T/AW087/21
Support and Coordination

2057699-TN-01-04

Hardware at the Exascale
Revision 4.0

Steven Wright, Ed Higgins, Ben Dudson, Peter Hill, and David Dickinson

University of York

Gihan Mudalige, Ben McMillan, and Tom Goffrey

University of Warwick

August 24, 2022

Contents

Glossary

AVX Advanced Vector eXtensions

CFD Computational Fluid Dynamics

DIMM Dual In-line Memory Module

DRAM Dynamic Random Access Memory

DSL Domain Specific Language

eDSL Embedded Domain Specific Language

FLOP/s Floating point operations per second

FPGA Field Programmable Gate Array

HBM High Bandwidth Memory

ILP Instruction Level Parallelism

ISA Instruction Set Architecture

JIT Just-in-time Compilation

MCDRAM Multi-Channel DRAM

N-1 N processes writing data to a single file

N-N N processes writing data to their own files

N-M N processes writing to M files

PCIe Peripheral Component Interconnect Express

SIMD Single-instruction, multiple-data

SMT Simultaneous multi-threading

SPMD Single-program, multiple-data

SSE Streaming SIMD Extensions

SVE Scalable Vector Extensions

Changelog

March 2022

- Reorganisation of document, combining elements of the previous four reports, 2047358-TN-01, 2047358-TN-02, 2047358-TN-03 and 2047358-TN-04 into a single report on hardware platforms.
- Updated some information regarding computational hardware to bring data up to date with developments as of March 2022.
- Updated listing of pre- and post-Exascale systems, specifically those planned in the US, Europe and the rest of the World.

July 2022

- Minor updates to the Summary.
- Restructure of Section 2, Hardware. This restructure means that each manufacturer has a dedicated section.
- Up to date information on Intel, AMD and NVIDIA architectures. Updates to other sections also.
- Update to section 3, with latest information on European and US systems

1 Summary

The end of CPU clock frequency scaling in 2004 gave rise to multi-core designs for mainstream processor architectures. The turning point came about as the current CMOS-based microprocessor technology reached its physical limits, reaching the threshold postulated by Dennard in 1974 [1]. The end of Dennard scaling has meant that further increases in clock frequency would result in unsustainably large power consumption, effectively halting a CPUs ability to operate within the same power envelope at higher frequencies.

More than a decade and a half has passed since the switch to multi-core, where we now see a golden age of processor architecture design with increasingly complex and innovative designs used to continue delivering performance improvements. The primary trend continues to be the development of designs that use more and more discrete processor "cores" with the assumption that more units can do more work in parallel to deliver higher performance by way of increased throughput. This has aligned well with the hardware industries' ambition to see the continuation of Moore's Law – exponentially increasing the number of transistors on a silicon processor.

As a result, on the one hand we see traditional CPU architectures gaining more cores, currently over 20 cores for high-end processors, and increasing vector lengths (e.g. Intel's 512-bit vector units) per core, widening their ability to do more work in parallel. On the other hand we see the widespread adoption of separate devices, called accelerators, such as GPUs that contain much larger numbers (over 1024) of low-frequency (power) cores, targeted at speeding up specific workloads.

More cores on a processor has effectively resulted in making calculations on a processor, usually measured by floating-point operations per second (FLOP/s), cheap. However feeding the many processors with data to carry out the calculations, measured by bandwidth (bits/sec), has become a bottleneck. As the growth in the speed of memory units has lagged that of computational units, multiple levels of memory hierarchy have been designed, with significant chunks of silicon dedicated to caches to bridge the bandwidth/core-count gap.

New memory technologies such as High Bandwidth Memory (HBM) has produced "stacked memory" designs where embedded DRAM is integrated on to CPU chips. New non-volatile memory technologies such as Intel's 3D X-Point (Optane) memory, which can be put in traditional DIMM memory slots, provides higher storage capacity but with lower bandwidth. The memory hierarchy has been further extended off-node, with burst buffers and I/O nodes serving as staging areas for scientific data en route to a parallel file system. Larger and more heterogeneous machines have also necessitated more complex interconnection strategies. Technologies such as NVLink allows GPUs to communicate point-to-point without requiring data to travel through the CPU. New high-speed interconnects have been developed that seek to minimise the number of *hops* required to move data between nodes and devices, potentially benefiting both inter-node communications and file system operations.

A decade ago, the vast majority of the fastest HPC systems in the world were homogeneous clusters based around the x86-64 architecture, with a few notable exceptions such as the IBM BlueGene architectures. Now, there is a diverse range of multi-core CPUs on offer, supported by an array of manycore co-processor

architectures, complex high-speed interconnects, and multi-level parallel file systems.

The underpinning expectation of the switch to multi-core and the subsequent proliferation of complex massively parallel hardware was that performance improvements could be maintained at historical rates. However, this has led to the need of a highly skilled parallel programming know-how to fully exploit the full potential of these devices and systems. The switch to parallelism and its consequences was aptly described by David Patterson in 2010 as a "Hail-Mary pass", an act done in desperation by the hardware vendors "without any clear notion of how such devices would in general be programmed" [2].

Nearly a decade later, industry, academia and stakeholders of HPC have still not been able to provide an acceptable and agile software solution to this issue. The problem has become even more significant with the current deployment of Exascale-capable HPC systems, limiting their use for real-world applications for continued scientific delivery. On the one hand, open standards have been slow to catch up with supporting new hardware, and for many real applications have not provided the best performance achievable from these devices. On the other hand, proprietary solutions have only targeted narrow vendor-specific devices resulting in a proliferation of parallel programming models and technologies.

In this report, we provide a survey of the hardware that is present, or likely to be present, in post Exascale systems.

The remainder of this report is organised as follows:

Section 2 reviews the current hardware landscape, and outlines the hardware expected in the coming five years.

Section 3 provides a summary of some of the pre- and post-Exascale machines currently being delivered, or expected to be delivered in Europe and the United States.

Section 4 describes the systems that are available for our evaluation, and why they are relevant to a post-Exascale world.

2 Hardware Roadmaps

In this section, we briefly introduce the architectures that are available, or likely to become available in the coming years.

