Proposal: CUDA Research Center at LSU

The Center for Computation and Technology (CCT) at Louisiana State University (LSU) proposes to become a CUDA Research Center (CRC) to help leverage its GPU programming efforts in support of existing and planned GPU clusters using NVIDIA technology. CRC@LSU-CCT will support elements of computational science research, education, training, and program development on GPU clusters.

The LSU CCT is in the process of upgrading the University's existing 360 node Linux cluster named Tezpur. The new machine will consist of 440 compute nodes - 50 of which will be GPU enabled with two NVIDIA M2090 GPUs each. LSU CCT has also purchased several 2-8 node GPU (NVIDIA Tesla S1070, C2050, M2050, M2070, M2090) clusters for pilot projects and proof-of-concept research. In addition, a proposal led by PI Honggao Liu to NSF has been recommended for funding to procure a cluster dedicated to development of production research software and the project has an anticipated start date in July 2012. The system, code named Shelob, will deploy 72 NVIDIA Kepler GPUs. These systems will support large-scale, multi-institutional projects together with numerous smaller research efforts whose goal is to develop open source codes, a GPU-enhanced run time environment, and new general programming frameworks to enable new discovery on the next generation of supercomputers. Primary focus will be placed on four projects: (i) the Cactus CaKernel – a general framework for the development of parallel scientific applications which hides the complexity of heterogeneous architectures in algorithmic modules; (ii) advancing the Pluto compiler with extensions to generate effective code for heterogeneous systems; (iii) ParalleX – a new parallel systems paradigm being developed by the STE| |AR group; and (iv) LA-SiGMA – a statewide NSF funded computational materials science project targeting new materials development through computation and simulation.

1 Vision

GPU acceleration is fast becoming the route of choice to exascale computing. The next generation supercomputers will very likely employ several multi-core CPUs and GPUs on each node to achieve the desired performance levels. NVIDIA has already won the hardware battle and established CUDA as the prime candidate for software development on GPUs. Unfortunately, development of an efficient scalable environment on these complex hybrid topology architectures is still a great challenge to the scientific community at large. This has led to a mid-term crisis in software development which can only be resolved by dedicated collaborative effort between domain scientists and computer engineers. Moreover, a reassessment of all requisite components including frameworks, runtime systems, algorithms and applications is critically required. The effort must involve substantial training initiatives at state and national level to develop a new generation of users and developers. To accomplish our goal of providing cutting edge research environments throughout Louisiana, we are keen to partner with NVIDIA. A CUDA research center at LSU would pave the way for the next generation of research software development.

The primary thrust of CCT research is on methods of hiding the complexity of heterogeneous system programming from the research developer. One approach envisions the use of programming frameworks which provide a way to write programs without actually requiring mastery of the low-level language details. The CACTUS framework is an exemplar of this approach, and LSU is lead organization developing thorns (i.e. modules) for numerical general relativity. By developing thorns which implement various algorithms using GPUs, a research will be able to quickly adapt programs from serial to parallel to GPU use simply by selecting the appropriate thorn. A second approach takes a more conventional compiler approach which generates GPU code as part of its native output without adding additional semantics to otherwise normal C code. The more conventional approach of allowing a core development group develop code which is used by other researchers is also supported, the main example being LA-SiGMA, which intends to develop new simulation tools for material design.

2 Quality

LSU and CCT are uniquely qualified to address the challenges of heterogeneous computing. We are involved in the full gamut of essential components including the GPU enhanced next generation run time system (STE| |AR), GPU accelerated code development frameworks (Cactus), and GPU C compilers (Pluto). The

impact of these efforts is greatly enhanced by two statewide virtual graduate research and education project (LA-SiGMA and LONI Institute) which leverage the Louisiana Optical Network Initiative (LONI) to bring together hundreds of researchers throughout the state and the region, creating a critical mass to ensure the success of this effort. We are involved in extensive outreach and education program ranging from boot camp for high school students to international workshops and conferences. Roughly 40-50 computational workshops and tutorials, ranging from high school Beowulf/GPU boot camps to international seminars on GPU computing, are organized at CCT every year. In addition, more than a dozen upper level courses on computational sciences were offered via distance learning throughout the region.

