DOING MORE WITH LESS: DEMOCRATIZING COMPUTING ACROSS SCIENTIFIC DOMAINS

Fernanda Foertter, GPU Developer Advocate Healthcare HPC + AI.
PLAY PC GAMES ON YOUR MAC

GeForce NOW™ brings your favorite PC games to your Mac at max settings, with smooth framerates—all powered by GeForce® GTX GPUs in the cloud. Get Game Ready drivers, cloud-sync saved games, and use express install to load games once, in seconds. Over 100 games are supported. Join the free beta today.

GEFORCE RTX™

GeForce RTX delivers the ultimate PC gaming experience. Powered by the new NVIDIA Turing™ GPU architecture.

THIN, FOR THE WIN.

Max-Q is an innovative approach to crafting the world’s thinnest, fastest, quietest gaming laptops.
USHERING A NEW-ISH ERA

- Phoenix X1
  - Dual core upgrade
  - Doubled size
  - X1e

- Jaguar XT3
  - Dual core upgrade

- Jaguar XT4
  - Quad core upgrade

- Jaguar XT5
  - 6 core upgrade

- 2004
- 2005
- 2007
- 2008
AI IS SPEEDING THE PATH TO FUSION ENERGY

Fusion is the future of energy on Earth. But it’s a highly sensitive process where even small environmental disruptions can stall reactions and damage multi-billion machines. Current models can predict the disruptions with 85% accuracy, but ITER will need something more precise.

Researchers at Princeton University have developed the Fusion Recurrent Neural Network (FRNN) using deep learning and NVIDIA GPUs with CUDA to predict disruptions and make adjustments to minimize damage and downtime. Even a 1% improvement in the prediction accuracy can be transformative considering the immense scale and cost of fusion science. FRNN has achieved 90% accuracy and is on the path to achieving its goal of 95% accuracy for ITER’s tests.

Visualization courtesy of Jamison Daniel, Oak Ridge Leadership Computing Facility
GPU ACCELERATED LIBRARIES
“Drop-in” Acceleration for Your Applications

DEEP LEARNING

- cuDNN
- TensorRT
- DeepStream SDK

SIGNAL, IMAGE & VIDEO

- cuFFT
- NVIDIA NPP
- CODEC SDK

LINEAR ALGEBRA

- cuBLAS
- cuRAND
- cuSOLVER
- CUDA Math library

PARALLEL ALGORITHMS

- nvGRAPH
- NCCL
- Thrust
LESSON: PERFORMANCE DOESN’T MATTER**

• Hypothesis: Performance is not the primary** concern of a domain scientist

Primary incentive to port

Correctness

Scientific Output

Performance

(Maybe)
“The next generation of supercomputers will most likely be similar to the last generation of supercomputers built in the early to mid-90s... but significantly faster.”

— Vincent Scarafino, Ford Motor Co

The government used to do just that, sponsoring development of high-end supercomputer architectures like the Cray vector machines. But now it seems to favor huge clusters of commodity microprocessors. Yes, in the mid-1990s they said that microprocessors were getting faster and faster, and we just need to put a whole bunch of them together and we’ve got a supercomputer. Well, it doesn’t work quite that way, microprocessors are fast at computing but in order to run real difficult problems, they have to have real fast access to memory and be able to do I/O quickly. And memory subsystems are extremely expensive.

If you look at the very large machines made up of off-the-shelf components, they get about 5% of their theoretical peak performance. But if you look at the Earth Simulator, you see numbers from the high 30s to mid-90s. To actually compute what kind of damage is done to human organs — the brain or liver, for example. Today’s analyses with test dummies are very crude. They find at a guess level whether that kind of crash is survivable. But occupant injury analysis takes much more computing power than is available now.

What else would you like to be able to do? Try to understand how exotic materials would work, well enough to understand if they’d work in vehicles. These composite materials are very strong, but understanding how they would react in a failure mode is a difficult problem to solve with today’s computers.

What will the next generation of supercomputers look like? The next generation of supercomputers will most likely be similar to the last generation of supercomputers built in the early to mid-1990s. But they will be significantly faster and able to execute difficult algorithms at speeds much closer to theoretical peak rates than commodity-based machines are able to do.