The HPC hardware landscape is dominated by a small number of manufacturers, and so in this section we will focus primarily on the roadmaps released by each of these vendors. Specifically, this section will focus on current and upcoming hardware from Intel, AMD and NVIDIA, and on products in the ARM family of processors. Alternative architectures and technologies will be discussed at the end of this section.

Recent trends in supercomputing show that reaching Exascale likely requires a heterogeneous approach, or at the very least the use of manycore architectures (i.e., processors with a high number of parallel cores) [3, 4]. There are already a number of systems in use or in active development that embody this principle – composed of computational nodes coupling a multi-CPU architecture with GPU accelerators.

2.1 Intel

Over the past decade, **Intel** has dominated large HPC installations. In November 2020, 90% of the Top500 were using Intel Xeon processors to provide some or all of their performance.

Between 2007 and 2016, Intel operated using a Tick-Tock production model, where each die shrink (tick) was followed by a microarchitecture change (tock). This has been succeeded by a Process-Architecture-Optimization model. Figure 1 shows Intel's current process roadmap.

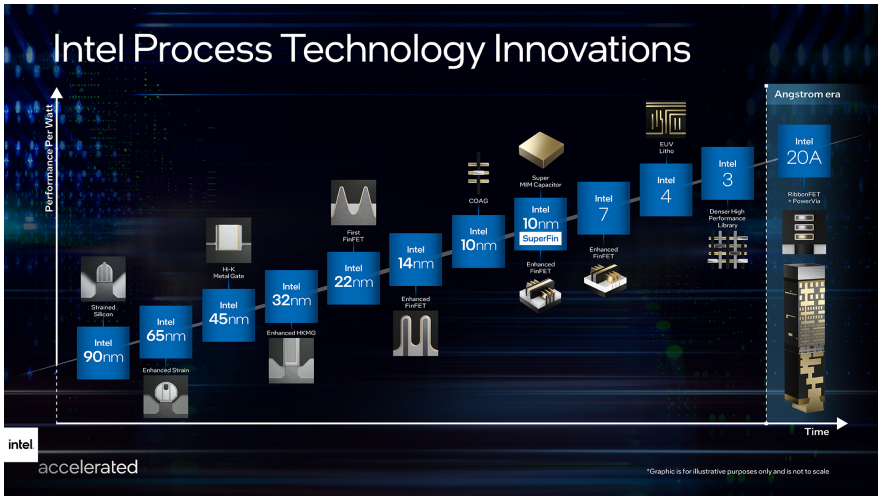

Figure 1: Intel's Process Roadmap

2.1.1 CPUs

The most widely used Intel Xeon CPUs currently are **Skylake** and **Cascade Lake**. The next generation, **Ice Lake** was released in 2021 and includes a number of enhancements including a new manufacturing process and additional memory channels. The upcoming **Sapphire Rapids** CPU includes a number of important architectural improvements over the current generation Xeon CPUs.

Key Features of Skylake and Cascade Lake

- Both manufactured on a 14 nm process.
- Skylake available with 2 to 28 cores, Cascade Lake with 4 to 56.
- Cascade Lake additionally supports **Intel Optane Persistent Memory** DIMMs
- Cascade Lake also adds some instructions targeted at Neural Networks.
- Skylake introduced **AVX-512** vector instructions from the **Intel Xeon Phi** product line.

[1] `https://www.servethehome.com/intel-details-sapphire-rapids-xeon-at-architecture-day-2021/`

Key Features of Ice Lake and Sapphire Rapids

- Ice Lake is manufactured using a new 10 nm production process.
- Sapphire Rapids will be manufactured using a new **"Intel 7"** process (10 nm Enhanced SuperFin).
- Both architectures feature two additional memory channels.
- Sapphire Rappids supports DDR5 memory[1].
- Ice Lake supports PCIe Gen 4.
- Sapphire Rapids adds support for PCIe Gen 5, and Intel's new **Compute eXpress Link (CXL)**, designed for connecting CPUs and accelerators.
- Sapphire Rapids adds **Advanced Matrix Extensions (AMX)** aimed at AI inference and training.
- Sapphire Rapids will have increased Last Level Cache (LLC), and support for HBM2.

2.1.2 Accelerators

Intel's first foray into computational accelerator was the now cancelled Intel Xeon Phi range. The first platform in the Phi range was the **Knights Corner**, which was available as a PCIe accelerator card. These accelerators provided much of the compute on China's Tianhe-2 system in 2015.

The second architecture, the **Knights Landing**, was available as a host platform and was present in the Stampede2, Cori and Trinity systems. Prior to its cancellation, Argonne's Exascale system, Aurora, was set to use an Intel Xeon Phi platform. This system will now be supported by Intel's new **Xe** discrete GPU.

Key Features of the Intel Xeon Phi range

- **Knights Corner (KNC)** was manufactured using a 22 nm process.
- **Knights Landing (KNL)** used a 14 nm process.
- Both KNC and KNL were capable of 4-way Simultaneous Multithreading (SMT).
- KNC introduced 512-bit wide vector instructions.
- KNL used **AVX-512**, now available in some Xeon CPUs.
- KNL also provided stacked **MCDRAM** (Multi-channel DRAM) high bandwidth memory.

Key Features of Intel's Xe Product Line[2]

- **Ponte Vecchio** Xe-HPC will be manufactured using a 7 nm process.
- PVC will feature a new Instruction Set Architecture (ISA).
- PVC will feature second generation high bandwidth memory (HBM2e).
- PVC will allow GPU-GPU and GPU-CPU communication via CXL.
- Prerelease A0 silicon achieved over 45 TFLOP/s in 32-bit precision.
- It will be succeeded by **Rialto Bridge**, which will also be part of an XPU codenamed **Falcon Shores** (CPU+GPU).

[2] https://www.nextplatform.com/2021/08/24/intels-ponte-vecchio-gpu-better-not-be-a-bridge-too-far/

2.2 AMD

Intel's main competitor in the x86_64 market comes from AMD. While once a mainstay of server architectures, AMD suffered a significant decline in popularity between 2010 and 2015. AMDs EPYC line of CPUs seems to be reversing this trend, with some of the largest systems currently being developed and deployed making use of their CPUs and GPUs.

