Science and Technology Facilities Council

Technical Report
DL-TR-2022-002

AI chips technical comparison report

A Shaikh, S Thorne

May 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.2022002

ISSN 1362-0207

AI Chips Technical Comparison Report

Miss Aiman Shaikh
Dr. Sue Thorne
Hartree Centre.

Contents

Introduction

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.
AI chips include graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs) that are specialized for AI. 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 (1).

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*	??	??

Graphcore

Graphcore has created an AI chip it calls an **i**ntelligence **p**rocessing **u**nit (**IPU**) that sacrifices a certain amount of number-crunching precision to allow the machine to tackle more math more quickly with less energy. This year it threw down the gauntlet to Nvidia when it released its latest IPU, the Colossus MK2, and packaged four of them into a machine called the IPU-M2000. About the size of a DVD player. The IPU-M2000 packs one petaflop of computing power. One petaflop is a quadrillion calculations per second, and the world's fastest supercomputer, Japan's Fugaku, is rated at a world-record 442 petaflops. Fortune reported earlier this year how 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, each of which contains eight of the company's top-of-the-line chips. 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 approach (2).

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, intelligent variable placement chooses when variables in the computation graph are in Streaming Memory and when they are in In-Processor Memory.

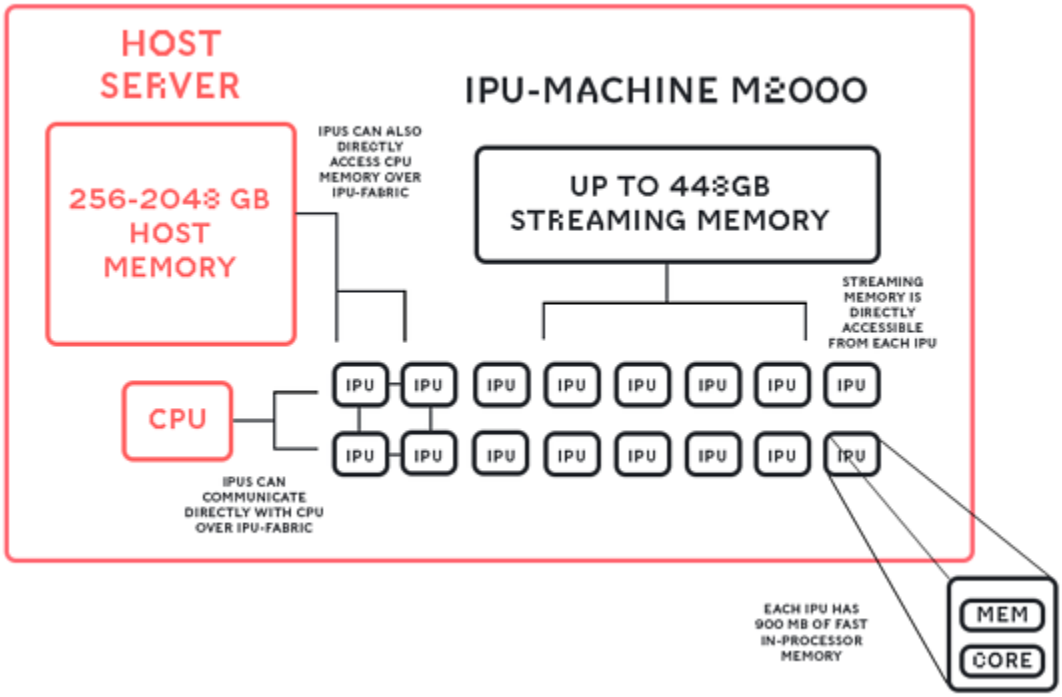


Fig 1. IPU-Machine Memory System (3).

SambaNOva

SambaNova has taken what's been called 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 startups that have stuck it out and continued to focus on the datacenter. While 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 couldn't tell us who else responded to the initial RFP for AI accelerators that could snap into HPC systems, but we have to figure that if the conversation turned to software as much as it did, it was far and away the differentiating factor over startups(4).

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 innovations at all layers including algorithms, compilers, system architecture and state-of-the-art 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, but instead is programmed specifically for each model resulting in a highly optimized, application-specific accelerator. 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 comprises a network of reconfigurable functional units. The architecture enables a broad set of highly parallelizable patterns contained within dataflow graphs to be efficiently programmed as a combination of compute, memory, and communication networks.
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 leveraging the RDA for workloads beyond machine learning.
SambaNova Systems DataScaleTM is a complete, rack-level, data-center-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 center 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. These are key advantages over fixed ASIC designs that can require years to develop and cannot be modified for algorithm changes or different workloads. At the other end of the spectrum, are FPGAs, which are highly reconfigurable. In contrast to the time-consuming, complex, low-level programming and long compilation times of FPGAs, RDUs can be reconfigured in microseconds. This level of flexibility and reconfigurability gives programmers the ability to work in high-level DSLs while providing enhanced execution efficiency, simplified compilation, and performance. (5)

Contrast that with software 2.0 (6), 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.

