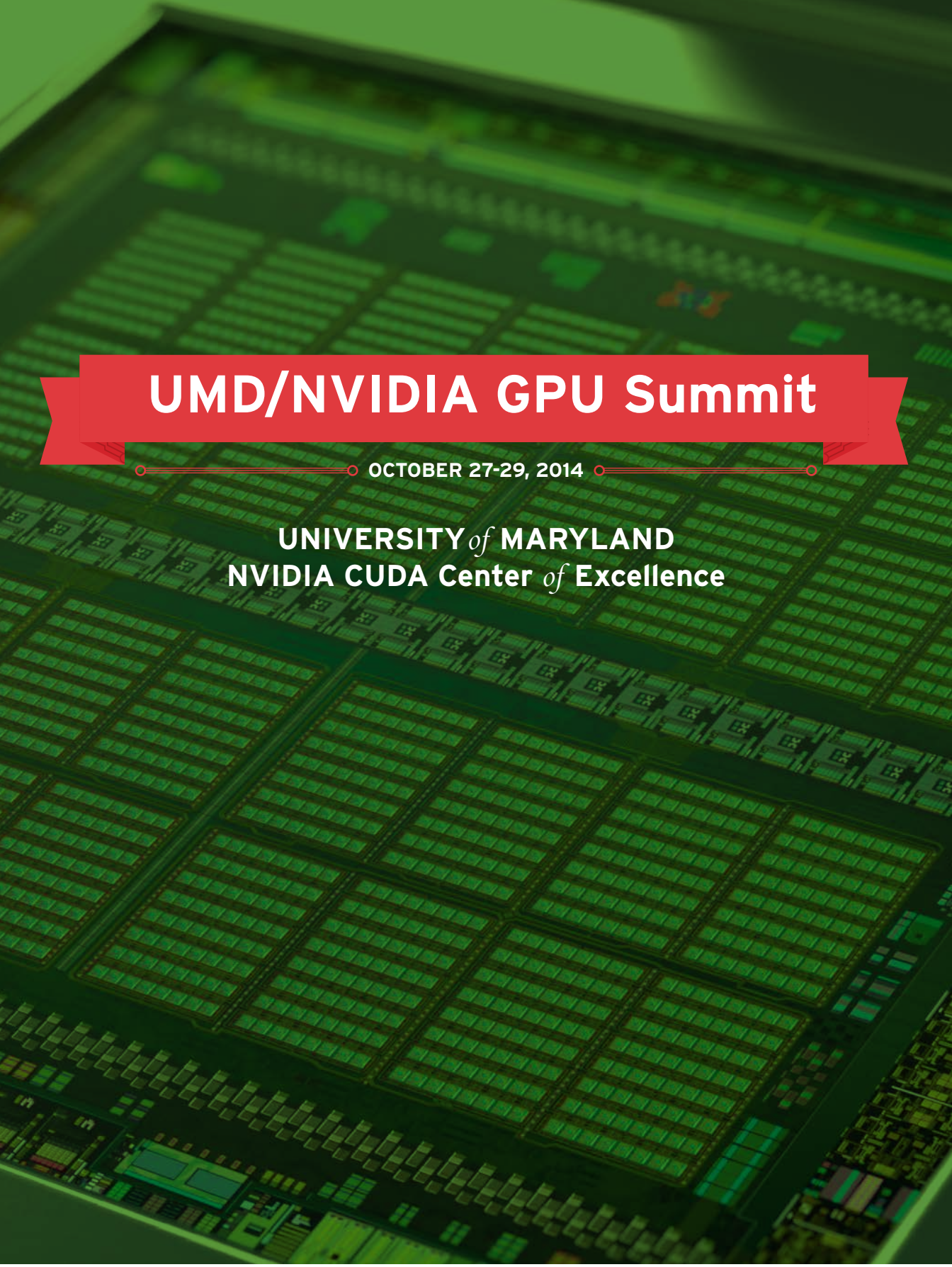

UMD/NVIDIA GPU Summit

OCTOBER 27-29, 2014

UNIVERSITY *of* MARYLAND
NVIDIA CUDA Center *of* Excellence

UNIVERSITY OF MARYLAND | NVIDIA

ADELE H. STAMP STUDENT UNION
UNIVERSITY OF MARYLAND
COLLEGE PARK, MD 20742

UNIVERSITY OF MARYLAND
MPOWERING THE STATE


McKELDIN LIBRARY
htc
McKELDIN LIBRARY

WELCOME TO THE GPU SUMMIT AT THE UNIVERSITY OF MARYLAND!