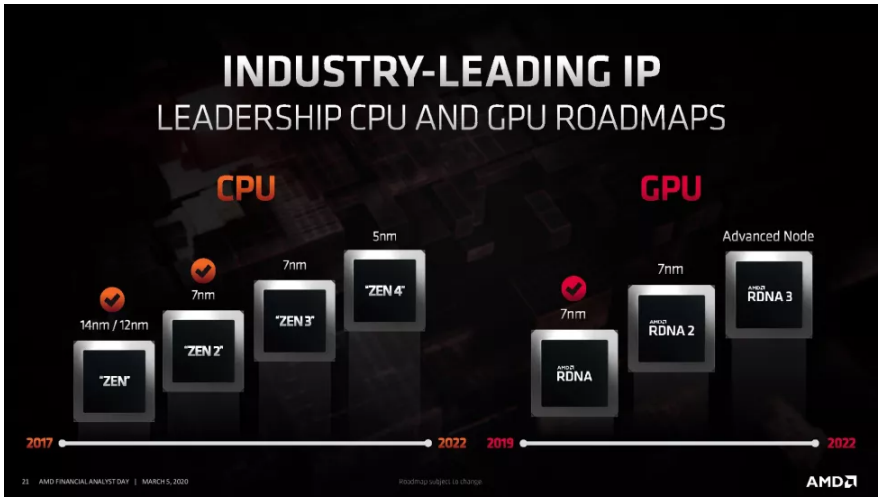

Figure 2: AMD's Process Roadmap

2.2.1 CPUs

AMD re-entered the server marking in 2017 with their EPYC line of processors based on their **Zen** microarchitecture. The first processor released in this series was codenamed **Naples**, based on the Zen 1 architecture. This was followed by **Rome** and **Milan**. The fourth iteration of the Zen architecture will be used for the forthcoming **Genoa** line.

Key Features of Rome and Milan

- Rome was released in 2019 and is based on the Zen 2 architecture.
- Milan was released in 2021 and is based on Zen 3.
- Both Rome and Milan are manufactured using a 7 nm process with 14 nm I/O components.
- Both generations are available with up to 64 cores, with SIMD extensions up to AVX2.
- Both generations support PCIe Gen 4.
- Both provide 8 memory channels.
- **Frontier** uses a Zen 3 variant, codenamed **Trento**.
- Trento is a Milan variant that is optimised for AI, with InfinityFabric 3.0, to allow a coherent memory interface with GPUs.

[3] `https://www.techradar.com/uk/news/gigabyte-hacker-spills-details-on-next-generation-amd-epyc-genoa-series`

Key Features of Genoa

- Genoa will be released in 2022, and will use the Zen 4 microarchitecture.
- It will be fabricated using a a 5 nm process.
- The Zen 4 microarchitecture will the first to include support for AVX-512.
- Genoa will provide 12 DDR5 memory channels.
- It will be available in up to 96 cores with 2-way SMT[3].

2.2.2 Accelerators

AMDs current line of Server GPUs is the Instinct series, launched in 2016. The first products in the Instinct line made use of AMDs Graphics Core Next (GCN) architecture. This was superseded in 2019 by the RDNA (Radeon DNA) and **CDNA (Compute DNA)** [5] architectures, targeted at Gaming and Compute, respectively.

AMD Instinct GPUs provide most of the power for Frontier, installed at Oak Ridge National Laboratory. The will also be a key component of the El Capitan supercomputer, to be installed at Lawrence Livermore National Laboratory.

Key Features of Vega and Arcturus Product Lines

- The **MI50** and **MI60** GPUs are from the Vega range and were released in 2018.
- They use the 5th generation of the GCN architectures.
- The **MI100** Arcturus GPU was the first to use the CDNA microarchitecture, and was released in 2020.
- Both Vega and Arcturus products lines are manufactured with a 7 nm process.
- The MI50 can provide 6.6 TFLOP/s double-precision performance, while the MI60 can provide 7.3 TFLOP/s.
- The MI100 can provide 11.5 TFLOP/s in double-precision.
- Both Vega GPUs offer up to 32 GB HMB2, while the Arcturus GPU offers up to 64GB.
- Each GPU can be connected with PCIe Gen 4, and can communicate GPU-GPU via **Infinity Fabric**.

Key Features of Aldebaran Product Line

- The Aldebaran line consists of the **MI210**, **MI250** and the **MI250X**, each released in 2021/22.
- The MI250X is the GPU used in the **Frontier** system.
- The Aldebaran line is manufactured using a 6 nm process and uses the CDNA 2 microarchitecture.
- The MI210 provides 64 GB of HBM2e, while the MI250 and MI250X provides up to 128 GB.
- The double-precision performance ranges from 22.6 TFLOP/s on the MI210, up to 47.87 TFLOP/s on the MI250X.
- Like the predecessors, each is connected with PCIe Gen 4, and allows communication via Infinity Fabric.

2.3 NVIDIA

GPUs have featured heavily in the top supercomputers in the world over the past decade. NVIDIA has been the dominant manufacturer of the GPUs in these systems for much of this time. Notable systems to employ NVIDIA accelerators include Tianhe-1A, Titan and Summit (all achieved #1 ranking).