3 Impact

The CRC@LSU-CCT, which will be unique to the south eastern region, will have a dramatic impact on the future of GPU computing as well as the education and economic development in the area. With an extensive research and outreach infrastructure already in place, we can catapult the use of CUDA in various fields ranging from physics, chemistry, material sciences and many engineering domains. The impact of our CUDA based research work could be further expanded via both national and international collaborations. On the economic development side, research and education on GPU programming is in line with the Louisiana Digital Media and Software Incentive. The incentive has already attracted several world's leading digital media companies including Electronic Arts Inc., whose North American quality assurance and testing center is in LSU's South Campus complex, and well known visual effects companies such as Pixomondo whose business operation will be opening here in Baton Rouge. The CRC@LSU-CCT will be helpful in establishing industry partnerships with digital media, visual effects, and software companies to work on CUDA based projects and prepare students for the ever challenging and exciting digital media and video game design business.

4 Ongoing Research

4.1 The Cactus CaKernel

Cactus [1, 2] is an international effort developing a general framework for programming parallel scientific applications. Using the framework approach, methodologies and best practices are captured in modules called thorns, allowing efficient programming via building blocks. The GPU support is enabled via a thorn, which hides the details of GPU programming while achieves high performance via automatic code generation. Cactus is heavily used at LSU in the development of the Numerical Relativity Tool Kit, but is by no means limited to this field. For instance there is a thorn which provides an adaptive mesh refinement driver that can be used in CFD programs.

4.2 Pluto Compiler Advancement

Pluto [3] is a compiler designed to generate effective code for heterogeneous systems. It functions as a source-to-source transformation system that can generate code for general-purpose multicore systems as well as for GPUs. The input source is basically standard C code with a few added directives. Pluto is being enhanced to handle heterogeneous environments with multiple GPUs and many-core processors. In addition, we are exploring the use of user directives to guide optimizations.

4.3 ParalleX

A new systems paradigm is being developed by the STE| |AR group using the ParalleX [4] execution model and its implementation in the experimental runtime system HPX (High Performance ParalleX) [5]. It is intended to improve the reliability and programmability of data-intensive massively parallel computing platforms using GPUs in a more user friendly way. The ParalleX notion of percolation provides for a perfect high level abstraction of massively data-parallel execution using GPUs. All applications built on top of HPX will immediately benefit from the results.

4.4 LA-SiGMA

LA-SiGMA [6] is a statewide NSF funded computational materials science project involving nearly 50 faculty at 7 Louisiana campuses. LA-SiGMA has several computational teams developing a variety of computa-

tional materials science applications of which the most active team is the GPU team. Concentration is being placed on computational quantum chemistry approaches to modeling novel materials and their properties. It will also make use of existing community codes that are GPU-enabled, such as NAMD and LAMMPS.

The CRC@LSU-CCT will be very helpful to the geometric and visual computing group, digital forensic facial reconstruction, medical imaging group, coastal modeling research, and the next-Generation GPU project at LSU. All these projects require efficient parallel processing and can greatly benefit from the proposed CUDA Research Center at LSU.

5 Outreach

HPC@LSU, a partnership between LSU's CCT and Information Technology Services (ITS), is the organization which provides services to the LSU, LONI, and NSF XSEDE (formerly TeraGrid) HPC user community. In addition to operational support for 16 clusters at 7 sites around the state, it also provides practical training via tutorials, workshops, and a variety of on-line resources. The LSU's CCT, provides tier 2 and 3 level support from expert researchers, and member faculty teach formal courses in computational science.

5.1 Description of courses

Several distance learning graduate level classes are shared via synchronous video and are available to students throughout the state and in national and international partner institutions. They are, for the most part, specialized graduate courses that emphasize computational methods and heterogeneous programming that could not be taught at the individual institutions due to the lack of a critical mass.

5.2 Training activities

Training and education at all levels, from primary school through graduate school and beyond, is viewed as an essential component of this project. Programs will be offered both to train a core set of users, who will develop codes for GPU clusters at LSU, and for the next generation of national level heterogeneous machines such as Titan and Blue Waters. The intent is to develop an expertise pipeline to sustain CRC@LSU-CCT in the future. The GPU systems will be used for summer workshops to train students to use PGI GPU accelerator compilers and CUDA. GPUs will be introduced as components of the machines built and used by high school students in our long-running Beowulf Bootcamp. To help establish a pipeline into our graduate programs, the GPU clusters will be used by undergraduate students in the CCT and LA-SiGMA REU programs. To establish a pipeline into our undergraduate programs, K-12 local teachers participating in LA-SIGMA summer RET programs will also use the GPU clusters.