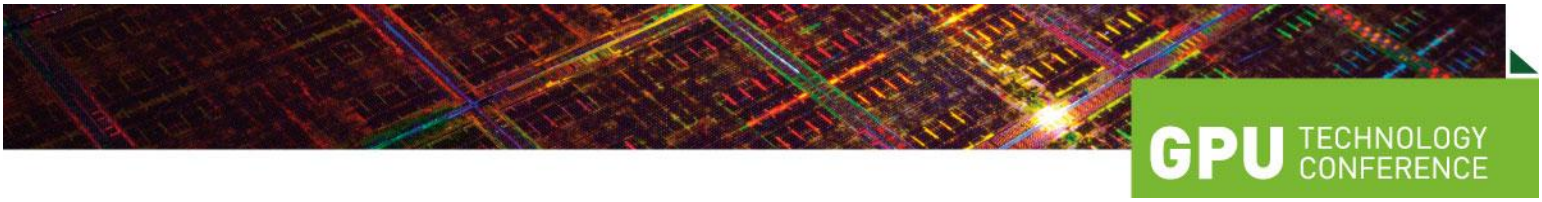

GPU Technology Conference, May 14-17, 2012
McEnery Convention Center, San Jose, California
www.gputechconf.com

Sessions on Application Design & Porting Techniques (subject to change)

IMPORTANT: Visit http://www.gputechconf.com/page/sessions.html for the most up-to-date schedule.

TUTORIALS

S0005 - Languages, APIs and Development Tools for GPU Computing
Will Ramey (NVIDIA)
Day: Monday, 05/14 | **Time:** 9:00 am - 10:30 am
Topic Areas: General Interest; Development Tools & Libraries; Application Design & Porting Techniques
Session Level: Beginner

Get a head start on the conference with this first-day introduction to key technologies for GPU Computing. This 90-minute tutorial session will cover the key features and differences between the major programming languages, APIs and development tools available today. Attendees will also learn several high level design patterns for consumer, professional and HPC applications, with practical programming considerations for each.

SESSIONS

S0248 - Excitements, Challenges, and Rewards in Optimizing GPGPU Kernels
Rajib Nath (University of California San Diego), Stanimire Tomov (University of Tennessee, Knoxville)
Day: Tuesday, 05/15 | **Time:** 9:00 am - 9:50 am
Topic Areas: Algorithms & Numerical Techniques; Application Design & Porting Techniques; Supercomputing
Session Level: Intermediate

Learn about the excitements and challenges in optimizing CUDA kernels for the last two generations of NVIDIA GPGPUs. Autotuning, although crucially important, is merely a silver bullet to port code from one generation of GPU to another. The process required many steps: (a) architecture specific algorithms, (b) tuning algorithms, (c) finding innovative tricks to handle generic cases, (d) tweaking GPU's internal scheduling to handle partition camping, and (e) above all, the dedication of many enthusiastic programmers. We will share our experiences and discoveries through the development of MAGMABLAS - a subset of CUDA BLAS, highly optimized for NVIDIA GPGPUs.

S0418 - High Productivity Computational Finance on GPUs
Aamir Mohammad (Aon Benfield Securities), Peter Phillips (Aon Benfield Securities)
Day: Tuesday, 05/15 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Finance; Application Design & Porting Techniques; Parallel Programming Languages & Compilers
Session Level: Beginner

Learn how Aon Benfield helps clients use GPUs to develop and accelerate Monte Carlo derivatives pricing models. We will present our PathWise software tools used by actuaries and quants in order to rapidly develop and deploy production quality, GPU grid enabled, Monte Carlo models, using only high-level languages and tools without

requiring any knowledge of CUDA or C/C++. We will describe our approaching of using Code Generation, Visual Programming, Domain Specific Languages and scripting languages to create a High Productivity Computing software stack for financial services applications.

S0379 - GPU-based High-Performance Simulations for Spintronics
Jan Jacob (University of Hamburg - Institute of Applied Physics and Microstructure Research Center)
Day: Tuesday, 05/15 | **Time:** 2:30 pm - 2:55 pm
Topic Areas: General Interest; Computational Physics; Application Design & Porting Techniques
Session Level: Intermediate

The joint utilization of the electron's charge and spin in "spintronics" represents a promising technology for data processing and storage in nanostructures. The complex quantum effects like the spin-Hall effect in these devices require demanding numerical simulations providing a convenient link between idealized analytical models to often very complex results from measurements. The simulations involving multiplications and inversions of large matrices provide an ideal showcase for performance gain by employing GPGPUs in the execution of the algebraic routines on these matrices in computing environments with shared execution of algorithms on multiple nodes with multiple GPGPUs and CPU cores.

