A NEW ERA 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
HOW A DEEP NEURAL NETWORK SEES

GPU DEEP LEARNING IS A NEW COMPUTING MODEL

Billions of Trillions of Operations
GPU train larger models, accelerate time to market
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

10s of billions of image, voice, video queries per day
GPU inference for fast response, maximize datacenter throughput

DATACENTER INFERENCING
NEURAL NETWORK COMPLEXITY IS EXPLODING
To Tackle Increasingly Complex Challenges

2015 - Microsoft ResNet
Superhuman Image Recognition

2016 - Baidu Deep Speech 2
Superhuman Voice Recognition

2017 - Google Neural Machine Translation
Near Human Language Translation
GEFORCE NOW
You listen to music on Spotify.
You watch movies on Netflix.
GeForce Now lets you play games the same way.
Instantly stream the latest titles from our powerful cloud-gaming supercomputers. Think of it as your game console in the sky.

ROAD TO EXASCALE
Volta to Fuel Most Powerful US Supercomputers

Summit Supercomputer
200+ PetaFlops
~3,400 Nodes
10 Megawatts

1.8X HPC Performance in 1 Year

V100 Performance Normalized to P100

<table>
<thead>
<tr>
<th>Category</th>
<th>Normalized Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>cuFFT</td>
<td>1.8</td>
</tr>
<tr>
<td>Physics (QUDA)</td>
<td>1.5</td>
</tr>
<tr>
<td>Seismic (RTM)</td>
<td>1.6</td>
</tr>
<tr>
<td>STREAM</td>
<td>1.5</td>
</tr>
</tbody>
</table>

System Config Info: 2X Xeon E5-2690 v4, 2.6GHz, w/ 1X Tesla P100 or V100. V100 measured on pre-production hardware.
NVIDIA SATURN V

124 DGX-1 Deep Learning Supercomputers
TESLA V100
THE MOST ADVANCED DATA CENTER GPU EVER BUILT

5,120 CUDA cores
640 NEW Tensor cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
120 Tensor TFLOPS
20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s
300 GB/s NVLink
NEW TENSOR CORE BUILT FOR AI
Delivering 120 TFLOPS of DL Performance

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

MATRIX DATA OPTIMIZATION:
Dense Matrix of Tensor Compute

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

VOLTA-OPTIMIZED cuDNN

ALL MAJOR FRAMEWORKS

Caffe2
mxnet
PyTorch
TensorFlow
REVOLUTIONARY AI PERFORMANCE

3X Faster DL Training Performance

Googlenet Training Performance (Speedup Vs K80)

- Over 80x DL Training Performance in 3 Years

LSTM Training (Neural Machine Translation)

- 3X Reduction in Time to Train Over P100

Multi-Node Training with NCCL2.0 (ResNet-50)

- 85% Scale-Out Efficiency
- Scales to 64 GPUs with Microsoft Cognitive Toolkit

ResNet50 Training for 10 Epochs with 1.28M images dataset | Using Caffe2 | V100 performance measured on pre-production hardware.
VOLTA DELIVERS 3X MORE INFRINGEMENT THROUGHPUT

Low Latency performance with V100 and TensorRT

3x more throughput at 7ms latency with V100 (ResNet-50)

Throughput @ 7ms (Images/Sec)

- 33ms
- 10ms
- 7ms
- 7ms

Trained Neural Network

TensorRT
Fuse Layers
Compact
Optimize Precision
(FP32, FP16, INT8)

Compiled Real-time Network

CPU Server: 2X Xeon E5-2660 V4; GPU: w/P100, w/V100 (@150W) | V100 performance measured on pre-production hardware.
NVIDIA TENSORRT 3
World’s Fastest Inference Platform

ResNet-50 Throughput @ 7ms Latency

Throughput @ 7ms (Images/Sec)

- CPU-only: 14ms
- Tesla V100 TensorFlow
- Tesla P4 TRT 3 (INT8)
- Tesla V100 TRT 3 (FP16)

40X

Workload: ResNet-50
Dataset: ImageNet
CPU: Skylake
GPU: Tesla P4 or Tesla V100