Southern Methodist University Task Force on High-Performance Computing: Final Report

Thomas Hagstrom (Mathematics) Chair,
Thomas Coan (Physics),
Tom Fomby (Economics),
Ira Greenberg (Meadows & Computer Science),
Sukumaran Nair (Computer Science and Engineering),
Sherry Wang (Statistics),
Indraneel Chakraborty (Finance),
Dieter Cremer (Chemistry),
Joe Gargiulo (OIT),
Nathan R. Huntoon (Electrical Engineering),
Justin Ross (OIT),
John Wise (Biological Sciences)

April 30, 2012

What is high-performance computing?

"High performance computing" (HPC) is the design, development and execution of software intended to run on specialized computing facilities whose processing speeds and data storage capabilities are typically at least $1,000$ times greater than what is capable using a desktop or laptop machine. This enormous increase in capability results from the concentration of processors and data storage devices in close physical and electronic proximity to permit the rapid exchange of data and intermediate results produced during an extended calculation. The processing power of such facilities also makes visualization of highly complex time varying physical structures (e.g., proteins and magnetic fields) practical. Typically, many users run their software simultaneously on the computing facility. Management of the submission and execution of user programs is done through sophisticated software that ensures that the computing hardware is utilized efficiently.

Rapid advancements in computing technology over the past decades have revolutionized research in science and engineering, leading to the concept that computational and data-enabled science should now be recognized as essential components of the scientific endeavor, in equal standing with theory and observation [1, 2]. In the second decade of this century, we expect further dramatic technological developments, enhancing the impact of computing not just in these traditional fields of application, but also reaching into business, the social sciences, the humanities, and the arts. A coarse measure of the processing power of an HPC facility is the number of arithmetic operations it can perform per second, termed flops. Currently, the world's largest computers are passing from the teraflop (10^{12} flops) to the petaflop (10^{15} flops) range, with the exaflop (10^{18} flops) barrier expected to fall before the end of the decade [3]. However, the needs of different applications have led to the development of varied HPC architectures which differ in the ways individual processing cores access the system memory and communicate data with each other. Moreover, we are now in the midst of a revolution in system design due to the exploitation of graphical processing units with thousands of cores for numerical computations.

Given the ongoing and often unpredictable developments in HPC technology, it is perhaps better to define it through its impact on society, and on the mission of the university in particular. For the first time the development of tornadoes and hurricanes, the prediction and the consequences of earthquakes, the folding of a protein, or the biological activity of an anticancer drug can be described with high reliability on the computer. Universities and research centers, which have the facilities to engage in this kind of computer-enabled research will create new knowledge which will have a direct impact on our society as a whole, and will dramatically improve their teaching and research possibilities. Interdisciplinary work will flourish and involve disciplines which without HPC would not be able to interact. Universities will be able to make the best use of research talent, and both graduate and undergraduate students who have access to HPC facilities will be able to carry out research early in their careers that will directly connect to questions about the sustainability of the world (health, food, water, environment, climate, democracy, war and peace).

What is the importance of high-performance computing?

High-performance computing, used both to simulate comprehensive models of complex systems and to rapidly and automatically analyze large datasets, has had and will continue to have a deep impact on science and engineering. It has not simply made the tasks of producing quantitative predictions from theories and extracting meaningful correlations from experiments more efficient, but has led to the emergence of new and innovative ways to understand the natural world. Prominent among these emerging approaches to performing scientific and engineering research is the development of multidisciplinary teams whose members can contribute authoritative submodels and extensive data, integrated via technology into hitherto unimaginable full system simulation tools. Many examples can be found of contemporary research which would have been inconceivable without the past decades rapid developments in computing technology. Below is a more discipline-centric discussion, with a focus on the interests of faculty in Dedman College.

Biology The complexity of biological systems inhibits our ability to design effective drugs. Using HPC it is possible to systematically develop large compound libraries for drug research and to simulate drug-receptor interactions. This process reduces the average development time of a drug (up to 14 years) in half, and may lead to the consideration of compounds which would otherwise have been neglected.

Chemistry Modeling of larger molecular systems employing ab initio (first principles) quantum chemical methods provides the possibility of investigating reactions and their consequences in the polluted atmosphere, intensifying the research on the chemical background of global climate change, studying the possibilities of water purification (especially from arsenic compounds), investigating protein similarity and protein folding, and solving the secrets of chirality in biomolecules.

Earth Sciences Using HPC, it is possible to simulate large scale geological events. For example, the Southern California Earthquake Centers community model, implemented on 223,000 cores at a supercomputer at Oak Ridge National Laboratory, was used to produce detailed simulations of a magnitude 8 earthquake on the San Andreas fault, enabling accurate assessments of seismic hazards in a highly populated region. The Center involves more than 600 scientists at 17 core and 47 participating institutions, combining expertise in computational algorithms, geophysics, structural dynamics, and numerous related fields.

Economics An emerging, computationally-intensive field is predictive analytics. Here the goal is to exploit the zettabytes (10^{21}) of digital data which can now be accessed to make better decisions [4, 5]. For example, as more and more companies go to text mining the sentiments of customers using live streaming data, the requirement for high performance computing will increase exponentially in applied economic analysis. See [6] for a discussion of the extensive computing demands required for sentiment analysis of Twitter streaming data.

Mathematics Computational mathematicians focus on the development, analysis, and testing of algorithms to exploit modern HPC technologies, and also participate in interdisciplinary teams which use these algorithms to solve challenging problems in a wide variety of fields. The rapid changes in HPC hardware and the size of problems which can be attacked has led to an explosion of new algorithm development, focused on finding methods which will allow the problem size to scale directly with the computational resources.

Physics Various faculty of the SMU physics department are members of two large international experiments that make extensive use of computing resources. One of these (ATLAS) is based at the CERN laboratory in Geneva, Switzerland while the other (NOvA) is based domestically at Fermi National Laboratory near Chicago. (See Figure 1 for a depiction of an event detection.) Both of these world class experiments receive extensive funding from the US Department of Energy and the NSF. HPC is required to analyze the massive data sets produced by these experiments, with the demands on computing power and storage continually on the rise. For example, the data sets collected by ATLAS at the LHC collider at CERN double every 6 months as ATLAS continues to phase in planned improvements both to the collider and to the detector.

Statistics Statistics by its nature is a data intensive discipline. Modern computer processors are now sufficiently powerful to make computations for many classical statistical problems seem instantaneous. However, recent advances in technology have allowed for simultaneous acquisition of enormous amounts of data in the basic sciences and other subjects. For example, high throughput technology has emerged over the last few years as an important tool in accelerating the pace of scientific discovery. The density and volume of data generated in a single high-throughput experiment continues to grow exponentially. Statistical and computational methods successful in dealing with regular-scale data are no longer effective for such data, posing new challenges for statisticians.

Business In finance, marketing and economics, many questions require a structural approach that is computation intensive. These include the modeling of how firms interact with each other during given economic conditions, how consumers may react to increased competition and choices, how much should households save over lifetime and business cycle etc. We solve for optimal policy (regarding savings or production) of agents (which may be firms or households) in the presence of state space variables (e.g. macroeconomic shocks and agent specific shocks). The resulting policies are simulated for a large number of periods (years) for a large number of heterogeneous agents who interact with each other, in a general equilibrium approach. The moments thus obtained from simulation are compared with real data moments. We iterate in terms of optimal policy calculation and moment generation through simulation until convergence, which can take months on even high performance computers.

