GPU Computing for HPC and AI

Mason Wu, Solutions Architect
Oct. 2019
EVOLUTION OF COMPUTING

1995
PC Internet
WinTel, Yahoo!
1 billion PC users

2005
Mobile-Cloud
iPhone, Amazon AWS
2.5 billion mobile users

2015
AI & IOT
Deep Learning, GPU
100s of billions of devices
NVIDIA
“THE AI COMPUTING COMPANY”

GPU Computing

Computer Graphics

Artificial Intelligence
RISE OF GPU COMPUTING

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp
HOW GPU ACCELERATION WORKS

Application Code

Compute-Intensive Functions

5% of Code

Rest of Sequential CPU Code

GPU

CPU
## BEYOND MOORE’S LAW

### Progress Of Stack In 6 Years

<table>
<thead>
<tr>
<th>2013</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>cuBLAS: 5.0</td>
<td>cuBLAS: 10.0</td>
</tr>
<tr>
<td>cuFFT: 5.0</td>
<td>cuFFT: 10.0</td>
</tr>
<tr>
<td>cuRAND: 5.0</td>
<td>cuRAND: 10.0</td>
</tr>
<tr>
<td>cuSPARSE: 5.0</td>
<td>cuSPARSE: 10.0</td>
</tr>
<tr>
<td>NPP: 5.0</td>
<td>NPP: 10.0</td>
</tr>
<tr>
<td>Thrust: 1.5.3</td>
<td>Thrust: 1.9.0</td>
</tr>
<tr>
<td>CUDA: 5.0</td>
<td>CUDA: 10.0</td>
</tr>
<tr>
<td>Resource Mgr: r304</td>
<td>Resource Mgr: r384</td>
</tr>
<tr>
<td>Base OS: CentOS 6.2</td>
<td>Base OS: Ubuntu 16.04</td>
</tr>
</tbody>
</table>

---

Accelerated Server With Fermi

Accelerated Server With Volta

---

**Measured performance of** Amber, CHROMA, GTC, LAMMPS, MILC, NAMD, Quantum Espresso, SPECfem3d
NVIDIA DATA CENTER PLATFORM
Single Platform Drives Utilization and Productivity

CUSTOMER USE CASES
- Speech
- Translate
- Recommender
- Healthcare
- Manufacturing
- Finance
- Molecular Simulations
- Weather Forecasting
- Seismic Mapping
- Creative & Technical
- Knowledge Workers

CONSUMER INTERNET & INDUSTRY APPLICATIONS

SCIENTIFIC APPLICATIONS

VIRTUAL GRAPHICS

APPS & FRAMEWORKS
- Python
- TensorFlow
- MXNet
- Chainer
- ONNX
- Amber
- NAMD
- +600 Applications
- RAPIDS
- PyTorch

CUDA-X & NVIDIA SDKs
- MACHINE LEARNING
  - cuDF
  - cuML
  - cuGRAPH
- DEEP LEARNING
  - cuDNN
  - CUTLASS
  - TensorRT
- HPC
  - OpenACC
  - cuFFT
- VIRTUAL GPU
  - vDWS
  - vPC
  - vAPPS

CUDA & CORE LIBRARIES - cuBLAS | NCCL

TESLA GPUs & SYSTEMS
- TESLA GPU
- EVERY OEM
- EVERY MAJOR CLOUD
NVIDIA CUDA-X
Software To Deliver Acceleration For HPC & AI Apps; 500+ New Updates

|----------------------------------|----------------------------------|-----------------------------|--------------------------------|---------------------|-----------------|-------------------------------|-----------------|-----------------------------|

600+ Apps

<table>
<thead>
<tr>
<th>Linear Algebra</th>
<th>Parallel Algorithms</th>
<th>Signal Processing</th>
<th>Deep Learning</th>
<th>Machine Learning</th>
<th>Visualization</th>
</tr>
</thead>
</table>

CUDA-X HPC & AI
40+ GPU Acceleration Libraries

CUDA

| Desktop Development | Data Center | Supercomputers | GPU-Accelerated Cloud |
MOST ADOPTED PLATFORM FOR ACCELERATING HPC

