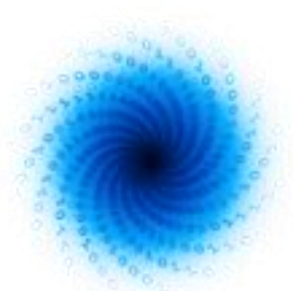




EuroHPC
Joint Undertaking



MAchinE Learning for Scalable meTeoROlogy and climate



MAELSTROM

Roadmap Analysis of Technologies Relevant for ML Solutions in W&C

Mats Brorsson (UL-SnT)

www.maelstrom-eurohpc.eu



D3.2 Roadmap Analysis of Technologies Relevant for ML Solutions in W&C

Author(s):	Mats Brorsson (UL-SnT)
Dissemination Level:	Public
Date:	3/12/2021
Version:	1.2
Contractual Delivery Date:	30/09/2021
Work Package/ Task:	WP3/ T3.1
Document Owner:	UL-SnT
Contributors:	Peter Dueben, Daniele Gregori
Status:	Final

MAELSTROM

Machine Learning for Scalable Meteorology and Climate

Research and Innovation Action (RIA)

H2020-JTI-EuroHPC-2019-1: Towards Extreme Scale Technologies and Applications

Project Coordinator: Dr Peter Dueben (ECMWF)

Project Start Date: 01/04/2021

Project Duration: 36 months

Published by the MAELSTROM Consortium

Contact:

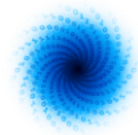
ECMWF, Shinfield Park, Reading, RG2 9AX, United Kingdom

Peter.Dueben@ecmwf.int

The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union's Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway

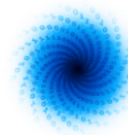


EuroHPC
Joint Undertaking



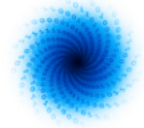
Contents

1	EXECUTIVE SUMMARY	6
2	INTRODUCTION	7
2.1	ABOUT MAELSTROM	7
2.2	SCOPE OF THIS DELIVERABLE	7
2.2.1	OBJECTIVES OF THIS DELIVERABLE	7
2.2.2	DEVIATIONS AND COUNTER MEASURES	8
3	COMPUTE TECHNOLOGIES.....	9
3.1	CPU ROADMAP	9
3.1.1	INTEL	9
3.1.2	AMD.....	12
3.1.3	ARM-BASED	13
3.1.4	RISC-V	16
3.2	ACCELERATOR TECHNOLOGIES.....	17
3.2.1	NVIDIA	17
3.2.2	AMD.....	20
3.2.3	FPGAS ACCELERATION OF ML APPLICATIONS.....	22
3.2.4	OTHER AI ACCELERATORS.....	22
3.3	SUMMARY.....	27
4	COMMUNICATION TECHNOLOGIES	29
4.1	ETHERNET	30
4.2	INFINIBAND	31
4.3	OMNI-PATH.....	32
5	CONCLUSION.....	33



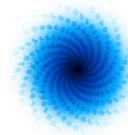
Figures

Figure 1: Intel aims to boost performance 1000x by 2025 © Intel.	9
Figure 2: The Golden Cove Performance core.	10
Figure 3: Introducing the forthcoming Sapphire Rapids.	10
Figure 4: Gen 4 EPYC processors are expected in 2022.	12
Figure 5: The ARM Neoverse N1 architecture reference design.	13
Figure 6: The Neoverse N1 core building blocks.	13
Figure 7: ARM Neoverse roadmap.	14
Figure 8: ARM Neoverse V1 architecture.	15
Figure 9: Nvidia announced the Grace processor in April 2021.	15
Figure 10: The first EPAC test samples.	16
Figure 11: The EPAC1.0 processor consists of four VPU cores and a stencil/tensor accelerator.	16
Figure 12: Nvidia has products that suit machine learning from training to inference, from desktop and datacentre to independent devices.	17
Figure 13: The Nvidia A100 GPU delivers 11 times higher performance on some important HPC applications compared to the Nvidia Pascal P100 architecture of 2016.	18
Figure 14: Probable roadmap of Nvidia chip products.	19
Figure 15: AMD's latest CDNA2 GPU architecture.	21
Figure 16: An example of tensor graph (from the Graphcore Poplar user guide)	23
Figure 17: The Colossus MK2 GC200 IPU.	24
Figure 18: The SambaNova reconfigurable dataflow unit (RDU).	25
Figure 19: The Cerebras Wafer Scale Engine compared to a "regular" GPU chip.	26
Figure 20: The individual core in the WSE-2 architecture.	27
Figure 21: Ethernet, Infiniband and Omnipath are the dominant internet technologies among the top 500 HPC systems.	29
Figure 23: Ethernet roadmap.	30
Figure 22: Infiniband roadmap.	31
Figure 25: Omni-path roadmap.	32



1 Executive Summary

In this report we survey the state-of-the-art in compute technology relevant for machine learning applications, and in particular deep learning algorithms. We cover CPU products from the major manufacturers/designers, Intel, AMD, ARM and with RISC-V as a special mention. Then we discuss the two leading GPU providers' (Nvidia and AMD) offerings for deep learning acceleration as well as looking into some detail on the three main specialised AI compute engine manufacturers, GraphCore, SambaNova and Cerebras.



2 Introduction

2.1 About MAELSTROM

To develop Europe's computer architecture of the future, MAELSTROM will co-design bespoke compute system designs for optimal application performance and energy efficiency, a software framework to optimise usability and training efficiency for machine learning at scale, and large-scale machine learning applications for the domain of weather and climate science.

The MAELSTROM compute system designs will benchmark the applications across a range of computing systems regarding energy consumption, time-to-solution, numerical precision and solution accuracy. Customised compute systems will be designed that are optimised for application needs to strengthen Europe's high-performance computing portfolio and to pull recent hardware developments, driven by general machine learning applications, toward needs of weather and climate applications.

The MAELSTROM software framework will enable scientists to apply and compare machine learning tools and libraries efficiently across a wide range of computer systems. A user interface will link application developers with compute system designers, and automated benchmarking and error detection of machine learning solutions will be performed during the development phase. Tools will be published as open source.

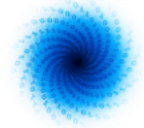
The MAELSTROM machine learning applications will cover all important components of the workflow of weather and climate predictions including the processing of observations, the assimilation of observations to generate initial and reference conditions, model simulations, as well as post-processing of model data and the development of forecast products. For each application, benchmark datasets with up to 10 terabytes of data will be published online for training and machine learning tool-developments at the scale of the fastest supercomputers in the world. MAELSTROM machine learning solutions will serve as blueprint for a wide range of machine learning applications on supercomputers in the future.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

This deliverable provides a survey of available hardware solutions and, where available, the roadmap of key technologies that are relevant for MAELSTROM. The analysis encompasses mostly commercial roadmaps but also the development in the European Processor Initiative which relates to the RISC-V ISA. A focus has been on accelerator technology, where we have connected with the leading vendors of accelerators to get heads up and previews of their roadmaps.

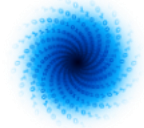
Deliverable 3.2 is one of four MAELSTROM deliverables that survey the state-of-the-art in terms of methods, tools, and developments in machine learning at the beginning of the project and aim to build additional links between the three workpackages that are involved in the MAELSTROM co-design cycle. Deliverable 1.2 is a survey of machine learning methods and tools that are currently used for weather and climate applications. Deliverable 2.1 is a survey of existing machine learning workflow tools and a summary of the MAELSTROM protocol and machine learning requirements.



Deliverables 3.1 and 3.2 provide a systematic analysis of the hardware requirements for the MAELSTROM applications and a roadmap analysis of hardware that will be relevant for machine learning in MAELSTROM.

2.2.2 Deviations and counter measures

However, the submission of the deliverable has been delayed due to delays in the recruitment of new staff at UL-SnT. The delay of this survey deliverable did not cause any impact on any of the tasks or other deliverables of MAELSTROM. The delay made it possible to include information from recent announcements at the Supercomputing and Nvidia GTC conferences.



3 Compute Technologies

3.1 CPU Roadmap

3.1.1 Intel

Intel has long been the workhorse of high-performance computing since they built the first supercomputer to achieve 1 TFLOPS with the ASCI Red¹. At the time of this writing, Intel is competing in the HPC domain with vendors like AMD and Nvidia, in particular (see below) and an interesting line-up in the roadmap was announced at the Intel Architecture Day 2021².

The plan of Intel of 2021 is to make their customers' workload 1000x faster (than today) by 2025. To achieve this, the technology roadmap includes improvements in all aspects of the computer system stack.

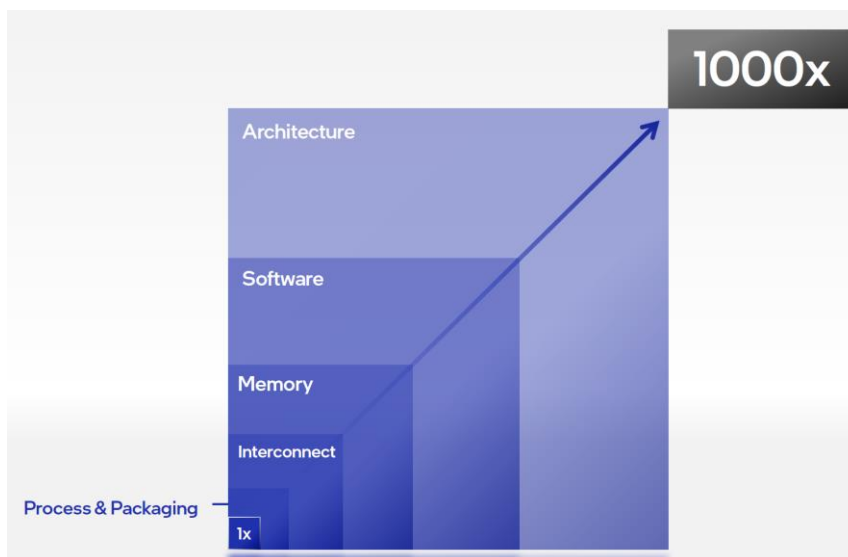


Figure 1: Intel aims to boost performance 1000x by 2025 © Intel.

