FOR AI, HARDWARE IS JUST THE BEGINNING

Ian Buck
September 18, 2019
SAFE HARBOR

Forward-Looking Statements

Except for the historical information contained herein, certain matters in this presentation including, but not limited to, statements as to: our strategies, growth, and market opportunities; future financial results, estimates and forecasts; scaling, shifting, growing, accelerating and democratizing artificial intelligence and its use cases; the performance, impact, abilities and benefits of our products and technologies and artificial intelligence; and other predictions and estimates are forward-looking statements within the meaning of the Private Securities Litigation Reform Act of 1995. These forward-looking statements and any other forward-looking statements that go beyond historical facts that are made in this presentation are subject to risks and uncertainties that may cause actual results to differ materially. These forward-looking statements and any other forward-looking statements that go beyond historical facts that are made in this presentation are subject to risks and uncertainties that may cause actual results to differ materially. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing products and technologies; market acceptance of our products or our partners’ products; design, manufacturing or software defects; changes in consumer preferences and demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.
SUPERCHARGED COMPUTING FOR THE DA VINCIS AND EINSTEINS OF OUR TIME

GPU computing serves the most demanding users — scientists, designers, artists, gamers

Fueled by insatiable demand for better AI, HPC and graphics.

GPU: Brain of VR, HPC, AI
AI INNOVATION IS SHIFTING, AND GROWING

Next-Level Use-Cases Require Gigantic Models

Number of Parameters by Network

- **Image Recognition**
  - 26M
- **NLP**
  - 340M
  - Q&A
  - Sentiment
  - Translation
  - Generative Tasks
    - Chatbots
    - Email auto-completion
    - Document summarization
- **Autonomous Vehicles**
- **Social Tagging**
- **Visual Search**
- **Q&A**
- **Sentiment**
- **Translation**
- **1.5Bn**

**Project Megatron**
- 8.3B parameters
- 8-way Model Parallel
- 64-way Data Parallel
- 24x larger than BERT

https://github.com/NVIDIA/Megatron-LM
AI TRAINING: A SUPERCOMPUTING CHALLENGE

Delivering a Time Machine to AI Researchers

<table>
<thead>
<tr>
<th>Year</th>
<th>Platform</th>
<th>Time Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>K80</td>
<td>36,000 Mins (25 Days)</td>
</tr>
<tr>
<td>2017</td>
<td>NVIDIA® DGX-1™ Volta</td>
<td>480 Mins (8 Hours)</td>
</tr>
<tr>
<td>2019</td>
<td>NVIDIA® DGX SuperPOD™ NVIDIA® NVSwitch™ Mellanox® InfiniBand</td>
<td>8 HRS TO 80 SECS</td>
</tr>
</tbody>
</table>

- 80 Secs (1.33 Mins) (ResNet-50, Image Classification)
- 1.59 Mins (Transformer, Non-Recruent Translation)
- 1.8 Mins (GNMT, Recurrent Translation)
- 2.23 Mins (SSD, Lightweight Object Detection)
- 13.57 Mins (Reinforcement Learning, MiniGo)
- 18.47 Mins (Mask R-CNN, Heavyweight Object Detection)
SCALING POSES TOUGH CHALLENGES

Data Center as “The Computer”

Driver
Ecosystem Support
Mem BW
System Interconnect

Multi-GPU
CPU/GPU Ratio
Thermals
Chassis

NVSwitch
Multi-NIC
Scaling SW

Network Topology
AI-Optimized FS
InfiniBand
NUMA
App-Pinning
Kernel Tuning
Scaling SW
EXAMPLE OPTIMIZATION: DALI

Easing Training Performance Bottleneck

NVIDIA DATA LOADING LIBRARY (DALI)

Training Data → Decode → Resize, Color, Augment → Ready-to-Train Data

DALI Training Speedups:
- DGX-1: 1.4x
- DGX-2: 1.9x

MXNet 19.02 NGC Container, batch size: 256
SCALING AI TO THOUSANDS OF GPUS

Key Enabling Technology: NCCL 2.4

<table>
<thead>
<tr>
<th>NCCL</th>
<th>Previous Version</th>
<th>New Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>Ring</td>
<td>Double Binary Tree</td>
</tr>
<tr>
<td>Latency</td>
<td>Linear</td>
<td>Logarithmic</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>Full</td>
<td>Full</td>
</tr>
<tr>
<td>Scale (#GPUs)</td>
<td>100’s</td>
<td>1,000’s</td>
</tr>
</tbody>
</table>

Automatically Switches Between Tree and Ring Algorithms

Open-Source – available on GitHub now

Testing done on ORNL’s Summit Supercomputer, scaling up to 24,576 GPUs
AI INFRASTRUCTURE DEFINING AI’S FUTURE

1.7 Tensor ExaFLOPS | 200+ AI Researchers | 170+ AI Papers
NVIDIA MEGATRON-LM

- World’s largest Transformer-based language model (8.3 billion parameters)
  - 24x the size of BERT (345M parameters)
  - 5.6x the size of GPT-2 (1.5B parameter)

