Report of the

High Performance Computing Task Force

May, 2015

RCA Members	
Joseph Comfort, Chair	Physics
Christy Lespron	Nutrition & Health Promotion
Hugh Mason	Biodesign (T)
Konstantinos Tsakalis	Elec., Computer & Energy Engin.
USFC Mmembers	
Yongming Liu	Matter, Transport, & Energy Engin.
Ann MacEachron	Social Work
Thomas Schildgen	GIT Program
David Wells	Letters & Sciences
Library	
Edward Oetting	Library
Representatives	
Charles Kazilek	Provost Office, VP for Technology
Frank Timmes	A2C2, Director

Executive Summary

In 2005, ASU initiated a University-wide supercomputing facility with a goal of bringing High Performance Computering (HPC) capabilities at minimal cost (including free) across the academic disciplines. It would join the growing number of other universities that were building HPC services which are needed as a foundation for research, and to be more competitive and successful in obtaining research grants. ASU hoped to emerge as a leader in HPC activities.

The initial growth of the HPC facility was strong. Several computer clusters across campus were moved to the controlled environment of the HPC center. Users could purchase cores and receive credit in units of CPU-hours (cpuh). They could also pay for use on the shared Saguaro cluster from accounts for the time used. Faculty and their students who did not have financial resources could obtain up to 10,000 cpuh/year through proposals, subject to review. The goal was to make resources available for innovative advanced computing projects broadly across ASU. The resources reached a peak of ∼6,000 cores in 2009.

The economic downturn of 2008 and later hit ASU hard, and the HPC lost ground. The budget decreased to less than $1 M/year, well below the more typical level of $2-3 M/year for facilities of comparable size at other universities. Some large users were able to buy their own supercomputers for full-time use. Others found other resources. The hardware aged, replacements were rare, and the number of available cores shrank to 1/3 to 1/4 of the peak. The total number of CPU hours used fell from ∼2,000,000 in the Spring 2012 to ∼500,000 in the Fall 2014. The staff size of 11 dwindled to the current level of 2, which also led to the much reduced ability to promote the use of the HPC. There are anecdotal accounts of faculty candidates declining interest in ASU due to its poor computing resources, and possible faculty who may leave for similar reasons. The A2C2 facility, as it came to be called, was in a downward spiral.

By the Summer 2014, it became clear that the current funding model was not sustainable. It was replaced with a new Community Cluster program in which users were required to purchase CPU cores, at $1000/core, for every core they would need simultaneously in their jobs. The University has initially provided an equal match for most of the first group of new cores. Users would have exclusive access to their cores for 5 years, and then would have to buy new cores as upgrades to the A2C2 facility. The Saguaro cluster with the pay-per-use option was to be terminated by the end of the year, but was restored after the passage of Senate motion 2015-44 to study the issues [US15]. However, it will be terminated in 3-4 years when the service contracts expire. There is no free use.

In December, the University Senate formed the HPC Task Force to study the use and funding issues. The Task Force obtained information on HPC use for current projects, new projects that are under competitive review for external funding, and new and bold opportunities where ASU could obtain a leadership role in the national academic communities. Examples are discussed in the body of this report.

In reviewing and discussing the issues, the Task Force recognized some deep deficiencies in the Computer Cluster model. Briefly, they include

1. The very high up-front cost of purchasing cores that is precluding people from using the HPC.
2. The high cost that comes from the lack of flexibility in running jobs that use a variable number of cores.
3. The absence of any guarantee that, after 5 years, funding agencies will be willing or able to invest in a new set of hardware.

The pay-per-use option can only partially alleviate some of these issues, but is planned to be terminated in a few years. The Task Force projects that the Computer Cluster model will itself be unsustainable. A specific set of recommendations for change are given.

The use of HPCs is growing very rapidly in the academic world and also expanding into businesses and government. Its use is becoming more and more integrated into the culture, much like the expansion of personal computers from the 1980s onward. Although supercomputers have been most closely associated with the STEM academic areas, like the mainframes of the past, HPC applications are moving rapidly into many more fields such as the humanities, social sciences, the arts, and others. Large data bases ('big data') are being built rapidly, and the academic world needs to develop the tools for working with them and extracting meaning information. Students need to be educated in HPC techniques and bring their skill into the world of jobs. ASU is already becoming engaged in some of the non-STEM fields. By incorporating an additional focus on them, the University can become a widely recognized leader in the HPC world.

The Vision for ASU adopted by the Task Force can be summarized by a statement that was found in a University of Houston document [UH14], which itself was taken from a previous A2C2 strategic plan:

> *HPC and cyber-infrastructure must be seen by academic institutions as fundamental tools and resources that are to be as ubiquitous and available as networks, telephones, and basic utilities. To that end, HPC must be a centrally managed and centrally funded activity on university campuses.*

In short, the role of HPC at ASU should be seen as part of the university infrastructure on par with libraries, buildings, utilities, and network connections.

An exciting possibility for expanding our HPC resources is the development of ASU's 100-Gb network connections through to Arizona Sun Corridor network to the very-high speed Internet 2 (I2) network. There are good opportunities for access to Amazon's Cloud Computing service at reduced costs, on a network that also provides access to the growing number of publicly available large data bases. This opportunity may initially be best suited for the less demanding non-STEM fields. It has the advantages of considerable user flexibility of CPU resources and data storage, without having to replace the hardware periodically. On the other hand, some cost estimates indicate that it is not a favorable option for the large power users.

