

# Advanced Computing Issues at Arizona State University

## A Faculty/Teaching Perspective

Arnold Maltz, Ph.D.  
Associate Professor, Supply Chain Management

### Executive Summary

This report covers the findings from a set of interviews I conducted in the summer of 2015 exploring the issues around advanced or “next generation” information technology needs at Arizona State University. The work was performed under the sponsorship of the Provost’s Office, pursuant to my position as incoming President of the Tempe Faculty Senate. The research was motivated by the increasing needs and expectations of information technology users both inside the University and among ASU partners. Some of the key findings include:

- 1) Several different aspects of the issue were identified including
  - a) Data Storage Available datasets beyond the “hard sciences” have increased dramatically in size over the past few years. For example, traffic sensors have been installed in many more places, sensors are being attached to production machinery, components, and shipments, and medical information is being generated at an increasingly granular level. ASU researchers are now asked to handle multi-terabyte datasets for some research and teaching/industry project efforts.
  - b) Data Composition Systematic, computer-assisted analysis of qualitative as well as quantitative data is becoming very common, and the two types of data are being connected in joint datasets in some cases. At ASU, content analysis is now part of the research and teaching tasks in Communications, Journalism and other social sciences, and matching those outputs with quantitative data will become an assumed skill for researchers and sophisticated students in many different areas. ASU industrial and government partners are also advancing in these kinds of projects.
  - c) Data Analysis As the availability of “Big Data” such as the above becomes commonplace, the tools to analyze this data are also progressing. Typically under the banner of “massively parallel” computing, these sophisticated algorithms and visualization tools are used to find patterns in the large datasets, display these patterns in a way that facilitates further investigation, and even automate many reactions to changes in environmental conditions. ASU researchers and teaching faculty are faced with understanding these newer tools, determining how to use them, and finally how to expose students to the newer analytical packages.

- 2) On the research side, these are not new concerns.

The physical and biological sciences, as well as Engineering, have been dealing with these issues for some time, especially in supporting research proposals and training advanced graduate students. Thus the Arizona Advanced Computing Center (A2C2) and other highly capable computing facilities are available on campus. For a summary of current issues please see the High Performance Computing Task Force Report of April, 2015 ([http://usenate.asu.edu/files/HPC\\_Final\\_Report\\_4-17-2015.pdf](http://usenate.asu.edu/files/HPC_Final_Report_4-17-2015.pdf)).

Even though High Performance Computing (HPC) resources are available on campus, they are not widely used in teaching nor are many faculty aware of their accessibility. Furthermore, a joint committee of the University Technology Office (UTO) and the Office of Knowledge Enterprise Development (OKED) is currently evaluating how to update and advance these resources which are important to research competitiveness for grants as well as working with external partners.

- 3) Notwithstanding 2) above, virtually the entire campus community recognizes the need to understand and address the issues around advanced computing (HPC).

Although the interviews underlying this report were not comprehensive, every interviewee commented on the need to better understand the situation and facilitate more effective faculty involvement. Furthermore, anecdotal reports have come from the Herberger College relative to “digital arts”, the Cronkite School which has reportedly been hiring data scientists, and the Health Solutions, Nutrition and Nursing programs which are faced with much more detailed patient and epidemiological data that they must process. Add to this the increasing digitization and online storage of historical and literary documents, current and past demographic data, video and audio performances, and transaction level sales data, and it seems clear that ASU needs to think about these issues from an enterprise perspective.

### Genesis of the Project

One of the rapidly growing programs in the W. P. Carey School of Business, where I am on faculty, is the Business Analytics Program jointly offered by the departments of Information Systems and Supply Chain Management. The program was started two or three years ago and has quickly grown to four cohorts of some fifty students each, as well as online offerings. Business Analytics for corporations and businesses is about better understanding and acting on the very large amounts of data that many companies routinely collect, whether that be sales data, financial data, operations data, or data about employee performance and satisfaction. As sensors are added to assets and processes are

increasingly electronic the amount of data (as opposed to usable information) is growing exponentially. Companies want to know what to do with their “Big Data” and how to act on it.

