THE AI COMPUTING COMPANY

Romuald Josien
NVIDIA — A LEARNING MACHINE

NVIDIA has continuously reinvented itself over more than two decades.

Our invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics, and revolutionized parallel computing. More recently, GPU computing ignited the era of AI.

NVIDIA is a “learning machine” that constantly evolves by adapting to new opportunities that are hard to solve, that only we can tackle, and that matter to the world.
For 30 years, the dynamics of Moore’s law held true. But now CPU scaling is slowing while the demand for computing power surges ahead.

With AI, machines can learn. AI can solve grand challenges that have been beyond human reach. But it must be fueled by massive compute power.

Accelerated computing is the path forward beyond Moore’s law, delivering 1,000X computing performance every 10 years.
NVIDIA is an accelerated computing company. It starts with a highly specialized parallel processor called the GPU and continues through system design, system software, algorithms, and optimized applications.

CUDA-X® is a suite of software libraries that accelerate applications for our growth markets — from gaming to transportation to healthcare — all based on a common CUDA architecture supported by more than 1.2 million developers today.
A RACE FOR PERFORMANCE
EXPONENTIAL GROWTH IN COMPUTING DEMAND

DATA SIZE GROWING

Zettabytes Generated

2012 2018 2025
0 40 80 120 160

160 ZB

AI RESEARCH GROWING

Papers Submitted to NeurIPS & CVPR

0 3K 6K 9K 12K

AI MODEL COMPLEXITY GROWING

Petaflop-Days (Training)

0.001 0.01 0.1 1 10 100 1,000

Source: IDC, GitHub, and OpenAI / NVIDIA
NETWORK COMPLEXITY IS EXPLODING

Petaflop/s-day (Training)

Source: OpenAI and NVIDIA
A CAMBRIAN EXPLOSION OF DL MODELS

CONVOLUTIONAL NETWORKS

RECURRENT NETWORKS

GENERATIVE ADVERSARIAL NETWORKS

REINFORCEMENT LEARNING

NEW SPECIES
AI INNOVATION IS SHIFTING, AND GROWING
Next-Level Use-Cases Require Gigantic Models

Number of Parameters by Network

Image Recognition
- Autonomous Vehicles
- Social Tagging
- Visual Search

NLP
- Q&A
- Sentiment
- Translation

ResNet-50: 26M
Transformer: 340M
GPT-1
BERT LARGE
GPT-2
GPT-2 8B
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 LEADERSHIP STARTS WITH AI COMPUTING LEADERSHIP

Researchers racing to advance AI for the world’s largest industries - auto, healthcare, manufacturing

Increasingly complex AI models and larger data size demand powerful computers

Iteration speed and time-to-train fuels innovation

NVIDIA created DGX SuperPOD to serve as the essential instrument of AI research
### NVIDIA BREAKS RECORDS IN AI PERFORMANCE
Both On 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

INDUSTRY WIDE BENCHMARK SUITE FOR AI PERFORMANCE

[https://mlperf.org/](https://mlperf.org/)
NVIDIA DGX-2
The World’s Most Powerful AI Computer

- 2 PFLOPS
- 512GB HBM2
- 10kW
- 350 lbs

- 16x Tesla V100 32GB
- 12x NVSwitch
- NVLink Plane Card
- 10x EDR IB/100 GigE
- 2x Xeon Platinum
- 1.5TB System Memory
- PCIe Switch Complex
- 30TB NVME SSDs
TIME MACHINE FOR AI
Smashing Time to Train From 8 Hours to 80 Seconds On V100

2015
K80 | CUDA®
36,000 Mins (25 Days)

2017
NVIDIA® DGX-1™ | Volta | Tensor Cores
480 Mins (8 Hours)

2019
NVIDIA DGX SuperPOD™ | NVIDIA NVSwitch™ | Mellanox InfiniBand
8 HRS TO 80 SECS

2019 MLPerf ID (in order from top to bottom of chart): ResNet-50: 0.6-30 | Transformer: 0.6-28 | GNMT: 0.6-14 | SSD: 0.6-27 | Mini-Go: 0.6-11 | Mask R-CNN: 0.6-23
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
UP TO 80% MORE PERFORMANCE ON SAME SERVER

