Cognitive Workloads on the Cloud: What Clouds Need To Do

Dr. Khoa Huynh, Dr. Larry Brown, Brian Wan – IBM Watson Cloud
• Introduction
• What cloud providers are offering today
• Cognitive workloads
• Performance data
• What cloud infrastructures need to do
• Q & A
## What Cloud Providers Are Offering Today

<table>
<thead>
<tr>
<th></th>
<th>IaaS Offering (customer-programmable)</th>
<th>SaaS Offering (accelerators used internally by offered services)</th>
<th>Custom Hardware</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GPU</td>
<td>FPGA</td>
<td>ASIC</td>
</tr>
<tr>
<td>AWS</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Google</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Azure</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>IBM Bluemix</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Nimbix</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>
Current Offerings in Major Clouds

as of July 2017

• IBM SoftLayer
  – Bare-metal server supports Nvidia K80 and P100 GPUs
    • From 2 Nvidia K80 GPUs (1 PCIe card) to 4 Nvidia K80 GPUs (2 PCIe cards)
    • From 1 Nvidia P100 GPU (1 PCIe card) to 2 Nvidia P100 GPUs (2 PCIe cards)

• Amazon AWS
  – EC2 P2 GPU instances for compute
    • From 1 Nvidia K80 GPU (p2.xlarge) to 16 Nvidia K80 GPUs (p2.16xlarge)
  – F1 EC2 instances available in N. Virginia zone (since 5/20/17)
    • From 1 Xilinx FPGA (f1.2xlarge) to 8 Xilinx FPGAs (f1.16xlarge)
  – EC2 instances with FPGA development images
    • Xilinx development tools only (no FPGA):
      • FPGA Development Kit
        • Support for custom logic (examples are provided in SystemVerilog, NOT OpenCL) with Amazon-provided FPGA Shell and AXI ports (for off-chip memory, PCIe, FPGA-to-FPGA communications)
        • Code: https://github.com/aws/aws-fpga

• Google Cloud
  – Compute Engine supports Nvidia K80 GPUs (GPU costs are additional to the compute node code’s cost)
    • From 1 Nvidia GPU (= ½ K80, 1 die) to 8 Nvidia GPUs (= 4 K80, 8 dies)
  – Google TPU is not directly exposed but used by several Google services (Vision, Image Search, Photos, Translate)

• Microsoft Azure
  – Linux VMs with GPUs for compute
    • From 1 Nvidia K80 GPU (NC6) to 4 Nvidia K80 GPUs (NC24)
  – Catapult V2
    • Datacenter hardware acceleration with Intel Stratix V FPGAs
    • Only used internally by Azure and Bing services, as well as Cortana
Cognitive Workloads
Machine learning
- Deep learning is a *subset* of machine learning
- Common approaches: linear regression, decision trees, association rules, etc.
- Some frameworks (Spark Mllib, Python SciKit, etc.) do *not* support GPU / accelerator options

Deep learning
- More and more in our lives (image recognition, Watson, Siri, Alexa, Cortana, language translation, self-driving cars, etc.)
- Convolutional neural networks require many multi-dimensional matrix multiplications
- Deep learning frameworks provide tools to implement neural networks

We focus more on Deep Learning
Deep Learning (DL) Neural Net Models

- Neural network models evaluated: AlexNet, GoogLeNet, VGG-16, InceptionV3
- DL frameworks evaluated: Caffe, TensorFlow

Deep Learning Neural Network Models

Model Training
- Must train with huge number of images to get sufficient accuracy
- Include forward and backward propagation passes (backward passes reduce loss function during training)
- Very compute-intensive – training could take many days and weeks
- Need parallel processing to reduce training time

Model Inferencing
- Only include forward propagation passes
- Less compute-intensive – usually only involve single GPU
Available Accelerators (GPUs, FPGAs)

Nvidia P4 GPU

FPGA hardware can map well to deep-learning neural networks
- Use all DSP blocks to enable parallelism for spatial processing
- Use embedded memory to keep all data on-chip between neural network layers (temporal processing) - exploiting high internal memory bandwidth

Nvidia K80 GPU

Intel Deep Learning Inference Accelerator (DLIA)