S0034 - Real-Time Risk Simulation: The GPU Revolution In Profit Margin Analysis
Gilles Civario (ICHEC), Renato Miceli (ICHEC)
Day: Tuesday, 05/15 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Finance; Application Design & Porting Techniques; Algorithms & Numerical Techniques
Session Level: Intermediate

Discover how ICHEC helped a world leading company in its sector, to dramatically speed-up and improve the quality of its real-time risk management tool chain. In this session, we present the method used for porting the core-part of the simulation engines to GPUs using CUDA. This porting was realized on two very different simulation algorithms and resulted in speed-ups of 2 to 3 orders of magnitude, allowing much greater accuracy of the results in a real-time environment.

S0075 - Oculus Real-Time Modular Cognitive Vision System
Jeremie Papon (University of Gottingen), Alexey Abramov (University of Gottingen)
Day: Tuesday, 05/15 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Computer Vision; Audio, Image and Video Processing; Application Design & Porting Techniques; Machine Vision
Session Level: Intermediate

This session will explore ways to integrate GPU processing into a real-time computer vision architecture. While there has been a rapid push to move vision algorithms onto GPUs, integration into an efficient vision system architecture remains elusive. We will discuss our development of a modular vision system architecture that enables rapid prototyping of complex pipelines using multiple GPUs. The system incorporates modules for segmentation, disparity mapping, optical flow and particle filter tracking on the GPU. Our talk will explore the various difficulties associated with developing such a system and will give a hands-on demonstration of Oculus, our vision platform.

S0067 - PIConGPU - Bringing large-scale Laser Plasma Simulations to GPU Supercomputing
Michael Bussmann (Helmholtz-Zentrum Dresden-Rossendorf) , Guido Juckeland (Technical University Dresden)
Day: Tuesday, 05/15 | **Time:** 3:00 pm - 3:50 pm

Topic Areas: Computational Physics; Algorithms & Numerical Techniques; Application Design & Porting Techniques; Supercomputing
Session Level: Advanced

With powerful lasers breaking the Petawatt barrier, applications for laser-accelerated particle beams are gaining more interest than ever. Ion beams accelerated by intense laser pulses foster new ways of treating cancer and make them available to more people than ever before. Laser-generated electron beams can drive new compact x-ray sources to create snapshots of ultrafast processes in materials. With PIConGPU laser-driven particle acceleration can be computed in hours compared to weeks on standard CPU clusters. We present the techniques behind PIConGPU, detailed performance analysis and the benefits of PIConGPU for real-world physics cases.

S0349 - Tree Accumulation on the GPU
Scott Rostrup (Synopsys Inc.)
Day: Tuesday, 05/15 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Algorithms & Numerical Techniques; Application Design & Porting Techniques
Session Level: Advanced

Learn how to map irregular tree structured computations to the GPU efficiently. See how extremely irregular data-dependent computations can be implemented by composing them out of regular data-parallel primitives. In particular we focus on the problem of tree accumulation, a generalization of the scan primitive to arbitrary tree data structures. We first show how tree orderings and properties can be computed using the Euler tour technique and standard scan primitives. Using these orderings we then develop our new approach to computing tree accumulations in parallel.

S0316 - Using GPUs to Accelerate Synthetic Aperture Sonar Imaging via Backpropagation
Thomas Benson (Georgia Tech Research Institute)
Day: Tuesday, 05/15 | **Time:** 3:30 pm - 3:55 pm
Topic Areas: Application Design & Porting Techniques
Session Level: Intermediate

This presentation describes our development of a GPU-accelerated backpropagation implementation for Synthetic Aperture Sonar systems that supports multiple nodes via MPI and multi-GPU nodes. This implementation can form a complex-valued gigapixel image in one hour on a single C2050. We further scale this implementation to the Keeneland system where we can form the same gigapixel image in 21 seconds on 48 nodes with 144 C2070 Tesla GPUs. Our talk will discuss the details of our implementation, including our optimizations and scaling results for various node and GPU configurations, as well as the applicability to other domains, including Synthetic Aperture Radar.

