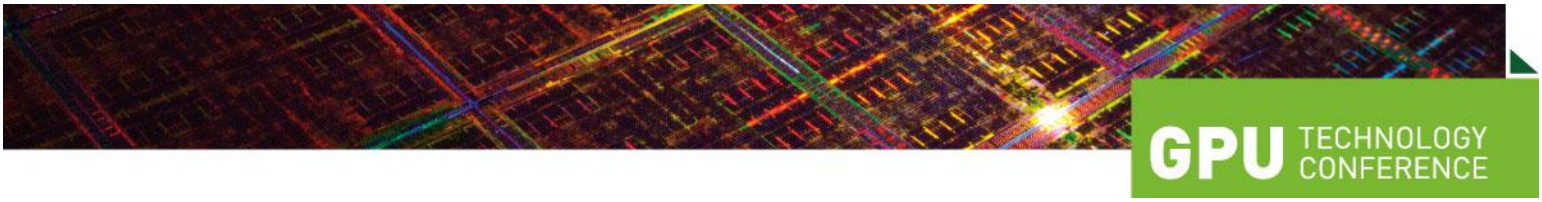


GPU Technology Conference, May 14-17, 2012
McEnery Convention Center, San Jose, California
www.gputechconf.com

Sessions on Development Tools & Libraries (subject to change)

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

TUTORIALS

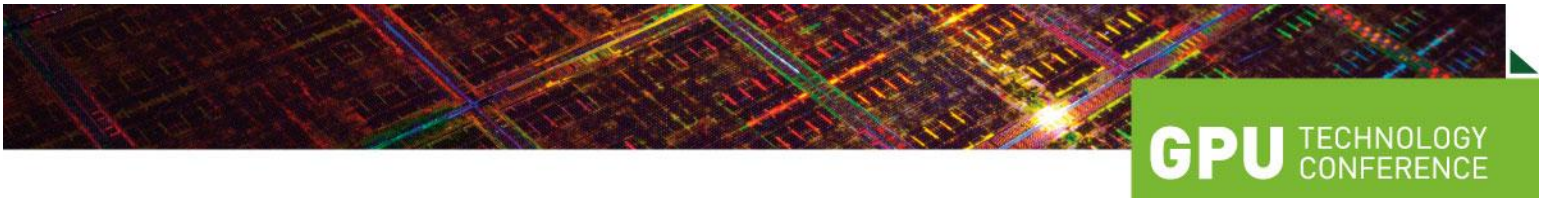
S0005 - Languages, APIs and Development Tools for GPU Computing
Will Ramey (NVIDIA)
Day: Monday, 05/14 | **Time:** 9:00 am - 10:30 am
Topic Areas: General Interest; Development Tools & Libraries; Application Design & Porting Techniques
Session Level: Beginner

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

S0614 - Part 1: Introduction to GPU Programming *(Presented by Acceleware)*
Chris Mason (Acceleware)
Day: Monday, 05/14 | **Time:** 9:00 am - 10:30 am
Topic Areas: Supercomputing; Development Tools & Libraries
Session Level: Beginner

Join us for an informative introduction to GPU Programming. The session will begin with a brief overview of CUDA and data-parallelism before focusing on the GPU programming model. We will explore the fundamentals of GPU kernels, host and device responsibilities, CUDA syntax and thread hierarchy. A programming demonstration of a simple CUDA kernel will be provided. Introduction to GPU Programming • CUDA overview • Data-parallelism • GPU programming model o GPU kernels o Host vs. device responsibilities • CUDA syntax • Thread hierarchy • Programming Demo: Simple CUDA Kernels

S0023 - NVIDIA OpenGL for 2012
Mark Kilgard (NVIDIA)
Day: Monday, 05/14 | **Time:** 9:00 am - 10:30 am
Topic Areas: Computer Graphics; Development Tools & Libraries; Visualization; Audio, Image and Video Processing
Session Level: Intermediate
Attend this session to get the most out of OpenGL on NVIDIA Quadro and GeForce GPUs. Topics covered include the latest advances available for Cg 3.1, the OpenGL Shading Language (GLSL); programmable tessellation; improved support for Direct3D conventions; integration with Direct3D and CUDA resources; bindless graphics; and more. When you utilize the latest OpenGL innovations from NVIDIA in your graphics applications, you benefit from NVIDIA's leadership driving OpenGL as a cross-platform, open industry standard.