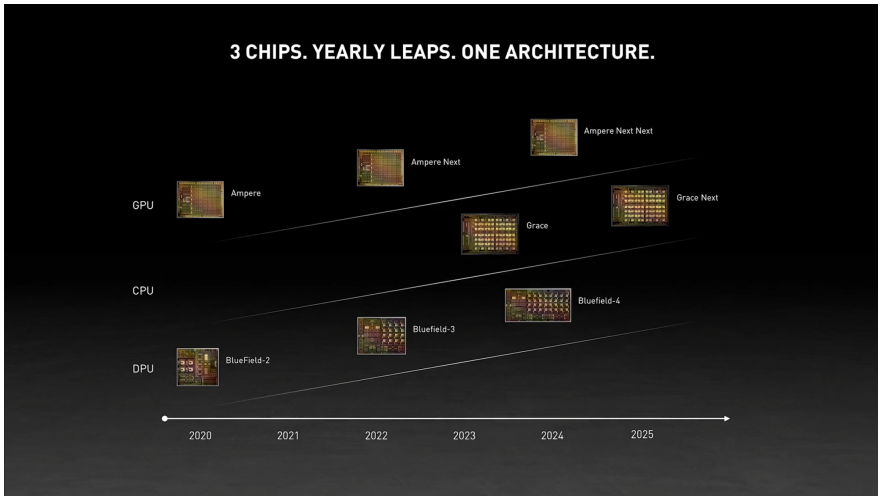

Figure 3: NVIDIA Process Roadmap

2.3.1 Accelerators

NVIDIAs dominance in the GPGPU market began with the launch of their **Tesla** range in 2007, alongside their **CUDA** programming model. The most recent HPC-focused architectures (each named after an eminent scientists) are **Volta** and **Ampere**; this will soon be followed by the **Hopper** architecture.

Key Features of Volta and Ampere

- **Volta** (V100) uses a 12 nm process and can provide 7.8 TFLOP/s of double-precision performance.
- The V100 is available with 16 or 32 GB HBM2 and can be connected via PCIe Gen 3.
- CPU-GPU and GPU-GPU communications can be achieved through NVLink 2.0.
- Volta extends on the previous Pascal architecture, introducing **Tensor cores** targetted at AI.
- **Ampere** (A100) uses an 7 nm process and can provide 9.7 TFLOP/s of performance.
- The A100 is available with 40 or 80 GB HBM2 and can be connected via PCIe Gen 4.
- The Ampere architecture also allows CPU-GPU and GPU-GPU communications via NVLink 3.0.
- The A100 expands on the Volta architecture with additional functionality for half-precision operations on the Tensor cores.

Key Features of Hopper

- **Hopper** (H100) will use a 5 nm process and will be released in 2022.
- The H100 will be available with 80 GB of HBM3 memory.
- The H100 can be connected via PCIe Gen 5 card and can communicate via NVLink 4.0.
- It will be NVIDIAs first multi-chip-module (MCM) design.
- The H100 will be able to provide 24 TFLOP/s in double-precision.

2.3.2 CPUs

While NVIDIA has traditionally focused on accelerator devices, they will enter the data-centre CPU market in 2023 with their **Grace** CPU. It will be available in two forms, both dubbed "Superchips".

Key Features of Grace[4]

- The **Grace CPU Superchip** combines 2 Grace CPUs using NVLink-C2C.
- This will provide up to 144 Arm v9 cores in a single socket.
- It will use LPDDR5x (low-power DDR) memory and provide up to 1 TB/s bandwidth.
- The **Grace Hopper Superchip** pairs a Grace CPU with a Hopper GPU.
- The NVLink-C2C connection is 7× faster than PCIe Gen 5.

2.4 Arm

While not a producer of CPUs or GPUs, Arm develop architectures that have long been successful in the mobile market. These architectures have been adopted by some manufacturers in making HPC-ready architectures. In particular, Marvell, NVIDIA and Fujitsu have been producing Arm-based CPUs for use in HPC systems.

Key Features of Arm architectures

- **Marvell** and **Fujitsu** produce chipsets based on the ARMv8.1-A and ARMv8.2-A ISAs
- Marvell **ThunderX2** powers Isambard in UK [6], and also the first Petascale Arm supercomputer, Astra, at Sandia National Laboratories.
- ThunderX2 is manufactured using a 16 nm process and has up to 32 cores with 4-way SMT
- **ThunderX3** is not available as a general-purpose architecture, which may limit its use in HPC.
- Fujitsu manufacture the **A64FX** chipset, that powers Fugaku (the current #2 system).
- A64FX uses a 7 nm process and contains 48 compute cores with 2-4 assistant cores.
- A64FX has 512-bit Scalable Vector Extensions (SVE).
- It contains 32 GB of on package HBM2.
- The **NVIDIA Grace** architecture is based on ARMv9 (details are given in Section 2.3.2.)

[4] `https://www.nvidia.com/en-us/data-center/grace-cpu/`

2.5 Other Architectures

Besides the CPUs and GPUs manufactured by NVIDIA, Intel and AMD, and architectures based on ARM's ISA, there are a number of other architectures featured in Top500 machines. We do not expect some of these architectures to see widespread adoption in Exascale machines, and some of these architectures are specific to Chinese systems. Nonetheless, they are discussed briefly here for completeness.

The recent Summit and Sierra machines were both primarily powered by NVIDIA GPUs connected to IBM Power9 CPUs. Although IBM have a long history in HPC architectures, it is not expected that next generation Power CPUs will be generally present at Exascale.

Key Features of IBM Power architectures

- All **IBM BlueGene** systems were driven by PowerPC architectures.
- **Sierra** and **Summit** were both powered by **Power9** CPUs with NVIDIA V100 GPUs.
- Power9 is manufactured on a 14 nm process, with up to 24 cores and 4-way SMT.
- Power9 is notable for its inclusion of NVLink (allowing GPU-GPU communication).
- IBM **Power10** was released in September 2021.
- It is manufactured using a 7 nm process and contains 15 cores, with 8-way SMT.
- It can support DDR5, GDDR6 or HBM2.
- It supports PCIe Gen 5, but has dropped support for NVLink.

Due to various export restrictions between the US and China, many Chinese systems now use locally-developed architectures. These architectures power some of the biggest and fastest supercomputers in the world, but it is unlikely these architectures will be adopted outside of China.

Architectures featured in Chinese machines

- The **Sunway SW26010** manycore processor powers the **TaihuLight** system.
- Each SW26010 CPU contains 260 cores, with 512-bit wide SIMD.
- Each core can deliver 3.06 TFLOP/s in double precision.
- NUDT's **Matrix2000** accelerators replaced Intel's KNC accelerators in **Tianhe-2A**.
- Each accelerator contained 128 RISC cores, with 256-bit wide SIMD. • Each card can provide 2.46 TFLOP in double precision.
- There is a join venture between AMD and China to licence their Zen architectures.
- The **Hygon Dhyana** processor is a variant of AMD's EPYC CPU for the Chinese market.

