Overview of the Cornell University Center for Advanced Computing Sustainable Funding Model

A Whitepaper Submitted to the "NSF Workshop on Sustainable Funding and Business Models for High Performance Computing Centers," May 3-5, 2010, Cornell University

David Lifka[1], Resa Alvord, Susan Mehringer, Paul Redfern
Cornell Center for Advanced Computing

Introduction

The Cornell Center for Advanced Computing (CAC) is a core facility at Cornell University providing advanced computational and data analysis services to Cornell faculty, staff, and students as well as, under select NSF, USDA, and DOD awards, to the national research community. CAC is a reinvention of what was formerly known as the Cornell Theory Center (CTC). The mission of CAC is to enable the success of the Cornell research community, its collaborators, and supporters whose work demands advanced computing solutions.

CTC began in 1985 as one of the five original NSF-funded supercomputing centers and as such had a primary mission of serving the national research community. Center director and Nobel Laureate Kenneth Wilson inspired the scientific community at that time with the notion that computation was equal with theory and experiment in scientific inquiry. Thousands of researchers from across the nation benefited from the Cornell NSF center. Cornell staff developed parallel tools, delivered high quality training, enabled scientific discovery, and advanced the science of high performance computing by deploying new HPC computing architectures such as the first IBM SP supercomputer. While some faculty preferred that the Cornell center was focused exclusively on Cornell research needs, other faculty members managed to take maximum advantage of the fact that a national resource was located on campus.

After Cornell's mission as a national center ended in 1997, CTC had to reinvent itself. The center decided to focus on the emergence of high volume, industry standard computing components and developed deep corporate partnerships. This resulted in several industry firsts, including the deployment of the nation's first Dell supercomputer. This and other partnerships brought in significant operational funding, but as CTC staff focused on vendor specific needs, some Cornell researchers felt their research needs were not being adequately addressed; instead, they preferred a multiplicity of platforms and staff focused exclusively on their needs. In time, as high volume, industry standard computing became more mainstream, the number of opportunities for emerging technology partnerships in this area began to decrease. This led to another drought in funding for CTC. By mid-decade it became clear that a new, more sustainable funding model was needed for research computing at Cornell.

In 2007 the Office of the Provost formed a team of vice-provosts, deans, senior faculty members, and select center management to develop a new vision and operational model for the center. The consensus was to focus the center on the needs of the Cornell research community and their national collaborators by building a sustainable cost recovery model that would ensure the center's financial health and minimize cyclical funding extremes. The center was renamed the Cornell Center for Advanced Computing (CAC) to signify the beginning of this new mission and recovery model. Today, CAC is approaching its second year of operation and, to date, has demonstrated steady growth in terms of the number of researchers served, computational resources supported, and federal grants awarded.

[1]Contact person: Cornell Center for Advanced Computing, Rhodes Hall, Ithaca, NY 14853-3801. Email: lifka@cac.cornell.edu

The purpose of this paper is to provide an overview of the center's new mode of operation, including motivating factors and lessons learned. It is intended to support discussions at the “National Science Foundation Workshop on Sustainable Funding and Business Models for High Performance Computing Centers” which will be held at Cornell University on May 3-5, 2010.

Funding Models

The goal of any sustainable funding model is to come up with a viable plan for an organization to provide services required by its “customers.” The price of those services should be one that they are not only willing to pay, but is also more attractive than any other solution available, including doing it themselves. CAC’s sustainable funding or “recovery model” was designed to do just that. One of the benefits of a successful recovery model is moving from a situation where large amounts of cyclical funding from one or two sources dictate the mission of a center and create a crisis when the funding ends (Figure A) to a model where many sources of more modest funding provide steady-state funding level and the loss of any portion of that funding does not create an organizational crisis (Figure B).

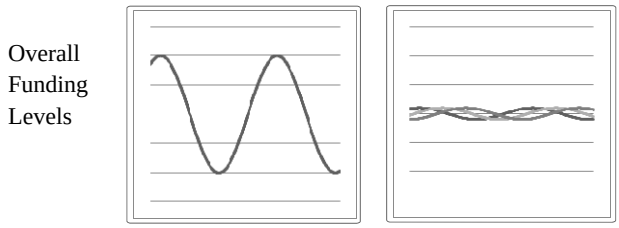


Figure A Figure B

The primary methodology used to design CAC’s sustainable funding model was to identify the advanced computing services that researchers at Cornell wanted and then to determine how to offer those services in a way that had clear economies of scale. A major challenge faced by CAC was that for 20+ years Cornell’s faculty was accustomed to getting free support and computing from the center.