Xilinx UltraScale FPGA
<table>
<thead>
<tr>
<th>Accelerator</th>
<th>Type</th>
<th>Availability</th>
<th>Supported Clouds</th>
<th>Hardware Form Factor</th>
<th>TDP</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA P100</td>
<td>GPU</td>
<td>Now</td>
<td>• SoftLayer (PCIe) • Nimbix • IBM Minsky (NVLink)</td>
<td>• Full-height, double-slot PCIe card • Socket (SXM2)</td>
<td>300W</td>
<td>DL model training &amp; large-scale inferencing</td>
</tr>
<tr>
<td>NVIDIA P40</td>
<td>GPU</td>
<td>Now</td>
<td>• SoftLayer (PCIe) • Nimbix • IBM Minsky (NVLink)</td>
<td>• Full-height, double-slot PCIe card</td>
<td>250W</td>
<td>Large-scale DL inferencing</td>
</tr>
<tr>
<td>NVIDIA P4</td>
<td>GPU</td>
<td>Now</td>
<td>• SoftLayer (PCIe) • Nimbix • IBM Minsky (NVLink)</td>
<td>• Half-height, single-slot PCIe card</td>
<td>75W</td>
<td>Scale-out DL inferencing</td>
</tr>
<tr>
<td>NVIDIA V100</td>
<td>GPU</td>
<td>2H2017</td>
<td>• SoftLayer (PCIe) • Nimbix • IBM Minsky (NVLink)</td>
<td>• Full-height, double-slot PCIe card • Socket (SXM2)</td>
<td>300W</td>
<td>Specifically designed for DL model training / inferencing</td>
</tr>
<tr>
<td>NVIDIA K80</td>
<td>GPU</td>
<td>Now</td>
<td>• SoftLayer (PCIe) • Nimbix • IBM Minsky (NVLink)</td>
<td>• Full-height, double-slot PCIe card</td>
<td>300W</td>
<td></td>
</tr>
<tr>
<td>Intel Stratix10</td>
<td>FPGA</td>
<td>2H2017</td>
<td>N/A</td>
<td>• Full-height PCIe card • Half-size PCIe card possible • Socket • MCP (Skylake+Arria10)</td>
<td>30-50W</td>
<td>DLIA is specifically tuned for AI</td>
</tr>
<tr>
<td>Intel Arria10/DLIA</td>
<td>FPGA</td>
<td>2H2017</td>
<td>N/A</td>
<td>• Full-height PCIe card • Half-size PCIe card possible • Socket • MCP (Skylake+Arria10)</td>
<td>30-50W</td>
<td>DLIA is specifically tuned for AI</td>
</tr>
<tr>
<td>Xilinx UltraScale+</td>
<td>FPGA</td>
<td>Now</td>
<td>• AWS (PCIe) • Nimbix</td>
<td>• Half-size PCIe card</td>
<td>30-50W</td>
<td></td>
</tr>
<tr>
<td>Knights Mill</td>
<td>Xeon Phi</td>
<td>4Q2017</td>
<td>N/A</td>
<td>• Socket (single)</td>
<td>280W</td>
<td></td>
</tr>
<tr>
<td>Intel Lake Crest/Spring Crest</td>
<td>Nervana (ASIC)</td>
<td>N/A</td>
<td>• PCIe cards</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Performance Data
### Deep Learning Model Training

**Number of Images Processed Per Second (Training)**

<table>
<thead>
<tr>
<th>System</th>
<th>VGG-16 on Caffe</th>
<th>InceptionV3 on TensorFlow</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x CPU cores (No GPU)</td>
<td>1.5</td>
<td>2.1</td>
</tr>
<tr>
<td>SoftLayer (8 Xeon E5-2683v4)</td>
<td>29.42</td>
<td>56.87</td>
</tr>
<tr>
<td>SL Bare-Metal (Dual Xeon ES-2690v4)</td>
<td>41.07</td>
<td>76.92</td>
</tr>
<tr>
<td>4 x K80 PCIe GPUs</td>
<td>SL Bare-Metal (Dual Xeon ES-2690v4)</td>
<td>SL Bare-Metal (Dual Xeon ES-2690v4)</td>
</tr>
<tr>
<td>1 x P100 PCIe GPU</td>
<td>83.27</td>
<td>160.31</td>
</tr>
<tr>
<td>2 x P100 PCIe GPUs</td>
<td>127.93</td>
<td>193.85</td>
</tr>
<tr>
<td>16 x K80 PCIe GPUs</td>
<td>328.21</td>
<td></td>
</tr>
<tr>
<td>4 x NVLinks P100 GPUs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- Input dataset: ImageNet (crop size=224x224)
- Highest batch sizes used to fully exploit GPU memory

**Running in Docker containers has less than 4% performance overhead**
Deep Learning Model Training

Deep-Learning Model Training: Price-Performance Comparison
VGG-16 on Caffe

Higher is better

Note:
- Input dataset: ImageNet (crop size=224x224)
Deep Learning Model Training

Deep-Learning Model Training - InceptionV3 on TensorFlow
16-Node Cluster with 32 x Nvidia P100 GPUs
(Each Node = Bare-Metal Server with 2 x Nvidia P100 PCIe GPUs)

- Each GPU runs a single worker server

Notes:
- IBM Research just announced Distributed Deep Learning (DDL) for PowerAI that could scale linearly across hundreds of Nvidia GPUs.
Deep Learning Model Inferencing