S0615 - Part 2: Introduction to the GPU Architecture and Memory Model *(Presented by Acceleware)*
Chris Mason (Acceleware)
Day: Monday, 05/14 | **Time:** 10:30 am - 12:00 pm
Topic Areas: Supercomputing; Development Tools & Libraries
Session Level: Beginner
Explore the memory model of the GPU. The first part of the session covers task parallelism and thread cooperation in GPU computing. The second part focuses on the different memory types available on the GPU. We will define shared, constant and global memory and discuss the best locations to store your application data for optimized performance. A programming demonstration of shared memory will be delivered. Introduction to the GPU Architecture and Memory Model • Task parallelism • Thread cooperation in GPU computing • GPU memory model o Shared memory • Constant memory • Global memory • Programming Demo: Shared Memory

S0616 - Part 3: Debugging GPU Programs *(Presented by Acceleware)*
Chris Mason (Acceleware)
Day: Monday, 05/14 | **Time:** 1:00 pm - 2:30 pm
Topic Areas: Supercomputing; Development Tools & Libraries
Session Level: Beginner

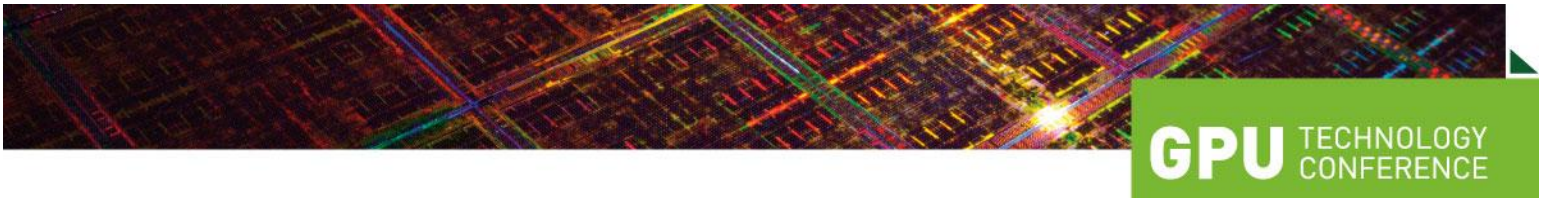
Get the low down on debugging your GPU program. This session includes discussion on debugging techniques and tools to help you identify issues in your kernels. The latest debugging tools provided in CUDA 4.1 including Parallel NSight, cuda-gdb and cuda-memcheck will be discussed. A programming demonstration of Parallel NSight will be provided. Debugging GPU Programs • Debugging tools and techniques • cuda-gdb • Parallel NSight • cuda-memcheck • Programming Demo: Parallel NSight

S0617 - Part 4: Introduction to Optimizations and Profiling *(Presented by Acceleware)*
Chris Mason (Acceleware)
Day: Monday, 05/14 | **Time:** 2:30 pm - 4:00 pm
Topic Areas: Supercomputing; Development Tools & Libraries
Session Level: Beginner

Learn how to optimize and profile your algorithms for the GPU. This session will cover the essentials of code optimization and will include: arithmetic optimizations, warps, branching efficiency, memory latency/occupancy and memory performance optimizations. Real life commercial examples will be discussed to highlight the critical aspects of GPU optimization techniques. A programming demonstration using the NVIDIA Visual Profiler will be included. Introduction to Optimizations and Profiling • Arithmetic optimizations • Warps • Branching efficiency • Memory latency/Occupancy • Memory performance optimizations • Programming Demo: Visual Profiler

S0027A - All-In-One Debugging Experience with CUDA-GDB and CUDA-MEMCHECK
Geoff Gerfin (NVIDIA), Vyas Venkataraman (NVIDIA)
Day: Monday, 05/14 | **Time:** 2:30 pm - 4:00 pm
Topic Areas: Development Tools & Libraries
Session Level: Advanced

CUDA Debugger tools CUDA-GDB and CUDA-MEMCHECK provide a whole new feature set to help improve your CUDA application development cycle. This session is a detailed walk-through of the key new features and advanced techniques on using CUDA-GDB and CUDA-MEMCHECK together to improve overall code productivity. This tutorial will also include live demos. This session will be repeated later during the conference.