- 13M CUDA Downloads
- 600+ Applications Accelerated
- 127 Systems on Top 500

4X IN 5 YEARS
ALL TOP 15 APPLICATIONS ACCELERATED
NEW HIGHS IN TOP 500 LIST

- World's #1 Summit: 144 PF
- Europe’s #1 Piz Daint: 21 PF
- Japan's #1 ABCI: 20 PF
- Industrial #1 ENI: 12 PF
- Taiwan’s #1 NCHC: 9 PF
MOST ADOPTED PLATFORM FOR ACCELERATING AI

8 MLPerf 0.6 Training Records

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Detection (Heavy Weight) Mask R-CNN</td>
<td>18.47 Mins</td>
</tr>
<tr>
<td>Translation (Recurrent) GNMT</td>
<td>1.8 Mins</td>
</tr>
<tr>
<td>Reinforcement Learning (MiniGo)</td>
<td>13.57 Mins</td>
</tr>
<tr>
<td>Object Detection (Heavy Weight) Mask R-CNN</td>
<td>25.39 Hrs</td>
</tr>
<tr>
<td>Object Detection (Light Weight) SSD</td>
<td>3.04 Hrs</td>
</tr>
<tr>
<td>Translation (Recurrent) GNMT</td>
<td>2.63 Hrs</td>
</tr>
<tr>
<td>Translation (Non-recurrent)Transformer</td>
<td>2.61 Hrs</td>
</tr>
<tr>
<td>Reinforcement Learning (MiniGo)</td>
<td>3.65 Hrs</td>
</tr>
</tbody>
</table>

RECORD-SETTING PERFORMANCE

END-TO-END SOFTWARE STACK

AVAILABLE EVERYWHERE
NVIDIA ENTERPRISE GPU PRODUCT FAMILY
Computing For Modern Enterprise Workloads

DATACENTER SERVERS

SPECIALIZED
(Max Performance)
AI & HPC
Model Development

MAINSTREAM
(Max Utility)
Rendering
Enterprise Application
Deployment

RTX 8000/6000
RT Core, Tensor Core
48/24 GB GDDR6
250W/300W

T4
Tensor Core, RT Core
16GB GDDR6, 70W

V100
Tensor Core, NVLink
32GB HBM2
250W/300W

WORKSTATIONS

DATA SCIENCE
AI Development
RTX 8000/6000
Tensor Core
48/24 GB GDDR6

VISUALIZATION
Design & Graphics
RTX 8000/6000/5000*/4000*
RT Core, Tensor Core

*Not designed & cannot be qualified for Servers
# End-to-End Product Family

## HPC / Training

<table>
<thead>
<tr>
<th>Deskto</th>
<th>Workstatio</th>
<th>Virtual Workstation</th>
<th>Data Center</th>
<th>Data Center</th>
<th>Automotive</th>
<th>Embedded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titan/GeForce</td>
<td>GPU Workstation</td>
<td>Virtual GPU</td>
<td>Tesla V100/T4</td>
<td>Tesla V100</td>
<td>Drive AGX Pegasus</td>
<td>Jetson AGX Xavier</td>
</tr>
</tbody>
</table>