Engineering Engineers use HPC to fully model the behavior of complex sytems whose size can range from the nanoscale to airplanes. Such simulations allow the direct optimization of design details.

Arts Artists use HPC to create algorithmically derived forms and animations, as well as to visualize massive data sets ("Big data"). Rendering, the process of converting code to the realistic effects seen in special effects or fully animated 3D films, is very computationally demanding. HPC will allow artists to render even more realistic virtual actors and environments.

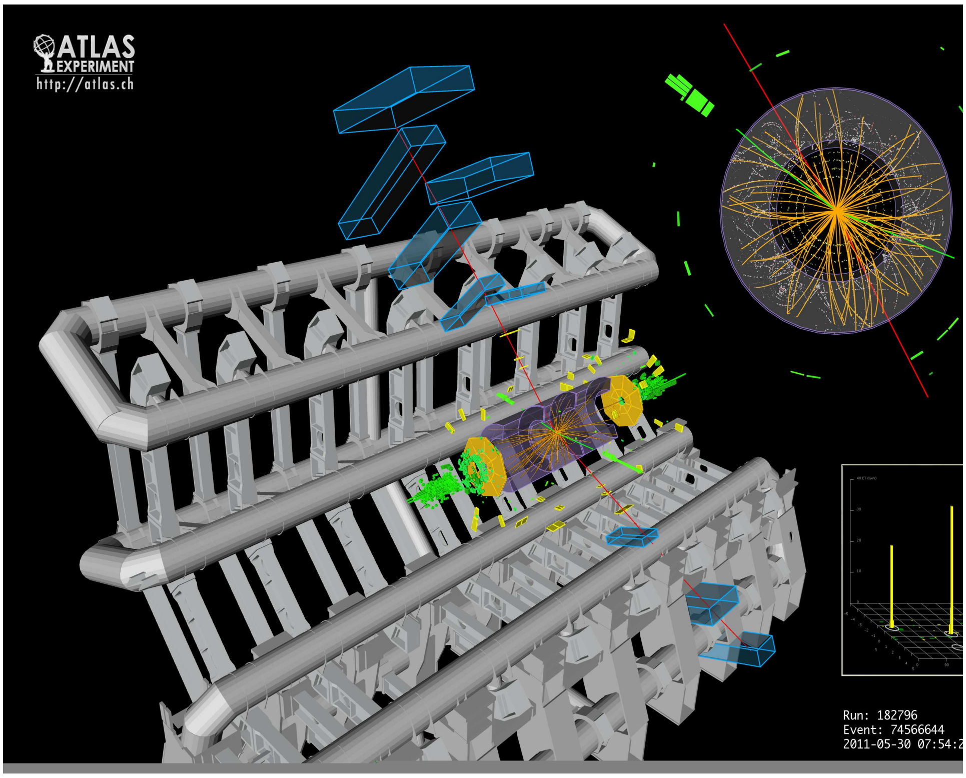

Figure 1: Depiction of the possible detection of a particle decay event at the ATLAS experiment. SMU faculty use HPC to find and analyze such phenomena.

This revolution is now moving beyond the sciences to essentially all fields of human inquiry. Besides the examples mentioned above related to large scale economic and financial forecasting and the exploitation of digital media in the arts, we note textual analysis in the humanities. These developments promise to produce even more radically new forms of collaboration, as already evidenced by SMUs Center for Creative Computation, centered in the Meadows school but engaging faculty from Dedman and Lyle.

How are SMU aspirational peer institutions investing in and utilizing high-performance computing?

A survey of high-performance computing at SMU's aspirational peer institutions reveals some diversity in approach. There are, however, two common themes. All but one have shared university-wide HPC facilities, the exception being Carnegie-Mellon which is a partner in the national Pittsburgh Supercomputing Center and which has extensive computational facilities supported within its Schools. (In fact Carnegie-Mellon offers Bachelor's, Master's and Doctoral degrees in Computational Biology within its School of Computer Science and in Computational Finance as a collaborative effort between Mathematics and Business.) In all cases staff support and basic infrastructure are provided by the University, and for a substantial majority of the institutions the computer hardware itself was purchased using University funds rather than external grants. Looking at the projects supported by the shared facilities, we see that the majority originate within the sciences (particularly Physics, Chemistry, and Biology) and engineering. However, we also see substantial use in the Social Sciences (particularly Economics), Management (particularly Finance), and some examples in the Humanities (for example the classics department at Tufts) and even Theology (at Boston College).

Among these peers, four stand out as having made the most substantial institutional investments in cyberinfrastructure for research: the University of Southern California, Notre Dame, Vanderbilt, and the University of Rochester. We will discuss these in reverse order. Rochester recently purchased a 4096-core IBM Blue Gene/P, which is the primary resource for its Health Sciences Center for Computational Innovation, and which, along with an earlier university-funded 1416-core Linux cluster, are maintained by the Center for Integrated Research Computing. (The cost of these systems to Rochester is not publicly listed.) Vanderbilt's Advanced Computing Center for Research and Education started as a joint project of Physics and Neuroscience. It was expanded into a University-wide center by an $8.3 million internal grant. Their primary cluster has more than 4000 cores, which have typically been purchased with external funds. Notre Dame's Center for Research Computing manages a variety of systems, with more than 10,000 total computational cores. Among five general-access clusters the largest has 5880 cores. The press release announcing its purchase describes it as part of a $1.8 million investment in computational hardware. The USC Center for High Performance Computing and Communications boasts the fifth largest academic supercomputer in the US. It has roughly 20000 cores and was built solely with local investments.

The total number of computational cores is a coarse measure of capability; additional and equally important components are storage, networking and support staff, where SMU has already made substantial investments using both internal and external funds, as well as visualization and other data analysis capabilities. These are all present at the four centers described above, with support staff, including executive directors, ranging from a half-dozen at Rochester to 30 at Notre Dame.

With the proposed investments in hardware and staff, SMU will vault towards the top of the list of its aspirational peers in terms of raw computational power, at least on a par with Vanderbilt and Rochester. (See Figure 2 below.) However, leveraging our traditional focus on undergraduate education and existing research groups in computational mathematics, physics, and chemistry, we believe that we can do more. In many aspects the centers listed above lack coherent educational programs in computational science, economics and the digital humanities, as well as the level of interdisciplinary collaboration we will achieve.