Deep-Learning Model Inferencing
Image Classification with VGG-16 on Caffe

<table>
<thead>
<tr>
<th>System</th>
<th>Number of Images Processed Per Second (Inferencing)</th>
<th>Higher is better</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x Nvidia K80 GPU</td>
<td>50.12</td>
<td></td>
</tr>
<tr>
<td>SL Bare-Metal (Dual Xeon ES-2690v4)</td>
<td>128.74</td>
<td></td>
</tr>
<tr>
<td>1 x Nvidia P4 GPU</td>
<td>289.28</td>
<td></td>
</tr>
<tr>
<td>Intel Bare-Metal (Dual Xeon ES-2650v3)</td>
<td>45.61</td>
<td></td>
</tr>
<tr>
<td>1 x Nvidia P100 GPU</td>
<td>11.7</td>
<td></td>
</tr>
<tr>
<td>SL Bare-Metal (Dual Xeon ES-2690v4)</td>
<td>1 x FPGA for DL</td>
<td></td>
</tr>
<tr>
<td>Intel Bare-Metal (Dual Xeon ES-2690v4)</td>
<td>1 x Nvidia K80 GPU</td>
<td></td>
</tr>
<tr>
<td>AWS p2.16xlarge (Dual Xeon ES-2686v4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Deep-Learning Model Inferencing
Image Classification with VGG-16 on Caffe

<table>
<thead>
<tr>
<th>System</th>
<th>Single-image classification latency (ms)</th>
<th>Shorter is better</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x Nvidia K80 GPU</td>
<td>50.12</td>
<td></td>
</tr>
<tr>
<td>1 x Nvidia P4 GPU</td>
<td>128.74</td>
<td></td>
</tr>
<tr>
<td>1 x Nvidia P100 GPU</td>
<td>289.28</td>
<td></td>
</tr>
<tr>
<td>1 x FPGA for DL</td>
<td>45.61</td>
<td></td>
</tr>
<tr>
<td>1 x Nvidia K80 GPU</td>
<td>11.7</td>
<td></td>
</tr>
</tbody>
</table>

Deep Learning Model Inferencing
Image Classification with VGG-16 on Caffe

Higher is better
Shorter is better
### Deep Learning Model Inferencing

#### Price-Performance Comparison

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Number of Images Processed Per US$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No GPU</td>
<td>2702.38</td>
</tr>
<tr>
<td>1 x Nvidia PCIe K80 GPUs</td>
<td>31737.50</td>
</tr>
<tr>
<td>1 x Nvidia PCIe P100 GPU</td>
<td>261948.42</td>
</tr>
<tr>
<td>1 x Nvidia PCIe K80 GPUs</td>
<td>206800.00</td>
</tr>
</tbody>
</table>

**Notes:**
- The Nvidia P4 GPU is comparable to many FPGAs in terms of power consumption (50-70W)

#### Deep-Learning Model Inferencing (Image Classification)

**Classification Power Efficiency (Images/Second/Watt)**

- **Nvidia P4 GPU**
- **FPGA for DL**

**VGG-16 Neural Net on Caffe Framework**

- Longer is better

What Cloud Infrastructures Need To Do?
What Cloud Infrastructures Need To Do

• **Deep Learning (DL) model training** – support the latest Nvidia GPUs (P100, V100)
  - SXM2 form factor (NVLink between GPUs, NVLink between GPUs and POWER CPUs, PCIe between GPUs and x86 CPUs)
  - PCIe form factor (PCIe between all GPUs and CPUs)
  - Need to support the largest possible GPU memory sizes (limiting factor for large batch sizes during model training)
  - *If costs are important*, consider Nvidia K80 GPUs
  - Our data flow analysis indicates that disk I/O traffic does not impact our performance measurements because of host system caching
  - Distributed model training & scaling will be important – IBM Research’s DDL can scale well to hundreds of GPUs
  - Support for both end users and DL services

• **DL model inferencing** – support cheaper, lower-powered Nvidia P4 GPUs as DL inferencing engine and P100 / V100 GPUs at the high end
  - Nvidia P4 GPU has better throughput, better latency, and better power efficiency than a DL FPGA available to us
  - FPGAs (e.g. Intel DLMIA) / ASICs (e.g. Google TPU, Hailo) could be also used as inferencing engine for DL services
    - FPGAs have more flexibility to support new features (e.g. low precision, binary/ternary weights/inputs) in hardware than GPUs

• **Need to support pre-built instance images for**
  - Accelerators (GPUs, FPGAs) for end-users
    - Intel has pre-built installation image for FPGAs that contains the DL entire stack and tools for easy setup
    - Debugging / profiling tools (e.g. nvidia-smi, nvprof, etc.)

• **Virtualization technologies**
  - One-to-one or many-to-one GPU to VM/server/container relationship
  - PCI pass-through (GPUs are dedicated to an instance/VM); Single-Root Virtual I/O (SR-IOV); Containers
    - *Not recommended* – Nvidia GRID vGPU technology (targeted for virtual desktop, video codec, gaming – not DL/HPC)
Q & A