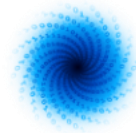
Improvements to reach this goal include:

- Improved micro-architecture to boost scalar performance
- Wider and higher performance vector units
- Improved accelerator support
- Hybrid architectures specific for anticipated workloads

When it comes to micro-architecture improvements, Intel claims an improvement of 19% “across a wide range of workloads” following advances in instruction level parallelism. The new performance-core, code named Golden Cove, has wider instruction decoding and execution (6-12 parallel micro-operations in parallel compared to previously 4-10). In order to fill the execution pipeline with useful instructions, the branch prediction has been improved and the L1 cache latency has been effectively reduced. Since there are more instructions in flight at any given time, hardware structures such as

¹ [ASCI Red - Wikipedia](#)

² [Intel Architecture Day 2021](#)



the number of physical registers and the size of the re-order buffer (the structure where instructions wait for their data dependencies to be filled) have been increased as well.



Figure 2: The Golden Cove Performance core.

Golden Cove also has full AVX-512³ support introducing a faster and more power efficient vector addition unit and 16-bit floating point number representation as well as support for complex numbers.

Based on the new performance-core, Intel has announced the next generation Xeon Scalable Processors with code name Sapphire Rapids designed for data centre and HPC applications. In order to increase scalability in terms of core count and other aspects, the Sapphire Rapids processors are based on a multi-tile design with multiple tiles in a single package. The novelty of this technology in comparison to Intel's first "multi-core" processors which also used multiple chips in the same package, is that all threads now have access to all resources irrespective of the tile they belong to.

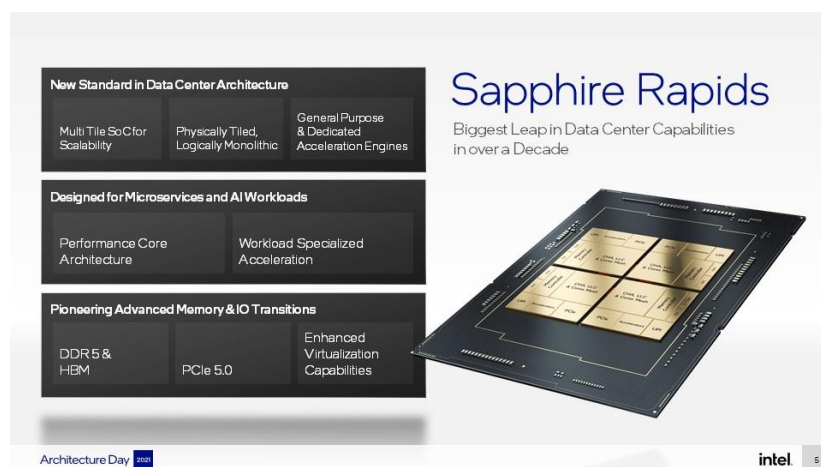
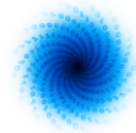


Figure 3: Introducing the forthcoming Sapphire Rapids.

³ 512 bit wide vector registers divided into 64, 32, or 16-bit floating point numbers.



Sapphire Rapids includes advanced virtualisation techniques to particularly enhance its uses in cloud data centres.

With Sapphire Rapids, Intel also embraces various accelerator technologies. The Accelerator Interfacing Architecture (AIA) brings a new instruction set specifically for integrating various accelerator technologies eliminating the needs to frequently switch between user and kernel space which is time and resource consuming. Sapphire Rapids will initially include the following accelerators using AIA:

- Data Streaming Accelerator (DSA) which optimises streaming data movement freeing up the CPU.
- Quick Assist Technology (QAT) which accelerates encryption, decryption, secure hashes and lossless data de-/compression.

Intel has added numerous features which are aimed to boost, in particular, AI applications such as machine learning which are relevant for the MAELSTROM project. These include:

- AMX – Advanced Matrix Extension,
- New hardware supported data types:
 - 8-bit integers, int8, with int32 accumulation
 - 16-bit floating point number, bfp16, with IEEE single precision accumulation

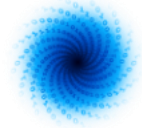
AMX is an acceleration engine designed to massively speedup tensor processing which is the cornerstone of deep learning algorithms. In the first incarnation, there are 8 new Tile registers, each 1 kB large. Paired with the Tile registers is a new instruction set, most notably the TMUL instruction which the matrix add multiplication ($C = C + A * C$) on Tile registers. It can perform 2000 int8 or 1000 bfp16 computations each cycle improving performance by a factor of 7 over using Intel AVX-512 VNNI which was the previous state-of-the-art for deep learning workloads on the CPU.

The processing side is thus significantly improved over previous generation processors but this is not worth much unless you can get data to and from the cores fast enough. Intel has improved on their UPI (Ultra-path interconnect), which is used to support a cache coherent shared memory model across multiple processors, with speeds up to 16 GT/s. Within a processor there is a shared last-level cache of up to over 100 MB size which has almost doubled over previous generations. The new Sapphire Rapids also includes the CLDEMOTE instruction which is used to proactively place cache contents, e.g. evicting specific L1 or L2 cache lines, giving place to more critical data. While integrated DDR5 technology in Sapphire Rapids brings increased memory performance, including improved bandwidth, Intel has integrated HBM (High-bandwidth memory) in package with an expected memory bandwidth of 1 TB/s⁴ compared to the 4.8 GB/s possible with the 8 channels of DDR5-channels⁵.

In summary, Intel has presented a step-up in capabilities for MAELSTROM-like workloads. Still, despite the AI-improvements, it is likely that dedicated accelerators for deep learning algorithms will perform best, at least for training.

⁴ [Intel to Launch Next-Gen Sapphire Rapids Xeon with High Bandwidth Memory \(anandtech.com\)](https://www.anandtech.com/show/15111/intel-to-launch-next-gen-sapphire-rapids-xeon-with-high-bandwidth-memory)

⁵ [DDR5 SDRAM - Wikipedia](https://en.wikipedia.org/wiki/DDR5_SDRAM)



3.1.2 AMD

The current highest performing AMD processor is the generation 3 EPYC processor (code named Milan) with 64 Zen3 cores. Looking at SPEC CPU17 benchmark scores⁶, it is consistently outperforming Intel's highest performing processors with the same configuration. The Zen3 core was introduced earlier in 2021 and improved on performance (measured in increase of instructions per clock cycle, IPC) with 19% over the previous generation Zen2 core.

The 3rd generation EPYC architecture has improved, in particular, the integration and L3 cache accessibility of the cores. AMD has also gone to great length of securing the processor for virtualisation workloads with: secure encrypted virtualisation, encrypted CPU registers, and secure nested paging.

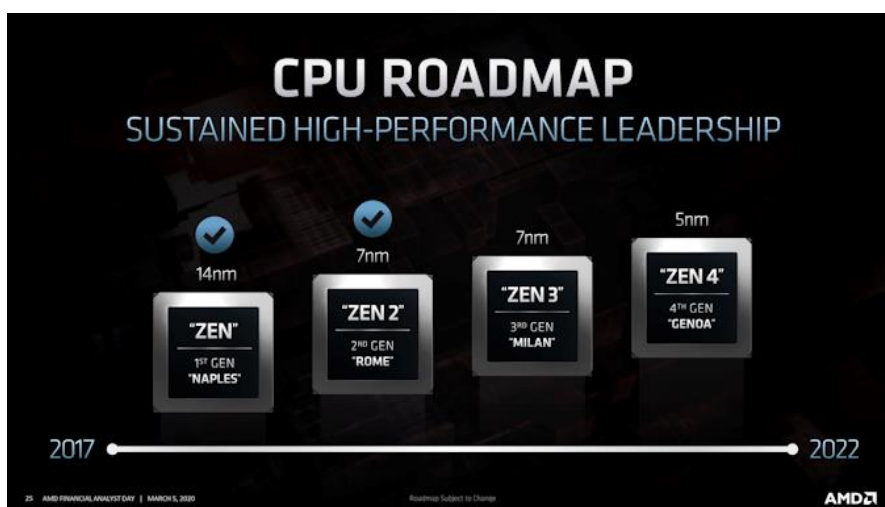


Figure 4: Gen 4 EPYC processors are expected in 2022.

According to the AMD roadmap, we can expect processors powered with Zen4 cores (code name Genoa), the successor to Zen3, to appear around Q2 2022. Details of the Genoa processor have not been officially announced yet, but what we can expect from the EPYC 7004 series of processors are⁷:

- Multi-chip processors with 12 CCDs, each with 8 cores for a total of 96 cores (192 threads with SMT⁸),
- The 4th generation EPYC processors will also have 12 DDR5 channels which should provide for a higher sustained memory bandwidth compared to the Sapphire Rapids, but there is no indication that it will contain HBM.

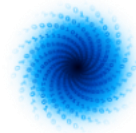
The performance of the Genoa processor, measured in IPC, is expected to be about 29% higher than the Milan processor which already is the highest performing server processor.

Just like Sapphire Rapids, the Genoa processor is expected to feature AVX-512 and bfloat16 operations. Besides that, there is no indication of AI-enhancing features on the processor itself, so that will have to be performed on accelerators such as the AMD Instinct accelerators.

⁶ [CPU2017 Results -- Query \(spec.org\)](#)

⁷ [AMD leaked roadmap teases monster EPYC CPU: 96C/192T on Zen 4 | TweakTown](#)

⁸ SMT – Simultaneous multi-threading.



Looking further ahead, Zen5-based Turin processor is rumoured to contain up to 256 cores (512 SMT), appearing some time in 2024. But AMD has not revealed any official information so this is highly unsure.