S0108 - An Innovative Massively Parallelized Molecular Dynamic Software
Thomas Guignon (IFPEN), Ani Anciaux Sedrakian (IFP Energie Nouvelles)
Day: Tuesday, 05/15 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Molecular Dynamics; Supercomputing; Application Design & Porting Techniques
Session Level: Intermediate

In this paper, we present how we improved the speedup of the electronic structure calculator VASP by more than an order of magnitude. Recently, the research works done (at IFP Energies Nouvelles) have shown that by coupling traditional clusters or High Performance Computing (HPC) machines with accelerators based on graphical processor units (GPUs), by recording the most time consuming parts of the codes (with programming languages like CUDA, OpenCL) and offloading them on the graphic chips, it is possible to reduce the computing time to ensure a speedup of a factor of 5 to 15.

S0021 - OptiX for DirectX Programmers - Eve Online's GPU-Raytraced Portraits
Bert Peers (CCP Games)
Day: Tuesday, 05/15 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Ray Tracing; Computer Graphics; Application Design & Porting Techniques
Session Level: Intermediate

By integrating NVIDIA's OptiX system for real-time GPU raytracing into a DirectX9 based engine, CCP Games enables high-quality raytraced player portraits for the single shard MMO Eve Online, reusing the game's assets and pipeline. We selectively add stochastic effects while closely maintaining the look of the DX9-based renderer that Art Direction aimed for. In this talk we approach OptiX from the point of view of a programmer familiar with DirectX, discuss integrating these two systems, and show how we reproduced some DirectX-based effects like transparency and subsurface scattering within OptiX.

S0317 - Compiling a Parallel Domain Specific Language to GPUs
Ramesh Narayanaswamy (Synopsys Inc.)
Day: Tuesday, 05/15 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Electrical Design and Analysis; Application Design & Porting Techniques
Session Level: Intermediate

Discuss techniques for compiling Parallel DSLs to GPUs. Verilog is a Domain Specific Language for Hardware Description. Verilog users express parallelism with guarded processes similar to Occamâ€™s guarded commands. Review Verilog semantics, and different approaches to compiling Verilog to parallel architectures and to GPUs. Discuss challenges with (a) Verilog descriptionâ€™s runtime behavior (b) managing process dependency. Discuss approaches and challenges in compiling a parallel DSL to CUDA C.

S0267A - Mixing Graphics and Compute with Multiple GPUs
Alina Alt (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:50 pm
Topic Areas: Computer Graphics; Application Design & Porting Techniques
Session Level: Beginner

In this session we will cover all the different aspects of interaction between graphics and compute. The first part of the session will focus on compute API interoperability with OpenGL (using CUDA and OpenCL APIs), while the second part of the session will delve into interoperability at a system level. In particular we will go through the challenges and benefits of dedicating one GPU for compute and another for graphics, how different system configurations affect data transfer between two GPUs, and how it translates into application design decisions helping to enable an efficient, cross-GPU interoperability between compute and graphics contexts. This talk is repeated on Thursday at 3:30 PM (session S0267B)

S0236 - Advanced Optimization Techniques on a CUDA Implementation of Conjugate Gradient Solvers
Eri Rubin (OptiTex)
Day: Wednesday, 05/16 | **Time:** 10:00 am - 10:25 am
Topic Areas: Algorithms & Numerical Techniques; Algorithms & Numerical Techniques; Computational Physics; Application Design & Porting Techniques
Session Level: Intermediate

Linear systems are at the heart of allot of compute problems. In large sparse systems, there are 2 distinct approaches, the direct and iterative solvers. After many years of researching and testing both approaches, on

CPU and GPU we have implemented a highly efficient CG solver on the GPU using a combination of unique techniques. In this talk we will go over these techniques and the improved performance they bring.

S0190 - Large-Scale Reservoir Simulation on GPU
Song Yu (Chemical & Petroleum Department, University of Calgary)
Day: Wednesday, 05/16 | **Time:** 2:30 pm - 2:55 pm
Topic Areas: Application Design & Porting Techniques; Algorithms & Numerical Techniques
Session Level: Intermediate

Develop highly parallel GPU-based GMRES solver and several precondtioners, and couple them with the in-house reservoir simulator to speedup large-scale reservoir simulation with over one million grid blocks. For those preconditioners, we develop the highly parallelized ILU(k), ILUT, and block ILU(k), block ILUT, with matrix partition by METIS on GPU. The excellent speedup and accurate results can demonstrate the great promising future of the GPU parallel device in parallel reservoir simulation.

