BEYOND MOORE’S LAW
Progress Of Stack In 6 Years

2013
- cuBLAS: 5.0
- cuFFT: 5.0
- cuRAND: 5.0
- cuSPARSE: 5.0
- NPP: 5.0
- Thrust: 1.5.3
- CUDA: 5.0
- Resource Mgr: r304
- Base OS: CentOS 6.2

Accelerated Server With Fermi

2019
- cuBLAS: 10.0
- cuFFT: 10.0
- cuRAND: 10.0
- cuSOLVER: 10.0
- cuSPARSE: 10.0
- NPP: 10.0
- Thrust: 1.9.0
- CUDA: 10.0
- Resource Mgr: r384
- Base OS: Ubuntu 16.04

Accelerated Server with Volta

Measured performance of Amber, CHROMA, GTC, LAMMPS, MILC, NAMD, Quantum Espresso, SPECFEM3D


Relative Performance

GPU-Accelerated Computing

Moore’s Law

CPU
AI, MACHINE LEARNING, AND DEEP LEARNING

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.

MACHINE LEARNING
Machine learning begins to flourish.

DEEP LEARNING
Deep learning breakthroughs drive AI boom.
THE BIG PROBLEM IN DATA SCIENCE

Manage Data

- All Data
- ETL
- Structured Data Store

Training

- Data Preparation
- Model Training

Evaluate

- Visualization

Deploy

- Inference

Slow Training Times for Data Scientists
DATA SCIENCE IS THE KEY TO MODERN BUSINESS

Use Cases in Every Industry

**CONSUMER INTERNET**
- Ad Personalization
- Click Through Rate Optimization
- Churn Reduction

**FINANCIAL SERVICES**
- Claim Fraud
- Customer Service Chatbots/Routing
- Risk Evaluation

**HEALTHCARE**
- Improve Clinical Care
- Drive Operational Efficiency
- Speed Up Drug Discovery

**RETAIL**
- Supply Chain & Inventory Management
- Price Management / Markdown Optimization
- Promotion Prioritization And Ad Targeting

**OIL & GAS**
- Sensor Data Tag Mapping
- Anomaly Detection
- Robust Fault Prediction

**MANUFACTURING**
- Remaining Useful Life Estimation
- Failure Prediction
- Demand Forecasting

**TELECOM**
- Detect Network/Security Anomalies
- Forecasting Network Performance
- Network Resource Optimization (SON)

**AUTOMOTIVE**
- Personalization & Intelligent Customer Interactions
- Connected Vehicle Predictive Maintenance
- Forecasting, Demand, & Capacity Planning

**RETAIL**
- Personalization & Intelligent Customer Interactions
- Connected Vehicle Predictive Maintenance
- Forecasting, Demand, & Capacity Planning
HOW GPU ACCELERATION WORKS

Application Code

GPU

5% of Code

Compute-Intensive Functions

Rest of Sequential CPU Code

CPU
NVIDIA BREAKS RECORDS IN AI PERFORMANCE
MLPerf Records Both At Scale And Per Accelerator

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

Per Accelerator comparison using reported performance for MLPerf 0.6 NVIDIA DGX-2H (16 V100s) compared to other submissions at same scale except for MiniGo where NVIDIA DGX-1 (8 V100s) submission was used. MLPerf ID Max Scale: Mask R-CNN: 0.6-23, GNMT: 0.6-26, MiniGo: 0.6-11 | MLPerf ID Per Accelerator: Mask R-CNN, SSD, GNMT, Transformer: all use 0.6-20, MiniGo: 0.6-10
4X MORE PERFORMANCE, SAME SERVER
Rapid Software Innovation Delivers Continuous Improvements

Comparing the performance of a single DGX-1 server at launch and MLPerf ID 0.6-8
DGX REFERENCE ARCHITECTURE SOLUTIONS
Growing Ecosystem of IT-approved Solutions for AI infrastructure

Benefits:

• No more design guesswork
• Faster, simpler deployment
• Predictable performance at scale
• Simplified, single-point of support

NVIDIA Confidential
ACCELERATING MAINSTREAM BUSINESS SERVERS
Modern Enterprise Computing Platform

T4 GPUs

Containers

NGC Containers
ML/DA/DL

NGC Ready Support
Direct Support from NVIDIA

CISCO UCS C240 M5
Dell PowerEdge R740
HPE Proliant DL380 Gen 10
Lenovo ThinkSystem SR670

Machine Learning
Virtual Graphics
Deep Learning
IVA/HPC/Others

NVIDIA Confidential
VAST WORLD OF AI INFERENCE
Creating A Massive Market Opportunity

General Purpose Computers

Embedded Computers

Embedded Devices
NVIDIA TensorRT 5 INFRINGEMENT PLATFORM
Accelerates Throughput On Leading Industry Platforms
NVIDIA DATA CENTER PLATFORM
Single Platform Drives Utilization and Productivity

CUSTOMER USE CASES
- Speech
- Translate
- Recommender
- Healthcare
- Manufacturing
- Finance

VIRTUAL GPU
- Molecular Simulations
- Weather Forecasting
- Seismic Mapping
- Creative & Technical
- Knowledge Workers

CONSUMER INTERNET & INDUSTRY APPLICATIONS

APPS & FRAMEWORKS
- python
- TensorFlow
- mxnet
- ONNX
- RAPIDS
- PyTorch
- Chainer
- Amber
- NAMD
- +550 Applications

VIRTUAL GRAPHICS

SCIENTIFIC APPLICATIONS
- Molecular Simulations
- Weather Forecasting
- Seismic Mapping
- Creative & Technical
- Knowledge Workers

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

CUDA & CORE LIBRARIES - cuBLAS | NCCL

GPUs & SYSTEMS
- TESLA GPU
- NVIDIA DGX FAMILY
- NVIDIA HGX
- SYSTEM OEM
- CLOUD