However, when I spoke to a faculty member of the Business Analytics program he was concerned about the limits he had run into in teaching in the program. An anonymized version of his concerns can be found in Appendix A. He and I both agreed that these issues might very well university-wide, hence my proposal for this research.

### Project Process

I interviewed a variety of people who were recommended to me as knowledgeable in this area, including faculty and staff from W. P. Carey, Engineering, CLAS, and UTO. Because of time constraints the interviews were all with Tempe located personnel with one exception (BioInformatics). Unfortunately, I did not talk to people from areas such as History, Psychology, Visual Arts, etc. who undoubtedly need to be included in further discussions.

Although I did not promise anonymity to those interviewed, I have elected not to include names in this report.

I did not gather “objective” information on resource use for the various university units, but that is or soon will be available from UTO. However, one of the findings in this research is that a number of researchers are not using the HPC resources on campus but have found other ways to get their work done. So even comprehensive UTO statistics may not identify all the “power users” at the university.

### Research

With one or two exceptions (see below) the various interviewees assumed our discussions should mainly concern research activities related to advanced computing. Many of these concerns were already covered in the High Performance Computing Task Force referenced above, so I will only cover the highlights.

- 1) Computing power and capability is one of the criteria used by funding agencies in evaluating research proposals. Both NSF and NIH, in particular require evidence of access to necessary computing resources as part of the proposal. At this time, ASU is actively seeking partnerships and/or working with outside providers (e.g. Amazon Web Services) if university resources are not adequate for the task at hand. This is one option for the future, but probably not the whole answer.
- 2) At this time computing resources are not adequate to support ASU aspirations. There are a number of reasons for this, and the issue is being addressed at the CTO/Vice-President Research level.

- 3) Funding models are being reconsidered (again, see the High Performance Computing Task Force Report) and changes, if any, have not been finalized as this is written.

Any options adopted in this area may have to recognize that much university research, especially in non-STEM fields, is not supported by outside funding. This kind of research assumes that necessary software is essentially a public good, available to the entire campus community. There are reports that this model is in place at Stanford, among other places.

### Teaching

When interviewed, W.P. Carey personnel chose to emphasize both teaching and research. This may be the result of the lesser role of funded research in the W. P. Carey business model, or it may have come from the acute issues in the Business Analytics program. In any case, both Carey and Engineering emphasized that college-level Citrix, the current method of sharing common software with students and classes, will have to be adapted to the new requirements, possibly with specialized “streaming tools.” Engineering is reportedly no longer using Citrix, while Carey is investigating options including moving instruction to the UTO Citrix environment. At a minimum we need to give many of our students exposure to and an appreciation of High Performance Computing. What is necessary for that is not clear.

As for teaching courses, professors for mainstream classes prefer to use small to moderate size problems that do not require massively parallel computing power to teach analysis methods. Since there is some open source software (e.g. R) available that runs on laptop/desktop machines, the professors I talked to had not utilized the HPC clusters for their classes. Although A2C2 historically has offered support for a few advanced parallel computing classes annually, there was no dedicated support for this effort.

This year, one professor has arranged for fifty Virtual Machines for his class to insure adequate performance and reasonable throughput times. I cannot evaluate whether this approach is scalable or whether the economics will work for classes without extra fees.

One more point came up in several interviews. Those responsible for providing IT resources and many of the faculty users do not even have the same vocabulary when trying to define needs and capabilities. In other words, UTO and other IT professionals often ask “What do you need?” for the classroom. But users do not really know what that question means or what is available in terms of resources/possibilities. Faculty are used to having resources such as Blackboard and Citrix simply provided, and they assume that necessary capacity will be forthcoming. When asked to define needs more carefully, many, other than those who teach information systems based classes, do not know where to start. This disconnect, common in all large organizations not just universities, is a continuing source of frustration. But when faculty are expected or trying to push the limits of analysis so as to expose their students to current practices, the issue becomes acute.

