Facebook Flexible GPU Expander Big Basin Refresh

Whitney Zhao/HW Eng/Facebook Inc.
Xiaodong Wang/SW Eng/Facebook Inc.
Agenda
Agenda

- Introduction
- Architecture
- Performance
- Questions
Impact

Facebook’s commitment to developing AI & advancing ML
Goal

- Open, full contribution to OCP
- Disaggregation/Modularity
- Serviceability

2016: Big Sur

2017: Leopard + Big Basin

2018: Tioga Pass + Big Basin V2
Big Basin V2 Overview

- 3 OU chassis
- Open Rack v2 compatible
- 8x Nvidia Tesla V100 GPUs; NVLink capable
- 300W TDP for each Tesla V100 GPU
- Facebook 2S Server Tioga Pass as Head node
A deeper look into Big Basin

- IO board
- Baseboard
- Midplane board

Baseboard on sliding tray
Serviceability

- Quick repairs at data center
- Telemetries accessible from head node
- Provisioning Big Basin with its head node is not much different from provisioning existing servers; these servers come with additional GPUs.
Agenda

- Introduction
- Architecture
- Performance
- Questions
Architecture (Headnode to Big Basin)

- MiniSAS HD cable (2 for each x16)
  - Standard PCIe x16
  - Present Pin
  - USB2.0
  - IPMB/I2C

Leopard + Big Basin (Tesla P100)

Tioga Pass + Big Basin V2 (Tesla V100)
Architecture (NVLINK)

Big Basin W/Nvidia Tesla P100

Big Basin V2 W/ Nvidia Tesla V100
Agenda
Performance

• Hardware Spec Improvement

• Application performance
  • Computer vision
    • Single-GPU
    • Multi-GPU scalability
  • TensorCore
  • Neural machine translation
## Performance

### Comparisons of GPU Hardware

<table>
<thead>
<tr>
<th>Metrics</th>
<th>NVIDIA V100</th>
<th>NVIDIA P100</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP-32</td>
<td>15 TFLOPS</td>
<td>10.6 TFLOPS</td>
<td>1.42x</td>
</tr>
<tr>
<td>FP-16</td>
<td>30 TFLOPS</td>
<td>21.2 TFLOPS</td>
<td></td>
</tr>
<tr>
<td>TensorCore</td>
<td>125 TFLOPS</td>
<td>NA</td>
<td>Up to 5x</td>
</tr>
<tr>
<td>Mem Bandwidth</td>
<td>900 GB/s</td>
<td>720 GB/s</td>
<td>1.25x</td>
</tr>
<tr>
<td>NVLink</td>
<td>300 GB/s</td>
<td>160 GB/s</td>
<td>1.88x</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>300 W</td>
<td>300 W</td>
<td></td>
</tr>
</tbody>
</table>
Performance

• Comparisons of GPU Hardware

• Head-node upgrade: Tioga Pass
  • New CPU architecture: Broadwell to Skylake
  • Double PCIe bandwidth
  • Upgraded 100G NIC

• CUDA 9 + cudnn 7: faster libraries, etc.
Impact - Computer Vision
Performance metrics in Computer Vision

- Computer Vision: resnet-50
  - 1-GPU training speed: use P100 + CUDA 8 as baseline
Computer Vision Performance

- Computer Vision
  - Multi-GPU speedup vs. 1 P100

![Graph showing speedup vs 1 P100](image)
Computer Vision Performance

- Computer Vision
  - High-bandwidth FP-16 TensorCore (WIP)
Machine Translation

Better Translation Quality

Phrase-based statistical approach

Neural network approach

Onlann, İzmir’ın neden hayır dediğini anlamalarını beklemiyoruz.
Their, İzmir’s why you said no we don’t expect them to understand.

Onlann, İzmir’ın neden hayır dediğini anlamalarını beklemiyoruz.
We don’t expect them to understand why İzmir said no.
Machine Translation Performance

- Neural Machine translation

- P100 + CUDA 8
  Training Throughput as Baseline

- V100 + CUDA 8
  1.45X

- V100 + CUDA 9
  2.2X
THIS JOURNEY
1% FINISHED
Questions?
OCP Marketplace

- [http://www.opencompute.org/products/specsanddesign?keyword=Big+basin](http://www.opencompute.org/products/specsanddesign?keyword=Big+basin)