2.6 Reconfigurable Architectures

For the past decade, accelerator architectures have demonstrated the benefit of hardware specialisation to achieving high performance. Field-Programmable Gate Arrays (FPGAs) may represent the next step towards application-specific hardware. At compile-time, entire algorithms can be synthesised as sequential logic circuits in hardware [7, 8].

The use of reconfigurable hardware in large HPC installations is currently rare, but there are signs that this may change as new programming models emerge. In particular, both OpenCL and Intel's Data Parallel

C++ can target FPGAs directly. Further, since FPGAs can synthesise circuitry specific to a computational kernel, they are able to eliminate computational units that would otherwise be powered but unused on CPU- and GPU-like architectures – potentially reducing energy wastage.

It should be noted that, while a number of recent studies [7, 8] have shown that FPGAs can achieve comparable performance to GPUs on some kernels, specialised non-trivial optimisations are required, coupled with long compilation times. The relative immaturity of the compiler toolchains, means that currently targeting FPGAs may significantly harm developer productivity.

Key Features of FPGA Architectures

- The market leaders are Xilinx (now part of AMD) and Intel (following their acquisition of Altera).
- Xilinx's flagship datacentre card is the **Alveo U280**.
- It is manufactured using 16 nm process and connected through PCIe Gen 4.
- The U280 provides 8 GB of HBM2 memory.
- Xilinx's new ACAP (Adaptive Compute Acceleration Platform) is the **Versal VCK5000**[5].
- It is manufactured using a 7 nm process and is available in multiple configurations.
- It is interconnected with PCIe Gen 5 and can provide HBM2e on-board memory.
- Alongside the FPGA units, there are two dual-core Arm Cortex CPUs (one big and one little).
- Intel's current datacentre FPGA offering is the **Stratix 10**.
- It is manufactured with a 14 nm process and connected via PCIe Gen 4.
- It can provide either 8 or 16 GB of HBM2.
- Intel's next generation FPGA will be the **Agilex M-Series**[6].
- It will be manufactured with the Intel 7 process, and will provide HBM2e on-board memory.
- It will also provide 400 G Ethernet network-on-chip, and can additionally communicate FPGA-FPGA through the Compute Express Link (CXL).

2.7 Comparison and Summary

The end of the "free lunch" [9] and the breakdown of Dennard scaling [10] has meant that today's performance improvements come from increasing parallelism rather than clock speed. Server-grade CPUs typically contain 10-50 cores (and some future CPUs may feature up to 100), and offer increasingly wide vector operations. GPUs and other accelerators, that offer hundreds of simple cores, now represent a significant proportion of the compute available on many of the world's biggest supercomputers.

Many of the architectures described in this section are either already present in pre- and post-Exascale systems, or are expected to feature in the near future. Tables 1 and 2 outline many of the features and differences between the CPUs and GPUs likely to be present at Exascale. Table 3 provides the same outline for reconfigurable architectures. Although reconfigurable architectures are not yet expected to feature heavily, improved programming models and compiler toolchains may make these technologies more viable in the future.

[5] https://www.hpcwire.com/2022/03/08/amd-xilinx-takes-aim-at-nvidia-with-improved-vck5000-inferencing-card/
[6] https://www.hpcwire.com/off-the-wire/intel-introduces-agilex-m-series-fpgas/

	Intel Sapphire Rapids	AMD Genoa	NVIDIA Grace	Fujitsu A64FX
Process	Intel 7 (10 nm)	7 nm	Unknown	7 nm
Cores	up to 48	up to 96	up to 144	up to 48
Threads	2 per core	2 per core	Unknown	4 per core
Memory Technology	8-channel DDR5	12-channel DDR5	LPDDR5x	HBM2
Interconnect	PCIe Gen 5	PCIe Gen 5	NVLink-C2C	Unknown
Instruction Set Architecture	x86_64, AVX-512	x86_64, AVX-512	Arm v9	Arm v8.2-A, 512-bit SVE

Table 1: Summary of CPU architectures likely to be present in Exascale Systems

	NVIDIA Hopper	AMD Radeon Instinct MI250X	Intel Xe Ponte Vecchio
Process	5 nm	6 nm	7 nm
Peak DP Performance	24 TFLOP/s	47.87 TFLOP/s	$\sim$ 20+ TFLOP/s
Memory Technology	80 GB HBM3	128 GB HBM2e	HBM2e
Interconnect	PCIe Gen 5, NVLink 4.0	PCIe Gen 4	CXL

Table 2: Summary of accelerator architectures likely to be present in Exascale Systems

	Xilinx Alveo U280	Xilinx Versal	Intel Stratix 10	Intel Agilex
Process	16 nm	7 nm	14 nm	Intel 7 (10 nm)
Memory Technology	8 GB HBM2	HBM2e	8/16 GB HBM2	HBM2e
Interconnect	PCIe Gen 4	PCIe Gen 5	PCIe Gen 4	CXL
Other Features	–	two dual-core Cortex CPUs	–	400 G Ethernet Network-on-Chip

Table 3: Summary of reconfigurable architectures

The diversity of architectures that are, or will be, available at Exascale represents a significant challenge for users of these systems – the majority of pre- and post-Exascale systems currently being installed will use both CPUs and Accelerators to achieve their stated performance. With this in mind, being able to develop applications and algorithms that can exploit the hierarchical parallelism likely to be available on Exascale systems will be vitally important. Even considering the likely prevalence of GPUs, the extensive use of GPU-GPU communication, and MPI-Aware programming models the architectures provided differ sufficiently such that a platform-agnostic approach will be vital to the success of any future-proofed Fusion simulation code.

3 Systems

As we enter the era of Exascale computing, it is clear that heterogeneity is going to play a part in most of the first generation of systems. This shift towards *accelerated* computing has been coupled with increasing diversity in the architectures available in HPC. Developing applications for these post-Exascale systems therefore requires careful consideration of and preparation for the systems they are expected to be executed on.

