GPU ACCELERATED COMPUTING IN HPC AND IN THE DATA CENTER

Peter Messmer, DATE 2019, March 27 2019
RISE OF GPU COMPUTING

APPLICATIONS
ALGORITHMS
SYSTEMS
CUDA
ARCHITECTURE

NVIDIA POWERS WORLD’S FASTEST SUPERCOMPUTERS

48% More Systems | 22 of Top 25 Greenest

ORNL Summit
World’s Fastest
27,648 GPUs | 144 PF

LLNL Sierra
World’s 2nd Fastest
17,280 GPUs | 95 PF

Piz Daint
Europe’s Fastest
5,704 GPUs | 21 PF

ABCI
Japan’s Fastest
4,352 GPUs | 20 PF

ENI HPC4
Fastest Industrial
3,200 GPUs | 12 PF
THE NEW HPC MARKET

SIMULATION

MACHINE LEARNING

DEEP LEARNING
NVIDIA POWERS 5 OF 6 GORDON BELL NOMINATIONS

GPU Acceleration Critical To HPC At Scale Today

- Genomics: 2.36 ExaOps
- Weather: 1.13 ExaOps
- Seismic: 1st Soil & Structure Simulation
- Material Science: 300X Higher Performance
- Quantum Chromodynamics: <1% of Uncertainty Margin
# TESLA UNIVERSAL ACCELERATION PLATFORM

Single Platform To Drive Utilization and Productivity

## CUSTOMER USECASES
- Speech
- Translate
- Recommender

## APPS & FRAMEWORKS
- Python
- TensorFlow
- MXNet
- Openacc
- Amber
- ANSYS
- ANSYS
- NAMD
- Simulia

## NVIDIA SDK & LIBRARIES
- **MACHINE LEARNING | RAPIDS**
  - cuDF
  - cuML
  - cuGRAPH

- **DEEP LEARNING**
  - cuDNN
  - cuBLAS
  - CUTLASS
  - NCCL
  - TensorRT

- **SUPERCOMPUTING**
  - CuBLAS
  - CuFFT
  - OpenACC

## TESLA GPUs & SYSTEMS
- TESLA GPU
- VIRTUAL GPU
- NVIDIA DGX FAMILY
- NVIDIA HGX
- SYSTEM OEM
- CLOUD
EXPANDING VALUE FOR HPC CUSTOMERS
Partnering With HPC Development Community

MORE PERFORMANCE WITH SAME GPU

ADDIMG NEW AND IMPROVED TOP APPLICATIONS

AMBER
CHROMA
GTC
LAMMPS
MILC
NAMD
QUANTUM ESP
SPECFEM3D

Cryogenics
Chemistry
Microscopy
Genomics
Weather

CRYOSPARC
FUN3D
GROMACS
MICROVOLUTION
PARABRICKS
WRF

2018
2019

CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ 4 NVIDIA V100 PCIe or SXM2 GPUs
CUDA DEVELOPMENT ECOSYSTEM

**GPU Users**

- Applications

**Domain Specialists**

- Frameworks

**Problem Specialists**

- Libraries

**New Algorithm Developers and Optimization Experts**

- Directives and Standard Languages

- Extended Standard Languages

**CUDA-C++**

- CUDA Fortran

---

**Ease of use**  **Specialized Performance**

**CUDA: Programming Model, GPU Architecture, System Architecture**
NEW PROGRAMMING MODEL FEATURES

**Execution**
- Asynchronous Task Graphs

**Precision**
- IEEE-754.2008 FP16 Specification
  - Binary representation: 0111001100100000
  - Mantissa: 0.707031

**Interop**
- Lightweight Graphics Interop

**Efficiency**
- NVCC Enhancements

**Turing**
- Multi-Precision Tensor Cores

**FP16 Operations**
- `atomicAdd(&h, (half)1.15f);`
- `half2 hvec(0.94f, -2.13f); atomicAdd(&h2, hvec);`
INDEPENDENT THREAD SCHEDULING

Communicating Algorithms

Pascal: Lock-Free Algorithms
Threads cannot wait for messages

Volta/Turing: Starvation Free Algorithms
Threads may wait for messages
ASYNCHRONOUS TASK GRAPHS
Execution Optimization When Workflow is Known Up-Front

Loop & Function offload

DL Inference

Linear Algebra

Deep Neural Network Training

HPC Simulation
DEFINITION OF A CUDA GRAPH

Graph Nodes Are Not Just Kernel Launches

Sequence of operations, connected by dependencies.

Operations are one of:

- **Kernel Launch**  CUDA kernel running on GPU
- **CPU Function Call**  Callback function on CPU
- **Memcopy/Memset**  GPU data management
- **Sub-Graph**  Graphs are hierarchical
WHAT IS OPENACC
Open Specification Developed by OpenACC.org Consortium

Directives-based programming model for parallel computing

Add Simple Compiler Directive
main()
{
    <serial code>
    #pragma acc kernels
    {
        <parallel code>
    }
}

POWERFUL & PORTABLE

Read more at www.openacc.org/about
WHO OPENACC IS FOR

The Main Focus

1. Want to do more science & less programming
2. Believe that GPUs are hard
3. Need help in learning how to easy start with GPUs
4. Mostly don’t have a computer science degree

Domain Scientists

Application Developers

Looking for:
1. easy code maintenance,
2. better efficiency,
3. portability

Mostly computer scientists
OPENACC GROWING MOMENTUM
Wide Adoption Across Key HPC Codes

Over 100 Apps* Using OpenACC

ANSYS Fluent  GTC
Gaussian      XGC
VASP          ACME
LSDalton      FLASH
MPAS          COSMO
GAMERA        Numeca

VASP
Top Quantum Chemistry and Material Science Code

For VASP, OpenACC is the way forward for GPU acceleration. Performance is similar to CUDA, and OpenACC dramatically decreases GPU development and maintenance efforts. We’re excited to collaborate with NVIDIA and PGI as an early adopter of Unified Memory.

