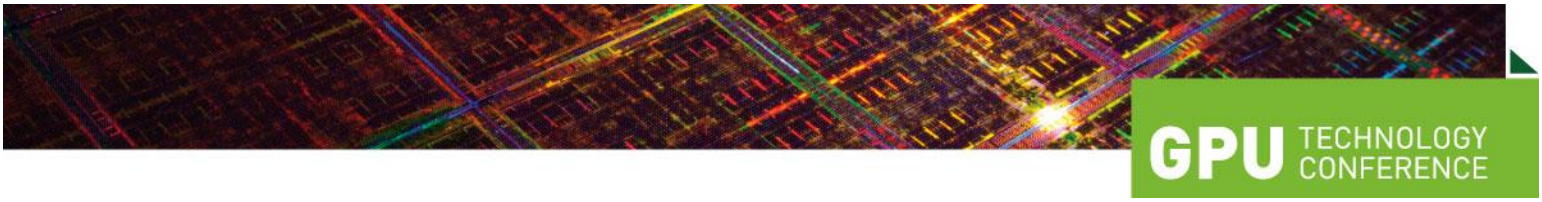


GPU Technology Conference, May 14-17, 2012
McEnery Convention Center, San Jose, California
www.gputechconf.com

Sessions on Algorithms & Numerical Techniques (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.

S0296 - A GPU-Enabled SPH Method for Micro and Nanofluidic Simulations
Daniel Gaudlitz (FluiDyna GmbH)
Day: Tuesday, 05/15 | **Time:** 9:00 am - 9:25 am
Topic Areas: Computational Fluid Dynamics; Algorithms & Numerical Techniques
Session Level: Intermediate

With SPH methods multi-phase flows within complex geometries can be efficiently investigated. Also physical effects present in micro- and nanofluidic applications are described with little effort using the SPH methodology. In order to investigate microfluidic applications relevant to industry, large domains and high spatial resolutions are required. Therefore, a SPH method for accelerated computations on GPUs is currently developed. The code features dynamic casting of computational data into blocks of appropriate size to fit the GPU memory layout. Also tree-like data structures for efficient manipulation of particle distributions help to obtain significant performance gains on GPU hardware.

S0268 - Virtual Process Engineering - Realtime Simulation of Multiphase Systems
Wei Ge (Institute of Process Engineering, Chinese Academy of Sciences)
Day: Tuesday, 05/15 | **Time:** 9:00 am - 9:50 am
Topic Areas: Computational Fluid Dynamics; Molecular Dynamics; Computational Physics; Algorithms & Numerical Techniques
Session Level: Advanced

Realtime simulation and virtual reality with quantitatively correct physics for industrial processes with multi-scale and multiphase system is once a remote dream for process engineering, but is becoming true now with CPU-GPU hybrid supercomputing. Numerical and visualization methods for such simulations on thousands of GPUs will

be reported with applications in chemical and energy industries.

S0255 - Telecom Systems Simulations Acceleration via CPU/GPU Co-Processing: Turbo Codes Case Study
Paolo Spallaccini (Ericsson), Stefano Chinnici (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 co-processing 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.

S0376 - Dynamic Programming on CUDA: Finding the Most Similar DNA Sequence
Grzegorz Kokosiński (IBM Poland), Krzysztof Zarzycki (IBM Poland)
Day: Tuesday, 05/15 | **Time:** 10:00 am - 10:25 am
Topic Areas: Bioinformatics; Algorithms & Numerical Techniques
Session Level: Intermediate

Learn a couple of techniques to speed up compute-heavy Dynamic Programming algorithms on the GPU. Our particular problem regarded DNA sequences: given a reference sequence, how to find the one most similar to it among a large database? The sequences are millions characters long, and their similarity is calculated with a (quadratic) DP algorithm, which makes the problem very tough even for the GPUs. We speed up both the theoretical and practical side: we present programming techniques that enable Dynamic Programming to be performed at the hardware speed, and improvements to the algorithm itself that drastically lower the execution time.

S0031 - Unstructured Grid Numbering Schemes for GPU Coalescing Requirements
Andrew Corrigan (Naval Research Laboratory), Johann Dahm (University of Michigan)
Day: Tuesday, 05/15 | **Time:** 10:00 am - 10:25 am
Topic Areas: Computational Fluid Dynamics; Algorithms & Numerical Techniques; Computational Physics
Session Level: Advanced

Learn how to achieve high performance for computational fluid dynamics (CFD) solvers over unstructured grids using numbering schemes tailored for GPU coalescing requirements. Using these techniques, unstructured grid CFD solvers can make more effective use of memory bandwidth, which is an otherwise significant performance bottleneck that has so far led to relatively limited performance gains on GPUs in comparison to structured grid CFD solvers. Performance benchmarks will be shown using the Jet Engine Noise Reduction (JENRE) code.

S0088 - Point Cloud Library (PCL) on CUDA
Radu Rusu (Willow Garage, Inc.), Michael Dixon (Willow Garage, Inc.)
Day: Tuesday, 05/15 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Computer Vision; Algorithms & Numerical Techniques; Stereoscopic 3D; Machine Vision
Session Level: Intermediate

The Point Cloud Library (PCL - http://pointclouds.org) is a large scale, open project for 3D point cloud processing. The PCL framework contains numerous state-of-the art algorithms including filtering, feature estimation, surface reconstruction, registration, model fitting and segmentation. Due to the massively parallel nature of many of the above algorithms, GPGPU accelerations holds great potential for achieving real-time performance in numerous applications. In this work we demonstrate some of the recent advances in GPGPU programming for 3D point cloud processing, and outline plans for future development.

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.

S0313 - Understanding and using Atomic Memory Operations
Lars Nyland (NVIDIA), Stephen Jones (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Algorithms & Numerical Techniques; Parallel Programming Languages & Compilers
Session Level: Advanced

Atomic memory operations provide powerful communication and coordination capabilities for parallel programs, including the well-known operations compare-and-swap and fetch-and-add. The atomic operations enable the creation of parallel algorithms and data structures that would otherwise be very difficult (or impossible) to express without them - for example: shared parallel data structures, parallel data aggregation, and control primitives such as semaphores and mutexes. In this talk we will use examples to describe atomic operations, explain how they work, and discuss performance considerations and pitfalls when using them.