3.1 Pre-Exascale Systems

3.1.1 The United Kingdom

In the UK, Supercomputing is focused around Universities, often funded by UKRI, and a small number of commercial sites. Currently, the biggest systems in the UK are those those found at research laboratories such as the Met Office (#75, #207 and #208 on Top500.org at the time of writing), ECMWF (#128 and #129) and AWE (#156). Each of these systems are homogeneous clusters using Intel CPUs, typically supporting applications that have been developed over a long period of time, in Fortran or C/C++, using MPI to distribute work across the cluster.

In 2021, the Met Office announced that its next system will also be a homogeneous cluster, but will be based on AMD Milan CPUs and delivered by Microsoft. It will deliver approximately 60 PetaFLOP/s of performance (i.e. 8× more powerful than their current XC40 system).

The HPC provision provided by UK Universities is structured in the form of a tiered system. The UK's national (Tier-1) supercomputer, **ARCHER2**, was recently installed at the Edinburgh Parallel Computing Centre (EPCC). Like many of the HPC systems in the commercial sector, ARCHER2 is an homogeneous system with AMD Rome CPUs and delivers approximately 20 PetaFLOP/s of performance.

It is at the regional (Tier-2) centres where there is a wealth of architectural diversity. The **Isambard** system, installed at the University of Bristol, is predominantly an ARM-based system, with one cabinet of Marvell ThunderX2 compute nodes and one cabinet of Fujitsu A64FX CPUs. Besides this, it contains the Multi-Architecture Comparison System (MACS), which consists of a range of alternative platforms for evaluation, including NVIDIA P100 and V100 GPUs, and CPUs from IBM, AMD and Intel.

The N8's **Bede** system is an IBM system that is similar in construction to the US Department of Energy Summit (#2) and Sierra (#3) systems. It consists of 32 nodes, each with two IBM Power9 CPUs and four NVIDIA V100 GPUs.

Besides these systems, York and Warwick also each have compute clusters for their own researchers. Each of these are predominantly homogeneous clusters with Intel CPUs, but both containing small GPU accelerated partitions.

Although the currently available UK systems are relatively small when compared to the European and US systems mentioned here, they are broadly representative of the hardware likely to be available at pre- and post-Exascale.

3.1.2 Europe

In Europe, PRACE (Partnership for Advanced Computing in Europe) provide access to a number of PetaFLOP-class HPC systems (Tier-0). The current Tier-0 systems are:

Marconi, a 30 PetaFLOP/s IBM Power9 and NVIDIA V100 system installed at CINECA;
Hawk, HLRS's 25 PetaFLOP/s homogeneous system using AMD Rome CPUs;
JUWELS, a 70 PetaFLOP/s system with AMD Rome CPUs and NVIDIA A100 GPUs installed at FZJ;
SuperMUC, a 26 PetaFLOP/s system installed at LRZ, using Intel Xeon Skylake CPUs.
Joliot-Curie, a 12 PetaFLOP/s homogeneous AMD Rome system at CEA;
Piz Daint, a Intel Xeon and NVIDIA P100 system, delivering 27 PetaFLOP/s at ETH Zurich;
and, **MareNostrum 4**, installed at BSC consisting of 4 separate systems: an Intel Xeon cluster, an IBM Power9 and NVIDIA V100 system, an AMD Rome and Radeon Instinct MI50 system, and an ARMv8 cluster.

In July 2019, the EuroHPC Joint Undertaking governing body selected 8 sites across the EU to host new HPC systems. Of these 8 sites, 3 will host pre-Exascale machines capable of at least 150 PetaFLOP/s.

The five Petascale systems are **Vega** in Slovenia, **MeluXina** in Luxembourg, **Discoverer** in Bulgaria, **Karolina** in the Czech Republic and **Deucalion** in Portugal.

The pre-Exascale systems are:

LUMI, which is installed in Kajaani, Finland, and is a **HPE/Cray Shasta** system comprising of AMD EPYC CPUs and AMD Radeon Instinct MI250X GPUs. Its GPU partition entered the Top500 in position #3 in June 2022 with an achieved peak of 151.9 PetaFLOP/s (from a theoretical peak of 214 PetaFLOP/s). Once fully installed and operational, it is expected that it will exceed 375 PetaFLOP/s on LINPACK.

LEONARDO, which is currently being installed at Cineca, Italy, and will be a Atos BullSequana system. It will be constructed of Intel Sapphire Rapids CPUs, coupled with 14,000 NVIDIA A100 GPUs, connected with Infiniband. It is expected to have a computational power of almost 250 PetaFLOP/s[7];

MareNostrum 5, which will be installed within the Barcelona Supercomputing Centre, and like its predecessor, will contain multiple partitions with alternative architectures. The largest partition will provide will provide 163 PetaFLOP/s and will be powered by Hopper GPUs and Sapphire Rapids CPUs. A second partition will be contain Intel's Rialto Bridge GPUs and Emeral Rapid CPUs. The final two partitions will be CPU partitions powered by Intel Sapphire Rapids CPUs and NVIDIA Grace Superchips, respectively[8].

[7] https://leonardo-supercomputer.cineca.eu/leonardo-first-components-arrived/
[8] https://www.hpcwire.com/2022/06/16/nvidia-intel-to-power-atos-built-marenostrum-5-supercomputer/

3.1.3 United States

In the United States, there is a long history of supercomputing within the Department of Energy. Currently the largest systems are both IBM Power9 and NVIDIA V100 systems installed at Lawrence Livermore National Laboratory and Oak Ridge National Laboratory, **Sierra** and **Summit**, respectively.

The first phase of the new **Perlmutter** system was recently installed at the Lawrence Berkeley National Laboratory, with 1,500 nodes, each with dual AMD Milan CPUs, coupled with four NVIDIA A100 GPUs. Perlmutter is an Cray Shasta system using AMD EPYC Milan CPUs and NVIDIA A100 GPUs. Its achieved performance of 70 PetaFLOP/s places it at #7 in the most recent Top500 list, with the second phase to be delivered at a later date, adding 3,000 more CPU-only nodes.