In summary, the AMD has not disclosed much information about their upcoming line of processors but has confirmed their current leading position. It is likely that AMD will keep the top position for single and parallel thread CPU intensive workloads.

3.1.3 ARM-based

ARM does not make their own processors but instead license their IP-cores to be integrated and manufactured by other companies. Most of these are in embedded consumer products such as mobile phones and are not known for the stunning performance characteristics. However, both AWS and Nvidia (partly in cooperation) are using ARM-based processors for price and power conscious customers.

The AWS Graviton2 processor, codesigned with Nvidia, is based on the ARM Neoverse N1 architecture shown in the figure below.

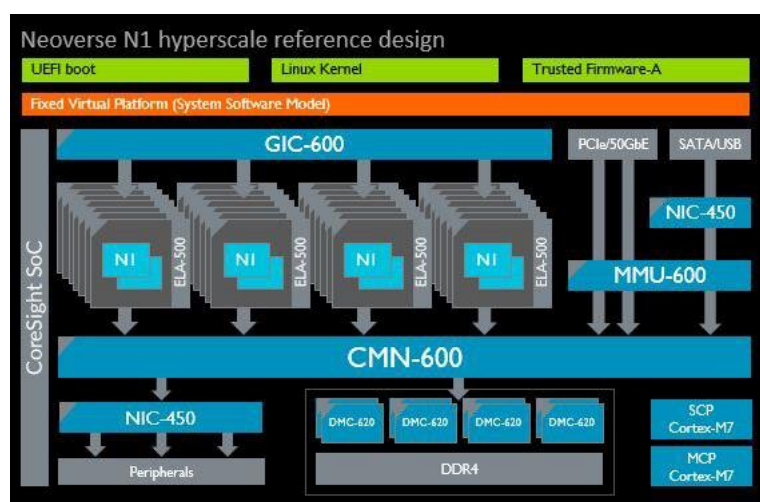


Figure 5: The ARM Neoverse N1 architecture reference design.

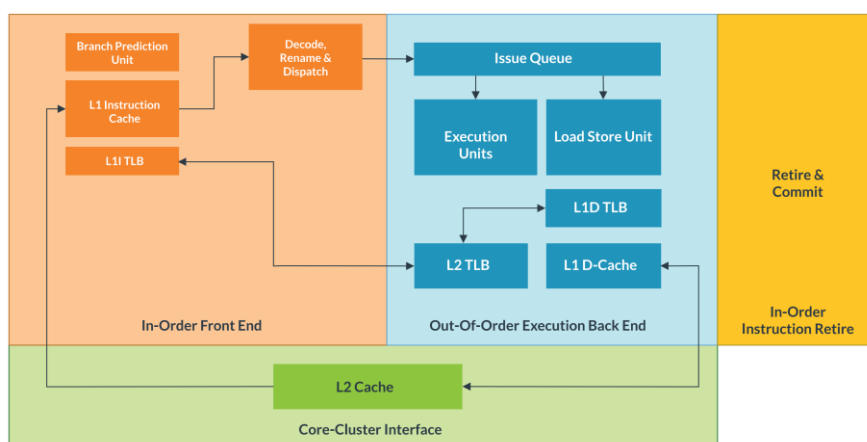
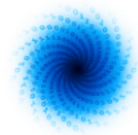


Figure 6: The Neoverse N1 core building blocks.



As shown in Figure 5, the N1 core is similar to other high-performance processors with an in-order front-end, a high-performance execution back-end with instruction level parallelism and an in-order instruction commit block.

AWS claims a 40% better price performance over x86-based instances when using the ARM-based instances. Still, when looking for machine learning workloads, specifically when training a model, many users would want to pair the ARM processor with an accelerator.

The not-yet-released AWS Graviton3 processor (rumoured to be presented at the next AWS re:invent event in December 2021), is likely to be based on the ARM Neoverse V1 (or N2) architecture.

In addition to the prevalence of ARM-processors in mobile and other low-power applications, they are also being used in some of the world's top-performing supercomputers. For instance, at the top of the top500 list of November 2021, is the Fugaku supercomputer at RIKEN with Fujitsu's A64FX processor which is an ARM v8.2 processor with ARM's 512-bit wide Scalable Vector Extension (SVE)⁹ and 48 cores.

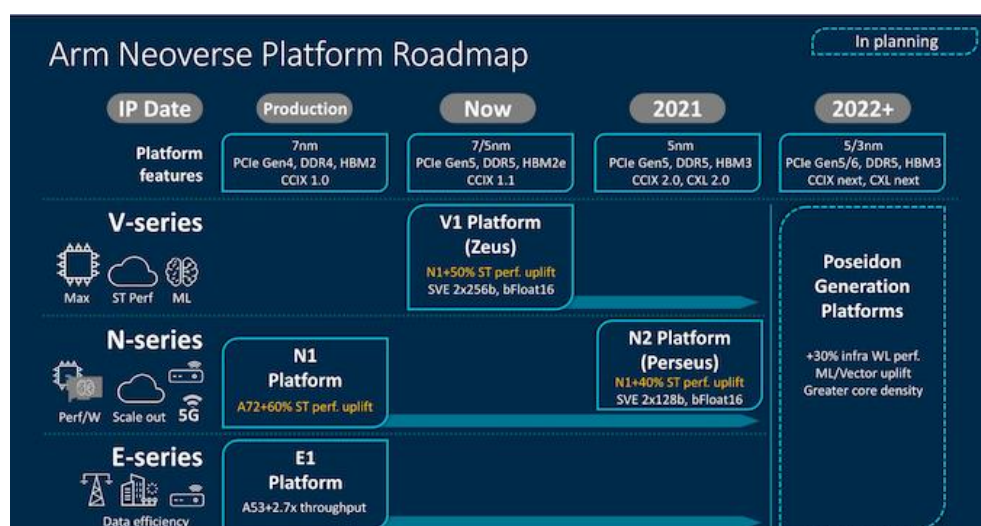


Figure 7: ARM Neoverse roadmap.

The current highest performing ARM processor architecture is the Neoverse V1, see Figure 8. Just like the AMD and Intel high-end processors, the V1 implements datatypes such as int8, bfloat16 and complex numbers to accelerate ML applications. It contains a high-performance memory interface with DDR5 support and HBM (something which AMD lacks as of now).

It remains to be seen who will be the first to implement Neoverse V1, but AWS seems to be a good contender.

⁹ The first implementation of SVE in the world, reportedly.

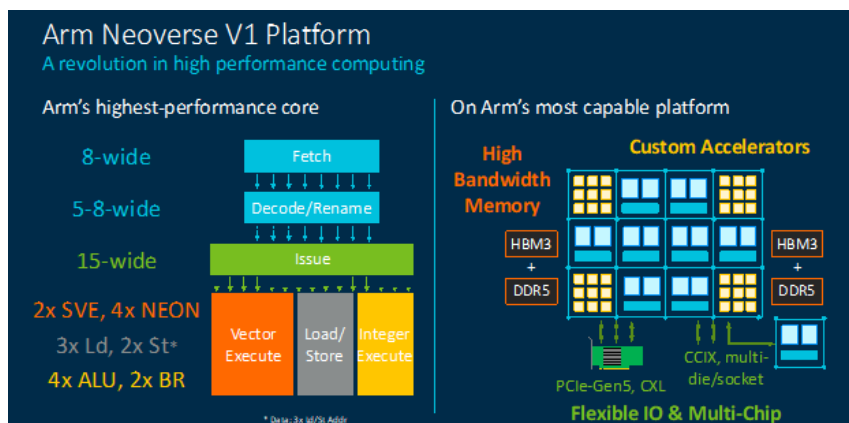
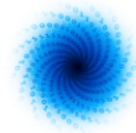
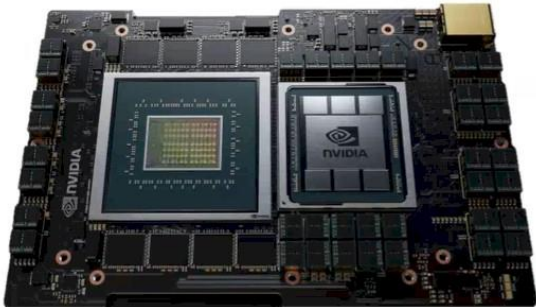


Figure 8: ARM Neoverse V1 architecture.

Nvidia not only co-designed the Graviton2 ARM processor with AWS. They are also developing their own line of ARM-based processors designed specifically for large-scale AI and HPC applications. In April 2021 they announced the Nvidia Grace processor. What distinguishes this processor from any other is the extreme speed by which it may be connected to Nvidia's GPUs. Cache-coherent NVlinks (see chapter 3.2.1 for more information) provide >900 GB/s between CPU and GPU and >600 GB/s between CPUs. This is completely unparalleled bandwidth between processors and GPUs.

ANNOUNCING NVIDIA GRACE

Breakthrough CPU Designed for Giant-Scale AI and HPC Applications



FASTEST INTERCONNECTS
>900 GB/s Cache Coherent NVLink CPU To GPU (14x)
>600GB/s CPU To CPU (2x)

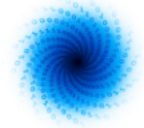
HIGHEST MEMORY BANDWIDTH
>500GB/s LPDDR5x w/ ECC
>2x Higher B/W
10x Higher Energy Efficiency

NEXT GENERATION ARM NEOVERSE CORES
>300 SPECrate2017_int_base
Availability 2023

Figure 9: Nvidia announced the Grace processor in April 2021.

The Grace processor will be available in 2023 and is expected to be based on the Neoverse V1 or N2 cores.

In summary, while AMD and Intel may beat the Grace processor in pure CPU performance (albeit with a significantly higher power envelope), for GPU-intensive workloads the CPU-GPU combination of Grace and Nvidia GPUs will be difficult to challenge thanks to the integrated NVLinks.