S0046 - Application of the GPU to a Two-Part Computational Electromagnetic Algorithm
Eric Dunn (SAIC)
Day: Tuesday, 05/15 | **Time:** 2:30 pm - 2:55 pm
Topic Areas: Computational Physics; Algorithms & Numerical Techniques; Ray Tracing
Session Level: Beginner

The shooting and bouncing ray (SBR) method is one way to simulate electromagnetic field radiation. Like all methods, there are certain problems where it does not yield accurate results. In this presentation, we will explain one such case that consists of an antenna resonating between two metal plates. We will discuss how we used the graphics processing unit (GPU) to separate the problem into two parts. Each part is simulated individually with SBR producing an improved result. Such a GPU-accelerated, two-part approach can be applied to other more general hybrid simulations.

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.

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.

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.

S0152 - Accurate Sequence Alignment using Distributed Filtering on GPU Clusters
Reza Farivar (University of Illinois at Urbana-Champaign) , Shivaram Venkataraman (University of Illinois at Urbana Champaign)
Day: Tuesday, 05/15 | **Time:** 3:30 pm - 3:55 pm
Topic Areas: Bioinformatics; Algorithms & Numerical Techniques
Session Level: Intermediate

Learn how GPUs enable new ways to rethink a complex bioinformatics problem: Accurate sequence alignment. What was once prohibitive to compute can become the basic block of novel GPU-based algorithms. Modern DNA sequencing machines generate enormous amounts of short sequences within minutes, and they should be aligned to a reference genome in real time. Most solutions only find a few locations that match a short sequence. We introduce a new technique to find all matching locations inside a reference sequence for a given number of mismatches. Our technique is based on a distributed filtering scheme and GPU based processing.

S0221 - 1024 Bit Parallel Rational Arithmetic Operators for the GPU
Robert Zigon (Beckman Coulter)
Day: Tuesday, 05/15 | **Time:** 4:00 pm - 4:50 pm
Topic Areas: Algorithms & Numerical Techniques; Computational Physics
Session Level: Intermediate

Learn how to create a set of rational arithmetic operators that manipulate 1024 bit operands on a Tesla C2050. These operators are used to create a numerically stable implementation for Bessel functions. Naive implementations of the Bessel functions produce unreliable results when they are used to solve Maxwell's equations by way of Mie theory. Maxwell's equations are used to model the scattering of light by small particles. Light scatter is used in Particle Characterization to measure the quality of materials like cocoa, cement and pharmaceuticals.

S0050 - High Performance Logic Simulation with GPUs
Yangdong Deng (Tsinghua University)
Day: Tuesday, 05/15 | **Time:** 4:00 pm - 4:50 pm
Topic Areas: General Interest; Algorithms & Numerical Techniques
Session Level: Advanced

Verification has become the bottleneck of IC design process due to its fast increasing complexity. The fundamental means of verifying digital circuits is logic simulation, which can be performed at both register-transfer level (RTL) and gate level. In this work, we developed GPU based logic simulation solutions. We implemented a Chandy-Misra-Bryant parallel simulation protocol on GPUs for sufficient parallelism. A dynamic GPU memory allocator was introduced to efficiently manage GPU memory resources. RTL simulation is performed in a compiled-code scheme by translating Verilog code into equivalent CUDA code. Experimental results proved that the GPU simulators significantly outperform their CPU counterparts.

S0314 - Efficient k-Nearest Neighbor Search Algorithms on GPUs
Nikos Pitsianis (Aristotle University),
Day: Tuesday, 05/15 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Machine Learning & AI; Databases, Data Mining, Business Intelligence; Algorithms & Numerical Techniques
Session Level: Beginner

Come see how to select the k smallest elements from an unsorted list. We present a selection and combination of different algorithms that perform exact k-nearest neighbors search (k-NNS) on GPUs and outperform the competition. In this session we present four different selection algorithms designed to exploit differently the parallelization of the GPU according to the relative size of the corpus data set, the size of the query set and the number of neighbors sought. We show the application of Logo Retrieval with SIFT vector matching on two different GPUs, the Tesla C1060 and the Fermi GTX480.

S0104 - GPU Implementation of Deep Learning for Intelligent Computer Vision
Ben Goertzel (Novamente LLC)
Day: Tuesday, 05/15 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Computer Vision; Algorithms & Numerical Techniques
Session Level: Advanced

Learn how to use GPU supercomputing for intelligent computer vision, via deep learning algorithms. We will focus on a case study of visual object and event recognition in a humanoid robotics context, involving a port to CUDA of the DeSTIN "compositional spatiotemporal deep learning network" vision processing algorithm (originally implemented at the University of Tennessee in Knoxville for conventional serial computers). The audience will learn how to use the open-source DeSTIN CUDA code, and also how to port other deep learning algorithms to CUDA.

S0219 - Efficient Top-Down Planning in Business Intelligence
Tobias Lauer (Jedox AG), Alexander Haberstroh (Jedox AG)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:25 pm
Topic Areas: Databases, Data Mining, Business Intelligence; Finance; Algorithms & Numerical Techniques
Session Level: Intermediate

In business intelligence, tasks like corporate planning or what-if analysis complement traditional reporting and analysis. One main difference is that while the latter only read data, the former require the change of possibly large numbers of existing and creation of new data records in the business model, preferably in real time. In this session, we describe the extension of an existing BI tool, Jedox OLAP, by GPU-based parallel algorithms for interactive planning scenarios. Compared to sequential in-memory algorithms, our CUDA approach yields tremendous speedups and can also cope with large amounts of data by using multiple GPUs.

S0247 - 3D ADI Method for Fluid Simulation on Multiple GPUs
Nikolai Sakharnykh (NVIDIA), Nikolay Markovskiy (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:50 pm
Topic Areas: Algorithms & Numerical Techniques; Computational Fluid Dynamics
Session Level: Intermediate

Find out about a multiple GPU implementation of the Alternating Direction Implicit method for large 3D domains. The ADI technique is applied towards direct numerical fluid simulation. Modeling complex flows demands extremely large grids and a distributed computation is required for sharing the memory among multiple GPUs. In this session a novel distributed tridiagonal solver as well as parallelization and load balancing strategies will be covered in detail. Finally, a comprehensive performance analysis and scaling studies for different input geometries and possible future improvements will be discussed.