## Inference

- Tesla T4
- Drive AGX Pegasus
- Jetson AGX Xavier
<table>
<thead>
<tr>
<th>Segment</th>
<th>Description</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Learning Training</td>
<td>Purpose built cluster for fastest training time</td>
<td>• 8-16 V100 NVLINK</td>
</tr>
<tr>
<td>High Perf. Computing</td>
<td>Scientific computing center or enterprise running HPC, AI, Data Science workload</td>
<td>• 4 V100 NVLINK (e.g. supercomputing center, enterprise)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 2-4 V100 PCIe (e.g. higher ed.)</td>
</tr>
<tr>
<td>Render Farms</td>
<td>Purpose built cluster for batch and real time rendering</td>
<td>• 4-8 RTX 6K/8K</td>
</tr>
<tr>
<td>Performance Graphics</td>
<td>Best graphics performance, professional graphics applications</td>
<td>• 8 RTX 6K/8K (e.g. remote workstations)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 4 RTX 6K/8K (e.g. virtual workstations)</td>
</tr>
<tr>
<td>Mainstream Graphics</td>
<td>Virtual desktop with cost effective TCO per user</td>
<td>• 2-8 T4 (balanced graphics and cost)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 2-4 M10 (absolute lowest TCO per user)</td>
</tr>
<tr>
<td>Enterprise Acceleration</td>
<td>Mixed Workloads - Graphics, ML, DL, analytics, training, inference</td>
<td>• 2-4 RTX 6K/8K (graphics Intensive)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 2-4 V100 (compute Intensive)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 4-8 T4 (balanced)</td>
</tr>
<tr>
<td>Edge Acceleration</td>
<td>Edge solutions with differing use cases and deployment location</td>
<td>• 1-8 T4 (inference and video decode intensive, e.g. IVA, industrial inspection)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 2-4 RTX 6K/8K (graphics intensive, e.g. augmented/virtual reality)</td>
</tr>
<tr>
<td><strong>GPU CHIP</strong></td>
<td><strong>P100 (SXMM2)</strong></td>
<td><strong>P100 (PCIE)</strong></td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------</td>
<td>----------------</td>
</tr>
<tr>
<td><strong>PEAK FP64 (TFLOPs)</strong></td>
<td>5.3</td>
<td>4.7</td>
</tr>
<tr>
<td><strong>PEAK FP32 (TFLOPs)</strong></td>
<td>10.6</td>
<td>9.3</td>
</tr>
<tr>
<td><strong>PEAK FP16 (TFLOPs)</strong></td>
<td>21.2</td>
<td>18.7</td>
</tr>
<tr>
<td><strong>PEAK TOPs</strong></td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Memory Size</strong></td>
<td>16 GB HBM2</td>
<td>16/12 GB HBM2</td>
</tr>
<tr>
<td><strong>Memory BW</strong></td>
<td>732 GB/s</td>
<td>732/549 GB/s</td>
</tr>
<tr>
<td><strong>Interconnect</strong></td>
<td>NVLINK + PCIe Gen3</td>
<td>PCIe Gen3</td>
</tr>
<tr>
<td><strong>ECC</strong></td>
<td>Internal + HBM2</td>
<td>Internal + HBM2</td>
</tr>
<tr>
<td><strong>Form Factor</strong></td>
<td>SXM2</td>
<td>PCIE Dual Slot</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>300 W</td>
<td>250 W</td>
</tr>
</tbody>
</table>
HIGH PERFORMANCE COMPUTING
PARALLEL COMPUTING DEVICES

CPU
Optimized for Serial Tasks

GPU Accelerator
Optimized for Parallel Tasks
### HOW TO START WITH GPUs

**Applications**

1. Review available GPU-accelerated applications
2. Check for GPU-Accelerated applications and libraries
3. Add OpenACC Directives for quick acceleration results and portability
4. Dive into CUDA for highest performance and flexibility

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy to use</td>
<td>Easy to Start</td>
<td>Most Performance</td>
<td>Most Performance</td>
</tr>
<tr>
<td>Most Performance</td>
<td>Portable Code</td>
<td>Most Flexibility</td>
<td></td>
</tr>
</tbody>
</table>
INDUSTRY EMBRACING GPU SUPERCOMPUTING

Oil and Gas Discovery
10X increase in data processing

Realtime Fleet Analytics
Streamline routes to save >$28M

Engineering Design + AI
Accelerate from 3-4 days to hours
NVIDIA POWERS TODAY’S FASTEST SUPERCOMPUTERS

22 of Top 25 Greenest

- **ORNL Summit**
  - World’s Fastest
  - 27,648 GPUs | 149 PF

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

- **Total Pangea 3**
  - Fastest Industrial
  - 3,348 GPUs | 18 PF

- **Taiwania 2**
  - Taiwan’s Fastest
  - 2,016 GPUs | 9 PF

- **Piz Daint**
  - Europe’s Fastest
  - 5,704 GPUs | 21 PF
## NVIDIA CUDA-X UPDATES

Software To Deliver Acceleration For HPC & AI Apps; 500+ New Updates

|----------------------------------|-----------------------------------|-------------------------------|--------------------------------|----------------------|------------------|-------------------------------|-------------------|-----------------------------|

### 600+ Apps

- Linear Algebra
- Parallel Algorithms
- Signal Processing
- Deep Learning
- Machine Learning
- Visualization

### CUDA-X HPC & AI
40+ GPU Acceleration Libraries

### CUDA
- Desktop Development
- Data Center
- Supercomputers
- GPU-Accelerated Cloud
EASE OF ACCELERATION WITH OPENACC

-200 APPS BEING ACCELERATED

ACROSS SCIENTIFIC DOMAINS
DRAMATICALLY MORE FOR YOUR MONEY

6X Better HPC TCO for Same Throughput

CPU-Only

160 Self-hosted Skylake CPU Servers
96 KWatts

MIXED HPC WORKLOAD:
Amber, CHROMA, GTC, LAMMPS, MILC, NAMD, Quantum Espresso, SPECFEM3D