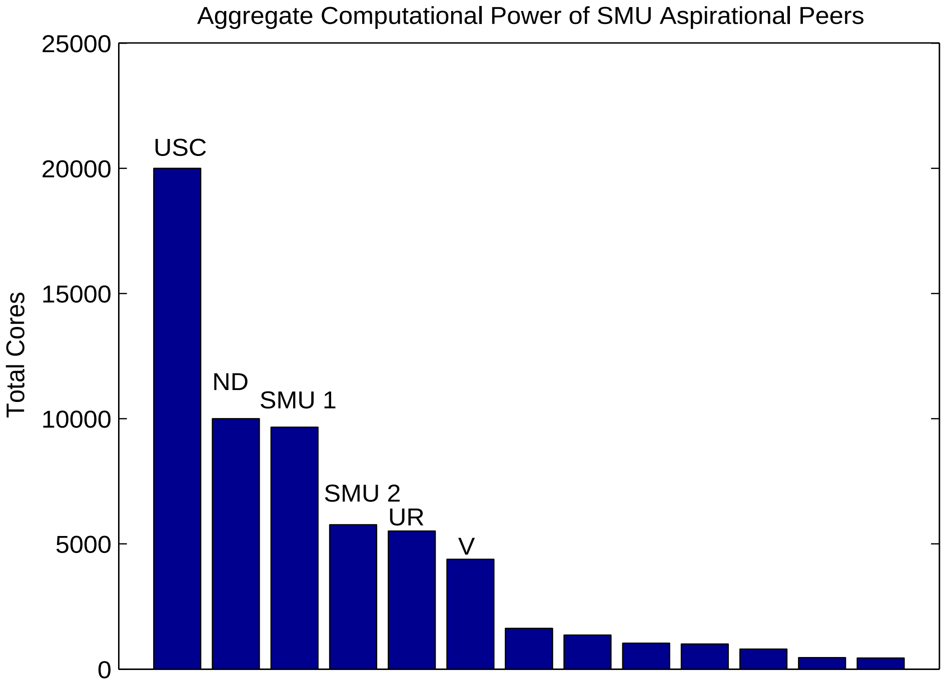

Figure 2: Available computational cores at aspirational peer institutions along with projections of SMU's facility. Here USC refers to the University of Southern California, ND to Notre Dame, UR to the University of Rochester, and V to Vanderbilt. SMU 1 and SMU 2 are projections of SMU's future capabilities under proposed investments.

How is high-performace computing utilized at Texas universities?

The Texas Advanced Computing Center (TACC) at UT Austin is among the world's largest academic HPC facilities. Their two current HPC clusters have in excess of 85000 cores; one will soon be replaced by a new cluster which will be capable of petaflop performance. NSF is the primary source of funding for these machines, with 90% of the compute time allocable to any academic researcher in the US via proposals to NSF's XSEDE program. The remaining time is allocable to Texas universities and industrial partners. TACC has more than 7PB of user storage, a local visualization facility, an extensive staff (more than 80), and offers a number of workshops, training sessions and for-credit courses. UT Austin's Institute for Computational Engineering and Sciences (ICES) is a world-class center for research and education. ICES has 39 core faculty, 93 additional associated faculty, 76 postdocs and other research associates, 71 graduate students, a 26-person administrative staff, and a large visitors program. It offers Doctoral and Master's degrees in Computational Sciences, Mathematics and Engineering, as well as undergraduate certificates and summer internships.

Besides UT, Rice and Texas A&M have made substantial investments in high-performance computing. Texas A&M maintains shared local clusters with more than 4000 cores, and offers graduate certificates in Computational Sciences through its Institute for Scientific Computation. The Ken Kennedy Institute for Information Technology at Rice maintains four large clusters with roughly 8000 cores, and has just announced a partnership with IBM to purchase a 25000 core Blue Gene/P, which will be shared with the University of Sao Paulo in Brazil. The Center provides staff support, but the existing compute hardware was purchased with external funds. A Master's Degree in Computational Science and Engineering is offered, and a separate department of Computational and Applied Mathematics, which is similar in configuration to SMU's Mathematics Department, offers degrees at all levels. Both the Rice and TAMU institutes boast extensive ties with the oil and gas industry. The Rice center sponsors a yearly workshop on HPC for this industry, and offers graduate fellowships sponsored by BP and Schlumberger.

In North Texas, the University of North Texas, the University of Texas at Arlington, and Baylor all have University-sponsored HPC facilities. Baylor invested $1.25 million in a 1024-core local cluster, UTA has an 840-core shared cluster as well as a larger machine devoted to Physics research, and UNT a facility with 2000 cores. Only UNT highlights its current research efforts; a Center for Advanced Scientific Computing and Modeling sponsored by the Department of Chemistry provides training workshops and classes in computational chemistry. Resources are set aside for computational research in the arts and humanities and the Texas Center for Digital Knowledge resides at the University.

With the proposed investment, SMU will become the premier center for research and education in high-performance computing in North Texas, with facilities exceeded in the state only by those available at UT Austin and Rice. The Rice model clearly shows that a high-performance computing center can catalyze engagement with local industry, and the SMU center will do the same in the Metroplex.

What are SMU investments to date, what resources do we have, and how do we use HPC?

SMU currently has a 3-component HPC facility targeting the three main configurations required by different user communities:

High-throughput cluster Purchased using a combination of external and internal funds, this is 1656-core machine used primarily for research in Physics and Biology. This purchase also provided 320TB of storage.

Tightly-coupled cluster Recently purchased via a grant from DOD to researchers in the Mathematics Department, this is a 384-core machine with fast interprocessor communications.

Shared memory cluster Purchased using SMU funds to support the Computational and Theoretical Chemistry group (CATCO).

Additional infrastructure investments which support HPC as well other computational resources include:

Patterson Renovation A temporary data center - this is where the current clusters are housed.

New Data Center Work will soon begin on a large, state-of-the-art data center, which can easily accommodate the proposed growth in our HPC facilities.

Networking and Fiber Plant Funds have been committed to enhance the campus network. Upgrading the campus network to 10Gbps is crucial for researchers needing to work with huge datasets.

The system is fully supported by staff from OIT.

There are currently 114 user accounts on the HPC systems, outside of the CATCO group. These include 32 faculty, 41 graduate students, 8 postdoctoral researchers, 14 undergraduates, and 19 outside collaborators. Though most users are from Dedman and Lyle, there are also users from Cox and Meadows. The system averages about one million core hours (or 114 years) of computation per month.

Below is a brief description of current research and teaching at SMU which is dependent on HPC or more generally on computing technology.

Biology

The current HPC computational resources at SMU have enabled John Wise and Pia Vogel to initiate a new research program aimed at resolving failed cancer chemotherapies. It has long been known that about 40% of chemotherapies fail due to the action of one specific protein in our bodies. For over 30 years, science has sought drug-like inhibitors for this protein so that patients with failed chemotherapies might have a final chance at defeating their disease. Since its inception, Vogel and Wise have used nearly 9 million hours of computation on the SMU HPC to screen over 8 million potential drugs. The best lead found so far inhibits the target protein by over 80%. This National Institutes of Health funded research would not have been possible and certainly would not have been externally funded without the resources available at the SMU HPC. An additional drug discovery project aimed at finding new antibiotics targeted to multidrug resistant bacterial pathogens has recently been initiated that also critically depends on the resources of the SMU HPC.

In addition to the direct advances of the research programs of Wise and Vogel into new drug discoveries at SMU, one outcome of the enabled research was the founding of the inter-disciplinary Center for Drug Discovery, Design and Delivery at Dedman College. This initiative has members from the departments of chemistry and biology working with their respective research specialties to further new drug development at SMU. Among the expertise of its members are drug discovery and optimizations which are being performed in both departments using HPC resources. These fundamentally predictive computational studies are now being combined with biochemical studies on target proteins, effectiveness and toxicity studies using cell-based assays, and the synthesis and use of new drug delivery vehicles by various members of the Center. Without the SMU HPC, a fundamental part of the synergistic whole would be missing. Arguably, this new Center in Dedman College would not have been started without the HPC and the research derived from it.

Four graduate students and two postdocs in Biology have carried out research enabled by the SMU HPC. In addition seven SMU undergraduate students are currently involved in some aspects of drug discovery research in the Vogel-Wise research group. Their research is ultimately derived from calculations performed on the SMU HPC that allow us to reasonably predict whether or not a given chemical may act to inhibit the problem protein. This research includes the purification of the target protein that causes the human chemotherapy failure from living cells and the biochemical testing of predicted inhibitor drugs on the isolated protein. We will expand this in the near future to cell based effectiveness and toxicity tests that could also be performed with undergraduates. It is worth noting that many of these students aspire to enter the medical field. Clearly taking part in this project exemplifies Engaged Learning. These students are thrilled to have a chance to make a difference and the SMU HPC has been instrumental in giving us this opportunity.