Initially, it was surprising that the University deans wanted the center to charge for services as opposed to establishing a full University subsidy based on, for example, a tax on the colleges that use the center. The deans preferred a more transparent, fee-based system that would provide them with detailed University-based accounting so that they could see exactly what they were getting for their money. This model allows deans and department chairs to invest in specific faculty and research projects as they see fit and helps to prevent faculty from over scoping resource requirements with the "I need the biggest computer in the world" argument. The sustainable funding model also provides a built-in system of checks and balances that ensures that the center only provides what the market will bear, limiting financial exposure for both the center and the University. In addition, because there is a cost for services, faculty are more motivated to write grant proposals to cover those costs. This keeps the faculty motivated and the University more active in pursuing external grants for computational resources and staff support.

Goals of the Sustainable Funding Model

Clearly, introducing a new model which charges researchers for services is going to cause some degree of angst. In anticipation of this, goals for the sustainable funding model were shared broadly with the faculty:

1. **Value:** There must be clear benefits to the faculty in working with the center as compared to creating their own HPC and data storage resources spread across the campus and staffed by graduate student labor.
2. **Transparency:** In order to make a fair and accurate evaluation of the value of the services being offered, faculty and researchers should have access to how costs for services are derived so that they can compare them to alternative solutions. Both the cost of the service (such as staff, resources, facilities, etc.) and the details of the service being provided shall be clearly conveyed and available for review at any time.
3. **Fairness:** All faculty and researchers must be treated equally. Cutting special deals or offering special services to particular groups will not be tolerated. All services must have the same price and quality standards for everyone. Any Provost subsidies received by the center shall be allocated to all faculty and researchers in an equitable and transparent manner.
4. **Economies of Scale:** Whenever possible, service offerings will leverage economies of scale in order to reduce the overall costs to the University.
5. **Cost Recovery:** The ultimate goal of any recovery model is self-sustainability. Customer focus and accountability are essential. As an academic core facility, however, a certain level of activities are provided at no fee, including introductory consulting support, strategic consulting on campus initiatives, collaborative proposal writing, etc.

Strategic Inventory and Cost Analysis of Services

Once these goals were established, a strategic inventory and detailed cost analysis was conducted for each service the center had to offer. The feasibility of financial recovery for each service had to be well understood, including the number of current customers, the belief that these customers would continue to use these services if they had to pay for them, and the potential for these services to serve a much larger user community over time. We also worked closely with the Division of Financial Affairs to ensure that we would be able to bill for these services in an efficient and timely manner.

Since its inception the center has offered five primary services (in order of demand):

1. Computational consulting services (programming, porting/tuning, portal development, database development, education, and training)
2. High performance computing resources (computers, software, and storage)
3. Support for dedicated or private research computing resources
4. Visualization
5. Outreach

It was determined that only services 1-3 had enough research users that would be willing to support the required costs (staff, hardware, software, facilities etc.) in an ongoing and, thus, sustainable fashion. The other services, while clearly important to some faculty, did not offer the potential for cost recovery, economies of scale, or provide unique capabilities that would be a strategic differentiator to the greater research community. A survey of faculty needs revealed that most researchers were visualizing their own results. Based on that, it was decided that our computational consulting staff would continue to test new visualization technologies, but not offer visualization consulting as a hands-on, fee-based service. Outreach activities were for the most part a vestige from previous contract awards that had since expired. It was decided that future outreach programs would be designed and staffed to meet the specific criteria and community needs of new contract awards. Currently, CAC operates a K-12 math and science gateway, supports outreach workshops such as “Expanding Your Horizons” which is designed to motivate young women to pursue careers in math and science, and develops education and outreach programs to meet the requirements of current NSF projects.