S0037 - SeqNFind™: Application Of CUDA GPU Technologies To Sequence Alignment Techniques
D. Andrew Carr (Accelerated Technology Laboratories, Inc.)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:25 pm
Topic Areas: Bioinformatics; Algorithms & Numerical Techniques
Session Level: Advanced

Explosive growth in the amount of genomic data has created a need for faster systems that align and compare nucleotide sequences. With the development of tools for leveraging the massively parallel architecture of NVIDIA GPUs it is a logical next step to construct algorithms for genomic analysis on GPU clouds/clusters. Although a seemingly simple task, there are a number of challenges to deploying the current algorithms. Every algorithm from Smith-Waterman to BLAST has its own unique set of barriers. Presented here some of the lessons learned and how ongoing genomic research projects have benefitted from the increased speed and accuracy.

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.

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.

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.

S0383 - Speedup Derivatives and Structured Products Pricing, Reduce TCO Using GPUs
Steve Karmesin (Numerix)
Day: Wednesday, 05/16 | **Time:** 9:00 am - 9:50 am
Topic Areas: Finance; Algorithms & Numerical Techniques
Session Level: Intermediate

Numerix will share its experience using GPU to significantly reduce its customers' Total Cost of Ownership (TCO) and accelerate forward Monte Carlo pricing methods and hybrid models of complex financial structured products and variable annuities. Numerix will describe how it combines complex financial and actuarial modeling with user scripting to drive GPU execution from a script interpreted at run time. This architecture is well suited to financial services firms with portfolios of many different types of structured products where deals are represented independently from the models used to price them.

S0289 - Fine-Grained Parallel Preconditioners for Fast GPU-based Solvers
Dimitar Lukarski (Karlsruhe Institute of Technology (KIT), Jan-Philipp Weiss (Karlsruhe Institute of Technology)
Day: Wednesday, 05/16 | **Time:** 9:00 am - 9:25 am
Topic Areas: Algorithms & Numerical Techniques
Session Level: Advanced

Leverage the power of GPUs for efficient parallel solution of large sparse linear systems of equations by means of fine-grained and scalable parallel preconditioners. In this session we describe parallel preconditioners for GPUs based on multicolor re-ordering for Gauss-Seidel-type and ILU-type preconditioners as well as approximate inverse (FSAI) preconditioners. With the power(q)-pattern method we detail a novel method for controlling the fill-in pattern of ILU(p) factorizations that introduces a high degree of parallelism in the preconditioning phase. We demonstrate significant improvements with respect to solver time for various problem scenarios and different Krylov-type solvers.

S0415 - An Accelerated Weeks Method for Numerical Laplace Transform Inversion
Patrick Kano (Acunum Algorithms and Simulations, LLC)
Day: Wednesday, 05/16 | **Time:** 9:30 am - 9:55 am
Topic Areas: Algorithms & Numerical Techniques
Session Level: Beginner

Mathematical methods based on the use of the Laplace transform are a standard component of undergraduate education. Real world problems however often yield Laplace space solutions which are too complex to be analytically inverted to expressions in physically meaningful variables. A robust numerical inversion approach is thus desirable. In this talk, I present one of the approaches to compute an approximate inverse, the Weeks method. I will also discuss the difficulties in performing numerical inversion. Finally, I will show how we have been able to utilize Jacket from AccelerEyes in MATLAB to more efficiently and robustly implement the Weeks method.

S0262 - GPU-Accelerated Model-Based Drug Development
Chee Ng (Children Hospital of Philadelphia/University of Pennsylvania)
Day: Wednesday, 05/16 | **Time:** 10:00 am - 10:25 am

Topic Areas: Life Sciences; Algorithms & Numerical Techniques; Bioinformatics
Session Level: Beginner

Explore how GPUs can be used to improve the efficiency of drug development. Drug development is a very time-consuming, complex and expensive process that has low successful rate. A model-based drug development paradigm has been proposed as a possible solution to overcome these problems. A key challenge is to develop computational intensive drug and disease-specific models from a large quantity of highly complicated preclinical and clinical data. This session will describe how GPUs can and will play a key role in shortening the model development times and improving the efficiency of model-based drug development.

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

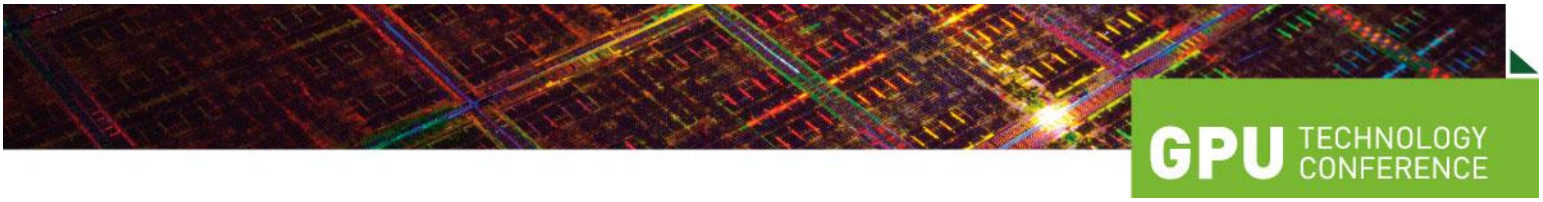
Linear systems are at the heart of allot of compute problems. In large sparse systems, there are 2 distinct approaches, the direct and iterative solvers. After many years of researching and testing both approaches, on CPU and GPU we have implemented a highly efficient CG solver on the GPU using a combination of unique techniques. In this talk we will go over these techniques and the improved performance they bring.