At this point, the Task Force believes that the growth and use of the A2C2 facility is very important for meeting the goals of our Vision. Included in this growth is the expansion of staff, some of whom will be proactive in seeking out opportunities for HPC applications in non-STEM areas and pursuing faculty to engage in it. The current Community Cluster program must be converted as soon as possible away from large up-front costs to multiple, shared clusters. Users can buy cores for these clusters in exchange for an allocation of CPU-hours and/or higher job priorities. A pay-per-use may be needed initially, at a minimal level. Free access for all faculty (and the students they sponsor) must be available in all cases.

In response to the evaluation of the weaknesses of the current HPC resources and the funding model, and in seeking to implement the Vision articulated in this report, the HPC Task Force makes the following recommendations. The context for each recommendation is given in the last section of the report.

> Recommendation 1.
> The University needs to move as rapidly as possible towards establishing HPC as part of its core infrastructure. The Vision is to include expansion into new academic areas such as the humanities, social sciences, and arts where the University can establish a special mark for itself.

> Recommendation 2.
> Studies leading to a possible implementation of an ASU Cloud Computing option must be actively pursued. In addition to the technical network issues, consideration is to be given to a structure for friendly and flexible user access. Network connections across campus must be enhanced. The Task Force recommends that the facility is to be at no cost to users.

> Recommendation 3.
> To be competitive in the academic world, the university-wide, shared HPC resources in A2C2 should be expanded as rapidly as possible towards 5,000–10,000 cores, comparable to the era of 2009 and many other universities.

> Recommendation 4.
> As is common at other universities, ASU should house clusters in A2C2 with a variety of processors as a shared facility. Users may buy cores for the shared clusters, and obtain an allocation of CPU-hours (possibly with scheduling priority)

to use as they see fit, including the use of more cores than purchased. A user will not own specific cores. The submission of a job from such a user shall not cause termination of other, incompleted jobs.

Recommendation 5.
A user or group that purchases a cluster from external funds may house it in a common University location to have utility services. The group may allow the use of the facility during idle times by other users, at Pay-per-Use rates. However, the submission of a job from the group shall not cause termination of other, incompleted jobs. Furthermore, if any portion of the cluster is subsidized by University funds (as in 2-for-1 deals or otherwise), the cluster shall be treated as a shared cluster as in Recommendation 4.

Recommendation 6.
The Pay-per-Use and Free options must be available for all users of the shared cluster facilities. The change in rates from $0.025 to $0.01 per cpuh after 2012 produced higher cost recovery, even as usage declined. The Task Force strongly recommends that the lower rate be used, and that in no case shall it be raised above $0.025/cpuh. A review of a proposal is an option that could be applied to the Free method.

Recommendation 7.
The funds for NGCC were provided by OKED, and the office must make a concerted effort to find a viable use for the cluster, either as part of the University-wide shared facilities or to a group for a specific purpose.

Recommendation 8.
To become comparable with the support provided at other universities, the operating budget for A2C2 must be increased to $2–3 million per year. The staff must be proactive in providing high-quality HPC services to users, with special attention towards expanding services into non-STEM academic areas. Cost recovery from core purchases and Pay-per-Use charges shall separately be used for HPC hardware including addition cores, core replacements, and data storage.

I. Introduction

Supercomputer capability at ASU was established in 2005 with the inauguration of the Fulton High Performance Computing (HPC) center, supported in part by a $500,000 Fulton grant. The initial director, Dan Stanzione, had great visions of growth across the entire university. The initial base of 400 CPU cores was expanded to ~5000 cores by late 2008 and to ~6000 cores in 2009. In 2009, Dan left to lead the Texas Advanced Computing Center, currently the largest university supercomputing center in the country.

Several computer clusters across campus were initially moved to the controlled environment of the HPC center. The platform for most users was the shared Saguaro cluster. Users could purchase cores and receive credit in units of CPU-hours (cpuh), and the allocation could be used for jobs that needed many more cores than those purchased. Users could also pay from accounts for time used. The initial rate of 2.5 cents/cpuh was later reduced to 1 cent/cpuh to stimulate broader use (and also brought in additional income). Very importantly, faculty and their students who did not have financial resources could obtain up to 10,000 cpuh/year through proposals, subject to review. The goal was to make resources available for innovative advanced computing projects broadly across ASU.

The economic downturn of 2008 and later hit ASU hard, and the HPC lost ground. It took a couple of years to find a new director, currently Frank Timmes of SESE. The facility was restructured and given the name A2C2. Yet, the budget remained constrained at less than $1 M/year, well below the more typical level of $2-3 M/year for facilities of comparable size at other universities. Some large users (*e.g.*, TGEN) were able to buy their own supercomputers for full-time use. Other users (*e.g.*, a cluster in engineering) saved significant costs from time and disk storage by housing their own facility. The hardware aged and replacements were very limited. A large number of cores, whose service contracts were long expired, were placed in salvage in the Summer 2014, reducing the remaining number of available cores to ~1700. The total number of CPU hours used fell from ~2,000,000 in the Spring 2012 to ~500,000 in the Fall 2014. The staff size of 11 dwindled to the current level of 2, which also led to the much reduced ability to promote the use of the HPC and to essential programs for educating new users.

