Deploying HPC for Interactive Simulation

Panelists:
Roger Smith, CTO, U.S. Army Simulation and Training
Brian Goldiez, Deputy Director, UCF Institute for Simulation and Training
Dave Pratt, Chief Scientist, SAIC Simulation
Robert Lucas, Division Director, USC Information Sciences Institute
Eng Lim Goh, CTO, SGI

Contact:
roger.smith14@us.army.mil

Introduction

The community of academia, industry, and government offices that are leading the development of new interactive simulations for training and analysis are reaching a point at which the application of traditional networks of computing assets are no longer able to support simulation scenarios of sufficient scope, breadth, and fidelity. Several organizations are turning to high performance computing in the form of clusters and shared memory machines to create a flexible computing platform that is powerful enough to run realistic models of military activities and very large scenarios as a matter of course. This BOF will discuss the problem-space and experiments that have been conducted in applying HPC's to this domain.

HPC Application to Interactive Simulation

Army Interactive Simulation (Roger Smith)

The Army training and simulation community, which includes TRADOC and PEO STRI are currently limited in their ability to provide systems and opportunities to units to train from locations that are remote from existing training facilities. These limitations are driven by historical limits on technology and our ability to design systems and training events that can be initiated at the request of a unit that wants to be trained. However, advances in networking, computing, and distribution services have created an opportunity for us to design systems which can be hosted at a powerful central facility, but which can be accessed, configured, and operated by remote units that need to be trained.

We are exploring the ability to configure an HPC as a central server for simulation-based training. A OneSAF On-demand HPC Training Center will be an "always on", network accessible, remotely configurable service for training. It will provide access to training in a manner similar to that delivered by vendors like Sun Microsystems' Grid Compute Utility or Amazon.com's Elastic Compute Cloud. Both of these make hardware available on demand as a service to business customers. Portions of an HPC will be configured to make the hardware available to up to 200 simultaneous units for training, but with the OneSAF software and scenario databases installed. The graphic display of the activities will run as clients at the customer's location. Configuring such a system will require tackling several issues regarding provisioning of machines to specific customers, loading or modifying scenarios for each customer, and providing interactive stimulation of a large number of external client machines. The training organizations listed above have been exploring these issues on a smaller scale for a number of years. The availability of HPC dedicated hardware coincides with the FY07 release of OneSAF 1.0 and the maturing of a number of smaller projects to allow us to take the next step in HPC-enabled, always-on, remotely accessible training.

Physics-Based Environment for Urban Operations (Dave Pratt)

As part of the DoD High Performance Computing Modernization Program's (HPCMP) efforts to provide support for the warfighter and demonstrate the effectiveness of HPC class resources, a Mini-Portfolio has been established to demonstrate the applicability of physics-based modeling in realistic mission planning and scenario analysis. The Mini-Portfolio sets a new direction for DoD M&S by integrating traditional high-fidelity computationally intensive models into operationally relevant scenarios. In doing so, we aim to advance the science by combining the physical, logical, and behavior models that enable us to better understand military-relevant operations and their consequences in context (e.g. IEDs, urban combat, smoke, loss of signal), at high resolution and fidelity. By showing the effects of realistic enhancements to the simulation of operationally relevant urban environments that are made possible through the introduction of first order physics models in to the simulation, we increase both the believability and usefulness of the models and simulations. Improved simulation accuracy is achieved by extending existing simulation architecture to support selected traditional HPC level models. We have demonstrated the relevance and effect of the scientifically valid models within the warfighter context. To date, we have integrated into the simulation context OneSAF the C4I (Scalable Urban Network Simulation (SUNS)) and aerosol particulate transport (CT-Analyst) high fidelity models. The end result of this portfolio will be a system where additional HPC researchers can demonstrate the effects of their computational advances in a warfighter relevant environment.

Critical Questions (Brian Goldiez)

High Performance Computing, characterized by 64 bit word length, MPI, high speed interconnects, large amounts of local and spinning memory, and appropriate operating system features (e.g., load balancing) has typically been reserved for batch processing. Also, HPC machines are typically procured for a specific class of problem that might need large amounts of cache, other memory, CPU cycles, or inter-processor communications. Interactive computing brings a potential new challenge with respect to end-to-end system latency and runtime parameter change where end-to-end implies an input from a user and output to a user in real time (say 30Hz) and parameter change implies changing input variables during runtime.

