Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

State-of-the-Art and Trends for Computing and Interconnect Network Solutions for HPC and AI

A. Tekin [a*1], A.Tuncer Durak [a*2], C. Piechurski [b*3], D. Kaliszan [c*4], F. Aylin Sungur [a*5], F. Robertsén [d*6], P. Gschwandtner [e*7]

[a]National Center for High Performance Computing – UHEM, [b]GENCI, [c]Poznań Supercomputing and Networking, [d]CSC, [e]Universität Innsbruck – UIBK

Abstract

Since 2000, High Performance Computing (HPC) resources have been extremely homogeneous in terms of underlying processors technologies. However, it becomes obvious, looking at the last TOP500, that new trends tend to bring new microarchitectures for General Purpose Processors (GPPs) and new heterogeneous architectures, combining accelerators with GPP, to sustain both numerical simulation and Artificial Intelligence (AI) workflows. The present report provides a consolidated view on the current and mid-term technologies (2019-2022+) for two important components of an HPC/AI system: computing (general purpose processor and accelerators) and interconnect capabilities and provides an outlook on future trends in terms of mid-term projections about what users may expect in the coming years.

[1] adem.tekin@be.itu.edu.tr
[2] a.tuncer.durak@uhem.itu.edu.tr
[3] christelle.piechurski@genci.fr
[4] damian.kaliszan@man.poznan.pl
[5] aylin.sungur@itu.edu.tr
[6] fredrik.robertsen@csc.fi
[7] philipp.gschwandtner@uibk.ac.at

Table of contents

1. Introduction 3
2. Key Factors in Processor Performance 4
   - 2.1. Manufacturing Process 4
   - 2.2. CPU Frequency 4
   - 2.3. Instruction Sets 4
   - 2.4. Vector Length 5
   - 2.5. Memory Bandwidth 5
   - 2.6. Simultaneous Multithreading (SMT) 6
   - 2.7. Processor Packaging 6
   - 2.8. Heterogeneous Dies 6
3. General Purpose Computing Capabilities 7
   - 3.1. X86_64 Processors 7
   - 3.2. Arm Processors 9
   - 3.3. POWER Processors 12
   - 3.4. Other Processor Technologies 13
4. GPU, Accelerator and FPGA 14
   - 4.1. GPUs 14
   - 4.2. Other Types of Accelerators 18
   - 4.3. FPGAs 21
   - 4.4. OCP Acceleration Module 22
5. Interconnects 22
   - 5.1. Inter-Node Connectivity 22
   - 5.2. Intra-Node Connectivity 25
6. Power efficiency 29
7. Conclusion: Major Trends 31
8. References 32
9. List of acronyms 34

1. Introduction

This technical report is part of a series of reports published in the Work Package "HPC Planning and Commissioning" (WP5) of the PRACE-6IP project. The series aims to describe the state-of-the-art and mid-term trends of the technology and market landscape in the context of HPC and AI, edge-, cloud- and interactive computing, Big Data and other related technologies. It provides information and guidance useful for decision makers at different levels: PRACE aisbl, PRACE members, EuroHPC sites and the EuroHPC advisory groups "Infrastructure Advisory Group" (INFRAG) and "Research & Innovation Advisory Group" (RIAG) and other European HPC sites. Users should refer to this series of reports as an overall view of HPC technologies and expect some of the solutions described to be available to them soon. The present report covers "State-of-the-Art and Trends for Computing and Network Solutions for HPC and AI". Further reports published so far are covering "Data Management Services and Storage Infrastructures" [1] and "Edge Computing: An Overview of Framework and Applications" [2]. The series will be continued in 2021 with further selected highly topical subjects.

Since 2000, High Performance Computing (HPC) resources have been extremely homogeneous in terms of underlying processors technologies being mostly based on clusters of nodes equipped with microprocessors. However, it becomes obvious, looking at the last TOP500 (June 2020) [3], that new trends tend to bring new microarchitectures for General Purpose Processors (GPP) and new heterogeneous architectures combining accelerators/GPUs with GPP to sustain both numerical simulation and Artificial Intelligence (AI) workflows.

While the GPP market was mostly led by Intel and its X86_64 processor family for more than 15 years and the GPU market was mainly sustained by NVIDIA until recently, there is a lot of existing companies and newcomers proposing new chips capable to satisfy application computing needs while been extremely efficient in terms of GFlops/Watt. With a large amount of information available on these technologies from various sources, the present report provides an overall and consolidated view on the current and mid-term technologies (2019-2022+) available for two important components of an HPC/AI system: computing (GPP and accelerators) and interconnect technologies. This report does not claim to be an exhaustive view of what is available today though covering the most popular and know current and future technologies.

Computing technologies are introduced first (Section 2) through key factors to consider for the analysis of processor performance and their relation to architectural choices. Section 2 aims to familiarise the reader with processor performance aspects, highlighting the most important problems and the proposed solutions. Besides covering competing technologies, it mentions trends and failed attempts in the past, define the state-of-the-art, and conclude with general projections into the future considering theoretical constraints and established trends. This section also, sparingly, draws on comparisons of coprocessor technologies.
Sections 3 and 4 discuss the current and near-future computing technology products for general purpose processors and accelerators/GPUs/FPGAs (Field Programmable Arrays). They both include technical specifications not discussed on purpose in Section 2.
The last section (Section 5) focuses on interconnects at two important levels: the first considers the high-speed and low-latency interconnects used to run massive MPI computations and the second focuses on local interconnects of computing components needed to improve data movement and ensure cache coherency within a single node.

Finally, building on top of the understanding of theoretical concepts and commercially available - or soon to be available - products, the conclusion section provides an outlook on future trends and summarises mid-term projections (3 to 5 years) about what users can expect to be available in the near future.

Readers may notice that some technologies are covered in more details than others. This is mainly due to the following reasons:

(1) Time elapsed since release on the market (the longer, the more information is available); as an example, x86 has been widely adopted by the market since the 2000s until today,
(2) Adoption by the market: while IBM Power and Arm technologies have been on the market for a while, both technologies are not so widespread nowadays; just looking at the June 2020 Top500, there are 10 supercomputers in total that are based on Power processors while 469 (93,8%) supercomputers are powered by x86 and only 4 are Arm-based.
(3) Size of the market (the larger, the more information is available) as information on widespread technologies is easier to find that information on niche/emergent technologies, for which technical details are generally less accessible.

2. Key Factors in Processor Performance

2.1. Manufacturing Process

The discussion about processor technologies should start with the manufacturing process, as it governs every aspect of processor capability. The basic element of processor design is the transistor, with the connection layout of transistors forming the architecture. Moore's Law [4] offers an observation on the industrial progress rate rather than a physical law and often is misquoted to claim that "processing power doubles every N months". Actually, this law refers to the number of transistors in a processor doubling every 18 months. The number of transistors in a processor is defined by the size of its die and the possible density of transistors. While Moore's Law was a good match for the actual development for the past decades, it has no more reflected the reality for the last 2 or 3 years. This observation is based on transistor density's limitation, closely linked to manufacturing process technology. The die area of processors has remained largely constant, as it is limited by the communication times across the die, in addition to manufacturing constraints (larger dies mean lower yield). The transistor density, in turn, is limited by the minimum feature size at which a transistor can operate reliably. Previous projections predicted a limit of manufacturing processes at 5nm, where they were expected to suffer from quantum tunnelling effects resulting in substantial power drain and heating effects which would both influence high-speed and reliability of results. However, a 1 nanometre wide transistor was shown to be possible in the past and Taiwan Semiconductor Manufacturing Company (TSMC) has already announced that 3-nanometre chips will come on the market in 2021. Furthermore, the junction – the actual gate that is switched on or off to create transistor functionality – could be as small as a single atom, but the gate contacts will remain relatively large on such designs. The possible breakthroughs that may overcome the density barrier are the use of photonics instead of electronic information transfer and utilisation of 3D (or pseudo-3D) stacked designs. The former is theoretical and the latter finds a limited use in memory design at the time of this technology report.

Currently, the most widely deployed processors in HPC, e.g. from Intel, are manufactured at 14nm level (Intel Cascade Lake), with a 10nm manufacturing process announced in Q4 2020 for Intel Ice Lake while other foundries or manufactures like TSMC and Samsung already offer 5-7 nanometre designs. However, it should be noted that the labels for commercial products, including marketing labels such as 7nm+ or the 10nm "refresh" iterations from Intel should not be taken to depict literal transistor sizes. Designs labelled as 14nm, for example, were "substantially" larger than 14 nanometres and Intel's 10nm is comparable to other foundries' 7nm, while Intel's 7nm is comparable to other 5nm processes. On this subject, Godfrey Cheng from TSMC, is quoted as follows: "I think we need to look at a different descriptor for the nodes beyond what we currently see today", here the term "nodes" refers to the transistors, currently described (inaccurately) by their size.

2.2. CPU Frequency

A dominant factor in processing power is the operational frequency. It strongly impacts the number of operations per second. However, the technological limits, most importantly power and temperature budgets, have also led to a stall in the increase in frequency. The 3.6+ GHz levels, common in the era of Intel's Pentium 4, have been replaced by multicore designs operating at modest 2.4 GHz, making up the reduced speed at which instructions complete by retiring more instructions per cycle. Another recent trend is the utilisation of "boost levels", where parts of the processor are able to operate at higher frequencies as long as the temperature and power remain in the limits of the processor. These "boost levels", coupled with the recent advancements in manufacturing processes, may signal the return to frequency levels of the previous era.

2.3. Instruction Sets

Besides the increased core counts to be discussed below, the main boost in recent processor performance comes from complex new instructions. These allow higher level operations to be expressed in denser instruction series (most of the time reducing the stress on the memory bandwidth).

A simple, yet powerful, example is the fused-multiply-and-add (FMA) instruction, which executes the a = b * c + d operation in a single step. Most modern architectures support this instruction in some form (e.g. FMA3 updating the target in place, or FMA4 writing the result to a new location), while the design decisions have evolved over time (i.e. Intel preferred FMA3 in AVX2, while AMD initially used FMA4 but now allows both). NVIDIA took this instruction a step further in its GPU architecture, executing FMA over 4x4x4 tensors in a single step. In the future we might expect CPU architectures to also incorporate such tensor operations as pressure from AI and Machine Learning (ML) users increases. Currently, as a result of these demands, Intel and some others (e.g. Arm and IBM) have included some additional ISA instructions for these specific workloads, such as Vector Neural

Network Instructions (VNNI) as part of AVX-512 for Intel. However, it should be noted that these instructions and the tensor operations provided by NVIDIA, operate on half precision floating point numbers for intermediate values as double precision implementation through Floating Point 64bits (FP64) Tensor Core. This precision level, sufficient for common AI applications, is also deemed acceptable for Molecular Dynamics (MD) simulations in some cases, covering another large portion of the HPC portfolio. The usage of reduced precision for numerical simulation is also an active area of research in terms of numerical methods. The half precision is formalised in the BFLOAT16 format, which instruction sets from Intel and Arm both supports. AMD's support for BFLOAT16 is currently limited to its GPU products, but future CPUs from AMD may also follow this trend.

2.4. Vector Length

The operations on lower precision floating point numbers also influence the vector instructions. The optimal length of vectors in vector operations is a subject of an ongoing debate. The length of vector operands in such instructions has been steadily increasing in the x86 architecture for the past few years, eventually reaching 512 bits width in AVX-512 extensions implemented by Intel. These 512 bits wide instructions debuted with the release of Intel's Xeon Phi architecture and were carried over to the Xeon architecture in the Skylake series. However, Intel's implementation came with the caveat that the operating frequency had to be reduced during the execution of these instructions which led to an AVX frequency much below the nominal frequency, allowing the processor to run under its specific TDP power. Regarding extensions to the x86 instruction sets, AMD has typically supported Intel's new instructions for compatibility, and subsequently AVX2 (Intel's 256 bits wide vector extensions) were implemented in AMD's Rome processors. AVX-512 has been an exception until now and since it did not reach the expected level of performances for most of the applications, it is not sure to be adopted in a future AMD design. Furthermore, Intel might also focus on 256 bits wide instructions, leaving 512 bits instructions for compatibility and special cases. It should be noted that other architectures, besides x86, have competing (yet not directly comparable) solutions, with SVE (Scalable Vector Extension) from Arm supporting a variable width ranging from 128 to 2048 bits as opposed to x86's fixed vector length.

The issue with introducing complex higher-level instructions, besides introducing complexity taking up valuable real estate on the die, is that such complex operations need a large number of cycles to complete. During the heated debate between Complex Instruction Set Computer (CISC) and Reduced Instruction Set Computer (RISC) design philosophies, the RISC proponents argued against the complex instructions for this very reason. Modern CPU architectures are CISC architectures, featuring a very diverse set of highly specialised functions. However, these CPUs handle these instructions by decoding them into smaller, RISC-like sub-operations and dispatching them into appropriate Execution Units (EUs). This allows software to be represented in a terse chunk of code in memory, reducing the cost in memory accesses. At the same time, it allows finer control over execution order inside the processor, thus increasing efficiency by hiding latency via pipeline exploitations or even multiplexing access to resources by Simultaneous Multithreading (SMT). Therefore, additional criteria for comparing architectures and their several iterations are pipeline depth and number of threads in SMT. The discontinued Netburst architecture from Intel took the pipeline depth to an extreme level by featuring 31 stages. The discontinuation of this architecture should indicate the caveats in implementing very long pipelines, and current architectures follow a more modest approach (e. g., 14-19 in Intel's Ice Lake and 19 in AMD's EPYC).

2.5. Memory Bandwidth

Memory bandwidth has been one of the key factors for CPUs to perform efficiently, both benefiting to memory bound applications as large-scale use cases to feed the CPU cycles as fast as possible and keep up with processors increased computing power. For the last few years, processor technologies have been designed with higher memory bandwidth: first, through a higher number of channels per processor; while in 2011, an X86 processor like the Intel Sandy Bridge had 4 memory channels, now X86 (Intel/AMD) and Arm processors have, in 2020, typically between 6 and 8 memory channels, allowing a theoretical global memory bandwidth performance improvement of +50% and +100%. The second criterion to take into account is the frequency of the DDR memories which has drastically improved over time. While in 2011, an X86 processor like the Intel Sandy Bridge supported DDR3 running at 1600 MT/s, an X86 processor and an Arm processor are, in 2020, supporting DDR4 running between 2933 MT/s and 3200 MT/s. Taking into account both the increase in the number of memory channels and the DDR technology improvement, the global memory bandwidth per processor for an X86 processor improved by a factor of 4, while the memory bandwidth per core remains nearly the same due to density growth on current available chips. A new trend is driving the market to High-Bandwidth memory (HBM) both within CPU and accelerators, providing high throughput memory access for applications (>= 1TB/s vs maximum 205 GB/s for the current AMD and Marvell ThunderX processors available on the market – minimum 5 times the memory bandwidth of DDR depending on the HBM type) and a more balanced byte-per-flop ratio than then one supported by DDR-only

processor technology. However, HBM provides a maximum capacity of 32GB (HBM2) today requiring transfers between DDR and HBM in case the data does not fit the HBM size.

2.6. Simultaneous Multithreading (SMT)

In SMT, multiple instructions are issued at each clock cycle, possibly belonging to different threads; thus increasing the utilisation of the various CPU resources and making it possible to reduce the effect of memory latency. SMT is convenient since modern multiple-issue CPUs have a number of functional units that cannot be kept busy with instructions from a single thread. By applying dynamic scheduling and register renaming, multiple threads can be run concurrently. Regarding SMT, it should be noted that the number of hardware threads in x86 is extremely low (i.e. 2), compared to other architectures: POWER from IBM supports up to 8 threads per core (SMT8) and various Arm-based processor implementations feature up to 4 threads per core (SMT4). It is possible to explain this discrepancy by multiplexing features such as register windows being absent in x86. GPU architectures prefer a simpler, straightforward approach to instruction complexity, spending die area real estate in multiplication of EUs to increase parallelism instead of featuring complex instructions. It is possible to implement this approach in CPUs by building semi-independent vector operation sections in future models.

