such as discrete energy levels and modified e-ph coupling in taking carbon based metal wires such as carbon nanotube (CNT) and graphene nano ribbon (GNR) from laboratory to production as next generation interconnects in chip design. The project, involving over 75 researchers across five universities and several industry collaborators, combines large scale quantum simulations with compact and circuit level modeling and experiments.

SciNet –Canada's most powerful supercomputer for global research collaboration. SciNet²⁷, a consortium of the Canadian university and government bodies, has a mandate to provide high-performance computing resources to their own academic researchers as well as other users across the country and international collaborations. Together with Compute Canada and IBM²⁸, SciNet has collaborated to create Canada's most powerful supercomputer and one of the most powerful and energy-efficient supercomputers in the world. The facility enhances SciNet's competitive position in globally important research projects.

How it works: With peak performance of more than 300 trillion calculations per second, the IBM System x iDataPlex system ranked 22nd in the Nov 2009 TOP500 list of world's most powerful supercomputers. It uses a total of 30,240 Intel processor 5500 series 2.53 GHz processor cores and is entirely water cooled. Adaptive Computing Moab and xCAT are used to administer, manage, and schedule a wide range of HPC applications and workloads on this system. The IBM Supercomputer is being used for ground-breaking research in aerospace, astrophysics, bioinformatics, chemical physics, climate change prediction, medical imaging and the global ATLAS project, which is investigating the forces that govern the universe.

Why IBM: The IBM System x iDataPlex server is specifically designed for data centers that require high performance, yet are constrained by floor space, power and cooling infrastructure. This system provides up to five times the compute density versus competitive offerings and a unique water cooled technology -- IBM's Rear Door Heat Exchanger -- extracts more heat than the systems actually generate. This new iDataPlex system adds to SciNet's existing supercomputing capability, which includes an IBM water cooled Power 575 supercomputer with 3,328 POWER6 cores with peak performance of more than 60 trillion calculations per second. This, combined with additional energy efficiency technologies, including dynamic provisioning software in xCAT that automatically turns off processors not currently in use, and the state-of-the-art data center design saves enough energy to power more than 700 homes yearly.

North Carolina State University - 24x7 access to general and research cloud computing for students, researchers, and faculty. IBM and North Carolina State University have collaborated to establish the Virtual Computing Lab (VCL)²⁹, a cloud computing-based technology, to provide students around the state and the University of North Carolina system 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 BladeCenter compute resources are dynamically loaded on demand with a choice of operating system images and predefined application sets in either bare metal or hypervisor 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 related license agreement, it was limited to university owned computers and was not allowed to be installed on the student's personal computer.

Why IBM: IBM has supported and helped fund VCL-related projects and initiatives through numerous hardware grants, software grants, faculty funded research grants, and one time donations. 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, etc.

Centering Research Computing Drives Collaboration and Innovation

Higher education institutions can optimize investments in research computing by centering their research computing initiatives for researchers through a wide range best practices to increase funding, improve advocacy, grow the user base, standardize the computing environment, improve reliability, and reduce costs and complexity. This drives increased collaboration, innovation, and scientific discovery. It also helps recruit outstanding faculty and students, improves the institution's brand equity, and enhances the long-term economic and competitive positions of all stakeholders. Many universities are also implementing cloud computing to realize these best practices. IBM academic initiatives and leading-edge high performance computing solutions are helping many research institutions implement these best practices.

¹ Judith A. Pirani, Donald Z. Spicer, Ronald Yanosky, "Supporting Research Computing Through Collaboration at Princeton University", ECAR Case Study 2, 2006, Case Study from the EDUCAUSE Center for Applied Research.

²National Science Foundation, "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure", January, 2003.

³ Sandra Braman, "What Researchers Want (from IT)", ECAR Symposium, Carefree, Arizona, December, 8, 2005.