S0115 - Specialized Sparse Matrix Formats and SpMV Kernel Tuning for GPUs
Alexander Monakov (ISP RAS), Arutyun Avetisyan (ISP RAS)
Day: Wednesday, 05/16 | **Time:** 10:30 am - 10:55 am
Topic Areas: Algorithms & Numerical Techniques
Session Level: Intermediate

This session is focused on optimizing sparse matrix-vector product for NVIDIA GPUs. This is a frequently studied kernel that appears in applications employing iterative methods for solving systems of linear equations. In the majority of cases the computation is memory bandwidth bound. Our study focuses on developing specialized sparse matrix storage formats and corresponding CUDA SpMV implementation that achieves high performance at the cost of additional start-up time required for conversion and tuning. The proposed storage formats allow to reduce required memory bandwidth by providing compact coding for locations of some frequently observed patterns of non-zero elements.

S0209 - Performance of 3-D FFT Using Multiple GPUs with CUDA 4
Akira Nukada (Tokyo Institute of Technology)
Day: Wednesday, 05/16 | **Time:** 10:30 am - 10:55 am
Topic Areas: Algorithms & Numerical Techniques; Development Tools & Libraries
Session Level: Advanced

Get the latest information on performance of 3-D fast Fourier transform using multiple GPU devices. CUDA 4.0 enables efficient data transfer between GPUs. It is really important in FFT computation since it requires a large amount of all-to-all data exchange between GPUs. The peer-to-peer communication feature of GPUDirect V2 improves the communication between the devices on same node. GPUDirect also accelerates the communication between GPUs on different nodes. We will present the latest performance results on a four-GPU system and up to 128 compute nodes of TSUBAME 2.0.



S0029 - Leveraging Matrix Block Structure In Sparse Matrix-Vector Multiplication
Steve Rennich (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 2:00 pm - 2:25 pm
Topic Areas: Algorithms & Numerical Techniques
Session Level: Intermediate

The commonly occurring block structure of sparse matrices can be effectively leveraged to improve the performance of Sparse Matrix-Vector multiplication (SpMV) on GPUs. This session will present one such algorithm and discuss both its design and its performance relative to other SpMV algorithms. In particular, aspects of GPU floating point performance, GPU memory use, and datastructure translation effort will be detailed.

S0142 - VMD: High Performance Molecular Visualization and Analysis on GPUs
John Stone (University of Illinois at Urbana-Champaign)
Day: Wednesday, 05/16 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Molecular Dynamics; Algorithms & Numerical Techniques; Computer Graphics
Session Level: Intermediate

This talk will present recent successes in the use of GPUs to accelerate interactive molecular visualization and analysis tasks on desktop computers, and batch-mode simulation and analysis jobs on GPU-accelerated HPC clusters. We'll present Fermi-specific algorithms and optimizations and compare with those for other devices. We'll also present performance and performance/watt results for VMD analysis calculations on GPU clusters, and conclude with a discussion of ongoing work and future opportunities for GPU acceleration, particularly as applied to the analysis of petascale simulations of large biomolecular complexes and long simulation timescales.

S0307 - New Advances in GPU Linear Algebra
John Humphrey (EM Photonics), Kyle Spagnoli (EM Photonics)
Day: Wednesday, 05/16 | **Time:** 2:00 pm - 2:25 pm
Topic Areas: Algorithms & Numerical Techniques
Session Level: Intermediate

Hear product experts explain how we have created two of the most widely used libraries in the GPU computing ecosystem. The CULA library for dense linear algebra has been expanding to multi-GPU and out-of-core applications, meaning that users are no longer limited by the onboard GPU memory for their work. In this field, effectively using multiple GPUs is significantly more challenging than a single GPU! The brand new CULA Sparse library tackles the tough world of sparse linear algebra and achieves 10x speedups. Learn more about what makes these two libraries work in this session.

S0085 - Floating Point and IEEE 754 Compliance for NVIDIA GPUs: Precision & Performance
Alex Fit-Florea (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 2:30 pm - 2:55 pm
Topic Areas: Algorithms & Numerical Techniques; Development Tools & Libraries
Session Level: Intermediate

As a result of continuing improvements, NVIDIA offers GPU-accelerated floating-point performance in compliance with IEEE 754. It is our experience that a number of issues related to floating point accuracy and compliance are a frequent source of confusion both on CPUs and GPUs. The purpose of this talk is to discuss the most common ones related to NVIDIA GPUs and to supplement the documentation in the CUDA C Programming Guide

S0143 - Fluid-Structure-Interaction Using SPH and GPGPU Technology
Jean Luc Lacome (IMPETUS Afea SAS), Jerome Limido (IMPETUS Afea SAS)
Day: Wednesday, 05/16 | **Time:** 2:30 pm - 2:55 pm
Topic Areas: Computational Structural Mechanics; Algorithms & Numerical Techniques; Computational Fluid Dynamics
Session Level: Intermediate

There are two goals when developing engineering analysis software, one is accuracy and the other is speed. In the area of Fluid-Structure Interaction (FSI) computational time has always been the major impediment to solving large realistic engineering problems. In our implementation the fluid/structural dynamics solver uses a combination of GPU/CPU processing. The added benefit of using a powerful GPU workstation is that it is roughly 10 times less expensive than a regular CPU cluster. In this paper, we present the use of GPU Technology as implemented in the explicit dynamic finite element software IMPETUS Afea Solver®.

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.

S0271 - Fast Adaptive Sampling Technique for Multi-Dimensional Integral Estimation Using GPUs
Pradeep Rao (Infosys Technologies Ltd.), Srinivasa Prasanna (Internation Institute of Information Technology Bangalore)
Day: Wednesday, 05/16 | **Time:** 2:30 pm - 2:55 pm
Topic Areas: Algorithms & Numerical Techniques; Finance
Session Level: Intermediate

Evaluating multi-dimensional integrals is a commonly encountered problem in many areas of science including Physics and Volume estimation of convex bodies. One of the widely used techniques for integral evaluation in large dimensions is the Monte Carlo method. Vanilla Monte Carlo methods of Integral Estimation use uniform sampling techniques. Variance of such uniform sampling reduces as 1/√Sample-size, which is too slow for most real life applications. In this study, we discuss about an adaptive sampling technique called VEGAS which reduces the variance at a much faster rate than uniform sampling. We present a new parallel implementation for VEGAS based on CUDA that can significantly reduce the computation time of multi-dimensional integrals. We show that our GPU based implementation of VEGAS achieves up to a 45x speed up over an equivalent CPU based implementation.

