THE LATEST IN CHIP DEVELOPMENT

JARED ROBBINS
GUIDE

- 8.i Intel
- 8.j Graphcore
- 8.k Habana
- 8.l Cerebras
INTEL: NEUROMORPHIC COMPUTING

• Attempts to emulate the structure and operation of the human brain

• Involves probabilistic computing
  • Deterministic computing has predictable outcomes, probabilistic does not because of random influences
  • Aims to output a result along with a measurement of uncertainty

• Spiking Neural Networks (SNN) are used to model the brain
  • Each neuron transmits data when a threshold is reached
INTEL: LOIHI

• Chip designed for neuromorphic computing
• Named after a rising volcano in Hawaii (“long”)
• Uses Intel’s 14nm process
• Contains 128 neuromorphic cores and 131,072 neurons
  • Each core has 1024 neurons
• Implements a fully integrated SNN

INTEL: LOIHI, MESH CAPABILITY

- Loihi chips have the interface to support a mesh
- The protocol supports 16,384 chips
  - Each chip in this protocol can have up to 4096 cores (current is 128)
  - This gives the potential to have nearly 69 billion neurons on one (large) board
- For reference, the human brain has around 100 billion neurons

https://en.wikichip.org/wiki/intel/loihi
INTEL: LOIHI, USE CASES

• Constraint satisfaction
  • Airline scheduling, package delivery planning

• Searching graphs and patterns
  • Just like how humans excel at finding paths and facial recognition

• Optimization problems
  • Anything involving getting a number to a desirable outcome
INTEL: LOIHI, EXAMPLES

• Professor Chris Eliasmith of Waterloo University found Loihi to beat GPU power usage by 105x and specialized hardware by 5x. Loihi kept real-time results when scaled up 50x while only increasing power 30%. The specialized hardware lost real-time results and increased power 500%.

• Foosball.

• Professor Konstantinos Michmizos of Rutgers compared a Loihi-run network with a CPU-run network and found Loihi achieved the same results with 100 times less power consumption while solving a SLAM problem (generating a map and navigating an unknown environment ex: self driving cars)

https://www.therobotreport.com/intel-pohoiki-beach-neuromorphic-chip-researchers/
INTEL: NAHUKU

- The Nahuku board has 32 Loihi chips (16 on each side)
- Intel uses this as a building block for larger systems
- The board has four large connectors with cardinal directions
  - These allow the boards to be arranged in a grid to operate as one

https://en.wikichip.org/wiki/intel/loihi
INTEL: PAHOIKI SPRINGS

• Includes 24 Nahuku boards

• If you have kept up with the math... that's 100,663,296 neurons
  
  • Fruit bats, mole rats, and hamsters have a similar number of neurons in their brains

• Only consumes 500W of power
  
  • For comparison, the new NVIDIA RTX 3080 is rated at 320W

• Currently only available for researchers

https://www.fierceelectronics.com/electronics/intel-shares-pohoiki-springs-chassis-ai-researchers
INTEL: WHAT’S NEXT?

• I have not found what’s next after Pahoiki Springs. It was announced in March 2020

• Intel signed a 3 year deal with Sandia National Laboratories in New Mexico (Fun fact “Sandia” = Watermelon in Spanish) and will be working with them using a 50 million neuron machine

• Intel’s delayed but coming 7nm process should further advance Loihi
QUESTIONS?
GUIDE

• 8.i Intel
• 8.j Graphcore
• 8.k Habana
• 8.l Cerebras
GRAPHCORE: INTRO

- Officially founded in 2016 in Bristol, England
- Received series A, B, C, D1, and D2 funding from venture capitalists
- Develops IPU units (a semiconductor company)
- As of 2019 IPU units are available on Microsoft Azure
- Rapidly growing with a valuation of around 2b
- Considered a “Unicorn” and plans to IPO eventually
GRAPHCORE: IPU