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.

Deployment of system

This is how the SambaNova DataScale platform has been deployed at one early customer, the Lawrence Livermore National Laboratory (LLNL). Here, the Corona supercomputing cluster, which boasts more than 11 petaflops of peak performance, has been integrated with a SambaNova DataScale SN10-8R system.

SambaNova said a single DataScale SN10-8R can train terabyte-sized models, which would otherwise need eight racks worth of Nvidia DGX A100 systems based on GPUs. This convergence of HPC and AI supports SambaNova's assertion that its architecture based on dataflow does not merely accelerate AI functions but represents the next generation of computing. (That said, the company does not expect this new breed of computing to replace CPU-based systems for more transaction-oriented applications.)

Cerebras

EPCC, the supercomputing center of the University of Edinburgh, has depoloyed 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 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 is a (5) measuring 46.2 sqcm (7.1 sq inches) 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 (6) 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. (7)

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.(8)

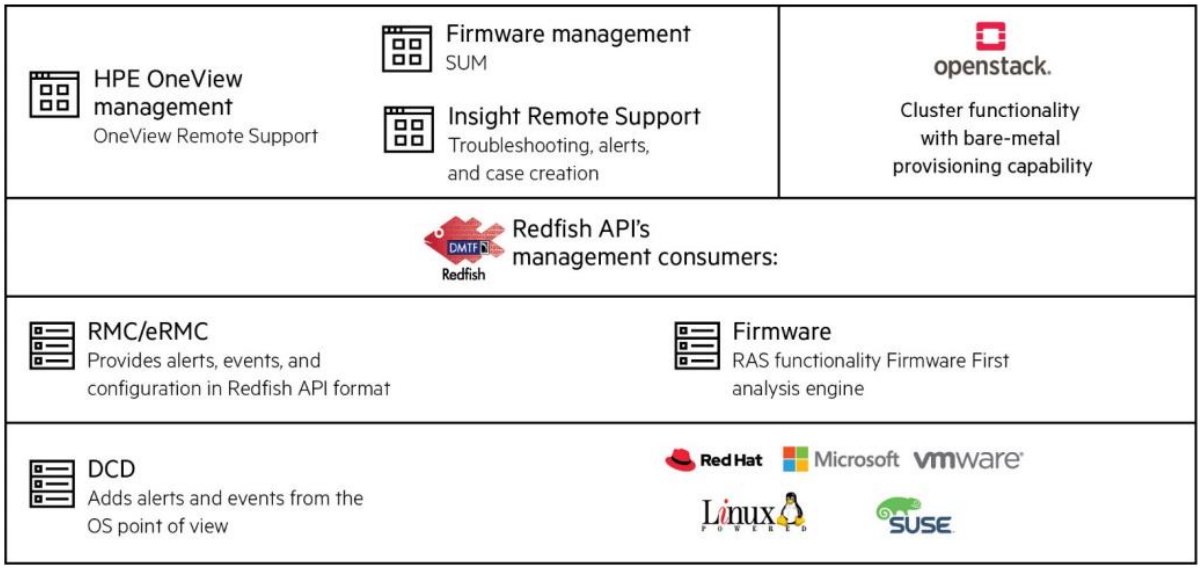
Argonne National Laboratory has been using the CS-1 since 2020 in COVID-19 research and cancer tumor research based on the world's largest cancer treatment database. A series of models running on the CS-1 to predict cancer drug response to tumors achieved speed-ups of many hundreds of times on the CS-1 compared to their GPU baselines.

The Lawrence Livermore National Lab'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 a compute breakthrough that can power critical applications, accelerate data analytics, and tackle high-performance computing (HPC) and artificial intelligence (AI) workloads holistically. It delivers an unmatched combination of flexibility, performance, and reliability for critical environments of any size. A unique modular architecture and unparalleled scale allow you to start small and grow at your own pace. 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 safeguards these vital workloads with superior RAS and end-to-end security. Meanwhile, HPE Pointnext, broad partner ecosystem, and mission-critical expertise complement the capabilities and value of the platform to help ensure your move to the HPE Superdome Flex is a success. (9)

