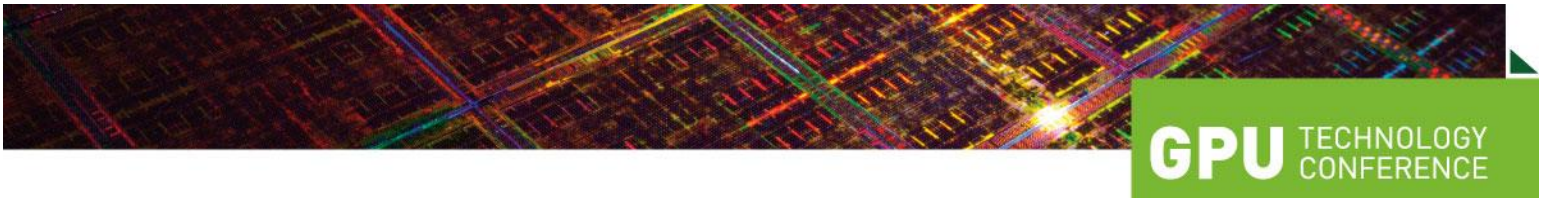

GPU Technology Conference, May 14-17, 2012
McEnery Convention Center, San Jose, California
www.gputechconf.com

Sessions on Supercomputing (subject to change)

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

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.

S0337 - High-Throughput Epistasis Screening Using GPUs
Mark Seligman (Insilicos LLC)
Day: Tuesday, 05/15 | **Time:** 9:00 am - 9:25 am
Topic Areas: Bioinformatics; Life Sciences; Supercomputing; Cloud Computing
Session Level: Intermediate

Epistasis is the interaction of two or more genes in coding for a biological property. Epistasis is believed to be an important factor in an individual's susceptibility to disease, and the search for epistasis is a major component in the development of personalized approaches to genomic medicine. Statistical tests for epistasis are typically confounded by the multiple-testing problem, that is, the aggregated loss of precision incurred through repeated hypothesis testing. One way to circumvent this problem is to simulate a false-discovery rate via resampling. We report success in using GPUs to accelerate these highly compute-intensive resampling techniques.

S0618 - Best Practices of a 800TFlop Hybrid Supercomputer Implementation *(Presented by Appro)*
Steve Lyness (Appro), Taisuke Boku (University of Tsukuba)
Day: Tuesday, 05/15 | **Time:** 9:30 am - 10:20 am
Topic Areas: Supercomputing; Astronomy & Astrophysics
Session Level: Intermediate

Learn about the "Frontier Computing System", deployed by Appro for the University Of Tsukuba Center Of Computational Sciences in Japan containing over half a million GPU cores. Learn how reliability, availability, manageability and compatibility were essential for this successful 800TF hybrid supercomputing implementation. Explore new techniques in how HA-PACS is accelerating large scale parallel code by combining CPU/GPU processing cluster configurations for scientific research, such as astrophysics and climate modeling. Learn how to improve data I/O performance and memory size limitations in hybrid systems configured with Lustre™ File System

offering the best performance per dollar and excellent memory capacity per/FLOP.

S0255 - Telecom Systems Simulations Acceleration via CPU/GPU Co-Processing: Turbo Codes Case Study
Paolo Spallaccini (Ericsson)
Day: Tuesday, 05/15 | **Time:** 10:00 am - 10:25 am
Topic Areas: Algorithms & Numerical Techniques; Audio, Image and Video Processing; Supercomputing
Session Level: Intermediate

Learn how the struggle for acceleration of simulations of a Serially Concatenated turbo code (SCCC) led to the knowledge of new techniques applicable to a broad range of non-natively parallel physical layer telecommunication systems simulations. The overall architectural features of CUDA became inspiring for newer parallelization techniques involving algorithm engineering; the simulation acceleration attained for iterative SCCC Decoder represents an example of efficiency of leveraging on heterogeneous GPU-CPU coprocessing concepts. The registrants will deep dive into data sets and tasks organization strategies as well as into results and insights, all widely presented and discussed.