SESSIONS

S0419A - Optimizing Application Performance with CUDA Profiling Tools
David Goodwin (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 9:00 am - 9:50 am
Topic Areas: Development Tools & Libraries
Session Level: Intermediate

NVIDIA provides two powerful profiling tools that you can use to maximize your application's performance. The NVIDIA Visual Profiler helps you understand your application's behavior with a detailed timeline and data from GPU performance counters. The Visual Profiler also provides an automatic, data-driven analysis engine that provides suggestions on potential optimization strategies for your application. Nvprof is a command-line profiler that provides gprof-like functionality for the GPU. Nvprof provides summary information about where your application is spending the most time, so that you can focus your optimization efforts. This session will provide a step-by-step walk through of both of these profiling tools, showing how you can use these tools to identify optimization opportunities at the application, kernel, and source-line levels.

S0258 - Sailfish: Lattice Boltzmann Fluid Simulations with GPUs and Python
Michal Januszewski (University of Silesia in Katowice; Google Switzerland)
Day: Tuesday, 05/15 | **Time:** 9:30 am - 9:55 am
Topic Areas: Computational Fluid Dynamics; Computational Physics; Development Tools & Libraries
Session Level: Intermediate

Learn how Run-Time Code Generation (RTCG) techniques allowed for fast development of a lattice Boltzmann (LB) fluid dynamics solver called Sailfish. Sailfish is completely open source, supports a wide variety of LB models (single and multiple relaxation times, the entropic model; single and binary fluids) and can take advantage of multiple GPUs. Even though the project is written predominantly in Python, no performance compromises are made. This talk will introduce the basic design principles of Sailfish and illustrate how RTCG allows to exploit the power of GPUs with minimal programmer effort.

S0528 - CUDA Debugger Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 9:30 am - 10:20 am
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a variety of powerful CUDA debugging feature set that enables developers to quickly spot bugs. From the memory checker to advanced breakpoints and variable warp watch panel, a developer can quickly isolate access memory errors, filter out the thousands of threads to a specific thread and quickly spot abnormal variable value ranges. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at developing CUDA code.

S0430 - Developing Next-Generation CUDA Acceleration in Wolfram's Mathematica with Nsight
Sebastien Domine (NVIDIA) , Ulises Cervantel-Pimentel (Wolfram) , Abdul Dakkak (Wolfram)
Day: Tuesday, 05/15 | **Time:** 9:30 am - 10:20 am
Topic Areas: Development Tools & Libraries
Session Level: Intermediate

Since version 8, Mathematica offers advanced support for GPU acceleration with optimized CUDA functions and a built-in framework for developing scientific CUDA kernel code. In this session, the Wolfram development team

will share their experience developing their next-generation CUDA support in Mathematica. From the unique ability of Parallel Nsight to attach its CUDA debugger to a running process, the new parallel Warp Watch for warp-wide variable views and expression evaluation, to the latest runtime CUDA profiling experiments; they will demonstrate how they were able to take advantage of Parallel Nsight to get the most out of CUDA and the GPU.

S0529 - CUDA Profiler Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a comprehensive set of performance analysis tools. From the ability to trace complete system multi-core CPU and multi GPU activities, to profile CUDA kernel with precise profiling experiments, developers can identify system level optimization opportunities as well as expensive and inefficient CUDA kernels requiring in-depth analysis with the CUDA profiler. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at optimizing complex CUDA applications.

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)

S0407 - A High Level Programming Environment for Accelerated Computing
Luiz DeRose (Cray Inc.)
Day: Tuesday, 05/15 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Development Tools & Libraries; Parallel Programming Languages & Compilers
Session Level: Intermediate

One of the critical hurdles for the widespread adoption of accelerated computing in HPC is programming difficulty. Users need a simple programming model that is portable and is not significantly different from the approaches used on current multi-core x86 processors. In this talk I will present Cray's strategy to accelerator programming, which is based on a high level programming environment with tightly coupled compilers, libraries, and tools. Ease of use is possible with compiler making it feasible for users to write applications in Fortran, C, C++, tools to help users port and optimize for accelerators, and auto-tuned scientific libraries.

S0049 - Using the GPU Direct for Video API
Thomas True (NVIDIA), Alina Alt (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Audio, Image and Video Processing; Development Tools & Libraries; Digital Content Creation & Film; Machine Vision
Session Level: Advanced

This tutorial will demonstrate how video I/O devices can take advantage of the GPU Direct for Video API to optimize the data transfer performance for digital video, film and broadcast applications and computer vision applications. The GPU Direct for Video API is a technology that permits the DMA transfer of data buffers between video I/O devices and the GPU through the use of a shared system memory buffer for immediate processing by OpenGL, DirectX, CUDA and OpenCL. This direct transfer can improve synchronization and eliminate latency between video capture, GPU processing and video output.

