Science and Technology Facilities Council

Technical Report
DL-TR-2022-003

AI chips technical comparison report

A Shaikh, S Thorne

August 2022

Enquiries concerning this report should be addressed to:

Chadwick Library
STFC Daresbury Laboratory
Sci-Tech Daresbury
Keckwick Lane
Warrington
WA4 4AD

Tel: +44(0)1925 603397
Fax: +44(0)1925 603779
email: librarydl@stfc.ac.uk

Science and Technology Facilities Council reports are available online at:
https://epubs.stfc.ac.uk

DOI: 10.5286/dltr.2022003

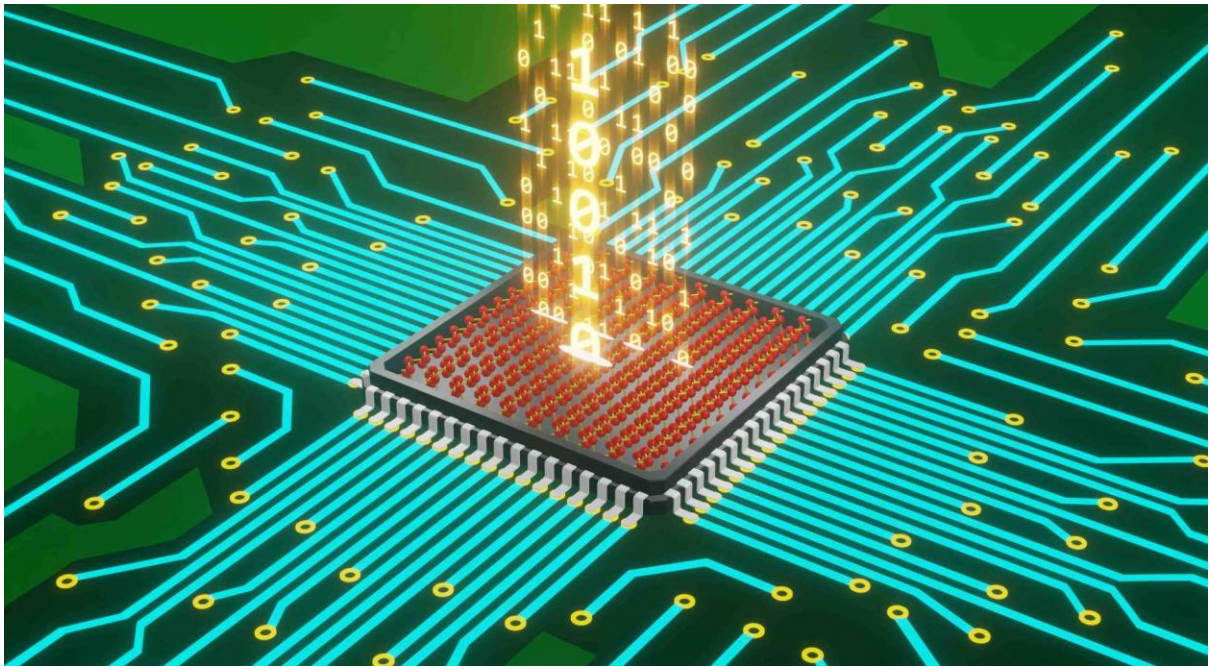
ISSN 1362-0207

AI Chips Technical Comparison Report

Aiman Shaikh

Sue Thorne

Hartree Centre



Pic credit : RMIT university.

Contents

Introduction

Artificial intelligence (AI) chips use semiconductors to provide powerful processors that can benefit areas needing high compute resource requirements such as climate, energy, health, and security. The term “AI chips” refers to a recent generation of microprocessors designed specifically to process artificial intelligence tasks faster. AI chips are comprehensive silicon chips, which integrate AI technology and are used for machine learning. (Viswanathan, 2020)

In the last decade, there have been numerous advancements in the field of deep learning technology. Since 2013, various novel AI chips have been developed along with products based on these chips (Momose, 2020).

General-purpose chips like central processing units (CPUs) can also be used for some simpler AI tasks, but CPUs are becoming less and less useful as AI advances (Saif M. Khan, 2020). AI chips include graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs) that are specialized for AI.AI chips include graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs) that are specialized for AI.

Graphical Processing Unit (GPUs)
GPUs were originally designed to process graphic intensive tasks such as games. GPUs are designed to handle parallelism and provide high performance, which is required for deep learning AI algorithms due to parallelism. GPUs makes a great AI hardware and are becoming popular for use in creative production and AI.