2.7. Processor Packaging

Another, largely orthogonal step towards achieving an increase in computational power is putting multiple individual cores onto a single die. Since the stall in frequency increase has begun to present itself, building multicore CPUs has been the preferred solution to continue increasing the processing power. However, the increase in core counts was limited by the need for communication and synchronisation between the cores and the difficulty in increasing the number of transistors on a single monolithic-die with a manufacturing process reaching nanometer level. This need arises largely from the constraint of maintaining a consistent shared memory to prevent race conditions. One approach to this subject is making the discrepancy and memory ownership explicit by presenting a Non-uniform Memory Architecture (NUMA), similar to the pre-existing multi-socket arrangements or chiplets. This is achieved through the MCM (Multi-Chip Module) concept. MCM is an electronic assembly of multiple chips or semiconductor dies, also called chiplets, that are integrated usually onto a unifying substrate, so that it can be treated as a larger integrated circuit. The chiplets are then connected through an intra-chiplet interconnect, as for example the Infinity Fabric (IF) interconnect for AMD zen2 and further AMD generation processors. While MCM has been early adopted by companies like IBM or AMD to increase core density (rather than clock speed) on a processor, Intel has decided so far, to remain with its monolithic chip architecture for general purpose processors (except for their Cascade Lake-AP processor) despite all the complexity of the manufacturing process faced at 14 and 10nm levels. MCM presents large advantages: It has helped AMD both to enter the market earlier than its main competition and reduce the price of their processor. MCM now underwent a broader adoption by the market and is a manufacturing process well mastered by foundries like TSMC. Some drawback at application level: As the processor presents several NUMA domains, it requires a strong knowledge of the processor micro-architecture to support suitable task placement.

2.8. Heterogeneous Dies

Planting heterogeneous cores that are specialised in different application areas in a single die is not a new approach. The Cell Broadband Engine Architecture from IBM, usually called Cell in short, has combined a two-way SMT general purpose PowerPC core with eight Synergistic Processing Elements (SPEs) specialised in vector operations on a single die. Unfortunately, despite being able to provide high performance, it has been a victim of its radical approach and has never become popular, suffering a fate similar to Itanium from Intel.

The general philosophy of combining heterogeneous compute elements, however, is not abandoned. In fact, it has been observed multiple times in the computing industry that the rise of co-processors has been inevitably followed by integrating them into the central processor, resulting in cyclical trends. The rise of Integrated Graphics Processing Units (iGPUs) could be given as a general example, and the fate of Xeon Phi architecture represents an interesting twist in the HPC world. The implementation of heterogeneous dies takes the form of big.LITTLE in the Arm architecture, combining low frequency energy efficient cores with high performance ones, but while reducing power consumption, the utilisation is limited in the HPC area. A different approach is exemplified in the efforts of AMD, where vector operation focused GPU cores are being moved into the unified address space and die space of the central processor.

In terms of the aforementioned observation of cyclical trends, the industry is at the early stages of the integration phase, where discrete co-processor products from NVIDIA dominate the market, but the demands for unified

memory address space and simpler programming models put pressure on the vendors. As an extreme example, the ominously named 'The Machine' from HPE has proposed a universal address space operated on by various specialised processors, connected not by a single continuous die, but a high-speed interconnect based on photonics. The future of this ambitious project, however, is unclear: widespread adoption is unlikely, based on the fate of such radical departures from the traditional model in the past.

3. General Purpose Computing Capabilities

3.1. X86_64 Processors

3.1.1.Intel X86_64

After years of delays, Intel's 10nm designs have finally seen the light of day in the Ice Lake series of the 10th generation Intel Core processors for desktops. However, even after this long delay, the new designs' yields and clock speeds have been generally unimpressive, resulting in the products of the ageing 14nm+ (and further iterations denoted by additional + symbols in their names) to continue being offered alongside with the newer, 10nm process-based ones for desktop and server platforms. This has also been the case for the HPC market for almost 4 years now, starting with the 5th Intel core processor generation code named Broadwell released in 2016 and ending with the Intel Copper Lake-SP processor for the HPC segment market that will be delivered at the earliest at the end of Q2 2020 for key customers in specific configurations (Cedar Island Platform only). Lack of satisfactory progress in this area has also been admitted by Intel with its CFO, George Davis, recognising that the 10nm process has not been as profitable as its long exploited 22nm and 14nm processes were [5].
The first 10nm CPU for HPC workloads, code named Ice Lake-SP (Whitley Platform) [6], should be available by the end of 2020. Its main purpose should be driving the path to 10nm mass production with the expected new Intel product called Sapphire Rapids [7] (together with a new LGA 4677 socket) that should be deployed in 2 phases: the first version might be based on Sapphire Rapids without HBM (High Bandwidth Memory), supporting up to 8 memory channels DDR5 and PCI-e gen5 as NVDIMM memory. The second step may add several important features such as the capability to interface with the new Intel GPU Called Intel Xe HPC "Ponte Vecchio" (PVC) (see the GPU section below for more details). The latter ensures a unified memory architecture between the Sapphire Rapids CPU and Intel PVCs GPU through the Xe links based on the CXL (Compute Express Link) standard.
In terms of the manufacturing process, the currently released plans from Intel state that 7nm might be available in 2021 and that Intel's 5nm should be released in 2023. When putting these improvements into perspective, Intel mentioned that its 10nm process is comparable to competing products labelled as 7nm in TSMC, also that their 7nm process is roughly equivalent to TSMC's 5nm process, with Intel's 5nm being similar to TSMC's 3nm [8]. As for frequency, the high-end products from Intel reach 3.0 GHz, but 2.5 to 2.7 GHz models are expected to be the common choice. As with the current trend, the focus is more on the boost levels, achieving 4.0 GHz for a single core. At least for the lifetime of the 10nm process, these frequency levels should not change radically. However, some of the improvements to the processor performance may come in the form of instructions that do require lower operating frequencies (as AVX-512 does today). In addition to extensions to AVX-512, widespread adoption of BFLOAT16 (starting within the Intel Copper Lake processor which should be delivered to large key AI customers) and Galois Field New Instructions make up the major changes to the instruction set in new products. A radical addition, however, is the inclusion of Gaussian Neural Accelerator v1.0 (GNA) in client version of Ice Lake. However, GNA is a low power, inference focused component, and it is unclear if it will be included in the server version.

3.1.2.AMD X86_64

The current AMD EPYC 7002 Series Processors (Rome) is the second generation of the EPYC x86 platform. Its implementation relies on an MCM (Multi-Chip Module) implementation and on the Zen2 core architecture built in a 7nm process technology to support up to 64 compute cores on a single CPU, enhanced IO capabilities through 128 lanes of PCI-e Gen4 I/O and 8 DDR4 memory channels (with up to 2DIMMs per channel) running at up to 3200MT/s, boosting memory bandwidth up to 204.8 GB/s peak. The chiplet of a MCM in AMD language is called CCD (Compute Core Die), with one CCD supporting up to 8 compute cores. The AMD Rome processor supports up to 280W TDP (Thermal Design Power) per socket and up to 3.4 GHz Turbo boost CPU clock speed.
The upcoming AMD Zen3 core is expected to enter mass production in Q3/Q4 2020 and might rely on TSMC's advanced 7nm processor, N7+. While the Zen2 cores are the main components of AMD Rome processor, the Zen3 cores will be the main components of the AMD Milan processor [9]. The main differences between Zen2 and Zen3 implementation should be the clock speed and micro-cache level architecture implementation on a CCD as the memory. The former means that at equivalent core counts the Zen3 core should be capable of operating at higher

frequencies targeting increased per core performance. The latter means reducing the number of NUMA domains inside a single CPU, having 8 CCD Zen3 cores sharing now 32 MB L3 cache on a single CCD (one core being capable of addressing 32 MB memory) while previously one single core was capable of addressing a maximum of 16 MB L3 cache. While the maximum memory bandwidth was sustained with 16 Zen2 cores on the Rome processors, the optimal number of Zen3 cores to achieve the best memory bandwidth might then be 8 cores. In addition, the Zen3 cores may have the capability to run Secure Encrypted Virtualisation-Encrypted State (SEV-ES) without any specific software changes. This feature would allow to encrypt all CPU register contents when a VM (Virtual Machine) stops running and prevent the leakage of information in CPU registers to components like the hypervisor. It can even detect malicious modifications to a CPU register state.
One key point to note is that, while the CPU frequency for Naples was far from the ones seen on Intel Skylake processor, the AMD Rome (and the future AMD Milan) clock speeds are now comparable to Intel's products, being slightly lower, 2.6 GHz for the highest core count top level variant. What makes the AMD Rome/Milan competitive with Intel is the density on the AMD processors, reaching up to 64 cores.
However, there will also be low-core count SKUs (Stock Keeping Units) (24, 16 and 8 cores variants) with higher frequencies. As for the instruction set, the main difference to Intel is that AMD Rome/Milan only support AVX-256 instead of AVX-512. However, the lack of wide vectors is made up by the fact of AMD having its own high-performance GPU line, and the plans for integrating them onto the same die. Furthermore, the higher core count also results in more vector processing units being available even without integrated co-processors. There are no disclosed plans from AMD to include a tensor processor, similar to Intel's GNA or VNNI feature, simply due to the fact that AMD now has its own high-performance GPU line (MIXX line).
The next generation of AMD general purpose processor should be based on Zen4 cores and should form the Genoa processor as announced [10]. This new microarchitecture is expected to be based on a 5nm process technology and might incorporate new features as DDR5, PCI-e gen5, HBM support and cache coherency (between CPU and GPU). It is already well-known to be the CPU that will power one of the three Exascale machines announced by the US DoE (Department of Energy), El Capitan (LLNL) in 2022. The processor on the Frontier machine (another of these three exascale machines) should be a custom Milan SKU, a Milan ++ probably with some of the Genoa capabilities. This Genoa processor should again enhance the strong I/O capability of the AMD processor providing again more IO capabilities and higher memory bandwidth which should benefit memory bound applications. While it is known that this new AMD processor might introduce a big step in the AMD roadmap, there is little public information available on the Genoa processor now [11].

3.1.3.Comparison of main technical characteristics of X86_64 processors

The Table 1 summarises the X86_64 processors main technical characteristics.

Chip maker	Intel			AMD		
Processor	Cascade Lake SP	Ice Lake	Sapphire Rapids	Naples	Rome	Milan
Platform	Purley	Whitley	Eagle Stream	EPYC	EPYC	EPYC
Core	Cascade Lake	Ice Lake	Sapphire Rapids	Zen	Zen 2	Zen 3
Manufacturer/Foundry	Intel	Intel	Intel	TSMC	TSMC	TSMC
Manufacturing Process (nm)	14	10	10	14	7	7
Status	Launched	Planned	Planned	Launched	Launched	Planned
GA or Estimated Availability	April 2019	Estimated Q4 2020	N/A	June 2017	August 2019	Estimated Q4 2020 - Q1 2021
Technology	Single-die	Single-die	N/A	MCM	MCM	MCM
Intra-node interconnect	UPI	UPI	UPI/CXL	PCI-e gen3	Infinity Fabric	Infinity Fabric
Extra-node interconnect	PCI-e gen3	PCI-e gen4	PCI-e gen5	PCI-e gen3	PCI-e gen4	PCI-e gen4
SMT	2	2	2	2	2	Min 2
ISA	AVX512	AVX512	N/A	AVX	AVX2	AVX2
Operations	2xFMA @512b	N/A	N/A	2x(ADD,FMA) @128b	2x(ADD,FMA) @256b	2x(ADD,FMA) @256b
Cores	Max 28	N/A	N/A	Max 32	Max 64	Max 64
channels/skt	6	8	8	8	8	8
DDR @ Memory Clock Speed	DDR4 @2933	DDR4	DDR5	DDR4 @2667	DDR4 @3200	DDR4 @3200
Theroritical Bandwidth (GB/s)	140,8	N/A	N/A	170,7	204,8	204,8
HBM @Memory BW (TB/s)	No	No	Maybe	No	No	No
Estimated Theoritical Gflops/Watt (Top bin)	11.8	N/A	N/A	3.13	9.51	9.30

Table 1: X86_64 Intel and AMD processors main technical characteristics

Note: N/A means the information is not available.

3.2. Arm Processors

3.2.1.EPI (European Processor Initiative)

The European Processor Initiative (EPI) [13] is in charge of designing and implementing a new family of low-power European processors designed to be used in extreme scale computing and high-performance Big Data applications as in the automotive industry.
EuroHPC [12] has an objective to power 2 European Exascale machines in 2023-2025, with at least one of them built with a European processor technology, hopefully a result of the EPI. In addition, EuroHPC also plans the acquisition of pre-Exascale systems (2021-2024/2025) and support for the first hybrid HPC/Quantum computing infrastructure in Europe.

The EPI product family will mainly consist of two computing products: an HPC general purpose processor and an accelerator. The first-generation of the general-purpose processor family named Rhea will rely on Arm's Zeus architecture general purpose cores (Arm v8.3/v8.4; up to 72 cores [14]) and on highly energy-efficient accelerator tiles based on RISC-V (EPAC – an open-source hardware instruction set architecture), Multi-Purpose Processing Array (MPPA), embedded FPGA (eFPGA) and cryptography hardware engine. First Rhea chips are expected to be built in TSMC's N7+ technology aiming at the highest processing capabilities and energy efficiency. The EPI Common Platform (CP) is in early development and may include the global architecture specification (hardware and software), common design methodology, and global approach for power management and security in the future. The Rhea chip will support Arm SVE 256 bits (Dual Precision, Single Precision, BFLOAT16), HBM2e, DDR memories and PCI-e gen5 as HSL (High Speed Links), which would support the interconnection of two Rhea dies or one Rhea die with an HPC accelerator like Titan Gen 1 (based on RISC-V instead of Arm). The Zeus cores and the memory subsystems (built on top of HBM, DDR and Last Level of Cache) will be connected through a Memory-coherent-on-chip network. The CP in the Rhea family of processors will be organised around a 2D-mesh Network-on-Chip (NoC) connecting computing tiles based on general purpose Arm cores with previously mentioned accelerator tiles. With this CP approach, EPI should provide an environment that can seamlessly integrate any computing tile. The right balance of computing resources matching the application needs will be defined through the carefully designed ratio of the accelerator and general-purpose tiles. The Rhea chip will support PCI-e and CCIX to interconnect and accelerators.
The second general purpose chip family is named Cronos (Rhea+) and should be based on the Arm Poseidon IP possibly with enhanced capabilities like Compute Express Link (CXL) built-in to ensure cache memory coherency between the CPU and the accelerator.

The Rhea chip and its next generation are designed and commercialised by SiPearl (Silicon Pearl). SiPearl is a European company that is using the results of the European Processor Initiative (EPI) project.

3.2.2.Marvell ThunderX

The current Marvell ThunderX processor on the market is the well-known ThunderX2 processor which has been available since 2018. The ThunderX2 is the second generation of Marvell 64-bit Armv8-A processors based on the 16nm process technology. It is also the first processor which has demonstrated the capability to compete with Intel and AMD. It is available with up to 32 custom Armv8.1 cores running at up to 2.5 GHz and supports Simultaneous Multi-threading (SMT) with 4 threads per core, so twice the number of threads compared to x86 processors. The ThunderX2 die is built on top of a monolithic implementation like all Intel processor generations up to Cascade Lake, in contrast to AMD with its Multi-Chip implementation (MCM). Each processor supports up to 8 memory channels or 8 memory controllers with up to 2 DPC (DIMM Per Channel), with DDR4 DIMMs running at up to 2667 MT/s (1DPC only). The processor has strong I/O capabilities with up to 56 PCI-e gen3 lanes and 14 PCI-e controllers along with integrated I/O and SATAv3 (Serial ATA) ports. Each ThunderX2 core has a dedicated L1 cache (32 KB instruction and data cache) and a dedicated 256 KB L2 cache. The L3 cache is 32 MB and is distributed among the 32 cores. In terms of computation, the Marvell ThunderX2 supports 128 bits NEON instructions (Arm ISA) and up to 2 FMA EUs, which means that each core is capable of executing 2 FMA instructions using a 128 bits vector during a single cycle. This has led to one core being capable of running 8 DP floating operations per second. The Marvell ThunderX2 socket is available as single or dual-sockets server with CCPI2 (Cavium Cache Coherent Interconnect) providing full cache coherency.
The following part of this section, regarding Marvell ThunderX3 processors, was written before the cancellation of ThunderX3 processors by Marvell and presents what was initially planned by Marvell before this cancellation. The authors of this report have decided to maintain the information regarding ThunderX3 processors for several reasons: (1) In case Marvell decides to sell their design to another chip maker, information will be known by users (2) To provide information about what could be achievable at the horizon of 2021.The next ThunderX line

processors should have been ThunderX3+ and ThunderX3 [15] [16], based on TSMC's 7nm lithography, again with monolithic chips rather than chiplets. Both CPUs should target different markets: ThunderX3+ and ThunderX3 should have focus on cloud and High-Performance Computing workloads, respectively, due to their internal properties. The ThunderX3 and X3+ were supposed to be based on ArmV8.3+ (+ means that it includes selected features of the ArmV8.4 ISA). The ThunderX3 was planned to be built on top of a single die, with up to 60 Armv8.3+ cores and 8 DDR4 memory channels running at 3200 MT/s supporting up to 2 DPC while the ThunderX3+ is planned to be built on top of 2 dies, each with 48 cores for a total of 96 cores, with also up to 8 aggregated channels (4 per die) DDR4 at 3200 MT/s, leading to the same global memory bandwidth on a SoC but a lower memory bandwidth per core, though giving penalty to this for memory bound applications. On the ThunderX3+, the 2 dies are interconnected through the new CCIP3 (Cavium Cache Coherent Interconnect) over PCI-e gen4. Each processor was designed with 64 PCI-e gen4 lanes over 16 PCI controllers and a 90 MB L3 shared cache while L1 and L2 caches remain single to each core.
Like their predecessor ThunderX2, ThunderX3/X3+ were expected to support SMT4, leading to 384 and 240 threads on ThunderX3+ and ThunderX3, respectively. These processors should have supported NEON (SIMD - Single Instruction Multiply Data) instruction sets with 4 FMA per cycle combined to 128 bits EUs and 16 operations per cycle. The native/base clock speed for ThunderX3+ should rather be around 2 to 2.2 GHz, while this would be increased by 2 bins (200 MHz) for ThunderX3, reaching 2,1 TFlops+ (minimum) peak performance for the HPC version of ThunderX3, for a TDP reaching 200W (minimum). The TDP was expected to depend on the core clock speed provided on the CPU, since while the clock speed will be higher, the TDP might also increase. The design should have come in both 1 and 2-socket configurations, and the inter-socket communication CCPI 3rd generation.