S0528 - CUDA Debugger Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 4:00 pm - 4:50 pm
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a variety of powerful CUDA debugging feature set that enables developers to quickly spot bugs. From the memory checker to advanced breakpoints and variable warp watch panel, a developer can quickly isolate access memory errors, filter out the thousands of threads to a specific thread and quickly spot abnormal variable value ranges. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at developing CUDA code.

S0062 - Histograms of Oriented Gradients with CUDA: Performance Analysis and Optimization Tips
Anton Obukhov (Consultant)
Day: Tuesday, 05/15 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Computer Vision; Machine Vision; Development Tools & Libraries
Session Level: Advanced

Computer Vision is becoming increasingly popular and important nowadays. With the advent of powerful mobile devices and increasing power of desktop PCs, it is important to improve user experience by tackling the hardest problems of real-time interaction with the user. These include body parts tracking, face, and gesture recognition. This talk discusses a well-known Histogram of Oriented Gradients approach to object detection in images and its implementation with CUDA. A detailed performance analysis of different algorithm parts is conducted and optimizations for various usage cases are proposed. The role of OpenCV GPU module is highlighted and implementation details are provided.

S0602 - An Introduction to the Thrust Parallel Algorithms Library
Nathan Bell (NVIDIA), Julien Demouth (NVIDIA)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:50 pm
Topic Areas: Parallel Programming Languages & Compilers; Development Tools & Libraries
Session Level: Beginner

Thrust is a parallel algorithms library which resembles the C++ Standard Template Library (STL). Thrust's high-level interface greatly enhances developer productivity while enabling performance portability between GPUs and multicore CPUs. Interoperability with established technologies (such as CUDA, TBB and OpenMP) facilitates integration with existing software. In this talk we'll walk through the library's main features and explain how developers can build high-performance applications rapidly with Thrust.

S0287 - Jacket for Multidimensional Scaling in Genomics
Chris McClanahan (AccelerEyes)
Day: Tuesday, 05/15 | **Time:** 5:30 pm - 5:55 pm
Topic Areas: Development Tools & Libraries
Session Level: Beginner

In this tutorial, we will present AccelerEyes' Jacket software which enables GPU computing in MATLAB through a user case study entitled "Multidimensional Scaling for Genomics". We show how Jacket enables developers to write and run code on the GPU in the native M-Language used in MATLAB. By simply casting data to Jacket's GPU data structure, MATLAB functions are transformed into GPU functions. Additionally, we will also include demos of running MATLAB code on the GPU for image and signal processing, life science, finance, and other applications. A Q/A session will enable audience members to ask specific questions about Jacket.

S0420 - NSight IDE for Linux and Mac
David Goodwin (NVIDIA) , Eugene Ostroukhov (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 9:00 am - 9:50 am
Topic Areas: Development Tools & Libraries
Session Level: Beginner

NSight IDE for Linux and Mac is an all-in-one development environment that lets you develop, debug and optimize CUDA code in an integrated UI environment. If you were waiting for an IDE on Linux and Mac then this session is for you. This session provides a detail usage walk-through of a fully CUDA aware source editor, build integration of the CUDA toolchain, graphical debugger for both CPU and GPU, and graphical profiler to enable performance optimization.

S0529 - CUDA Profiler Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 9:00 am - 9:50 am
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a comprehensive set of performance analysis tools. From the ability to trace complete system multi-core CPU and multi GPU activities, to profile CUDA kernel with precise profiling experiments, developers can identify system level optimization opportunities as well as expensive and inefficient CUDA kernels requiring in-depth analysis with the CUDA profiler. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at optimizing complex CUDA applications.

S0325 - ArrayFire Graphics: A Tutorial
Chris McClanahan (AccelerEyes)
Day: Wednesday, 05/16 | **Time:** 10:00 am - 10:25 am
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Learn how to use the graphics primitives for GPU computing available in ArrayFire, a new C and C++ library for GPU computing in both CUDA and OpenCL. In this session, we will cover the capabilities of ArrayFire's graphics primitives and show how to build fast, visual computing applications. The tutorial centers around the construction of an application for the computation of optical flow on the GPU and will illustrate how to couple graphics with compute using ArrayFire's graphics primitives. We will also show how the graphics primitives can be composed to result in scalable, fast graphics that complement GPU applications.