Field Programmable Gate Array (FPGA)
FPGAs are programmable arrays that can be reprogrammed based on the requirements. FPGAs are integrated circuit silicon chips which have an array of logic gates: this array can be programmed in the field i.e., the user can overwrite the existing configurations with their new defined configurations and can create their own digital circuit. FPGAs are costly due to their flexibility. (Pandit, 2019)

Application Specific Integrated Circuits (ASIC)
ASIC chips are exclusively designed for AI applications and are integrated with AI algorithms. There are different types of ASIC-based AI chips.

This report presents a technical comparison and programming model specification for AI chips such as Graphcore, Cerebras , SambaNova , and GPUs by Nvidia , Intel, AMD and also Google TPUs. This is a continuous ongoing work with an aim to evaluate as many AI chips as possible. Only Cerebras and Graphcore and Nvidia GPus are available at the time of writing this. **This report by no means favours any vendor and is vendor agnostic.**

Technical Comparison Table

AI Chips

Server	Chip	Programming model/Software	Memory	Networking
IPU Machine: M2000	4 x Mk2 GC200 IPU	Poplar SDK	450 GB exchange memory	IPU Fabric
CS-2	WSE2	ML frameworks compiled by Cerebras Graph Compiler (CGC)	40 Gb	SwarmX Interconnect
SN10-8R sytem	8 RDU	SambaFlowTM	12 TB / 48 DDR4	32 PCIe-Gen4

HPE Superdome	Intel Xeon	Red Hat Enterprise Linux (RHEL) • SUSE Linux Enterprise Server (SLES) • Oracle Linux/Oracle UEK • Oracle VM • VMware •	• 48 DIMM slots of DDR4 memory 768Gb – 48TB of shared memory	InfiniBand EDR/Ethernet 100Gb; Infiniband HDR

GPUs

GPU	GPU Type	Software	Programming	Memory	Networking
NVidia	DGX-A100 320	NVIDIA CUDA-X and DGX software stack	CUDA, C++, OpenCL, OpenACC	320 GB HBM2	Mellanox Infiniband PCIe 4+ support
NVidia	DGX-A100 640GB	NVIDIA CUDA-X and DGX software stack	CUDA, C++, OpenCL, OpenACC	640 GB HBM2	Mellanox Infiniband PCIe 4+ support
NVidia	DGX-SuperPod	NVIDIA CUDA-X and DGX software stack	CUDA, C++, OpenCL, OpenACC	49 TB HBM2	Mellanox Infiniband PCIe 4+ support
AMD	Radeon Instinct M125 AMD Vega10	Rocm	ISO C++, OpenCL™, CUDA (via AMD's HIP conversion tool) and Python5 (via Anaconda's NUMBA)	16GB HBM2	PCIe 4.0 + infinity Fabric
AMD	Radeon Instinct M150 Vega20	Rocm	ISO C++, OpenCL™, CUDA (via AMD's HIP conversion tool) and Python5 (via Anaconda's NUMBA)	32 GB HBM2	PCIe 4.0 + infinity Fabric
Intel	Ponte Vecchio Xe HPC	OneAPI	OneAPI supported*	/	/

/ More information will be added once available.

Graphcore

Graphcore has created an AI chip it calls an **i**ntelligence **p**rocessing **u**nit (**IPU**) that sacrifices a certain amount of arithmetic precision to allow the machine to be able to do more energy efficient mathematics. They recently released the Colossus MK2, and packaged four of them into a machine called the IPU-M2000, which is about the size of a DVD player. The IPU-M2000 is quoted as providing one petaflop of computing power. We note that one of the world's fastest supercomputers, Japan's Fugaku, is rated at as 442 petaflops. Some users believe that Graphcore could compete with Nvidia in the supercomputer chip race. In a test using a state-of-the-art image classification benchmark, eight of Graphcore's new IPU-M2000 clustered together could train an algorithm at a cost of $259,000 compared to $3 million for 16 of Nvidia's DGX clusters. IPU-Machine M2000 has off-chip DDR memory. However, there is no cache or anything in the hardware to automatically control at runtime the moving or buffering of data between the external streaming memory and on-chip in-processor memory. It is all controlled in software based on the computation graph. Memory management is just one of the parts of the software stack where optimisation of the hardware is based on advanced analysis. This is key to their approach (Lacey, 2020).

Types of memory

1. Streaming memory
2. In-Processor memory

In-Processor Memory is on the IPU and can access to Streaming Memory outside of the chip. Like many modern processors, the IPU can deal with a layered memory hierarchy. Since IPUs work together in many-chip computations, the best way to think about it is the memory specification of a system, for example the IPU-Machine M2000 Each IPU-Machine has up to 450GB of memory addressable by the 4 IPUs. This is split into the 900MB per IPU of In-Processor Memory and up to 112GB per IPU of Streaming Memory. The Streaming Memory is contained in DDR4 DIMMs on the IPU-Machine 1U Server (in the same server as the IPUs) and the exact amount of memory will depend on the size of those DIMMs (dual in-line memory module or RAM) (Fig 1)

Intelligent Variable Placement in ML Frameworks

At the ML Framework (e.g. TensorFlow) level, Graphcore uses intelligent variable placement in the computation graph when they are in Streaming Memory and when they are in In-Processor Memory.