In preparation for Aurora, Argonne National Laboratory has recently installed the 24 PetaFLOP/s **Polaris** system. Like Perlmutter, Polaris consists of AMD CPUs alongside NVIDIA A100 GPUs, interconnected through Slingshot. The systems support for the MPI, OpenMP and SYCL programming models was a key factor in the commissioning of Polaris as a testbed for Aurora[9].

NVIDIA's new Eos system will replace the Selene system, also installed at Argonne National Laboratory[10]. Eos will be a DGX system based on Hopper H100 GPUs. Its 4608 GPUs will provide 275 PetaFLOP/s of double precission compute, with 18 ExaFLOP/s of AI compute.

3.1.4 World Wide

In 2020, the **Fugaku** system became the fastest supercomputer in the world with a theoretical peak double-precision performance in excess of half an ExaFLOP. The system consists of 160,000 Fujitsu A64FX CPUs and is connected with a 6-dimensional torus interconnect (Torus Fusion). In addition to topping the Top500, Fugaku also tops the Graph500, HPC-AI and HPCG lists – being the first supercomputer to achieve this feat.

Also of note, **Sunway TiahuLight** is a 93 PFLOP/s supercomputer powered by 41,000 Sunway SW26010 manycore processors. Each node is connected to 255 other nodes via PCIe Gen 3.0 to form a *supernode*; each supernode is connected via an infiniband interconnect [11].

3.2 Post-Exascale Systems

There are on going efforts towards Exascale happening around the world, and the first Exascale system appeared on the June 2022 Top500 list. There are a number of other systems in production that will also break the ExaFLOP threshold, and there are a small number of systems installed in China that likely exceed this mark but are not present on the Top500 list.

[9] `https://www.hpcwire.com/2021/08/25/argonnes-44-petaflops-polaris-system-will-be-testbed-for-aurora-exascale-era/`

[10] `https://www.hpcwire.com/2022/03/22/nvidia-announces-eos-supercomputer/`

3.2.1 The United Kingdom

The UK's Exascale strategy is currently focussed around the ExCALIBUR (Exascale Computing ALgorithms & Infrastructures Benefiting UK Research) project – a £46m project led by the Met Office and EPSRC.

Alongside the ExCALIBUR programme, it is UKRI's intention to deploy an Exascale supercomputer by 2025. To support this, the UK Government will be investing up to £1.2bn in new supercomputing infrastructure for the Met Office[11].

3.2.2 Europe

The first Exascale system in Europe is likely to be acquired and installed in the next two years (2022-2024). The EuroHPC Joint Undertaking issued a call for proposals for an Exascale-capable system in December 2021 (alongside a number of mid-sized systems). The large-scale system should be capable of achieving 1 ExaFLOP/s on LINPACK.

The deadline for the CFP was in February 2022. This report will be updated when more information is made available.

3.2.3 United States

The DoE are currently in the process of building and installing their first three Exascale systems, Aurora, Frontier and El Capitan. Each of them are heterogeneous systems, consisting of mixture of CPUs and GPUs.

Frontier was deployed at Oak Ridge National Laboratory in 2022 and has a peak performance in excess of 1.5 ExaFLOP/s. The system achieved 1.1 ExaFLOP/s on LINPACK, and did so using its AMD EPYC Trento CPUs with AMD Radeon Instinct MI250X GPUs.

Argonne's 1+ ExaFLOP/s **Aurora** system will follow in 2022/23 and will be constructed with Intel CPUs and GPUs – with each node being two Sapphire Rapids CPUs, with six Ponte Vecchio GPUs.

These systems will be followed in 2023 by **El Capitan** at LLNL, which is expected to exceed 2 ExaFLOP/s. Like Frontier, El Capital will consist of AMD hardware, with EPYC Genoa CPUs and a next generation Radeon Instinct architecture.

3.2.4 Worldwide

While there is a single Exascale machine listed in the Top500, there are at least two other machines capable of an ExaFLOP/s in double precision. These machines were not benchmarked for the most recent Top500

[11] `https://www.datacenterdynamics.com/en/news/uk-research-innovation-fund-exascale-supercomputer-software-algorithms-excalibur/`

list, but details of the machines was revealed in a number of paper submissions at the Supercomputing conference[12].

OceanLight is the successor to the TaihuLight system, installed in Qingdao, China. The system is reportedly capable of 1.05 ExaFLOP/s LINPACK performance, from a theoretical peak of approximately 1.3 ExaFLOP/s. Like TaihuLight, the system has been designed and manufactured by Sunway, based on the SW26010Pro CPUs. Each processor is capable of 14 TFLOP/s in double precision, and 55 TFLOP/s in half precision. According to the Gordon Bell Prize-winning research paper, the largest run was conducted on 107,520 SW26010Pro CPUs (when multiplied by 14 TFLOP/s, this suggests a possible peak of 1.5 ExaFLOP/s) [12].

The **Tianhe-3** system is reportedly capable of 1.3 ExaFLOP/s on LINPACK, out of an estimated 1.7 ExaFLOP/s. It is based on the Phytium 2000+ FTP Arm chip, coupled with a Matrix 2000+ MTP accelerator.

A third Exascale system is reportedly under construction at the National Supercomputing Center in Shenzen. The system is being developed by Sugon, and will be capable of 2 ExaFLOP/s. It was intended to be build using Sugon's Hygon CPUs (part of the AMD-Chinese joint venture), but due to restrictions imposed by the U.S. Government it is no longer clear what platform will ultimately be used.

3.3 Summary

The shift towards accelerated computing has made the task of efficiently programming these systems much more difficult. For homogeneous platforms, standard programming models (i.e. Fortran, C/C++, etc) along with well maintained compilers is sufficient for developing complex physics simulations. For accelerated platforms, hierarchical parallelism is usually exposed through a custom API and compiler developed specifically for the accelerator in use. For NVIDIA, this is the CUDA programming model; for AMD, this is HIP; and, for Intel, this will be SYCL/DPC++.