Dramatic increases in computing performance, enabled by harnessing the power of the graphics processing unit (GPU), are changing the way we live, work, educate, train and assimilate data. With the evolution of GPUs into powerful, programmable and inexpensive parallel processors, software developers, scientists and researchers are deploying these units in an ever-expanding array of uses.

In this three-day program, we will highlight this evolution, and illustrate its most promising applications. Our speakers will detail their visionary, cutting-edge research, ranging from nano-robot manufacturing to mapping brain networks, from modeling the evolutionary relationships of life on Earth to developing better impact-resistant materials for aviation, from understanding how proteins interact at microscopic scales to designing new interfaces for exploring virtual and augmented worlds. We'll also introduce the University of Maryland's own Deepthought2 high-performance computing cluster: with 200,000 GPUs, it is one of the top 10 clusters among U.S. universities. Our panels will feature discussions on the GPU state-of-the-art in both high-throughput computing and medicine. We're particularly pleased about networking opportunities that will be available across a wide spectrum of colleagues spanning academia, industry and federal institutions.

We're also excited to offer an opportunity to train with NVIDIA CUDA GPU programming experts. Training sessions will include an introduction to GPU computing, basics of CUDA programming, and fundamental performance optimizations. These sessions will provide an excellent portal to harnessing the performance of GPUs in your own work.

We hope this is an exciting and productive event for all!

Sincerely,

AMITABH VARSHNEY
Director, University of Maryland Institute for Advanced Computer Studies
Director, UMD-NVIDIA CUDA Center of Excellence

MONDAY, OCT. 27
STAMP STUDENT UNION – ATRIUM

8 - 8:30 A.M. Registration, Breakfast

8:30 - 10 A.M. KEYNOTE SESSION

8:30 A.M. **Fran LoPresti**, Welcome Address: "GPUs at Maryland"

8:40 A.M. **David Luebke**, "Computational Displays: How GPU Horsepower and Novel Optics Enable Thin, Light, Wide-Angle Virtual and Augmented Reality"

9 A.M. **Amitabh Varshney**, "Augmented Reality Made More Real"

9:20 A.M. **GPU Panel for Medicine:** Moderator, **Terry Yoo**; Panelists: **Peter Bajcsy**, **Raj Shekhar** and **Oleg Kuybeda**

10 - 10:30 A.M. Coffee Break

10:30 A.M. - 12 P.M. RESEARCH OVERVIEW SESSION I

10:30 A.M. **Alex Szalay**, "Data-Intensive Science Using GPUs"

10:50 A.M. **Michael Cummings**, "GPU Computing and the Tree of Life"

11:10 A.M. **Jeff Hollingsworth**, "Automatically Tuning Performance and Power for GPUs"

11:30 A.M. **Satyandra K. Gupta**, "GPU-Enabled Computing in Robotics and Advanced Manufacturing Applications"

11:50 A.M. **Joseph JaJa**, "Mapping Biomedical Applications onto GPU Platforms"

12:10 - 1:30 P.M. Lunch

1:30 - 3 P.M. RESEARCH OVERVIEW SESSION II

1:30 P.M. **Norman Wereley**, "Particle Simulations in Magnetorheological Flows"

1:50 P.M. **Jeffery Klauda**, "Molecular Modeling of Biomolecules: How Can GPUs Advance Research?"

2:10 P.M. **Shuvra Bhattacharyya**, "Vectorization and Mapping of Software Defined Radio Applications on GPU Platforms"

2:30 P.M. **Lorena Barba**, "PyGBe for Probing Protein Orientation Near Charged Surfaces"

3 - 3:30 P.M. Coffee Break

3:30 - 5 P.M. RESEARCH OVERVIEW SESSION III

3:30 P.M. **Ramani Duraiswami**, "Fast Heterogeneous Computing"

3:50 P.M. **Laura Monroe**, "Probabilistic Computing on the GPU"

4:10 P.M. **GPU Panel for High-Throughput Computing:** Moderator, **Jimmy Lin**; Panelists, **Raju Namburu**, **George Stantchev** and **R. Jacob Vogelstein**

TUESDAY, OCT. 28
STAMP STUDENT UNION – PRINCE GEORGE'S ROOM

8 A.M. - 5:30 P.M. TUTORIALS

8 A.M. **Intro to GPU Computing** [1 hr]
High-level discussion of GPU computing

9:15 A.M. **Programming with OpenACC** [2 hrs]
Using simple directives to accelerate code

11:30 A.M. **Basics of CUDA Programming** [Part 1, 1 hr]
CUDA syntax, memory allocation, launching simple kernels