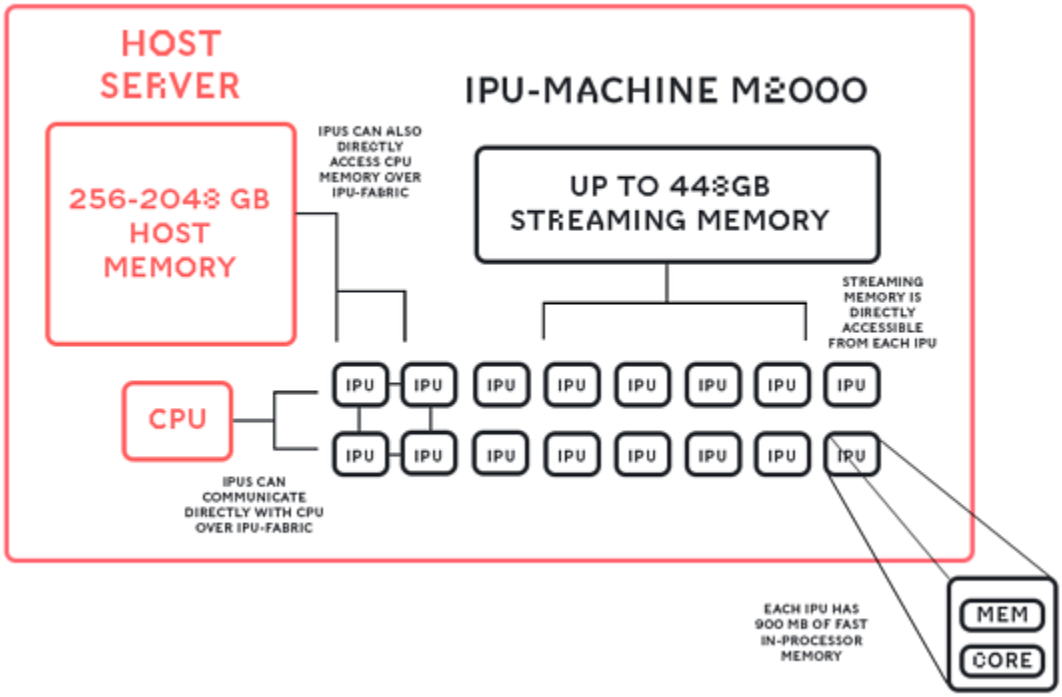


Fig 1. IPU-Machine Memory System (Lacey, 2020)

Photo credit : STH (GC200 chip at launch).

SambaNOva

SambaNova has taken a hybrid approach to AI chip design by giving equal emphasis to both the hardware and software. There are only a certain number of AI chip start-ups that have stuck it out and continued to focus on the datacentre. Lawrence Livermoore National Laboratory recently purchased a SambaNova system and, whilst hardware performance is paramount (and gets all the attention) for the LLNL team, it is SambaNova's software environment that appears to have made the difference in the competitive bid. The team could not tell us who else responded to the initial request for procurement for AI accelerators that could be incorporated into HPC systems.

SambaNova Reconfigurable Dataflow ArchitectureTM

The SambaNova Reconfigurable Dataflow Architecture™ (RDA) is a computing architecture designed to enable the next generation of machine learning and high-performance computing applications. The Reconfigurable Dataflow Architecture is a complete, full-stack solution that incorporates innovative algorithms, compilers, system architecture and silicon. The RDA provides a flexible, dataflow execution model that pipelines operations, enables programmable data access patterns and minimizes excess data movement found in fixed, core-based, instruction set architectures. It does not have a fixed Instruction Set Architecture (ISA) like traditional architectures. The Reconfigurable Dataflow Architecture is composed of the following:

SambaNova Reconfigurable Dataflow UnitTM is a next-generation processor designed to provide native dataflow processing and programmable acceleration. It has a tiled architecture that contains a network of reconfigurable functional units.

SambaFlowTM is a complete software stack designed to take input from standard machine-learning frameworks such as PyTorch and TensorFlow. SambaFlow automatically extracts, optimizes, and maps dataflow graphs onto RDUs, allowing high performance to be obtained without the need for low-level kernel tuning. SambaFlow also provides an API for expert users and those who are interested in using the RDA for more classical workloads.