Although both AMD and Intel provide source-to-source translators that can take already developed CUDA code, and generate equivalent code for their accelerators, there are a number of efforts aimed at developing platform-agnostic applications from the outset. Whether applications developed using these platform-agnostic frameworks can be both *performant* and *portable* remains an open question.

[12] `https://www.hpcwire.com/2021/11/24/three-chinese-exascale-systems-detailed-at-sc21-two-operational-and-one-delayed/`

4 Possible Evaluation Platforms

A focus of this project is to provide an assessment of the options available when developing *performance portable* Exascale-capable software. This project will use existing data where possible to provide this analysis, but where data is missing or incomplete, we can use UK-based platforms to evaluate the performance portability of applications of interest to this work package.

Broadly speaking, we can divide UK-based HPC platforms into two categories, *homogeneous systems* and *heterogeneous systems*. The UK's only Tier-1 system, ARCHER2, is an homogeneous system with an estimated peak performance of 20 PFLOP/s. It is across the UK's Tier-2 systems that there is a significant degree of diversity, offering a wide range of homogeneous and heterogeneous platforms/partitions for evaluation where needed.

4.1 Homogeneous Systems

ARHCER2

The national supercomputer, ARCHER2, is installed at the Edinburgh Parallel Computing Centre (EPCC). ARCHER2 is a Cray Shasta system interconnected with Cray Slingshot fabric. It consists of 5,848 nodes, each with two AMD EPYC Rome CPUs.

Avon

Avon will be a homogeneous cluster of 180 nodes, containing dual Intel Xeon Cascade Lake CPUs installed at the University of Warwick (expected mid-2021). It will be interconnected with Infiniband.

Isambard

The Isambard Tier-2 service is predominantly composed of Marvell ThunderX2 ARM cores, connected by a Cray Aries interconnect. Beside the ThunderX2 cabinet, Isambard also contains a cabinet of Fujistu A64FX processors.

Viking

Viking is a large Linux compute cluster supporting research needs at the University of York. It consists of approximately 170 compute nodes, each with Intel Xeon Skylake CPUs, connected via Infiniband.

Cirrus

The Cirrus cluster, installed at EPCC, consists of 280 compute nodes, each with dual Intel Xeon Broadwell processors. The cluster is connected via Infiniband fabric.

4.2 Heterogeneous Systems

Viking

The Viking cluster, at the University of York, is further bolstered by two GPU nodes, providing a small heterogeneous compute capability. The two GPU nodes each contain four NVIDIA V100 GPUs.

Bede

The Bede system, installed at the University of Durham, has an architecture similar to that found on Summit and Sierra. Bede is a single cabinet of IBM POWER9 CPUs each supporting four NVIDIA V100 GPUs.

Isambard

Alongside the two ARM-based partitions on the Isambard system is the Multi-Architecture Comparison System (MACS). MACS contains four nodes each with two NVIDIA P100 GPUs, and four nodes each with an NVIDIA V100 GPU. It also contains four nodes with AMD EPYC Rome CPUs, and four nodes with Intel Xeon Cascade Lake CPUs. Finally, there are also eight Intel Xeon Phi nodes, and two IBM Power9 nodes with NVIDIA V100 GPUs.

CSD3

The Cambridge Service for Data Driven Discovery (CSD3) provide two supercomputers under EPSRC Tier-2. Peta4 is a system comprising predominantly of Intel Xeon Skylake CPUs, with a small number of Intel Xeon Phil nodes. Wilkes2 provides the largest GPU enabled system in the UK, comprising of 90 nodes each with four NVIDIA P100 GPUs.

Baskerville

The Baskerville system will be the University of Birmingham's Tier-2 cluster. There are 46 compute nodes, each with four NVIDIA A100 GPUs alongside Intel Xeon Ice Lake CPUs.

References

[1] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc. Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268, 1974.

[2] David Patterson. The trouble with multi-core. *IEEE Spectrum*, 47(7):28–32, 53, 2010.

[3] Thiruvengadam Vijayaraghavan et al. Design and analysis of an apu for exascale computing. In *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 85–96, 2017.

[4] Jack Dongarra, Steven Gottlieb, and William T. C. Kramer. Race to exascale. *Computing in Science and Engg.*, 21(1):4–5, January 2019.

[5] AMD. Introducing AMD CDNA Architecture. `https://www.amd.com/system/files/documents/amd-cdna-whitepaper.pdf` (accessed April 27, 2021), 2020.

[6] Simon McIntosh-Smith, James Price, Tom Deakin, and Andrei Poenaru. A performance analysis of the first generation of hpc-optimized arm processors. *Concurrency and Computation: Practice and Experience*, 31(16):e5110, 2019. e5110 cpe.5110.

[7] K. Kamalakkannan, Gihan R. Mudalige, Istvan Z. Reguly, and Suhaib A. Fahmy. High-level fpga accelerator design for structured-mesh-based explicit numerical solvers. In *35th IEEE International Parallel & Distributed Processing Symposium*. IEEE, May 2020.

[8] Tan Nguyen, Samuel Williams, Marco Siracusa, Colin MacLean, Douglas Doerfler, and Nicholas J. Wright. The performance and energy efficiency potential of fpgas in scientific computing. In *2020 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*, pages 8–19, 2020.

[9] H. Sutter. The Free Lunch is Over: A Fundamental Turn Toward Concurrency in Software. *Dr. Dobb's Journal*, 30(3):202–210, March 2005.

[10] W. Haensch, E. J. Nowak, R. H. Dennard, P. M. Solomon, A. Bryant, O. H. Dokumaci, A. Kumar, X. Wang, J. B. Johnson, and M. V. Fischetti. Silicon CMOS Devices Beyond Scaling. *IBM Journal of Research and Development*, 50(4.5):339–361, 2006.

[11] Jack Dongarra. Report on the Sunway TaihuLight System. Technical report, University of Tennessee, June 2016.

[12] Yong Liu et al. Closing the "Quantum Supremacy" Gap: Achieving Real-Time Simulation of a Random Quantum Circuit Using a New Sunway Supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, New York, NY, USA, 2021. Association for Computing Machinery.