Software Innovation Delivers Continuous MLPerf Improvements

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
HOW TO REACH THIS PERFORMANCE
MIXED PRECISION ACCELERATION AND TENSOR CORES
TENSOR CORE GPU FUSES HPC & AI COMPUTING

VOLTA TENSOR CORE GPU

MULTI-PRECISION COMPUTING

FUSION OF HPC & AI

HPC (Simulation) - FP64, FP32

AI (Deep Learning) - FP16, INT8
Mixed Precision Accelerator - Delivering Up To 5X Throughput of FP32

Mixed Precision Accelerator - Delivering Up To 5X Throughput of FP32

Memory Savings
- Half Storage Requirements (larger batch size)
- Half the memory traffic by reducing size of gradient/activation tensors

FP16 Reduced Precision
Lower Latency
Range: +/- 65,504

FP32
Higher Precision
Range: +/- 3.402823x10^{38}

4x4 Matrix
16 FP16 values

5X Throughput of FP32

4x4 Product and Accumulate
FP32 = FP16 x FP16 + FP32

D = A*B+C

1Fastest Tensor Core Speedup by Facebook on NMT (Arxiv paper Sep 2018)
TENSOR CORE AUTOMATIC MIXED PRECISION

3x Speedup With Just One Line of Code

**Training Speedup Over 3X**

**Inference Speedup Over 4X**
MIXED PRECISION MAINTAINS ACCURACY
Benefit From Higher Throughput Without Compromise

Mixed Precision - Same hyperparameters and learning rate schedule as FP32 ILSVRC12 classification top-1 accuracy.
(Sharan Narang, Paulius Micikevicius et al., "Mixed Precision Training", ICLR 2018)
ACTIVATING MIXED PRECISION WITH EASE

Use NVIDIA Optimized Models...

Computer Vision  Speech  Recommender  GANs

...Available From:

NVIDIA GPU Cloud
Docker Pull Containers

Tensor Core Journey Page

NVIDIA Deep Learning Examples

Github Repository

Or Accelerate Your Own Models

Webinar

Nvprof
Nsight Compute

Github

Automatic Mixed Precision
Just Add 2 Lines of Code

Profiler Tools
Individual Kernel Optimization

Use NVIDIA Optimized Models...

Computer Vision  Speech  Recommender  GANs

...Available From:

NVIDIA GPU Cloud
Docker Pull Containers

Tensor Core Journey Page

NVIDIA Deep Learning Examples

Github Repository

Or Accelerate Your Own Models

Webinar

Nvprof
Nsight Compute

Github

Automatic Mixed Precision
Just Add 2 Lines of Code

Profiler Tools
Individual Kernel Optimization
NGC
NGC: GPU-OPTIMIZED SOFTWARE HUB
Ready-to-run GPU Optimized Software, Anywhere

50+ Containers
DL, ML, HPC

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

60 Pre-trained Models
NLP, Image Classification, Object Detection & more

Industry Workflows
Medical Imaging, Intelligent Video Analytics

On-prem
Cloud
Hybrid Cloud
Multi-cloud
SIMPLIFYING APPLICATION DEPLOYMENTS

Driving Productivity and Faster Discoveries

Data Scientists & Developers

Superior Performance - Continuous optimizations
Pre-trained Models & Scripts - Speed up AI workflows
On-demand Software - Higher productivity
Scalable - on multi-GPU, multi-node systems
Run Anywhere - On-Prem, Cloud, Hybrid
Designed for Enterprise & HPC - Docker & Singularity

Sysadmins & DevOps
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, PC workstations, and servers
- From Core to the Edge
- On-Prem to Hybrid to Cloud
DALI
Eliminating CPU Bottleneck for DL Workflows

CPU Bottleneck Waste GPU Cycles
- Complex I/O pipelines
- Multi-pipeline frameworks
- Decreasing CPU:GPU ratio

DALI Shifts Workloads to GPUs
- Full input pipeline acceleration including data loading and augmentation
- Integrated in PyTorch, TF, MxNET
- Supports Resnet50 & SSD
GET STARTED WITH NGC
Explore the NGC Registry for DL, ML & HPC

Deploy containers: ngc.nvidia.com
Learn more about NGC offering: nvidia.com/ngc
Technical information: developer.nvidia.com
GPU ACCELERATED SERVER PLATFORMS
TESLA V100
TENSOR CORE GPU