HPE Superdome Flex provides numerous management tools that work in concert to configure resources, monitor the system, and send alerts. These tools help identify issues and resolve problems through numerous interfaces—from intuitive interactive graphical user interfaces (GUIs) to comprehensive command line interfaces (CLIs) that can be automated using a range of scripting languages. HPE Superdome Flex communicates management actions through industry-standard Redfish APIs, enabling industry-standard OpenStack cluster management. These APIs also enhance the built-in, mission-critical functionality of the HPE Superdome Flex Server and HPE standard environments using HPE OneView and Insight Remote Support.(10)



Nvidia

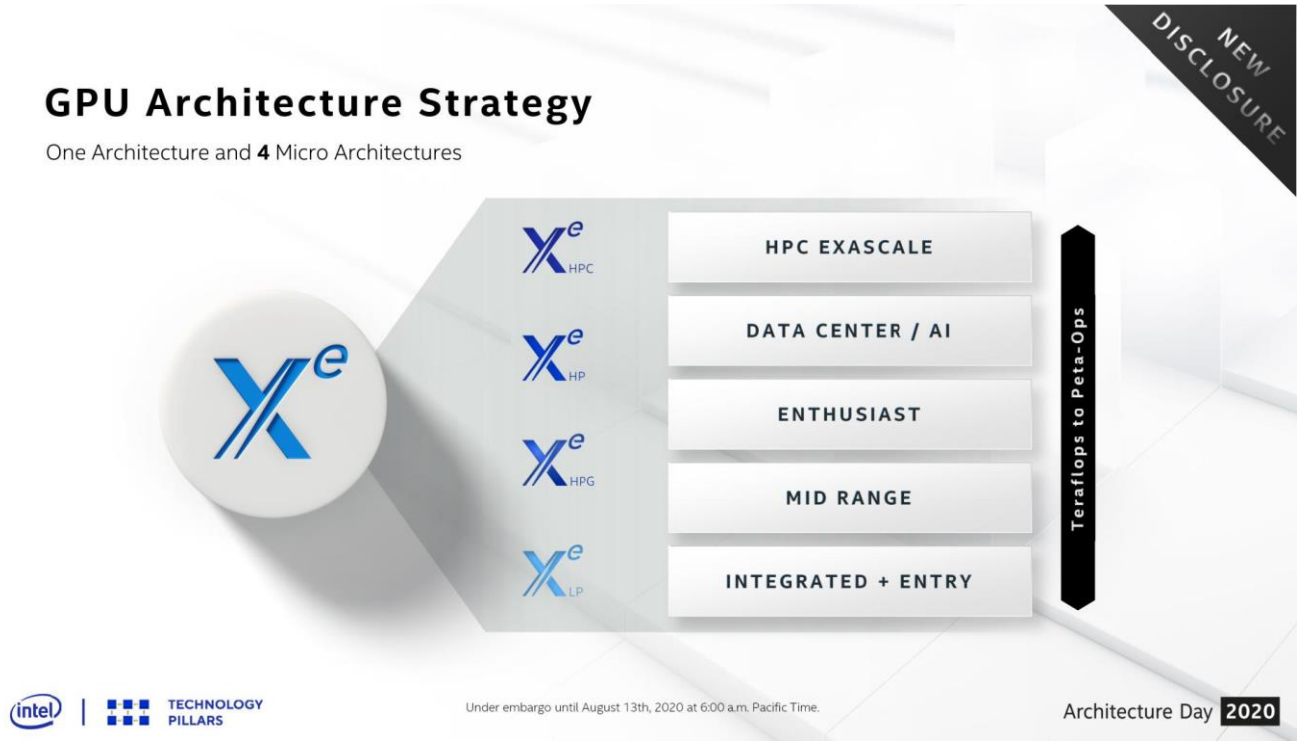
NVIDIA DGX™ A100 is the universal system for all AI workloads, offering unprecedented compute density, performance, and flexibility in the world's first 5 petaFLOPS AI system. NVIDIA DGX A100 features the world's most advanced accelerator, the NVIDIA A100 Tensor Core GPU, enabling enterprises to consolidate training, inference, and analytics into a unified, easy-to-deploy AI infrastructure that includes direct access to NVIDIA AI experts.(11)

NVIDIA DGX SuperPOD™ with NVIDIA® BlueField® data processing units (DPUs) and NVIDIA Base Command™ ushers in the era of cloud-native supercomputing, bringing together leadership-class infrastructure with agile, scalable performance for the most challenging AI and high-performance computing (HPC) workloads. DGX SuperPOD gives modern enterprises a secure, multi-tenant data center platform on which IT can deliver performance without compromise for every user and workload. (12)

Intel

Intel Xe (stylized as Xe and pronounced as two separate letters), earlier known unofficially as Gen12, is GPU architecture developed by Intel.