GPU-Accelerated

6 Accelerated Servers w/4 V100 GPUs
10 KWatts

SAME THROUGHPUT

1/6 THE COST

1/10 THE POWER

1/11 THE SPACE

6X Better HPC TCO for Same Throughput
INTERSECTION OF HPC & AI
TRANSFORMING SCIENCE

HPC
> Algorithms based on first principles theory
> Proven models for accurate results

AI
> Neural networks that learn patterns from large data sets
> Improve predictive accuracy and faster response time

SPEEDING PATH TO FUSION ENERGY
EXASCALE WEATHER MODELING
IDENTIFYING CHEMICAL COMPOUNDS
O&G FAULT INTERPRETATION

90% Prediction Accuracy
Publish in Nature April 2019
Tensor Cores Achieved 1.13 EF
2018 Gordon Bell Winner
Orders Of Magnitude Speedup
3M New Compounds In 1 Day
Time-to-solution Reduced From Weeks To 2 Hours
AI FOR SCIENCE
Transformative Tool To Accelerate The Pace of Scientific Innovation

Improves Accuracy
Enabling realization of full scientific potential

Accelerates Time to Solution
Unlocking the use of science in exciting new ways

90% accuracy
Fusion Sustainment
Clean Energy

33% Faster
Track Neutrinos
Particle Physics

5,000X Faster
Process LIGO Signal
Understanding Universe

300,000X Faster
Predict Molecular Energetics
Drug Discovery

70% accuracy
Score Protein Ligand
Drug Discovery

11% higher accuracy
Monitor Earth’s Vital
Climate

Weeks to 10 milliseconds
Analyze Gravitational Lensing
Astrophysics

14X Faster
Generate Bose-Einstein
Condensate (Physics)
Mixed-Precision Computing

TENSOR CORES FOR SCIENCE

FP64+ MULTI-PRECISION

PLASMA FUSION APPLICATION

EARTHQUAKE SIMULATION

MIXED PRECISION WEATHER PREDICTION

FP16 Solver 3.5x faster

FP16-FP21-FP32-FP64 25x faster

FP16/FP32/FP64 4x faster
ARTIFICIAL INTELLIGENCE
DEEP LEARNING

TRAINING
Learning a new capability from existing data

INFEERENCE
Applying this capability to new data

Untrained Neural Network Model

Deep Learning Framework

TRAINING DATASET

Trained Model
New Capability

NEW DATA

App or Service
Featuring Capability

Trained Model
Optimized for Performance
“SUPERHUMAN” RESULTS
SPARK HYPERSCALE ADOPTION

ImageNet — Accuracy %

<table>
<thead>
<tr>
<th>Year</th>
<th>Hand-coded CV</th>
<th>Deep Learning</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>72%</td>
<td>74%</td>
<td>74%</td>
</tr>
<tr>
<td>2011</td>
<td>74%</td>
<td>74%</td>
<td>74%</td>
</tr>
<tr>
<td>2012</td>
<td>76%</td>
<td>84%</td>
<td>88%</td>
</tr>
<tr>
<td>2013</td>
<td>86%</td>
<td></td>
<td>93%</td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td>96%</td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cloud Services with AI Powered by NVIDIA

Alibaba/Aliyun  Amazon  Baidu  eBay  Facebook
Flickr  Google  iFLYTEK  iQIYI  JD.com
Orange  Periscope  Pinterest  Qihoo 360  Shazam
Skype  Sogou  Twitter  Yahoo Supermarket  Yandex  Yelp
WORLD RECORDS FOR CONVERSATIONAL AI

BERT Training and Inference Records
Largest Transformer Based Model Ever Trained

EXPLODING MODEL SIZE

Complexity to Train

CONVERSATIONAL AI RECORDS

Code Available on Github

Training GPUs - Near Linear Scaling
Requires Leading AI Infrastructure

BERT<sub>LARGE</sub> Training Record: 1472 Tesla V100-SXM3-32GB 450W GPUs | 92 DGX-2H Servers | 8 Mellanox Infiniband Adapters per node
BERT<sub>BASE</sub> Inference Record: SQuAD Dataset | Tesla T4 16GB GPU | CPU: Intel Xeon Gold 6240 & OpenVINO v2
Scaling Training Performance on: BERT<sub>LARGE</sub> | Speedups show performance scaling on 1x, 16x, 64x and 92x DGX-2H Servers with 16 NVIDIA V100 GPUs each