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.

S0528 - CUDA Debugger Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a variety of powerful CUDA debugging feature set that enables developers to quickly spot bugs. From the memory checker to advanced breakpoints and variable warp watch panel, a developer can quickly isolate access memory errors, filter out the thousands of threads to a specific thread and quickly spot abnormal variable value ranges. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at developing CUDA code.

S0419B - Optimizing Application Performance with CUDA Profiling Tools
David Goodwin (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Development Tools & Libraries
Session Level: Intermediate

NVIDIA provides two powerful profiling tools that you can use to maximize your application's performance. The NVIDIA Visual Profiler helps you understand your application's behavior with a detailed timeline and data from GPU performance counters. The Visual Profiler also provides an automatic, data-driven analysis engine that provides suggestions on potential optimization strategies for your application. Nvprof is a command-line profiler that provides gprof-like functionality for the GPU. Nvprof provides summary information about where your application is spending the most time, so that you can focus your optimization efforts. This session will provide a step-by-step walk through of both of these profiling tools, showing how you can use these tools to identify optimization opportunities at the application, kernel, and source-line levels.

S0027B - All-In-One Debugging Experience with CUDA-GDB and CUDA-MEMCHECK
Geoff Gerfin (NVIDIA), Vyas Venkataraman (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Development Tools & Libraries
Session Level: Advanced

CUDA Debugger tools CUDA-GDB and CUDA-MEMCHECK provide a whole new feature set to help improve your CUDA application development cycle. This session is a detailed walk-through of the key new features and advanced techniques on using CUDA-GDB and CUDA-MEMCHECK together to improve overall code productivity. This tutorial will also include live demos. This session will be repeated later during the conference.

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

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

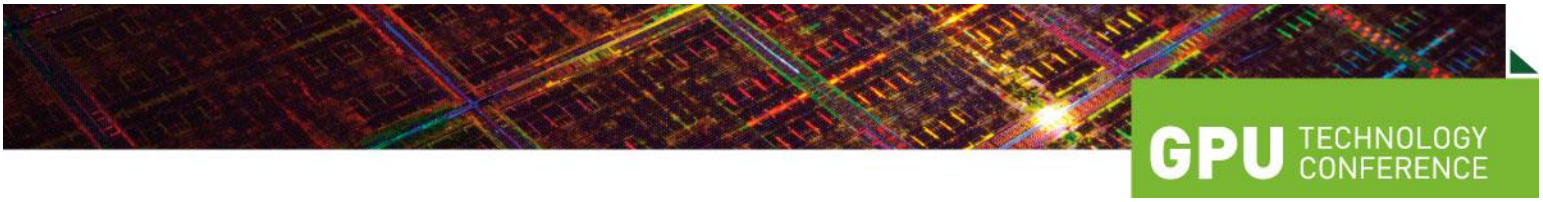
See the newest features integrated in MAGMA (Matrix Algebra on GPU and Multicore Architectures) to tackle the multiple GPU-based systems for numerical linear algebra. In this talk, we describe how we leveraged MAGMA to solve existing and new challenging numerical problems on multiple hardware accelerators. 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.

S0099 - Debugging GPU Applications For Correctness and Performance
David Lecomber (Allinea Software)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:25 pm
Topic Areas: Development Tools & Libraries
Session Level: Intermediate

This session reveals how debugging CUDA applications is made straightforward with the powerful Allinea DDT debugger. New features enabling greater understanding of performance optimizations will be explored, showing how they can be used to produce better, faster CUDA code. Coupled with newly released support for multiple languages and compilers we will also show how Allinea DDT is enabling developers on desktops and the largest supercomputers to achieve both correct and efficient GPU applications.

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.



S0529 - CUDA Profiler Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 4:00 pm - 4:50 pm
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a comprehensive set of performance analysis tools. From the ability to trace complete system multi-core CPU and multi GPU activities, to profile CUDA kernel with precise profiling experiments, developers can identify system level optimization opportunities as well as expensive and inefficient CUDA kernels requiring in-depth analysis with the CUDA profiler. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at optimizing complex CUDA applications.

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.