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),[5] to enthusiast/high performance gaming (Xe-HPG), datacenter/high performance (Xe-HP) and high performance computing (Xe-HPC) (12)



Intel has powered on the “Ponte Vecchio” Xe HPC GPU accelerator A0 silicon and it seems to be working perfectly. Like the built-for-gaming Xe HPG, the base building block for Xe HPC starts with the Xe-core. There are still eight vector engines and eight matrix engines in the Xe-core, but this Xe-core is fundamentally very different from Xe HPG. The vector engine uses a 512-bit register (for 64-bit floating point), and the XMX matrix engine has been expanded to 4096-bit chunks of data. That's double the potential performance for the vector engine and quadruple the FP16 throughput on the matrix engine. L1 cache sizes and load/store bandwidth have similarly increased to feed the engines.
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), which has gained popularity in the machine learning community. The vector engine also adds support for FP64 data, though only at the same rate as FP32 data.

With eight vector engines per Xe-core, the total potential throughput for a single Xe-core is 256 FP64 or FP32 operations, or 512 FP16 operations on the vector engine. For the matrix engines, each Xe-core can do 4096 FP16 or BF16 operations per clock, 8192 INT8 ops per clock, or 2048 TF32 operations per clock. But of course, there's more than one Xe-core in Ponte Vecchio. (13)

The first test is the “Aurora” A21 supercomputer being installed at Argonne National Laboratory, which will have tens of thousands of these Ponte Vecchio GPU accelerators installed and representing the vast majority of the floating point processing power in the system, which was expected to have at least 1.1 exaflops and which is rumoured to have 1.3 exaflops of aggregate compute. (15)

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, for 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.5 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 don't require full FP32 precision. 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™" (pronounced "rock 'em") programming and software environment, making applications more portable to future generations of processors, regardless of where they come from. (16)

AMD ROCm is the first open-source software development platform for HPC/Hyperscale-class GPU computing. AMD ROCm brings the UNIX philosophy of choice, minimalism, and modular software development to GPU computing. (17) ROCm™ has HIP programming model. Heterogeneous-Computing Interface for Portability (HIP) provides a C++ syntax that is 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.(18)

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. (19)

Google TPUs

Tensor Processing Units (TPUs) are Google's custom-developed application-specific integrated circuits (ASICs) used to accelerate machine learning workloads. TPUs are designed from the ground up with the benefit of Google's deep experience and leadership in machine learning.

Cloud TPU enables you to run your machine learning workloads on Google's TPU accelerator hardware using TensorFlow (20). Cloud TPU is designed for maximum performance and flexibility to help researchers, developers, and businesses to build TensorFlow compute clusters that can leverage CPUs, GPUs, and TPUs. High-level TensorFlow APIs help you to 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. (21)

Following are some salient features of the TPU programming model :

- All model parameters are kept in on-chip high bandwidth memory.
- The cost of launching computations on Cloud TPU is amortized by executing many training steps in a loop.
- Input training data is streamed to an "infeed" queue on the Cloud TPU. A program running on Cloud TPU retrieves batches from these queues during each training step.
- The TensorFlow server running on the host machine (the CPU attached to the Cloud TPU device) fetches data and pre-processes it before "infeeding" to the Cloud TPU hardware.
- Data parallelism: Cores on a Cloud TPU execute an identical program residing in their own respective HBM in a synchronous manner. A reduction operation is performed at the end of each neural network step across all the cores.

References

Google TPU
https://www.nextplatform.com/2021/05/21/google-hints-about-its-homegrown-tpuv4-ai-engines/
Intel GPUs
https://edc.intel.com/content/www/us/en/products/performance/benchmarks/architecture-day-2021/

GraphCore
https://docs.graphcore.ai/projects/ipu-overview/en/latest/index.html
https://github.com/graphcore/poplibs
https://www.graphcore.ai/developer
https://docs.graphcore.ai/projects/ipu-overview/en/latest/programming_model.html
https://docs.graphcore.ai/en/latest/getting-started.html

Sambanova
https://www.servethehome.com/sambanova-sn10-rdu-at-hot-chips-33/
https://sambanova.ai/wp-content/uploads/2021/06/SambaNova_RDA_Whitepaper_English.pdf
Cerebras
https://f.hubspotusercontent30.net/hubfs/8968533/CS-2%20Data%20Sheet.pdf
HPE Superdome Flex Server architecture
https://www.hpe.com/us/en/collaterals/collateral.a00036491enw.html
AMD
https://www.amd.com/system/files/documents/amd-cdna-whitepaper.pdf