THE POWER OF GPU-ACCELERATED DEEP LEARNING

Gabriel Noaje
Senior Solutions Architect, NVIDIA APAC South
gnoaje@nvidia.com
ACCELERATED DATA SCIENCE

DATA ANALYTICS
Extracting insights from big data

MACHINE LEARNING
Learning from examples in the data

DEEP LEARNING
Automating feature engineering
# GPU-Accelerated Data Science

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
BEYOND DEEP LEARNING

Opportunities to Accelerate Data Science

ARTIFICIAL INTELLIGENCE

Data Analytics  Machine Learning (Regressions, Decision Trees, Graph)  Deep Learning

Tabular/Sparse Data  Dense Data

2.2 exabytes (2.2B GB) of data created daily - McKinsey

$260B annual revenue by 2022 for big data and business analytics - IDC
## WHAT PROBLEM ARE YOU SOLVING?

Defining the AI/DL task

<table>
<thead>
<tr>
<th>INPUTS</th>
<th>BUSINESS QUESTIONS</th>
<th>AI / DL TASK</th>
<th>EXAMPLE OUTPUTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Is “it” present or not?</td>
<td>Detection</td>
<td>Healthcare</td>
</tr>
<tr>
<td>Text Data</td>
<td>Cancer Detection</td>
<td>Targeted Ads</td>
<td>Cybersecurity</td>
</tr>
<tr>
<td>Images</td>
<td>What type of thing is “it”?</td>
<td>Classification</td>
<td>Image Classification</td>
</tr>
<tr>
<td>Video</td>
<td>To what extent is “it” present?</td>
<td>Segmentation</td>
<td>Tumor Size / Shape Analysis</td>
</tr>
<tr>
<td>Audio</td>
<td>What is the likely outcome?</td>
<td>Prediction</td>
<td>Survivability Prediction</td>
</tr>
<tr>
<td></td>
<td>What will likely satisfy the objective?</td>
<td>Recommendations</td>
<td>Therapy Recommendation</td>
</tr>
</tbody>
</table>
DEEP LEARNING APPLICATION DEVELOPMENT

**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

App or Service Featuring Capability

NEW DATA

Trained Model Optimized for Performance
## CHALLENGES

<table>
<thead>
<tr>
<th>DEEP LEARNING NEEDS</th>
<th>WHY</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Skills</td>
<td>New computing model</td>
</tr>
<tr>
<td>Latest Algorithms</td>
<td>Rapid evolving</td>
</tr>
<tr>
<td>Fast Training</td>
<td>Impossible -&gt; Practical</td>
</tr>
<tr>
<td>Deployment Platforms</td>
<td>Must be available everywhere</td>
</tr>
</tbody>
</table>
RISE OF GPU COMPUTING

AI PLATFORM CONSIDERATIONS
Factors Impacting Deep Learning Platform Decisions

DEVELOPER PRODUCTIVITY

“Must get started now, line of business wants to deliver results yesterday

SCALING PERFORMANCE

“I want the most GPU bang for the buck

TOTAL COST OF OWNERSHIP

“I have limited budget, need lowest up-front cost possible
CHALLENGES WITH COMPLEX SOFTWARE

Current DIY GPU-accelerated AI and HPC deployments are complex and time consuming to build, test and maintain.

Development of software frameworks by the community is moving very fast.

Requires high level of expertise to manage driver, library, framework dependencies.
THE TRUE TCO OF AN AI PLATFORM

1. Designing and Building an AI Compute Platform - from Scratch
COMPARING AI COMPUTE ALTERNATIVES
Looking Beyond the Spec Sheet

Evaluation Criteria

- AI/DL Expertise & Innovation
- AI/DL Software Stack
- Operating System Image
- Hardware Architecture
PURPOSE-BUILT AI SUPERCOMPUTERS

DGX SOFTWARE STACK

- Universal SW for AI
- Predictable execution across platforms
- Pervasive reach

AI WORKSTATION

AI DATA CENTER

DGX Station

DGX-1

DGX-2

AI Workstation for Data Science Teams

The Essential Instrument for AI Research

The World’s Most Powerful AI System for the Most Complex AI Challenges
NVIDIA TENSOR CORE GPU - BUILT FOR AI
Delivering 125 TFLOPS of DL Performance