12:30 - 1:30 P.M. Lunch

1:30 P.M. **Basics of CUDA Programming** [Part 2, 1 hr]
CUDA syntax, memory allocation, launching simple kernels

2:30 P.M. **Fundamental GPU Performance Optimizations** [3 hrs]
Using the performance profiler; global and shared memory optimizations

WEDNESDAY, OCT. 29
STAMP STUDENT UNION – PRINCE GEORGE'S ROOM

9 A.M. - 12 P.M. TUTORIAL

9 A.M. **Intermediate CUDA Optimizations** [3 hrs]
Overlapping communication with computation; streams and concurrency

KEYNOTE SESSION

Fran LoPresti
Deputy CIO for Cyberinfrastructure and Research IT, Division of Information Technology, University of Maryland
MONDAY, 8:30 A.M.
WELCOME ADDRESS: "GPUS AT UMD"

Deepthought2, the University of Maryland central cluster, was launched in July 2014. Deepthought2 will be utilized by workshop participants. This talk will provide an introduction to Deepthought2 and other research support activities by the Division of Information Technology.

David Luebke
Senior Director of Research, NVIDIA
MONDAY, 8:40 A.M.
"COMPUTATIONAL DISPLAYS: HOW GPU HORSEPOWER AND NOVEL OPTICS ENABLE THIN, LIGHT, WIDE-ANGLE VIRTUAL AND AUGMENTED REALITY"

In this talk, I will discuss the broad and exciting subject of GPU-powered computational displays—a new approach that combines the computational power of the GPU with novel optics to produce new displays suitable for augmented and virtual reality.

Amitabh Varshney
Director, University of Maryland Institute for Advanced Computer Studies (UMIACS)
MONDAY, 9 A.M.
"AUGMENTED REALITY MADE MORE REAL"

Augmented reality is the next logical leap forward in the ever-expanding information revolution. By overlaying, or augmenting, digital information on top of real-world settings, augmented reality allows people from all walks of life—physicians, educators, industrial workers, artists and everyday citizens—to see and to use the information that matters most to them. In this talk, I will discuss some of the computational challenges in the augmented reality pipeline that GPUs are well-positioned to address.

GPU PANEL FOR MEDICINE
MONDAY, 9:20 A.M.

Moderator: Terry Yoo
Office of High Performance Computing and Communications, National Library of Medicine, National Institutes of Health

This panel discussion will include **Peter Bajcsy** (computer scientist, Information Technology Laboratory, NIST), **Raj Shekhar** (founder, IGI Technologies), and **Oleg Kuybeda** (image processing scientist, Office of High Performance Computing and Communications, National Library of Medicine, National Institutes of Health). The discussion will center on how GPUs are enabling and improving a host of medical applications, including image-guided interventions and surgical planning, cancer detection and diagnosis, single-particle microscopy, and cardiac stress-testing.

RESEARCH OVERVIEW I

Alex Szalay
Alumni Centennial Professor, Department of Physics and Astronomy, Johns Hopkins University
MONDAY, 10:30 A.M.
"DATA-INTENSIVE SCIENCE USING GPUS"

In this talk, I will present an overview on how data is changing science. This also has profound implications for the computational architectures used to perform computations and analyses. We will discuss several challenging science cases, and show how a combination of high-speed I/O and GPUs can achieve remarkable results.

Michael Cummings
Associate Professor, Department of Biology and UMIACS, University of Maryland
MONDAY, 10:50 A.M.
"GPU COMPUTING AND THE TREE OF LIFE"

Phylogenetic inference is fundamental to our understanding of the tree of life — the evolutionary relationships of life on earth. There is a recent concentration of interest in statistical approaches, such as maximum likelihood estimation and Bayesian inference. Yet, for large data sets and realistic or interesting models of evolution, these approaches remain computationally demanding. High-throughput sequencing can yield data for thousands of taxa, but scaling to such problems using serial computing often necessitates the use of non-statistical or approximate approaches. The emergence of GPUs provides an opportunity to leverage their excellent floating-point computational performance to accelerate statistical phylogenetic inference. A specialized library for phylogenetic calculation would allow existing software packages to make more effective use of available computer hardware, including GPUs. We have developed BEAGLE, an application programming interface (API) and library for high performance statistical phylogenetic inference, which provides a uniform interface for performing phylogenetic likelihood calculations on a variety of computer hardware platforms.