S0035 - GPU Parallelization of Gibbs Sampling: Abstractions, Results, and Lessons Learned
Alireza Mahani (Sentrana)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Algorithms & Numerical Techniques; Databases, Data Mining, Business Intelligence
Session Level: Intermediate

Monte-Carlo-Markov-Chain (MCMC) estimation of Hierarchical Bayesian (HB) models is not only time-consuming, but also difficult to parallelize due to its sequential (Markovian) nature. We present an abstraction of a widely-

used MCMC algorithm, called Gibbs sampling. We define a taxonomy of variable blocks, and for each type of variable block we offer suitable parallelization strategies, along with their corresponding CUDA implementations. For large problems where model estimation may take several hours or days using a single-threaded software, we see speedups in the 30x-100x range, thereby reducing estimation time to a few hours. In addition to lower computation cost relative to MPI-based parallelization, the reduction in estimation time allows for a more interactive modeling experience. We offer an extensive discussion of lessons learned for the broader scientific computing field, including an analysis of tradeoffs between computation costs and development costs, implications of our tradeoff analysis for optimal software development and parallelization, and some practical tips and gotcha's for rookie GPU programmers.

S0042 - Solving Challenging Numerical Linear Algebra Algorithms using Multiple GPU Accelerators
Hatem Ltaief (KAUST Supercomputing Laboratory), Stanimire Tomov (University of Tennessee)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Algorithms & Numerical Techniques; Development Tools & Libraries
Session Level: Intermediate

See the newest features integrated in MAGMA (Matrix Algebra on GPU and Multicore Architectures) to tackle the multiple GPU-based systems for numerical linear algebra. In this talk, we describe how we leveraged MAGMA to solve existing and new challenging numerical problems on multiple hardware accelerators. Using a hybridization methodology, the new multiGPU-enabled MAGMA is characterized by a representation of linear algebra algorithms as directed acyclic graphs, where nodes correspond to tasks and edges to data dependencies among them, and a dynamic runtime system environment StarPU used to schedule various computational kernels over hybrid architectures of GPUs and homogeneous multicores.

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.

S0269 - Accelerating 3D-RISM Calculations using GPUs
Yutaka Maruyama (Institute for Molecular Science), Fumio Hirata (Institute for Molecular Science)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Life Sciences; Algorithms & Numerical Techniques; Computational Physics
Session Level: Intermediate

The three-dimensional reference interaction site model (3D-RISM) theory, is a powerful tool to investigate biomolecular processes in solution. Unfortunately, 3D-RISM calculations are often both memory intensive and time-consuming. We sought to accelerate these calculations using GPUs. To work around the problem of limited memory size in GPUs, we modified the less memory-intensive Anderson method for faster convergence of 3D-RISM calculations. Using this method on C2070, we reduced the computational time by a factor of eight compared to

Intel Xeon (8 cores, 3.33GHz) with the conventional method.

S0214 - GPU Based Stacking Sequence Optimization For Composite Skins Using GA
Sathya Narayana K. (Infosys Ltd.), Ravikumar G.V.V. (Infosys Ltd, Bangalore)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Computational Structural Mechanics; Algorithms & Numerical Techniques; Parallel Programming Languages & Compilers; Algorithms & Numerical Techniques
Session Level: Advanced

The goal of this session is to showcase how GPUs can be used to achieve high performance in a Genetic algorithm based optimization. The particular domain applied is stacking sequence optimization of Aircraft wing skins. The concepts illustrated use CUDA but are generic to any other GPU language. It is assumed that the registrants have exposure to optimization in engineering domain.

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.

S0432 - New Ideas for Massively Parallel Preconditioners
John Appleyard (Polyhedron Software Ltd.), Jeremy Appleyard (Polyhedron Software Ltd.)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Algorithms & Numerical Techniques; Computational Fluid Dynamics; Energy Exploration
Session Level: Advanced

Linear Solvers on serial machines tend to be highly recursive, but that's not an option on GPUs. In this paper we describe a new preconditoner for GMRES and similar Krylov subspace linear solvers that is highly parallel, but also provides effective mechanisms to reconcile remote driving forces in a spatially discretized system. We will present results, taken from some real-world studies using a commercial oil reservoir simulator, showing how it compares with a state of the art serial solver, and showing how performance scales in a domain decomposition formulation run on a multiple CPU+GPU cluster.

S0293 - Culises - A Library for Accelerated CFD on Hybrid GPU-CPU Systems
Daniel Gaudlitz (FluiDyna GmbH), Bjorn Landmann (FluiDyna GmbH)
Day: Wednesday, 05/16 | **Time:** 3:30 pm - 3:55 pm
Topic Areas: Computational Fluid Dynamics; Algorithms & Numerical Techniques
Session Level: Intermediate

The vast majority of CFD simulations relies on the solution of large-scale systems of linear equations (SLE), where the solution of a system can consume most of the total CPU time. We have developed a library (Culises) for state-

of-the-art solution of SLE that is targeted on hybrid GPU-CPU platforms. Culises can be connected to MPI-parallelized CFD codes (e.g. OpenFOAM) via an application-specific interface. In this talk, we focus on efficient implementation of preconditioned Krylov subspace methods. Using the computing power of GPUs, Culises can significantly accelerate pure CPU computations for a multitude of industrial CFD applications.

S0207 - GPU Enabled Macromolecular Simulation: Challenges and Opportunities
Michela Taufer (University of Delaware), Sandeep Patel (University of Delaware)
Day: Wednesday, 05/16 | **Time:** 3:30 pm - 3:55 pm
Topic Areas: Molecular Dynamics; Algorithms & Numerical Techniques
Session Level: Advanced

GPU enabled simulation of fully atomistic macromolecular simulation is rapidly gaining momentum, enabled by the massive parallelism and due to parallelizability of various components of the underlying algorithms and methodologies. The massive parallelism in the order of several hundreds to few thousands of cores, presents opportunities as well poses implementation challenges. In this talk dive deep into the various key aspects of simulation methodologies of macro molecular systems specifically adapted to GPUs. Learn some of the underlying challenges and get the latest solutions devised to tackle them in the FEN ZI code for fully atomistic macromolecular simulations.

