Workshop on the Frontiers of Computational Science and Engineering at Caltech (CSE@CIT)

November 18, 2006

Executive Summary

The goal of this workshop, organized and sponsored by the Center for Advanced Computing Research (CACR), was to help develop a strategic vision for further developments of computationally enabled research, Institute-wide. This summary was prepared by Mark Stalzer and George Djorgovski, with input from the speakers and panelists. The workshop website is http://www.cacr.caltech.edu/cse2006/. About 40 people from the Caltech/JPL community attended the workshop. Proceedings will be published electronically in the form of an interactive DVD and a web-cast, that will be linked from the website.

The Workshop Rationale and Motivation: Computation is playing an increasingly important and enabling role in all fields of science and engineering. We see a need to develop a coherent approach, share experiences, and leverage multiple investments already made by many groups. What are some specific actions to help build and support the campus CSE community? What can we do together to assume leadership in some approaches to e-Science? How can CACR best serve the Caltech community as a catalyst, matchmaker, or forum for exchange of ideas and experiences across many different disciplines in CSE? To this effect, the speakers were asked to cover some of the following specific questions:

- Briefly describe their CSE activities,
- What is going well and not so well in their CSE approach,
- What are some outstanding challenges,
- Do they see any potential interactions with other CSE groups on campus, and
- What are the impacts of CSE on Caltech's education mission?

These questions were then addressed in more details in two panel discussions, asking the broader questions: "What are the commonalities among different disciplines?" and "Strategic Trends and Future Directions". Several prominent themes emerged from these discussions, which should guide our strategic and tactical developments in the future:

High-Performance Computing Facilities: This included questions of optimal organization, sharing, and management of such systems campus-wide, and the proper balance between facilities owned by individual groups and consortia of groups (e.g. Division-wide, but also cross-cutting). Common themes are how to propagate information on the best choices and configuration of the hardware and systems software and tools to serve many fields on campus, and where and when we need to participate in the development of these systems and methods to fulfill general or particular needs of science. Access to and participation in larger external and national efforts (e.g. the NSF Petascale computing initiative) is another important consideration.

Multi-Scale, Physics-Based, and Chemistry-Based Simulations: This is one of our major strengths, with projects like ASC, numerical relativity (an essential component of LIGO), geophysics, the Materials and Process Simulation Center (MSC), computational protein design, etc., representing leadership on an international scale. We should build on this strength and transfer the skills and expertise to other fields.

Knowledge Extraction From Massive and Complex Data Sets, Including Visualization: This is becoming a critical need across all disciplines, as the exponential growth of data volumes *and* complexity renders the traditional methods and software tools inadequate for the job, both for experimental measurements and numerical simulations. It represents the intellectual core of the modern scientific exploration process in many forms, from the multi-scale problems mentioned above to the large scale modeling, reconstruction, and extraction of rare discovery signals in high energy physics, leading from

vast quantities of bits into knowledge and understanding of the physical world. Visualization of complex and high dimensional data sets is an especially pressing, universal need.

Design and Implementation of Novel Scientific Research Software Environments: This is another area where Caltech groups are leading the world in their respective fields, including ASC, CMS, DANSE, MSC, NVO, and other data-intensive astronomy projects (e.g. GRIST), TeraGrid science gateways, etc. There is a real opportunity for advances in this arena, through introduction of novel data mining, exploration, and visualization tools, coupled with the already existing data grid environments and massive numerical simulations. Basic design principles and toolkits for construction of such software environments, both locally and for distributed environments will have universal applicability in all fields.

Our Educational Mission in Teaching the Tools and Methods of Computational Science: As the importance of advanced computational and data analysis techniques, and their ongoing development, increases in all fields of science and engineering, we must make a greater effort to educate the next generation of science and technology leaders – now our students – in the understanding and use of these practical and intellectual tools. A general curriculum along these lines should be developed, and then deployed across the Institute in the manner consistent with the educational mission and needs of different Divisions and Options.

In addition, a number of other interesting topics were discussed at least briefly, including these:

- How can we connect better the expertise and efforts in Computer Science (including theoretical ones) at Caltech with the practical needs posed by other scientific and engineering disciplines; for example, in data mining or visualization?
- How can we incorporate into the teaching programs the integration of multiple paradigms and multiple scales that are becoming essential to the coupling of fundamental physics and chemistry to the macroscopic behavior observed in experiments.
- How can we exploit better the advances in fast networking and data transfer, as well as large scale distributed systems for collaboration and grid operations and monitoring pioneered by High Energy Physics working together with Computer Science at Caltech?
- How can we expand further the constructive interactions with JPL and IPAC, in a way consistent with their mission? For example, in the area of intelligent software systems and sensor networks.
- How can the Institute library's understanding of standards, interoperability requirements, digital repositories, and collaboration with other university libraries facilitate innovative projects?

As the next step, we will initiate the following efforts with selected faculty and other participants, in order to better define the paths forward in these areas:

- A task force to summarize the state of the many diverse HPC efforts on campus, and explore the ways of reorganizing and unifying them for greater efficiency.
- An ad hoc group of interested faculty and CSE researchers to outline a path towards an enhanced, Institute-wide curriculum in applied computational science and engineering.
- A working group to explore the ways in which further JPL–Campus collaborations in this general area can be set up, having in mind a long-term strategic positioning of JPL.
- A working group to explore ways that the fundamental work being done at Caltech on information science can be applied to the model representation and knowledge extraction problem.

Other follow-up venues will be developed as a result of these discussions.

We conclude that CACR can play a significant enabling and stimulating role in CSE in the years to come, and serve as a focal point for computationally enabled or data driven science (e-Science) on campus.