3.1.4 RISC-V

RISC-V is an open instruction set architecture (ISA)¹⁰ and is most well-known as an ARM rival in the embedded space. The European Processor Initiative¹¹, was established in 2018 with an explicit goal to build an EU Exascale machine based on an EU processor by 2023. In September 2021 the first test chips, based on the RISC-V ISA were delivered¹². The processor is named EPAC (European Processor Accelerators).



Figure 10: The first EPAC test samples.

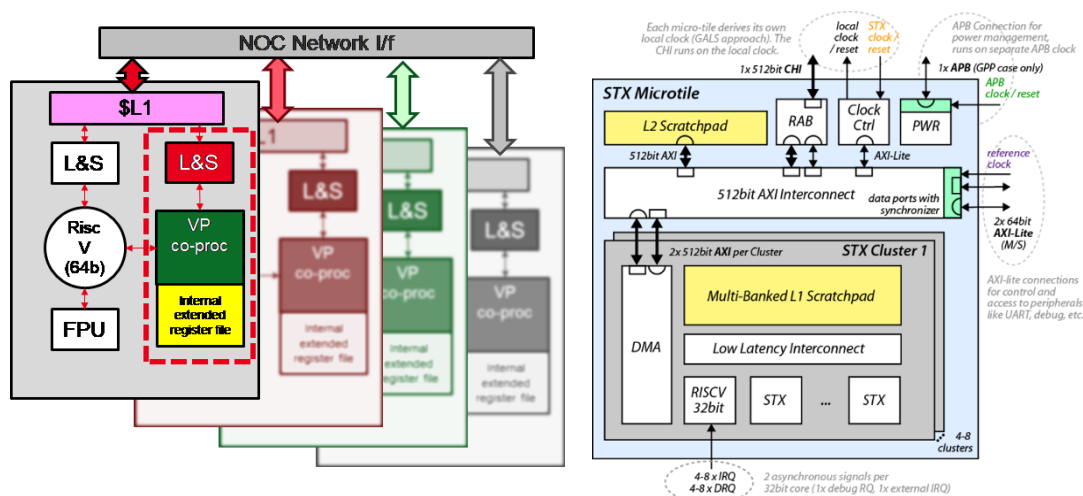


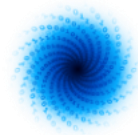
Figure 11: The EPAC1.0 processor consists of four VPU cores and a stencil/tensor accelerator.

As seen in the figure above, the EPAC 1.0 has four Variable Precision Units (VPUs) and each of them consists of a RISC-V core augmented with a variable precision vector unit and a stencil/tensor accelerator which will accelerate specifically machine learning applications. This accelerator will be

¹⁰ [RISC-V International \(riscv.org\)](https://riscv.org)

¹¹ [Home - European Processor Initiative \(european-processor-initiative.eu\)](https://european-processor-initiative.eu)

¹² [European Processor Initiative EPAC1.0 RISC-V Test Chip Samples Delivered \(hpcwire.com\)](https://hpcwire.com)



programmed using OpenMP which has the potential to significantly reduce the complexity of software development.

In summary, the only current HPC initiative using the RISC-V ISA is the European Processor Initiative and it is too early to draw any conclusions about performance while the EPAC processor has many of the same components as displayed in the high-end Intel, AMD or ARM processors. Information on performance numbers from EPI or more details about memory bandwidth etc. have not been released yet. It is surely a technology to monitor, and the first systems should be available during the lifetime of MAELSTROM.

3.2 Accelerator technologies

3.2.1 Nvidia



Figure 12: Nvidia has products that suit machine learning from training to inference, from desktop and datacentre to independent devices.

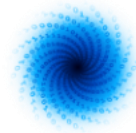
3.2.1.1 The A100 Ampere Tensor Core GPU

The current top-of-the-line GPU architecture from Nvidia is the Ampere embodied as the A100 Tensor Core GPU. While GPUs originally were developed to accelerate the repetitive tasks associated with visual rendering of 2D and 3D scenes on screen, the GPU is now the workhorse of machine learning and other computational tasks which consist of many similar operations on the same data set. The high-end products from Nvidia do not even have display connectors but are specifically dedicated for accelerating compute bound workloads.

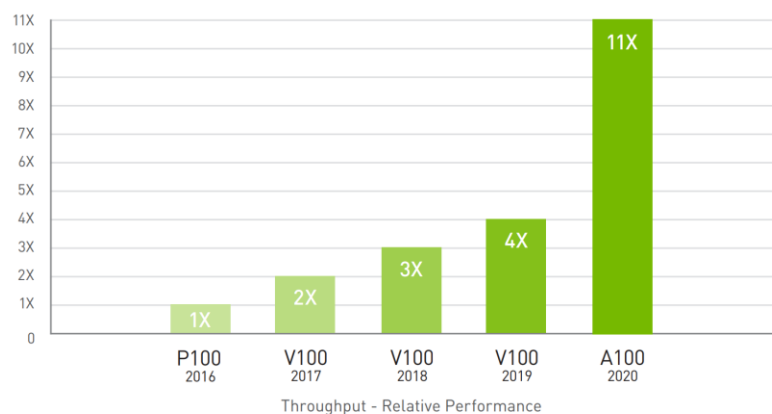
Introduced in 2020, the A100 has delivered a remarkable performance increase over the previous top-of-the-line GPUs from Nvidia as shown in Figure 13 below^{13 14}.

¹³ [NVIDIA Ampere Architecture In-Depth | NVIDIA Developer Blog](#)

¹⁴ [Nvidia A100 Tensor Core GPU Architecture In-Depth \(PDF\)](#)



Throughput for Top HPC Apps



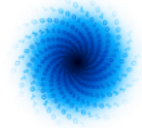
Geometric mean of application speedups vs. P100: Benchmark application: Amber [PME-Cellulose_NVE], Chroma [szsc121_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch [BERT-Large Fine Tuner], Quantum Espresso [AUSURF112-jR]; Random Forest FP32 [make_blobs [160000 x 64: 10]], TensorFlow [ResNet-50], VASP 6 [Si Huge] | GPU node with dual-socket CPUs with 4x NVIDIA P100, V100, or A100 GPUs.

Figure 13: The Nvidia A100 GPU delivers 11 times higher performance on some important HPC applications compared to the Nvidia Pascal P100 architecture of 2016.

The following features of the A100 streaming multiprocessor (SM), which is at the core of every Nvidia GPU, contribute to the remarkable performance increase:

- **Updated third-generation Tensor Cores.** Each SM contains four tensor cores which are specialised for matrix multiplications. Each tensor core can perform 64 mixed-precision fused multiply-add operations per clock cycle. The new features over previous generations' tensor cores are acceleration of all datatypes (FP16, BF16, TF32, FP64, Int8, Int4 and binary) and the sparsity feature which doubles the performance for matrices which have half of their elements being zero.
 - The TF32 is a new innovative 19-bit datatype which provides the range of FP32 (8 bits for exponent) with the precision of FP16 (10 bits mantissa).
- **Increased combined shared memory and L1 data cache.** The fused L1 cache/shared memory, first introduced with the Volta V100 GPU, provides higher flexibility to use equally fast memory for programmer managed shared memory and as L1 cache.
- **New asynchronous copy instruction and associated hardware barrier.** This instruction will initiate transfer from global memory to the SM-shared memory, optionally bypassing the L1 data cache. The barrier can be used to make sure data is not used until it has been loaded.
- **Software improvements.**
 - New instructions to support warp-level reductions.
 - Additional programmability improvements reducing software complexity.

At a system level, the A100 also provides a Multi-instance GPU architecture. This means that it can be divided into, up to, 7 instances which are virtually divided GPU instances reserving GPU resources, such as the compute clusters, L2 cache and DRAM. This capability is particularly important for multi-tenant HPC and Cloud providers and effectively enables the division of a GPU into smaller instances. Previous generations could only support multiple compute instances which would share L2 cache and memory which could have adverse effects when multiple applications shared the GPU.



The shared L2 cache is 40 MB and new instructions have been added for L2 cache management and residence controls. These instructions make it possible to control L2 cache residency in order to optimise capacity utilisation. The L2 capacity is almost sevenfold increased compared to the previous generation. A100 is available with 40 GB or 80GB HBM2 memory interface and the 80 GB variant can sustain up to 2 TB/s memory bandwidth which is the current world record in memory speed. Still, compared to the competition, 80 GB might be proven to be too small for large workloads.

The A100 also introduces the third-generation NVLink for multi-GPU configurations. The total communication bandwidth is doubled compared to V100 for an impressive 600 GB/s capacity.

The A100 also introduced a new hardware-based JPEG decode feature to support higher throughput for deep learning applications (both inference and training) working on images. Nevertheless, the most compelling feature of the A100 for deep learning applications is high memory bandwidth and the reduced and mixed precision arithmetic.

3.2.1.2 Roadmap

Nvidia presented the Ampere A100 in 2020. In Spring 2021, they announced new ARM-based CPU architectures to go with and complement their GPU products, as described in section 3.1.3. In contrast to Nvidia GTC in Spring 2021, they did not reveal any new hardware products at Nvidia GTC in November 2021. However, the previously announced roadmap (see Figure 14) projects the “Ampere Next” to come in 2022.

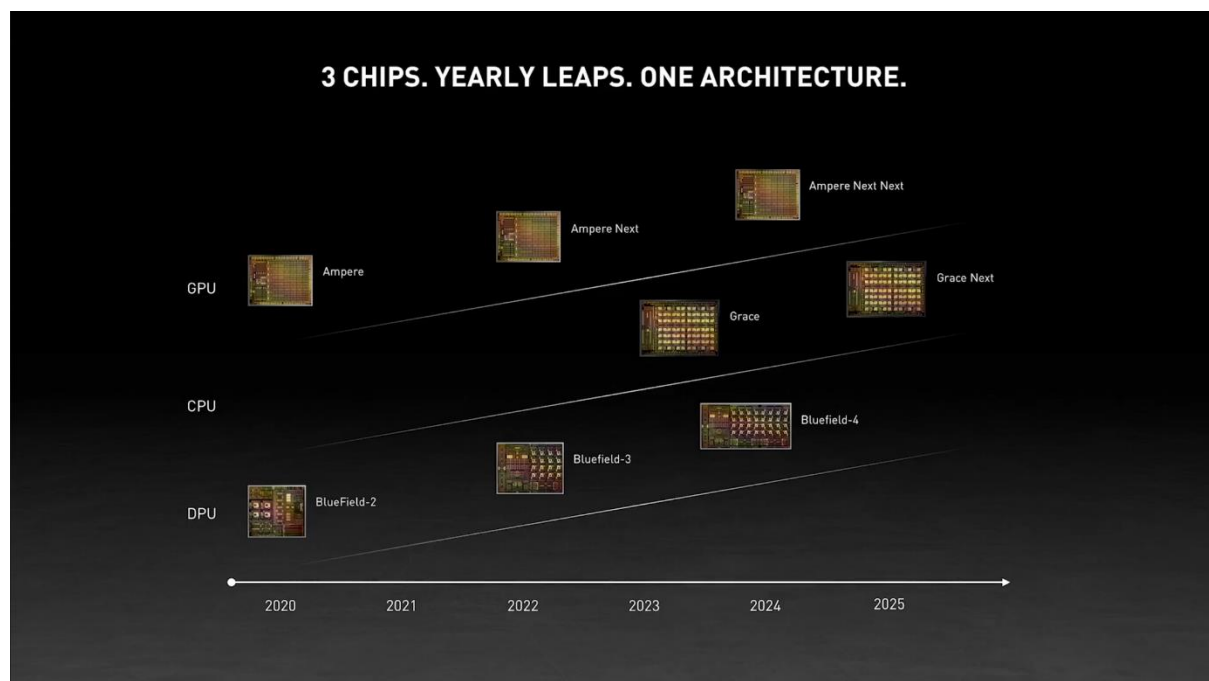
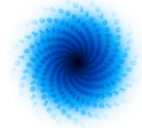


Figure 14. Probable roadmap of Nvidia chip products.



What “Ampere Next” will be is not disclosed yet, but according to Tom’s hardware¹⁵ it will be codenamed Hopper (to go well together with the Grace ARM CPU previously announced in memory of rear-admiral Grace Hopper¹⁶):

“Based on past releases, we expect the Ampere Next / Hopper GPUs to launch with RTX 4000 series branding, and the top model (ie, RTX 4090) would likely outperform the current offerings by at least 25–30 percent, in situations that are GPU limited. We’ll probably see fourth-gen Tensor cores and third-gen RT cores, each boosting performance over the current solutions.”

In summary, the Nvidia A100 Ampere GPU can deliver a peak performance of 19.5 TFLOPS of FP32 operations, common in ML training applications, or up to peak FP16 performance of 312 TFLOPS using the Tensor cores, or even 624 TFLOPS when using the sparsity functionality. If the predictions hold, we could see performance of 25 TFLOPS for FP32 and 400 (800) TFLOPS for FP16 with the Ampere Next (Hopper) GPU.

In addition to the line of hardware developed by Nvidia they have an impressive range of software libraries supporting the life of developer, in particular the CUDA X deep learning software stack¹⁷.

3.2.2 AMD

3.2.2.1 Instinct MI250X

The AMD Instinct MI250X is claimed to be the world’s fastest accelerator for HPC and machine learning workloads with a peak performance of 47.9 TFLOPS for FP32 (95.7 TFLOPS Matrix computations) and 383 TFLOPS for FP16. Comparing to the Nvidia A100, this is indeed higher (not counting using the sparsity feature of the Tensor cores of the A100), although it does not say much about the performance of the application at system level.

At the core of the AMD Instinct MI200 series of GPUs is the CDNA 2 architecture¹⁸. Just like the high-end AMD CPUs, it consists of multiple chip dies (GCD – Graphics Compute Dies) integrated in the same package, the MI250X consists of two chips in a package, see Figure 15.

¹⁵ [Nvidia's 'Ampere Next' GPU Is Coming in 2022 | Tom's Hardware \(tomshardware.com\)](https://www.tomshardware.com/news/nvidia-ampere-next-gpu-is-coming-in-2022)

¹⁶ [Grace Hopper - Wikipedia](https://en.wikipedia.org/wiki/Grace_Hopper)

¹⁷ [Deep Learning Software | NVIDIA Developer](https://developer.nvidia.com/deep-learning-software)

¹⁸ [amd-cdna2-white-paper.pdf](https://www.amd.com/en/na/instinct/white-papers/amd-cdna2-white-paper.pdf)

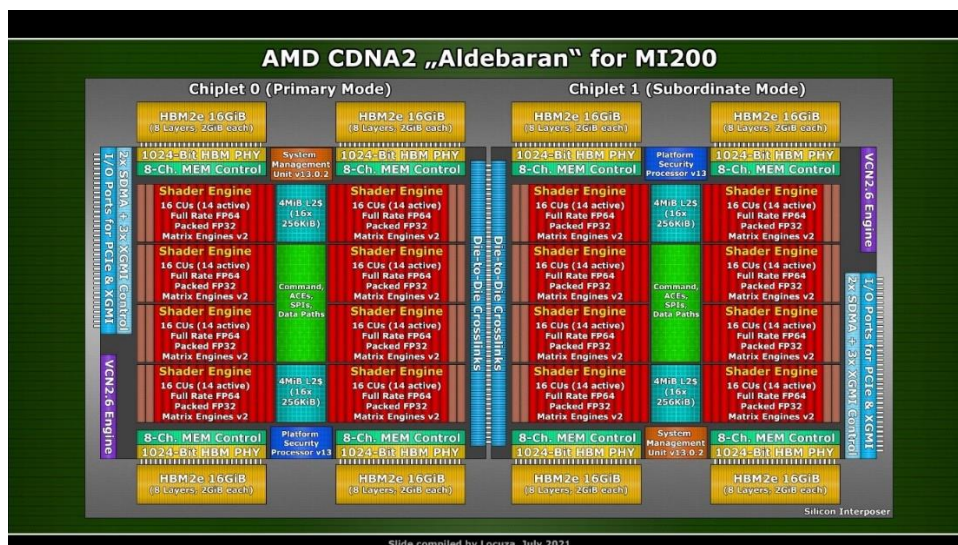
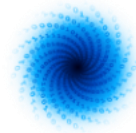


Figure 15. AMD's latest CDNA2 GPU architecture¹⁹.

At the shader engine level (corresponding roughly to a streaming multiprocessor in Nvidia A100), the architecture of AMD CDNA2 is very similar to the Nvidia A100. They both support multiple floating-point representations and specific acceleration of matrix multiplications which is the core of deep learning training and inference computations. AMD have a little more emphasis on FP64, which is important for many HPC applications while Nvidia emphasises machine learning applications more which favour trading precision for higher throughput. This is also evident in that the Matrix Core of AMD CDNA2 does not have support for sparse matrices.

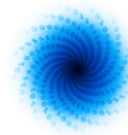
The AMD Infinity Fabric is a distinguishing technology compared to Nvidia. Previous generations of Infinity Fabric could be used to connect multiple GPUs together, while with the CDNA2 architecture and the new EPYC processors, the CPUs and the GPUs can now be part of the same Infinity Fabric providing coherent access to memory across chip boundaries.

The L2 cache architecture of CDNA2 is not too different from the Ampere GPU from Nvidia. It is, however, much smaller which makes you wonder about the efficiency. For streaming data, and machine learning applications tend to stream data, the 8 MB per chip (or GCD for Graphics Compute Die), might be enough. The memory capacity of the Instinct MI250X is 64 GB/GCD which results in an impressive amount of up to 128 GB per device at an aggregate bandwidth of 6.96 TB/s.

As mentioned before, the Instinct MI250X contains two chips which are nodes of the Infinity Fabric. Because the proximity, these two chips may communicate with up to 400 GB/s between the chips in the same package. Each package contains 8 external Infinity Fabric links for a total theoretical communication bandwidth of 800 GB/s. This is more than the double bandwidth of the previous generation.

Compared to Nvidia, the AMD Instinct GPUs lack the sophisticated virtualisation support for multi-tenant workloads. This may not be so important for HPC applications but might be crucial in a cloud environment setting.

¹⁹ Figure from [AMD Instinct MI200 CDNA 2 MCM GPU Is A Beast: 1.7 GHz Clocks, 47.9 TFLOPs FP64 & Over 4X Increase In FP64/BF16 Performance Over MI100 \(wccftech.com\)](https://www.wccftech.com)



On the software side, AMD has an almost drop-in replacement of the industry standard CUDA in the HIP programming model²⁰. However, by far the easiest way of making use of both Nvidia as well as of AMD devices is by means of higher-level abstractions. There are back-ends of both Tensorflow and PyTorch for either architecture and the recent oneAPI from intel contains pathway for C++ programming for both HIP and CUDA devices (as well as for FPGAs)²¹.

In summary, the MI250X performance is impressive on the HPC side requiring mostly high-precision arithmetic. The DL performance, mostly based on 16-bit precision, is not that much ahead of the competition so it remains to be seen how well it will do in this domain.

3.2.2.2 AMD Roadmap

AMD has not yet publicly disclosed any roadmap of specifically the CDNA-line (Instinct) of GPUs. The Instinct MI250X just recently became available and no word has been said about the successor. However, given the head-to-head competition with Nvidia, it is fair to believe that besides “more of everything”, the successor to the MI250X will introduce sparsity support and possibly also virtualisation support to be able to sub-divide the GPU for multiple tenants.

3.2.3 FPGAs acceleration of ML applications

When it comes to using Field programmable gate arrays (FPGA) as accelerator for machine learning applications, there are two dominant players: Intel and Xilinx. Both Intel and Xilinx provide solutions to design custom made hardware accelerators, including a complete software stack, primarily for machine learning inference as the back-propagation of weights is difficult to achieve in an FPGA.