By the Summer 2014, it became clear that the current funding model was not sustainable. It was replaced with a new Community Cluster program in which users were required to purchase CPU cores at $1000/core, for every core they would need simultaneously in their jobs. To alieviate some funding issues, the University has initially provided an equal match for most of the first group of computers. The users would have exclusive access to their cores for 5 years, and then would have to buy new cores as upgrades to the A2C2 facility. (Other users could use the cores during idle periods, but the jobs would be immediately terminated when the owner needed the cores again.) The shared Saguaro cluster with the pay-per-use option was to be terminated by the end of the year, but was restored after protests. However, it will be terminated in 3-4 years when the service contracts expire. There is no allocation of free time based on reviews of proposals. Also, there is no option to buy into a shared cluster and receive an unrestricted time allocation.

In December 2014, the University Senate adopted a motion "to study the role of supercomputing at ASU, possible needs for expansion, and options for funding scenarios," to be followed by a report to the Senate. A HPC Task Force was appointed. Because there ware also concerns at that time about inadequate University support for some research facilities used among a broad range of academic units for education, the University Academic Council asked the Task Force to study these issues as well.

II. Computers in Academia

Computers in academia have had a long history. The interplay between the research faculty and industry (*e.g.*, Digital Computer Corp.; Cray Computers, *etc.*) can be represented as: <technology development ⟷ new capabilities for research>. Processing speeds multiplied, memories expanded, new storage media were created, networks were designed, and user-friendly computer languages were developed. The World Wide Web was born when faculty needed to monitor and control their experiments at CERN from their home institutions.

Most universities had central computing facilities. In the 1980's, the main ASU computer for research was an Amdahl mainframe. Users were charged to run a job, the CPU time, the memory used, every computer card read, every printout page, tape mounts, *etc.*. It was soon recognized that improving ASU's prospects for research grants, attracting quality faculty, and strongly improving its status in the academic community would require broader computer access with reduced financial burden. A special task force recommended a program of 'distributed computing' leading to a personal computer (PC) in every office, along with phasing out the central computer charges.

Supercomputing in universities is now at a similar status. Over 84% of public universities have centralized supercomputing services. An extensive list of over 100 public and private U.S. universities is given in a 2014 report from the University of Houston [UH14]. [1] In addressing the necessity for HPC in Tier One institutions that are focused on producing frontier research, the report quotes a statement [UH14]:

> *HPC and cyber-infrastructure must be seen by academic institutions as fundamental tools and resources that are to be as ubiquitous and available as networks, telephones, and basic utilities. To that end, HPC must be a centrally managed and centrally funded activity on university campuses.*

It is pertinent to note that this statement was taken from a previous Strategic Plan for the ASU A2C2 facility. While it is not in the most recent strategic plan, we can take it as a guide for the growth and extension of HPC at ASU.

[1] The document *The Case for High Performance Computing at the University of Houston* provides a recent and very comprehensive study of HPCs for research along with business models, *etc.*

III. HPC Use at ASU

Supercomputing is necessary for highly complex computational problems that require large memory space and/or execution times, certainly exceeding anything that can be done on a single desktop computer (even one with multiple computer cores). As with previous computer generations, their primary applications are in the areas of science, technology, engineering, and mathematics (STEM). The computations may involve searches, comparisons, and analyses of very large data bases, modeling and simulations (often with Monte Carlo techniques) of physical processes, or sophisticated algorithms.

Such has been the case at ASU, although there are signs of new buds poking up through the surface. Between 2012–2014, the academic areas with the largest HPC use were SESE (26%) TGEN (25%), Engineering (19%), Math (10%), and Physics (9%) [AR14]. Between 2005–2011, the fraction of research funding from grants that incorporated HPC to total research funding increased from about 1/7 to 1/4 [HPC11]. Over the past 5 years, ASU grants with HPC support have brought in over $123 million to ASU. Such increases represent the rapidly growing trend to tackle highly sophisticated issues at the forefronts of academic research. It will continue to grow.

III.A. Some Current Examples

Robert Farmer, a Res. Assoc. in SESE and in collaboration with Prof. Frank Timmes, is using models of stellar evolution to study the 'death' of a class of stars, 7–10 times as massive as the Sun. Some can become white dwarfs, while others can have a supernova event and become a black hole. What determines the differences? The evolution of stars involves complex stages of nuclear physics from hydrogen burning, to helium burning, to carbon burning. The models track all stages and provide continuous descriptions of the temperature, density, and composition distributions throughout the entire star. The goals are to elucidate how the mechanisms contribute to the data from astronomical observations and to obtain fundamental understandings of the highly complex mysteries of the universe [QR_Wi15].

Oliver Beckstein, Asst. Prof. of Physics, and his group are using sophisticated simulations to study molecular mechanisms in biological systems. Of particular interest is the role of large, 3-dimensional, folded proteins within the membranes of living cells for moving nutrients, signal molecules, and toxic compounds into or out of the cells. The molecular dynamics calculations can include 1 million or more atoms and require hundreds of computer cores running in parallel. The studies of the weaknesses or breakdowns of such transport processes could eventually lead to methods for combating hypertension, heart disease, or epilepsy [QR_Wi14].

III.B. Education

A2C2 provides 100,000–200,000 cpu-hours per semester of resources and support for many courses in which HPC programming and applications are taught. Most of the courses are in the units of Computer Science, Mathematics, and Engineering. These courses provide the basic knowledge needed by students, undergraduate and graduate, for HPC work on ASU research projects. Even more importantly, they provide education and develop skills that our students can bring to the world of jobs, where HPC is expanding rapidly. In addition to direct education, A2C2 provides periodic seminars and other educational materials for faculty and students to learn how to access the facilities and run jobs productively.