S0127 - Petascale Molecular Dynamics Simulations on GPU-Accelerated Supercomputers
James Phillips (University of Illinois)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Molecular Dynamics; Application Design & Porting Techniques; Parallel Programming Languages & Compilers; Supercomputing
Session Level: Intermediate

The highly parallel molecular dynamics code NAMD was chosen in 2006 as a target application for the NSF petascale supercomputer now know as Blue Waters. NAMD was also one of the first codes to run on a GPU cluster when G80 and CUDA were introduced in 2007. How do the Cray XK6 and modern GPU clusters compare to 300,000 CPU cores for a hundred-million-atom Blue Waters acceptance test? Come learn the opportunities and pitfalls of taking GPU computing to the petascale and the importance of CUDA 4.0 features in combining multicore host processors and GPUs in a legacy message-driven application.

S0405 - New Generation GPU Accelerated Financial Quant Libraries
Daniel Egloff (QuantAlea GmbH)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Finance; Application Design & Porting Techniques; Algorithms & Numerical Techniques; Cloud Computing
Session Level: Advanced

Learn from industry experts how new generation GPU accelerated solutions for derivative pricing, hedging, and risk management can be build more efficiently with modern technology and functional programming languages like F# on .NET or Scala on the Java VM. As a concrete example we report from a large derivative pricing project developed in F# on .NET. We will introduce the key design concepts and parallelization strategies, which lead to an efficient and transparent GPU acceleration. Several examples will illustrate the benefit of the functional as compared to the classical object oriented approach.

S0286 - Scaling Applications to a Thousand GPUs and Beyond
Alan Gray (The University of Edinburgh), Roberto Ansaloni (Cray Italy)
Day: Wednesday, 05/16 | **Time:** 4:00 pm - 4:50 pm
Topic Areas: Supercomputing; Computational Fluid Dynamics; Parallel Programming Languages & Compilers; Application Design & Porting Techniques
Session Level: Intermediate

Discover how to scale scientific applications to thousands of GPUs in parallel. We will demonstrate our techniques using two codes representative of a wide spectrum of programming methods. The Ludwig lattice Boltzmann package, capable of simulating extremely complex fluid dynamics models, combines C, MPI and CUDA. The Himeno three-dimensional Poisson equation solver benchmark combines Fortran (using the new coarray feature for communication) with prototype OpenMP accelerator directives (a promising new high-productivity GPU programming method). We will present performance results using the cutting-edge massively-parallel Cray XK6 hybrid supercomputer featuring the latest NVIDIA Tesla 2090 GPUs.

S0377 - C++ Data Marshalling Best Practices
Cliff Woolley (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Finance; Application Design & Porting Techniques
Session Level: Intermediate

When integrating CUDA C++ kernels into existing C++ applications, it is at times desirable to migrate a C++ object instance from the host to the device or vice versa. Given variations among host compilers regarding structure layout, accomplishing this data marshalling in a manner that is reliable, simple, and efficient is a complex issue. cudaMemcpy is our primary means to transfer data to the GPU, but memcpy-style operations are more readily amenable to C-style structures and arrays than to C++ objects or collections of objects. In this session, we will cover the caveats and best practices for marshalling C++ data.

S0298 - Performance Tools for GPU-Powered Scalable Heterogeneous Systems
Allen Malony (University of Oregon)
Day: Wednesday, 05/16 | **Time:** 5:00 pm - 5:50 pm
Topic Areas: Development Tools & Libraries; Parallel Programming Languages & Compilers; Application Design & Porting Techniques
Session Level: Intermediate

Discover the latest parallel performance tool technology for understanding and optimizing parallel computations on scalable heterogeneous platforms. The session will present the TAU performance system and its support of measurement and analysis of heterogeneous platforms composed of clusters of shared-memory nodes with GPUs. In particular, TAU's integration of the CUPTI 4.1+ technology will be described and demonstrated through CUDA SDK examples and the SHOC benchmarks. Attendees will be provided LiveDVDs containing the TAU toolsuite and many pre-installed parallel tool packages. It will also include the last CUDA driver, runtime library, and CUPTI.