Will there be any breakthroughs in software over the next five years? There has been significant progress in the area of parallel processing during the last eight years. I would expect continued evolution. I am not aware of any specific areas that seem ripe for breakthroughs, but these things are difficult to predict. Software cannot substitute for raw processing speed.

Best hardware/technology? Nothing.
The firm also announced that it has inked a pact with IBM and Sun Microsystems to develop a programming standard that will allow the same application written in Fortran to run unmodified on workstations, mainframes and supercomputers. The joint effort to develop common software standard helped clinch the sale...The firm had seriously considered Intel Corp until learning of the pact”—Steve Cone, Senior VP Amex
HACKATHONS

2016 GPU HACKATHONS

Five Days + Two Mentors + HPC App + Your Team

don’t swim alone, bring friends!

bit.ly/2016GPUHack

Events

- TU-Dresden
  Deadline: Dec 15th
  Date: 29 Feb

- University of Delaware
  Deadline: Mar 4th
  Date: 2 May

- CSCS
  Deadline: Apr 9th
  Date: 4 Jul

- ORNL
  Deadline: Aug 1st
  Date: 17 Oct

- TBD
  Coming Soon!

OAK RIDGE NATIONAL LABORATORY

COMPUTING FACILITY
# GPU ACCELERATED LIBRARIES

“Drop-in” Acceleration for Your Applications

## DEEP LEARNING
- **cuDNN**
- TensorRT
- DeepStream SDK

## LINEAR ALGEBRA
- **cuBLAS**
- cuRAND
- cuSPARSE
- cuSOLVER
- CUDA Math library

## SIGNAL, IMAGE & VIDEO
- cuFFT
- NVIDIA NPP
- CODEC SDK

## PARALLEL ALGORITHMS
- nvGRAPH
- NCCL
- Thrust
NVIDIA® JETSON AGX XAVIER™ DEVELOPER KIT

Powering AI in Autonomous Machines

BUY NOW
DEEP LEARNING IN DRUG DEVELOPMENT

- DISCOVERY
- VARIANT CALLING
- HIGH CONTENT SCREENING
- PRE & CLINICAL PATHOLOGY
- REAL WORLD EVIDENCE DATA ANALYTICS

- detector
- molecular systems biology
- Deep learning for high-content imaging screens
- PathAI
- CLOUDMEDX
PATH OF FUTURE ARCHITECTURES SHOWS INCREASING PARALLELISM

- Hierarchical parallelism
- Hierarchical data spaces

Jaguar XT5
- 6 core upgrade
  2008

Titan XK7
- 16 cores
- GPU upgrade
  2012

Summit
- Hybrid Accelerated
  2017

OLCF5
- ?
  2022
With the Earth’s population at 7 billion and growing, understanding population distribution is essential to meeting societal needs for infrastructure, resources and vital services. Using GPUs with CUDA and deep learning, Oak Ridge National Laboratory can quickly process high-resolution satellite imagery to map human settlements and changing urban dynamics. With the ability to process a major city in minutes, ORNL can provide emergency response teams critical information that used to take days to create.
COLLABORATION IS KEY

- **teams**: Application teams get dedicated assistance from 2 mentors during week. *Their own apps, their own team members*

- **partners**: Partners benefit from direct user interaction. *Builds goodwill and trust, exposes centers*

- **tools**: Tools are improved through live bug identification and tool + developer observation

- **centers**: Centers increase application readiness from current and future users.
CUDA TOOLKIT 10.0

TURING AND NEW SYSTEMS
New GPU Architecture, Tensor Cores, NVSwitch Fabric

CUDA PLATFORM
CUDA Graphs, Vulkan & DX12 Interop, Warp Matrix

LIBRARIES
GPU-accelerated hybrid JPEG decoding, Symmetric Eigenvalue Solvers, FFT Scaling

DEVELOPER TOOLS
New Nsight Products - Nsight Systems and Nsight Compute
WHAT ARE YOU WORKING ON?

TRYING TO FIX THE PROBLEMS I CREATED WHEN I TRIED TO FIX THE PROBLEMS I CREATED WHEN I TRIED TO FIX THE PROBLEMS I CREATED WHEN...

Try many ideas and APIs; avoid making costly “singular roadmap” decisions.
"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov