ACCELERATING SMART CITIES WITH GPU INFRASTRUCTURE

Dr. Leo K. Tam
MEGATRENDS ARE DRIVING WORLD CITIES

URBANIZATION

↑ 2.5Bn
Urban pop. growth
United Nations DESA

DIGITIZATION

50Bn
Connected things by 2020
Cisco

INDUSTRIALIZATION

↑ 50%
Energy consumed
IESA
World class cities deserve world class infrastructure

Urbanization

- Urban pop. growth
- United Nations, DESA

- 2.5Bn
- Connected things by 2020
- Cisco

Digitization

- 50Bn
- Energy consumed
- IESA

Industrialization

- 50%

- Energy consumed
- IESA
TEN YEARS OF GPU COMPUTING

- 2006: CUDA Launched
- 2008: World’s First GPU Top500 System
- 2010: Stanford Builds AI Machine using GPUs
- 2012: Discovered How H1N1 Mutates to Resist Drugs
- 2012: AlexNet beats expert code by huge margin using GPUs
- 2014: World’s First 3-D Mapping of Human Genome
- 2014: Stanford Builds AI Machine using GPUs
- 2016: World’s First Atomic Model of HIV Capsid
- 2016: Google Outperforms Humans in ImageNet
- 2016: GPU-Trained AI Machine Beats World Champion in Go
HARDWARE AND DATA DRIVES DEEP LEARNING

facebook
350 millions images uploaded per day
2.5 Petabytes of customer data hourly

Walmart

You Tube
300 hours of video uploaded every minute

Image

“Volvo XC90”
MOST PERVERSIVE HPC PLATFORM EVER BUILT

ACCESS ANYWHERE

BUY ANYWHERE

LEARN EVERYWHERE

78 Countries

1000 Universities Teaching CUDA

400K CUDA Developers

+ 240 Resellers Worldwide
SCALING DL
ALPHAGO

Training DNNs: 3 weeks, 340 million training steps on 50 GPUs

Play: Asynchronous multi-threaded search

- Simulations on CPUs, policy and value DNNs in parallel on GPUs
- Single machine: 40 search threads, 48 CPUs, and 8 GPUs
- Distributed version: 40 search threads, 1202 CPUs and 176 GPUs

Outcome: Beat World Go champion in best of 5 matches

http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
http://deepmind.com/alpha-go.html
TESLA BUILT FOR THE DATA CENTER

- 24/7 Uptime
  Maximize reliability

- Scalable Performance
  Boost data center throughput

- Data Center Ready
  Simplify system operations
END-TO-END DESIGN FOR SYSTEM UPTIME

- 24/7 Uptime
- Scalable Performance
- Data Center Ready

Differentiated Engineering
- Low Operating Voltage for Long Term Reliability
- Large Guard-band for Guaranteed Quality
- Error Correction Code (ECC) for Data Integrity

Extensive Qualification & Testing
- Long Burn-in Testing
- Zero Error Tolerance at Aggressive Clocks

Guaranteed Quality
- System Qual. Tests: Thermal, Stress, Airflow rate, Shock & Vibe
- System Monitoring and Management for Tesla only
- Dedicated Technical Staff for Failure Analysis
DATA CENTER QUALIFIED BY SERVER OEMS

24/7 Uptime
Scalable Performance
Data Center Ready

Server with Tesla GPU
- Designed for max airflow through GPU
- Supports airflow front-to-back & back-to-front
- Lower power consumption
- GPU Temp Running Linpack: 54C

Server with Unqualified GPU
- Works against server airflow
- Higher power consumption
- Lower reliability
- GPU Temp Running Linpack: 71C

Temp: 54C
Temp: 71C
SCALE-OUT PERFORMANCE IN THE DATA CENTER

GPUDirect RDMA
- Direct transfers between GPUs
- 67% Lower GPU-to-GPU Latency
- 5x Higher GPU-to-GPU MPI Bandwidth

Up to 2x Faster
Application Performance at Scale with GPUDirect RDMA

Hoomd-Blue Application
LJ Liquid Benchmark, 256K Particles

<table>
<thead>
<tr>
<th># of Nodes</th>
<th>without RDMA</th>
<th>with RDMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NVLINK DELIVERS SCALABLE PERFORMANCE