- IPU's or intelligence processing units aim to provide compute power for AI workloads.
- Graphcore has optimized its IPU to excel in massively parallel low precision floating point operations.
- Developed because Graphcore found the most limiting factor in AI for developers is the hardware.

https://www.graphcore.ai/products/ipu
COLOSSUS MK2 IPU CHART COMING UP… PREPARE FOR INFO OVERLOAD
IPU-Tile™
1472 independent IPU-Tiles™ each with an IPU-Core™ and In-Processor-Memory™

IPU-Core™
1472 independent IPU-Core™
8832 independent program threads executing in parallel

In-Processor-Memory™
900MB In-Processor-Memory™ per IPU
475TB/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,
3200GB/s chip to chip bandwidth

https://www.graphcore.ai/products/ipu
GRAPHCORE: COLOSSUS MK2, MORE DETAILS

- 59.4 billion transistors on a 823mm\(^2\) die
  - In comparison, NVIDIA’s latest A100 has 54 billion on a 826mm\(^2\) die

- Developed using TSMC’s 7nm process
  - Same as NVIDIA and AMD

- Supports Stochastic Rounding which allows the MK2 to keep all arithmetic in a 16 bit format
  - This is unique, saves power, and saves time

https://www.graphcore.ai/posts/introducing-second-generation-ipu-systems-for-ai-at-scale
• Graphcore is a young company, but is already making huge improvements
• MK1 came out in 2017, MK2 in 2020
• In just three years speed has increased many times over
• It can be expected that more is on the way

https://www.graphcore.ai/posts/introducing-second-generation-ipu-systems-for-ai-at-scale
GRAPHCORE: IPU-MACHINE M2000

• IPU-Machine M2000 is the form in which Graphcore sells its IPUs
• Powered by 4 Colossus MK2s
• Delivers 1 petaflop of computing power
• Can be scaled to 64,000 IPUs
  • This translates to 16 exaflops of computing power

https://www.graphcore.ai/posts/introducing-second-generation-ipu-systems-for-ai-at-scale
GRAPHCORE: A CHALLENGER ARISES!

- Recall the recent mention of the NVIDIA A100? It’s back!
- The IPU-Machine M2000 competes with the NVIDIA DGX A100
  - 1 petaflop
  - $32,450 (these can be stacked)
  - 4 Colossus MK2s
  - 5 petaflops
  - $199,000
  - 8 A100 Tensor Core GPUs
GRAPHCORE: WHAT’S NEXT?

• According to Bloomberg as of Nov 2020 Graphcore is in talks to raise another $200 million in funding
  • Sources are “people familiar with the matter”

• TSMC announced in August 2020 that they are working with Graphcore to implement a new processor based on their 3nm process
  • The Colossus MK1 is 16nm and MK2 is 7nm
  • TSMC plans to have mass production of 3nm in the second half of 2021
QUESTIONS?

GUIDE

- 8.i Intel
- 8.j Graphcore
- 8.k Habana
- 8.l Cerebras
HABANA: INTRO

• Semiconductor company founded in 2016 in Tel Aviv, Israel
• Develops neural processors for training and inference tasks
• Habana, but actually just Intel again
  • Habana Labs was acquired in late 2019 by Intel for $2 billion

https://www.graphcore.ai/posts/introducing-second-generation-ipu-systems-for-ai-at-scale
HABANA: TPC

• Habana develops two main products, but both use the same base components
• The Tensor Processor Core was designed by Habana from the ground up
• Fully C programmable
• Has local memory, but no standard cache
• Each core is a VLIW DSP design that is optimized for AI
  • “Very long instruction word digital signal processor” Beyond being a mouthful, this helps the core process instructions in parallel
HABANA: GOYA

- Named for Francisco Goya, a famous Spanish artist

https://www.theartstory.org/artist/goya-francisco/artworks/
HABANA: GOYA (FOR REAL)

• The Goya processor aims to provide fast and efficient inference for AI workloads

• The next slide will be a deep dive into the architecture
HABANA: GOYA, ADDITIONAL INFO