## Outside Projects

Many ASU programs make extensive use of outside partners for classroom projects as well as research partnerships. As our outside partners improve their capabilities, we somehow need to keep pace. This arena seems to be the leading indicator of where we are falling short.

I have already indicated that outside funded research opportunities are at risk because our HPC infrastructure is middling, at best. But on the teaching side, the ability of our students to learn through working with outside agencies and companies is also at risk. Some easily imagined scenarios:

- 1) The datasets our partners want analyzed cannot be brought into the university environment either easily or securely. One professor reported that the partner company wanted the students to look at a 4 terabyte dataset. After determining that the outside electronic connections would not accommodate this much data, the company physically delivered two two terabyte hard drives so that they could be physically added to the relevant servers. Another company that has potentially useful data for my own work has 337,000,000 transactions on an annual basis. I'm not sure I know where to start.
- 2) Traffic flows and information are increasingly monitored on a real-time basis by networks of sensors. As this data gets more fine grained, the ability of our geography and civil engineering students to work with it could be compromised by our limited computer infrastructure. Similarly, large agent-based simulations are certainly possible now, but how large a problem ASU can teach people to handle, even in the Decision Theater, is not clear to me as a faculty member who works in transportation.
- 3) In the political science area, micro-targeting for campaign purposes is clearly data-intensive, as is the messaging on social media that goes with it. Can our students work with this much data? How scalable are ASU capabilities in the area? Clearly, that is a discussion to have with experts, but it is reported that data scientists are in demand for these problems.
- 4) Visual and performance artists are even now exploring the possibilities that go with massive numbers of sensors and immediate audience feedback. Do we need more power, or better access, or both, for students in these disciplines, especially if we are cooperating with nationally known institutions or major film studios?

## Conclusion

This investigation into ASU advanced computing capabilities has been fascinating but is also clearly incomplete. I have not spent enough time with most of the professional schools, nor the specialized programs that are not on the Tempe campus. Since many of

those have even closer connections with outside partners, they should be given a greater opportunity to add their opinions and needs to this effort.

Furthermore, let me emphasize again that mine is but one of several efforts underway to address the computing capacity issues to support ASU's aspirations. UTO, OKED, W. P. Carey, Fulton Engineering, and various departments in the College of Liberal Arts and Sciences (CLAS) are all participating in and sometimes leading multiunit task forces to think about the issues raised in this report.

In synthesizing the interviews, the following themes stood out:

- 1) Although research and research funding are the current drivers of HPC needs, everyone agreed that teaching either HPC techniques, or at least HPC appreciation was an important consideration as we upgrade infrastructure. In particular, I am concerned that advances among our partners will make it difficult or impossible for our students to successfully execute projects with some of our major outside participants, including Fortune 500 companies, local and state governments, regional planning commissions, etc. We may want to try to determine what a minimum knowledge base of HPC is for the various disciplines currently involved.
- 2) Technical support personnel and faculty need some vehicle for an ongoing dialog on both research and teaching expectations and advances. Right now, in many cases the two groups are speaking "different languages." This is exacerbated by current faculty acceptance of computing power as a public good, and technical people trying to keep up with escalating faculty expectations. Currently there is a group led by OKED working to make this a reality.
- 3) ASU needs to decide if we are going to make computing power "ubiquitous" as it apparently is at many leading universities, and if so what that means and how to find the resources to support that goal.
- 4) In connection with 3) above, ASU needs to formulate a funding model that will support faculty efforts aligned with the goals of the university and the individual units. See the HPC Task Force Report for a summary of previous efforts in this area, but it is unquestionably an open issue at this time.
- 5) The balance between centralizing efforts throughout the university and leaving initiatives at the unit or even subunit level needs to be thought through. Several interviewees thought it unwise and unrealistic to expect nearly complete centralization, while others believed that centralizing efforts would maximize efficiency.