S0343 - A Quantum Chemistry Domain-Specific Language for Heterogeneous Clusters
Antonino Tumeo (Pacific Northwest National Laboratory), Oreste Villa (Pacific Northwest National Laboratory)
Day: Tuesday, 05/15 | **Time:** 10:00 am - 10:25 am
Topic Areas: Quantum Chemistry; Supercomputing
Session Level: Intermediate

This talk discuss the development of a Domain-Specific Language (DSL), the tools and the related runtime for efficiently generating Tensor Contractions (generalized matrix multiplications), an important part of many quantum chemistry methods (e.g. Coupled Cluster Theory). Starting from a high level description of the computation, the tool analyses it and generates optimized C, OpenCL or CUDA implementations. The runtime, supporting a task based computation model, is then able to execute the generated code on GPU-accelerated heterogeneous large scale clusters, maximizing the utilization of the processing elements and minimizing communication costs.

S0606 - GPU-accelerated Science on Titan: Tapping into the World's Preeminent GPU Supercomputer to Achieve Better Science
Jack Wells, Ph.D. (Oak Ridge National Laboratory)
Day: Tuesday, 05/15 | **Time:** 2:00 pm - 6:00 pm
Topic Areas: Supercomputing
Session Level: Beginner

This year, the leadership-class computing facility at Oak Ridge National Labs is upgrading its largest supercomputer for open science, "Jaguar", to employ high-performance, power- efficient GPUs. Once the transition is complete, the machine will be known as "Titan." In this extended GTC session, we will feature a range of presenters showcasing research codes that will run computational science on the GPU at scale. Through these selected presentations, we will investigate the progress and anticipated results of GPU-acceleration of these significant codes. In this session, we will also explain how research scientists interested in tapping into the immense capabilities of Titan can do so, through programs such as the Incite program sponsored by the US Department of Energy.

S0412 - A 2-Petaflops Stencil Application with Stereoscopic 3D Visualization - Gorden Bell Prize 2011
Takayuki Aoki (Tokyo Institute of Technology)
Day: Tuesday, 05/15 | **Time:** 2:00 pm - 2:25 pm
Topic Areas: Supercomputing; Computational Fluid Dynamics; Climate & Weather Modeling; Stereoscopic 3D
Session Level: Intermediate

Most stencil applications such as CFD and structure analysis are memory-bound problems. GPU has high performances in both computation and memory bandwidth suitable for them. The TSUBAME 2.0 supercomputer with 4224 GPUs has started since November 2010. We study a metal dendritic solidification by solving the phase-field model. The performance of 2.0 Petaflops was achieved for 4,096x6,500x1,0400 mesh on 4000 GPUs and we received the ACM Gordon Bell Prize in 2011. We also demonstrated several large-scale stencil applications (Lattice Boltzmann, weather prediction and so on) with stereoscopic 3D visualization.

S0519 - GPU Accelerated Bioinformatics Research at BGI
BingQiang Wang (BGI)
Day: Tuesday, 05/15 | **Time:** 2:00 pm - 2:25 pm
Topic Areas: Bioinformatics; Life Sciences; Algorithms & Numerical Techniques; Supercomputing
Session Level: Intermediate

After digitizing DNA double helix by sequencing, computation is the key connecting raw sequences with life science discoveries. As massive data is generate, how to process and analysis as well as storage them in an efficiently manner turns out to be a major challenge. By developing GPU accelerated bioinformatics tools and integrate them into pipelines, BGI researchers now run analysis pipelines in several hours instead of several days. These tools include SOAP3 aligner, SNP calling and tool for population genomics. The speed up is generally around 10-50x comparing with traditional counterparts.

S0620 - VSIPL++: A High-Level Programming Model for Productivity and Performance ***(Presented by Mentor Graphics Corporation)***
Brooks Moses, Ph.D. (Mentor Graphics Corporation)
Day: Tuesday, 05/15 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Supercomputing
Session Level: Beginner

Learn how VSIPL++ can improve your productivity and provide software portability, without sacrificing performance. We will describe how VSIPL++'s open-standard high-level programming model addresses the challenges of writing high-performance embedded software on GP-GPUs and other heterogeneous hardware, using advanced C++ techniques and data abstraction -- and how we make this work in the real world. We will also present a comparison of performance results from various configurations of CPU and GP-GPU processing engines for a signal processing application developed using VSIPL++.