SambaNova Systems DataScaleTM is a rack-level, data-centre-ready accelerated computing system. Each DataScale system configuration consists of one or more DataScale nodes, integrated networking, and management infrastructure in a standards-compliant data centre rack, referred to as the SN10-8R.

Flexibility and Reconfigurability with Dataflow

The SambaFlow optimizations described above and programmability of the RDU allow it to be optimized and configured for a variety of workloads across machine learning, scientific computing, and other data-intensive applications. Rapid reconfiguration enables the architecture to be quickly repurposed for new needs or to adapt to the latest algorithm breakthroughs. SambaNova claims that the RDUs can be reconfigured in microseconds, which is a significantly lower amount of time than a FPGA. (SambaNova, 2021)Contrast that with software 2.0 (Karpathy, 2017) where the idea is that you train neural networks As an example, Olukotun cites the Google Translate service, which Google reduced from 500,000 lines using training data, and the program is written in the weights of the neural network. This has several advantages, and the key one is that you have a reduced number of lines of

code that must be explicitly developed by the programmer. (Karpathy, 2017) SambaNova Systems has developed a new computing architecture called 'Reconfigurable Dataflow Architecture' (RDA), built to support Software 2.0 and bring machine learning to all types of dataflow computation problems.

Cerebras

EPCC, the supercomputing centre of the University of Edinburgh, has deployed a Cerebras CS-1 supercomputer for AI-based research. The Cerebras CS-1 system uses the large Wafer Scale Engine (WSE) processor alongside an HPE Superdome Flex Server system for front-end storage and pre-processing, which, when combined, the company says will greatly reduce training time for AI models. The first deployment of a CS-1 system in Europe will be used for natural language processing and data science research across public, private, and academic organizations.

"We are proud to announce this audacious infrastructure investment and partnership with the world leaders in AI computing," said EPCC Director Professor Mark Parsons. "This installation will enable massive breakthroughs in our vision for data science and greatly accelerate our research across genomics and public health, including time-sensitive and pressing issues such as leveraging AI across large models to advance Covid-19 therapeutic research."

"We are excited to bring our industry-leading CS-1 AI supercomputer, coupled with HPE's advanced memory server, to EPCC and the European market to help solve some of today's most urgent problems," said Andrew Feldman, CEO and co-founder of Cerebras. "Our vision with the CS-1 was to reduce the cost of curiosity, and we look forward to the myriad experiments and world-changing solutions that will emerge from EPCC's regional data center."

The WSE measures at 46.2 cm^2 (7.1 $inches^2$), which Cerebras claims is 56 times larger, has 54 times more cores, 450 times more on-chip memory, 5,788 times more memory bandwidth and 20,833 times more fabric bandwidth than the leading graphics processing unit (GPU). The HPE Flex Server on this supercomputer will reportedly be provisioned with 18TB of memory, 102TB of flash storage, 24 Intel Xeon CPUs, and 12 network interface cards to deliver 1.2 Tbps of data bandwidth to the CS-1.

"HPE has a long-standing collaboration with EPCC to develop solutions to some of the most challenging computational problems, and we are excited to be working at this time to provide a highly productive AI platform," said Mike Woodacre, HPE CTO of HPC.

"By tightly coupling a Cerebras Wafer Scale Engine with an HPE Superdome Flex Server In-Memory host, we are aiming to enable researchers to tackle complex AI workloads at unprecedented rates."