This report should be an opening leading to an ongoing dialog throughout the university. As I noted, some units are already moving toward this, others are essentially "doing their own thing" and others don't yet know what their concerns are. The faculty, even those of us who are not directly affected by High Performance Computing requirements, need to stay involved. We may not teach sophisticated analysis in our courses, but our students

almost certainly will be, at least in evaluating the results of such analyses performed by their mechanical servants.

## Appendix A - A Report from the Front Lines

I'm not sure how broadly relevant the following info is so I'm sending it just to you and you can manage it from there.

I met twice in the last month with WPC admins re support for computing in the MSBA program. My intent was to convey to them that this isn't an MSBA problem but a growing university problem and should be addressed at that level (as well as administered at the college levels, too). IT is as necessary a part of the university infrastructure as having classroom space for students. IT infrastructure is organic to being the new American University and should be viewed as being as indispensable as having classroom space for students when they show up to be educated.

The first meeting was with xxx, xxx joined that meeting, too. The second meeting was with xxx, xxx, xxx, xxx, andxx.

I think the broader context ("problem") is that, in electric utility terms, the 'base load' has risen substantially and is rising faster each year. This means that there is less 'peak load' capability when larger assignments hit the network, so things choke. The reason the base load is rising so fast is that all the colleges are incorporating more computing and larger datasets into their routine large-section classes. Textbooks all come with access to easy-interface software now, homework files come with larger and larger datasets because data is essentially free, and more computing is required across a large proportion of university courses. As an example, the journalism college has acquired SABR, the baseball stat group that has been around for decades (HQ in Phx), the school hired three data scientists for its faculty starting this Fall and is incorporating stats and analytics into its curriculum because that is a key part of journalism now (the Nate Silver effect, so to speak). This same thing is occurring in literature (textual analysis – high demands on infrastructure) where they have a nexus lab for textual analysis, and similar things are going on at all the colleges. Similar trends in music, etc. It isn't just science, engineering, stat, and analytics classes.

In WPC, xxx has said a number of times that the college wants to include more analytics in all the curricula. Recruiters are assuming familiarity, at least, with some analytics and facility for some majors.

While the IT challenges have arisen for MSBA program in the WPC context, the resources needed to address this rising tide should be a university issue, I believe, not just college-by-college budgeting gymnastics each year.

Example details of challenges: University (and college) IT is moving more software to servers (fewer laptop licenses). This makes sense administratively and software providers like it but it only works if the servers can support the traffic. For the second semester of the MSBA program (and some the first semester) as student computing capabilities increase, problems that should require maybe 15 minutes of computing time take up to 4 hours to run. Sometimes they get bumped off during that time and



everything starts again. Sometimes the students can't even get on Citrix; other times they can but can't get access to the software they want (some software is used across many programs and there may be license issues on number of simultaneous users). Assignments had to be modified, some dropped altogether due to inability of students to do homework.

There are three aspects of infrastructure support that need to be addressed (in combination, of course):

1. Storage: this, so far, has been the least of the problems for us, but it has arisen. The size of datasets and databases, esp for programs like the MSBA, have increased dramatically in the past three years. We did some textual analysis projects for the nexus lab and there was a project with Intel that involved 4 terabytes of text – just one project. When Amazon provides unlimited storage for anything for \$60/year for an individual and google pictures provides unlimited storage for free for pictures (as long as each is less than 16MB), publishers and companies are showing up with larger and larger datafiles. We can store data on the cloud but the issue is data that has to be behind the firewall or accessible to Citrix server software. I believe this issue is being actively addressed for next year.

There is a second aspect to data management that has been problematic: how to get data from off-site sources. The ftp ports were mostly shut down last academic year. One project company had a small (50 MB) datafile for us and we eventually had to drive to their office with a USB drive to get it – the university didn't have a way to transfer the data to us. When Intel gave access to the 4 TB of text, they just brought two 2 TB portable harddrives and left them here. There was no other way to get the data. Companies often are not allowed to put their data directly on the cloud (dropbox, etc.), so if they can't wire transfer it, physical pickup is the only alternative.