S0121 - Software Architecture to Facilitate CUDA Development
Peter Shenkin (Schrodinger), K. Patrick Lorton (Schrodinger)
Day: Wednesday, 05/16 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Development Tools & Libraries; Life Sciences
Session Level: Intermediate

We describe workflow architecture and its use in developing Schrödinger's core-hopping application. The application supplies the stages as callbacks. A stage may have multiple implementations; for example, CUDA and CPU. An implementation can be assigned a maximum number of simultaneous threads. When any stage completes, a scheduling algorithm determines which implementation of which stage will be launched next. The application may detect "special" environments, such as CUDA, and set up its stages accordingly, or it may allow specification of which implementation of each stage to run. This makes it easy to develop and debug CUDA stages flexibly and incrementally.

S0257 - Trace Based Performance Analysis for GPU Accelerated Multi-Hybrid Applications
Guido Juckeland (TU Dresden - ZIH)
Day: Wednesday, 05/16 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Development Tools & Libraries
Session Level: Intermediate

Get in contact with performance tuning experts for multi-hybrid applications and see first hand how VampirTrace/Vampir can significantly speed up application porting and development.

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.

S0100 - Mathematica as a Practical Platform for GPU-Accelerated Finance
Dylan Roeh (Wolfram Research Inc.), Abdul Dakkak (Wolfram Research Inc.)
Day: Wednesday, 05/16 | **Time:** 5:00 pm - 5:25 pm
Topic Areas: Finance; Development Tools & Libraries
Session Level: Intermediate

With the introduction of GPU support in version 8, Mathematica has become an excellent environment for integrating CUDA with high level code for interpretation or visualization. In this presentation, we will show the usefulness of Mathematica in the venue of computational finance. In addition to demonstrating the GPU-accelerated financial computations which can be readily performed within Mathematica, we will show that these calculations can easily be integrated with third-party data sources including Microsoft Excel and databases. Furthermore, we will cover the UnRisk Mathematica package written by MathConsult, which seamlessly adds GPU-accelerated complex model calibration algorithms to Mathematica's repertoire.

S0242 - Harnessing GPU Compute with C++ AMP (Part 1 of 2)
Daniel Moth (Microsoft)
Day: Wednesday, 05/16 | **Time:** 5:00 pm - 5:50 pm
Topic Areas: Parallel Programming Languages & Compilers; Development Tools & Libraries
Session Level: Intermediate

C++ AMP is an open specification for taking advantage of accelerators like the GPU. In this session we will explore the C++ AMP implementation in Microsoft Visual Studio 11. After a quick overview of the technology understanding its goals and its differentiation compared with other approaches, we will dive into the programming model and its modern C++ API. This is a code heavy, interactive, two-part session, where every part of the library will be explained. Demos will include showing off the richest parallel and GPU debugging story on the market, in the upcoming Visual Studio release.

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

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

S0605 - cudaDMA: Emulating DMA engines on GPUs for Performance and Programmability
Brucek Khailany (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 5:00 pm - 5:25 pm
Topic Areas: Development Tools & Libraries
Session Level: Intermediate

The CudaDMA library is a collection of DMA objects that support efficient movement of data between off-chip global memory and on-chip shared memory in CUDA kernels. CudaDMA objects support many different data transfer patterns including sequential, strided, gather, scatter, and halo patterns. The library encapsulates efficient synchronization and data transfer implementations to achieve high memory bandwidth utilization. Programmer productivity is achieved by avoiding the need for thread array shapes to match data layout. Using CudaDMA, speedups of up to 1.37x on synthetic micro-benchmarks and 1.15x-3.2x on kernels from scientific applications have been demonstrated.

S0280 - MATLAB and GPU: American Exercise Options Monte Carlo 3 Ways
John Ashley (NVIDIA)
Day: Wednesday, 05/16 | **Time:** 5:30 pm - 5:55 pm
Topic Areas: Finance; Development Tools & Libraries
Session Level: Beginner

The same Longstaff-Schwartz Monte Carlo implementation done 3 ways -- using GPUArray, arrayFun, and custom kernels -- to illustrate the trade-offs between ease of use and performance. Side by side examination of segments of code and their relative performance will help build intuition about the different ways to leverage GPU acceleration of projects using Matlab.