The choice of FPGA for machine learning is usually motivated by power and efficiency (favouring FPGA solutions) over ease of use (favouring GPU solutions). In general, it usually does not make sense using FPGAs for regular deep learning applications developed by the standard frameworks, Tensorflow or PyTorch. However, when certain optimisations like unorthodox number representations, dropout strategies, etc. are used, GPUs become less efficient or not even feasible and, in these cases, an FPGA solution might be relevant. It is also highly relevant for edge inference, for example like in smart cameras to make quick and power efficient decisions based on a trained machine learning model.

MAELSTROM applications are not foreseen to run in edge devices or having special arithmetic data type needs making FPGA an unlikely option.

3.2.4 Other AI accelerators

There are too many startups in the area of hardware AI acceleration to cover the entire field, but we have chosen three representatives which provide a snapshot: Graphcore, SambaNova and Cerebras. A common characteristic of these accelerators is that they are graph compute engines meaning that the computation to execute is represented as a graph where vertices are computational steps and the edges between the vertices are data dependences.

²⁰ [Fundamentals of HIP Programming - AMD](#)

²¹ [GitHub - illuhad/hipSYCL: Multi-backend implementation of SYCL for CPUs and GPUs](#)

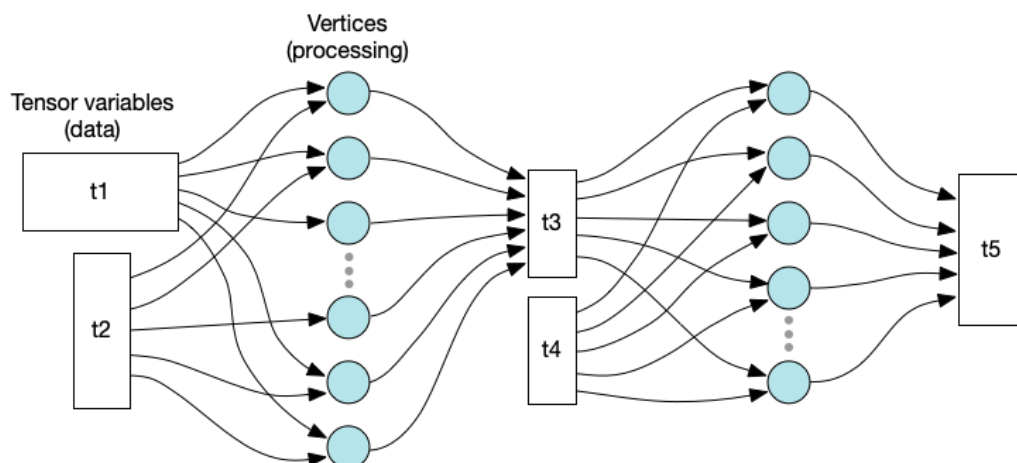
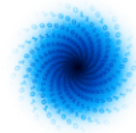


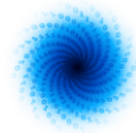
Figure 16. An example of tensor graph (from the Graphcore Poplar user guide)²²

This style of representing computations naturally fits the deep learning model which is now dominating machine learning algorithms.

Another common feature of these accelerators is that they seamlessly integrate with the standard deep learning frameworks such as Tensorflow and PyTorch. This makes the adoption of these accelerators easy but maybe less useful for other kinds of workloads.

All three solutions are suitable for both training and inference.

²² [2. Programming with Poplar — Poplar and PopLibs User Guide \(graphcore.ai\)](https://graphcore.ai/poplar-and-poplibs-user-guide)



3.2.4.1 Graphcore

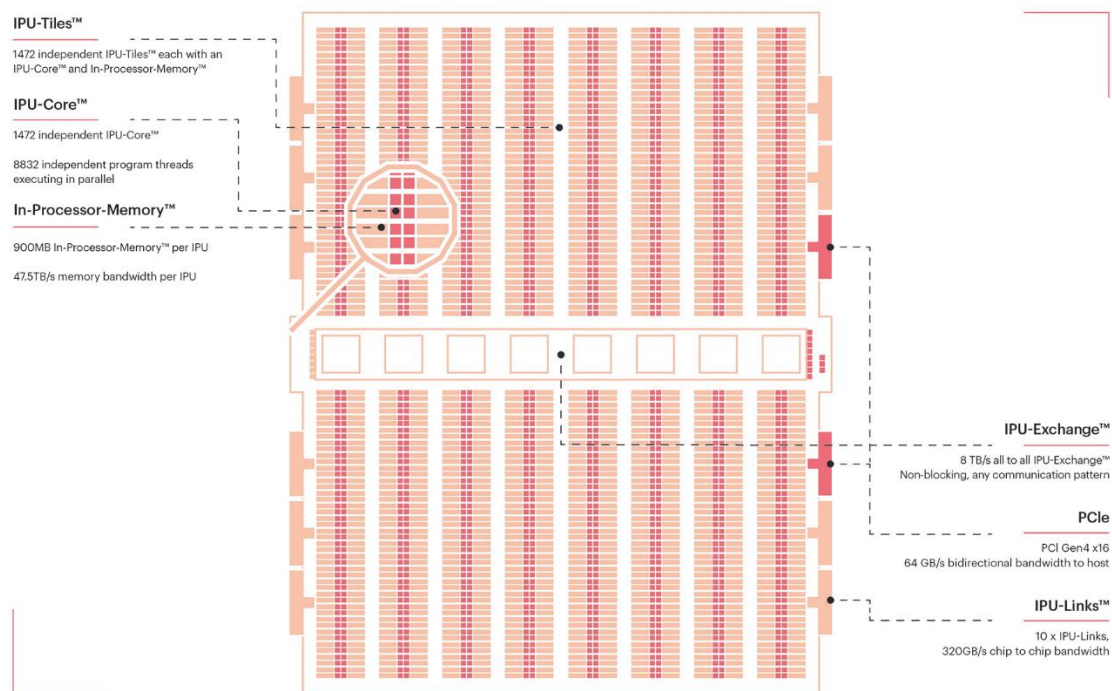


Figure 17. The Colossus MK2 GC200 IPU²³.

At the core of Graphcore's offering is the Colossus MK2 GC200 IPU (Intelligent Processing Unit). The main innovations of the latest IPU are in:

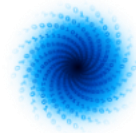
- Compute
- Data handling
- Communication

When it comes to compute, the MK2 IPU contains 1472 independent cores hosting 8832 threads resulting in a peak performance of 250 TFLOPS at FP16 precision, which is the most common in deep learning algorithms.

The data handling deals with the memory on and off chip. There's 900 MB memory on chip (compared to Nvidia's L2 cache 40 MB). The Exchange Memory technology can put 100s of GB next to the processor. The IPU-M2000 is a full IPU-Machine system with four IPU's, up to 450 GB Exchange Memory yielding a peak 180 TB/s memory bandwidth using the system-native Poplar programming model. The IPU-Machine has 1 PFLOPS compute capacity in one 5 cm rack chassis.

With the IPU Fabric, a 3D ring communication technology, systems of up to 64000 IPU's can be built for large scale model processing.

²³ [IPU Processors \(graphcore.ai\)](https://www.graphcore.ai)



With the Poplar SDK one can connect to popular ML frameworks such as PyTorch and Tensorflow. Graphcore's engine builds the runtime to execute the workload across as many IPU machines as needed. The GraphCore Virtual IPU is a software layer to support multi-tenant workloads on a GraphCore IPU system. GraphCore has started to benchmark performance for MAELSTROM datasets²⁴.

3.2.4.2 SambaNova

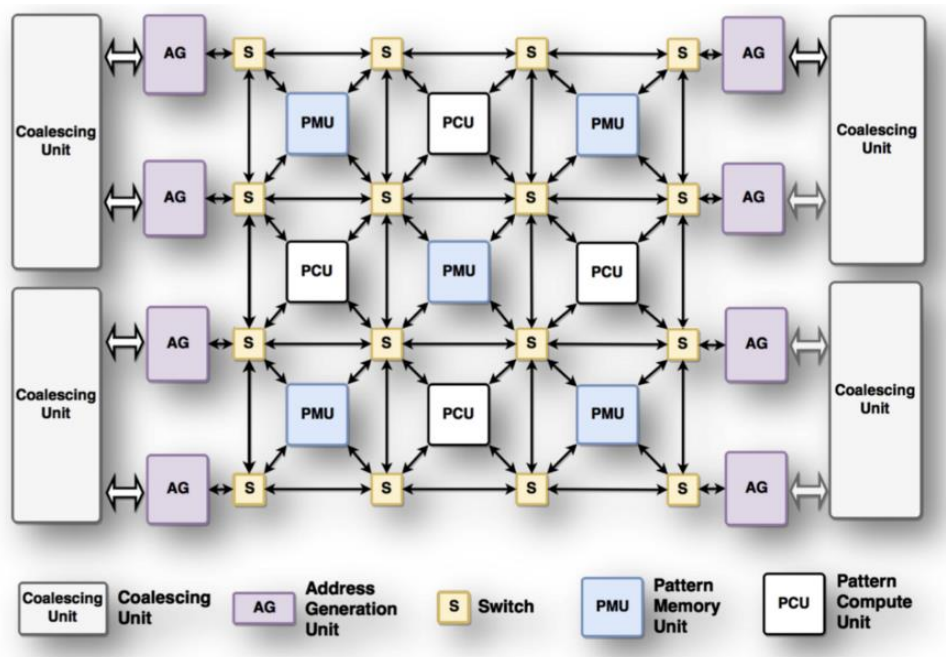


Figure 18. The SambaNova reconfigurable dataflow unit (RDU)²⁵.

Similar to the Graphcore, the SambaNova Reconfigurable Dataflow Architecture is a dataflow engine, see Figure 18 above. The SambaFlow software environment maps graphs expressed in a language called Spatial to hardware tiles, or complete models in Tensorflow or PyTorch. In contrast to GraphCore's device, the RDU is not directly programmable, but reconfigurable which is akin to the process done in FPGA's to create new logic, but much more light-weight. The process of reconfiguring the RDU takes about 10 – 40 μ s.