Chemistry

The CATCO group, led by Professors Dieter Cremer and Elfi Kraka, has four major research thrusts, all of which are dependent on HPC.

The effective and permanent removal of pollutants such as arsenic (from drinking water) or heavy metals such as mercury from the environment requires new chemical strategies. In collaboration with experimental chemists organic tweezers molecules are designed that selectively bind a contaminant such as arsenic or mercury and then precipitate in aqueous solution so that the precipitate can be mechanically removed. The design of the tweezers molecules requires hundreds of calculation with quantum chemical methods and computer programs that have been developed at SMU in the CATCO group.

The successful synthesis of new materials requires the control of chemical reactions. This control can only be obtained if the understanding of the mechanism of chemical reactions would be improved by a factor 10^6. A step in this direction has been accomplished by the Unified Reaction Valley Approach (URVA) of Kraka and Cremer that provides for the first time the needed insight into the mechanism by a stepwise monitoring of reaction path direction and curvature.

The huge variation of chemical bonds from very weak to very strong is the prerequisite for the huge manifold of different chermical materials. We have derived from the normal vibrational modes of a molecule local vibrational modes which can describe any bonding interaction in a molecule or molecular complex. This approach is based on accurate quantum chemical calculations of the molecular vibrations and is used to determine the bond strength of hundreds of different chemical bonds. In this way H-bonds, dihydrogen bonds, halogen bonds, agostic bonds, and especially transition metal bonds are investigated.

Microorganisms produce enediynes to destroy the cells of viruses and bacteria. There have been attempts to use the biological activity of enediynes for medical purposes and to develop more powerful anticancer drugs. All these attempts have been unsuccessful because enediynes attack also normal cells and accordingly they are highly toxic. Our approach is based on the Kraka-Cremer concept to modify the active part of naturally occurring enediynes in such a way that they become only active in cancer cells. All modifications are tested first on the computer with the help of quantum chemical investigations.

Currently five Ph.D. students are involved in this research. In addition undergraduate research plays a key role in the education of our students. The students are exposed to real research problems and thus they more easily develop an understanding for the importance of knowledge and the possibilities of applying this knowledge. This has been shown by two undergraduate research projects which have been completed in Computational Chemistry.

Economics

The Economics Department has a very active MA in Applied Economics program which has a predictive analytics track. Students take courses at the confluence of economics, computing, and machine learning that typically work with very large data sets. These courses cover the full gamut of predictive analytical techniques and require a wide variety of software products.

We currently have forty students in the Applied MA program, approximately half of them focusing on predictive analytics. We have had a very good placement rate for the predictive analytics MA graduates over the last several years. Our Ph.D. students are also heavy users of computing techniques, including simulations on Dynamic Stochastic General Equilibrium (DSGE) models, and large state space models.

The Economics Department is beginning a collaboration with IBM and a local alcohol beverages distribution company, Andrews Distributing, as part of IBM's SMART program. This program is buttressed by our Master Internship course that results in participating students receiving credit for completing the SMART program internship with a local company. The internship program invariably involves the use of predictive analytics tools found in IBM's SPSS Modeler program.

Mathematics

The computational mathematics group currently consists of five faculty and eleven doctoral students, with plans for expansion, and most other members of the department use computing at some level in their research. Research is focused on the solution of large scale models of physical phenomena, and requires tightly-coupled clusters. As SMU's tightly-coupled cluster has just come on line, most work requiring HPC has been performed at NSF or DOE supercomputing centers.

Teams including SMU Mathematics Professor Dan Reynolds have developed simulations of complex multiscale magnetohydrodynamic flows, with applications in cosmology and to the simulation of fusion reactors. In collaboration with researchers at nine other universities, he has worked on ENZO, a community-developed code for astrophysical applications, and used implementations on 40,000 cores on the Oak Ridge machine to simulate reionization in the early universe. See Figure 3 for a visualization of these computations.

Other HPC projects include direct simulations of turbulence mixing noise in jets, electromagnetic scattering for military applications, the fluid dynamics of insect flight, and the solution of density function models of materials.

Two current undergraduate research students in mathematics are doing computational work: one making use of facilities at TACC. We have also used TACC facilities to teach a graduate course in HPC.

Physics

Since each project on ATLAS produces data formats specific to their needs, this often means that each researcher has to download their own special versions of the data in order to conduct their research. In 2011, these data sets were each typically 20 TB in size, with 4-5 independent projects ongoing at any time within the SMU ATLAS group. This means that 100 TB was required just to meet the needs and demands of this high-profile research. Increasing storage capacity and processing power allows SMU physicists to stay ahead of the growth of these data sets. The research is highly competitive on an international scale, and being first (as well as best) matters. Access to substantial computing resources ensures SMU ATLAS physicists can compete effectively.

As an example of the invaluable nature of a local HPC computing system, several of the analyses done by SMU ATLAS physicists on 2011 data were done entirely at SMU because the turn-around time to rerun the entire analysis was hours to a day; using the GRID, turnaround times varied from a day to days, simply because user priority is volatile on a globally distributed system.

A comparable situation occurs for SMU NOvA physicists. Although the data sets for this experiment are smaller than those for ATLAS, since neutrinos interact feebly with matter computational requirements are similar. Indeed, because the effects looked for in so-called neutrino oscillation

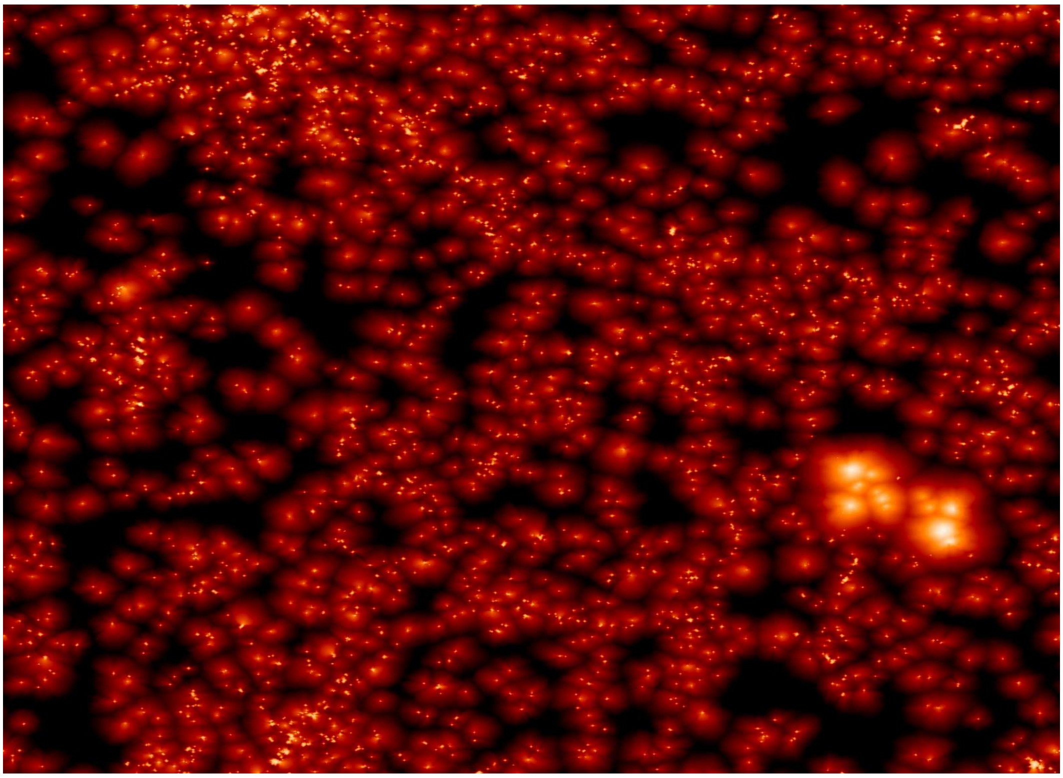

Figure 3: Radiation energy density emanating from early galaxy clusters as simulated by ENZO. The calulation used 40,000 processors on a supercomputer at the National Institute for Computational Sciences in Tennessee.

experiments are subtle, it is necessary to run large and extensive simulations to ensure that the neutrino interactions sought are not mimicked by a far less interesting interaction.