At this point in time, it is not clear if Marvell will pursue their ThunderX line.

3.2.3.Fujitsu A64FX

The Fujitsu A64FX is an Arm V8.2 64bits (FP64) processor designed to handle a mixed processing load, including both traditional numerical simulation HPC and Artificial Intelligence, with a clear target to provide an extremely high energy-efficient performance (performance/watt) and a very good efficiency for a large spectrum of applications. Built on top of a 7nm TSMC process technology, its design has been HPC optimised being the first general purpose processor supporting 32 GB HBM2 (around 1 TB/s aggregate - 4x 256 GB/s) and native hardware SVE, while considering various AI instruction set extensions, such as supporting half precision (FP16) and INT16/INT8 data types. The A64FX has 48 compute cores and 4 additional assistant cores to process the OS and I/O. Like AMD, Fujitsu has chosen to build its processor based on MCMs. These modules are called CMG (Core Memory Group) in the Fujitsu design. The compute and assistant cores are split into 4 CMGs, each with 12 compute cores and 1 assistant core sharing one 8MiB L2 cache (16-way) through a cross-bar connection and accessing 8GB HBM2 through a dedicated memory controller (maximum 256 GB/s between L2 and HBM2). In addition, each core has its own 64 KiB L1 cache and supports 512-bit wide SIMD SVE implementation 2x FMAs, leading to around 2.7 TFlops/s DP on a single A64FX processor. The 4 CMGs are connected by a coherent NoC capable of supporting Fujitsu's proprietary Tofu interconnect and standard PCI-e gen3 [17] [18] [19].

The Fujitsu A64FX is provided as a single socket platform only, while most of its competitors have chosen to provide single- and dual-socket platforms. The processor powers Fugaku, the first Exascale class machine in Japan and worldwide in 2020 timeframe. The machine has been built on top of more than 150,000 (158,976) compute nodes in more than 400 racks. The nodes are connected by a Tofu-D network running at 60Pbps, reaching 537 PF+ FP64 peak performance with access to more than 163 PB/s theoretical memory bandwidth. Data is stored on a hierarchical storage system with 3 levels: the 1st layer relies on high throughput NVMe (Non-Volatile Memory Express) GFS cache/scratch file systems (>30 PB), the 2nd layer is a high capacity Lustre-based global file system (150 PBs) based on traditional magnetic disks, and the last layer is currently planned to be an archive system stored off-site on a cloud storage. The cooling infrastructure relies on DLC to reach a 1.1 PUE. This system required around 30 MW during benchmarking as reported in June 2020 TOP500 list. Fugaku's installation started in December 2019 and was completed by mid-May 2020. The system should be fully operational and open to the user community in 2021. A key point was the capability of Fujitsu to power a Fugaku like prototype for SC19, which was ranked number 1 in the November 2019 Green500, with a machine based on the General-Purpose Processors only, while most of the other systems at the top of the Green500 list mainly rely on GPUs. The prototype was powered with 768 A64FX CPUs supporting the Arm SVE instructions for the first time in the world. This performance measurement demonstrated that Fugaku technologies have the world's highest energy efficiency, achieving 1.9995 PFlops sustained performance against 2.3593 PFlops as peak performance, and 16.876 GFlops/W (Gigaflops per watt).

In addition, early February 2020, Fujitsu announced that it would supply the Nagoya University Information Technology Center with the first commercial supercomputer powered by the Arm-based A64FX technology. The new system will have 2,304 Fujitsu PRIMEHPC FX1000 nodes, offering a theoretical peak performance of 7.782 PFlops and a total aggregated memory capacity of 72 terabytes. In the meantime, other customers have acquired Fujitsu A64FX systems mostly as test beds for now, e.g. the Isambard 2 system from University of Bristol and the Wombat cluster at Oak Ridge National Laboratory. The 4th Petascale EuroHPC supercomputer, the Deucalion machine at Minho Advanced Computing Centre (Portugal) should be equipped at least with a large partition relying on Fujitsu A64FX processor.

Two other Arm processors, the Graviton from Amazon and the Altra from Ampere, are described in the next sub-sections, even though these 2 processors are more dedicated to compete with AMD and Intel x86 processors for the data centre market rather the HPC market. These 2 platforms are based on Arm Neoverse (ArmV8.1/2) microarchitecture, which is the same Arm platform than the Fujitsu A64FX processor.

3.2.4.Amazon Graviton

As the dominant cloud service provider, Amazon has a keen interest in cost-efficient cloud services and started working on an Arm-based SoC in 2015. Amazon recently announced its second-generation Graviton processor based on Arm's new Neoverse N1 microarchitecture implemented on top of Arm v8.2 and a CMN-600 mesh interconnect. This second generation offers a monolithic die of 64 cores running at 2.5 GHz along with 64 KB of L1 and 1 MB L2 private data caches and a shared L3 cache of 32 MB; it is manufactured in TSMC's 7nm process. Clearly, Amazon also targets accelerated platforms given that the Graviton provides 64 PCI-e 4.0 lanes compared to, for example, the 16 PCI-e 3.0 lanes of the A64FX. Further characteristics of the chip show that it is designed for rather compute-intensive workloads (or co-scheduled compute- and memory-intensive workloads), with 8-16 cores already able to max out the available peak memory bandwidth (which is also the case for the AMD processor). Also, Machine Learning workloads are in focus with explicit support for INT8 and FP16 data types in order to accelerate AI inference workload. Given its single-NUMA design, it is also optimised for low core-to-core latencies across the entire chip compared to more conventional architectures of comparable core numbers that rely on multi-die designs (AMD) or multi-socket designs (Intel), both of which show a significant increase in latency for inter-NUMA communication. This is further illustrated by high scalability results, for example for most SPEC 2017 benchmarks that are not memory-bound. Nevertheless, also sequential performance is competitive with x86-based systems, as the second-generation Graviton shows a large boost over the first generation and preliminary benchmarks reach between 83% and 110% of sequential compute performance compared to systems employing the Intel Xeon Platinum 8259 (Intel Skylake Lake, the microarchitecture previous to Intel Cascade Lake) or AMD EPYC 7571 (AMD Naples) processors, while doubling or tripling the achievable single-core memory bandwidth performance of these Intel and AMD processors.
Both GCC 9.2.0 as well LLVM 9 currently offer support for the Neoverse N1 microarchitecture and hence facilitate a fast adoption rate of software ecosystems for the Graviton. Given Amazon's integration of Graviton in their new M6g instance types, current benchmark results and a "40% better performance per dollar than its competition" claim, this chip might introduce a massive change to Amazon's traditionally x86-heavy cloud solutions and hence the entire data centre market [20].

3.2.5.Ampere Altra

Another Arm CPU architecture is built by Ampere and aims at challenging Intel and AMD for the data centre market. Ampere's new Altra CPU is a TSMC-manufactured 7nm chip with up to 80 cores clocked at a maximum of 3 GHz and a TDP of 210 Watts. Similar to Amazon's second-generation Graviton it implements the Arm v8.2 architecture (along with some additional features not yet part of this standard), supplies each core with dedicated 64 KB of L1 and 1 MB of L2 cache and allocates a total of 32 MB shared L3 cache for all cores (yielding slightly less L3-per-core than the Graviton and falling well below Arm's recommendation of 1 MB per core). It also features 8 DDR4-3200 memory controllers with the same peak memory bandwidth (204.8 GB/s) than AMD Zen2/Zen3 processor. In contrast to the Graviton, it offers a much larger contingent PCI-e 4.0 lanes per CPU (128) and enables dual-socket setups. The chip reserves 32 of these lanes for inter-socket communication, leaving up to 192 lanes and hence a maximum throughput 378 GB/s for accelerator communication in a fully stacked node. This exceeds the 160 lanes offered by the current leader in this field, AMD, in dual-socket configurations (128 lanes per CPU with at least 48 dedicated to inter-socket configuration). While reliable, independent benchmark data is not yet available, Ampere claims a performance matching that of contemporary AMD systems (AMD Rome EPYC 7742) and outperforming Intel (Cascade Lake SP Xeon Platinum 8280) by a factor of 2. The chip targets applications in data analytics, AI, databases, storage, Edge Computing and cloud-native applications. Ampere plans to continue this line of processors with the next model Mystique following in 2021, using the same socket as Altra, and Siryn following in 2022 at a planned 5nm process [21] [22] [23] [24].

3.2.6.Summary of Arm processors' main technical characteristics (dedicated to HPC)

The Table 2 summarises the Arm HPC processors main technical characteristics.

Architecture	Arm					
Chip maker	Marvell			Fujitsu	EPI	
Processor	ThunderX2	ThunderX3	ThunderX3+	A64FX	Rhea	Chronos
Platform	N1	Zeus	Zeus	N1	Zeus	Poseidon
Core	ARMv8.1	ARMv8.3+	ARMv8.3+	Armv8.2	ARMv8.3/8.4	x
Manufacturer/Foundry	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
Manufacturing Process (nm)	16	7	7	7	7/6+	5
Status	Launched	Cancelled	Cancelled	Launched	Planned	Planned
GA or Estimated Availability	May 2018	None	None	Q4 2019	Estimated 2022	Estimated 2023-2024
Technology	Single-die	Single-die	Dual-Die	CMG	Chiplet	Chiplet
Intra-node interconnect	CCPI	CCPI	CCPI	NOC	CCIX	CCIX/CXL
Extra-node interconnect	PCI-e gen3	PCI-e gen4	PCI-e gen4	PCI-e gen3	PCI-e gen5	PCI-e genx
SMT	4	4	4	4	N/A	N/A
ISA	NEON	NEON	NEON	SVE-512	SVE-256	N/A
Operations	2xFMA @128b	4xFMA @128b	4xFMA @128b	2xFMA @512b	2xFMA @256b	N/A
Cores	Max 32	Max 60	Max 96	Max 48	72	N/A
channels/skt	8	8	8	NA	4 - 6 DDR5	N/A
DDR @ Memory Clock Speed	DDR4 @2667	DDR4 @3200	DDR4 @3200	NA	DDR5 @Min 4800	N/A
Theroritical Bandwidth (GB/s)	171	205	205	NA	230	N/A
HBM @Memory BW (TB/s)	No	No	No	32GB (4 x 8GB) @ 1 TB/s	Maybe	Maybe
Estimated Theoritical Gflops/Watt (Top bin)	3.2	9.2	N/A	>10	N/A	N/A

Table 2: Main technical characteristics of Arm processors dedicated to HPC

Note: N/A means the information is not available.

3.3. POWER Processors

Despite being a contender for the highest slots in TOP500 in recent years, the amount of publicly available information about the current plans for the POWER architecture is scarce at the moment. The June 2020 TOP500 has ranked Summit and Sierra, 2 supercomputers based on POWER9 architecture, at the second and third place. At the third place on that list with 94.64 PFlops, Sierra (introduced in the June 2018 edition of the TOP500) boasts a very similar performance level compared to the fourth contender, Sunway TaihuLight (introduced in the June 2016 edition of the TOP500) 93.01 PFlops, despite having almost an order of magnitude less cores (1.5 million vs 10.6 million) and requiring almost half the power (7.4 MW vs 15.4 MW). However, both of these contenders have been surpassed by Fugaku on the most recently published June 2020 TOP500 list.
The bulk of the processing power in these systems comes from the accelerators, namely, NVIDIA GPUs. In fact, most of the marketing material from IBM has focused on the architectural advantages of the POWER system as a whole, instead of focusing on the raw processing power of the POWER CPU alone. In particular, the early adoption of the NVLink communication protocol from NVIDIA has given the architecture a significant advantage over competitors when combined with NVIDIA GPGPUs.

Another area, where IBM had a leading edge over competitors, was the manufacturing process, which did not pan out as expected. In 2015, IBM announced that they were able to produce transistors operational at the 7nm level using silicon-germanium gates, but declined to give a product delivery date at the time. However, in 2018, Globalfoundries announced that they would be ceasing 7nm Extreme-Ultraviolet Lithography (EUV), due to the

lack of demand. This led to uncertainty in both AMD's and IBM's products and, since then, AMD has decided to work with TSMC for their Zen2 line. In late 2018, after considering all three major producers (i.e., TSMC, Samsung and, interestingly, their rival Intel), IBM opted to partner with Samsung Foundry for using their 7nm EUV process. POWER10 is available in 2 versions: 15 SMT8 cores or 30 SMT4 cores per processor while Power 9 was either 24 SMT4 cores or 12 SMT8. It supports PCI-e Gen5, wider sustained memory bandwidth (800+ GB/s as opposed to 650 GB/s in POWER9), double I/O signalling speed (50 GT/s, as opposed to 25 GT/s in POWER9) and a new microarchitecture, in addition to the 7nm manufacturing process (down from 14nm in POWER9). As for POWER11, even fewer details are available to the public, but William Starke, IBM's POWER10 architect has reiterated their preference for the chiplet design for the best utilisation of the manufacturing process in future products, in a recent interview [25]. It is to be noted also that, while support for NVLink on-chip was part of POWER8 and POWER9 architecture, it is no more the case on POWER10 with PCI-e Gen5 providing the suitable bandwidth to feed GPUs.
In the meantime, IBM has also released a new iteration of their current, 14nm based POWER9 line, featuring the new Open Memory Interface (OMI) for decoupling the memory interface from the core CPU design, in order to exploit the advances in memory technology without waiting for the release of their next generation architecture.

The Table 3 summarises the POWER processors main technical characteristics.

Architecture	POWER	
Chip maker	IBM	
Processor	POWER9	POWER10
Platform	Power	Power
Core	POWER9	POWER10
Manufacturer/Foundry	Globalfoundries	Samsung
Manufacturing Process (nm)	14	7
Status	Launched	Launched
GA or Estimated Availability	2017	2020
Technology	MCM	MCM
Intra-node interconnect	CAPI2.0/NVLink	openCAPI
Extra-node interconnect	PCI-e gen4	PCI-e gen5
SMT	12 SMT8 cores or 24 SMT4 cores	15 SMT8 cores or 30 SMT4 cores
ISA	POWER ISA V3.0	POWER ISA V3.1
Operations	2xFMA @64b	N/A
Cores	Max 24	Max 30
channels/skt	Max 8	N/A
DDR @ Memory Clock Speed	DDR4 @ 3200	DDR5 @ Min 4800
Theroritical Bandwidth (GB/s)	205	N/A
HBM @Memory BW (TB/s)	No	No
Estimated Theoritical Gflops/Watt (Top bin)	N/A	N/A

Table 3: Power processors' main technical characteristics

Note: N/A means the information is not available.

3.4. Other Processor Technologies

China plans to build several Exascale systems using their own manufactured CPUs and GPUs. The first one is NRCPC, a CPU-only machine equipped with ShenWei 26010 (SW26010) processors which is the one used in Sunway TaihuLight (Rank 4 in June 2020 TOP500) [26]. The SW26010 contains 260 cores which produce nearly

3.06 TFlops of 64 bits floating point peak performance per CPU. In that respect, with an expected number of dual-sockets nodes larger than 100,000 in their Exascale system, NRCPC should reach a peak performance over 0.6 EFlops. However, it is most probable that the CPU for the future Tianhe-3 Exascale system will be the next Sunway CPUs which should deliver a peak performance above 10 TFlops. If this scenario comes into reality, NRCPC can reach an aggregate peak performance above 2 EFlops.
The second system in China is the Shuguang Exascale machine relying on two Hygon x86 CPUs and two DCUs [26]. While Hygon's CPU is licensed from AMD's first-generation Zen architecture, DCU is a domestic accelerator produced by Hygon delivering 15 TFlops.