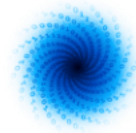
The SambaNova Cardinal SN10 RDU is part of a complete system for deep learning acceleration in the DataScale SN10-8R product.

The Cardinal SN10 RDU contains 640 pattern compute units with an aggregate peak performance of 300+ TFLOPS at bfloat16 precision. Additionally, it has 640 Pattern memory units with an aggregate of 300+ MB on-chip memory.

The DataScale systems come in racks with 1 to 4 SN10-8R units.

²⁴ [Climate Change: Foreseeing the Unexpected with Graphcore IPU's](#)

²⁵ [Accelerating Scientific Applications With SambaNova RDA](#)



3.2.4.3 Cerebras

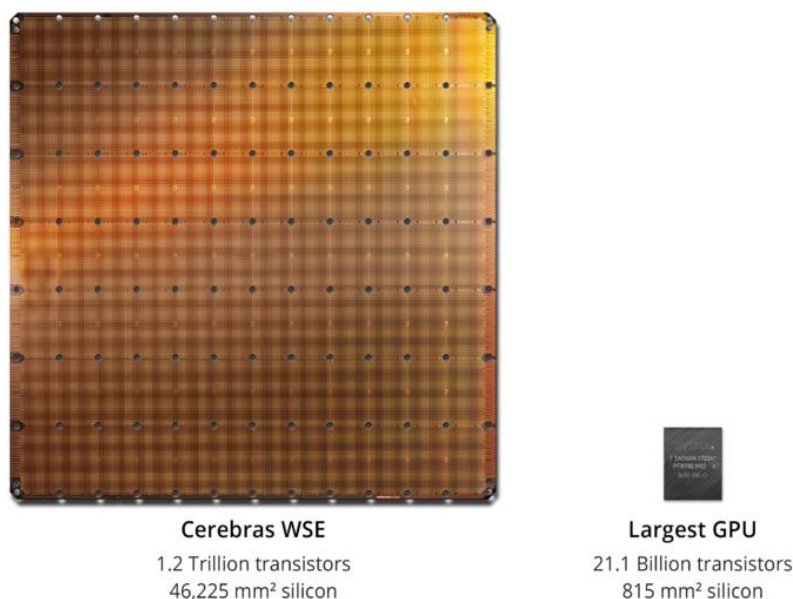
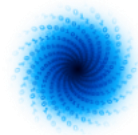


Figure 19. The Cerebras Wafer Scale Engine compared to a "regular" GPU chip²⁶.

The most important claim to fame for Cerebras is its giant wafer-scale die which is by far the largest in the world with its 46 225 mm² compared to the 800+ mm² size of GraphCore Mk2, SambaNova Cardinal SN10 RDU and Nvidia A100. The Cerebras WSE-2 contains 850 000 programmable cores (called Sparse Linear Algebra cores, or SLA cores for short) optimised for the mathematical operations specifically found in deep neural networks. Integrated with the cores is 40 GB of fast memory (compare this to the 40 GB of external RAM of the Nvidia A100 systems). The aggregated memory bandwidth is 20 PB/s and the aggregated communication bandwidth on-chip is 220 Pbits/s.

Although the big WSE-2 chip enables processing of big models without the need to move parts of the models out to external memory, the Cerebras CS-2 system can be scaled to up to 192 CS-2 with the SwarmX interconnection technology to build even larger systems.

²⁶ [Why We Need Big Chips for Deep Learning | Cerebras](#)



Cerebras Architecture is Designed for Sparse Compute

- Fine-grained dataflow cores
 - Triggers compute only for non-zero data
- High bandwidth memory
 - Enables full datapath performance
- High bandwidth interconnect
 - Enables low overhead reductions

Only architecture capable of accelerating **all types of sparsity**, including dynamic and unstructured sparsity.

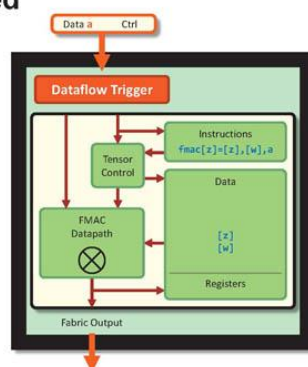


Figure 20. The individual core in the WSE-2 architecture²⁷.

As seen in Figure 20, each individual core is very simple, just as in the SambaNova and GraphCore architectures. But with over 800-thousand cores the aggregate performance becomes immense (although not disclosed by Cerebras).

Programming the WSE-2 is as simple as using Tensorflow and PyTorch. Everything works seamlessly and the Cerebras Graph compiler (CGC not to be confused with GCC) translates the neural network to an CS-2 executable.

3.3 Summary

The table below summarises the current state-of-the-art of each chip discussed above (except FPGAs). The numbers are per chip but for some devices, you can only buy them as complete systems.

System	FP16 peak/chip	L2 cache /on-chip mem	Memory	Mem BW	Comm BW	TDP	Est price
Nvidia A100	312 TFlops	40 MB	40 GB	1.5 TB/s	600 GB/s	400 W	\$15 000-\$20 000
AMD MI250X	380 TFlops	16 MB	128 GB	7 TB/s	800 GB/s	500 W	N/A
Graphcore	250 TFlops	900 MB	450 GB	180 TB/s	2.8 Tbit/s per IPU	375 W ²⁸	\$8 113 ²⁹
SambaNova	300 TFlops	300 MB	1.5 TB	153 GB/s	N/A	N/A	N/A ³⁰
Cerebras	N/A	40 GB	N/A	20 PB/s	220 Pbit/s	15 kW	"Several million" ³¹

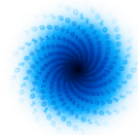
²⁷ [HC33 Cerebras WSE 2 Architecture For Sparsity - ServeTheHome](#)

²⁸ A system with four IPUs is the smallest system you can buy which has a TDP of 1.5 kW. Divided by four it becomes 375 W.

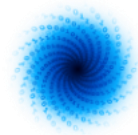
²⁹ For the M2000 system with four MK2 IPUs the price is \$32450.

³⁰ The purchase price is not disclosed, but you can rent a system for \$10 000 per month.

³¹ Can be rented for \$180 000 per month.



The price of the AMD MI250X is not yet known at the time of this writing, and is often a case for negotiation anyway. But it is not likely to be cheaper than the Nvidia A100. The tradeoffs are not entirely clear. You most likely need to understand your workloads better and engage in discussions with the individual vendors to be able to decide on the best price/performance/power design point.



4 Communication technologies

For a large-scale high-performance computing, the communication system between the individual nodes is one of the more crucial components for a high-performing system. Based on information from the top 500 HPC systems in the world as of November 2021, see Figure 21, we have chosen to take a closer look at the development in Ethernet, Infiniband and Omni-path.

Interconnect Family System Share

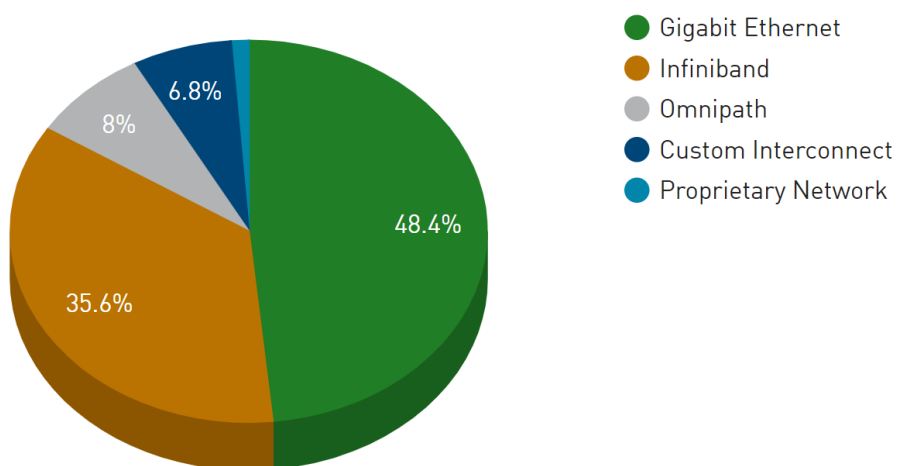
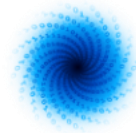


Figure 21. Ethernet, Infiniband and Omnipath are the dominant internet technologies among the top 500 HPC systems³².

³² [List Statistics | TOP500](#)



4.1 Ethernet

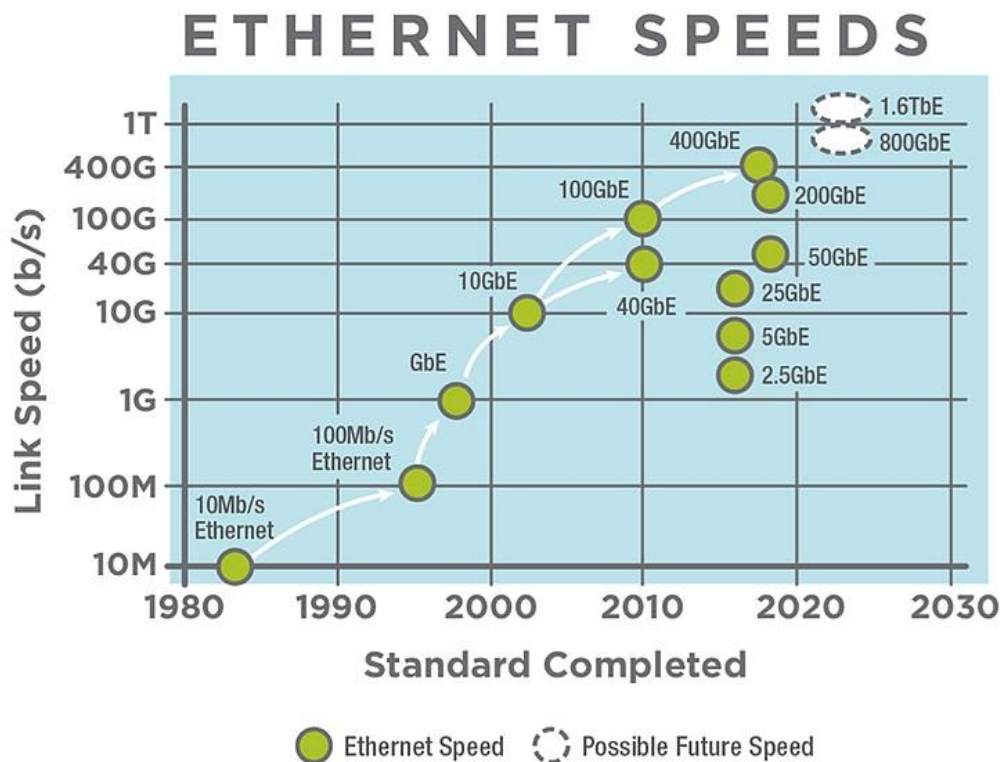
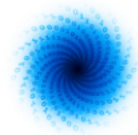


Figure 22. Ethernet roadmap³³.