In the past two years, the SMU HPC cluster has been extremely useful to the physics department theory group. Four out of the 5 members of our team (Professor Pavel Nadolsky, post-docs Marco Guzzi and Jun Gao, plus graduate student Zhihua Liang) use SMU HPC on a continuous basis. With the influx of high-precision LHC and Tevatron experimental data, even basic calculations must be carried out at a higher level of accuracy, which means more resource-extensive computations. Every week, this group submits 10-50 daylong jobs on SMU HPC to fit “CTEQ NNLO parton distribution” functions and perform Monte-Carlo integration of multi-loop scattering cross sections. These calculations are used, for example, for benchmarking analyses of key LHC measurements (jet production, W & Z production, etc.). Given the large number of the scientific groups that participate in such analyses, it helps our group greatly to have access to a fast and reliable high-performance cluster that allows us to keep the competitive edge.

Seven SMU physics Ph.D. students and six post-docs have already used extensively the SMU cluster for their research.

Electrical Engineering

Electrical Engineering Professor Marc Christensen is the leader of a DARPA-funded University Focused Research Center in Neurophotonics. The Center's goal is the rapid development and prototyping of a novel all-optical photonic sensor capable of directly detecting the action potentials of a neuron or nerve cell. To model these sensors, multiple techniques are needed that span the mechanical and electromagnetic regimes. Modeling and simulations have been performed on SMU's current cluster. These simulations required the full computational resources available and individual runs took several days to complete. A Ph.D. student, Lively, has developed an implementation on GPUs, which has shown promise in reducing the turnaraound time.

Two undergraduate students have also worked with Professor Nathan Huntoon on HPC. Both students were able to contribute to exisiting research programs by developing and running simulations on HPC resources at SMU. The simulations were performing calculations that the students did not have a full understanding of, but the analysis of results led to learning opportunities that would not have presented themselves otherwise.

Both undergraduates have since continued on to work towards higher degrees in EE, one at USC and the other here at SMU, and continue to use HPC in their research.

What would expanded high-performance computing enable at SMU?

Expanded high-performance computing at SMU would create new opportunities in research, graduate and undergraduate education, and engagement with the community. Our focus below is on specific programs within departments as well as more general educational efforts.

Enhancing research and graduate education

Biology

The present state of operations of the SMU HPC has enabled the group led by Professors John Wise and Pia Vogel to carry out single searches for new drug candidates for important problems in chemotherapy; research that otherwise would not have been possible. These searches for new drugs occur in two main phases: (1) The exploration of the structure and dynamics of the actual intended target of the drug, which is in almost all cases a protein structure, and (2) the systematic and massively parallel search for drug candidates that effectively interact with the drug target. Both of these phases of research represent some of the most computationally intensive calculations that are performed today on any high performance instrument.

The first phase of our efforts, the elucidation of the dynamics with which these protein structures move in their natural "nano-scale" environments by computer simulations, marks one area in which expanded HPC capabilities at SMU would greatly accelerate the efforts of ongoing research projects and enable new projects and experiments to be initiated. With the present instrumentation, wait times and calculation times for these simulations are rate-limiting. Expansion of the resources would enable a fuller exploration of the chemotherapeutic drug target dynamics that would greatly accelerate the research progress. This acceleration would be approximately proportional to the expansion of the tightly-coupled cluster. Currently, supplemental computation has been performed on TACC and local GPU-processor machinery, but both of these sources do not offer the access that expanded SMU HPC would provide.

The second phase of the project, the massively parallel search for new effective drug candidates would be accelerated linearly with the increased capacities of the SMU HPC. These studies typically run whenever any resource on the HPC would otherwise be idle. The number of potential drug interaction simulations we have pending for these searches runs into the tens of millions per individual experiment, with each simulation taking up to one hour of core time. This means that any expansion of the HPC capacity will allow more experiments to be performed. It should be noted that any increased capacity in the SMU HPC will not be wasted, since our drug discovery projects will always use available computational cycles. If the capacity is there, we will use it.

In a pragmatic light, the added capabilities of the SMU HPC would allow better than proposed performance on our NIH-funded drug discovery project, which would clearly be positive for future granting success. In a more fundamental way, however, better investigation of the drug targets will undoubtedly increase our chances of finding that "one in 100 million" compound that might solve this fundamentally critical human health problem in chemotherapy. Significant expansion of the SMU HPC would allow us to pursue additional projects such as our preliminary investigation for new classes of antibiotics to combat multidrug resistances in bacterial pathogens as well as our chemotherapeutic work.

Chemistry

CATCO's research projects mentioned earlier would all benefit from access to more computational power. Computational quantum chemistry is most efficiently carried out using shared-memory architecture, which will be greatly expanded as part of the proposed purchase.

Removal of heavy metal, arsenic, and other contaminants from the polluted environment An increase of the available computational power by a factor 1000 would make an extension of the investigation to 10 different contaminants possible where now just two are investigated. (Note that the investigation of each additional contaminant requires testing of the new tweezers molecule against all previous contaminants)

Generation of a chemical reaction library for transition metal catalysts For the investigation of 160 catalytic reactions using URVA about 200,000 hours of computer time are required, which will be the prerequisite for the design of powerful transition metal catalysts in chemical synthesis. Such a computer load is not available currently, but would become possible and would even lead to an increase of the reactions studied if the calculational efficiency would increase by a factor of 10^3.

Description of unusual chemical bonds An investigation of the 60 most common H-bonds required about 500 cpu hours. Considering that a set of about 20,000 different chemical bonds leads to the most common chemical compounds in inorganic, organic, and biochemistry, it becomes clear that a description of these bonds by the current approach requires an increase of the currently available computer power by a factor of $10^3 - 10^4$.

Computer-assisted drug design of an enediyne based antitumor drug The design of new enediyne-based antitumor drug requires about 4000 cpu-hours. An increase in computational power would make it possible to explore drug modifications in a way that would reduce drug development time from now about 12-14 years to half of this time.

