Background

Cerebras

Graphcore

Habana

Comparisons

References
CPU VS GPU
Cerebras
Currently ML accelerators consist of several GPUs wired together and attached to an external memory.

Aims to condensing components into a single chip (Wafer-Scale processor).

The Cerebras CS-1 boasts 1.2 trillion transistor and 400,000 AI optimized cores onto 46,225 square millimeters of silicon.

This chip is not compared to other chips in the market but to the entire cluster (CS-1 vs DGX-1).
Entire neural network can be placed onto the chip at once instead of running one layer at a time.

Common ML frameworks like TensorFlow and Pytorch are supported.

All 400,000 cores in CS-1 are AI-optimized.

Native sparse processing unlike NVIDIA's approach with cuSPARSE library.
• In von Neumann architecture memory is kept separate from the core

• Memory is slow and has high latency

• Caches are used in traditional memories

• Low data reusability makes caches less effective in ML operations

• Better idea is to have the memory on the same chip
- Larger chips -> highly impossible to have zero defects
- Trilogy created ‘Redundancy of every logic gate’
- Cerebras included redundant cores fabrics links on the chips to replace defective cores
- Power efficiency is an issue in larger chips
- If power is delivered from one edge, voltage might be dropped before reaching middle of the chip
- So, the chip is powered from above and not from any sides
• Dense packing made heat density a challenge

• CS-1 uses upto 20 kilowatts where as V100 uses upto 300 watts

• Direct air cooling can’t be used as a cooling technique

• The silicon chip is stacked on top of a watercooled plate

• Every section of the chip can be cooled this way
• Power delivery, the chip and cooling system expand at different rate

• All the components may be disconnected

• In extreme conditions, the chip can be broken

• Cerebras invented a material to address this issue
• Pittsburgh Supercomputing Center has deployed a unique high performing AI system

• Partnered with CMU, Cerebras and HPE

• Neocortex has two extremely powerful Cerebras CS-1 servers

• HPE superdome Flex, user friendly front-end HPC solution for CS-1

• HPE superdome Flex has 24 TB of memory, 205 TB of flash storage, 32 powerful Intel Xeon CPUs and 24 network interface cards for 1.2 Tbps bandwidth to each CS-1
Graphcore
• After Moore’s law, parallel computing over many chips is the trend

• Graphcore introduced Intelligent Processing Unit(IPUs) to accelerate ML applications

• Graphcore C2 is a PCI express card containing two IPUs

• Can be programmed using Graphcore’s Poplar SDK
IPU-Core™
1472 independent IPU-Core™
8832 independent program threads executing in parallel

In-Processor-Memory™
900MB In-Processor-Memory™ per IPU
47.5TB/s memory bandwidth per IPU

IPU-Exchange™
8 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern

PCIe
PCI Gen4 x16
64 GB/s bidirectional bandwidth to host

IPU-Links™
10 x IPU-Links,
320GB/s chip to chip bandwidth
• After Moore’s law, parallel computing over many chips is the trend

• Graphcore introduced Intelligent Processing Unit(IPUs) to accelerate ML applications

• Graphcore C2 is a PCI express card containing two IPUs

• Can be programmed using Graphcore’s Poplar SDK
• GC2 “Colossus Mk1” IPU [2018 power-on]
• 23,647,173,309 active transistors in TSMC N16
• 1216 processor tiles @ 256KiB
• Total 125Tflop/s + 304MiB SRAM
• 62TB/s memory, 7.8TB/s inter-tile, 320GB/s inter-chip
• GC200 “Colossus Mk2” IPU [2020 power-on]
• 59,334,610,787 active transistors in TSMC N7
• 1472 processor tiles @ 624KiB
• Total 250Tflop/s + 896MiB SRAM
• 62TB/s memory, 7.8TB/s inter-tile, 320GB/s inter-chip
<table>
<thead>
<tr>
<th>IPUS</th>
<th>16 * GC200 IPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPU-M2000s</td>
<td>4</td>
</tr>
<tr>
<td>Memory</td>
<td>1038 GB</td>
</tr>
<tr>
<td>Performance</td>
<td>4 petaFLOPS FP16.16</td>
</tr>
<tr>
<td></td>
<td>1 petaFLOPS FP32</td>
</tr>
<tr>
<td>IPU cores</td>
<td>23552</td>
</tr>
<tr>
<td>Division</td>
<td>Model</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Closed</td>
<td>Resnet 50 v1.5</td>
</tr>
<tr>
<td>Closed</td>
<td>BERT</td>
</tr>
<tr>
<td>Open</td>
<td>BERT</td>
</tr>
</tbody>
</table>
Habana
• Habana’s Gaudi compares itself with Nvidia’s V100
• Gaudi uses 140W which is half of what V100 uses
• Goya focuses on int computation and Gaudi has FP computation
• Habana sells a PCI accelerator containing one Goya based chip
• Habana HLS-1 has 4 Gaudi HL-205 cards (4 * 2 Gaudi chips)
• Total memory of 128 GB (4 * 32)
• Max power usage 2.3 KW
• HLS-1H pod contains 16 Gaudi HLS-1H systems
• Uses Resnet 50 from MLPERF benchmarks
• Single Gaudi chip delivers 1590 images per second
• Scaling this to eight Gaudi chips gave 12008 images per second
• Ability to scale the training (Claim that NVIDIA’s GPUs don’t perform great after 16 GPUs)
Habana – HLS-1 Architecture
Comparison
• Might not be apple to apple comparisons
• MLPerf is a recent effort for standardizing benchmarks (only training time)
• Several other measurements like hardware utilization, power efficiency, cost efficiency are ignored
• Habana claims that Gaudi exceeds 1,650 images per second (IPS) with a batch size of 64 compared to 1,360 IPS with an unspecified batch size for NVIDIA’s Tesla V100.
• Habana Gaudi consumes 140w power which is half of NVIDIA’s V100
• Cerebras claims there is 100-1000 fold improvement in terms of performance
• Graphcore C2 has two IPUs which draw 300 W similar to single chip NVIDIA’s V100
<table>
<thead>
<tr>
<th>Metric</th>
<th>Nvidia V100</th>
<th>Cerebras CS1</th>
<th>GraphCore IPU1</th>
<th>Graphcore IPU2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area(mm²)</td>
<td>815</td>
<td>46,225</td>
<td>900</td>
<td>823</td>
</tr>
<tr>
<td>Transistor #</td>
<td>11 B</td>
<td>1.2T</td>
<td>23.6 B</td>
<td>59.4 B</td>
</tr>
<tr>
<td>SRAM</td>
<td>36 MB</td>
<td>18 GB</td>
<td>300 MB</td>
<td>900 MB</td>
</tr>
<tr>
<td>Cost ($)</td>
<td>10,664</td>
<td>2,000,000</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

- Graphcore’s IPUPOD16 costs 150,000 whereas near comparable NVIDIA’s DGX A100 costs 300,000
- The Nvidia system scored better on BERT than the IPU system, roughly 21 minutes versus 34 minutes for Graphcore

2) https://habana.ai

3) https://www.cerebras.net/

4) https://www.eetasia.com/graphcore-ipu-vs-nvidia-gpus-how-theyre-different/


6) https://www.binarytides.com/graphics-card-specs-explained/

Thank you!