MATRIX DATA OPTIMIZATION:
Dense Matrix of Tensor Compute

TENSOR-OP CONVERSION:
FP32 to Tensor Op Data for Frameworks

VOLTA TENSOR CORE
4x4 matrix processing array
Optimized For Deep Learning

VOLTA-OPTIMIZED cuDNN

ALL MAJOR FRAMEWORKS

Caffe2, mxnet, DL4J, PyTorch, TensorFlow
NVIDIA DEEP LEARNING SOFTWARE PLATFORM

TRAINING

- Data Management
- Training
- Model Assessment

Trained Neural Network

INFERENCEN

- Data center
- Embedded
- Automotive

- TensorRT
- JETPACK SDK
- DriveWorks SDK

NVIDIA DEEP LEARNING SDK and CUDA

- cuDNN
- NCCL
- cuBLAS
- cuSPARSE
- TensorRT
- DeepStream SDK

developer.nvidia.com/deep-learning-software
COMMON SOFTWARE STACK ACROSS DGX FAMILY

Single, unified stack for deep learning frameworks
Predictable execution across platforms
Pervasive reach
NGC: GPU-OPTIMIZED SOFTWARE HUB
Simplifying DL, ML and HPC Workflows

- **Containers**: DL, ML, HPC
- **Model Training Scripts**: NLP, Image Classification, Object Detection & more
- **Helm Charts**: AI applications, K8s cluster, Registry
- **Pre-trained Models**: NLP, Classification, Object Detection & more
- **Industry Workflows**: Medical Imaging, Intelligent Video Analytics

- **Simplify Deployments**
- **Innovate Faster**
- **Deploy Anywhere**
CONTINUOUS PERFORMANCE IMPROVEMENT

Developers’ Software Optimizations Deliver Better Performance on the Same Hardware

Monthly DL Framework Updates & HPC Software Stack Optimizations Drive Performance

256 Batch Size | ResNet-50 Training v1.5 | 16x V100 | DGX-2

512 Batch Size | ResNet-50 Training v1.5 | 16x V100 | DGX-2

512 Batch Size | ResNet-50 Training v1.5 | 16x V100 | DGX-2

Speedup across Chroma, GROMACS, LAMMPS, QE, MILC, VASP, SPECFEM3D, NAMD, AMBER, GTC, RTM | 4x V100 v. Dual-Skylake | CUDA 9 for Mar '18 & Nov '18, CUDA 10 for Mar '19
THE IMPACT OF DGX SYSTEMS ON TIMELINE

Wasted time/effort - eliminated

Day 1

Month 3

2. Deploying an Integrated, Full-Stack AI Solution using DGX Systems
2. Deploying an Integrated, Full-Stack AI Solution using DGX Systems
NVIDIA DGX SUPERPOD
Instant AI Compute Infrastructure

Highest-Performance AI Research Supercomputer
  – #20 on TOP500 list | Top AI Performance Records

Instant AI Infrastructure
  – Modular and scalable architecture
  – Integrated and optimized compute, networking, storage, and software

NVIDIA-Optimized Software Stacks
  – Freely available on NGC

November 2019 TOP500 list: https://www.top500.org/lists/2019/11/
UP TO 80% MORE PERFORMANCE ON SAME SERVER

Software Innovation Delivers Continuous Improvements

MLPerf On DGX-2 Server
7 Month Improvement

Comparing the throughput of a single DGX-2H server on a single epoch (single pass of the dataset through the neural network) | MLPerf ID 0.5/0.6 comparison: ResNet50 v1.5: 0.5-20/0.6-30 | Transformer: 0.5-21/0.6-20 | SSD: 0.5-21/0.6-20 | GNMT: 0.5-19/0.6-20 | Mask R-CNN: 0.5-21/0.6-20
THE VALUE OF AI INFRASTRUCTURE WITH DGX REFERENCE ARCHITECTURES

SCALABLE PERFORMANCE
Reference architectures from NVIDIA and leading storage partners

FASTER, SIMPLIFIED DEPLOYMENT
Simplified, validated, converged infrastructure offers

TRUSTED EXPERTISE AND SUPPORT
Available through select NPN partners as a turnkey solution
TRADITIONAL HYPERSCALE CLUSTER

300 Dual-CPU Servers
$3M
180 kW
NVIDIA DGX-2 FOR DEEP LEARNING

1 DGX-2
Big Savings for Deep Learning
10 kW

1/8 the Cost
1/60 the Space
1/18 the Power
DGX POD MANAGEMENT SOFTWARE (AKA DEEPOPS)
For Large-Scale Multi-User AI Software Development Teams
DGX SYSTEMS — THE VALUE OF INTEGRATED SOLUTIONS AND AI EXPERTISE