- Developed with TSMC's 16nm process
  - If you have been following along you can probably see why TSMC is a 500 billion dollar company
- Designed with datacenters in mind and can withstand 200W, but usually consumes half of that

https://habana.ai/
**GOYA™ IMAGE CLASSIFICATION ON RESNET-50**

- **GOYA**
  - Latency: 1.0ms
  - Images per second: 15,453

- **T4**
  - Latency: 26.0ms
  - Images per second: 5,013

**Goya Measurement**:
- Hardware: Goya HL-100 PCIe Card; CPU XEON E5
- Software: Ubuntu v18.04; SynapseAI v0.1.6
- Workload Implementation: Precision INT8; Batch size 10

**GPU Measurement**:
- Hardware Configuration: T4; Host Supermicro SYS-4029GP - TRT T4
- Software Configuration: TensorRT 5.1; Synthetic dataset; Container - 19.03-py3
- Workload Implementation: Precision INT8; Batch size 128

https://habana.ai/inference/
GOYA™ PERFORMANCE ON BERT

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Batch Size</th>
<th>Latency (ms)</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOYA</td>
<td>12</td>
<td>9.4</td>
<td>1,273</td>
</tr>
<tr>
<td>T4</td>
<td>16.3</td>
<td>736</td>
<td></td>
</tr>
<tr>
<td>GOYA</td>
<td>24</td>
<td>16.3</td>
<td>1,527</td>
</tr>
<tr>
<td>T4</td>
<td>32.4</td>
<td>738</td>
<td></td>
</tr>
</tbody>
</table>


**Goya Configuration**:
- Hardware: Goya HL-100; CPU Xeon Gold 6152@2.10GHz
- Software: Ubuntu v16.04.4, SynapseAI v0.2.0-1173
- Precision: 16-bit

**GPU Configuration**:
- Hardware: T4; CPU Xeon Gold 6154@3GHz/16GB/4 VMs
- Software: Ubuntu 18.04.2, x86_64-gnu, CUDA Ver 10.1, cudnn7.5, TensorRT 5.1.5.0
- Precision: 16-bit

**Source**: https://github.com/NVIDIA/TensorRT/tree/release/5.1/demo/BERT

https://habana.ai/inference/
HABANA: GAUDI

• Named after Antoni Gaudi, a famous Spanish architect

• (What’s up with all the Spanish names from an Israeli company)

HABANA: GAUDI

• The Gaudi processor released in June 2019
• Instead of focusing on inference like the Goya, Gaudi focuses on training
• It’s built on the same base architecture as Goya
  • Goya used TPC 1.0, Gaudi uses TPC 2.0
HBM2

10×100 GbE (RDMA)

GEMM Engine

Shared Memory Pool

TPC

TPC

TPC

TPC

TPC

TPC

PCle Gen 4×16

DMA

HBM2

https://en.wikichip.org/wiki/habana/microarchitectures/gaudi
HABANA: GAUDI, ADDITIONAL INFO

- High Bandwidth Memory 2 (HBM2)
  - Each one holds 8GB and combined provides 1TB of bandwidth
- Connect multiple Gaudi processors over ethernet
  - Not proprietary like competitors (NVLink from NVIDIA)
- Implements RDMA over converged ethernet (RoCE) directly on die along with the ports
  - Get faster speeds without using a RDMA controller and PCIe switch

https://habana.ai/training/
Gaudi® Performance
A single Gaudi card dissipating 140 Watts, delivers **1,650 images/second training throughput**. And, Gaudi’s training performance scales — from small scale servers to large-scale deployments— with record-breaking performance.

Gaudi® vs. V100
ResNet-50 Training Throughput at Scale

![Graph showing Gaudi vs. V100 performance](https://habana.ai/training/)

Based on nVIDIA reported MLPerf V0.5 performance metrics
HABANA: GAUDI, WAIT A MOMENT!