Mathematics

As mentioned above, research of the computational mathematics group, with funding from NSF, DOE, ARO, and the Israel-US BSF, is focused directly on the development and application of algorithms to exploit modern HPC architectures. A sampling of projects which would greatly benefit from an enhanced local capability includes:

Development of advanced algorithms for simulating wave propagation This work is carried out by Prof. Thomas Hagstrom in collaboration with researchers in Mechanical Engineering at Caltech, Civil Engineering at Carnegie-Mellon, Aerospace Engineering at the Israel Institute of Technology (Technion), and Hypercomp Inc. Current applications include the simulation of broadband radar, turbulence mixing noise from jet engines, and seismic waves, with new work on focused ultrasound for medical applications in the planning stages;

Efficient solvers for the time evolution of multiscale systems This work, led by Prof. Dan Reynolds, includes the development of enhanced methods for the treatment of radiation transport and chemical ionization within simulations of cosmological reionization, which form a core module in the open-source *Enzo* code (http://enzo.googlecode.com). Research into general integration methods for multi-rate simulations is being developed into new components within the *SUNDIALS* suite of portable solvers for time-dependent, nonlinear and linear systems of equations (https://computation.llnl.gov/casc/sundials).

Simulation of fluid flows with moving surfaces In collaboration with researchers at Cornell University, Prof. Sheng Xu uses simulation to understand how insects attain flight stability and maneuver with great precision. With the enhanced facility, high-fidelity simulation of the free flight of a virtual dragonfly will be in reach.

Efficient and robust algorithms for eigen-related problems Eigen-related problems arise in diverse applications ranging from data mining to materials science. Prof. Yunkai Zhou's research is focused on the development of robust but scalable algorithms for solving such problems. His work is an essential component of the PARSEC software, which is being used by dozens of research groups in materials science around the world. Current research is centered on making the algorithms run faster and on treating new applications.

In many of these cases, besides the direct applications being addressed, the goal is to release open-source software for use by the scientific community. Although, as mentioned above, many in the group already use NSF and DOE supercomputer centers for large scale simulations, our progress is held back by the lack of a large, modern, on-campus tightly-coupled cluster. The group recently purchased a 384-core machine, but to produce good software we must test and optimize algorithm performance on a range of configurations. In addition we need access to state-of-the-art accelerators such as GPUs or Intel's planned MIC, which will be an important part of next-generation supercomputers at TACC and elsewhere. The proposed purchases would allow routine use of cutting edge hardware, significantly accelerating our work.

As the research of all of our graduate students is similarly focused on algorithm development, the proposed purchases will allow us to provide them a world-class educational program in computational mathematics. At present, due to our limited access to appropriate hardware, we only offer one graduate course in high-performance computing, and don't make use of HPC in the rest of our extensive sequence of courses in computational mathematics. This would change if a large local facility were available. Lastly, there is now often a considerable delay between the time students begin their thesis work, and when they start to use HPC at national facilities. With a good local machine and expert staff, this delay can be eliminated.

Physics

A nascent collaboration between SMU physics faculty and the Stanford Linear Accelerator Center is to simulate crystal lattice vibrations (phonons) in the detector crystals used for the SuperCDMS experiment that is looking for dark matter in a deep underground mine. This type of matter is far more prevalent in our universe than the type of matter that comprises stars, moons and people. Understanding these vibrations is important as a way to distinguish a real dark matter signal in the experiment's detectors from a fake one caused by a myriad of mundane sources that have nothing to do with dark matter particles. Substantially larger computing resources, of the scale envisioned for SMU's HPC center, make the crystal studies possible in a timely fashion as well as more comprehensive scientifically.

The ATLAS opto-electronics group in the physics department uses the current cluster to simulate the behavior of three integrated "serializer" circuits for the readout of the upgraded ATLAS detector at CERN in Geneva, Switzerland. A wide variety of operating modes are simulated with an eye toward optimizing performance before the design is turned into actual chips. Simulation of an entire chip, although highly desirable, is not practical at SMU's current computing cluster because it is much too small, so simulations take much too long. A substantially larger facility would permit simultaneous simulation of the entire chip with a concomitant increase in the chip's performance and reliability, since subtle, correlated effects among different parts of the integrated circuit could be examined.

The golden objective objective for the SMU physics theory group in next few years is to apply the Monte-Carlo sampling method to the analysis of "CTEQ parton distribution" functions, by following an approach that is similar to the method developed by the Neural Network PDF analysis group. Understanding these functions is necessary to understand the proton-proton collisions at the LHC collider in Geneva. We have a well-defined work plan for implementation of the Monte-Carlo sampling of CTEQ parton distribution functions and this computation is well suited for parallel computations on a cluster of the SMU HPC-type. The usage of the cluster will increase substantially when this project ramps up to full steam.

The Ph.D. thesis is the capstone educational experience of a physics graduate student. Extensive computing (writing of software and practical utilization of hardware) is an important component of Ph.D.-related research in the physics department, somewhat independent of the actual Ph.D. topic. Direct experience with a high performance computing center and interaction with the technical staff that runs it will enhance markedly the quality of the graduate student's research expertise as the student will learn the practical aspects of computing on a large machine.

Statistics

The statistics department is on the cutting edge of data analysis and modeling. One of the encouraging trends in Statistics has been our increasing engagement with front-line science. As a result, we are dealing with extremely large datasets and more complicated and computationally demanding statistical models and algorithms, which all rely heavily on high performance computing.

Currently, many active research areas of our stat faculty have demanding computational requirements, including Bayesian statistics, complex hierarchical modeling, spatial modeling and optimization, model selection, statistical learning and data mining, Monte Carlo methods, resampling methods, simulation methods, etc. More importantly, our faculty members have been actively involved in multidisciplinary research projects, all involving large, complex, multi-dimensional datasets. Below are several examples of such projects where we have collaborated intensively with UT Southwestern researchers.

- Richard Gunst, William Schucany, and Wayne Woodward have worked for several years developing statistical methodology for analyzing large collections of brain imaging data. This work has been done with Dr. Robert Haley (epidemiologist in the Department of Internal Medicine) in support of a major study of Gulf War syndrome involving veterans from the 1991 Gulf War. This work has produced major statistical advances in the complex problem of detecting brain activation and differences in brain activation between impaired Gulf

War veterans and normal controls. Collaboration is continuing with colleagues at UTSW on better characterizations of the differences in brain activation using advanced parametric and nonparametric statistical modeling approaches.

- Sherry Wang and her Ph.D. students have been collaborating with the Quantitative Biomedical Research Center, Department of Clinical Sciences, UTSW, on statistical methodological development in Biostatistics and Bioinformatics. The research effort involves several NIH and NSF funded projects for development of Bayesian hierarchical methods for modeling three-dimensional chromosomal structures, Bayesian spatial modeling for efficient analysis of high-throughput data, and integrative Bayesian analysis for exploring molecular mechanisms of cocaine addiction. The research has produced novel statistical methods to study changes of chromosome structures in response to experimental conditions, and enriched Bayesian spatial modeling for high-density and high-volume data. Further, it has a direct impact on cutting-edge biomedical research, by improving the efficiency of high-throughput data analysis.

- Monnie McGee has been working with Dr. Richard Scheuermann (John H. Childers, Director, Division of Translational Pathology and Division of Biomedical Informatics, UTSW) to develop methods for determining whether genes are significantly overrepresented in biological pathways of interest. Dr. McGee is also working on the development of methods for preprocessing microarray data. Along with the development of new methods for preprocessing of gene expression microarray data, Dr. McGee and Dr. Scheuermann have also been involved in comparative assessment of the performance of new and existing methods using real microarray data. Other future projects include statistical methods for assembly of fragments of gene sequences from next generation sequencing technology, and genetic linkage analysis considering space and time of genetic migration, with application to influenza virus.