Value to IT and users

Value to IT and users

GPUs

NVLink Server & GPUs

NVLink Server & GPUs DIY stack

DIY stack

From Design to Support: End-to-End AI Expertise

NVIDIA AI Experts

NGC DL SW

NVLink

NGC DL SW

NVLink

NVLink

GPU

DGX Systems
In business, ensuring equipment uptime and meeting safety and regulatory compliance is non-negotiable. Using deep neural networks developed on NVIDIA DGX-1 in the data center that can easily extend to NVIDIA DGX Station in the field, Avitas Systems delivers inspection services using robotic-based autonomous inspection and advanced data analytics.

In addition to safeguarding workers, Avitas Systems AI solutions can reduce inspection costs by 25% and reduce maintenance downtime by 15%.
IMPROVING DEMAND FORECASTS

With >100,000 different products in its 4,700 U.S. stores, the Walmart Labs data science team predicts demand for 500 million item-by-store combinations every week.

By performing forecasting with the open-source RAPIDS data processing and machine learning libraries built on CUDA-X AI on NVIDIA GPUs, Walmart speeds up feature engineering 100x and trains machine learning algorithms 20x faster, resulting in faster delivery of products, real-time reaction to shopper trends, and inventory cost savings at scale.
Delivering impeccable quality is a great opportunity for high precision manufacturers to differentiate but raises the bar for accurately detecting the smallest micron-scale product defects.

Foxconn Interconnect Technology Group (FIT) is seizing the opportunity. Its new AI-based inspection systems, powered by NVIDIA HGX-1, Tesla V100/P4, and Jetson TX2, reach greater than 95% accuracy, approaching the level of quality and efficiency of professional human inspection.
READY TO GET STARTED?

Project Checklist

What problem are you solving, what are the AI/DL tasks?

What data do you have/need, how is it labeled?

Which tools & environment will you use?

On what platform(s) will you train and deploy?
Gabriel Noaje  
Senior Solutions Architect  
NVIDIA APAC South  
gnoaje@nvidia.com
Fujitsu Leasing Options

Paul Floro
Senior Sales Manager
National Pre Sales
Product offering

Fully Integrated Appliance and Ready to run DGX Systems

Data Centre and Edge GPUs (Titan, RTX, M10, Tesla V100, Tesla T4, Quadro, Xavier)

OEM Systems

VDI Systems (Fujitsu, HPe, Lenovo)

Data Centre Systems - NVLink (Fujitsu, HPe)

Data Centre Systems (Fujitsu, HPe, Lenovo)

Edge and Workstations (Fujitsu, Lenovo)
Fujitsu Australia Leasing Options

- Avoids up front Capital Expenditure
- Aligns costs with usage over time
- Smooth Cash Flow with regular payments
- Lease terms from 24 to 60 months

- For transactions > $60K
- Subject to Credit Approval and Lease documentation
## Rental ("Operating Lease")

<table>
<thead>
<tr>
<th>Term</th>
<th>Committed Initial Term (24, 36, 48 or 60 Months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who owns the equipment?</td>
<td>The Lessor; the client manages equipment during lease term.</td>
</tr>
<tr>
<td>Effect on cashflow</td>
<td>Smooth, fixed payments over useful working life.</td>
</tr>
<tr>
<td>Effect on TCO</td>
<td>Cost of <strong>useful working life + interest</strong> Reduced rental cost from equity investment</td>
</tr>
<tr>
<td>Who bears depreciation &amp; obsolescence costs?</td>
<td>The Lessor</td>
</tr>
<tr>
<td>Capex or Opex budget</td>
<td>Opex</td>
</tr>
<tr>
<td>Impact on asset refresh strategy</td>
<td>Supportive – planned refresh cycle</td>
</tr>
<tr>
<td>Ability to upgrade Infrastructure?</td>
<td>Flexible - simple upgrade process at any point during lease term</td>
</tr>
<tr>
<td>Who manages assets?</td>
<td>The client</td>
</tr>
</tbody>
</table>
| End of lease options          | • Continue to rent  
• Upgrade  
• Return the equipment  
• Offer to purchase |
Fujitsu

shaping tomorrow with you