S0308 - Recent Trends in Hierarchical N-body Methods on GPUs
Rio Yokota (King Abdullah University of Science and Technology)
Day: Tuesday, 05/15 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Algorithms & Numerical Techniques; Supercomputing; Development Tools & Libraries
Session Level: Intermediate

See the newest developments in the area of hierarchical N-body methods for GPU computing. Hierarchical N-body methods have O(N) complexity, are compute bound, and require very little synchronization, which makes them a favorable algorithm on next-generation supercomputers. In this session we will cover topics such as hybridization of treecodes and fast multipole methods, auto-tuning kernels for heterogenous systems, fast tree construction based on prefix sums, fast load balancing of global trees, and more. Examples will be given using ExaFMM --an

open source hierarchical N-body library for heterogenous systems developed by the speaker. (released at SC11)

S0067 - PIConGPU - Bringing large-scale Laser Plasma Simulations to GPU Supercomputing
Michael Bussmann (Helmholtz-Zentrum Dresden-Rossendorf), Guido Juckeland (Center for Information Services and High Performance Computing, 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.

S0089 - Accelerator Directives, OpenACC and OpenMP4ACC
James Beyer (Cray Inc.), David Oehmke (Cray Inc.)
Day: Tuesday, 05/15 | **Time:** 4:00 pm - 4:50 pm
Topic Areas: Parallel Programming Languages & Compilers; Supercomputing
Session Level: Intermediate

Rather than require the programmer to rewrite code for accelerators several directive sets have been created and proposed to support non-cache coherent and cache coherent accelerators. This talk will present the OpenACC specification and its implementation for Cray developers, as well as touch on a similar proposal being evaluated by the OpenMP language committee. The presentation will start by discussing the Memory and Execution model needed to allow a programmer to write codes that will run effectively on both distinct memory systems and unified memory systems. Once a proper background has been set the directives will be examined via usage examples.

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.

S0156 - Towards Computing the Cure for Cancer
Wu Feng (Virginia Tech)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:50 pm
Topic Areas: Bioinformatics; Life Sciences; Supercomputing; Algorithms & Numerical Techniques
Session Level: Intermediate

Learn about how to create "designer" genomic analysis pipelines as part of the "Compute the Cure" for cancer initiative from NVIDIA Foundation. Get an overview of an open-source framework that enables the creation of customized genomic analysis pipelines. Discover how different plug-ins from the "mapping/realignment/discovery" repositories, respectively, can be composed to form a genomic analysis pipeline. Learn to use next-generation sequencing data to characterize previously undetectable genetic changes between normal and malignant cells. Find out how you can contribute to the "Compute the Cure" cause.

S0427 - Intra-Day Risk-Management with Parallelized Algorithms on GPUs
Partha Sen (Fuzzy Logix)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:50 pm
Topic Areas: Databases, Data Mining, Business Intelligence; Finance; Algorithms & Numerical Techniques; Supercomputing
Session Level: Advanced

The challenge with intra-day risk management is that a very large number of calculations are required to be performed in a very short amount of time. Typically, we may be interested in calculating VaR for 100 to 1000 securities per second based on 100 million potential scenarios. The magnitude of these calculations is not Utopian but it reflects the reality of modern financial institutions and exchanges. In this presentation, we outline how the complex problem of intra-day risk management can be solved using parallelized algorithms on GPUs. The methodology has been proven in a POC at 2 financial institutions.

S0531 - Exascaling Your Apps
Mike Bernhardt (The Exascale Report)
Day: Wednesday, 05/16 | **Time:** 9:00 am - 10:30 am
Topic Areas: Supercomputing
Session Level: Beginner

In the global exascale race, hardware often takes center stage. But the race might ultimately be won or lost based on how well the industry optimizes new and existing applications for extreme parallelism. Todayâ€™s apps will not just run on tomorrow's systems, so we must think strategically and creatively about how to design applications that take maximum advantage of the first power-efficient, accelerator-driven exascale systems. This panel of HPC, software and computer science experts will discuss what we can, and should be doing, including a review of new scientific and commercial HPC requirements, programming model options and how to best align architecture and software design processes.