Though different in statistical approaches and application areas, the above research projects have all generated significant impacts in advancing knowledge discoveries in the sciences, and received funding from the NSF, NIH, VA, and DoD. They also share a common feature; all deal with extremely high dimensional, and hence storage and memory intensive, datasets. For example, one mRNA sequencing sample that consists of about 60 million reads of length 100 base pairs will take 15GB disk space to store, and to efficiently process this single sample, we need to load the data into memory once. Thus the minimum memory requirement is greater than 24GB, and it takes over 12 hours using a 12-core server just to finish the initial preprocessing steps.

Continuing research in these areas, and others, involve complex statistical models and/or advanced computational procedures that push, and sometimes exceed, the current computing resources of SMU. Undoubtedly, high performance computing is key to successes in such high-impact research projects, by alleviating the computational boundaries and limitations and improving our productivity.

Computer Science and Engineering

The Computer Science and Engineering Department in the Lyle School of Engineering will be one of the biggest beneficiaries of the SMU HPC Initiative. In addition to supporting several research projects that require high-end computing, there are multiple research projects, led by Professors Suku Nair, Mitch Thornton, Maggie Dunham, and Michael Hahsler, that investigate reliability, security and performance of high-speed computing systems. Further, the department offers several courses on Data Mining, Web Analytics, Information Retrieval, Cyber and Information Security, and Graphics that have projects requiring the use of HPC systems.

- The High Assurance Computing and Networking (HACNet) Lab, founded in 1996 and backed by more than 20 years of research in security, reliability, and systems engineering is SMU's premier security laboratory for education and research on information security, secure communications, disaster tolerance and healthcare safety and security. It is a multidisciplinary Lab, established as part of the Department of CSE and designated as a DHS/NSA Center of Excellence in Information Assurance Education. Currently there are 14 Ph.D. students and several M.S. and undergraduate students in HACNet conducting research under the guidance of 5 full-time faculty members. In the Lab, there is heavy use of HPC systems in projects that deal with Cyber Analytics, Automated Intel Systems, and Large Scale Intrusion Detection and Prevention Systems.
- The National Institutes of Health (NIH) have identified Sequence Analysis as the number one challenge for the genome research community. The Intelligent Data Analysis (IDA@CSE) Lab in the CSE department at SMU has taken up the challenge and is designing a new approach to analyze and compare DNA/RNA sequences. The approach is based on an extension of the Extensible Markov Model (EMM) which can be used to model the distribution of frequencies of sub-patterns within a DNA/RNA strand. Given a set of sequences belonging to a known class, an EMM is constructed that models that class. The EMM can then be used to predict the membership of a test sequence to that class. Multiple EMMs can be created, one for each sub-pattern length, and then used as the basis for a meta-classification (MCM). The IDA group has developed an online tool for this purpose (http://lyle.smu.edu/ida). As this approach is quite CPU intensive for longer sequences and sub-patterns, the ability to use the proposed HPC would greatly benefit their research.

Civil and Environmental Engineering

Professor Usama El Shamy studies geotechnical systems involving the complex interaction of elements such as the porous medium, composed of soil grains and pore-fluids, with a structural system such as a foundation or a retaining structure. He is particularly interested in the behavior of such systems subjected to extreme loading conditions, such as those encountered during earthquakes or hurricanes.

The availability of an HPC system will enable the next generation of model-based and scenario-based simulations at unprecedented resolution, thereby providing a comprehensive understanding of failure mechanisms during several flooding scenarios and geotechnical system characteristics. Moreover, it will provide an opportunity to effectively model the timescale of the physical processes leading to failure, which may be hours in the case of a hurricane, or days in the case of an intense rainstorm.

Electrical Engineering

Enhanced HPC capabilities will greatly accelerate work at the University Focused Research Center in Neurophotonics. As mentioned above, the Center's goal is the rapid development and prototyping of a novel all-optical photonic sensor capable of directly detecting the action potentials of a neuron or nerve cell. Such optical systems combine large spatial domains with high resolution and thus challenge even modern supercomputers. To accurately model the this sensor's design requires even higher than normal resolution, which the proposed increases in computational resources will enable.

Formal programs

As noted above, many of SMU's existing graduate degree programs have significant computational components, including M.S. and Ph.D. programs in Computational and Applied Mathematics and Computer Science and Engineering, as well as the predictive analytics track in the Applied Economics Master's program. In addition, as offered at institutions such as Brown (www.brown.edu/Departments/CCV/hpc-cert) and Texas A&M (isc.tamu.edu/research-education/CSCP/), the potential for a new certificate program, open to both graduate and undergraduate students, will be assessed. Such a program would serve both full-time students at SMU and professionals employed in local industry.

Undergraduate education

Enhanced HPC capabilities will lead to substantially expanded research possibilities for undergraduate students. We believe that computing provides a unique platform for involving students in internationally-relevant research problems at an early stage of their career. Although the examples cited above show that such research activities are underway at SMU, through easier access and better support we can make them the rule rather than the exception. The execution of the initiatives discussed here will allow SMU to promise every prospective student that they will be able to use state-of-the-art computational facilities to pursue their research interests, and they will receive the training and support necessary to use them. Thus HPC will be an integral component for the development of the Unbridled Projects component of the Engaged Learning initiative as well as projects which are part of the Big Ideas program.

In support of undergraduate education, we would offer an intensive summer training program, where students would learn from expert faculty and proessional staff, getting their feet wet with the possibilities of HPC. These students would typically be juniors, and will have completed an undergraduate course preparing them for the summer session. For most students we expect this will be the joint Math/CS course in scientific computing, which has a current annual enrollment of 144, but which is being expanded to an annual enrollment of 192 to meet demand. Math and CS will review the content of this course and others in the undergraduate sequence to make certain that students are well-prepared. Other departments such as Physics offer comparable undergraduate courses which will also serve as a prerequisite.

Concerning formal programs, we note that a double major in the sciences, engineering or economics with Mathematics, concentrating in computational mathematics, is an option today. This program will be reviewed for possible enhancements in conjunction with growth in HPC on campus. (For an example of an interdisciplinary undergraduate degree program in computational science, see Stanford's, which has engineering, biology as well as management-oriented tracks (www.stanford.edu/group/mathcompsci/).) As mentioned above, a certificate program would also be open to undergraduates. Such programs exist, for example, at UT Austin (www.ices.texas.edu/programs/cse-certificate/) and Princeton (www.pacm.princeton.edu).

Finally, undergraduate courses can benefit from HPC on an occasional basis through the delivery of simulations to the classroom. See, for example, the NCSA's cybereducation initiatives (education.ncsa.illinois.edu). The center's staff will work with interested SMU faculty to enable both the use of HPC for class projects and the development of classroom presentations based on realitstic simulations.

Community outreach

Expanded activities in HPC will lead to new outreach programs in at least three areas.

Area Universities There are a number of Universities and Community Colleges in the Metroplex serving disadavantaged communities. SMU's HPC center will engage faculty and students at these institutions to explore possible collaborations. Examples could range from access to simulations for the classroom to student research.

Precollege Education Researchers currently associated with the CSC have already participated in programs exposing high school students to computational science. These include programs in Chemistry mentioned earlier, as well as others carried out in collaboration with the Lyle School's extensive outreach programs. The proposed initiatives will enable the expansion of such activities. In addition, the NCSA has already developed simulations for delivery to undergraduate classrooms and have archived various instructional materials. We will engage local educators to see if we can provide useful services of this type.

Industry A primary interaction with local industry will be the training of their employees, either through our degree programs or through specialized programs such as a certificate. In addition, as the center's staff grows, one-day workshops could be offered and made open to the public on a fee basis. The model of Rice's Ken Kennedy Institute shows that more extensive ties are possible, through industrial partnership programs, sponsored graduate fellowships, and industry-specific workshops. Such programs will be explored as the center develops.