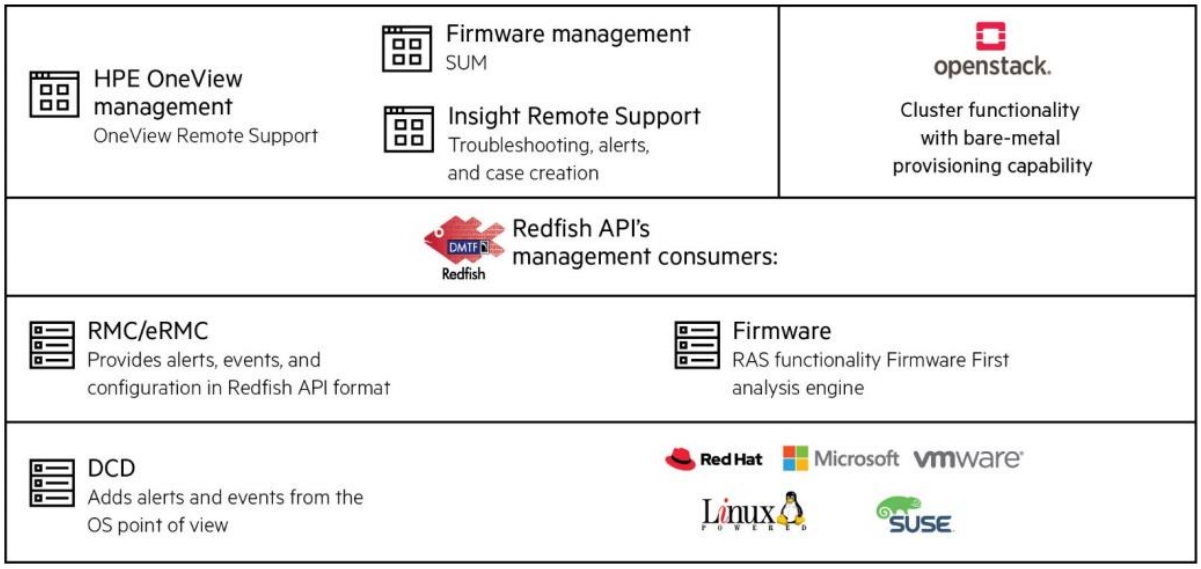
Last year, the Pittsburgh Supercomputing Center chose Cerebras and HPE for its Neocortex supercomputer. In 2020, GlaxoSmithKline (GSK) began using the Cerebras CS-1 AI system in their London AI hub, for neural network models to accelerate genetic and genomic research and reduce the time taken in drug discovery. (Ray, 2020) The GSK research team was able to increase the complexity of the encoder models they could generate, while reducing training time. Other pharmaceutical industry customers include AstraZeneca, who was able to reduce training time from two weeks on a cluster of GPUs to two days using the Cerebras CS-1 system. (Hansen, 2021)The GSK research team was able to increase the complexity of the encoder models they could generate, while reducing training time. Other pharmaceutical industry customers include AstraZeneca: it is claimed that they were able to reduce training time from two weeks on a cluster of GPUs to two days using the Cerebras CS-1 system. (Hansen, 2021)

Argonne National Laboratory has also been using the CS-1 since 2020 in COVID-19 research and cancer tumour research based on the world's largest cancer treatment database. A

series of models running on the CS-1 to predict cancer drug response to tumours achieved speed-ups of many hundreds of times on the CS-1 compared to their GPU baselines.
The Lawrence Livermore National Laboratory's Lassen supercomputer incorporated the CS-1 in both classified and non-classified areas for physics simulations.

HPE Superdome Flex

HPE Superdome Flex Server is described as a compute breakthrough that can power critical applications, accelerate data analytics, and tackle high-performance computing (HPC) and artificial intelligence (AI) workloads holistically. It aims to provide a combination of flexibility, performance, and reliability for environments of any size by using a large-scale modular architecture. Leveraging its in-memory design and ground-breaking performance, your business can process and analyse growing quantities of data at extraordinary speed. HPE Superdome Flex provides management tools that work together to configure resources, monitor the system, and send alerts. (Hewlett Packard Enterprise Development LP, 2020)



Nvidia

NVIDIA claim that the DGX™ A100 is a “universal system for all AI workloads, offering unprecedented compute density, performance, and flexibility in the world’s first 5 petaFLOPS AI system”. NVIDIA DGX A100 features the NVIDIA A100 Tensor Core GPU, “enabling enterprises to consolidate training, inference, and analytics into a unified, easy-to-deploy AI infrastructure that includes direct access to NVIDIA AI experts.”