S0171 - Numerical Modeling Of 3D Anisotropic Seismic Wave Propagation On MultiGPU Platforms
Denis Sabitov (Schlumberger)
Day: Wednesday, 05/16 | **Time:** 9:00 am - 9:50 am
Topic Areas: Energy Exploration; Algorithms & Numerical Techniques; Supercomputing; Molecular Dynamics
Session Level: Intermediate

We present an efficient and accurate numerical algorithm for the simulation of seismic experiments. The basis of the approach is a heterogeneous spectral element method implemented on MultiGPU applied to anisotropic elastic wave equation. The approach was designed to simulate wave propagation in 3D arbitrary anisotropic elastic media. Due to the use of an unstructured grid, the spectral element algorithm enables handling

complicate geometries of the layers. We discuss results and computational efforts of simulation on MultiGPU platform. Several aspects of the code implementation are considered: optimal domain decomposition, data transfers between GPU by means of P2P and UVA, etc.

S0633 - Learn about new Hewlett Packard GPU Systems, Solutions, and Applications! ***(Presented by Hewlett Packard)***
David Korf (Hewlett Packard) , John Brown (Hewlett Packard)
Day: Wednesday, 05/16 | **Time:** 10:00 am - 10:50 am
Topic Areas: Supercomputing
Session Level: Intermediate

Learn how to shorten time to discovery, gain faster insight, and beat the barriers to innovation, with performance, efficiency and agility! Hear the latest on how you can do this and more with HP's purpose built SL server line. Servers are specifically designed for GPUs with HP ProActive Insight Architecture. Discover what a new generation of workstation desktop GPU computing technology from HP and NVIDIA can do for you! HP will compare and contrast GPU compute performance on the PCI Express Gen2 architecture available in HP's Z800 Workstation to the PCI Express Gen3 architecture in HP's latest Z820 Workstation.

S0304 - Large Scale Computational Fluid Dynamics Simulations on Hybrid Supercomputers
John Humphrey (EM Photonics), Eric Kelmelis (EM Photonics)
Day: Wednesday, 05/16 | **Time:** 10:30 am - 10:55 am
Topic Areas: Computational Fluid Dynamics; Supercomputing
Session Level: Intermediate

Learn how to approach the all-too-common program of trying to retrofit a major application for speed in the modern era of the hybrid supercomputer. In this talk, we will focus on computational fluid dynamics (CFD) codes that are run on Top500 Supercomputers. Many of these applications have existed for 20 or more years, so the process of adding the GPU and getting wall-clock improvements in performance can be very challenging! Our talk will discuss how to properly target your effort, the impact of directives-based coding, and how to maintain efficiency across a hybrid cluster.

S0342 - Volumetric Processing and Visualization on Heterogeneous Architecture
Wei Li (Siemens Corporation)
Day: Wednesday, 05/16 | **Time:** 2:00 pm - 2:25 pm
Topic Areas: Visualization; Supercomputing
Session Level: Advanced

Volumetric data is typically very large and involves intensive computation for processing and visualization. We have developed an OpenCL-based framework that can utilize all available resources in a system or a cluster of systems. The framework manages one or more OpenCL devices. A large volume is partitioned into bricks. Each OpenCL device is associated with a set of brick producers that generates the contents of bricks while optionally utilizing other bricks as input. The framework is also composed of a scheduler that distributes brick workloads to different devices and chooses an optimized processing order aiming at certain criteria.

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 known 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.

S0259 - A High Performance Platform for Real-Time X-Ray Imaging
Suren Chilingaryan (Karlsruhe Institute of Technology)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: General Interest; Supercomputing; Audio, Image and Video Processing; Algorithms & Numerical Techniques
Session Level: Intermediate

We will share our experience on development of the GPU-based platform for synchrotron-based X-ray imaging aimed to analysis of dynamic processes. The complete data flow from the camera to the data storage will be discussed with a special focus on I/O issues, hardware platform, and ways to utilize the available system resources. An efficient GPU-implementation of filtered back projection will be presented highlighting differences of implementations for GT200, Fermi, and AMD Cypress architectures. We will introduce our software platform used to abstract current configuration of the imaging station and to simplify the development of parallel image processing algorithms.