Ethernet, originally developed as a pure shared communication medium for interconnected computers using collision detection and exponential back off for collision resolution, now rivals the competition in pure bandwidth. This is mostly thanks to the evolution from shared physical medium to point-to-point links and switches. 1.6 Tbits/s transfer speed is on the horizon.

Bandwidth is not the only concern and Ethernet has been lagging behind Infiniband because of the legacy reliance on the TCP/IP software stack. However, with technologies such as RDMA (remote direct memory access), bypassing the operating system, similar latencies as Infiniband can be achieved. On the other hand, machine learning applications, relevant for MAELSTROM, are not as latency sensitive as numerical simulation and therefore Ethernet is likely a cost-effective solution.

³³ [Ethernet Roadmap | Ethernet Alliance](#)



4.2 Infiniband

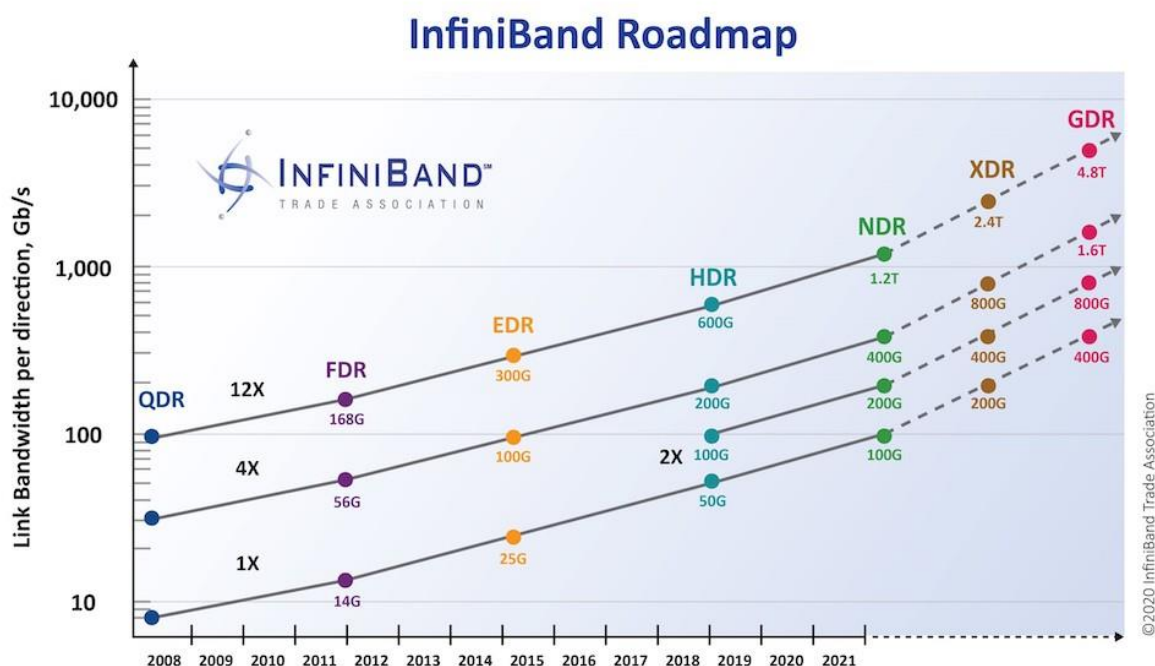
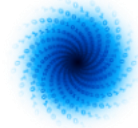


Figure 23. Infiniband roadmap³⁴.

Infiniband was for a long time the choice of network in most HPC systems but has since 2016 been surpassed by Ethernet. However, if you weigh the systems by performance it still holds the first place. On a single link, the speed of Infiniband is similar to that of Ethernet but since you can couple links together, the aggregated bandwidth becomes very high and with sub μ s end-to-end latency.

³⁴ [To Infini\(Band\)ty and Beyond \(hpcwire.com\)](https://www.hpcwire.com/2019/09/10/infiniband-roadmap/)



4.3 Omni-path

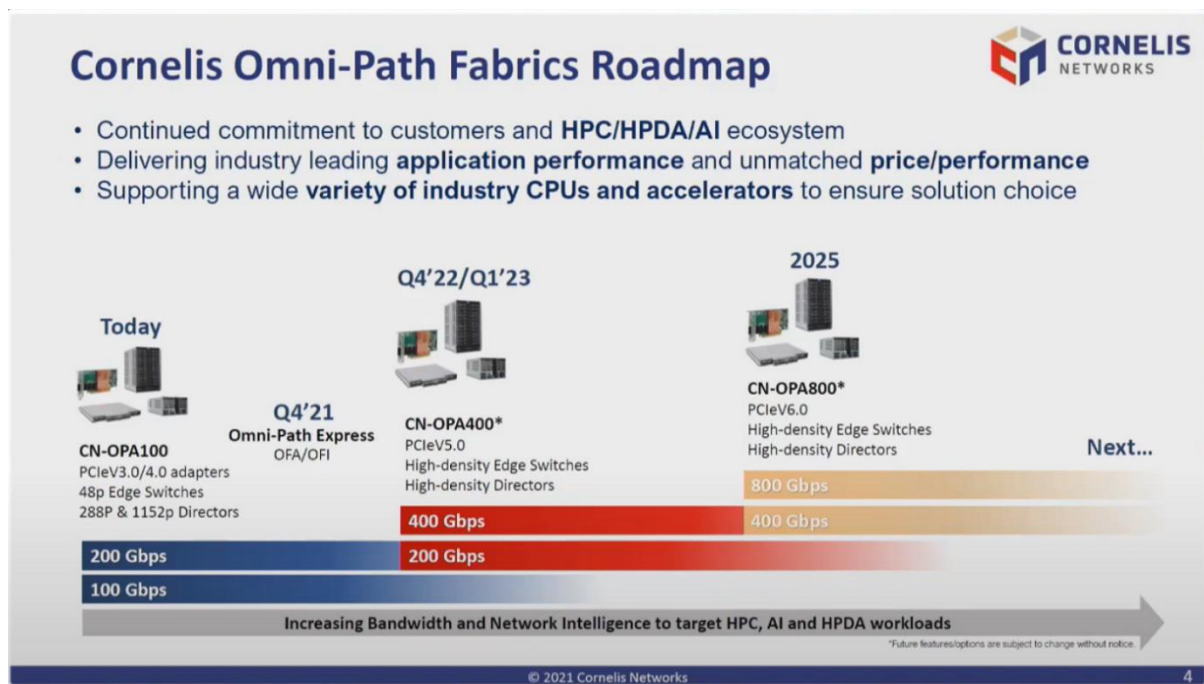
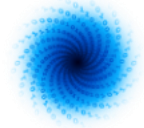


Figure 24. Omni-path roadmap³⁵.

Since Intel decided to leave Omni-path, its future was uncertain, but under its new custodian, it might be making a come-back. Cornelis networks announced this summer new products aiming to close in to the Tbit/s speed range.

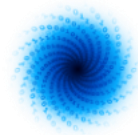
³⁵ [With New Owner and New Roadmap, an Independent Omni-Path Is Staging a Comeback \(hpcwire.com\)](https://hpcwire.com/2021/08/24/with-new-owner-and-new-roadmap-an-independent-omni-path-is-staging-a-comeback/)



5 Conclusion

We are living in exciting times. For a computer architect, the systems covered in this report represent a remarkable effort. It is clear that, although all CPU providers are making great strides to support AI applications, they will not be enough for the enormous compute capacity needed when coming generations of deep learning algorithms will be trained. The traditional method has been to use multiple GPUs for these jobs, and all the covered vendors have products that fit the bill but to a high power and price cost, as these systems are still quite general. FPGA's are currently not the solution for core training jobs.

The more custom-made architectures have a high price point, although not per floating point operation as they are so efficient. The choice of system to use is, however, not straight forward. A more thorough study on the applications at hand is needed and probably also direct discussions with the vendors themselves.



Document History

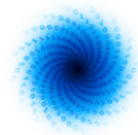
Version	Author(s)	Date	Changes
0.1	Mats Brorsson (UL-SnT)	2021-07-20	Creation
1.0	Mats Brorsson (UL-SnT)	2021-12-03	First complete version
1.1	Mats Brorsson (UL-SnT)	2021-12-12	Revision after review
1.2	Mats Brorsson (UL-SnT)	2021-12-13	Revision after review 2
1.3	Mats Brorsson (UL-SnT)	2021-12-17	Revision 2 after review 2

Internal Review History

Internal Reviewers	Date	Comments
Peter Dueben (ECMWF)	2/12/2021	
Daniele Gregori (E4)	11/12/2021	
Andreas Herten (FZJ)	13/12/2021	

Estimated Effort Contribution per Partner

Partner	Effort
UL-SnT	1
Total	1



This publication reflects the views only of the author, and the European High-Performance Computing Joint Undertaking or Commission cannot be held responsible for any use which may be made of the information contained therein.