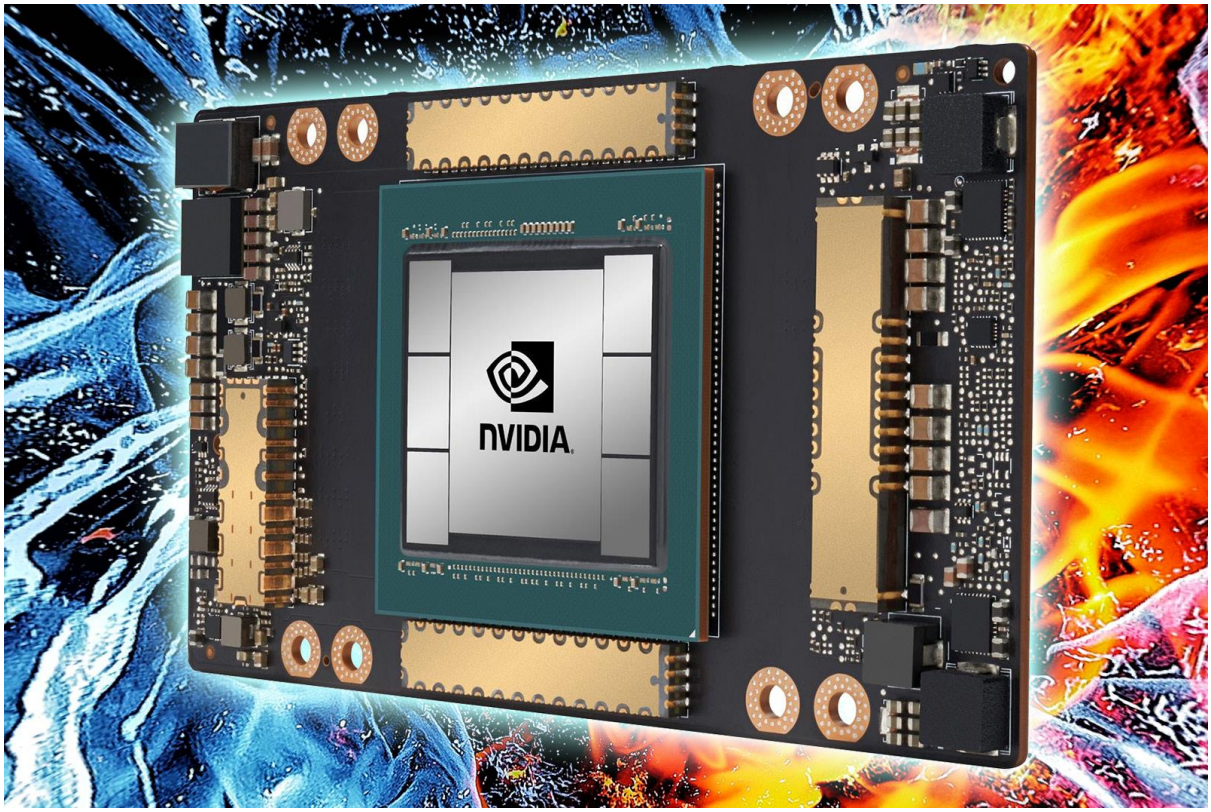
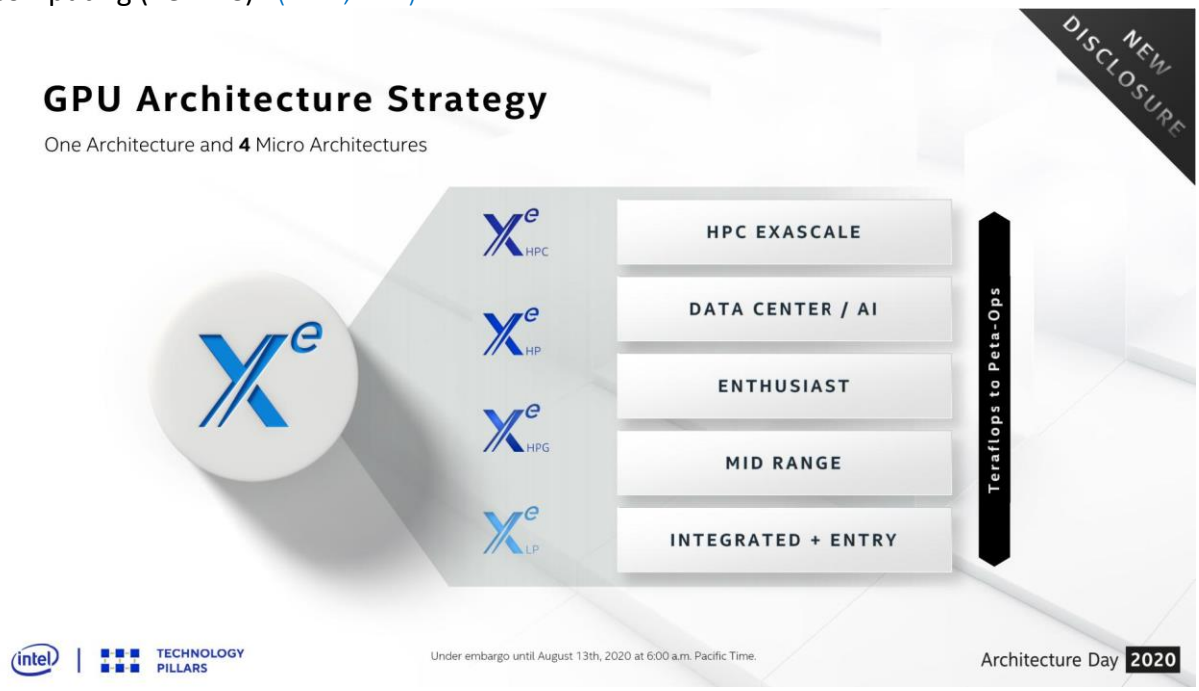


Photo credit : Spiria. Nvidia DGX A100

Intel

Intel has recently entered the GPU market with the Intel Xe, earlier known unofficially as Gen12. The Intel Xe includes a new instruction set architecture. The Xe GPU family consists of a series of microarchitectures, ranging from integrated/low power (Xe-LP) to high performance gaming (Xe-HPG), datacentre/high performance (Xe-HP) and high-performance computing (Xe-HPC) . (Intel, n.d.)