Number of Parameters by Network

- ResNet-50
- Transformer
- GPT-1
- BERT<sub>BASE</sub>
- GPT-2
- GPT-2 8B

- NLP - Generative Tasks (Chatbots, Auto-completion)
- Image Recognition
- NLP (Q&A, Translation)

- 26M
- 340M
- 1.5Bn
- 8.3Bn

- 53 minutes
- BERT<sub>LARGE</sub>
- Speed Training Record

- 8.3Bn
- GPT-2 8B
- Largest Transformer Based Model Trained

- 2.2ms
- BERT<sub>BASE</sub>
- Fastest Inference (18X Faster Than CPU)
POWERING THE DL ECOSYSTEM
NVIDIA SDK accelerates every major framework

COMPUTER VISION
OBJECT DETECTION
IMAGE CLASSIFICATION

SPEECH & AUDIO
VOICE RECOGNITION
LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING
RECOMMENDATION ENGINES
SENTIMENT ANALYSIS

DEEP LEARNING FRAMEWORKS
Caffe
Caffe2
Chainer
Cognitive Toolkit
KALDI
mxnet
PaddlePaddle
PYTORCH
TensorFlow

NVIDIA DEEP LEARNING SDK and CUDA
cuDNN
NCCL
cuBLAS
TensorRT
cuSPARSE
DeepStream SDK

developer.nvidia.com/deep-learning-software
NVIDIA DL SOFTWARE STACK

TRAINING

Data Management
Training
Model Assessment

Trained Neural Network

INFEERENCE

Data center
TensorRT

Embedded
JETPACK SDK

Automotive
DriveWorks SDK

NVIDIA DEEP LEARNING SDK and CUDA

cuDNN
NCCL
cuBLAS
TensorRT
cuSPARSE
DeepStream SDK

developer.nvidia.com/deep-learning-software
AMAZING ACHIEVEMENTS IN AI

- Healthcare
- Security
- Product Design
- Chat bot
END-to-END GPU-ACCELERATED CRYO-EM
Single Compute Platform for AI, High-Performance Computing, and Visualization
Intro to Cryo-Electron Microscopy (CryoEM)

3D protein structures drive Disease research

Drug development

Fundamental biology
Intro to CryoEM

how?

Macromolecular structures at near-atomic resolutions?

using

Electron microscope
Intro to CryoEM

1. Purified proteins in solution
2. Apply
3. Metal wafer ‘EM grid’
4. Freeze
5. Freeze thin layer of proteins in solution
Intro to CryoEM

Metal wafer ‘EM grid’

EM images of $10^4 - 10^7$ proteins

Dimension $\approx$ wavelength of visible light

Combine to resolve 3D structures
Difficulties in CryoEM Object Detection

- Very low SNR; ~0.1
- Non-globular proteins
- Small proteins
- Diverse false positives
Highlights of Topaz

1. **Pick more real particles** with less hand labeling
   a. Topaz enables picking more representative particles (i.e. all particle views)
   b. More particles leads to better downstream results

2. Topaz picks particles of **any shape & size**

3. Reduce bias due to iterative filtering of particles - Topaz particle sets do not require post-processing

4. No need to fully label micrographs - PU framework does not assume micrographs are fully labeled

5. Micrograph normalization in Topaz makes **grid masking a thing of the past**

---

[https://github.com/tbepler/topaz](https://github.com/tbepler/topaz)
Single-particle cryo-electron microscopy

Sample preparation

2D Imaging

Particle detection

2D averaging

3D reconstruction

Particle picking

2D averages

3D reconstruction (not the same protein)

EMPIAR-10096
Tan et al., 2017
Structure determination with Topaz particles

Dataset has micrographs and particle coordinates associated with the published structure

1. Holdout 20% of micrographs for evaluation (test set)
2. Simulate labeling 1000 particles by sampling from the published particle set on the training micrographs
   - Topaz accepts very sparsely labeled micrographs - this is the beauty of PU learning!
MORE INFO

AI FOR CRYOEM DATA SELECTION AND CLASSIFICATION