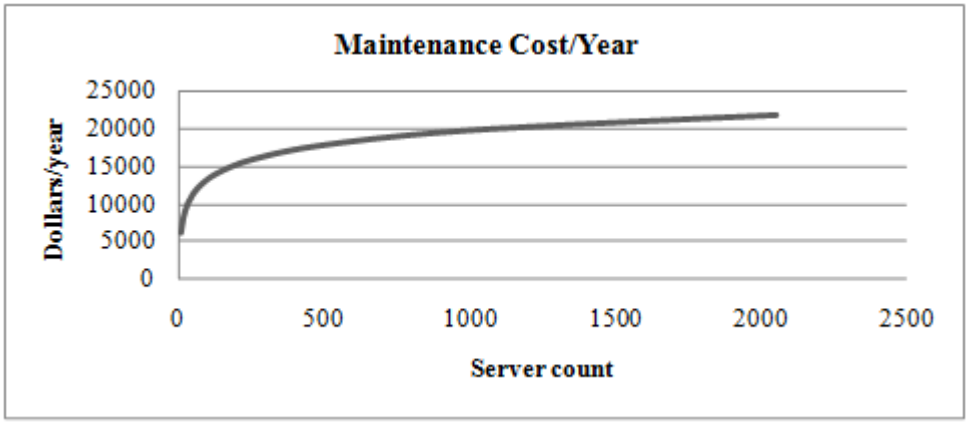
After identifying the three primary services CAC would provide as its initial service offering, a complete cost analysis for providing each service was performed. Rates for each service based on costs and anticipated level of recovery were formulated. The University Provost provides the center with an annual subsidy that is approximately 1/2 of its total budget and the goal is to shrink this subsidy to 1/3 of the budget over time. In order to ensure this subsidy is shared equally among all Cornell faculty and researchers, it is applied across all of our service rates. The result of this subsidy allocation is that the Cornell community procures professional services at rates that are competitive with graduate student labor. This is extremely compelling to faculty who may otherwise look to build their own group or departmental resource. The incentive to use a centralized core-facility has significant cost-avoidance benefits for the University. The Provost's subsidy ends up saving the University money by discouraging the creation of distributed campus resources that create an additional financial burden for each department that would have to provide their own data center space, power, cooling, and staff to support their systems. The services that faculty receive are also more reliable and sustainable than those provided by graduate students who are not available year round and have frequent turnover.

Initial system installations, configuration and testing are billed hourly at the CAC hourly consulting rate as are complete system reinstallations. However, it was clear that a systems maintenance rate that increased linearly with the number of nodes in an HPC parallel system, much like many administrative IT units would charge, would not be appropriate. Given current systems administration tools that allow a systems administrator to "push" system images and software packages to sets of identical resources, the amount of staff time necessary to manage a large HPC cluster was not significantly more than managing a small HPC cluster. In order to recover the right amount of effort, we devised the following formula which increases the rate for cluster maintenance based on the number of nodes in a way that was considered fair to all.

N = Number of nodes: includes compute node(s), head node(s) and dedicated file server(s)
S = Subsidized hourly staff rate
D = Working hours in a day
Y = 12 months in a year

Annual cost = (0.085 + 0.6 * log (N)) * (S * D) * Y

Given an hourly staff rate of S = $100 (example hourly rate, not actual subsidized rate), the cost recovery curve for private cluster maintenance is as shown in the following graph:



Cost Recovery Curve for Private Cluster Maintenance

Other Funding Sources

In addition to providing these services to campus, CAC also competes for federal grants as a lead organization, as a partner to other centers, and as a partner on proposals led by Cornell faculty. There are three important practices we have adopted when it comes to competing for grants:

1. Compete only for those grants that will have a direct or indirect benefit to Cornell. As a University core-facility, pursuing grants that do not align with the priorities of Cornell and its faculty are a poor use of the center's finite resources.
2. Find a niche where the center can provide world class expertise to the national community and form partnerships with other institutions, maximizing the likelihood of success. While this takes time and effort, it is an essential part of developing and delivering a strong value proposition.
3. Help Cornell researchers be successful in their grant proposals. This is the best way to generate new business on campus. Word-of-mouth amongst faculty and researchers is the best way to promote the center.

Corporate partnerships are another important revenue stream, but they require careful consideration. Ensuring that a partnership leads to mutual benefits is essential. Understand your expertise and brand value and be sure to assess the strategic, technical, and financial benefits when considering a corporate relationship. Be careful! A gift may actually be "the gift that keeps on taking." Hardware may be "free," but burdened by excessive annual maintenance fees. Software may be "free," but take more staff time than it's worth because it's half-baked or a low priority for the ISV. "Gives" and "gets" should be well defined and include metrics that are clearly understood by both parties.

Flexibility: Adapting Services to Meet Faculty Needs

The services researchers used in the past for free were not necessarily good indicators of the types of services desired by researchers going forward under a charge back model. While some services were popular, others proved non-starters. For example, consulting services are a huge success. Faculty appreciate being able to get computational, programming, and database support from a staff consultant (anywhere from 25% to 100% of their time) for weeks or months at a time. This enables them to get quick access to a highly skilled professional to meet a grant or project objective without having to hire a full-time person and without having to worry about how to fund that person long term. Within several months of implementing the cost recovery model, more than half of our staff were devoting 50 to 100% of their time each week to funded projects.