Besides being bigger, Xe HPC also supports additional data types. The Xe HPG MXM only works on FP16 and BF16 data, but the Xe HPC also supports the TF32 (Tensor Float 32). The vector engine also adds support for FP64 data, though only at the same rate as FP32 data. For the matrix engines, each Xe-core can do 4096 FP16 or BF16 operations per clock cycle, 8192 INT8 ops per clock cycle, or 2048 TF32 operations per clock cycle. (Walton, 2021)
The first test is the “Aurora” A21 supercomputer at Argonne National Laboratory, has tens of thousands of these Ponte Vecchio GPU accelerators installed and representing the vast majority of the floating-point processing power in the system. (Intel, 2020)

AMD (Advanced Micro Devices)

AMD’s Vega 10 GPU chips will be packaged in AMD Radeon Instinct graphics add-in board (AIB) form factors that plug into PCIe Gen 3 x16 slots. AMD’s Vega 10 GPU chip has 64 graphics cores, each core containing 64 stream processors, giving a total of 4,096 stream processors. AMD Vega 10 and the Radeon Instinct family of AIBs will continue support for AMD’s Multiuser GPU (MxGPU) virtualization technology. The Vega 10 architecture is notable for its implementation of packed 16-bit floating point (FP16) instructions. Packing means that AMD implemented a little extra logic to allow two FP16 operations in the same execution pipeline as a single 32-bit floating point (FP32) operation. This can double the performance of algorithms that do not require full FP32 precision. It is claimed that a single

Radeon Instinct MI25 (“MI” stands for “machine intelligence”) server GPU AIB can execute a total of 12 FP32 TFLOPS or 25 FP16 TFLOPS.
AMD is addressing heterogeneous programming in two ways. First, with coherent memory, programming is more straightforward, with fewer lines of code or custom calls. Second, AMD supports an open “ROCm™” programming and software environment, which aims to make applications more portable to future generations of processors, regardless of where they come from. (AMD, 2021)
ROCm™ uses HIP programming model. (AMD, 2022) The Heterogeneous-Computing Interface for Portability (HIP) provides a C++ syntax that aims to be suitable for compiling most code that commonly appears in compute kernels, including classes, namespaces, operator overloading, templates and more. Additionally, it defines other language features designed specifically to target accelerators.
Porting the CHOLLA code over from existing NVIDIA CUDA platform to AMD ROCm™ open software platform and both AMD and Radeon Instinct™ GPUs proved easy and provided immediate performance gains. “We got most of the porting to HIP to run on AMD hardware done in a few hours,” says Budiardja. “With the AMD Radeon Instinct MI50, we've got relatively similar performance to the NVIDIA Tesla V100. With the AMD Instinct MI100, we realized about 1.4x speed up without doing anything at all. We just changed to compile to ROCm to get the code to run. There is a lot of benefit in HIP using similar function calls. If you're already familiar with CUDA, you can take one look at the HIP function and realize what it's doing.” (AMD + ORNL, 2021)

Google TPUs

Tensor Processing Units (TPUs) are Google's custom-developed application-specific integrated circuits used to accelerate machine learning workloads. Cloud TPU enables the user to run machine learning workloads on Google's TPU accelerator hardware using TensorFlow. Google provide high-level TensorFlow APIs to help the user get models running on the Cloud TPU hardware.

Cloud TPU programming model

Cloud TPUs are very fast at performing dense vector and matrix computations. Transferring data between Cloud TPU and host memory is slow compared to the speed of computation—the speed of the PCIe bus is much slower than both the Cloud TPU interconnect and the on-chip high bandwidth memory (HBM). Partial compilation of a model, where execution passes back and forth between host and device causes the TPU to be idle most of the time, waiting for data to arrive over the PCIe bus. To alleviate this situation, the programming model for Cloud TPU is designed to execute much of the training on the TPU—ideally the entire training loop. (Google TPU, n.d.)