World’s Most Powerful Data Center GPU

5,120 CUDA cores
640 NEW Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS
| 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32 GB HBM2 @ 900GB/s |
300GB/s NVLink
TESLA PLATFORM ENABLES DRAMATIC REDUCTION IN TIME TO TRAIN

ResNet-50 v1.5, 90 epochs to solution | CPU Server: dual socket Intel SKX 8168 | Refer to MLPerf 0.6 for convergence criteria
UP TO 50% PERFORMANCE IMPROVEMENT
32GB Benefits for AI and HPC

**FASTER RESULTS**

- 1.5X Faster Language Translation
- 1.5X Faster Calculations

- Neural Machine Translation (NMT)
- 3D FFT 1k x 1k x 1k

**HIGHER ACCURACY**

- 40% Lower Error Rate
- Accuracy (152 layers)

- Accuracy (16 layers)

The additional memory improved our ability to handle higher definition images on a larger ResNet-152 model, reducing error rates by 40 percent on average. This results in accurate, timely and auditable services at scale.

Michael Kemelmakher
Vice President
SAP Innovation Center, Israel

---

Dual E5-2698v4 server, 512GB DDR4, Ubuntu 16.04, CUDA9, cuDNN7. NMT is GNMT-like and run with TensorFlow NGC Container 18.01 (Batch Size= 128 (for 16GB) and 256 (for 32GB) | FFT is with cufftbench 1k x 1k x 1k and comparing 2 V100 16GB (DGX1V) vs. 2 V100 32GB (DGX1V)

R-CNN for object detection at 1080P with Caffe | V100 16GB uses VGG16 | V100 32GB uses Resnet-152
# HGX PLATFORMS

<table>
<thead>
<tr>
<th></th>
<th>HGX-1</th>
<th>HGX-2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topology</strong></td>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>1 petaFLOP tensor operations 125 teraFLOPS single-precision 62 teraFLOPS double-precision</td>
<td>2 petaFLOPS tensor operations 250 teraFLOPS single-precision 125 teraFLOPS double-precision</td>
</tr>
<tr>
<td><strong>GPUs</strong></td>
<td>8x NVIDIA Tesla V100</td>
<td>16x NVIDIA Tesla V100</td>
</tr>
<tr>
<td><strong>GPU Memory</strong></td>
<td>256GB total</td>
<td>512GB total</td>
</tr>
<tr>
<td><strong>Communication Channel</strong></td>
<td>Hybrid cube mesh powered by NVLink 300GB/s bisection bandwidth</td>
<td>NVSwitch powered by NVLink 2.4TB/s bisection bandwidth</td>
</tr>
</tbody>
</table>
NVLINK AND MULTI-GPU SCALING
For Data Parallel Training

- Data loading over PCIe
- Gradient averaging over PCIe and QPI
- Data loading and gradient averaging share communication resources: Congestion

- Data loading over PCIe (red)
- Gradient averaging over NVLink (green)
- No sharing of communication resources: No congestion
NVSWITCH
World’s Highest Bandwidth
On-node Switch

7.2 Terabits/sec or 900 GB/sec
18 NVLINK ports | 50GB/s per port bi-directional
Fully-connected crossbar
2 billion transistors | 47.5mm x 47.5mm package
SCALING-UP PERFORMANCE WITH NVSWITCH

![Graph showing performance scaling with NVSwitch vs. PCIe for V100 GPUs. The graph plots tokens/second against the number of V100 GPUs. The green line represents V100 with NVLink and NVSwitch, while the gray line represents V100 with PCIe.](Image)
THE TESLA BRAND PROMISE

Backed by NVIDIA Products, People, & Processes

- **Shorten Time-to-Insight**
  - NVLink, NCCL, NVSwitch
  - GPUDirect RDMA

- **Run Large Jobs With Confidence**
  - ECC, DPR, Testing & Qualification
  - Sys Monitoring and Management

- **Complex Stacks Just Work**
  - Direct Technical Support for Developers
    - GPU Optimized Apps
    - NGC
    - Top Apps
  - App Testing Infrastructure

- **Enterprise Grade Reliability**
  - Access to NVIDIA Experts Defined Escalation Paths
  - Reliable Software for Mission Critical Applications
  - NVIDIA NVONLINE Proactive Software Bug Fix Process