S0382 - Hybrid System Architectures for High-Speed Processing in Optical Coherence Tomography
Brian Applegate (Texas A&M University Department of Biomedical Engineering), Brian Applegate (Texas A&M University Department of Biomedical Engineering)
Day: Thursday, 05/17 | **Time:** 10:00 am - 10:25 am
Topic Areas: Medical Imaging & Visualization; Life Sciences; Application Design & Porting Techniques; Development Tools & Libraries
Session Level: Intermediate

Several factors are spurring the development of hardware and software to accomplish high-speed processing for Optical Coherence Tomography (OCT), e.g. ultrahigh speed (>1 MHz) volumetric imaging and clinical applications (e.g. intravascular imaging). The computation power of GPUs ensures that it will be an essential part of the solution. We are exploring the development of a hybrid system in which the computational burden is shared between GPUs and other processors. This will make it possible to extract crucial diagnostic information in real or near real time. Technical challenges and recent progress will be discussed.

S0039 - Data-Driven GPGPU Ideology Extension
Alexandr Kosenkov (University of Geneva), Bela Bauer (Microsoft Research)
Day: Thursday, 05/17 | **Time:** 10:00 am - 10:25 am
Topic Areas: Application Design & Porting Techniques; Computational Physics; Parallel Programming Languages & Compilers; Development Tools & Libraries
Session Level: Advanced

In this session we will demonstrate how the GPGPU ideology can be extended so that it can be used on a scale of Infiniband hybrid system. The approach that we are presenting combines delayed execution, scheduling techniques and, most importantly, casts down the CPU multi-core ideology to the streaming multiprocessor's one enforcing full fledged "GPGPU as a co-processor" way of programming for large-scale MPI hybrid applications. Staying compatible with modern CPU/GPGPU libraries it provides more than a fine grained control over resources - more than you wanted that is.

S0107 - Acceleration of Long-Wave Rapid Radioactive Transfer Model on GPGPU
Mahesh Khadtare (I2IT, Pune University), Prakalp Somawanshi (CRL India)
Day: Thursday, 05/17 | **Time:** 10:30 am - 10:55 am
Topic Areas: Climate & Weather Modeling; Application Design & Porting Techniques; Climate & Weather Modeling
Session Level: Intermediate

The WRF model is a next-generation mesoscale numerical weather prediction system designed to serve both operational forecasting and atmospheric research communities. WRF offers multiple physics options, one of which is the Long-Wave Rapid Radiative Transfer Model. We found, porting rtrn() subroutine to the CUDA challenging. It has couple of recursive loops, for which GPGPUs are actually not suitable. We developed a new technique called loop inversion, which helped us in getting 7.7x speed up for the individual, rtrn() subroutine without memory transfer, and in turn 10x speed up for overall RRTM module including initialization and memory transfer.

S0217 - Efficient Implementation of CFD Algorithms on GPU Accelerated Supercomputers
Ali Khajeh-Saeed (University of Massachusetts, Amherst), Blair Perot (University of Massachusetts, Amherst)
Day: Thursday, 05/17 | **Time:** 10:30 am - 10:55 am
Topic Areas: Computational Fluid Dynamics; Computational Physics; Supercomputing; Application Design & Porting Techniques
Session Level: Intermediate

The goal of this session is to introduce the concepts necessary to perform large computational fluid dynamic (CFD) problems on collections of many GPUs. Communication and computation overlapping schemes become even more critical when using fast compute engines such as GPUs that are connected via a relatively slow interconnect (such as MPI on InfiniBand). The algorithms presented are validated on unsteady CFD simulations of turbulence using 192 graphics processors to update half-a-billion unknowns per computational timestep. The performance results from three different GPU accelerated supercomputers (Lincoln, Forge, and Keeneland) are compared with a large CPU based supercomputer (Ranger).

S0291 - LAtoolbox: A Multi-platform Sparse Linear Algebra Toolbox
Dimitar Lukarski (Karlsruhe Institute of Technology (KIT), Jan-Philipp Weiss (Karlsruhe Institute of Technology)
Day: Thursday, 05/17 | **Time:** 10:30 am - 10:55 am
Topic Areas: Application Design & Porting Techniques
Session Level: Intermediate

Find out about an easy way for building sparse linear solvers for GPUs and multi-/many-core platforms. Based on data abstraction and virtualization of the hardware, the LAtoolbox supports several platforms such as GPUs, multi-core CPUs, and accelerators. The various backends (CUDA, OpenCL, OpenMP, ...) utilize optimized and platform-specific routines and allow seamless integration of GPUs into scientific applications. By means of unified interfaces across all platforms the library enables you to build generic linear solvers and preconditioners on a single code base without specific information of your hardware. We demonstrate portability and flexibility of our open-source approach on heterogeneous platforms.