S0206 - Monte-Carlo Pricing Under a Hybrid Local Volatility Model
Sebastien Gurrieri (Mizuho International)
Day: Wednesday, 05/16 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Finance; Algorithms & Numerical Techniques
Session Level: Intermediate
This session shows how to calculate the prices of several financial products, vanilla and exotic, under Dupireâ€™s Local Volatility model. We start with vanilla options on the foreign exchange rate and explain how to rescale the Local Volatility matrix in order to take advantage of the fast texture memory interpolation. We then extend this framework to two factors by including stochastic interest rates following Hull-White model, and show how to price Power-Reverse Dual Coupon swaps with an exotic TARN feature. We provide details of the algorithms and compare accuracy and speed with typical performances of single-core production implementations.

S0131 - Multi-GPU Real-Time Ptychographic X-ray Image Reconstruction
Filipe Maia (Lawrence Berkeley National Laboratory)
Day: Wednesday, 05/16 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Audio, Image and Video Processing; Algorithms & Numerical Techniques
Session Level: Intermediate

Learn how a new imaging technique, combined with the computational power of GPUs and the brightness of modern X-ray synchrotrons can quickly and easily produce images with nanometer level resolution. Ptychography is a recent X-ray imaging technique in which overlapping regions of a sample are exposed in quick succession and the resulting scattering is used to reconstruct a high resolution image of the sample. Discover why GPUs can substitute for the lack of X-ray lenses and how they enabled a dramatic reduction in the feedback time for users of the technique from days to seconds.

S0149 - On the Parallel Solution of Sparse Triangular Linear Systems
Maxim Naumov (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Algorithms & Numerical Techniques; Development Tools & Libraries
Session Level: Intermediate

A parallel algorithm for solving a sparse triangular linear system on the GPU is proposed. It implements the solution of the triangular system in two phases. The analysis phase builds a dependency graph based on the matrix sparsity pattern and groups the independent rows into levels. The solve phase obtains the full solution by iterating sequentially across the constructed levels. The solution elements corresponding to each level are obtained in parallel. The numerical experiments are presented and it is shown that the incomplete-LU and Cholesky preconditioned iterative methods can achieve a 2x speedup on the GPU over their CPU implementation.

S0332 - Efficient Graph Matching and Coloring on the GPU
Patrice Castonguay (NVIDIA), Jonathan Cohen (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Algorithms & Numerical Techniques
Session Level: Intermediate

The goal of this session is to compare the performance of graph matching and graph coloring algorithms on massively parallel devices such as GPUs. We present novel algorithms, which produce superior results for certain graphs and also discuss the techniques used to efficiently implement these algorithms on the GPU.

S0273 - Fast JPEG Coding on the GPU
Fyodor Serzhenko (Fastvideo)
Day: Wednesday, 05/16 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Audio, Image and Video Processing; Algorithms & Numerical Techniques
Session Level: Advanced

The goal of this session is to demonstrate how high speed JPEG compression and decompression can be efficiently implemented on the GPU using CUDA. In this session we will present: detailed analysis of Baseline JPEG compression and decompression processes and its constituent parts (such as Huffman Coding, RLE, Differential Coding, Quantization, Discrete Cosine Transform) and their suitability for the GPU architecture, analysis of achieved results and comparison with existing implementations, applications to high-speed imaging.

S0241 - Large Graphs on Multi-GPUs
Enrico Mastrostefano (Sapienza Università di Roma)
Day: Wednesday, 05/16 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Algorithms & Numerical Techniques
Session Level: Intermediate

The goal of this session is to propose new paradigms to explore large graphs on GPUs. Graphs with billions of edges don't fit within the memory of a single GPU. A possible solution is to resort to multiple GPUs. Most of common graph algorithms show low arithmetic intensity and irregular access patterns. These features lead to a poor load balance among threads and un-coalesced access to memory. We show how to balance the load to exploit as much as possible all threads and then how to use fast algorithms, as radix-sort and scan, to rearrange data before process them.

S0070 - GPU-Friendly Preconditioners for Thin Structure Analysis
Krishnan Suresh (University of Wisconsin)
Day: Wednesday, 05/16 | **Time:** 5:00 pm - 5:25 pm
Topic Areas: Computational Structural Mechanics; Algorithms & Numerical Techniques
Session Level: Intermediate

The goal of this session is to identify a niche area within computational structural mechanics, namely finite element analysis of thin structures (such as beams, plates and shells) where GPU computing can make a significant impact. We will discuss methods for: (1) extracting equivalent lower-dimensional models from complex 3-D structures, and (2) using these lower-dimensional models as pre-conditioners for a full 3-D finite element analysis on the GPU. It is assumed that registrants are familiar with finite element analysis, and some of the underlying challenges.

S0410 - Computing Hausdorff Distances between Freeforms on the GPU
Sara McMains (UC Berkeley), Adarsh Krishnamurthy (UC San Diego)
Day: Wednesday, 05/16 | **Time:** 5:00 pm - 5:25 pm
Topic Areas: Algorithms & Numerical Techniques; Computer Graphics; Computer Vision
Session Level: Intermediate

We present new GPU algorithms for computing the directed Hausdorff distance between freeform surfaces, with applications in shape matching, mesh simplification, and geometric approximation and optimization. Our algorithms run in real-time with very small error bounds for parametric models defined by complex NURBS surfaces and can be used to interactively compute the Hausdorff distance for models made of dynamic deformable surfaces. We discuss implementation decisions and tradeoffs between OpenGL, Cuda, and Thrust, and the advantages and disadvantages of parallel hierarchical culling methods for this application.

S0096 - Summed Area Ripmaps
Gernot Ziegler (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 5:30 pm - 5:55 pm
Topic Areas: Algorithms & Numerical Techniques; Computer Vision; Computer Graphics
Session Level: Intermediate

In this presentation, we show how ripmaps can replace Summed Area Tables (SATs) for the purpose of computing a large number of spatially varying box filter kernels throughout the input data, providing both higher accuracy and higher speed for typical use cases. For this purpose, we demonstrate an implementation of ripmap generation in CUDA C (accelerated by shared memory usage), and a texture-cache based box filter for spatially varying kernel sizes, which can be implemented in both CUDA C and graphics-based APIs (e.g. OpenGL and DirectX).