Offering "pay as you go" computing services was not received as well, even with aggressive pricing based on wall clock node hours. The center's general purpose cluster had 128 nodes, each with 8 cores, 16GB RAM, 70GB local disk space, and a layer 3 non-blocking gigabit Ethernet interconnect. In the first 6 months, hardly any compute time was sold. It turned out that "pay-as-you-go computing" was a concern to our faculty even though our online accounting system allowed them to put limits on resource utilization by any person in their project account. Faculty expressed, instead, a desire for "fixed-rate" computing services. Given this input, we established a lease program where faculty may pay for a set of nodes that is dedicated solely to their group. These are one year leases that are billed monthly through internal Cornell accounts. The cost of a lease is based on the original cost of hardware spread over 3 years plus the cost of staff time required to maintain the number of nodes leased. Only 80% of the general purpose cluster was made available for lease. The remaining 20% was made available to pay-as-you-go customers at the established node-hour rate and, as a bonus, to lease customers at no cost. Within 3 months of establishing the lease option over 75% of our nodes were leased.

Another lesson learned was in the area of disk storage services. Our rates for disk storage were not competitive compared with what research groups could purchase for themselves. We talked with faculty researchers to better understand their requirements and their target price point. We learned that if the storage offering was not $1,000 per terabyte per year or less with better reliability than the USB drives that they had in their labs, they were not interested. The only way to provide that level of price/functionality was to look at enterprise solutions that would drive the price per terabyte below $1000/year and still offer great reliability. The challenge was getting an entry level enterprise solution that we could afford. DataDirect Networks (DDN) offered us 50TB with RAID 6 and unique on-the-fly read and write error correction at the target price of $1000/terabyte/year including hardware maintenance. A key feature of the DDN unit was that another 1150 drives could be added to the initial configuration of any size or type (SAS, SATA, SSD etc.) as needed. Disks are the lowest cost component of the solution, so as disks are added, the price/terabyte goes down for all customers. When the faculty was given this information, they were excited; however, a commitment from them to pay for at least 1/2 of the initial 50 terabytes was required to minimize the financial risk to the center. The faculty agreed, the solution was purchased, and this service has been extremely popular. In under a year, customers have elected to pay for over 200 terabytes of storage. In addition, an astronomy group purchased their own DDN unit with 200 terabytes storage and pays the center to house and maintain it for them.

A willingness to invent and reinvent services based on faculty needs and emerging technologies is essential for a center operating under a cost recovery model. Flexibility and adaptability are required to ensure long term relevance and center survival.

Organizational Model & Staffing

The director of CAC reports to the Senior Vice Provost for Research. Unlike some institutions, research computing is not part of administrative IT and does not report to the University CIO. Where it makes sense, CAC leverages economies of scale provided by central IT rather than inventing their own solutions. Tape backup is a good example. CAC also shares data center space with administrative IT. As a research function, CAC is not required to recover for power and cooling. As the center supports more and more computational resources, adequate data center space is becoming an issue.

CAC has a Faculty Oversight Committee that assists the Senior Vice Provost for Research and the center director by providing strategic advice. CAC currently has 14 core staff members. All staff, including the director and the assistant directors for systems, consulting, and strategic partnerships bill their time through CAC faculty services, corporate partnerships, and grants. The core staff will grow or shrink based on cost recovery for the services that they provide. When large grants or corporate projects are secured, new hires will be term rather than permanent appointments. If a term appointee wishes to stay with the center beyond the term of their grant or project funding, they will need to work with center management to identify and successfully secure follow-on or new funding sources. This is an important concept for succession planning, discussed later in this paper.

Vendor Relationships

A positive relationship with a variety of high performance computing industry vendors is essential. Vendors tend to be attracted to established centers that can demonstrate a value proposition for their products and serve as a customer reference site for new strategic technologies. Often the most successful relationships are based on identifying technical challenges that the center has in common with the vendor’s current or desired customer base. Jointly defining and developing prototypes is an effective way to define and develop new technology features that may be used as reference product implementations. Analyzing and understanding a vendor’s technology roadmap, implementation roadblocks, and market objectives is important in order to identify mutual opportunities and to propose unique solutions.