It is not clear what type of architecture best supports interactivity and allows accurate physical and behavioral representations of ever growing numbers of interacting entities in a virtual environment. More specific issues that need to be addressed include;

1. How can inter-core and inter-node communications be mapped with various types of interactive simulation needs?
2. What strategies exist or should be created to partition various interactive simulations such as models of terrain or computer controlled avatars that interact require interactions with human users?
3. Are special I/O devices and interconnects needed distribute user inputs and integrate system outputs to facilitate interactivity?
4. Are existing operating systems used in HPC's appropriate for interactivity especially where fixed update rates may be needed?
5. How will interactive application scale with various HPC architectures and operating systems?
6. Most batch processing users of HPC build from existing commercially available HPC applications (e.g., MatLab for parallel machines). Are groups working on porting existing interactive applications to HPC platforms? If so, what techniques are being used? If not, how should the process be catalyzed?

Addressing these issues are relevant to the military, homeland security, events where large numbers of crowds of people are expected to interact, massively multi-player game environments, etc.

Opportunities in HPC (Robert Lucas)

There is a continued call for better training, evaluation and analysis, where better means faster, cheaper, more available, and of improved realism and validity. Those of us in the High Performance research community see disparate groups working on issues of common concern and universal utility. Experiments such as Urban Resolve have successfully used Linux clusters to host large ensembles of agents interacting in real time with human users. Now we seek to use forces modeling technology and HPC techniques to enhance both real time analysis of intelligence information and to apply the lessons learned from our simulations to the real World. All of these uses will only be truly effective if interactive HPC becomes a readily available tool. Such use will require a number of adjustments and concessions to implement an interactive environment in a "batch-processing" world. An outline of the challenges facing, as well as the opportunities afforded to, interactive HPC will be presented

HPC Architectures (Eng Lim Goh)

"Today's methods of scientific and engineering investigation range from theoretical, experimental to computational science. In computational science, the classical approach has been modeling and simulation. The concern here is the growing gap between actual applications and peak compute performances. We believe one major solution to this growing performance gap is the new multi-paradigm computing architecture. It tightly integrates, what were previously, disparate computing architectures into a highly scalable single system and, thus, allows them to cooperate on the same data residing in scalable globally-addressable memory. Enabling scientists to focus on science, not computer science.

“Additionally, with globally-addressable memory growing to Terascale sizes, a plethora of new, huge-memory applications that profoundly improve scientific and engineering productivity will come on line. From these, may emerge a new branch of computational science called data intensive methods. It includes the traditional method of query, to the more abstract methods of inference and even interactive data exploration. The availability of such a powerful range of interactive methods, for operation on Terascale data sets, all residing in monolithic globally-addressable memory, is a novel combination that will not only facilitate intended discoveries but may also give rise to a new complement which I will call 'planned serendipity'. The latter will be of growing significance in intelligence, science and engineering. And as the amount of data generated by faster and more productive systems grows, visualization will increasingly become an essential tool. The recent advances in display and related technologies, could pave the way for revolutionary new ways of visual, interactive and collaborative communications.”

“SGI’s focus on memory management stems from us seeing a rising concern from our government customers with the deluge of data they are getting. For various reasons, they are not able to exploit that data effectively. So what we started on is how we could leverage our current architecture to accelerate knowledge discovery.

“What we did was tinker with the idea of putting an entire database in memory. NUMAlink allows multiple nodes to be tied tightly together, so that all the memory pieces are seen as one. Once the processors can see all the memory across all nodes as a single memory, then they can load a large database entirely into that memory. So a complex query that would normally take seconds to return a response—because the disk query takes some time to scan the database—could be returned in under a second. When we went out with the idea, we got enthusiastic responses. We heard how it could fundamentally change the discovery process. When you ask questions with complex queries, you sit and wait for a response. It breaks the thinking process, because you might want to converge on an idea by quickly firing off questions and getting quick responses. You want to have a conversation with the data.”

SC07 Web Site Link:
http://sc07.supercomputing.org/schedule/event_detail.php?evid=11317