2. Computing muscle: more programs are doing things like ensemble modeling where a large number of analytic programs are launched simultaneously on relatively large datasets. Even single analytical procedures (say, for example, cluster analysis) that are run iteratively require a lot of computing support. The current server configuration has been inadequate for this. Using the MSBA as an example, the number of students has increased from 45 in Fall 2013 to nearly 90 in Fall 2014 and will be 150 in Fall 2015; in addition to more bodies, the software we are using is more sophisticated, the students are more sophisticated, the problems are larger and more sophisticated.

3. Bandwidth: there is 'data at rest' (as they say) in datafiles – that is the storage problem – and 'data in motion' issues as communication over the infrastructure increases in volume and speed. Running programs remotely in a server-client arrangement puts large demands on the bandwidth and sockets available, etc.

There is a final issue that further complicates things and that is vocabulary: we (faculty, for example, both IS and SCM) don't know the terms that the IT people use for characterizing capacity. We can't give them descriptions of what we need in the terms

they use. We can describe our demand in the ways I have described them above but I don't know how many portals or sockets or cable specs or even total storage capacity needed to adequately support academic operations.

We have talked some to industry partners and xxx has suggested getting targeted donations for servers, etc., but all this presupposes a structured way to identify and obtain relevant information. So far, the interactions have been rather ad hoc. A committee (such as yours, perhaps) could set up a small group who could take a more systematic approach to gathering info – maybe along the lines that are taken when info is gathered before building a new building.

Talking to industry folks has the same vocabulary issues: IT specialists (infrastructure people) not only use different vocabulary but they typically are not free to discuss a lot of detail without some higher authorization and often NDAs or institutional agreements. I spent a fair amount of time this past spring getting NDAs in place just so we could talk to people and share some info, but I'm not an infrastructure person so it was primarily just to facilitate our own projects with them. Even then, ASU security issues prevented things as simple as ftp transfers of small datafiles.

MSBA did not use any massively parallel computing this past year. This was something that was looked at a couple of years ago but is outside (or above?) the programmatic level for planning purposes.

I'm glad this is getting addressed at the university level. Technology that industry uses (analytical software, large datasets) or wants to use is outpacing many universities' ability to provide hands-on training on that technology. Mesa and Maricopa community colleges have historically had a better reputation with some companies in the Valley for hands-on experience than ASU has. You might at some point consider talking to the IT counterparts at MCC or Maricopa Tech to see what they are doing.

One other quick private comment: in talking with xxx and xxx, it has been important to keep clarity on what kind of 'capacity' you are talking about – this may or may not be an issue at the university level.

It is not an easy conversation to have: there are (1) multiple capacities (storage, computing, bandwidth), (2) there are vocabulary issues (they use technical terms that are not familiar), (3) there are scale issues (we are talking about the MSBA issues to illustrate the challenges but they have to think on the college or university levels), (4) there are timing issues: they are always talking about what will be in the future and how certain issues will be solved then and we are talking about right now and the ability to support activities, and (5) there are a lot of turf issues – budget, authority, resource access, etc., and this really chops things up and makes discussion progress very difficult at times. Finally, it is really easy to have very lengthy discussions that get wrapped around one particular event or problem or one particular initiative (a new server, more off-site storage, etc.) and meetings run out of time and it has again been primarily a floating discussion of current particular experiences, future plans, and rather vague next steps.

I would suggest that your meetings at some point have a strict agenda that is enforced time-wise, that war stories and personal experiences be kept to a minimum, that future things that are being considered be included only to the extent that they impact current agenda items, that initially matters of turf be kept secondary at least until the scope of the challenges are better defined, that there be some clearly defined meeting outcomes, and that there be a person whose primary responsibility is to manage the agenda and keep the group on schedule and on topic.

Anyway, my point is that even though the issues seem plain (we need upgraded IT infrastructure and everyone working in any of these areas has a horror story or two about which they are quite passionate), it is surprisingly difficult to have productive discussions about the issues, let alone identify paths forward.