Why not rely on TACC at UT?

The facilities at TACC are a tremendous resource for experienced computational scientists, and a number of SMU researchers already use TACC machines as well as supercomputers at the National Center for Supercomputing Applications (NCSA - Illinois), the National Institute for Computational Sciences (NICS - Tennessee), and the National Energy Research Scientific Computing Center (NERSC - California). However, these centers cannot substitute for a robust local facility.

To use TACC, potential users are required to submit proposals for competitive review during one of four annual submission intervals. Winners are notified two months after the close of each interval. Successful proposals, as identified by TACC, should include, among other items, preliminary results previously obtained at smaller scale computing facilities. There are online tutorials explaining how the researcher interfaces with TACC but these are completely technical. Consistent with its mission as a research tool, TACC has no provision for teaching programming or teaching core principles of writing simulation software.

Below is a summary of critical tasks which require a local facility.

- Training of students in this strategically critical area for the cyber-age is a fundamental mission for the University. As such, undergraduate and graduate education must be a focus of our developments in cyberinfrastructure. National centers such as TACC are not accessible to inexperienced users. Baldly speaking, "a student doesn't do homework on TACC" since homework is not cutting edge research. Even an undergraduate participating in SMU's undergraduate research programs would not gain access to TACC resources since they would not realistically be able to submit a competitive proposal. Integrating high performance computing into coursework and responding to myriad issues associated with classroom instruction requires flexibility. A local facility provides that flexibility. Moreover, training in advanced computing requires intensive interaction with both the machine and with expert staff. The developments proposed here will enable such intensive interactions, and can help make SMU a leader in HPC education and research among its peers.

- As high-performance computing plays a more important role not only in traditional fields in science and engineering, but also in the humanities, social sciences, business, and the arts, the need for training extends to faculty and students throughout these disciplines, creating even greater demand for local infrastructure and support.

- A local facility will help SMU recruit outstanding faculty and students interested in the application of HPC to challenging problems. Reliance on TACC will not.

- As a small University, interdisciplinary collaborations are crucial to the development of internationally-recognized research programs. A comprehensive interdisciplinary local center can catalyze such collaborations. The sum of a disjoint set of computational scientists relying on outside resources will mean much less to SMU than a collaborative whole.

- Similarly, SMU researchers need to be involved in larger scale national and international collaborations part of "virtual organizations" as discussed in the NSF Campus Bridging report [7]. Local resources will better enable our participation in such organizations work on our current cluster as part of the ATLAS project provides an excellent example.

- Relying on TACC, SMU cannot become the focal point for computational research and education in the Metroplex. With a robust local capability, we will be able to reach out to industry, schools, community colleges, and historically black and hispanics-serving institutions.

- Although researchers carrying out medium-scale simulations (relative to "grand-challenge problems) can access sufficient resources at national centers, preferred scheduling to simulations at the largest scale can result in throughput times for medium-scale jobs which are significantly higher than could be attained with a local facility.

- Even for users whose simulations are primarily completed at national facilities, the crucial step of data analysis often requires direct interaction with huge volumes of data, perhaps requiring compute-intensive analysis and visualization. Limitations to network bandwidth necessitate that such analyses be performed locally.

Why not rely on the cloud?

At this time commercial cloud services are not a competitive option for research and education in high-performance computing. The points in the discussion of why NSF/DOE facilities such as TACC are insufficient to meet our goals all apply to commercial cloud services. But in addition, many cloud services are not specialized to HPC applications, do not have user portals and instructional aids at the same level as the national centers, and have direct costs to the University and to individual researchers' grants. Moreover, computing services as direct costs may not be supported by many federal agencies, and will be impossible to predict when developing research budgets. Lastly, even if these barriers are overcome, a local facility which is almost fully utilized is more cost-effective. In a recent NSF-sponsored workshop on campus computing facilities, V. Agarwala, the Senior Director for Research Computing and Cyberinfrastructure at Penn State, estimated the total of cost (power, cooling, system support, and hardware refresh) per core hour of a fully utilized on-campus facility to be approximately \$.025-\$.04, which compares favorably with the prices charged by the Amazon Elastic Compute Cloud [8, 9]. Given our experience with the existing HPC facility on campus and our extensive plans for further development of computationally-intensive programs in research and education, we are confident that utilization of the hardware will exceed the point where cloud services are a cost-saver.

How would high-performance computing interface with resources at TACC and elsewhere and further inter-institutional collaboration?

With the training and experience gained at SMU, faculty and students with large-scale applications will be well-positioned to exploit national facilities, such as those provided by NSFs XSEDE network, which encompasses 16 supercomputers and data analysis resources nationwide, as well as by the DOE and DOD. Our proximity to TACC is a great benefit in this regard, as users can participate in hands-on workshops onsite. However, many projects will be better served by the resources at other centers, and advanced collaboration facilities to be supported by this proposed investment will enable direct participation by the SMU community in the workshops they provide.

A point of emphasis in the report of the NSF Task Force on Campus Bridging [7] is the importance of supporting virtual organizations of researchers across many institutions and across the world. The proposed investment will enhance participation in virtual organizations at SMU, which will develop according to the interests of the faculty and students, unlimited by disciplinary or geographical boundaries. Within North Texas, a number of possibilities exist for enhanced collaborations built around computational research. These include joint efforts in computational chemistry and the digital humanities with groups at UNT, as well as collaborations with experimentalists at UT Dallas and UT Southwestern.

We note that there is a great deal of contemporary effort in the application of HPC to the health sciences and biology (see, e.g., the University of Rochester's Health Sciences Center for Computational Innovation and Carnegie-Mellon's Lane Center for Computational Biology). High performance computing would greatly facilitate the long-desired establishment of an inter-institutional graduate program in Biostatistics offered jointly through the Department of Statistical Science at SMU and the University of Texas Southwestern (UTSW) Medical Center Department of Clinical Sciences.

With the UTSW Medical Center in close proximity to SMU, the medical researchers from UTSW provide natural research partners for members of our statistics department. Strong collaborations have existed between the two respective departments over years. Establishing a joint program in Biostatistics would solidify this collaborative spirit and provide SMU students with an interdisciplinary research environment, which is vital to their academic growth and future success.

Lastly, following the Rice model, we expect that programs will be developed which are of interest to local industry; at a minimum we expect to attract local students to our postgraduate educational programs.

References

[1] President's Information Technology Advisory Committee, *Computational Science: Ensuring America's Competitiveness*, 2005.

[2] T. Hey, S. Tansley and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009.

[3] www.top500.org

[4] McKinsey Global Institute, *Big data: The next frontier for innovation, competition, and productivity*, 2011.

[5] The Economist *Data, data everywhere: A special report on managing information*, February, 2010.

[6] A. Bifet and E. Frank, *Sentiment Knowledge Discovery in Twitter Streaming Data*, Lecture Notes in Computer Science, vol. 6332/2010, pp. 1-15, 2010.

[7] National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging, Final Report, 2011.

[8] V. Agarwala, *Research Computing and Cyberinfrastructure - The Sustainability Model at Penn State*, NSF Workshop on Sustainable Funding and Business Models for HPC Centers, Cornell, 2010.

[9] aws.amazon.com/ec2/

[10] SMU *Vision Statement for the Dedman College Institute for Interdisciplinary Research and Teaching*, 2011.