An example is a course taught by Assoc. Prof. Violet Syrotiuk in Computer Science and Engineering. The students first learn about serial programming, then how to use tools to convert their programs to a parallel environment. Finally, they apply their skills to realistic research applications. A recent case was to use a very large data base of sea ice coverings, from meteorological satellites, to track changes in the fraction of the oceans that are covered by ice. The trends are disturbing, suggesting that the artic seas may be seasonally ice free by 2050-2100. [QR_Su13].

III.C. Near-Term Expansion

Research in STEM areas will continue to grow, bringing ever increasing pressure to increase ASU's HPC resources. With very tight budgets in the funding agencies, the climate for proposals has become increasingly competitive. The success rate for NSF funding is ∼20%. In addition, reviews and award decisions generally consider the resources provided by the institution, especially including computing resources, to ensure the success of the research. Even though NSF, DOE, and others are providing national HPC facilities, these may not provide sufficient edge compared to local, easily accessed resources.

The following sections highlight some under-utilized or new areas where proposals are being written or submitted. Estimates of the HPC resources needed are listed.

III.C.1. Materials and Bio-systems

Several proposals and projects are underway involving faculty from the College of Letters and Sciences.

- Xihong Peng and Meng Tao (Fulton Engineering) have submitted a $2 million proposal for 4 years to develop precision and reliable chemical vapor processes for depositing monolayers of particlar compounds on silicon wafers. Part of the work requires extensive Monte Carlo simulations and first principles. The work will lead to new and valuable knowledge of electronic properties of very thin layers. The simulations are expected to require 200,000 cpuh each year on an ASU HPC system. The NSF has approved a pre-proposal and invited a full proposal, which is pending.
- Xihong Peng and Robert Nemanich (Physics) are seeking to establish an ASU Materials Research Science and Engineering on Precision Interfaces program. The goal is to synergistically predict, fabricate, and characterize materials with precision interfaces, and provide solutions for advanced technology and energy applications. There is considerable overlap with the previous item. Again, the NSF has invited a proposal for competitive review. The HPC requirements are also for 200,000 cpuh/year. In both of

these cases, the budget can not support the purchase of cores. A rate of $0.01/cpuh reduces the cost by at least a factor of 4 over the period of the grants.

- Yun Kang specializes in gene regulatory and social interaction networks from insects to humans. A new project is being developed that requires very large data sets that can not be handled on a desktop PC. It is estimated that 60,000 cpuh/year will be needed.

III.C.2. Genome Sequencing and Computing in Biology

For several years, faculty from Biodesign, SOLS, SHESC, the Dept. of Biomedical Informatics in the College of Health Solutions, and some other units, have been engaged in a Next Generation DNA Sequencing project. It is a rapidly growing area of biology and represents a fundamental change in how biologists do research. Much of the activity has involved development of a large collection of plasmids at ASU, acquisition of genome sequencing apparatus, and analysis of genomes. It is possible to obtain the entire genomic sequence of an individual or species, to examine it for mutations, and to determine all of the genes expressed in a biospecimen. Applications of the information extend to biomedicine, health outcomes, bioengineering, informatics, and biosecurity.

The data collections are in the realm of 'big data' with tens to hundreds of Gigabytes. Analyzing the data to identify associations between genomic variants and chronic diseases such as cancer, diabetes, and obesity requires HPCs. Several faculty and students have been pursuing such studies for several years. There are plans to hire another two or more faculty for HPC work over the next few years.

III.C.3 Animation and 3D Modeling

The Graphic Information Technology program at the Polytechnic campus offers focus areas in animation and 3D modeling. Because of the large data sets being processed, the use of HPC to aggregate computing power is much more efficient than the use of typical desktop computers or workstations.

The production of 3D animated movies involves the rendering of motion picture frames, which requires a very large amount of data crunching. Each frame must be generated from images of 3D model through software applications. A rendering engine computes resolution, textures, lighting, and other details. The process can require 90–100% of the CPU resources and memory.

III.D. Breaking New Ground: Humanities

The academic world is much more than the STEM disciplines. As computer mainframes gave way to personal computers, opportunities were opened for other subject areas. A common theme in several of the humanities areas was textual analysis to find distinguishing patterns in writings that might convey additional insights and meanings (or even authenticity).

More and more books, documents, or fragments have been converted to digital form, and very large repositories created. While simulations are a hallmark of STEM computations, ‘big data’ analytics and image analysis are the primary approaches in the humanities. Some examples can be identified.

- Large collections of ancient glyphs can be searched to identify all the sites for a particular one, or to identify the historical era and origin of an artifact that is recently ‘discovered.’ By modifying the criteria of pattern recognition, the evolution of a glyph might be studied.
- Large collections of digitized images of 3D objects can be searched to look for items that have similar shapes, or might go together in some way.
- Very large collections of images from paintings can be studied to learn about features of local cultures (*e.g.*, farm life, how people are represented in pictures, *etc.*).
- Pattern recognition can be used to resolve questions such as “Who wrote this book?” or document, “Who painted this picture?” *etc.*
- Details of historical events can be traced through digitized collections of newspapers.
- Some sophisticated mathematical applications of network theory have been applied to the analysis of genres in literature.