- Achieved 15.1 PetaFLOPs per second sustained performance over the entire application using 512 GPUs at 76% scaling efficiency

- 12 ZettaFLOPs to converge in 9.2 days

- SOTA for Lambada accuracy (66.5% compared to 63.2%) and Wikitext-103 perplexity (10.81 compared to 16.4)

Figure from Huggingface DistilBERT blog post (https://medium.com/huggingface/distilbert-8cf3380435b5)

https://github.com/NVIDIA/Megatron-LM
VAST WORLD OF AI INFERENCE

Creating A Massive Market Opportunity

General Purpose Computers

Embedded Computers

Embedded Devices
MODERN USE-CASES: REAL TIME & MULTI-MODEL
Enabling AI Conversational with Performance and Versatility

What are different types of lighting for a living room?

Ambient, task and accent

20-30 containers end-to-end | RNN, CNN, MLP in INT8, FP16, FP32 | Latency <300ms
CRITICAL COMPONENT: ACCURATE NLP
Achieving Superhuman Accuracy in Real Time

TensorRT 6
Now Available

- API for dynamic input shapes
- Dynamic input batch size
- New layers for 3D convolutions
- Optimizations for 2D U-Net

Automatic Speech Recognition
NLP
Text to Speech

![Graph showing latency comparison between BERT-Base and BERT-Large on NVIDIA T4]

- BERT-Base: 2.1 ms
- BERT-Large: 5.8 ms

Automatic Speech Recognition
Text to Speech

NVIDIA T4
INFERENCNE NEEDS A VERSATILE PLATFORM
To Accelerate Images, Speech, Recommenders, and…
CUSTOMERS SEEING REAL RESULTS
From Fraud Detection to Speech Recognition to Recommender Systems
GPU INFERENCE IS DEPLOYED TODAY

Services

- Speech Recognition
- Natural Language Processing
- Text to Speech
- Recommender Systems
- Object Detection
- Translation

Usages

- Services
- Usages
- T4 Now Available
THE NEXT AI FRONTIER: INDUSTRIAL ADOPTION

Requires Verticalized AI Platforms

ROBOTICS
SMART CITY
AUTONOMOUS VEHICLES

HEALTHCARE
MANUFACTURING
3D RENDERING
THE NEXT AI FRONTIER: INDUSTRIAL ADOPTION
Requires Verticalized AI Platforms

ROBOTICS → ISAAC
SMART CITY → METROPOLIS
AUTONOMOUS VEHICLES → DRIVE

HEALTHCARE → CLARA
MANUFACTURING → JETSON
3D RENDERING → OMNIVERSE
NVIDIA DRIVE

Vertical Platform for Autonomous Vehicles
WORLD OF RADIOLOGY
Millions of Instruments + Thousands of AIs

Acquire → Image Processing → Image Analysis → Read & Report
AI-ASSISTED ANNOTATION
APIs to Plug into Any Existing Medical Viewer

Auto Segmentation: Allows for 1 click organ segmentation
Interactive Annotation: A user can apply Auto-segmentation and then correct the extreme point in interactive mode.

Annotation Client APIs hosted on GitHub and Annotation Server is part of Clara Train Docker container

![AI-Assisted Annotation Speedup](image-url)
CLARA AI TOOLKIT

Unlabeled Data → Annotated Data → Training → Transfer Learning → Clara Application

- AI-Assisted Annotation
  - Pre-Trained Models
  - Annotation Client APIs
  - Annotation Server

- Training
  - Pre-Trained Models
  - Transfer Learning Toolkit

- Clara Train SDK
  - Containers and built-in support for DICOM communication
  - Services for workflow inference and orchestration
  - Pipeline Definition Language to define custom workflow
  - Tools that simplify the setup of Kubernetes clusters

- Clara Deploy SDK
CLARA TECHNOLOGY STACK

Toolkit for Medical Imaging

Clara Train SDK

- SAMPLE TRAINING PIPELINES
- DICOM 2 NIFTI
- AI-ASSISTED ANNOTATION
- TRANSFER LEARNING
- TRAIN
- PRE-TRAINED MODELS

Clara Deploy SDK

- DICOM ADAPTER
- SAMPLE DEPLOYMENT PIPELINES
- PIPELINE MANAGER
- DATA MEMORY OPTIMIZATION
- AI INFERENCE
- STREAMING RENDER

KUBERNETES

CUDA

LINUX

SOFTWARE STACK

CUDA X LIBS

HARDWARE

CUDA

- cuBLAS
- cuFFT
- NPP

COMPUTE

- cuDNN
- DALI
- TRT

ARTIFICIAL INTELLIGENCE

- OPTIX
- INDEX
- NVENC

VISUALIZATION

- TESLA GPU
- NVIDIA DGX FAMILY
- SYSTEM OEM
- CLOUD

Toolkit for Medical Imaging
FOR AI, HARDWARE IS JUST THE BEGINNING

Accelerating AI Acceleration

AI Training is a Supercomputing Challenge

AI Inference is Diverse And “Part” of a Broader Application

Industry Adoption Needs Verticalized Platforms