AI REVOLUTIONIZING OUR WORLD

- Search, Assistants, Translation, Recommendations, Shopping, Photos...
- Detect, Diagnose and Treat Diseases
- Powering Breakthroughs in Agriculture, Manufacturing, EDA
NEURAL NETWORK COMPLEXITY IS EXPLODING
To Tackle Increasingly Complex Challenges

- **2015 - Microsoft ResNet**
  - Superhuman Image Recognition
  - 7 ExaFLOPS
  - 60 Million Parameters

- **2016 - Baidu Deep Speech 2**
  - Superhuman Voice Recognition
  - 20 ExaFLOPS
  - 300 Million Parameters

- **2017 - Google Neural Machine Translation**
  - Near Human Language Translation
  - 100 ExaFLOPS
  - 8700 Million Parameters
EXPONENTIAL DEMAND FOR COMPUTE CYCLES

Growing Usage of Deep Learning

Internet Services

Growing Use of Deep Learning at Google

P2 instance is one of the fastest growing instance in AWS history.”

Andrew Jassy, AWS CEO

Cloud Services

Growing Usage of GPU Cloud

Supercomputing

Growth In Total FLOPS

Google

Growing Usage of Deep Learning

Amazon Web Services

TOP 500

Growth In Total FLOPS
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

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
Microsoft Cognitive Toolkit
mxnet
PyTorch
TensorFlow
3X Faster DL Training Performance

- Over 80x DL Training Performance in 3 Years
- 85% Scale-Out Efficiency Scales to 64 GPUs with Microsoft Cognitive Toolkit
- 3X Reduction in Time to Train Over P100

**Googlenet Training Performance (Speedup Vs K80)**

- 1x K80 cuDNN2
- 4x M40 cuDNN3
- 8x P100 cuDNN6
- 8x V100 cuDNN7

**LSTM Training (Neural Machine Translation)**

- 1X V100
- 1X P100
- 2X CPU

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

- 8X P100
- 8X V100
- 64X V100

**Performance measured on pre-production hardware.**
VOLTA DELIVERS 3X MORE INFERANCE THROUGHPUT

Low Latency performance with V100 and TensorRT

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

Throughput @ 7ms (Images/Sec)

- CPU: 33ms
- Tesla P100 (TensorFlow): 10ms
- Tesla P100 (TensorRT): 7ms
- Tesla V100 (TensorRT): 7ms

CPU Server: 2X Xeon E5-2660 V4; GPU: w/P100, w/V100 (@150W) | V100 performance measured on pre-production hardware.
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.

Gaming is now easy and instant.

ROAD TO EXASCALE
Volta to Fuel Most Powerful US Supercomputers

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

1.5X HPC Performance in 1 Year

System Config Info: 2X Xeon E5-2690 v4, 2.6GHz, w/ 1X Tesla P100 or V100. V100 measured on pre-production hardware.
**SINGLE UNIVERSAL GPU FOR ALL ACCELERATED WORKLOADS**

<table>
<thead>
<tr>
<th>Workload</th>
<th>Boost Factor</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPC</td>
<td>1.5X</td>
<td>Vs P100</td>
</tr>
<tr>
<td>AI Training</td>
<td>3X</td>
<td>Vs P100</td>
</tr>
<tr>
<td>AI Inference</td>
<td>3X</td>
<td>Vs P100</td>
</tr>
<tr>
<td>Virtual Desktop</td>
<td>2X</td>
<td>Vs M60</td>
</tr>
</tbody>
</table>

**V100 UNIVERSAL GPU**
OPTIMIZED FOR DATACENTER EFFICIENCY

40% More Performance in a Rack

80% Perf at Half the Power

ResNet-50 Networks Trained Per Day

13 KW Rack
4 Nodes of 8xV100
13 ResNet-50 Networks Trained Per Day

13 KW Rack
7 Nodes of 8xV100
18 ResNet-50 Networks Trained Per Day

ResNet-50 Training, Max Efficiency run with V100@160W | V100 performance measured on pre-production hardware.
## TESLA V100 SPECIFICATIONS

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compute</strong></td>
<td>7.5 TF DP • 15 TF SP • 120 TF DL</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>HBM2: 900 GB/s • 16 GB</td>
</tr>
<tr>
<td><strong>Interconnect</strong></td>
<td>NVLink (up to 300 GB/s) + PCIe Gen3 (up to 32 GB/s)</td>
</tr>
<tr>
<td><strong>Availability</strong></td>
<td>DGX-1: Q3 2017</td>
</tr>
<tr>
<td></td>
<td>OEM : Q4 2017</td>
</tr>
</tbody>
</table>