• You may have noticed the last slide compared Gaudi to a V100. What about an A100 (specifically, the new 80GB version released just a few days ago on 11/16/2020)

• Running ResNet-50 v1.5 the A100 processed 2,048 images per second
  • This is compared to Gaudi’s 1,650
  • It appears NVIDIA caught up to Habana this year on single processing unit computing power
HABANA: MORE THAN JUST A CHIP

• Habana aims to provide more than just Gaudi and Goya

• Habana developed SynapseAI, a suite of tools for compiling and running programs on their chips.
  • Native integration with Tensorflow (and more to come)
  • Python API to load any existing framework including ONNX
HABANA: WHAT’S NEXT?

• Intel announced they would stop moving forward with another AI chip company that they purchased called Nervana to focus on Habana.

• Graphcore (and our next company Cerebras) have a huge amount of positions available, while Habana does not have any on their website.

• Intel has started volume production and shipment of Habana products, but doesn’t have a new one in the pipeline.
QUESTIONS, DOC?

GUIDE

• 8.i Intel
• 8.j Graphcore
• 8.k Habana
• 8.l Cerebras
Cerebras was founded in 2016 in Menlo Park, California.

- Received a total of $112 million from series A, B, and C funding.
- An emerging semiconductor company.
- Around 200 employees.
- Has a lot of open positions on their website.
CEREBRAS: WSE

• 400,000 programmable Sparse Linear Algebra Compute cores
• Cores are connected in a 2D mesh with a total bandwidth of 100PB/s
• 18GB of SRAM all accessible within one clock cycle
  • Memory bandwidth of 9.5PB/s

https://www.cerebras.net/product/
CEREBRAS: WSE, DID YOU THINK I FORGOT?

- Yes, it was manufactured with TSMC’s 16nm process

https://robinhood.com/stocks/TSM
CEREBRAS: WSE, DID I MENTION IT’S BIG?

- 8.5in x 8.5in !!!!!!!!
- 1.2 trillion transistors
- Cerebras realized that to keep up with the rapidly increasing size of models, processors also had to increase the number of cores
  - Between 2012 and 2018 the compute power needed increased 300,000x
- NVIDIA has invested in NVLink to use multiple chips to increase the number of cores
- Cerebras took a different approach and tried to do more with just one chip because on chip communication is significantly faster

https://www.anandtech.com/show/15838/cerebras-wafer-scale-engine-scores-a-sale-5m-buys-two-for-the-pittsburgh-supercomputing-center
CEREBRAS: CS-1

- Comes with proprietary software that works with PyTorch and Tensorflow, among others
- Contains one WSE
- 15 rack units, 20kW power consumption
- 12 gigabit ethernet connections for 1.2TB/s IO
- 18GB memory can be a limit

CEREBRAS: CS-1, SALES

• The Pittsburgh Supercomputing Center bought a pair of CS-1’s in June 2020 for $5 million.
• Argonne National Laboratory deployed the first CS-1 in November 2019
• Cerebras claims to have sold dozens
CEREBRAS: CS-1, PERFORMANCE

• Cerebras claims the system has more compute power than 1,000 NVIDIA V100 GPUs at 1/40th the size and 1/50th the power
  • This hasn’t been put to the test as Cerebras has not run MLPerf on its system yet
  • Regardless of speed, it’s certainly easier to plug and play one device rather than 1,000

• Cerebras claims it has 3 times the performance of a Google TPU 2 Pod.
  • Google’s system takes up 10 racks and 5 times more power

• Definitely a competitor
CEREBRAS: WHAT’S NEXT?

• Cerebras teased the WSE 2 at the Hot Chips conference in August 2020

• Interesting to note that they are working with TSMC on the 7nm process rather than the 3nm process as Graphcore (a competitor) is

• My (personal) theory is that they will introduce a CS-2 with multiple WSE 2 chips.

https://www.anandtech.com/show/16006/hot-chips-2020-live-blog-cerebras-wse-programming-300pm-pt
QUESTIONS?

https://animaniacs.fandom.com/wiki/Yakko_Warner