4. GPU, Accelerator and FPGA

While NVIDIA has led the GPU market for the HPC world over the last 10 years, new players like AMD and Intel are entering the game. However, while AMD is still at an early stage to deliver their MI GPUs to the HPC market to support both HPC and AI workloads, Intel is working on a 2021 timeframe to launch the Intel Xe Ponte Vecchio GPU. Overall, it is evident that efforts to include more accelerator performance into HPC nodes at large scale continue to be intensified with specialised units for AI covering not only training but also inference, ray tracing and other use cases to be included in newer generations, enabling their use for new applications and supporting convergence of all workloads. Also, the traditional gap between separate memory spaces and device architectures will decrease thanks to new hardware implementations as well as software solutions (e.g. Intel OneAPI), shielding the user from diverging host and device code.

4.1. GPUs

4.1.1.NVIDIA GPUs

In the past few years, the history of the fastest supercomputers worldwide has shown a steady increase of accelerated systems, the majority being equipped with GPUs. Today, 34% of the 50 fastest systems (TOP500 [3], June 2020) are GPU-powered by NVIDIA. NVIDIA has a strong history of GPU accelerators in the context of HPC, with only 2% among those accelerated systems using non-NVIDIA accelerator hardware in 2020. With Piz Daint and Marconi, Europe is a prominent marketplace: Piz Daint (Swiss National Supercomputing Centre) is equipped with 5,704 Tesla NVIDIA P100 nodes providing a theoretical peak performance of 27 PFlops, mainly dedicated to numerical simulation while Marconi (CINECA) is built on top of 980 V100 nodes, each node with 4 GPUs Volta100. With the French national supercomputer Jean Zay built to answer to the French AI plan for humanity [27], the machine is built with up to 28 PFlops (one scalar partition and one hybrid partition with 2696 GPU NVIDIA V100) dedicated to both HPC and AI with the capability to run HPC/AI combined simultaneously for science. Following the Pascal (P100) and Volta (V100) generation, the new generation of NVIDIA GPUs released is the Ampere GPU A100 announced by Jensen Huang, NVIDIA CEO, on 14 May 2020. The Ampere A100 is built on TSMC's 7nm process and is both delivered in an SXM form factor (400W TDP) and as a PCI-e card (250W TDP). While the FP64 performance of A100 compared to V100 only increases from 7.8 TFlops to 9.7 TFlops (+25% performance improvement per chip) and FP32 similarly by the same ratio from 15.7 TFlops to 19.5 TFlops, the most important added value for numerical simulations is the memory bandwidth improvement (+75% compared to V100) with a higher HBM2 capacity (40GB) and a higher number of NVLINK3 links allowing to double the global performance capability of A100 to 600 GB/s theoretically. NVLink3 has a data rate of 50 Gbit/s per signal pair, nearly doubling the 25.78 Gbits/s rate compared to V100. A single A100 NVLink provides 25 GB/s bandwidth in each direction, using only half the number of signal pairs per link compared to V100. The total number of links is increased to 12 in A100, vs. 6 in V100, yielding 600 GB/s total bandwidth vs. 300 GB/s for V100.
The A100 will support 6912 FP32 cores per GPU (vs 5120 on V100) and 432 tensor cores per GPU (vs 640 on V100).
The other big jump is for the AI workloads that can leverage instructions using the BFLOAT16 format with performance improving by 2.5x. Furthermore, there are new instructions that enable the use of tensor cores using INT8/4 and TF32 (TensorFloat-32), FP64 and FP32 data. While Volta 100 was mainly focussing on training, the A100, with the support of multiple high precision floating-point data formats as well as the lower precision formats commonly used for inference will be a unique answer to training and inference.
Another important aspect of the A100 for sustainability is the capability of supporting Multi-Instance GPU (MIG) allowing the A100 Tensor Core GPU to be securely partitioned into as many as seven separate GPU instances for CUDA applications, providing multiple users with separate GPU resources to accelerate their applications. This new feature will help optimise resource utilisation knowing that not all the applications are taking advantage of a single GPU while providing a defined QoS (Quality of Service) and isolation between different clients, such as VMs, containers, and processes. Due to its implementation, it ensures that one client cannot impact the work or

scheduling of other clients, in addition to providing enhanced security and allowing GPU utilisation guarantees for each workload. Effectively, each instance's processors have separate and isolated paths through the entire memory system. The on-chip crossbar ports, L2 cache banks, memory controllers, and DRAM address busses are all assigned uniquely to an individual instance. This ensures that an individual user's workload can run with predictable throughput and latency, with the same L2 cache allocation and DRAM bandwidth, even if other tasks are thrashing their own caches or saturating their DRAM interfaces.
In addition, as A100 supports PCI-e gen4 with SR-IOV (Single Root Input/Output Virtualisation), allowing to share and virtualise a single PCI-e connection for multiple processes and/or virtual machines to support a better QoS for all over services (I/O, etc.) [28]
In addition, NVIDIA has announced a new software stack including new GPU-acceleration capabilities coming to Apache Spark 3.0. The GPU acceleration functionality is based on the open source RAPIDS suite of software libraries, built on CUDA-X AI. The acceleration technology, named the RAPIDS Accelerator for Apache Spark, was collaboratively developed by NVIDIA and Databricks. It will allow developers to take their Spark code and, without modification, run it on GPUs instead of CPUs. This makes for far faster ML model training times, especially if the hardware is based on the new Ampere-generation GPU due to its characteristics.

4.1.2.AMD GPUs

Although less visible in the HPC market, AMD is taking a position in the landscape with its planned CDNA GPU architecture at an efficient 7nm fabrication process [29]. Optimised for ML and HPC, AMD envisions these architectures to pioneer the road to Exascale by specifically focusing on the CPU-GPU interconnect. This general trend also adopted by other vendors is further detailed in Section 5.2. With the recent acquisition of Mellanox by NVIDIA showing continued interest in interconnects, also higher bandwidth connections such as AMDs Infinity Fabric will manifest themselves in future large-scale HPC systems, offering around 100 GB/s of full-duplex bandwidth for fast data movement among CPUs and GPUs. Coupled with AMDs aggressive roadmap for X3D packaging, this is expected to lead to more tightly integrated intra-node components, partially mitigating the current relative cost of moving data as computational power increases and limiting the responsibility of programmer and software stack to provide efficient software. Furthermore, specialised hardware units such as ray tracing units have also been confirmed, showing AMDs ambition to continue to compete with NVIDIA in that regard. AMD's successful development is evident in part also due to recently awarded supercomputer contracts, namely Frontier at a planned 1.5 EFlops (ORNL) and El Capitan at 2 EFlops (LLNL). Both systems are planned with AMD CPUs and GPUs, and will be one of the first benchmarks of closely coupled CPU-GPU technologies. The awarding of these contracts shows the commitment of part of the HPC community to AMDs technologies for the next couple of years, with new generations of devices to be released approximately once per year [30]
The Radeon Instinct MI50 compute card, available now, is and designed to deliver high levels of performance for deep learning, high performance computing (HPC), cloud computing, and rendering systems. The MI50 is designed with deep learning operations (3.3 TFlops FP32; 26.5 TFlops FP16; 53.0 TOPS INT8) and double precision performance (6.6 TFlops FP64) with access to up to 32GB HBM2 (ECC) memory delivering 1 TB/s theoretical memory bandwidth. In addition, the Infinity Fabric Link (AMD technology) can be used to directly connect GPU to GPU with 184 GB/s peer-to-peer GPU communication speeds, GPU/CPU communication being run on PCI-e gen3 and 4 with up to 64 GB/s between CPU and GPU. While AMD is coming back in the GPU world, one of the key points is maturity of the software stack with its ROCm (Radeon Open Compute) open ecosystem. The current ROCm3.0 (2019) is more focused on ML which includes MIOpen libraries supporting frameworks like TensorFlow PyTorch and Caffe 2. On the HPC side, AMD is working on programming models like OpenMP which is still not supported in ROCm3.0 though it should be in the next generation ROCm software stack currently under development to have Frontier running optimally in 2021/2022. Another important feature was the AMD capability of providing developers tools on ROCm to help translate the CUDA code automatically into codes capable of running on AMD GPUs. For this reason, HIP (Heterogeneous-Computing Interface for Portability) was created. It is a C++ Runtime API that allows developers to create portable applications for AMD and NVIDIA GPUs from a single source code, removing the separation into different host code and kernel code languages.
The next generation MI Radeon Instinct should be built on the 7nm+ process, and based on the names MI100 GPU whereas its successor should be the MI200.

4.1.3.Intel GPUs

As announced at SC19 in Denver, Intel plans to release a dedicated GPU sometime in 2021. The new Intel GPU, called Intel Xe HPC PVC [31] [32] [33], is built on a 7nm manufacturing process. It will be hosted by the Intel Sapphire Rapids CPU which facilitates Xe use through a unified memory architecture between the host and the device through an Xe link which should be based on CXL standards, layered on top of PCI-e Gen5. Intel plans to make the Xe GPUs as adaptable as needed to accommodate as many customers as possible. Hence, there could be

several versions of GPU Xe either to accommodate HPC needs (double-precision performance FP64 & run high-performance libraries) or AI needs (equivalent of tensor accelerators for AI; flexible data-parallel vector matrix engine; BFLOAT16). Intel's Xe link should also be chosen to interconnect the Intel Xe HPC GPUs together, similarly like NVLINK does between NVIDIA GPUs.
It will feature an MCM package design based on the Foveros 3D packaging technology. Each MCM GPU will be connected to high-density HBM DRAM packages through EMIB (Embedded Multi-Die Interconnect) joining all chiplets together. The Xe HPC architecture should also include a very large unified cache known as Rambo cache which should connect several Xe HPC GPUs together on the same interposer using Foveros technology. This Rambo cache should offer a sustainable peak FP64 compute performance throughout double-precision workloads by delivering huge memory bandwidth. Similar to the Xeon CPUs, Intel's Xe HPC GPUs will come with ECC memory/cache correction and Xeon-Class RAS.
Intel has mentioned that its Xe HPC GPUs could feature 1000s of EUs, each capable of performing eight operations per clock and therefore sometimes seen as 8 cores. The EUs are connected with a new scalable memory fabric known as XEMF (Xe Memory Fabric) to several high-bandwidth memory channels. 16 EUs are grouped into a subslice within a Gen 12 GPU (the first generation of Xe GPUs), with the subslice being similar to the NVIDIA SM unit inside the GPC or an AMD CU (Compute Unit) within the Shader Engine. Intel currently features 8 EUs per subslice on its Gen 9.5 and Gen 11 GPUs. Each Gen 9.5 and Gen 11 EU also contain 8 ALUs which are expected to remain the same on Gen 12. A 1000 EU chip will hence consist of 8000 cores. However, this is just the base value and the actual core count should be much larger than that.
In terms of vector length, Intel Xe GPUs would feature variable vector width as mentioned below:

- SIMT (GPU Style)
- SIMD (CPU Style)
- SIMT + SIMD (Max Performance).

This architecture (Intel Sapphire Rapids + Intel Xe HPC Ponte Vecchio GPU) will power the future Aurora Supercomputer which will be launched sometime in 2021 at the Argonne National Laboratory and should be one of the first Exascale machines in the world.
Intel is backing the new hardware development with software support, aiming to provide a stack of hassle-free programming tools and increase their market share by ensuring a large user base. To that end, Intel is focusing their effort on OneAPI [34], an initiative that tries to combine many software projects under one roof in order to facilitate programming CPUs, GPUs, or FPGAs. OneAPI follows the core principle of a single-entry point into the ecosystem, no matter what the underlying hardware base is. OneAPI's [35] main player here is Distributed Parallel C++ (DPC++), which is essentially a mixed language of C++ and SYCL, enhanced with a few Intel flavours, targeting a single-source approach for programming multiple devices. Beyond DPC++, Intel is also working on OneAPI support in OpenMP in both their Fortran and C++ compilers, as any new programming language entails possibly large porting efforts of legacy code bases. Beyond that, Intel also intends to offer debugging and analysis tools with OneAPI, including existing solutions such as vTune, Trace Analyzer, and Intel Advisor, but also third-party tools such as GDB. Finally, Intel intends to offer migration tools that facilitate smooth porting of legacy codes that do require adaption to e.g. new hardware features – a crucial aspect, given that the issue of migration tools is traditionally a difficult one.

4.1.4.Summary of main technical characteristics of GPUs

The Table 4 summarises the main technical characteristics of GPUs.

	Intel	AMD		NVIDIA					
Name	**Ponte vecchio**	**MI50**	**MI100**	**P100**	**P100**	**V100**	**V100**	**A100**	**A100**
Architecture	Intel Xe	Vega20	Arcturus CDNA1.0	Pascal	Pascal	Volta	Volta	Ampere	Ampere
Form Factor	OAM	PCIe	PCIe	PCIe	SXM2	PCIe	SXM2	PCIe	SXM4
Manufacturing Foundry	N/A	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
Manufacturing/Process (nm)	7	7	7	16	16	12	12	7	7
Status	Planned	Launched	Launched	discontinued	discontinued	discontinued	discontinued	Launched	Launched
Availability	N/A	November 2018	November 2020	April 2016	April 2016	March 2018	March 2018	May 2020	May 2020
Accelerator	yes	yes	yes	yes	yes	yes	yes	yes	yes
Standalone	no	no	no	no	no	no	no	no	no
Frontend CPU	yes	yes	yes	any	any	any	any (CC support with IBM POWER)	any	any (CC support with IBM POWER)
Cache coherent link support	N/A	Not supported	Not supported	NVLink 1.0	NVLink 1.0	(For PCI-e GPU to connect via NVLink 2.0 bridge)	NVLink 2.0	(For PCI-e GPU to connect via NVLink 3.0 bridge)	NVLink 3.0
graphic capability	N/A	yes	N/A	yes	yes	yes	yes	yes	yes
AI /HPC application support	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes	yes/yes
Mixed precision	yes	yes	yes	yes	yes	yes	yes	yes	yes
Tensor core support	N/A	no	no	no	no	yes	yes	yes	yes
PCI-e gen	5.0	4.0	4.0	3.0	3.0	3.0	3.0	4.0	4.0
Propriatery inter links per GPU/Accelerator	Xe	XGMI (IF2)	XGMI (IF2)	NVLink 1.0	NVLink 1.0	NVLink 2.0	NVLink 2.0	NVLink 3.0	NVLink 3.0
Inter links Support	yes	yes	yes	yes	yes	yes	yes	yes	yes
Link Speed (Unidir) (GB/s)	N/A	46	46	20	20	25	25	50	50
BW interco (GB/s) bidir	N/A	184	276	160	160	300	300	600	600
Cores	N/A	3840 (60 CUs)	7680 (120 CUs)	3584	3584	5120	5120	6912	6912
Tensor cores	Not Supported	Not Supported	Not Supported	Not supported	Not supported	640	640	432	432
Multi-instances GPUs/Accelerator	N/A	N/A	N/A	Not supported	Not supported	Not supported	Not supported	7	7
HBM or other (GB)	Supports HBM	up to 32 (HBM2)	32 (HBM2)	16 (HBM)	16 (HBM)	16 / 32 (HBM2)	16 / 32 (HBM2)	40 (HBM2e)	40 (HBM2e)
HBM Aggregate Theoritical BW (GB/s)	N/A	1000	1200	732	732	900	900	1555	1555
Software stack	OneAPI	ROCm	ROCm	CUDA	CUDA	CUDA	CUDA	CUDA	CUDA
FP64/32/16 (Tflops)	N/A	6.6 / 13.3 / 26.5	11.5 / 23.1 / 184.6	4.7 / 9.3 / 18.7	5.3 / 10.6	7.0 / 14.0 / 112.0	7.8 / 15.7 / 125.0	9.7 / 19.5 / N/A	9.7 / 19.5 / N/A
FP64/32/16 Tensor Core (Tflops)	N/A	Not supported	Not supported	Not Supported	Not Supported	Not Supported / Not Supported / 112	Not Supported / Not Supported / 125	19.5 / 156 / 312	19.5 / 156 / 312
INT8/4 (Tflops)	N/A	Not Supported	184.6	Not supported	Not supported	130 / 260	130 / 260	1248 / 2496	1248 / 2496
Bfloat 16	N/A	Not Supported	92.3	Not Supported	Not Supported	N/A	N/A	312	312
TDP(W)	N/A	300	300	250	300	250	300	250	400
Peak GFLOP/s/Watt (FP64 DP)	N/A	22.00	38.33	18.8	17.7	28.00	26.00	38.8	24.25

Table 4: main technical characteristics of GPUs

N/A means Not Available.