More than 45x Faster with 8x P100 Interconnected with NVLink

<table>
<thead>
<tr>
<th>Application</th>
<th>2x K80 (M40 for Alexnet)</th>
<th>2x P100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe/Alexnet</td>
<td>45x</td>
<td>45x</td>
</tr>
<tr>
<td>VASP</td>
<td>30x</td>
<td>45x</td>
</tr>
<tr>
<td>HOOMD-Blue</td>
<td>25x</td>
<td>45x</td>
</tr>
<tr>
<td>COSMO</td>
<td>20x</td>
<td>45x</td>
</tr>
<tr>
<td>MILC</td>
<td>15x</td>
<td>45x</td>
</tr>
<tr>
<td>Amber</td>
<td>10x</td>
<td>45x</td>
</tr>
<tr>
<td>HACC</td>
<td>5x</td>
<td>50x</td>
</tr>
</tbody>
</table>

2x Haswell CPU
DATA CENTER GPU MANAGEMENT
Enterprise-Grade Management Tool for Operating the Data Center

24/7 Uptime
Scalable Performance
Data Center Ready

Device Management

Per GPU Configuration & Monitoring
• Device Identification
• Board Monitoring
• Clock Management

All GPUs Supported

Data Center GPU Manager

Active Health Monitoring
Runtime Health Checks
Prologue Checks
Epilogue Checks

Diagnostics & System Validation
Deep HW Diagnostics
System Validation Tests

Policy & Group Config Management
Pre-configured policies
Job level accounting
Stateful configuration

Power & Clock Mgmt.
Dynamic Power Capping
Synchronous Clock Boost

*Prerelease Now, GA Q3
SCALING NEURAL NETWORKS

Data Parallelism

Notes:

Need to sync model across machines
- Requires P-fold larger batch size
- Works across many nodes – parameter server approach – linear speedup

Adam Coates, Brody Huval, Tao Wang, David J. Wu, and Andrew Ng
SCALING NEURAL NETWORKS

Model Parallelism

Notes:
- Allows for larger models than fit on one GPU
- Most commonly used within a node – GPU P2P
- Effective for the fully connected layers
- Requires much more frequent communication between GPUs

Adam Coates, Brody Huval, Tao Wang, David J. Wu, and Andrew Ng
PARTNER RESULTS - Baidu

Near linear scaling - synchronous training

PARTNER RESULTS - DGX-1 TENSORFLOW

Training: NVIDIA® DGX-1™ synthetic data (1, 2, 4, and 8 GPUs)

https://www.tensorflow.org/performance/benchmarks#methodology
METROPOLIS PARTNER PROGRAM
## NVIDIA METROPOLIS PARTNERS

<table>
<thead>
<tr>
<th>AllGoVision</th>
<th>AnyVision</th>
<th>BriefCam</th>
</tr>
</thead>
<tbody>
<tr>
<td>herta</td>
<td>iCetana</td>
<td>MotionDSP</td>
</tr>
<tr>
<td>OMNI AI</td>
<td>OpenALPR</td>
<td>SENSETIME</td>
</tr>
<tr>
<td>VisionLabs</td>
<td>VOCORD</td>
<td>XJERA LABS</td>
</tr>
</tbody>
</table>
Real-Time Multistream Analytics
ADVANCED MODELS MAY ERODE PRIVACY

The Target Dilemma

- Using a basket of 25 product features, Target generated classification score
- This resulted in empathetically recommending baby-related promotions
- In the literature, Youyou et. al. worked with Facebook likes

Duhigg 2016, New York Times
THE ROLE OF PUBLIC INFRASTRUCTURE
**TESLA V100**

- 21B transistors
- 815 mm²
- 80 SM
- 5120 CUDA Cores
- 640 Tensor Cores
- 16 GB HBM2
- 900 GB/s HBM2
- 300 GB/s NVLink

*full GV100 chip contains 84 SMs*
ROAD TO EXASCALE
Volta to Fuel Most Powerful US Supercomputers

Volta HPC Application Performance

<table>
<thead>
<tr>
<th>Application</th>
<th>DGEMM</th>
<th>FFT</th>
<th>Physics (QUDA)</th>
<th>Seismic</th>
<th>STREAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative to Tesla P100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGEMM</td>
<td>1.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFT</td>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physics (QUDA)</td>
<td>1.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seismic</td>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STREAM</td>
<td>1.35</td>
<td></td>
<td></td>
<td></td>
<td></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.
Thank you!
leot@nvidia.com