Test Case Used

I have been evaluating a test case provided by UKAEA. More information on the test case will be added once the test case is published and a license is in place. This work of evaluating AI chips is part of collaboration with UKAEA.
The test case provided belongs to the category of Physics Informed Neural Network (PINN). The specific test case uses a 2D approach. PINN models are crafted to solve the 2D wave equation without using any data but just the PDE constraints.
Deep learning has been shown to be an effective tool in solving partial differential equations (PDEs) through physics-informed neural networks (PINNs). PINNs embed the PDE residual into the loss function of the neural network, and have been successfully employed to solve diverse forward and inverse PDE problems. (Jeremy Yu, 2022)

Investigating Cerebras

We have been working closely with vendor. After initial evaluation, we did find some bugs that were fixed by their internal systems team. To use the Cerebras system, we found that we needed to make many changes to the test problem.
The test case is available in both Tensorflow and PyTorch. Using the Tensorflow version, the test case model is not able to be executed on Cerebras system due to the Cerebras SDK not supporting the gradient function: without the gradient functions the PINN falls apart as it relies on the gradient function to approximate the derivatives. Cerebras SDK has custom gradient functions which are not helpful as using them would require drastic changes in the given test case.
Using the PyTorch version, the test case makes use of torch.autograd.grad. Again, the gradient functions are not supported, which is an issue.
According to the Cerebras team, there is a future roadmap to support these important mentioned functions. Considering the drastic changes that are required to make the model work, the Cerebras investigation has been paused here.

Investigating Graphcore

With Graphcore IPUs, we are working closely with the vendor. We had direct bare metal access to IPUs and are now sorting out access through cloud under new agreement. With Graphcore, the model has not needed many drastic changes.

The initial results from Graphcore looks quite promising and will be published once test case license is in place.

Investigating GPUs

We will publish GPUs comparison with Graphcore once the test case license is in place.

Bibliography

AMD + ORNL, 2021. *Cholla CAAR team plans to simulate the entire Milky Way using AMD Instinct™ accelerators,* s.l.: s.n.

AMD, 2021. *NEW AMD CPUs and GPUs CONTINUE,* s.l.: Intersect360.

AMD, 2022. *New AMD ROCm™ Information Portal - ROCm v4.5 and Above,* s.l.: Sphinx.

Google TPU, n.d. *Cloud TPU.* [Online]
Available at: https://cloud.google.com/tpu

Hansen, L. L., 2021. *Accelerating Drug Discovery Research with New AI Models: a look at the AstraZeneca Cerebras collaboration.* [Online]
Available at: https://larslynnehansen.medium.com/accelerating-drug-discovery-research-with-new-ai-models-a-look-at-the-astrazeneca-cerebras-b72664d8783
[Accessed July 2022].

Hewlett Packard Enterprise Development LP, 2020. [Online]
Available at: https://www.hpe.com/psnow/doc/a50000335enw?jumpid=in_lit-psnow-red
[Accessed 2022].

Intel, 2020. *Intel.com.* [Online]
Available at: https://www.intel.com/content/www/us/en/newsroom/news/argonne-national-laboratory-quantum-research.html#gs.6wnxdm
[Accessed 2022].

Intel, n.d. [Online]
Available at: https://www.intel.com/content/www/us/en/high-performance-computing/hpc-products.html

Jeremy Yu, L. L. X. M. G. E. K., 2022. Gradient-enhanced physics-informed neural networks for forward and inverse PDE problems. *Scince Direct,* Volume 393.

Karpathy, A., 2017. *Karpathy.* [Online]
Available at: https://karpathy.medium.com/software-2-0-a64152b37c35
[Accessed July 2022].

Lacey, D., 2020. *Intelligent Memory for intelligent computing.* [Online]
Available at: https://www.graphcore.ai/posts/intelligent-memory-for-intelligent-computing
[Accessed July 2022].

Lacey, D., 2020. *The future of AI chips lies in software.* [Online]
Available at: https://www.graphcore.ai/posts/the-future-of-ai-chips-lies-in-software
[Accessed July 2022].

Momose, H., 2020. Systems and circuits for AI chips and their trends. *Japanese Journal of Applied Physics,* Volume 59, p. 16.

Pandit, A., 2019. *Circuit Digest.* [Online]
Available at: https://circuitdigest.com/tutorial/what-is-fpga-introduction-and-programming-tools

Ray, T., 2020. *Glaxo's biology research with novel Cerebras machine shows hardware may change how AI is done.* [Online]
Available at: https://www.zdnet.com/article/glaxos-biology-research-with-novel-cerebras-machine-shows-hardware-may-change-how-ai-is-done/
[Accessed July 2022].

Saif M. Khan, A. M., 2020. AI Chips: What They Are and Why They Matter. April, p. 72.

SambaNova, 2021. *AI is changing everything,* s.l.: Situation Publishing.
Viswanathan, S. M., 2020. AI Chips: New Semiconductor Era. *Research Gate,* 7(8), p. 8.
Walton, J., 2021. *Intel Ponte Vecchio and Xe HPC Architecture: Built for Big Data.* [Online]
Available at: https://www.tomshardware.com/features/intel-ponte-vecchio-and-xe-hpc-architecture-built-for-big-data
[Accessed 2022].