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.

S0302 - Accelerating miniFE: A Finite Element Mini-application
Justin Luitjens (NVIDIA)
Day: Thursday, 05/17 | **Time:** 9:00 am - 9:25 am
Topic Areas: Algorithms & Numerical Techniques
Session Level: Intermediate

The Mantevo performance project is a collection of self-contained proxy applications that illustrate the main performance characteristics of important algorithms. miniFE is intended to be and approximation to an unstructured implicit finite element or finite volume application. Our work investigated algorithms for assembling a matrix on the GPU. Parallelization algorithms using both 1 thread and 8 threads per element were investigated. Using these approaches a significant speedup (over 60x for double precision) compared to the serial algorithm.

S0264 - CU++: An Object-Oriented Framework for Computational Fluid Dynamics (CFD) Applications
Dominic Chandar (University of Wyoming)
Day: Thursday, 05/17 | **Time:** 9:30 am - 9:55 am
Topic Areas: Computational Fluid Dynamics; Algorithms & Numerical Techniques
Session Level: Intermediate

In this session, I will elucidate the power of blending C++ expression templates and CUDA which has resulted in a smart framework - CU++ for solving Computational Fluid Dynamics problems on structured and unstructured meshes. Briefly, CU++ allows a code developer with just C/C++ knowledge to write computer programs that will execute on the GPU with minimal knowledge of specific programming techniques in CUDA. It allows the user to reuse existing C/C++ CFD codes with minimal changes. Codes written in CU++ can also be compiled in serial mode to be executed on a CPU using the tool ugc.

S0290 - Algorithm Acceleration for Geospatial Analysis
James Goodman (HySpeed Computing LLC), Matthew Sellitto (Northeastern University)
Day: Thursday, 05/17 | **Time:** 9:30 am - 9:55 am
Topic Areas: Algorithms & Numerical Techniques; General Interest
Session Level: Intermediate

Learn how the power of GPU computing is being leveraged to accelerate algorithms in the field of geospatial image analysis. The data volume and computation requirements associated with geospatial imagery are rapidly expanding as a result of the increasing number of satellite and airborne sensors, greater data accessibility, and expanded utilization of data intensive technologies. This equates to a growing need for high-performance computing in this field. We demonstrate the capacity for GPU computing to meet this need by accelerating a complex non-linear optimization algorithm used for the mapping and assessment of coral reef ecosystems.

S0324 - Content Generation and Real-Time Hologram Computation for Holographic 3D-Displays
Enrico Zschau (SeeReal Technologies GmbH)
Day: Thursday, 05/17 | **Time:** 10:00 am - 10:25 am
Topic Areas: Visualization; Stereoscopic 3D; Algorithms & Numerical Techniques; Audio, Image and Video Processing
Session Level: Beginner

This session will introduce SeeReal's sub-hologram technology to massively reduce hologram computation effort in comparison to classic holography and how SeeReal implemented those still compute intensive algorithms to execute on the GPU to enable viewing of interactive, rich 3D-content on holographic 3D-displays using off-the-shelf graphics hardware. In contrast, you will explore why classic holography does not suit well for interactive applications. Furthermore guidelines to create appropriate 3D-content are presented, including aspects regarding

transparency in holograms. Finally the specification and some impressions of SeeReal's 20 holographic prototype will be presented, which allows viewing of live computed holograms showing 3D-content and 3D-video.

S0305 - Classical Algebraic Multigrid for CFD with CUDA
Simon Layton (Boston University)
Day: Thursday, 05/17 | **Time:** 10:00 am - 10:25 am
Topic Areas: Computational Fluid Dynamics; Algorithms & Numerical Techniques
Session Level: Intermediate

Classical algebraic multigrid (AMG) is one of the most popular algorithms used in engineering, and the engine in many successful commercial packages. Among sparse linear solvers, it is known for being fast, parallel and scalable, yet it maps to GPU architecture with some considerable difficulty. We have tackled these difficulties and currently have a full CUDA implementation of classical AMG, which has been validated against the gold-standard, Hypre. Significant effort was dedicated to reducing thread divergence and optimizing memory access, and we continue to work on performance improvements. We are aiming for a competitive AMG code for fluid dynamics applications.

S0508 - Faster Finite Elements for Wave Propagation Codes
Max Rietmann (Institute for Computational Science / USI Lugano, Switzerland)
Day: Thursday, 05/17 | **Time:** 10:00 am - 10:25 am
Topic Areas: Algorithms & Numerical Techniques; Computational Physics
Session Level: Intermediate

Learn how to develop faster and better finite-element codes for wave propagation using GPUs and MPI combined with overlapping techniques to hide the cost of communications and of host/device memory copies. Different options based on mesh coloring or on atomic operations will be presented. The difficulty to define speedup will also be discussed (speedup versus what? using what definition of "cost"?). Examples will be given using SPECFEM3D, a highly optimized spectral finite-element code that has won the Gordon Bell SuperComputing award and the BULL Joseph Fourier award, and that can run on CPU or GPU clusters.

S0079 - Warped Parallel Nearest Neighbor Searches using KD-Trees
Andrei Tchouprakov (D4D Technologies), Roman Sokolov (D4D Technologies)
Day: Thursday, 05/17 | **Time:** 10:30 am - 10:55 am
Topic Areas: Algorithms & Numerical Techniques
Session Level: Intermediate

We propose a nearest neighbor search algorithm for a set of closely located query points that utilizes GPU parallelism and is optimized for a single CUDA warp. Instead of each query point traversing its own distinct path, a combined non-divergent path suitable for the entire query set can constructed. Therefore, for a single warp a single stack can be maintained for the entire set of query points, allowing for efficient utilization of the shared memory and a number of simultaneous queries equal to the number of threads in a warp.