Jeff Hollingsworth
Professor, Department of Computer Science and UMIACS, University of Maryland
MONDAY, 11:10 A.M.
"AUTOMATICALLY TUNING PERFORMANCE AND POWER FOR GPUS"

Getting the most performance at the lowest power consumption is a challenging, and often time-consuming, problem. Historically, such projects required extensive manual effort and needed to be repeated each time new hardware was released. In this talk, I will describe our work on the Active Harmony system and its NEMO algorithm, which provides a way to automatically tune complex multi-objective criteria, such as obtaining the best performance possible subject to a power constraint. In addition to describing our approach, I will also present results from utilizing the tool on benchmark codes running actual GPUs.

Satyandra K. Gupta
Professor, Department of Mechanical Engineering and Institute for Systems Research, University of Maryland
MONDAY, 11:30 A.M.
"GPU-ENABLED COMPUTING IN ROBOTICS AND ADVANCED MANUFACTURING APPLICATIONS"

Robotics and Advanced Manufacturing applications utilize extensive geometric and physical simulations. These simulations are needed to enable automated planning and optimization. High simulation fidelity is very important in these applications. High simulation speed is needed to solve planning and optimization problems in a reasonable amount of time. GPUs can be used to speed up computations needed to enable high-fidelity, high-speed simulations, and, hence, significantly improve the performance of the automated planning and optimization. This presentation will describe how GPU-enabled computing is being used in planning for autonomous boats, automated mold design, and automated optical micromanipulation.

Joseph JaJa
Professor, Department of Electrical and Computer Engineering and UMIACS, University of Maryland
MONDAY, 11:50 A.M.
"MAPPING BIOMEDICAL APPLICATIONS ONTO GPU PLATFORMS"

GPUs are currently playing a major role in driving advances in high performance computing due to their advantages in performance/cost ratio, energy consumption, and programmability. Our work has been aimed at developing optimization techniques for mapping algorithms and applications onto GPUs and heterogeneous CPU-GPU platforms. During the past year, we have focused on biomedical applications, including agent based modeling of vocal fold inflammation and wound healing, development of connectivity-based brain parcellations using diffusion tensor imaging, and construction and analysis of brain networks using resting state fMRI. In this presentation, we will give an overview of our recent work focusing on some of the optimization methods used for core scientific computations.

RESEARCH OVERVIEW II

Norman Wereley
Department Chair, Minta Martin Professor of Aerospace Engineering, University of Maryland
MONDAY, 1:30 P.M.
"PARTICLE SIMULATIONS IN MAGNETORHEOLOGICAL FLOWS"

Recent work in the development of crashworthy crew seats in helicopters, blast resistant crew seats for ground vehicles, and other shock and impact problems has focused on the use of magnetorheological fluids (MRFs) in novel adaptive shock absorber designs. MRFs change their viscosity in response to an applied field, when the magnetic particles align their dipoles parallel to the field lines and form chain structures. By varying the magnetic field, the shock absorber can adjust its stroking load without using an active valve system, so it is more reliable in impact and shock mitigation systems. A key problem is the formation of microstructures in the valve of the shock absorber. This talk will show how the use of a GPU code allows for simulation of magnetorheological flows at device scales.

Jeffery Klauda
Associate Professor, Department of Chemical and Biomolecular Engineering, University of Maryland
MONDAY, 1:50 P.M.
"MOLECULAR MODELING OF BIOMOLECULES: HOW CAN GPUS ADVANCE RESEARCH?"

Advances in computational resources have allowed researchers focusing on molecular-scale simulations to begin to fully probe timescales of lipid membrane phase changes, protein folding and ligand binding. In this talk, I will discuss how GPUs have allowed for improved molecular dynamics (MD) simulations on proteins and membranes. The speedup on even a single node allows us to simulate parameter space in an efficient manner. Sample scaling will be shown with the use of the NAMD simulation package. Some sample applications will also be presented on bilayer phase changes and protein-ligand interaction.

Shuvra Bhattacharyya
Professor, Department of Electrical and Computer Engineering and UMIACS, University of Maryland
MONDAY, 2:10 P.M.
"VECTORIZATION AND MAPPING OF SOFTWARE DEFINED RADIO APPLICATIONS ON GPU PLATFORMS"

As the variety of off-the-shelf processors expands, traditional implementation methods of systems for digital signal processing and communication are no longer adequate to achieve design objectives in a timely manner. It is necessary for designers to easily track the changes in computing platforms, and apply them efficiently while reusing legacy code and optimized libraries that target specialized features in single processing units. In this context, we propose a new system design workflow to schedule and implement Software Defined Radio (SDR) applications that are developed using the GNU Radio environment, and targeted to GPU platforms. We present a design flow that extends the popular GNU radio environment, lays the foundation for rigorous analysis based on formal dataflow models, and provides a stand-alone library of GPU-accelerated actors that can be integrated efficiently into existing applications.