A specific application of HPC in the humanities, already present at ASU, is the Large-Scale Video Analysis project of Michael Simeone and collaborators. Its aim is to establish a software workbench for video analysis, annotation, and visualization. Although videos can currently be scanned to identify specific objects of security, military, or commercial interest, tools do not exist to help us better understand our past and present culture and history. Instead of serially scanning videos one after another, cross-cuts can be done on the digitized data from multiple videos.

ASU is in a good position to move aggressively into HPC applications in the humanities. A new course (CDH 501 Digital Humanities: Critical Theory and Methods) is going through the Senate approval process. It is intended to be the first core course leading to a Digital Humanities Certificate. At least one faculty member (M. Simeone) is strongly engaged in HPC research. Other faculty are also engaged in digital technologies, recognizing that they are intricately woven into our lives. We can look forward to additional faculty with HPC expertise.

It should be noted that HPC can also have applications in the arts and social sciences. Indeed, statistical methods in the social sciences were a very common component of computer work in the days of the mainframes. Later, much of the work could be done on PCs. Yet, as the world becomes increasingly flooded with digitized data, huge data bases involving social patterns (in a very broad sense) are being generated. The opportunities seem almost unlimited. The challenge in each case will be to identify the academic themes and goals and to develop the analytical tools needed to achieve them.

Animation and 3D modeling was mentioned in Sec. III.C.3. Graphics are often an integral part of STEM HPC projects. These tools are available for the creative imagination of artists to explore new directions.

IV. External HPC Options

A possible option for HPC projects is the use of external resources, either to supplement campus resources or as a replacement for them. Some large facilities are provided and supported by the major federal funding agencies DOE, NSF, and NIH, and there is little or no cost to university researchers for using them. For projects that are less demanding, several commercial enterprises provide cloud computing services. There are also possibilities of developing regional networks.

IV.A. National Supercomputer Facilities

Table I shows a number of facilities at the national level that can provide hundreds to thousands of CPU cores (often in combination with GPU acceleration such as Titan). Allocation awards range from a few hundred-thousand CPU hours (for small XSEDE awards) upwards to 50 million core hours on the large leadership systems such as Titan or BlueWaters.

Table 1: National supercomputing facilities. Access is by proposal.

Facility	Resource	Sponsor	Problem Size	URL
XSEDE	various	NSF	medium, large	https://www.xsede.org
PSC Anton	PSC Anton	NSF/NIH	specialized (molecular dynamics only)	https://www.psc.edu/index.php/computing-resources/anton
INCITE, ALCC	Tital, Mira	DOE	very large (peta scale)	http://www.doeleadershipcomputing.org/guide-to-hpc
Petascale Computing Resource Allocation (PRAC)	BCSA, BlueWaters	NSF	very large (peta scale)	https://bluewaters.ncsa.illinois.edu http://www.nsf.gov/funding/pgm_summ.jsp?plms_id=503224

Access to these machines is in general through peer-reviewed proposals. Officially published proposal success rates are not readily found. However, at least for XSEDE proposals, it is known that capacity is always oversubscribed so that not all proposals can be accommodated. Even for the awarded proposals, the allocated CPU-hours are typically reduced by 10-30%. Some awards (such as PRAC, with about 10-14 awards per year) also provide up to $40,000 in budget for travel *etc.*

No prior funding is required to apply for any of these facilities, although peer review typically takes feasibility into account. Startup and Educational allocations have easy application procedures, but they are usually limited in time and/or other resources. For Research allocations on XSEDE, researchers are normally expected to have NSF funding. Their applications only undergo technical peer review, without assessing the merit and impact of the science, thus making awards more likely. A user must also wade through a plethora of options for computing clusters, short- and long-term disk space, and data transfer methods to the home institution.

IV.B. Individual Private and Public Cloud Computing

Another alternative for some computer tasks (especially for many independent smaller jobs) are cloud computing platforms. Private cloud computing services are paid services where resources can be rented (e.g. Amazon AWS EC2/HPC [Amazon], Microsoft Azure [Azure]. or Google Compute Engine [Google]. For smaller, infrequent computations that do not generate large amounts of data (<100 GB), this approach can be cost-effective compared to purchasing a cluster. However, for researchers with a continuous stream of calculations, it does not appear to be cost effective. Public cloud computing such as Boinc [Boinc] provides a platform to run calculations on the computers of volunteers. Access is usually based on review of proposals.

IV.C. Cloud Computing Option for ASU

An interesting option that is being explored is to have Amazon cloud computing services be provided at a discount to ASU by DLT Solutions. This option has several advantages including the availability of a variety of CPU resources, charging for only the cores used during program execution (which saves the cost of idle cores), and a regular upgrade of capabilities without the need to replace a whole cluster every 3–5 years. User support services would be provided through DLT Solutions, augmented by ASU staff.

The service would be provided through 100-Gb connections, now available to ASU on Arizona's Sun Corridor R&E network to the very-high-speed Internet2 (I2) network. Short- and long-term storage of data will be possible. Connections with large, publicly available 'big-data' data bases can be made from user programs at I2 network speeds. Of course, transfer of data between ASU and the cloud facility (at no charge on the Sun Corridor network) will be limited by the internal ASU network speeds, which are being upgraded.