S0279 - Optimization Techniques for GPU and GPP
Lionel Lacassagne (Institute for Fundamental Electronics), Antoine Pedron (Institute for Fundamental Electronics)
Day: Thursday, 05/17 | **Time:** 2:00 pm - 2:25 pm
Topic Areas: Application Design & Porting Techniques; Computer Vision; Audio, Image and Video Processing
Session Level: Intermediate

We present and evaluate optimizations techniques for GPU and GPP. We describe High level transforms targeting algorithm refactoring and low-level optimizations targeting hardware. The algorithm used to present theses optimizations comes from image processing but is also representative of algorithms of different areas with same kind of local computations like stencils in computer science or finite difference methods in numerical analysis. We evaluate the impact of optimization on three generations of GPU and GPP, both mobile and desktop. We show that significant speedup can be achieved on GPP but also on GPU. Finally we compare together these two architectures.

S0378 - VASP Accelerated with GPUs
Maxwell Hutchinson (University of Chicago)
Day: Thursday, 05/17 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Quantum Chemistry; Application Design & Porting Techniques; Computational Physics
Session Level: Intermediate

This session will detail the performance and capabilities of GPU-accelerated VASP, explain design decisions made in porting VASP to CUDA, and present a roadmap for GPU accelerated VASP development. We've achieved performance improvements up to around 20x on systems of around 100 ions and have implemented exact-exchange. We are working on ports of more conventional functionality.

S0106 - GPU Based Numerical Methods in Mathematica
Ulises Cervantes-Pimentel (Wolfram Research), Abdul Dakkak (Wolfram Research)
Day: Thursday, 05/17 | **Time:** 2:30 pm - 3:20 pm
Topic Areas: Algorithms & Numerical Techniques; Visualization; Application Design & Porting Techniques; Development Tools & Libraries
Session Level: Intermediate

A fast way of developing, prototyping and deploying numerical algorithms that can take advantage of CUDA capable systems is available in Mathematica 8. Over the past year, educators, scientists, and business users have taken advantage of the benefits that the support of GPU programming in Mathematica. By integrating and implementing CUDA/OpenCL in their programs, users make use of a hybrid approach, combining the speed-up that GPUs offer and a powerful numerical development system. In this presentation several examples describing numerical applications ranging from deconvolution of MRI imaging, linear solvers for FEM, systems of ODEs, line integral convolution visualization are presented.

S0231 - Levenberg-Marquardt using Block Sparse Matrices on CUDA
Tetsuo Tawara (Koozyt, Inc.)
Day: Thursday, 05/17 | **Time:** 2:30 pm - 2:55 pm
Topic Areas: Application Design & Porting Techniques; Algorithms & Numerical Techniques
Session Level: Intermediate

This session describes the experiences of constructing GPU based matrix-vector functions for block sparse matrices having multiple block sizes and a domain-specific numerical Jacobian generation function. The bundle adjustment algorithm is an optimization procedure which attempts to refine the relative camera pose, and 3D structure location variables, estimated from multiple sets of images. The Conjugate Gradient algorithm is used to solve the normal equations which appear in the inner loop to the non-linear least squares problem.

S0368 - Unraveling the Mysteries of Quarks with Hundreds of GPUs
Ronald Babich (NVIDIA)
Day: Thursday, 05/17 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Computational Physics; Application Design & Porting Techniques; Algorithms & Numerical Techniques; Supercomputing
Session Level: Intermediate

Dive into the world of quarks and gluons, and hear how GPU computing is revolutionizing the way many calculations in lattice quantum chromodynamics (lattice QCD) are performed. The main computational challenge in such calculations is to repeatedly solve large systems of linear equations arising from a four-dimensional finite-difference problem. In this session, we'll discuss strategies for parallelizing such a solver across hundreds of GPUs. These include techniques and algorithms for reducing memory traffic and inter-GPU communication. The net result is an implementation that achieves better than 20 Tflops on 256 GPUs, realized in the open-source "QUDA" library.