⁴ Developing a Coherent Cyberinfrastructure from Local Campus to National Facilities: Challenges and Strategies - A Workshop Report and Recommendations , EDUCAUSE Campus Cyberinfrastructure Working Group and Coalition for Academic Scientific Computation, February 2009.

⁵ Francine Berman, "Beyond Branscomb", <http://www.cs.utk.edu/~dongarra/cegsc2006/Slides/talk06%20Fran%20Berman.ppt>

⁶ National Science Foundation, "Investing in America's Future", Draft National Science Foundation Strategic Plan, www.nsf.gov/about/performance/nsfplandraft.pdf, 2006.

⁷ Malcolm Atkinson, "Progress with e-Science", NCeSS International Conference, Manchester, June, 2006.

⁸ Harvey B. Newman, "Global Networks for High Energy Physics", California Institute of Technology, ICFA Standing Committee on Interregional Connectivity (SCIC), Daegu, September, 30, 2005.

⁹ Patrick Thibodeau, "IT to Help Astronomical Armageddon", Computerworld Storage, September, 6, 2004.

¹⁰ Microsoft Corporation Press Release, "Working Together: Researchers Unite Web Services and Grid Computing to Enhance Scientific Study", <http://www.microsoft.com/presspass/features/2003/oct03/10-06GridComputing.mspx>, October 6, 2003.

¹¹ Office of Science, U. S. Department of Energy, "A Science-Based Case for Large-Scale Simulation – Volume 1", July, 30, 2003.

¹² Michael A. Robbie, "Re-centering the Research Computing Enterprise", Viewpoints, EDUCAUSE Review, May/June 2006.

¹³ Dempsey, Jed, Robert E. Dvorak, Endre Holen, David Mark, and William F. Meehan III, "A Hard and Soft Look at IT Investments." [McKinsey Quarterly](http://www.mckinseyquarterly.com/126-137) 1: 126-137, 1998.

¹⁴ Larry Goldstein, "Making the Case for TCO", EDUCAUSE

¹⁵ Bradley Wheeler and Thomas Hacker, "Centralize Research Computing to Drive Innovation...Really", <http://net.educase.edu/ir/library/powerpoint/EDU05163.pps>

¹⁶ Srini Chari, "Engineered for a Difference in High Performance Computing (HPC): Why IBM Power Systems Lead in Performance, Reliability, Availability, and Serviceability, December, 2009, http://www-03.ibm.com/systems/resources/systems_deepcomputing_IBMPower-HPC-RAS_Final-1.pdf

¹⁷ Peter Kogge, et. al., "Exascale Study: Technology Challenges in Achieving Exascale Computing, September, 2008, http://www.darpa.mil/ipto/personnel/docs/ExaScale_Study_Initial.pdf

¹⁸ "The data deluge", the Economist, February, 2010.

¹⁹ "Enabling and Sustaining Campus-to-Campus Cyberinfrastructure, January, 2010, http://www.sura.org/programs/docs/CI_White_Paper_Final.pdf

²⁰ Wikipedia

²¹ Tim Jones, "Cloud Computing with Linux", <http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux/l-cloud-computing/l-cloud-computing-pdf.pdf>

²² Jeffrey Rayport and Andrew Hayward, "Envisioning the Cloud: The Next Computing Paradigm", March, 2009.

²³ Ashar Baig, "A Cloud Guide for HPC", May 2009.

²⁴The IBM Deep Computing Portfolio, <http://www-03.ibm.com/systems/deepcomputing/index.html>

²⁵ <https://www.ibm.com/developerworks/university/research/index.html>

²⁶ Srini Chari, "A Total Cost of Ownership Study (TCO) Comparing the IBM Blue Gene/P with Other Cluster Systems for High Performance Computing", November 2008, http://www-03.ibm.com/systems/resources/tcopaper_finalfinal_2008.pdf

²⁷ www.scinet.utoronto.ca

²⁸ <http://www-03.ibm.com/press/us/en/pressrelease/27755.wss>

²⁹ <http://vcl.ncsu.edu/>