S0133 - Improving Mars Rover Image Compression Via GPUs And Genetic Algorithms
Brendan Babb (University of Alaska Anchorage), Frank Moore (University of Alaska, Anchorage)
Day: Thursday, 05/17 | **Time:** 9:00 am - 9:25 am
Topic Areas: Machine Learning & AI; Audio, Image and Video Processing; Development Tools & Libraries
Session Level: Beginner

Learn how to use Jacket to accelerate genetic algorithm (GA) image compression. Our research uses a GA to optimize lossy compression transforms that outperform state-of-the-art wavelet-based approaches for a variety of image classes, including fingerprints, satellite, medical, and images transmitted from the Mars Exploration Rovers. A typical training run evolves a population of transforms over many generations; since each transform must be applied to each image from the training set, each run entails thousands of independent, parallelizable fitness evaluations. By using MATLAB, and Jacket to perform 2D convolution on the GPU, we have greatly reduced the total computation time needed.

S0528 - CUDA Debugger Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Thursday, 05/17 | **Time:** 9:00 am - 9:50 am
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a variety of powerful CUDA debugging feature set that enables developers to quickly spot bugs. From the memory checker to advanced breakpoints and variable warp watch panel, a developer can quickly isolate access memory errors, filter out the thousands of threads to a specific thread and quickly spot abnormal variable value ranges. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at developing CUDA code.

S0333 - GMAC-2: Easy and Efficient Programming for CUDA-Based Systems
Javier Cabezas (Barcelona Supercomputing Center), Isaac Gelado (University of Illinois at Urbana Champaign)
Day: Thursday, 05/17 | **Time:** 9:00 am - 9:50 am
Topic Areas: Development Tools & Libraries
Session Level: Intermediate

In this talk we introduce GMAC-2, a framework that eases the development of CUDA applications and tools while achieving similar or better performance than hand-tuned code. The new features implemented in GMAC-2 allow programmers to further fine-tune their code and remove some limitations found in the original GMAC library. For example, memory objects can be now arbitrarily mapped on several devices without restrictions and a host thread can launch kernels on any GPU in the system. Moreover, GMAC-2 transparently takes advantage of the new features offered by the hardware like the GPUDirect 2 peer-to-peer communication.

S0256 - A Stencil Library for the New Dynamic Core of COSMO
Tim Schroeder (NVIDIA), Tobias Gysi (SCS)
Day: Thursday, 05/17 | **Time:** 9:00 am - 9:50 am
Topic Areas: Climate & Weather Modeling; Development Tools & Libraries
Session Level: Advanced

We will present a stencil library used in the heart of the COSMO numeric weather prediction model. During the talk we'll show how we implemented an abstraction that allows easy development of new stencils and solvers on top of a framework allowing execution on both CPU and GPU. The library makes efficient use of GPU resources and we will show how to structure memory accesses and computation optimally. Developers involved in porting or writing fully-featured C++ libraries for CUDA will also be interested in attending.

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

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

near real time. Technical challenges and recent progress will be discussed.

S0244 - Harnessing GPU Compute with C++ AMP (Part 2 of 2)
Daniel Moth (Microsoft)
Day: Thursday, 05/17 | **Time:** 10:00 am - 10:50 am
Topic Areas: Parallel Programming Languages & Compilers; Development Tools & Libraries
Session Level: Intermediate

C++ AMP is an open specification for taking advantage of accelerators like the GPU. In this session we will explore the C++ AMP implementation in Microsoft Visual Studio 11. After a quick overview of the technology understanding its goals and its differentiation compared with other approaches, we will dive into the programming model and its modern C++ API. This is a code heavy, interactive, two-part session, where every part of the library will be explained. Demos will include showing off the richest parallel and GPU debugging story on the market, in the upcoming Visual Studio release.

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

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

S0078 - Panoptes: A Binary Instrumentation Framework for CUDA
Christopher Kennelly (D. E. Shaw Research)
Day: Thursday, 05/17 | **Time:** 10:00 am - 10:50 am
Topic Areas: Development Tools & Libraries
Session Level: Advanced

Traditional CPU-based computing environments offer a variety of binary instrumentation frameworks, while the instrumentation and analysis tools available to date for GPU environments have been more limited. Here we present Panoptes, a binary instrumentation framework for CUDA that targets the GPU. By exploiting the GPU to run modified kernels, Panoptes allows computationally intensive programs to be run at the native parallelism of the device during analysis. To demonstrate the instrumentation capabilities of Panoptes, we will present our work on a memory addressability and validity checker that targets CUDA programs.

S0054 - PFAC Library: GPU-Based String Matching Algorithm
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/).