4.2. Other Types of Accelerators

4.2.1.EPI Titan

The first generation of the EPI accelerator relying on RISC-V is called Titan (gen 1). It might support (but not be limited to) VPU (Vector Processing Unit), STX (Stencil/Tensor Accelerator - BF16) and VRP (Variable Precision accelerator).
EPI plans to design two different accelerators: one is based on RISC-V instruction set architecture and the other one is based on Kalray's intellectual property (IP) core. While the former will be used for HPC and Al, the latter will function in automotive computation.

4.2.2.Graphcore IPU

Graphcore is a young UK-based company which has designed a specific chip called IPU (Intelligence Processing Unit) dedicated to intensive ML algorithms. The IPU is a fine-grained parallel processor designed to deliver high performance for a wide range of computationally intensive algorithms in ML. The IPU design goal was to solve problems beyond the capabilities of current acceleration architectures found in most ASICs and GPUs. The first Graphcore product is the Colossus GC2 built upon a 16nm manufacturing process and is illustrated in Figure 1 [36].

One IPU is built with 1216 interconnected IPU-tiles. Each tile has one IPU core tightly coupled with 300 MB In-Processor-Memory (SRAM) local to each die to enable the model and the data to reside on the IPU to improve memory bandwidth and latency. The 1216 tiles are interconnected through an 8 TB/s on-die fabric (the "IPU-Exchange"), which also connects through "IPU-Links" running at 320 GB/s to create a chip-to-chip fabric. Each IPU core is capable of supporting 8 threads so one IPU can execute 7296 executions in parallel.

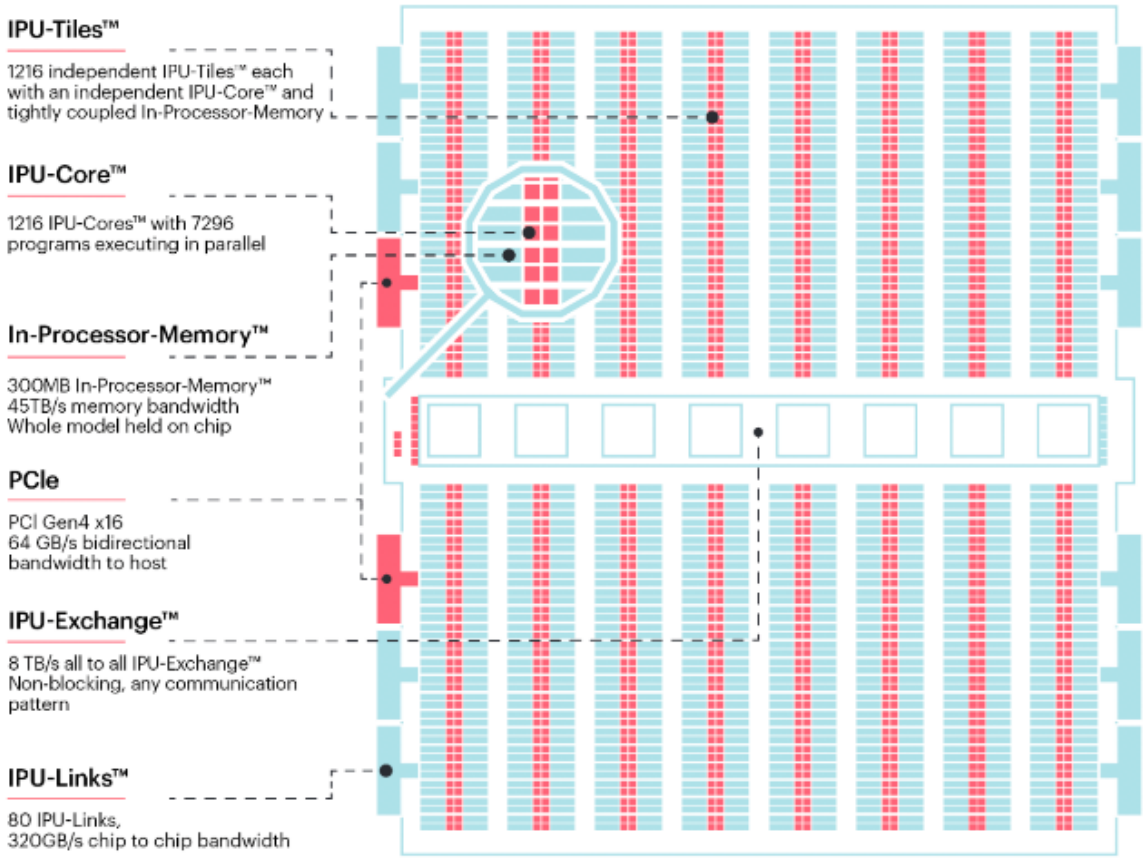

Figure 1 : Graphcore GC2 Intelligent Processor micro-architecture (https://moorinsightsstrategy.com/wp-content/uploads/2020/05/Graphcore-Software-Stack-Built-To-Scale-By-Moor-Insights-And-Strategy-2.pdf)

The GC2 is delivered in a Gen4 16x PCI-e card form factor, C2 PCI-e card. It embeds two GC2 IPU processors providing 600 MB aggregated In-Processor-Memory, 200 TFlops peak mixed precision FP16.32 IPU compute with a 315W TDP and an IPU link running at 2.5Tbps. The entire system (composed of many IPUS) executes in 2 synchronous phases: computation and communication. Applications that target the IPU are expressed as computational graphs. Computations are performed at the vertices of the graph, and the results are communicated to adjacent vertices according to the edges interconnecting the graph. The communication phase is implemented as a Bulk Synchronous Parallel (BSP) operation, which efficiently transfers data from each tile's on-die SRAM memory to connected tiles' memory. In addition to computation instructions, each IPU core features a dedicated tile-level instruction set for communication phases of the BSP model. The integrated exchange-communication fabric is designed to support BSP for both data and model parallelism — enabled by the graph compiler — potentially scaling to thousands of nodes. An important distinction for the IPU architecture, according to Graphcore, is the ability to efficiently process sparse data and graphs, which improves performance while reducing the total memory requirements. The Graphcore software stack is Poplar. Poplar supports users addressing the main challenges of ML, such as deep neural networks, providing the capability to optimise and efficiently run ML algorithms as research and development of entirely new fine-grained parallel workloads to run on an IPU infrastructure. The current ML frameworks supported by the Graphcore software platform are the most popular ones like TensorFlow, Pytorch, Mxnet, etc. Graphcore has also taken the next step in management software, providing containerisation, orchestration, security, and virtualisation. These strategic choices should ease the adoption as more applications are deployed on the Graphcore platform.

Graphcore has announced their new GC200 processor (Figure 2) in July 2020 built upon a TSMC 7nm FinFET manufacturing process. The GC200 processor features now 1472 independent IPU-tiles (+20% compared to the 1st GO2 generation), each IPU-Tile is built on top of an IPU-core coupled with 900 MB In-Processor-Memory (x3 times compared to the previous generation). The new processor is now capable to execute 8832 programs in parallel. The GC200 is delivered in a Gen4 16x PCI-e card form factor. The PCI-e card is as C200 PCI-e card and embeds two GC200 IPU processors providing 1.8 GB aggregated In-Processor-Memory, 600 TFlops (FP16) peak performance with a 425W TDP with an IPU link running at 2.5Tbps. It can also be provided as an IPU server with an x86 or Arm host CPU and 8 C200 PCI-e cards (16 GC200) in a 2D ring topology with a 256 GB/s card (C200) to card (C200), providing up to 4 PFlops (FP16) peak performance. The last form factor is an IPU-POD providing up to 32 PFlops (FP16) peak performance. The design relies then on the integration of several IPU-Machine (16 to 32), each one being based on 4x GC200 IPU + 1x IPU GATEWAY providing access to the other IPU-Machine through 2x 100Gbps links for inter-communication. The IPU-Machines are all connected through a 2D-Torus topology.

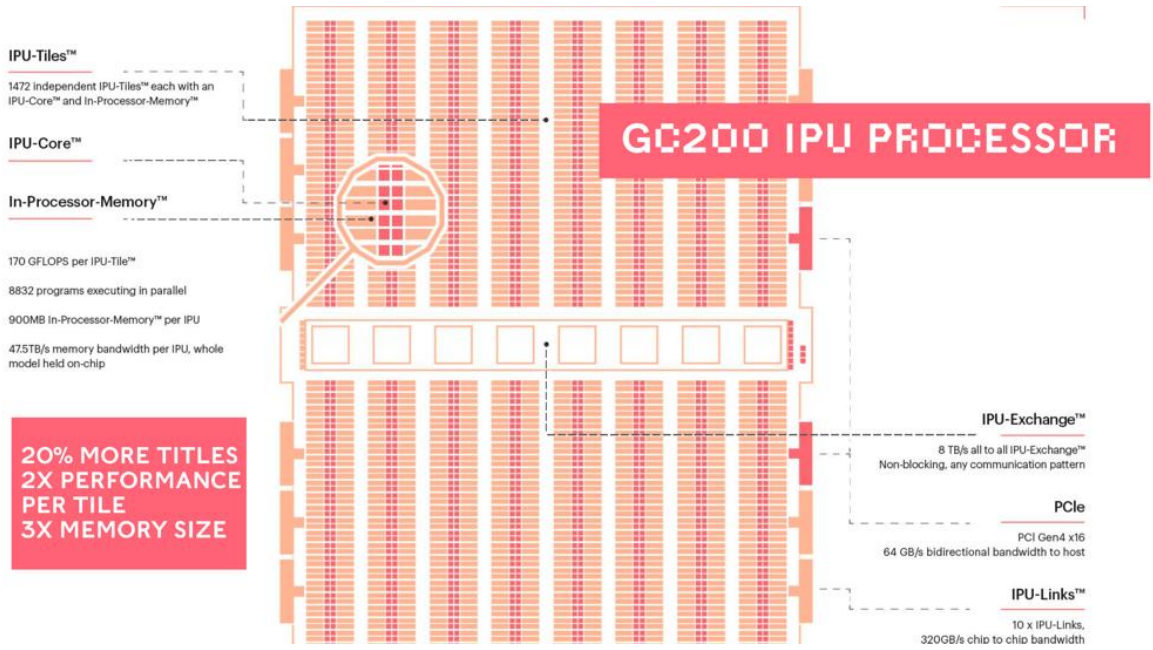

Figure 2: Graphcore GC200 processor microarchitecture

The Table 5 summarises the Graphcore accelerators' main technical characteristics.

	Graphecore	
Name	GC2	GC200
Architecture	GC2	GC200
Form Factor	PCI-e	PCI-e
Manufacturing Foundry	TSMC	TSMC
Manufacturing/Process (nm)	16	7
Status	Launched	Launched
Availibility	2018	July 2020
Accelerator	yes	yes
Standalone	no	no
Frontend CPU	yes	yes
Cache coherent link support	Not supported	Not supported
graphic capability	no	no
AI /HPC application support	yes/no	yes/no
Mixed precision	yes	yes
Tensor core support	no	no
PCI-e gen	4.0	4.0
Propriatery inter links per GPU/Accelerator	IPU link	IPU link
Inter links Support	yes	yes
Link Speed (Unidir) (GB/s)	2	16
BW interco (GB/s) bidir	320	320
Cores	1216 IPU cores	1472 IPU cores
Tensor cores	Not supported	Not supported
Multi-instances GPUs/Accelerator	Not supported	Not supported
HBM or other (GB)	No HBM - 300 MB (in-processor memory)	No HBM - 900 MB (in-processor memory)
HBM Aggregate Theoritical BW (GB/s)	45000	47500
supported instruction sets	Poplar	Poplar
FP64/32/16 (Tflops)	Not Supported / 120 FP16.32	Not Supported / 250 FP16.16
FP64/32/16 Tensor Core (Tflops)	Not supported	Not supported
INT8/4 Tensor core (Tflops)	Not supported	Not supported
Bfloat 16 Tensor core	Not supported	Not supported
TDP(W)	>=150	>=200
Peak GFLOP/s/Watt (FP64 DP)	Not supported	Not supported

Table 5: Graphcore Accelerators' technical characteristics

4.2.3.NEC SX Aurora

NEC has invested since the 80s in vector supercomputers and has recently innovated a new hybrid architecture called SX-Aurora TSUBASA. This hybrid architecture consists of a computation part and an OS function part. The heart of the new SX architecture is the vector engine (VE) contained in the vector host (VH) with the VE executing complete applications while the VH mainly provides OS functions for connected VEs.
The SX Aurora vector processor is based upon a 16 nm FinFET process technology and is available as a standard PCI-e card (Figure 3) which can be hosted on any x86 server host environment. Each vector CPU has access to six HBM2 memory modules, leading to a theoretical memory bandwidth of 1.53TB/s.

VE20, the second generation of SX vector processor, has two VE types, Vector Engine Type 20A and Type 20B. VE Type 20A is 3.07TF peak performance with 10 vector cores, and VE Type 20B is 2.45TF peak performance with eight vector cores. Each vector core and 16MB of shared cache are connected by a two-dimensional mesh network providing a bandwidth per vector core of 400GB/s maximum. The vector core on the VE achieves 307 GFlops peak performance per core and an average memory bandwidth of 150 GB/s per core for the 10 cores configuration per the VE processor. Each vector core mainly consists of 32 vector pipelines, and three FMA units are implemented into each vector pipeline. The vector lengths of 256 is processed by 32 vector pipelines with eight clock cycles. The 64 fully functional vector registers per core – with 256 entries of 8 bytes width each – can feed

the functional units with data, or receive results from them, thus being able to handle double-precision data at full speed.

Depending on how much the application is vectorised, the VE card can be used in 2 modes. A native mode to run full application on VE card (for vector codes) or an offload mode to run part of the application on VE Card, the remaining on VH X86 scalar system (for scalar & vector codes) with a VE engine supporting standard programming languages and parallelisation paradigms (Fortran, C, C++, MPI, OpenMP, etc.).

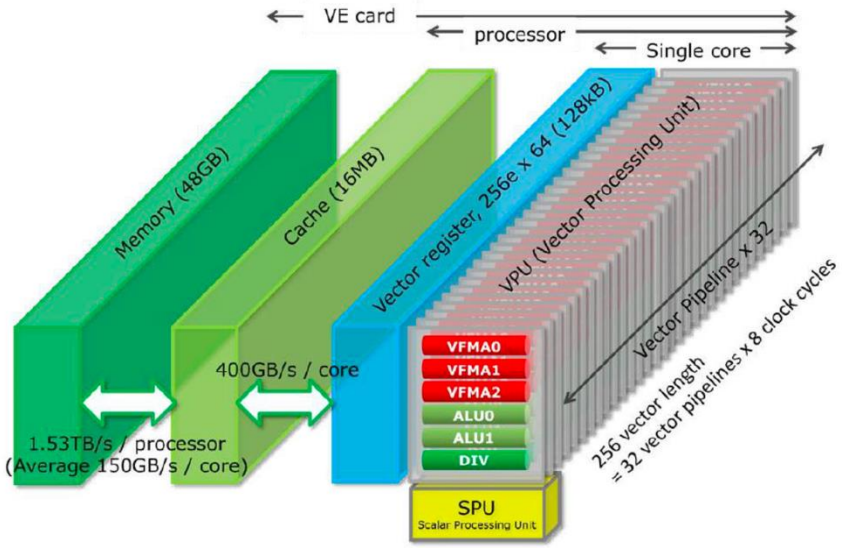

Figure 3 : VE card architecture and workflow

The next generation, named VE30, is planned to be released as the successor of the VE20 generation sometime in 2022. The main improvement from the predecessor is a memory bandwidth of 2+TB/s and a memory subsystem achieving higher sustained performance with lower power consumption.

