Effectively Using a Large GPGPU-Enhanced Linux Cluster

Gene Wagenbreth, Robert F. Lucas and Dan M. Davis
Information Sciences Institute, Univ of So. Calif.
Marina del Rey, California
{genew, rflucas,ddavis}@isi.edu

ABSTRACT

Computing power units per dollar, per kilowatt and per square meter of computer floor footprint may be increased by using heterogeneous computing. The Joint Forces Command (JFCOM) has an urgent and continuing need for large-scale simulations. Both the training and the experimentation directorates must be able to effectively portray the battlespace of the future, often an urban setting with elaborate infrastructure, vast collateral damage possibilities and up to ten million civilians and vehicles. The compute power required is substantial. The authors report on their role in and work with the largest General Purpose Graphics Processing Unit (GPGPU)-enhanced Linux cluster of which they are aware: Joshua at JFCOM, which was awarded as a Dedicated High Performance Computing Project Investment (DHPI) project in 2007. Joshua's 256 nodes are enhanced with an NVIDIA 8800 GPU, each with two 2.33 GHz AMD dual-core Opterons and 16 GB of memory. The authors discuss the theoretical underpinnings that led them to propose such a computer, the process of acquiring it, its installation, early experience, and characterization. They then discuss the creation and their presentation of a course for users and programmers in the new Compute Unified Device Architecture (CUDA), and report on the success of this course. They will give a short precise of the course for those who may be inclined to seek out such and opportunity. They finally compare this programming model with several alternate programming models and compare the ease of programming GPGPUs with that of programming FPGAs and Cell processor chips. In this process the benchmarking and characterization approaches for several types of code are laid out and the results of the experiments is set forth. Several codes were considered, *e.g.* the traditional Linpack, the Multi-Frontal Sparse Matrix Solver, Route Planning algorithms, Line of Sight (LOS) and other agent-based simulation algorithms. The reasons for the final selection of codes for extensive characterization will be discussed. Performance data and optimization techniques used will be laid out in sufficient detail to assist others who are interested in the approach and assessing how effective it may be, if implemented in their environment. Future and expanded uses of the GPGPU acceleration technique and a description of logical programming candidates for this method are also considered in the conclusion section.

ABOUT THE AUTHORS

Dan M. Davis is Director, JESPP Project, Information Sciences Institute (ISI), University of Southern California, and has been active in large-scale distributed simulations for the DoD since 1988 at Caltech, the Maui High Performance Computing Center and at ISI. Prior to that, he was a Software Engineer on the All Source Analysis System project at the Jet Propulsion Laboratory and worked on a classified project at Martin Marietta, Denver. An active duty Marine Cryptologist, he recently retired as a Commander, USNR, Cryptologic Specialty. He received a B.A. and a J.D., both from the University of Colorado in Boulder.

Robert F. Lucas is the Director of the Computational Sciences Division of the University of Southern California's Information Sciences Institute (ISI). There he manages research in computer architecture, VLSI, compilers and other software tools. He has been the principal investigator on the JESPP project since its inception in 2002. Prior to joining ISI, he was the Head of the High Performance Computing Research Department at Lawrence Berkeley National Laboratory, the Deputy Director of DARPA's Information Technology Office, and a member of the research staff of the Institute for Defense Analysis's Center for Computing Sciences. Dr. Lucas received his BS, MS, and PhD degrees in Electrical Engineering from Stanford University in 1980, 1983, and 1988 respectively.

Gene Wagenbreth is a Systems Analyst for Parallel Processing at the Information Sciences Institute at the University of Southern California, doing research in the Computational Sciences Division. Prior positions have included Vice President and Chief Architect of Applied Parallel Research and Lead Programmer of Pacific Sierra Research, where he specialized in tools for distributed and shared memory parallelization of Fortran programs. He has also been active in benchmarking, optimization and porting of software for private industry and government labs. He received a BS in Math/Computer Science from the University of Illinois in 1971