S0091 - Sustainable Hybrid Parallelization of an Unstructured Hydrodynamic Code
Raphaël Poncet (Commissariat à l'Energie Atomique et aux Energies Alternatives)
Day: Thursday, 05/17 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Application Design & Porting Techniques; Algorithms & Numerical Techniques; Computational Fluid Dynamics; Computational Physics
Session Level: Advanced

The goal of this presentation is to share our methodology for porting a numerical code to hybrid supercomputing architectures using MPI coupled with directive-based languages (OpenMP for multicore CPUs, and HMPP for GPUs). Our code, VOLNA, is an unstructured partial differential equation hydrodynamic solver developed for the simulation of tsunamis. Our results demonstrate that using directive-based languages such as HMPP for GPU programming, one can retain good performance (e.g. speedup of 15 compared to 1 CPU core, 3 compared to 8 CPU cores) with minimal modifications of the original CPU source code (about 30 lines of directives in our case).

S0267B - Mixing Graphics and Compute with Multiple GPUs (Repeat Presentation)
Alina Alt (NVIDIA)
Day: Thursday, 05/17 | **Time:** 3:30 pm - 4:20 pm
Topic Areas: Computer Graphics; Application Design & Porting Techniques
Session Level: Beginner

In this session we will cover all the different aspects of interaction between graphics and compute. The first part of the session will focus on compute API interoperability with OpenGL (using CUDA and OpenCL APIs), while the second part of the session will delve into interoperability at a system level. In particular we will go through the

challenges and benefits of dedicating one GPU for compute and another for graphics, how different system configurations affect data transfer between two GPUs, and how it translates into application design decisions helping to enable an efficient, cross-GPU interoperability between compute and graphics contexts. This talk is repeated on Tuesday at 5:00 PM (session S0267A)

S0138 - GPU Task-Parallelism: Primitives and Applications
Stanley Tzeng (University of California, Davis), Anjul Patney (University of California, Davis)
Day: Thursday, 05/17 | **Time:** 3:30 pm - 3:55 pm
Topic Areas: Application Design & Porting Techniques; Development Tools & Libraries; Computer Graphics
Session Level: Intermediate

We explore how a task-parallel model can be implemented on the GPU and address concerns and programming techniques for doing so. We discuss the primitives for building a task-parallel system on the GPU. This includes novel ideas for mapping tasking systems onto the GPU including task granularity, load balancing, memory management, and dependency resolution. We also present several applications which demonstrate how a task-parallel model is more suitable than the regular data parallel model. These applications include a Reyes renderer, tiled deferred lighting renderer, and a video encoding demo.

S0111 - An Efficient CUDA Implementation of a Tree-Based N-Body Algorithm
Martin Burtscher (Texas State University)
Day: Thursday, 05/17 | **Time:** 3:30 pm - 4:20 pm
Topic Areas: Application Design & Porting Techniques; Astronomy & Astrophysics; Molecular Dynamics; Supercomputing
Session Level: Advanced

This session presents a complete CUDA implementation of the irregular Barnes-Hut n-body algorithm. This algorithm repeatedly builds and traverses unbalanced trees, making it difficult to map to GPUs. We explain in detail how our code exploits the architectural features of GPUs, including lockstep operation and thread divergence, both of which are commonly viewed as hurdles to achieving high performance, especially for irregular codes. On a five million body simulation running on a Tesla C2050, our CUDA implementation is 30 times faster than a parallel pthreads version running on a high-end 6-core Xeon.

S0220 - Enabling faster material science modeling using the accelerated Quantum ESPRESSO
Filippo Spiga (Irish Centre for High-End Computing)
Day: Thursday, 05/17 | **Time:** 4:30 pm - 5:20 pm
Topic Areas: Quantum Chemistry; Supercomputing; Application Design & Porting Techniques
Session Level: Intermediate

The goal of this session is to present the advantages of mixing CUDA libraries and CUDA kernels to deliver a robust community package for material science modeling that fully exploits multi-core systems equipped with GPUs. The Plane-Wave Self-Consistent Field (PWscf) code of the Quantum ESPRESSO suite is the focus of this work. During the session the main computation-dependent components, that also represent fundamental building blocks for many other quantum chemistry codes, will be discussed and analyzed. Subsequently an in-depth performance assessment of several realistic scientific cases will be presented, starting from single workstations to large clusters equipped with hundreds of GPUs.