4.3. FPGAs

Usage of FPGAs in HPC was very limited in the past, mainly due to the adoption barrier of porting applications (e.g. using VHDL – Very High-Speed Integrated Circuits Hardware Description Language) and a lack of support in domain-specific libraries (e.g. FFT, BLAS, LAPACK) or predominant parallelisation models (e.g. OpenMP). In addition, the relatively high performance of GPUs for large classes of floating-point-heavy applications provided a more feasible solution for increasing per-node computational power. For this reason, FPGAs were mainly used in business-centric applications such as high frequency trading and only selectively offered in data centres such as Amazon's Web Service (AWS) F1 instances (which employ Xilinx Virtex UltraScale+ VU9P devices). However, the recent rise in integer and fixed-point math fields such as deep learning, bioinformatics, cryptanalysis or data mining has opened a new market for FPGAs. The predominant company is Intel, having acquired both Omnitek and Altera, with their line of Agilex, Arria, and Stratix devices. With the recent acquisition of Omnitek, Intel plans to continue to offer devices tailored especially towards AI inference and visual applications [37]. Furthermore, Intel's recent advances in the programming software stack (e.g. OneAPI) will likely ensure programmability and sustainability of these devices. The Stratix 10 DX is specifically designed as a cache-coherent accelerator for servers and offers a logic capacity of up to 2.7 million elements along with four Arm Cortex-A53 cores on a 14nm monolithic die. Memory options include HBM2 up to 8 GB, Intel's persistent memory technology Optane up to 4 TB, or a combination of both. The interconnect to host systems will be established through Ultra Path Interconnect (UPI), the successor of QPI, which provides a peak bandwidth of 28 GB/s and enables adoption of future technologies such as CXL and PCI-e 5.0. Coupled with a 100 Gbps Ethernet network interface, these devices can be tailored to many application-specific use cases by end-users able to provide close to 10 TFlops in single precision IEEE 754 math and multiple 100 GBit Ethernet interfaces [38]. The next generation FPGAs is the Agilex product line, with a number of samples already produced. It is manufactured at 10nm, increasing performance or decreasing power consumption by up to 40%, respectively. Specifically, the I and M series of these devices are intended as cache-coherent accelerators for Xeon processors and focus on high-performance interfaces

and bandwidth-intensive use cases (I series) and compute-intensive applications (M series). These FPGAs are designed to deliver up to 40 TFlops of signal processing performance and include support for INT2 through INT8, FP16, and BFLOAT16 data types, improving their applicability to AI workloads. All these features are exposed through Intel's OneAPI initiative, facilitating fast adoption by end-users. Fabrication sizes are currently topping out at 10nm for their newest Agilex product [39], but are expected to shrink to keep up with the increasing energy efficiency of competing technologies.

4.4. OCP Acceleration Module

A new trend to consider in order to accelerate the integration of any kind of accelerator technologies emerging on the market within existing systems is the OAM (OCP Acceleration Module) specification - where OCP stands for Open Compute Project - which defines an open-hardware compute accelerator module form factor and its interconnect. New technologies frequently come with different sizes, different thermal characteristics, a variety of board wiring schemes, and, in some cases, unique sockets. This leads to many form factors that impact the whole system, that in turn need to be accommodated or redesigned for only a few add-ins, thus delaying time-to-market. Since many of these accelerators have similar design requirements and specifications - inter-module communication (to scale up) and high input/output bandwidth (to scale out) - enabling a common form factor specification that can be used in different types of hardware accelerators would definitely help the accelerators' integration. In that respect, the Open Compute Project has been created to reimagine hardware, making it more efficient, flexible, and scalable to support a common design on OAM available to those who want to use it. The OAM form factor could be a standard chosen by integrators and chip makers to ease integration of their new hardware in existing platforms reducing the amount of time needed to push for a new technology onto customer production.

5. Interconnects

Interconnects can be roughly divided into two categories: Inter-Node Connectivity (interconnects used between compute nodes) and Intra-Node Connectivity (interconnects used within a compute node).

5.1. Inter-Node Connectivity

The inter-node networks are one of the key building blocks for HPC systems, a fundamental requirement to connect any number of nodes to form a single, large system. Up to now, for medium and large-scale systems, the inter-node connectivity has usually been split into two physically independent network implementations: on the one hand a low-latency, high throughput network (e.g. Mellanox InfiniBand, Cray Aries, Intel Omni-Path Architecture (OPA), etc.) that is used for user traffic, i.e. MPI communication between nodes and I/O and, on the other hand, a management network (usually Ethernet) to support administrative traffic. This distinction is also made for security reasons. For some small-scale systems, Ethernet is also used for MPI communication. Over the years, the networks used in HPC systems have gradually increased in performance, both in terms of throughput and latency. Most of the network types currently used are switched networks deployed in different topologies in order to optimise the global bandwidth and keep the latencies low.

5.1.1.Ethernet

For a long time, the standard single lane multi-gigabit Ethernet was based on 10 Gbps, with 4 such links being used to create a 40 Gbps link. The corresponding standards were introduced well over 10 years ago and while they first saw slow adoption due to cost, they have nowadays become widespread. Some manufacturers, for instance Intel, have recently integrated 10 Gbps connectivity into chipsets and SoCs with 10 Gbps becoming available even in consumer devices.

5.1.1.1. 100/25 Gigabit Ethernet

The 25G Ethernet Consortium was formed in 2014 to develop a single lane 25 Gbps Ethernet standard, approved as IEEE 802.3by in 2016. The 25 Gbps standard is based on one lane from the 100 Gbps 802.3bj approved in 2014, which uses 4 lanes running at 25 Gbps, similar to 10/40 Gbps Ethernet. This provides the ability to use 100 Gbps switches and fan-out cables to get a large number of 25 Gbps ports from a single switch.

5.1.1.2. 200 and 400 Gigabit Ethernet

The next step for Ethernet will be 200 and 400 Gbps ports, ratified as a standard at the end of 2017. 200 Gbps Ethernet uses four 50 Gbps lanes per port while initial 400 Gbps standards will use eight 50 Gbps lanes and simply double the number of lanes in the port. Products are starting to become available now: Mellanox for instance has made both 200 Gbps network adapters and 400 Gbps switches available, with the 400 Gbps switches offering ports that can be split into two 200 Gbps ports.

5.1.1.3. RoCE

RoCE is a protocol for enabling RDMA accesses/transfers over regular Ethernet. With direct memory access (DMA) the network adapter can read and write from the host memory directly bypassing the CPU cores, thus lowering the CPU load. RoCE packets also bypass part of the regular ethernet stack and combined with DMA RoCE can have a significantly lower latency than traditional Ethernet. The basic idea for RoCE is to encapsulate an InfiniBand transport packet into a regular Ethernet packet on the link layer. RoCE comes as version 1 and version 2. V1 is a link layer protocol, allowing communication between two hosts in the same broadcast domain. V2 extends the RoCE packet with necessary IP headers to make it effectively a normal UDP (User Datagram Protocol) packet, making it routable.

RoCE capable hardware is available from multiple vendors and bandwidth ratings, from 10 Gbps up to 200 Gbps. While it is unlikely that RoCE will completely replace high performance interconnects such as InfiniBand, it does have some important use cases. For instance, for smaller or more cost optimised systems that need lower latency networks but not the extreme bandwidth, RoCE over 25 Gbps would be a good compromise offering lower latencies than regular Ethernet and reducing the network load on the CPUs of the system.

5.1.2.InfiniBand

InfiniBand is the most widely used HPC interconnect. It is based on a standard that is maintained by the InfiniBand trade association. While there have been multiple vendors manufacturing InfiniBand products, currently the only adapters and switches available are from NVIDIA Networking (formerly Mellanox). InfiniBand focuses on low latency and high bandwidth connections, providing RDMA capability to lower CPU overhead.

5.1.2.1. High Data Rate (HDR/HDR100)

The last generally available InfiniBand products are based on the HDR (High Data Rate) standard. With HDR, a regular 4-lane port theoretically reaches 200 Gbps with a physical port that can be split into two 100 Gbps ports, referred to as HDR100. HDR100 allows HDR to be efficiently used in servers providing only 16x PCI-e gen 3 ports while the 200 Gbps port requires 16x PCI-e gen4 to provide its full capability. Since HDR100 splits a single HDR port into two, it effectively doubles the port count per switch, building fat trees with hosts based on HDR100. One can thus use 200 Gbps HDR to connect and build the network reducing the cabling and switches needed for the network.

The updated version 2 of SHARP (Scalable Hierarchical Aggregation and Reduction Protocol) was introduced with HDR. Before that, SHARP version 1 was introduced in the Mellanox software stack with the previous version of the InfiniBand standard, EDR. SHARP allows some collective operations to be offloaded from the compute nodes to the network by the switches and adapters. SHARP v2 enables offloading to work with large sets of data. HDR also supports hardware-based congestion control and adaptive routing.

5.1.2.2. Next Data Rate (NDR)

The next InfiniBand standard will be NDR, offering a per port bandwidth of 400 Gbps. Availability of NDR products have not been announced yet. As a single 400 Gbps port would require 16x PCI-e gen5 in order not to be limited by the PCI-e bus. It is likely that NDR will be able to split the port into at least two 200 Gbps ports to allow it to be used optimally in systems before PCI-e gen5 processors become available.

5.1.3.Omnipath

In 2015, Intel introduced Omnipath, an HPC interconnect running at 100 Gbps per port. It was available both as standalone PCI-e adapters but also integrated into certain Skylake and Knights Landing CPUs, as an extra chip on the CPU substrate. However, the integrated version was not available on Cascade Lake CPUs and despite a strong

roadmap for 200 Gbit Omnipath 2, Intel ceased development of Omnipath in 2019. End of September 2020, Intel has announced that the company spun off Omni-Path Architecture Business to Cornelis Network, an independent company from Intel [40].

5.1.4.Bull eXascale Interconnect (BXI)

The Bull eXascale Interconnect (BXI) is an HPC network designed by Bull, Atos' HPC division, integrated in the Atos BullSequana XH platforms and also available in standard form factors for rack mountable servers. One should notice that, while EPI is the only EU-processor, BXI is the only EU-interconnect network. Like several other interconnects, BXI supports multiple lanes per port, with currently available BXI adapters using four 25 Gbps lanes to provide access to 100 Gbps unidirectional bandwidth. Each BXI switch has forty-eight 100 Gbps ports. The BXI implementation relies on the Portals 4 architecture and has hardware features that map directly to many MPI and PGAS functions. While BXI initially only supported a fat-tree topology, it is now also capable of supporting DragonFly+ topology.

Despite 100 Gbps BXI having been available for quite some time, its deployment was mainly limited to Atos strategic partners. Atos has been selling XH2000 machines with InfiniBand network and not aggressively pushing their own alternative up to now. The largest BXI deployment has been the Tera-1000 machine (CEA - 8000 KNL nodes), a 23 PFlops peak performance supercomputer; in addition to this there are some smaller machines that also use BXI as low latency interconnect network to glue their supercomputing resources.

5.1.5.HPC Ethernet/Slingshot

Now part of HPE, Cray has been developing their own interconnects for a long time. Even after selling the interconnect division to Intel in 2012, Cray did not stop developing their own interconnects. The latest one introduced is branded Slingshot and it is based on Ethernet but uses features that make it more appropriate for HPC workloads, targeting lower latency and better scalability, while at the same time maintaining compatibility with commodity Ethernet. Slingshot is the network that will be used in the 3 US Exascale supercomputers announced recently.
Slingshot switches have 64 ports and each port is running at 200 Gbps. Switches either come integrated into Shasta supercomputers or as regular 1U rack switches. Slingshot can use regular 100 Gbps Ethernet adapters to connect to the switches or dedicated Slingshot network adapters based on HPEs own NIC (Network Interface Card). The dedicated NICs will operate at full 200 Gbps speed with ports capable of some additional capabilities.
While the Shasta machines are built with a Dragonfly network topology, Slingshot does support other topologies such as the more traditional fat-tree as well. For large Dragonfly topologies, the first level switch uses 16 ports for connecting to compute nodes, 31 to do all-to-all connections between all the switches in the same group and then 17 to do the global network.

One of the more advertised features of the Slingshot network is its congestion control mechanisms which is supposed to provide significant improvement compared to the previous Aries network. These congestion control mechanisms should minimise the impact one job can have on other jobs running in the machine at the same time, especially trying to guarantee low latency for latency sensitive applications. This is done by identifying the sources of the congestion and throttling them at the source ports.

5.1.6.Summary of Inter-node interconnect main technical characteristics

The Table 6 summarises the main Inter-node interconnect main technical characteristics.

	Infiniband		Low Latency Ethernet	BXI	Omnipath	Commodity Ethernet		
Name	**HDR**	**NDR**	**Slingshot**			**RoCE**	**25-100Gbps**	**200-400Gbps**
Manufacturer	Mellanox	Mellanox	HPE	Atos	Intel	Multiple	Multiple	Multiple
Avaibility	Available	Future	Available	Available	Discontinued*	Available	Available	Available/Future
Open/Proprietary	Proprietary	Proprietary	Proprietary	Proprietary	Proprietary	N/A	Open	Open
Unidirectional Bandwidth (Gbps)	100/200 gbit/s per port	200/400 gbit/s per port	100/200 gbit/s per port	100 gbit/s per port	100 gbit/s per port	N/A	25-100 gbit/s per port	200/400 gbit/s per port
End to End Latency (micro-second)	<1 usec	N/A	<2 usec	<1 usec	<1 usec	~1 usec	N/A	N/A
PCI-e card (gen)	Gen 3/Gen 4	N/A	Gen 3/Gen 4	Gen 3	Gen 3	N/A	Gen 3/Gen 4	Gen 3/Gen 4
Switch	40 ports @ 200gbit/s	N/A	64 ports @ 200gbit/s	48 ports @ 100 gbit/s	48 ports @ 100 gbit/s	N/A	Various	Various
Topology supported	Fat-Tree, Dragonfly+	N/A	Dragonfly	Fat-Tree, torus, flattened butterfly ...	Fat-Tree	N/A	Various	Various
Lanes Throughput	4 lanes @ 50 gbit	4 lanes @ 100 gbit	4 lanes @ 50 gbit	4 lanes @ 25 gbit	4 lanes @ 25 gbit	N/A	1-4 lanes @ 25 gbit	4-8 lanes @ 50 gbit
RDMA support	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Hardware Features embedded (collectives,etc.)	Offloaded collectives, tag matching	N/A	Some			N/A	N/A	N/A

Table 6: Inter-node Interconnect main technical characteristics

N/A means Not Available.
*Omnipath is discontinued by Intel. Business activities are ensured by Cornelis Network.

5.2. Intra-Node Connectivity

The intra-node connectivity is one of the major challenges to improve efficiency of heterogenous architectures (GPP combined to accelerators) since they became more widely adopted. A few years ago, the only open-standard was PCI-e, which does not feature cache coherency. This has led to the development of many new intra-node interconnects.

One could further divide the intra node interconnects between those used to connect different devices within the node and those used to connect different parts of the processor in silicon or on substrate. Modern servers include a large number of different devices that need to communicate with each other in order for the system to work. A typical HPC node may include multiple CPUs, GPUs, network adapters and SSD storage that all need to be connected to each other to make it work. A typical server will include multiple high-speed internal interconnects. For instance, a node with multiple Intel CPUs will use UPI to connect the CPUs together, PCI-e to connect storage, network and GPUs, while providing an additional high bandwidth link to connect all accelerators to each other such as NVIDIA NVLink.

5.2.1.PCI-e gen 3, 4, 5 and 6

PCI express (PCI-e) is the primary internal bus used today to connect devices to CPUs within a node. It is used to connect GPUs, network adapters, NVMe storage and other additional devices. Usually PCI-e starts out from the CPU and devices are directly connected to the CPU through PCI-e; however, there can also be switches introduced along the path allowing multiple devices to share a single connection to the CPU.

PCI-e is usually deployed in multi-lane configurations, 4 lanes (4x) being commonly used for NVMe SSDs and I/Os and 16x being used for more bandwidth intensive applications such as GPUs, and now - with the advent of 100 Gbps or multi 100 Gbps networks - also used for network adapters. CPUs come with a specific number of lanes: Intel server CPUs are using 56 lanes whereas the AMD CPUs come with 128 lanes. Using PCI-e switch chips servers provide more lanes than those that are available from the CPU. However, all devices connected to the same switch will share the same uplink connection to the CPU. While devices connected to the same switch can communicate at whatever speed they are connected, if multiple devices communicate with the CPU, they will share the link going to the CPU and thus may not get the full bandwidth to the CPU that the switch could provide. The PCI-e standard has been rapidly evolving since it has finally moved on from staying at Gen 3 for a long period. Currently the industry is quickly moving to PCI-e Gen 4 with a 16 GT/s (Giga Transfers/s) transfer rates, for a theoretical throughput of 31.5 GB/s for a 16x port with the same encoding as PCI-e Gen 3 (128/130bits). Both

AMD and IBM are currently offering server integrated CPUs with PCI-e Gen 4, and Intel is expected to introduce Gen 4 with the Ice Lake CPUs expected later in 2020.
The next evolutionary step is PCI-e Gen 5, the standard was finalised in 2019. The transfer rate is doubled from Gen4 to 32GT/s, yielding roughly 63 GB/s of bandwidth from a 16x port. CPUs and GPUs supporting gen 5 PCI-e are expected in between 2021 and 2022.
The final version of the PCI-e Gen 6 specification is expected to be released in 2021. It is again expected a doubling of the per lane transfer rate to 64GT/s, yielding roughly 124 GB/s of bandwidth for a 16x port. The new standard will switch to PAM4 to allow 2 bits to be transferred per transfer, and the new standard will also use forward error correction.