S0635 - How to Bake Portable Many-Core Programs *(Presented by CAPS)*
François Bodin (CAPS Entreprise)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Supercomputing
Session Level: Intermediate

A legacy code, a cool many-core accelerator and a directive-based programming environment are the main ingredients of the recipe to transform your legacy code into a portable many-core one. This presentation shows by the example how to exploit accelerators in legacy code without sacrificing portability. We describe a methodology and the use of directives, such as HMPP and OpenACC, to exploit the massive parallelism provided by many-core devices. During the presentation we illustrate using numerous illustrations how to analyze performance, tune accelerator code, reduce data transfers, deal with libraries, exploit multiple accelerators, etc.

S0340 - Debug Multi-GPU Applications on CUDA-Accelerated Clusters with TotalView
Chris Gottbrath (Rogue Wave Software)
Day: Wednesday, 05/16 | **Time:** 3:30 pm - 4:20 pm
Topic Areas: Development Tools & Libraries; Supercomputing
Session Level: Intermediate

Learn how TotalView can help you develop CUDA applications on single servers, multi-GPU servers, and HPC-style clusters. For more than 20 years the TotalView debugger has set the standard for parallel and multi-core

debugging on Linux, HPC clusters and custom supercomputers such as the Cray XT/XE/XK series. CUDA developers deal with the same types of complexity and can realize the same productivity benefits. This talk will introduce TotalView for CUDA and show how you can program more easily with CUDA 3.2, 4.0 and 4.1.

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.

S0367 - Physis: An Implicitly Parallel Framework for Stencil Computations
Naoya Maruyama (Tokyo Institute of Technology)
Day: Wednesday, 05/16 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Parallel Programming Languages & Compilers; Supercomputing; Development Tools & Libraries; Computational Fluid Dynamics
Session Level: Intermediate

This session presents how to implement finite difference methods in a concise, readable, and portable way, yet achieving good scalability over hundreds of GPUs, using the Physis high-level application framework. Physis extends the standard C language with a small set of custom declarative constructs for expressing stencil computations with multidimensional structured grids, which are automatically translated to CUDA for GPU acceleration and MPI for node-level parallelization with automatic domain-specific optimizations such as overlapped boundary exchanges. We demonstrate the programmability improvement and performance of Physis using hundreds of GPUs on TSUBAME2.0.

S0217 - Efficient Implementation of CFD Algorithms on GPU Accelerated Supercomputers
Ali Khajeh Saeed (University of Massachusetts, Amherst), Blair Perot (University of Massachusetts, Amherst)
Day: Wednesday, 05/16 | **Time:** 5:30 pm - 5:55 pm
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).

S0057 - GPU-Accelerated Molecular Dynamics Simulation of Solid Covalent Crystals
Chaofeng Hou (Institute of Process Engineering, Chinese Academy of Sciences)
Day: Thursday, 05/17 | **Time:** 9:00 am - 9:25 am
Topic Areas: Molecular Dynamics; Algorithms & Numerical Techniques; Supercomputing
Session Level: Intermediate

An efficient and highly scalable algorithm for molecular dynamics (MD) simulation (using sophisticated many-body potentials) of solid covalent crystals is presented. Its effective memory throughput on a single C2050 GPU board reached 102 GB/s (81% of the peak), the instruction throughput reached 412 Ginstr/s (80% of the peak), and 27% of the peak flops of a single GPU was obtained. Parallel efficiency of the algorithm can be as high as 95% on all 7168 GPUs of Tianhe-1A, reaching possibly a record in high performance of MD simulations, 1.87Pflops in single precision.