Lorena Barba
Associate Professor, School of Engineering and Applied Science, George Washington University
MONDAY, 2:30 P.M.
"PYGBE FOR PROBING PROTEIN ORIENTATION NEAR CHARGED SURFACES"

PyGBe is a code that uses Python, GPUs and boundary elements to solve problems in protein electrostatics. We released it last year, showing how it compares with a well-known finite-difference code to compute protein solvation energies. This is a quantity used by biologists in various situations governed by protein electrostatics. This year, we've worked on an extension of PyGBe to study the preferred orientation of proteins near charged surfaces. The target application is computational modeling of biosensors.

RESEARCH OVERVIEW III

Ramani Duraiswami
Professor, Department of Computer Science and UMIACS, University of Maryland
MONDAY, 3:30 P.M.
"FAST HETEROGENEOUS COMPUTING"

We detail two applications of GPU-based heterogeneous computing developed in my group. The first allowed the development of a real-time computational acoustical imaging device, since spun out as a company, VisiSonics. This device combines a spherical microphone array and a co-located array of HD video cameras, and works with algorithms implemented on a GPU-enabled laptop computer. The GPU performs acoustical beamforming and video image-stitching, while the CPUs provide control. The second uses heterogeneous architectures for speedup of Fast Multipole Methods (FMM). The FMM is an approximation algorithm allowing the fast computation to specified accuracy of dense matrix vector products that arise in fluid mechanics, acoustical and EM wave scattering, molecular dynamics, statistics, and other fields. This algorithm has particular promise on distributed parallel architectures, as it has good communication complexity, which is important on parallel architectures.

Laura Monroe
Project Leader, ASC Production Visualization Project, Los Alamos National Laboratory
MONDAY, 3:50 P.M.
"PROBABILISTIC COMPUTING ON THE GPU"

As Moore's Law fades and new computing architectures, such as the GPU, come into wide use, we will correspondingly need to develop new ways to exploit them. This includes not only meeting programming challenges, but also addressing power and resiliency issues. Probabilistic methods underlie many new and emerging paradigms, and may be useful in addressing challenges that arise. In this talk, we discuss aspects of the GPU that align well with probabilistic paradigms, and present a case study illustrating probabilistic computing on a GPU.

GPU PANEL FOR HIGH-THROUGHPUT COMPUTING
MONDAY, 4:10 P.M.

Moderator: Jimmy Lin
Associate Professor, College of Information Studies and UMIACS, University of Maryland

This panel will include **Raju Namburu** (Computational Sciences Division, Army Research Laboratory), **George Stantchev** (Naval Research Laboratory), and **R. Jacob Vogelstein** (ODNI/IARPA). GPU characteristics, like numerous simple yet energy-efficient computational cores, thousands of simultaneously active fine-grained threads, and large off-chip memory bandwidth, have motivated their deployment into a range of high-performance computing systems. This discussion will center on applications of these units in high-throughput computing.

TUTORIALS

TUESDAY AND WEDNESDAY

INTRO TO GPU COMPUTING

High-level overview of GPU hardware and software with an emphasis on how to use GPUs via applications, libraries and programming languages. [1 hr]

PROGRAMMING WITH OPENACC

Overview of the OpenACC programming model with some introductory hands-on exercises. [2 hrs]

BASICS OF CUDA PROGRAMMING

Introduction to programming in CUDA C/C++ with hands-on exercises. [2hrs]

FUNDAMENTAL GPU PERFORMANCE OPTIMIZATIONS

Optimization techniques for global memory and shared memory access, including using profiling tools to identify program hot spots and how to optimize them. [3 hrs]

INTERMEDIATE CUDA OPTIMIZATIONS

More use of profiling tools and using asynchronous communication to further optimize the entire compute workflow. [3 hrs]

GPU SUMMIT ORGANIZING COMMITTEE

Summit Co-Chairs
Kimberly Powell (NVIDIA)
Barbara Brawn-Cinani (UMD)

Summit Marketing Chair
Pauline Essalou (NVIDIA)

Summit IT Chair
Fritz McCall (UMD)

Summit Communications
Melissa Brachfeld (UMD)

WE GRATEFULLY ACKNOWLEDGE FUNDING PROVIDED BY THE UNIVERSITY OF MARYLAND/ MPOWERING THE STATE THROUGH THE CENTER FOR HEALTH-RELATED INFORMATICS AND BIOIMAGING.

NVIDIA CUDA CENTER OF EXCELLENCE
UNIVERSITY OF MARYLAND
2119 A. V. WILLIAMS BUILDING
COLLEGE PARK, MD 20742