5.2.2.CXL

The Compute Express Link (CXL) is an open standard for a high-speed interconnect intended to be used as an interconnect between CPUs and devices and CPUs and memory. The initial consortium was made up of Alibaba, Cisco, Dell EMC, Facebook, Google, HPE, Huawei, Intel, and Microsoft; however, since then the consortium has grown to now include all major HPC hardware vendors, such as AMD, NVIDIA, Mellanox and Arm.
CXL is based on top of PCI-e, specifically CXL v1.0 is based on PCI-e Gen 5 using the same physical and electrical interface. CXL provides protocols in three areas: CXL.cache, CXL.memory and CXL.io. CXL.io is the simplest one, it is very PCI-e like and supports the same features as PCI-e. The more interesting protocols are CXL.memory and CXL.cache, which provide cache and memory semantics. CXL.cache is used for devices that want to cache data from the CPU memory locally, allowing for instance network drives to have their own cache. CXL.memory is used to provide processor access to the memory of attached devices.
The first practical HPC implementation of CXL is expected to be the CPU and GPU combination used in the US Aurora supercomputer that is to be installed in 2021, where it will be used to connect the accelerators to the host CPU system and create a coherent memory space between accelerators and the CPUs. So far, the only product supporting CXL has been some FPGA models from Intel.
With CXL being an open standard and with the participation of all of the large vendors there is hope that CXL could see widespread adoption. However, Intel has been the only major manufacturer that has been making commitments to create products using CXL so far.

5.2.3.Infinity fabric

Infinity Fabric (IF) is AMDs proprietary system interconnect. It is used on multiple levels within AMD systems, to connect different parts of the processor and even multiple GPUs together. The following will focus mostly on the variants not used within the processor.
AMD CPUs come with 128 PCI-e lanes, in dual socket configurations 48 to 64 of them are used to connect the CPUs to each other, and these lanes then switch over to running as IF lanes, providing a coherent shared memory space between the CPUs.
The MI50 and MI60 cards feature an IF connector allowing multiple GPUs to be connected to each other. Up to 4 GPU can be connected in a ring topology, providing up to 184 GB/s of peer to peer GPU bandwidth.

Considering that IF is already used to connect multiple CPUs and between multiple GPUs, the next logical step for AMD would be to extend it to work between CPUs and GPUs. Such a configuration is to be used in the Frontier supercomputer to provide a coherent memory space between accelerators and the CPU.

5.2.4.CAPI/OpenCAPI

The Coherent Accelerator Processor Interface (CAPI) is IBM's proprietary coherent interconnect for directly connecting accelerators to CPUs and forming a coherent shared memory space. CAPI was introduced with the POWER8 processors in 2014. It works over PCI-e, with version 1 using PCI-e gen 3 and version 2, introduced with the POWER9 processors, using PCI-e gen4. One practical implementation of CAPI has been Mellanox network cards in POWER9 systems, such as the ones used on the Summit and Sierra supercomputers.

OpenCAPI is the evolution of the CAPI protocol but instead of being a proprietary standard it is an open standard published by the OpenCAPI consortium. Unlike CAPI, OpenCAPI does not run on top of PCI-e but uses its own protocol. OpenCAPI continues the version numbering from CAPI, meaning the initial version is OpenCAPI 3.0. OpenCAPI 3.0 uses 25 GT/s transfer speeds, with the common deployment being 8 lanes, yielding a bandwidth of 25 GB/s. While OpenCAPI can be seen to use PCI-e slots, these are only used for power feed and as fixtures for the modules. The actual OpenCAPI connector is a slimline SAS cable carrying 8 lanes. Thus far the only support for OpenCAPI seen from a CPU is from IBMs POWER9 series of processors.

One interesting OpenCAPI product that has been shown is a serially attached memory through OpenCAPI. Here the additional system memory can be attached to the CAPI ports instead of the regular DIMM slots. This would alleviate the ever-increasing footprint and pin count needed for the increasing number of memory channels used in servers. This will be an open alternative to IBM's own Centaur memory controllers used until POWER9. The updated version of POWER9 has already supported this new Open Memory Interface, which will also be used for POWER10 [41].

5.2.5.CCIX

Cache coherent interconnect for accelerators (CCIX) is an open standard interconnect, designed to provide a cache coherent interconnect between the CPU and accelerators. CCIX operates on top of PCI-e. The 1.0 specification utilises the standard 16 GT/s transfer rate of PCI-e gen 4, but can also run in an extended speed mode of 25 GT/s. Version 1.1 of the specifications supports PCI-e gen 5 and the 32 GT/s transfer speeds that introduces.
CCIX allows devices to be connected in different flexible topologies: each CCIX device has at least one port that can be used either as a direct connection to another CCIX device or to a CCIX switch. Devices with multiple ports can be used to build more complex topologies, such as daisy changing multiple devices or creating mesh or all-to-all topologies.
With only some HPC vendors participating in the consortium (including Arm, AMD and Mellanox), and some (like Intel and NVIDIA) being absent, and considering also that all major vendors of the CCIX consortium are part of CXL, the future of CCIX is uncertain.

5.2.6.UPI/QPI

UPI (UltraPath Interconnect) is the evolution of QPI (QuickPath Interconnect), Intel's proprietary interconnect used to connect multiple processors in a node together. The previous QPI link was running at 9.6GT/s whereas the new UPI link, introduced with the Skylake architecture is running at 10.4 GT/s. Skylake and newer CPUs have 2 to 3 UPI links per processor. Intel has also used UPI to connect CPUs to FPGAs, creating a cache coherent domain between the CPU and FPGA.
Dual-socket systems, depending on motherboards and CPU types, can have either 2 or 3 links between the CPUs. Quad-socket systems can be built using some CPUs with just 2 UPI links, but in that case each CPU would only be directly connected to 2 other CPUs, and communication with the third one would have to traverse one of the other two. Most Xeon Scalable CPUs that support 4 sockets have the full 3 UPI links so a quad-socket system with just 2 links per CPUs is very uncommon. Since the current maximum number of lanes per CPUs is 3, systems with 8 sockets will not have direct connection between all CPUs in the system.
With Intel embracing CXL, the future of the UPI interconnect is uncertain. Intel will use CXL to connect the GPUs in the upcoming Aurora system to the CPUs, so the question is whether UPI will still be used between the CPUs or it will also be transitioned to CXL. With CXL being based on PCI-e Gen 5, it would potentially offer a higher bandwidth, but possibly a higher latency than solutions based on UPI.

5.2.7.NVLink

NVLink is NVIDIA's proprietary interconnect primarily used for connecting multiple GPUs together. The exceptions are POWER8+ and POWER9 processors that can use NVLink to connect the GPUs directly to the CPUs. NVLink is designed to offer a significantly faster link than what can be achieved with regular PCI-e in order to improve GPU to GPU communication in multi-GPU systems. NVLink allows multiple GPUs into one unified memory space, i.e. allowing a GPU to work on memory local to another GPU with support for atomic operations. With the second generation NVLink introduced with the Volta generation of GPUs, the NVLink connection between GPUs and CPUs was improved to support atomic operations, coherence operations, and allowing data reads from the GPUs memory to be cached in the CPUs cache hierarchy.

NVLink was introduced with the Pascal generation of GPUs, where each GPU had 4 NVLink ports, with 40 GB/s of bandwidth per port. This was expanded to 6 ports running at 50 GB/s for the Volta generation of GPUs. The latest Ampere generation keeps the same 50 GB/s bandwidth per port but increases the number of ports from 6 to 12, yielding a total of 600 GB/s of bandwidth for the GPU.

NVIDIA has also developed a separate NVLink switch chip, with multiple ports that can be used to connect GPUs with each other. The first generation was introduced for the Volta architecture, where the switch used 16 ports, and the NVSwitch was updated for the Ampere architecture to provide double the bandwidth.
GPU-to-GPU connectivity using NVLink can be implemented either as direct connection or using NVSwitch chips. Direct connection with NVLink is primarily used to connect 4 GPUs to each other, and with the Ampere generation this means 4 links per GPU in an all-to-all configuration yielding 200 GB/s of bandwidth between each

pair of GPUs. Initially the direct connection was also used to build 8 GPU systems, now completely replaced by 8 GPU systems featuring NVSwitches. NVSwitches can be used to build a 16 GPU system where each GPU can use its full bandwidth to communicate with any other GPU in the system.
The first system to employ NVSwitches was the Volta DGX-2 system. In these systems, 12 NVSwitches are used and the GPUs are in two planes, each using 6 switches with half the links going to the GPUs in that plane and the other half going to the switches of the other plane. For the Ampere generation, the 8 GPU DGX-1 systems also moved to feature NVSwitches. In this case, 6 switches are used, and the system baseboard is similar to one of the planes of a DGX-2 system.

5.2.8.Gen-Z

Gen-Z is an open standard, built on memory semantic operation over a fabric. The goal is to be able to efficiently move data between memories located in different devices with low latency. The fabric supports basic operations such as loads and stores with the idea that all devices in a system will natively support the same operations, replacing both existing buses between GPUs and different devices, and possibly also between the CPU and its own memory. It lacks hardware coherency features, instead relying on software and atomic operations to avoid race conditions.

While not strictly an intra node interconnect, it provides similar functionality as the other intra node interconnects covered here. It has the capability of providing this feature also between hosts. The goal is to provide composable systems, for instance having a pool of memory that can be allocated to multiple different nodes based on their needs.
As of April 2020, the Gen-Z and CXL consortiums announced a Memorandum of Understanding describing how they are intending to cooperate in the future. In essence, CXL will focus on connectivity within the node whereas Gen-Z will focus on rack level fabric connectivity. Large HPC hardware manufacturers which are members in the Gen-Z consortium include NVIDIA, IBM and AMD, along with system integrators such as Dell and HPE. The latter vendors both having shown Gen-Z hardware used to connect multiple systems together, hinting that Gen-Z may evolve into a competitor to Ethernet and InfiniBand.

5.2.9.Summary of Intra-node interconnect main technical characteristics

The Table 7 summarises the main Intra-node interconnect main technical characteristics.

Name	PCI-e gen3	PCI-e gen4	PCI-e gen5	CXL	GEN-Z	CCIX	CAPI/openCAPI	Infinity Fabric	QPI	UPI	NVLINK
Manufacturer	N/A	N/A	N/A	N/A	N/A	N/A	N/A	AMD	Intel	Intel	NVIDIA
Version	Gen 3	Gen 4	Gen 5	V1.0		V 1.1		N/A	N/A	N/A	V 3.0
Avaibility	Available	Available	No hardware support yet	No hardware support yet	Available	Available	Available	Available	Superseded by UPI	Available	Available
Open/Proprietary	Open	Open	Open	Open	Open	Open	Open	Proprietary	Proprietary	Proprietary	Proprietary
CPU-CPU interconnect	No	No	No	No	N/A	Yes	No	Yes	Yes	Yes	No
CPU-Accelerator interconnect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes (POWER only)
Accelerator - Accelerator Interconnect	Yes	Yes	Yes	N/A	N/A	N/A	No	Yes	No	No	Yes
Switch	Switches available	Switches available	Switches available	Unknown	Switches available	Switches planned	No switches	No switches	No switches	No switches	Switches available
Lanes per port	1 to 16 lanes	1 to 16 lanes	1 to 16 lanes	16 lanes	Variable	16 lanes	8 lanes	16 to 64 lanes	16 lanes	16 lanes	8 lanes
Lanes Throughput	8 GT/s	16 GT/s	32 GT/s	32 GT/s	8 to 50 GT/s	16, 25, 32 GT/s	25 GT/s	16 GT/s	9.6 GT/s	10.4 GT/s	50 GT/s
Typical deployment	One 16 lanes port	One 16 lanes port	One 16 lanes port	One 16 lanes port	Various	One 16 lanes port	One 8 lanes port	Four 16 lanes ports in CPU-CPU connection	Two ports with 16 lanes	2 to 3 ports with 16 lanes	4 ports for GPU-GPU , 12 ports to switch
Typical combined unidirectional bandwidth	~16 GB/s	~32 GB/s	~64 GB/s	~64 GB/s		32, 50, 64 GB/s	25 GB/s	~128 GB/s in four port CPU-CPU conenction	38.4 GB/s	3 ports 62.4 GB/s	~100 GB/s GPU-GPU, ~300 GB/s to switch
Hardware Features embedded				Cache coherence		Cache coherence	Cache coherence	Cache coherence	Cache coherence	Cache coherence	Cache coherence

Table 7: Intra-node interconnect main technical characteristics

N/A means Not Available.

6. Power efficiency

Over the past 40 years, and until recently, the performance of HPC systems has grown following Moore's law. This was possible thanks to progress in semiconductor technology with continuously shrinking manufacturing processes as explained in Section 2 of this document.
While computer capacity has been evolving, the power consumption of supercomputers has also increased over the time leading to energy being one of the most expensive recurrent costs to run a supercomputing facility. As a reminder, the DOE power target was to run an Exascale machine at 20MW, which translates into 50 GFlops/W.

Looking back to 1997, the Sandia National Laboratory held the first TFlop/s computer, taking the No.1 spot on the 9th TOP500 list in June with a power consumption of 850kW resulting in 0.001 GFlops/W (HPL). Eight years later, in 2005, highest performing systems in Top10 were around 100 TFlops/s sustained performance with a power efficiency between 0.01 (CPU based) and 0.2 GFlops/W (Hybrid). In 2008, the first PFlops HPL performance was achieved on Jaguar (ORNL) and RoadRunner (Los Alamos National Laboratory - LANL) with an efficiency of 0.45 GFlops/W for the RoadRunner hybrid machine-based AMD Opteron processor and PowerXcell 8i, 45 times better than general purpose systems in 2005.

Over time, the compute capacity per watt kept increasing to reach around 17 GFlops/W in 2019 for the most energy efficient system, the Fujitsu prototype machine based only on general purpose A64FX processor, an Arm native SVE implementation. The other systems leading the Green500 list are mostly hybrid machines based on general purpose processors and NVIDIA GPUs. In June this year, the Fugaku machine was ranked number one in the Green500 and the GFlops/W was nearly 9% above the record achieved in November 19, reaching 21 GFlops/W.

Looking at the recent DOE announcements (Figure 4) or the US Exascale machines to be installed in 2021/2022 timeframe, the maximum GFlops per watt considering 100% HPL efficiency would be between 35 and 50 GFlops/W.

US Department of Energy Exascale Supercomputers			
	El Capitan	Frontier	Aurora
CPU Architecture	AMD EPYC "Genoa" (Zen 4)	AMD EPYC (Future Zen)	Intel Xeon Scalable
GPU Architecture	Radeon Instinct	Radeon Instinct	Intel Xe
Performance (RPEAK)	2.0 EFLOPS	1.5 EFLOPS	1 EFLOPS
Power Consumption	<40MW	~30MW	N/A
Nodes	N/A	100 Cabinets	N/A
Laboratory	Lawrence Livermore	Oak Ridge	Argonne
Vendor	Cray	Cray	Intel
Year	2023	2021	2021

Figure 4: US Department of Energy Exascale Supercomputers main characteristics

Considering a more realistic approach based on a minimum of 68% efficiency, the energy efficiency would be minimum 35 GFlops/W with a target of 50 Gflops/W (sustainable) which could be reached somewhere in 2023 (Figure 5). GFlops/W data from 2015 up to 2020 have been collected from Green500 [42].