S0347 - Accelerating Radio Astronomy Cross-Correlation beyond 1 Tflops using Fermi
Michael Clark (NVIDIA)
Day: Thursday, 05/17 | **Time:** 9:00 am - 9:50 am
Topic Areas: Astronomy & Astrophysics; Supercomputing
Session Level: Intermediate

Radio astronomy is a signal processing application that requires extreme supercomputing. While today's radio telescopes require 10-100 Tflops of computational power, by the end of the decade this will increase to 1 Exaflops. The most compute intensive part of this problem is the so-called cross-correlation algorithm, which is a linear-algebra problem. In this session we demonstrate that the Fermi architecture is ideally suited to this problem, and through exploiting the Fermi memory hierarchy it is possible to achieve close to 80% of peak performance in a real application.

S0362 - Maximizing Performance on Multi-GPU Systems
Kenneth Czechowski (Georgia Tech)
Day: Thursday, 05/17 | **Time:** 9:00 am - 9:25 am
Topic Areas: Supercomputing
Session Level: Advanced

Are 512 CUDA Cores not enough? This session is for power users that are looking to scale applications to multi-GPU systems. We will take a holistic approach towards optimization. Rather than just focusing on CUDA programming, this session will cover techniques for reducing pressure on the PCIe bus, using CUDA Streams to improve load balance, dealing with NUMA impacts, and taking advantage of CPU threads. This talk will also cover strategies for developing applications that run on clusters with 100 or more GPUs.

S0044 - A Massively Parallel Two-Phase Solver for Incompressible Fluids on Multi-GPU Clusters
Peter Zaspel (University of Bonn)
Day: Thursday, 05/17 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Computational Fluid Dynamics; Supercomputing; Algorithms & Numerical Techniques; Digital Content Creation & Film
Session Level: Intermediate

Join our presentation of a multi-GPU fluid solver for high performance GPU compute clusters. We use high-order scientific techniques to simulate the interaction of two fluids like air and water. Scientists, engineers and even the computer animation industry will profit from the enormous compute power of tens or hundreds of GPUs. A major focus in this talk will be on the applied GPU implementation techniques and the performance results including performance per Watt and performance per dollar results. We also highlight the lessons we learned

from porting the complex CPU CFD code NaSt3DGPF to the GPU.

S0119 - Best Practices for Architecting and Managing High-Performance GPU Clusters
Dale Southard (NVIDIA)
Day: Thursday, 05/17 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Cluster Management; Supercomputing
Session Level: Intermediate

An overview of designing, deploying, and managing GPU clusters for HPC. Learn to build and operate top500-class GPU computing resources that provide users with the latest CUDA features.

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.

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.

S0038 - Designing Killer CUDA Applications for X86, multiGPU, and CPU+GPU
Robert Farber
Day: Thursday, 05/17 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Machine Learning & AI; Supercomputing; Databases, Data Mining, Business Intelligence; Computer Vision
Session Level: Intermediate

CUDA redefined software development with 10 to 1000-times faster GPU applications. Now a single CUDA source tree can support the x86 mass market (no GPU required) and 1/3 billion CUDA-enabled GPUs. MultiGPU and CPU+GPU apps utilize all system resources. GPUdirect, UVA, caches, prefetching, ILP (Instruction level

Parallelism), automated analysis tools and more offer ease, capability, and performance. The overall impact on software investment, scalability, balance metrics, programming API, and lifecycle will be considered. Working real-time video and other examples from my book, "CUDA Application Design and Development" provide practical insight to enable augmented reality and your killer apps.

S0282 - Leveraging NVIDIA GPUDirect on APEnet+ 3D Torus Cluster Interconnect
Davide Rossetti (Italian National Institue for Nuclear Physics)
Day: Thursday, 05/17 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Supercomputing; Computational Physics
Session Level: Intermediate

APEnet+ is a novel cluster interconnect, based on a custom PCI card which features a PCI Express Gen2 X8 link and a re-configurable HW component (FPGA). It supports a 3D Torus topology and has special acceleration features specifically developed for NVIDIA Fermi GPUs. An introduction to the basic features and the programming model of APEnet+ will be followed by a description of its performance on some numerical simulations, e.g. High Energy Physics simulations.

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.