Prof. Georg Kresse
Computational Materials Physics
University of Vienna

* Applications in production and development
SINGLE CODE FOR MULTIPLE PLATFORMS

OpenACC - Performance Portable Programming Model for HPC

OpenPOWER
Sunway
x86 CPU
x86 Xeon Phi
NVIDIA GPU
AMD GPU
PEZY-SC

AWE Hydrodynamics CloverLeaf mini-App, bm32 data set
http://uk-mac.github.io/CloverLeaf

Systems: Haswell: 2x16 core Haswell server, four K80s, CentOS 7.2 (perf-hsw10), Broadwell: 2x20 core Broadwell server, eight P100s (dgx1-prd-01), Broadwell server, eight V100s (dgx07), Skylake 2x20 core Xeon Gold server (sky-4).
Compilers: Intel 2018.0.128, PGI 18.1
Benchmark: CloverLeaf v1.3 downloaded from http://uk-mac.github.io/CloverLeaf the week of November 7 2016; CloverLeaf_Serial; CloverLeaf_ref (MPI+OpenMP); CloverLeaf_OpenACC (MPI+OpenACC)
Data compiled by PGI February 2018.
NSIGHT SYSTEMS
System-wide Performance Analysis

Observe Application Behavior: CPU threads, GPU traces, Memory Bandwidth and more

Locate Optimization Opportunities: CUDA & OpenGL APIs, Unified Memory transfers, User Annotations using NVTX

Ready for Big Data: Fast GUI capable of visualizing in excess of 10 million events on laptops, Container support, Minimum user privileges

https://developer.nvidia.com/nsight-systems
Processes and threads
CUDA and OpenGL API trace
cuDNN and cuBLAS trace
Kernel and memory transfer activities
Multi-GPU
Thread/core migration
Thread state
CONTAINERS: SIMPLIFYING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments
- Eliminates complex, time-consuming builds and installs

Get started in minutes
- Simply Pull & Run the app

Portable
- Deploy across various environments, from test to production with minimal changes
NGC CONTAINERS: ACCELERATING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments
- Eliminates complex, time-consuming builds and installs

Get started in minutes
- Simply Pull & Run the app

Portable
- Deploy across various environments, from test to production with minimal changes

WHY NGC CONTAINERS

Optimized for Performance
- Monthly DL container releases offer latest features and superior performance on NVIDIA GPUs

Scalable Performance
- Supports multi-GPU & multi-node systems for scale-up & scale-out environments

Designed for Enterprise & HPC environments
- Supports Docker & Singularity runtimes

Run Anywhere
- Pascal/Volta/Turing-powered NVIDIA DGX, PCs, workstations, servers and top cloud platforms
THE NEW NGC
GPU-optimized Software Hub. Simplifying DL, ML and HPC Workflows

50+ Containers
DL, ML, HPC

50+ Pre-trained Models
NLP, Classification, Object Detection & more

10+ Model Training Scripts
NLP, Image Classification, Object Detection & more

Industry Workflows
Medical Imaging, Intelligent Video Analytics

Simplify Deployments
Innovate Faster
Deploy Anywhere

ngc.nvidia.com
NGC-READY ECOSYSTEM
Now Over 50 GPU-Optimized Containers
RE-IMAGINING DATA SCIENCE WORKFLOW
Open Source, End-to-end GPU-accelerated Workflow Built On CUDA

data → cuDF → cuML → Visualization

- cuDF: Data preparation / wrangling
- cuML: Optimized ML model training
- Visualization: Data visualization libraries

insights
RAPIDS — OPEN GPU DATA SCIENCE

Software Stack Python

- Data Preparation: cuDF
- Model Training: cuML
- Visualization: cuGRAPH

Diagram:

- DASK
- CUDA
- APACHE ARROW on GPU Memory
- CUDNN
- DEEP LEARNING FRAMEWORKS
- RAPIDS
- cuDF
- cuML
- cuGRAPH

Python
ACCELERATING MACHINE LEARNING
The RAPIDS Ecosystem

Open Source Community
- Anaconda
- Apache Arrow
- gunmen
- PyData
- Ursa Labs

Enterprise Data Science Platforms
- IBM
- Oracle
- SAS
- SAP

Startups
- BlazingDB
- FastData
- Graphistry
- H2O.ai
- Kinetica
- MapR
- OmniSci
- Chainer
- PyTorch

Deep Learning Integration

GPU Servers
- Cisco
- DELL EMC
- IBM
- Lenovo
- Hewlett Packard Enterprise

Storage Partners
- DELL EMC
- IBM
- NetApp
- Pure Storage
SUMMARY

GPUs are established in HPC and Datacenter

Full stack optimization, not just selling silicon

Improvements and simplification on multiple fronts

- HW: chip, node and system level
- SW: low- and high-level languages, libraries, frameworks, apps

Convergence of HPC and accelerated machine learning in the data center