Many details about this option are still being studied. Costs are difficult to estimate, but a comparison of a 'large' computational job is available. This job used 256 cores for 12.5 days, or 76,800 cpuh. The cost for the pay-by-user method ($0.01/cpuh) is $768; at A2C2 the prorated cost of an initial $256K investment in cores (at $1,000/core) is $1,750; the estimated cost on Amazon is ∼$3,000 (reserved) or ∼$5,000 (on-demand). The on-demand rate is charged only during execution, while the reserved rate requires a long-term commitment of the cores (1–3 years). The results are consistent with the statements in Sec. IV.B.

Initially, this Cloud Computing option may be focused towards the 'small job' and unfunded users who can not subscribe into A2C2 or run jobs irregularly, possibly with the University picking up most or all of the charges. It could help in an expansion of HPC across campus and at least partially resolve the costs of replacing the A2C2 computers periodically. A potential downside is that both DLT Solutions and Amazon are both commercial entities and are free to change their business plans at any time, leaving ASU without service.

V. A Vision for ASU HPC

Ever since the seers and schools of ancient cultures, through the medieval guilds, and through the American colleges for training ministers and leaders of society, the education of students has been the primary purpose of universities. In recent centuries, and especially starting in the 19^{th}, some faculty, mostly in the sciences, were able to made great advancements in knowledge about nature. They had no funding agencies. Yet, their feats have led directly to the wonders of our modern civilization.

Research in American universities began to flourish much more in the 20^{th} century. Donors often provided support. But the main funding agencies we know today, such as the NSF, DOE, and NIH, were not created until after WWII.

Universities saw great value in pursuing research. Projects supported by national funding agencies could be seen as serving national and community needs. The faculty who pursued them often had exceptional skills and brought stature to the university. Most importantly, they were at the leading edge of knowledge in their fields and could both enhance classroom instruction as well as open great opportunities for students to develop skills for use after graduation. In turn, the public would have increased pride in the university, and alumnae and alumni, along with other donors, would be inspired to contribute ever more to supporting the original and primary goal of student education.

Research costs money. As soon as a university decides to make research an integral part of its core mission, it commits to providing *its own resources* to support it. Research can not make a 'profit' directly. Grants from funding agencies are designed to provide cost recovery for actual expenses; at best, they are a zero-sum process. Even so, universities are usually expected to share substantially in the costs. In fact, they do by providing buildings, utilities, libraries, and auxiliary services, as well as faculty and staff for the research activities. Such support is, of course, under-written by tuition, endowments, and donations along with legislative funding for public universities.

The growth rate of HPC facilities, both at the national level and at universities, is very high. Certainly in terms of the capabilities they provide, the rate is larger than that of the PCs in the '80s and '90s. The growth is not at all limited to the STEM fields. In fact, Digital Humanities currently has the highest HPC growth rate in the NSF. HPC is growing in academia, businesses, and government agencies, and will continue to permeate more and more of society.

Our vision is one in which HPC at ASU becomes embedded across all of the academic fields. Based on current resources, we are behind the curve. But we are in a good position. We do not have to unmake the patterns and practices of other universities, but can develop our own leadership in the academic community.

As stated in Section II:

> *HPC and cyber-infrastructure must be seen by academic institutions as fundamental tools and resources that are to be as ubiquitous and available as networks, telephones, and basic utilities. To that end, HPC must be a centrally managed and centrally funded activity on university campuses.*

We need to think of HPC as being part of the university infrastructure on par with libraries, buildings, utilities, and network connections.

V.A. HPC Resources

The current CPU core resources are well below the peak of ∼6,000 in 2009, but growing. The 2014 Saguaro cluster has about 1700 cores of mixed type. The new 2015 community cluster Ocotillo has about 900 cores (including about 800 purchased under ASU's 2-for-1 deal). The 2016 cluster Agave is anticipated to be similar.

There is substantial disk storage available, with high-speed Infiniband connections to the clusters. There are also high-speed ethernet links out to the campus, although there have been some severe bottlenecks across campus.

A group representing the Life Sciences has provided a list of needs for their activities. Included are 500 cores with 16 MB of memory per core, and two 'fat' nodes with 32 cores and 1 TB of memory each. An additional 4 TB of disk storage is needed for user home directories, and 32 TB dedicated to host public data bases. Regular backup schedules are required.

These resources are not large on the scale of supercomputers and can serve most of all of the activities discussed in Sec. III.C.2.

A supercomputer known as the Next Generation Cyber Capability computer (NGCC) is housed in the same room as the A2C2 clusters, but is not part of the A2C2 operations. It is associated with the Complex Adaptive Systems Initiative. It was purchased for $4.7M and is currently mostly idle with few or no dedicated users. It is a large and flexible cluster that can adapt to many types of jobs. However, due to a somewhat unusual design, it is not clear whether the NGCC can be adapted to the needs of the Life Science users.

Requirements for the humanities and similar areas can not be easily estimated at this time. A pilot version of the video analysis project of M. Simeone, with a small number of files, needed about 50 GB, but ran into I/O saturation with 12 cores. One can expect these numbers to grow rapidly as the project scales up. Projects with 'big data' will certainly have large needs.

V.B. HPC Access

In general, access to HPC facilities can be obtained by several methods.