Building and maintaining industry relationships is a key component in developing competitive infrastructure grant proposals. Trust must be established in order to share architectural plans that are competitively sensitive and under NDA. It takes time and effort to develop these relationships and, once they are established, a center staff that knows how to deal with the pressure of meeting deliverables or research milestones in a timely manner. In addition to staff technical expertise, good project management skills and communication skills are required.

Mission Alignment and Management

The mission of the center must reflect the overall mission of the university. At Cornell, enabling the success of the Cornell research community is a core value shared all the way to the top of the University's administration. It is important to ensure through timely meetings and reports that the Senior Vice Provost for Research is in agreement with the center's strategic decisions and the execution of its recovery model. The center director must not only be an astute technical manager, but also a business manager whose primary focus is the success of the center rather than his or her personal research agenda. Faculty directors may be passionate about computational science, but may not have the time, patience, or desire to manage the business aspects of a cost-recovery center.

As mentioned earlier, hiring term appointments for large grants or corporate projects is an important staffing strategy. It enables the center's core staff to keep their focus on their main customer, the Cornell research community, while providing a way to hire staff with the specialized skills needed for short term projects. There are two advantages to this approach: (1) the center has a constant influx of new blood with fresh ideas that keeps the center alive; (2) center management does not have to worry about retaining and funding these positions indefinitely. When a term appointment expires, the most talented appointees will have the highest probability of finding a new project to work on; other personnel will move on rather than become a long term financial burden to the center and the university.

Metrics of Success

Operating with a recovery model means that the metrics of success are: (a) obvious; (b) easy to track; and, (c) enable realistic growth projections. If the center continually provides services that are in demand, it will recover financially and have a satisfied user community. If the center fails to stay relevant, it will shrink in size or even disappear. While this may sound harsh, it is exactly the model under which businesses in our country operate. Learning to operate with business metrics has helped the center to better understand its vendors and what motivates them and to provide center transparency and accountability. Besides actual dollars and cents, other metrics of success to be considered are the number of faculty and researchers supported, the number of grants that faculty partner with the center on, and the number of grants and industry gifts/funding that the center secures on its own.

Today, CAC supports over 50 faculty-led research groups on the Ithaca campus and at the Weill Cornell Medical College in New York City. Customers pay for over 200 terabytes of disks storage and lease over 70% of our general computing resource. CAC also maintains over 15 private research clusters ranging from several nodes to 492 nodes. The NSF-funded MATLAB on the TeraGrid grant is an example of a recent federal award won by CAC serving as the Principal Investigator (PI). Recent faculty grant awards that CAC participated in as a Co-PI and/or service provider include the NSF Large Casual Databases award (Co-PI), the Defense University Research Instrumentation Program award (Co-PI) and the NSF Social Science Gateway to TeraGrid award (service provider). Prior to instituting the center's recovery-model, the center only supported ~15 research groups on the Ithaca campus and ran only 2 small private clusters.

Industry Relationships

Industry interest in the center is another indication that the services the center provides and the technical solutions the center uses to provide those services are relevant beyond the academic research enterprise. Industry values our staff expertise and the ability to test solutions using our resources rather than purchasing their own. Sometimes there is a desire to enter into a joint agreement to do development work for an industry partner. In this case, clearly intellectual property is something universities work to protect. CAC tends to focus on industry relationships that do not require the legal overhead of intellectual property transfer unless the funding is significant or the project is of such strategic importance or industry impact that it clearly makes sense. CAC has an established corporate membership program which allows industry partners to enter into technical exchange with the center staff or use center resources without the promise of specific deliverables or the exchange of intellectual property.

Identifying Emerging Collaboration and Outsourcing Opportunities

The center is currently investigating whether some services could be delivered through collaborations with other organizations or through outsourcing. Examples are sharing a joint data center space with local or regional universities and/or government agencies and using cloud computing to augment the services currently offered to the Cornell research community. Ideas such as sharing a machine room or outsourcing computing will be carefully analyzed to see if the cost savings are real and, if so, how significant they are. Obviously, performance, security, and ease of use are important metrics to consider in addition to cost savings.

Conclusion

The Cornell Center for Advanced Computing recovery model will have been in effect for almost two years at the time of this workshop. Changing the culture at Cornell and the center during this period has been both challenging and rewarding. We are proud of our steady growth and the new NSF, DOD, USDA, and corporate awards that we have secured. We would also like to express our gratitude to the Cornell administration and university researchers who have supported the center and helped us make it a valuable resource for everyone associated with it.