S0529 - CUDA Profiler Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Thursday, 05/17 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a comprehensive set of performance analysis tools. From the ability to trace complete system multi-core CPU and multi GPU activities, to profile CUDA kernel with precise profiling experiments, developers can identify system level optimization opportunities as well as expensive and inefficient CUDA kernels requiring in-depth analysis with the CUDA profiler. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at optimizing complex CUDA applications.

S0320 - PTask: OS Support for GPU Dataflow Programming
Christopher Rossbach (Microsoft Research Silicon Valley), Jon Currey (Microsoft Research Silicon Valley)
Day: Thursday, 05/17 | **Time:** 2:00 pm - 2:50 pm
Topic Areas: Development Tools & Libraries; General Interest; Parallel Programming Languages & Compilers
Session Level: Advanced

This session considers the PTask API, OS-level abstractions that support GPUs as first-class computing resources, and supports a dataflow programming model. With PTask, the programmer specifies where data goes, rather than how and when it should get there, allowing the system to provide fairness and isolation guarantees, streamline data movement in ways that currently require direct programmer involvement, and enable code portability across diverse GPU-based platforms. Our experience building the PTask APIs shows that PTask can provide important system-wide guarantees and can enable significant performance benefits, for example improving the throughput of hand-tuned CUDA programs by up to 2x.

S0032 - Teraflop GPU Acceleration of Large Matrix Algebra
Ronald Young (Multipath Corporation)
Day: Thursday, 05/17 | **Time:** 2:30 pm - 2:55 pm
Topic Areas: Development Tools & Libraries; General Interest
Session Level: Beginner

Learn how Multipath's Fast Matrix Solver (FMS) is setting performance records using multiple GPU's solving large matrices in production applications. By (1) leveraging NVIDIA's CUBLAS library, (2) operating multiple GPU's in parallel and (3) overlapping data transfers with computation, FMS averages over 2 teraflops of performance, even on jobs lasting for days. The presentation also includes a description of what problems FMS solves and how it is incorporated into applications programs.

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.

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.

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

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

S0428 - Panini: A GPU Aware Array Class
Priyanka Sah (NVIDIA), Santosh Ansumali (JNCASR, Bangalore)
Day: Thursday, 05/17 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Development Tools & Libraries
Session Level: Beginner

We present a new templated C++ class library, PANINI, for use in the development of large-scale scientific simulations in an hetrogeneous computing environment. The key feature of this new library is a generic parallel array class built on advanced generic programming methodologies where details of parallelization is hidden inside the array class itself. This library will be used for Poison Solver, Advection Diffusion and other equation.

S0528 - CUDA Debugger Training on Windows
NVIDIA Developer Tools Team (NVIDIA)
Day: Thursday, 05/17 | **Time:** 4:00 pm - 4:50 pm
Topic Areas: Development Tools & Libraries
Session Level: Beginner

Nsight offers a variety of powerful CUDA debugging feature set that enables developers to quickly spot bugs. From the memory checker to advanced breakpoints and variable warp watch panel, a developer can quickly isolate access memory errors, filter out the thousands of threads to a specific thread and quickly spot abnormal variable value ranges. Through a set of comprehensive exercises, the attendee will be able to utilize these features to become fully proficient at developing CUDA code.

S0218 - ASI Parallel Fortran: A General-Purpose Fortran to GPU Translator
Rainald Lohner (George Mason University)
Day: Thursday, 05/17 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Development Tools & Libraries; Computational Fluid Dynamics; Computational Physics; Parallel Programming Languages & Compilers
Session Level: Advanced

Over the last 3 years we have developed a general-purpose Fortran to GPU translator: ASI Parallel Fortran does. The talk will detail its purpose, design layout and capabilities, and show how it is used and implemented. The use of ASI Parallel Fortran will be shown for large-scale CFD/CEM codes as well as other general purpose Fortran codes.

S0074 - Techniques for Designing GPGPU Games
Mark E. S. Joselli (UFF), Esteban Clua (Universidade Federal Fluminense)
Day: Thursday, 05/17 | **Time:** 5:00 pm - 5:25 pm
Topic Areas: Development Tools & Libraries
Session Level: Intermediate

Learn how to develop faster and better games with the use of GPGPU thought the use of Game GPU tricks. Normally, games process most of its tasks in the CPU, using the GPU only for graphics processing. This session shows some techniques on how to better use the GPGPU power to process all the game logic, achieving speedups when compared to CPU, and traditional GPU models. This session also shows some examples of this technique in practice.