S0054 - PFAC Library: GPU-Based String Matching Algorithm
Lung-Sheng Chien (National Tsing Hua University), Cheng-Hung Lin (National Taiwan Normal University)
Day: Thursday, 05/17 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Development Tools & Libraries; Algorithms & Numerical Techniques
Session Level: Beginner

In this section, we first propose an exact string matching algorithm, called Parallel-Failureless Aho-Corasick (PFAC) algorithm which is used to match input texts against a set of string patterns on GPUs. The string patterns are compiled into a finite state machine similar to the well-known Aho-Corasick algorithm. Furthermore, to accommodate large number of patterns, we present two kinds of hash functions which are adopted to compress the state transition table. The experimental results show that the PFAC library achieves significant performance on NVIDIA GPUs. Finally, the PFAC library has been released on Google code (http://code.google.com/p/pfac/).

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.

S0285 - Optimization of a Sparse Matrix-Matrix Multiplication on the GPU
Julien Demouth (NVIDIA)
Day: Thursday, 05/17 | **Time:** 2:00 pm - 2:25 pm
Topic Areas: Algorithms & Numerical Techniques
Session Level: Advanced

The goal of this session is to present advanced techniques to optimize CUDA code on the GPU. In particular, we will demonstrate the use of advanced CUDA instructions (inline PTX, warp instructions, "extended" syncthreads) and load-balancing strategies to improve the performance of a sparse matrix-matrix multiplication on the GPU.

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.

S0071 - The High-Level Linear Algebra Library ViennaCL And Its Applications
Karl Rupp (TU Wien)
Day: Thursday, 05/17 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Development Tools & Libraries; Algorithms & Numerical Techniques; Computational Physics
Session Level: Intermediate

Get to know ViennaCL, an OpenCL high-level linear algebra software, which allows to get the speed of GPU computing at the convenience level of the C++ Boost libraries. Decrease the development and execution time of applications by utilizing our well-tested and widely used library, instead of spending days on learning details of GPU architectures and debugging. We provide examples that demonstrate not only how quickly existing applications are ported efficiently from single-threaded execution to fully utilizing multi-threaded environments, but also how to utilize the rich set of functionalities ranging from common BLAS routines to iterative solvers.

S0087 - GPU Acceleration of Dense Stellar Clusters Simulation
Bharath Pattabiraman (Northwestern University), Stefan Umbreit (Northwestern University)
Day: Thursday, 05/17 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Astronomy & Astrophysics; Computational Physics; Algorithms & Numerical Techniques
Session Level: Intermediate

Computing the interactions between stars within dense stellar clusters is a problem of fundamental importance in theoretical astrophysics. This paper presents the parallelization of a Monte Carlo algorithm for simulating stellar cluster evolution using programmable Graphics Processing Units. The kernels of this algorithm exhibit high levels of data dependent decision making and unavoidable non-contiguous memory accesses. However, we adopt various parallelization strategies and utilize the high computing power of the GPU to obtain substantial near-linear speedups which cannot be easily achieved on a CPU-based system. This acceleration allows to explore physical regimes which were out of reach of current simulations.

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).

S0334 - The Fast Multipole Method on CPU and GPU Processors
Toru Takahashi (Nagoya University), Cris Cecka (Harvard University)
Day: Thursday, 05/17 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Computational Physics; Molecular Dynamics; Algorithms & Numerical Techniques
Session Level: Advanced

The fast multipole method (FMM) is a widely used numerical algorithm in computational engineering. Accelerating the FMM on CUDA-enabled GPUs is challenging because the FMM has a complicated data access pattern, mostly during the so-called multipole-to-local (M2L) operation. We have created several schemes to optimize the M2L and have attained a performance of over 350 (resp. 160) Gflop/s for single (double) precision arithmetic. The optimal algorithm was incorporated into a complete FMM code, which can accept any smooth kernel as specified by the user, making it very flexible. We have also developed a highly efficient CPU version.

S0063 - Robust Preconditioned Conjugate Gradient for the GPU and Parallel Implementations
Rohit Gupta (Delft University of Technology)
Day: Thursday, 05/17 | **Time:** 4:00 pm - 4:50 pm
Topic Areas: Computational Fluid Dynamics; Algorithms & Numerical Techniques
Session Level: Intermediate

Get a closer look on how parallel conjugate gradient (CG) method can get an edge over it's optimized CPU implementation. We have developed preconditioning techniques for CG which are suited to the GPU and match Block-IC in terms of numerical performance. We present our results for two level preconditioned CG on the GPU and also compare it with multi-CPU implementations. Our results show that for large problem sizes(1 million unknowns and above) it is possible to achieve an order of magnitude and higher speedups for the two level preconditioned CG method.

S0292 - MultiGPUs Simulation of Lattice QCD with Domain-Wall Fermion
Ting-Wai Chiu (National Taiwan University)
Day: Thursday, 05/17 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Computational Physics; Algorithms & Numerical Techniques
Session Level: Advanced

To understand the nature of the strong interaction in the subatomic regime is a grand challenge in science. Now we know that the fundamental theory for the strong interaction is Quantum Chromodynamics (QCD). However, starting from the action of QCD, it is very computationally demanding to extract physical observables in QCD, which always requires the state-of-the-art supercomputers. In this talk, I outline the salient features of QCD which are relevant to HPC, and explain how GPU can serve as a vital device for large-scale QCD simulations.

S0411 - Artifact-Free Cloud-Based CAD Rendering
Sara McMains (UC Berkeley), Sushrut Pavanaskar (UC Berkeley)
Day: Thursday, 05/17 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Algorithms & Numerical Techniques; Computer Graphics; Cloud Computing; Visualization
Session Level: Beginner

Cloud computing for mechanical CAD provides centrally stored and synchronized models for concurrent engineering. For compactness, trimmed parametric NURBS surface representations are optimal for data transfer to client devices, which must evaluate and render models locally. Direct GPU rendering without pre-tessellation is an attractive solution in this context, both for speed and to preserve fidelity to the original geometry. However, existing data-parallel direct rendering approaches for NURBS suffer from rendering artifacts at trim boundaries. This talk proposes a solution to address these rendering artifacts that are still preventing wide-scale adoption of all such direct rendering algorithms for trimmed parametric models.