- **Available Everywhere**
  - Every OEM and Cloud Provider
    - Alibaba Cloud
    - Microsoft Azure
    - Google Cloud
    - Bull
    - CISCO
    - Dell
    - IBM
    - Inspur
    - Sugon
    - 240+ Resellers Worldwide

- **Certainty of Supply**
  - Purchase unlimited quantities
  - 3-year warranty
  - 10-month advanced EOL notification
  - 3-year SKU life

**NGC Top Apps**

**THE TESLA BRAND PROMISE**

Backed by NVIDIA Products, People, & Processes

- **Improved throughput**
  - Lower TCO
  - Increased productivity

- **Lower TCO**
  - Increased throughput
  - Deploy with confidence

- **Maximum Uptime**
  - Lower TCO
  - Choice of supplier
  - Lower acquisition costs
  - Avoid supply shortages

- **Increased productivity**
  - Available Everywhere
  - Certainty of Supply
SHORTEN TIME TO INSIGHT
Up to 2.5X Faster with NVLink and GPUDirect RDMA

Application Performance Within a Node

<table>
<thead>
<tr>
<th>Application</th>
<th>Without NV Link</th>
<th>NVLink</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUDA</td>
<td>1.2x</td>
<td>2.3x</td>
<td>1.83</td>
</tr>
<tr>
<td>PyTorch ResNet50</td>
<td>1.3x</td>
<td>1.5x</td>
<td>1.23</td>
</tr>
<tr>
<td>HOOMD-Blue</td>
<td>1.5x</td>
<td>1.6x</td>
<td>1.07</td>
</tr>
<tr>
<td>NAMD</td>
<td>1.6x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMMPS</td>
<td>2.3x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MILC Performance At Scale

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Time to Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.5X reduced TTS</td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

NVLink: GPU Servers: Dual Xeon Gold 6140@2.30GHz or E5-2698 v4@3.6GHz for PyTorch with 8 V100 PCIe vs V100 NVLink
GPUDirect RDMA: HW: DGX-1 with 8xV100 SXM2, SW: MILC (APEX large benchmark). This is measured on DGX-1 systems.
RUNNING LARGE JOBS WITH CONFIDENCE
Enterprise Reliability, Management and Live Migration

**ERROR CORRECTION CODE (ECC)**
- GPU MEMORY
- Uncorrectable Data Error causes application to crash

**DYNAMIC PAGE RETIREMENT (DPR)**
- Tesla GPU with DPR
- Weak memory page is retired

**SYSTEM QUALIFICATION AND TESTING**
- Double-bit error mitigation
- Removes suspect memory locations with simple reset
- No physical work required for IT
- <0.01% of memory is retired
- Long burn-in testing
- Zero error tolerance at aggressive clocks
- Large guard-band for guaranteed quality
- 5% of GPUs are screened out

**SYSTEM MONITORING AND MANAGEMENT**
- Active health monitoring
- Diagnostics and system validation
- Policy and group config management
- Power and clock management

**LIVE MIGRATION FOR NVIDIA VGPUS**
- Keep large jobs running during patches and upgrades
- Maximize infrastructure investment
DRAMATICALLY MORE FOR YOUR MONEY

CPU-Only Cluster

300 Self-hosted Broadwell CPU Servers
180 KWatts

Deep Learning Training
Image training
Resnet 50

GPU-Accelerated

1 DGX-2
10 KWatts

SAME THROUGHPUT

1/8 THE COST
1/18 THE POWER
1/30 THE SPACE

Deep Learning Training
Image training
Resnet 50
DEEP LEARNING INSTITUTE
University Ambassador Program

Preparing today’s students and researchers for tomorrow’s AI computing challenges

Want to bring DLI to your campus?

DLI can award qualified academics as certified DLI Ambassadors, enabling them to bring ready-made, free DLI content exclusively to university students and staff.

DLI University Ambassadorship is an additional status on top of DLI Instructor Certification with additional benefits.

Candidates should have relevant teaching and research experience, and can apply [here](#) for an invitation to an on-site instructor certification event.
GET STARTED WITH FUNDAMENTALS

New to deep learning or accelerated computing?

Fundamentals training is the place to start. Content is designed for a technical audience of developers, researchers, and data scientists.