- Colocation: A user or user group purchases a system for exclusive use, but housed in the central facility.
- Condo: A user can purchase nodes (blades or boards), typically with multiple cores, for a common-access system. There are several variations.
    1. The user has exclusive 'ownership' of the number of purchased cores.
    2. The user receives an allocation of time (cpuh) that can be used for jobs even with many more cores than purchased.
    3. The user's jobs are given higher queues or priority for running.
- Pay-per-Use: A base rate in terms of CPU-hours is applied. Higher priority can possibly be obtained for higher rates.
- Proposal: An allocation of time (up to some maximum) can be granted based on a review of a proposed project.
- Free: A job can be submitted by anyone without cost. It may be subject to low priority, restricted queues, or other restrictions.

The Community Cluster model is the first option of the Condo method. For A2C2, the other Condo options are not available; neither are the Proposal, or Free methods. The Pay-per-Use method is available for the general-use Saguaro cluster, until it is retired.

The Community Cluster program has the advantages of ease of use and high throughput. It also has several disadvantages:

1. The user will have a very large first-year cost to provide the maximum number of cores that are expected to be used. As a result, some users have opted to use other resources. The concern could be reduced if a choice of the time period, say 1 to 5 years, could be made as is done at some institutions.
2. There is a loss of flexibility (or a cost inefficiency) if the user plans to run jobs with a variable number of cores (*e.g.*, 1–N), or to run jobs irregularly.
3. After 5 years, there is no guarantee that a funding agency may wish or is able to make another large investment in new cores. Some academic fields have projects that are fully active over a decade or more, and their computing ability could suddenly be terminated.

The second Condo option, to receive a CPU time allocation for core purchases, allows for both cost recovery to the facility and wide flexibility for the user. It is also more easily extendable. In effect, it is a Pay-per-Use method built on core purchases. It is a common option at other universities, often combined with priority scheduling, and seems superior to the current Community Cluster method.

The Free method has very wide use at other institutions. In fact, 77 of the 108 institutions listed in Appendix *A* of the University of Houston report include free access in meaningful forms [UH14]. A typical statement, from Michigan State: "Accounts on the HPCC are available free of charge to all MSU researchers." At UofA, "The base systems are purchased and supported with central funding so there is no charge for using the systems." NAU does not charge users. Normally, buy-in options are also available, and the free use may have various limitations. Even if a 'free' option is not directly available, access may be available through department or other user accounts, possibly through proposals.

The Free method provides for (1) beginning faculty who are building up a program; (2) faculty in units without computing infrastructure; (3) faculty between grants who may have difficulties getting funding re-established; and (4) faculty who wish to change research areas. The method can be combined with the Proposal method. It is an essential feature of an institution that has a vision to expand the role of HPC to new and exciting areas.

Altogether, the best approach for access would be to change the Community Cluster model to the second option of the Condo method, and include both the Pay-per-Use and Free methods.

V.C. Education, Training, and Outreach

A crucial part of a revitalized ASU HPC facility is restoration of a strong training program for users. Supercomputer operations are usually based on Unix or Linux operating systems, which differ substantially from Windows systems. They are also typically batch-job systems rather than GUI-based interactive systems. This latter distinction makes sense when computation jobs can run for many days or weeks. Expanding the use of HPCs to many more academic areas will thus require substantial technical assistance.

A strong outreach program (within ASU) will be necessary to build HPC activities across the full range of academic disciplines. Opportunities will need to be identified and potential users cajoled into bringing them to fruition. Similar to how new and faster methods for genome sequencing suddenly changed what biologists do in research, the digital revolution can transform what historians, social scientists, faculty in the arts, and others can do. It will take some time but, with dedication, ASU can obtain very high recognition.

Education of students is always at the heart of what a university does. As mentioned in Sec. III.B, HPC courses provide expanded opportunities to prepare themselves for their careers. Being able to work directly on HPC systems is essential. In short, the vision presented here advances the primary goal of the university as well as the integrated goals of research.

VI. Recommendations

Ten years ago, ASU initiated a University-wide supercomputing facility with a goal of bringing HPC capabilities at minimal cost (including free) across the academic disciplines. It would join the growing number of other universities that were building HPC services which are needed as a foundation for research, and to be more competitive and successful in obtaining research grants ASU hoped to emerge as a leader in HPC activities.

Resources and users grew, and the University got recognition in the academic community. Then the recession hit. ASU did not have the deep resource base that other universities had, and even lost a critical part of its base. University support decreased; eventually, some users took other options, and the hours of use drastically declined. There are reports of faculty candidates who declined interest in ASU due to limited computation support, and faculty who have or may move away for similar reasons.

The hardware aged beyond the end of service contracts, and would need to be replaced. With resources down, concerns about the sustainability of A2C2 increased. As in the old days of mainframes and research in STEM areas, the University model changed to one in which only those who bring in external funds can continue to use the facilities. As discussed in Sec. V, the Community Cluster model works against the growth of HPC in the University. It is contrary to increasing the research profile across disciplines that underlies student education, and may itself be unsustainable.

It should be noted that the broad PC culture of the University (Computer Commons, offices, laboratories, *etc.*.) also requires periodic hardware extensions and replacements. Except for the fraction of PCs that are obtained on research grants, the University must provide them from operating funds.

Research is costly and the university must cover its bills. External funding agencies will not cover all of the costs. So the tension is to find the hazy line that provides the necessary resources while also distinguishing an environment that allows faculty to thrive, pursue front-line and innovative research, and become leaders in HPC applications beyond the STEM areas. At the moment, we appear to be on the wrong side.

Section V described a vision in which HPC is to be seen as part of the core infrastructure of the University, analogous to libraries. It has grown into other universities, and must be the future here.