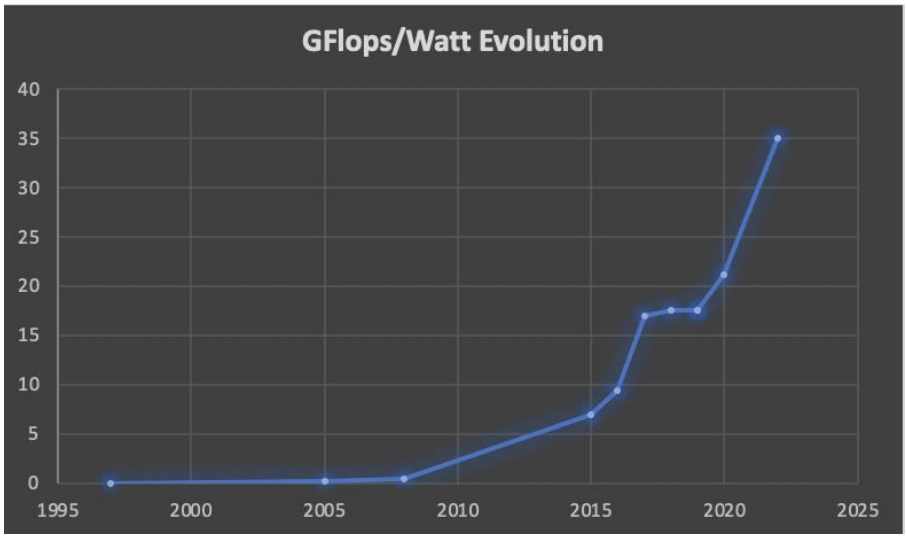

Figure 5: GFlops/Watt evolution based on collected data and Green500

Until end of 2019, there was no debate regarding which type of system could achieve the best performance-per-watt at the horizon of 2021/2022 (some systems are announced/chosen 2 years before they will be deployed): only a system built at least with a major special purpose partition (vs general purpose processor only) would have the capability to be as energy efficient as expected as both capable to support both AI and HPC workloads efficiently. In that respect, the three DOE Exascale supercomputers have already been announced based on hybrid infrastructure (AMD Genoa & ADM MI200 GPU - AMD Milan+ & AMD MI200 GPU – Intel SR & Intel Ponte Vecchio GPU) in the timeframe of 2021/2022.

End of 2019, the A64FX General Purpose Processor has demonstrated the highest Performance-Per watt (17 GFlops/W) on a prototype and mid-2020, A64FX efficiency was confirmed at large-scale on the Fugaku machine positioning Arm-based systems as serious candidates for Exascale, at least to challenge power efficiency.

7. Conclusion: Major Trends

The current section provides an outlook on future trends and summarises mid-terms projections about what users may expect in the coming years; The conclusions are based on all information gathered to build that report.

1. The big CPU market players are still Intel and AMD with their X86_64 processors with a strong competition between the two since AMD increased its market share at the end of 2019. While the X86_64 microarchitecture is still the most adopted in the HPC market, new players are taking the ARM path, while information on IBM POWER is scarce.
2. Market focus is to be part both of HPC and AI markets, with the capability to provide both CPU and GPU and to improve the CPU-GPU interconnect performance as well as the global memory bandwidth, either on pure DDR technology or by using HBM, in addition to DDR or in a DDR-less mode. As an example, upcoming Intel products feature the support of BFLOAT16 for Machine Learning applications and high-performance interfaces to their Xe GPU. The acquisition of ARM by Nvidia is also part of this trend.
3. ARM processors are continuing to expand their market share through Fujitsu (A64FX), Marvell (ThunderX – until their cancelation), Amazon (Graviton), Ampere (Altra) and other chip makers like SiPearl, the company which will design and commercialise the EPI Rhea processor, the first and only European Arm-based processor. While Amazon and Ampere might rather target the cloud sector with an extremely high core count per socket (up to 96), SiPearl (EPI Rhea) and Fujitsu (A64FX) both offer the most relevant features for HPC with less cores per socket but both support HBM and SVE capabilities.
4. It is more and more obvious that most of the high-end computing capabilities would, at least, partially, rely on accelerator technologies, with CPU technologies hosting either GPUs, accelerators (EPI Titan, NEC SX) and/or FPGAs. NVIDIA's new chip, the A100 optimised for AI and AMD GPUs will also implement machine-learning-specific optimisation with their new CDNA line, supporting AI-tailored data types. Intel will release their new Xe "Ponte Vecchio" GPU aimed at accommodating both HPC and AI users through flexible vector engines, supporting both GPU-style and CPU-style vector parallelism with enhanced CPU/GPU cache memory coherency.
5. FPGAs might see their use increased in the future for HPC and AI, mostly through Intel's acquisition of Omnitek and Altera and offer of their Stratix 10 DX product line, among others. AMD has recently also acquired Xilinx, the inventor of FGPA and adaptative SoC, to expand on the datacentre market. Similar to competitor technologies, it focuses on high interconnect and memory bandwidth and offers data types suitable for AI workloads.
6. While the landscape on the low-latency interconnect network is wider compared to a few years ago where most of the systems were InfiniBand-based, there are only a few players capable to power large scale supercomputers (> 5000 nodes): Mellanox (IB), Atos (BXI, the only European Inter-Node Interconnect), HPE Former Cray (Aries) and HPE Cray (Slingshot). Low latency Ethernet is a promising technology that will have to demonstrate its capabilities on the field.
7. Intra-node interconnect is an exciting area to follow in the future as it will allow building a tighter integration between CPU and GPU to ensure cache memory coherency and minimise data movement between CPU and GPU, also allowing to potentially rethink the design of an accelerated node with DDR-less CPU and memory consumed directly from GPU/accelerators' HBM. It will also help to support the adhesion of MCM design to go beyond the current process manufacturing limits and more powerful chips. Openness of intra-node interconnects will be key so it can see a wide adoption and ensure hardware interoperability as ease software programming.
8. Future Exascale system target efficiency of 50 GFlops/W system in order to sustain a high energy efficiency. While the first ½ Exascale class system is based on Arm architecture and has a very good power efficiency (Figure 5) due a balanced design, the future announced Exascale systems at 2021/2022 should reach around 35 sustainable GFlops per watt. 50 GFlops per watt should be achievable in the 2023 timeframe either through heterogeneous architectures based on accelerator computing capabilities combined with Arm processors or through a future well balanced Arm processor design with enhanced capabilities (AI, edge computing, etc.).
9. Further near-terms developments might include merging CPUs and GPUs onto a single die to build APU dedicated to HPC, both improving latency and memory coherency. Longer-term investment could be quantum computing: Lithography approaching the size of silicon atoms might entail incorporating quantum computing for suitable workloads, giving birth, in a first approach, to another type of hybrid systems based on current computing technologies and quantum accelerators and/or simulator.

8. References

1. A. Johansson D. Pleiter, C. Piechurski, K. Wadówka. *Data Management Services and Storage Infrastructures.* 2020. PRACE Technical Report.
2. E. Krishnasamy S. Varrette, M. Mucciardi. *Edge Computing: An Overview of Framework and Applications.* 2020. PRACE Technical Report.
3. [Online] https://www.top500.org.
4. [Online] http://www.mooreslaw.org.
5. IntelCFO. [Online] https://www.anandtech.com/show/15580/intel-cfo-our-10nm-will-be-less-profitable-than-22nm.
6. [Online] https://www.nextplatform.com/2020/08/17/the-ticking-and-tocking-of-intels-ice-lake-xeon-sp/.
7. [Online] https://adoredtv.com/exclusive-intel-sapphire-rapids-to-feature-chiplets-hbm2-400-watt-tdp-coming-in-2021/ .
8. IntelAMD. [Online] https://www.techradar.com/news/intel-admits-it-wont-catch-up-with-amds-7nm-chips-until-2021.
9. [Online] https://www.tomshardware.com/news/leaked-amd-epyc-milan-specifications-tease-possible-64-zen-3-cores-at-3-ghz.
10. [Online] https://www.tomshardware.com/news/amd-zen-3-zen-4-epyc-rome-milan-genoa-architecture-microarchitecture,40561.html .
11. AMD2022. [Online] https://www.tomshardware.com/news/amd-ddr5-usb-4-next-gen-cpus-2022.
12. EuroHPC. [Online] https://eurohpc-ju.europa.eu/.
13. EPI. [Online] https://www.european-processor-initiative.eu/project/epi/.
14. [Online] https://www.anandtech.com/show/16072/sipearl-lets-rhea-design-leak-72x-zeus-cores-4x-hbm2e-46-ddr5.
15. [Online] https://www.anandtech.com/show/15621/marvell-announces-thunderx3-96-cores-384-thread-3rd-gen-arm-server-processor .
16. [Online] https://www.nextplatform.com/2020/08/18/taking-a-deeper-dive-into-marvells-triton-thunderx3/.
17. HPCWIRE. [Online] https://www.hpcwire.com/2020/02/03/fujitsu-arm64fx-supercomputer-to-be-deployed-at-nagoya-university/.
18. ANANDTECH. [Online] https://www.anandtech.com/show/15169/a-success-on-arm-for-hpc-we-found-a-fujitsu-a64fx-wafer.
19. NEXTPLATFORM. [Online] https://www.nextplatform.com/2019/09/24/europeans-push-fpga-powered-exascale-prototype-out-the-door/.
20. ANANDTECH-2. [Online] https://www.anandtech.com/show/15578/cloud-clash-amazon-graviton2-arm-against-intel-and-amd.
21. EXTRTECH. [Online] https://www.extremetech.com/extreme/306951-ampere-altra-arm-cpus-launch-with-up-to-80-cores-to-challenge-xeon-epyc.
22. AMPERECOMP. [Online] https://amperecomputing.com/ampere-altra-industrys-first-80-core-server-processor-unveiled/.
23. VENTUREBEAT. [Online] https://venturebeat.com/2020/03/03/ampere-altra-is-the-first-80-core-arm-based-server-processor.
24. ANANDAMP. [Online] https://www.anandtech.com/show/15575/amperes-altra-80-core-n1-soc-for-hyperscalers-against-rome-and-xeon.
25. https://www.nextplatform.com/2019/08/06/talking-high-bandwidth-with-ibms-power10-architect/. [Online]
26. TIANHE-3. [Online] https://www.nextplatform.com/2019/05/02/china-fleshes-out-exascale-design-for-tianhe-3.
27. [Online] https://www.aiforhumanity.fr/en/.
28. NVIDIAAMP. [Online] https://devblogs.nvidia.com/nvidia-ampere-architecture-in-depth/.
29. AMD. [Online] https://www.anandtech.com/show/15593/amd-unveils-cdna-gpu-architecture-a-dedicated-gpu-architecture-for-data-centers.
30. AMDEpyc. [Online] https://www.servethehome.com/amd-cdna-gpu-compute-architecture-5nm-epyc/.
31. [Online] https://www.hpcwire.com/2020/07/30/intels-7nm-slip-leaves-questions-about-ponte-vecchio-gpu-aurora-supercomputer/.
32. [Online] https://www.anandtech.com/show/15119/intels-xe-for-hpc-ponte-vecchio-with-chiplets-emib-and-foveros-on-7nm-coming-2021.
33. [Online] https://www.anandtech.com/show/15188/analyzing-intels-discrete-xe-hpc-graphics-disclosure-ponte-vecchio/2.
34. oneAPI. [Online] https://software.intel.com/content/www/us/en/develop/tools/oneapi.html.
35. oneAPI-2. [Online] https://www.anandtech.com/show/15188/analyzing-intels-discrete-xe-hpc-graphics-disclosure-ponte-vecchio/4.

36. [Online] : https://moorinsightsstrategy.com/wp-content/uploads/2020/05/Graphcore-Software-Stack-Built-To-Scale-By-Moor-Insights-And-Strategy-2.pdf.
37. FPGA. [Online] https://techcrunch.com/2019/04/16/intel-acquires-uks-omnitek-to-double-down-on-fpga-solutions-for-video-and-ai-applications/.
38. Stratix. [Online] https://www.intel.com/content/www/us/en/products/programmable/fpga/stratix-10.html.
39. Agilex. [Online] https://www.anandtech.com/show/14149/intel-agilex-10nm-fpgas-with-pcie-50-ddr5-and-cxl.
40. [Online] https://www.hpcwire.com/off-the-wire/intel-omni-path-business-spun-out-as-cornelis-networks/.
41. *IBM's POWER10 processor.* W.Starke B. Thompto. brak miejsca : Hot Chips 32, 2020.
42. [Online] https://www.top500.org/lists/green500/.

9. List of acronyms

2D	2-Dimension
AI	Artificial Intelligence
AI	Artificial Intelligence
AVX	Advanced Vector Extensions
AWS	Amazon Web Service
BFLOAT	Brain floating-point format
BLAS	Basic Linear Algebra Subprograms
BSP	Bulk-synchronous parallel
BXI	Bull eXascale Interconnect
CAPI	Coherent Accelerator Processor Interface
CCD	Compute Core Die
CCD	Compute Core Die
CCIX	Cache Coherent Interconnect for Accelerators
CCPI	Cavium Cache Coherent Interconnect
CDNA	GPU architecture for data centre compute
CEO	Chief Executive Officer
CFO	Chief Financial Officer
CISC	Complex Instruction Set Computer
CMG	Core Memory Group
CP	The EPI Common Platform
CPU	Central Processing Unit
CU	Compute Unit
CXL	Compute Express Link
DDR	Double Data Rate
DIMM	Dual in-line memory module
DMA	Direct Memory Access
DoE	Department of Energy
DP	Double Precision
DPC	DIMM Per Channel
DPC++	Distributed Parallel C++
DRAM	Dynamic Random Access Memory
EFLOPS	Exa Flops
eFPGA	embedded FPGA
EPI	European Processor Initiative
EU	European Union
EUs	Execution Units
EUV	Extreme-Ultraviolet Lithography
FFT	Fast Fourier Transform
FMA	Fused Multiply Add
FP16	Floating Points 16 bits
FP32	Floating Points 32 bits
FP64	Floating Points 64 bits
FPGA	Field Programmable Array
GB	Gigabyte
GB/s	Gigabyte per second
Gbits/s	Gibabits per second
Gbps	Gigabit per second
GCC	GNU Compiler Collection
GFS	Global File System
GHz	GigaHertz
GNA	Gaussian Neural Accelerator
GPP	General Purpose Processor
GPU	Graphics Processing Unit
GT/s	Giga Transfers per second
HBM	High Bandwidth Memory
HDR	High Data Rate
HIP	Heterogeneous-Computing Interface for Portability
HPC	High Performance Computing
HSL	High Speed Links
HSM	Hardware Security Modules

IDS	Intrusion Detection System
IF	Infinity Fabric
iGPU	Integrated Graphics Processing Unit
INT16	Integer 16 bits
INT8	Integer 8 bits
IO	I/O Input/Output
IP	Intellectual Property
IP	Internet Protocol
IPU	Intelligence Processing Unit
ISA	Instruction Set Architecture
KB	Kilobytes
KiB	Kibibyte
kW	Kilowatt
L	level
LANL	Los Alamos National Laboratory
LAPACK	Linear Algebra Package
LGA	Land Grid Array
LLNL	Lawrence Livermore National Laboratory
MB	Megabytes
MCM	Multi-Chip Module
MiB	Mebibyte
MIG	Multi-Instance GPU
ML	Machine Learning
MPI	Message Passing Interface
MPPA	Multi-Purpose Processing Array
MT/s	Mega Transfers per second
MW	Megawatt
N/A	Not Available
NDR	Next Data Rate
NIC	Network Interface Card
nm	Nanometre
NoC	Network-on-Chip
NUMA	Non-uniform Memory Architecture
NVDIMM	Non-volatile dual in-line memory module
NVLink	NVIDIA Link
NVMe	Non-Volatile Memory Express
NVSwitches	NVLink switches
OAM	OCP Acceleration Module
OCP	Open Compute Project
OMI	Open Memory Interface
OPA	Omni-Path Architecture
ORNL	Oak Ridge National Laboratory
OS	Operating System
PAM-4	Pulse Amplitude Modulation with 4 levels
PB	Petabyte
Pbps	Petabytes per second
PCI-e	Peripheral Component Interconnect Express
PFLOPS	Peta FLOPS
PUE	Power Usage Effectiveness
PVC	Ponte Vecchio
Q	Quarter
QoS	Quality of Service
QPI	QuickPath Interconnect
RDMA	Remote Direct Memory Access
RISC	Reduced Instruction Set Computer
RoCE	RDMA over Converged Ethernet
ROCm	Radeon Open Compute
SAS	Serial Attached SCSI
SATA	Serial ATA
SEV-ES	Secure Encrypted Virtualisation-Encrypted State
SHARP	Scalable Hierarchical Aggregation and Reduction Protocol
SIMD	Single Instruction Multiply Data

SKU	Stock Keeping Unit
SMT	Simultaneous Multithreading
SoC	System on Chip
SP	Scalable Platform
SPE	Synergistic Processing Element
SR-IOV	Single Root Input/Output Virtualisation
SSD	Solid State Drive
STX	Stencil/Tensor Accelerator
SVE	Scalable Vector Extension
TB	Terabyte
Tbps	Terabits per second
TDP	Thermal Design Power
TF32	TensorFloat-32
TFLOPS	Tera FLOPS
TSMC	Taiwan Semiconductor Manufacturing Company
UDP	User Datagram Protocol
UPI	Ultra Path Interconnect
VE	Vector Engine
VH	Vector Host
VHDL	Very High Speed Integrated Circuits Hardware Description Language
VM	Virtual Machine
VRP	Variable Precision accelerator
W	Watt
XEMF	XE Memory Fabric

Acknowledgements
This work was financially supported by the PRACE project funded in part by the EU's Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 823767.