> **Recommendation 1.**
> **The University needs to move as rapidly as possible towards establishing HPC as part of its core infrastructure. The Vision is to include expansion into new academic areas such as the humanities, social sciences, and arts where the University can establish a special mark for itself.**

The challenge for the administration will be to provide the resources, including faculty. The challenge for the faculty will be to identify creative fore-front projects and develop the tools to pursue them.

Initially, the expansion into non-STEM areas may not require substantial resources and might be well served by the ASU Cloud Computing option discussed in Section IV.C. The implementation of such an option is under active study, and needs to be brought to fruition.

> **Recommendation 2.**
> **Studies leading to a possible implementation of an ASU Cloud Computing option must be actively pursued. In addition to the technical network issues, consideration is to be given to a structure for friendly and flexible user access. Network connections across campus must be enhanced. The Task Force recommends that the facility is to be at no cost to users.**

Because the Cloud Computing option may not be available very soon, and is likely to be much too costly for the large-scale power users, the standard HPC capabilities must be continued and even expanded.

> **Recommendation 3.**
> **To be competitive in the academic world, the university-wide, shared HPC resources in A2C2 should be expanded as rapidly as possible towards 5,000–10,000 cores, comparable to the era of 2009 and many other universities.**

As discussed in Section V, the current Computer Cluster model has clear deficiencies that prevent some users from buying in or using it at all. There are indications that the resources and operation model are having detrimental effects on recruitment and retention.

> **Recommendation 4.**
> **As is common at other universities, ASU should house clusters in A2C2 with a variety of processors as a shared facility. Users may buy cores for the shared clusters, and obtain an allocation of CPU-hours (possibly with scheduling priority) to use as they see fit, including the use of more cores than purchased. A user will not own specific cores. The submission of a job from such a user shall not cause termination of other, incompleted jobs.**

A good example of such an arrangement is at Michigan State University (see Sec. V.B). The HPC has about 14,000 cores spread over 11 or more architectures for free use. Users can also buy into 4 of the architectures in units of 20 cores for costs between $3,000–9,000 (compared to $20,000 full cost at A2C2). They get higher priority for their jobs. The University of Arizona and many other schools have similar policies.

A user or group may purchase a cluster from external funding and house it in A2C2 for the group's exclusive use (Colocation model). Their advantage is that they have all of the needed utilities and no conflicts with other users. The advantage for the University is that utilities do not have to be installed in individual labs.

> **Recommendation 5.**
> **A user or group that purchases a cluster from external funds may house it in a common University location to have utility services. The group may allow the use of the facility during idle times by other users, at Pay-per-Use rates. However, the submission of a job from the group shall not cause termination of other, incompleted jobs. Furthermore, if any portion of the cluster is subsidized by University funds (as in 2-for-1 deals or otherwise), the cluster shall be treated as a shared cluster as in Recommendation 4.**

Some users may not be able to invest in cores or may not have funds for HPC computing, but nevertheless need to develop projects for proposals, teaching, or other reasons. As with the majority of universities, they should not be excluded.

> **Recommendation 6.**
> **The Pay-per-Use and Free options must be available for all users of the shared cluster facilities. The change in rates from $0.025 to $0.01 per cpuh after 2012 produced higher cost recovery, even as usage declined. The Task Force strongly recommends that the lower rate be used, and that in no case shall it be raised above $0.025/cpuh. A review of a proposal is an option that could be applied to the Free method.**

The substantial and flexible NGCC cluster is currently located in the A2C2 facility, with little or no use.

> **Recommendation 7.**
> **The funds for NGCC were provided by OKED, and the office must make a concerted effort to find a viable use for the cluster, either as part of the University-wide shared facilities or to a group for a specific purpose.**

A HPC can not be successful without having a substantial expert staff. Implementing the Vision in this report requires that, in addition to technical software and hardware support, the staff also engage in dedicated outreach to the faculty in a broad range of academic areas. The staff must provide training seminars for users and also support courses for student education.

> **Recommendation 8.**
> **To become comparable with the support provided at other universities, the operating budget for A2C2 must be increased to $2–3 million per year. The staff must be proactive in providing high-quality HPC services to users, with special attention towards expanding services into non-STEM academic areas. Cost recovery from core purchases and Pay-per-Use charges shall separately be used for HPC hardware including addition cores, core replacements, and data storage.**

Finally, although the Task Force was asked to also explore the issues of usage and costs regarding other user facilities, such as clean rooms, microscopes, *etc.*, it could not be completed. Another task force will be needed.

References

Amazon http://qws.amazon.com/hpc

AR14 A2C2 Annual Report, 2014.

Azure http://asure.microsoft.com

Boinc http://boinc.berkeley.edu

Google https://cloud.google.com/compute

HPC11 Slide presentation A competitive and sustainable model for Advanced Computing at ASU, Frank Timmes, 2011. (Document is out of date and no longer posted on the A2C2 web site.)

QR_Su13 A2C2 Quarterly Report, Summer 2013.

QR_Wi14 A2C2 Quarterly Report, Winter 2014.

QR_Wi15 A2C2 Quarterly Report, Winter 2015.

UH14 Report The Case for High Performance Computing at the University of Houston, 2014. Available at `www.cacds.uh.edu/sites/default/files/business_case_hpc_at_UH.pdf`

US15 University Senate motion 2015-44, `http://usenate.asu.edu/node/5133`