T/AW088/22
Software Support Procurement

Report 2067270-TN-01-06

D1.1 – Hardware at the Exascale
Revision 8.0

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

University of York

Gihan Mudalige, Ben McMillan, and Tom Goffrey

University of Warwick

February 9, 2024

Contents

Glossary

AVX Advanced Vector eXtensions

DIMM Dual In-line Memory Module

DRAM Dynamic Random Access Memory

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

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

February 2024

- Refactored all data in boxes in Section 2, to use Tables to summarise relevant data.
- Updated data in Section 2 to use current process roadmap diagrams, and update product information where appropriate.
  - Updated information on next Intel Xe GPUs
  - Updated information on next AMD CPU (Bergamo)
  - Updated information on AMD MI300 GPUs
  - Corrected some minor previously erroneous data (e.g. PCI generation support etc.)
- Removed the comparison tables from the end of Section 2, since this was repeated information.
- Refactored Section 3, to use tables to list supercomputers, with their architectures, performance and rankings, where available.
- Updated information about UK Exascale System (following announcement that one would be hosted at EPCC).
- Updated information about EU Exascale System (JUPITER).
- Updated information regarding US Exascale Systems after Aurora's debut at SC23.
- Removed the Evaluation Platforms section at the end of this report, as it is no longer relevant at the end of the project.

September 2023

- Added information about Intel Granite Rapids and Sierra Forest.
- Added information about AVX10 and APX instruction sets.
- Updated information regarding AMD Genoa, Bergamo and Genoa-X CPUs.
- Updated information regarding AMD MI300 GPUs.
- Updated information regarding NVIDIA Grace Hopper CPU/GPUs.
- Updated information regarding the installation of Aurora (likely to be #1 in November).
- Added Viking 2 to platforms. The system should be online in October for use in evaluations.

March 2023

- Updated information about the Intel roadmap, including the cancellation of Rialto Bridge, and the repackaging of the Falcon Shores XPU (as a GPU).
- Added information about AMD Bergamo CPU and the AMD Instinct MI300 APU.
- Updated information regarding MareNostrum 5 (following cancellation of Rialto Bridge).
- Updated information regarding the MI300s in LLNL's El Capitan system.
- Added note regarding UK Exascale system following the 2023 March Budget.
- Added information regarding Isambard Phase 3.

November 2022

- Minor grammatical fixes.
- Removed mentions of tech no longer being developed (Optane)
- Added information on Graviton2/3 CPUs from AWS.
- Updated information in light of announcements at Supercomputing 2022.

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

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.

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.

Memory technologies such as High Bandwidth Memory (HBM) has produced "stacked memory" designs where embedded DRAM is integrated on to CPU chips. 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 the UK, Europe, and the United States.

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 (within an acceptable power envelope) currently 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.

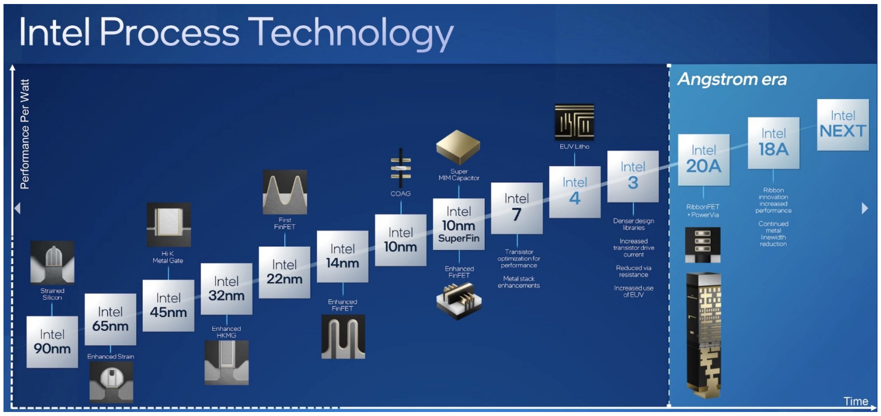


Figure 1: Intel's Process Roadmap

2.1.1 CPUs

The most widely used Intel Xeon CPUs currently are **Skylake**, **Cascade Lake**, and **Sapphire Rapids** (released in 2023). With Sapphire Rapids came a number of important architectural improvements over previous generation Xeon CPUs. Sapphire Rapids powers the Aurora supercomputer at Argonne National Laboratory (#2 in the November 2023 Top500).

Table 1: Key Features of Cascade Lake and Sapphire Rapids

	Cascade Lake	Sapphire Rapids
Node Technology	14 nm	Intel 7 (10 nm Enhanced SuperFin)
Configurations	up to 56 cores	up to 60 cores
Instruction Sets	AVX2, AVX-512	AVX2, AVX-512, AMX
Memory Support	6-channel DDR4	8-channel DDR5, HBM2e
Connectivity	PCIe gen 3	PCIe gen 5, Compute eXpress Link (CXL)

The next Intel CPU products will be **Emerald Rapids**, an upgrade on Sapphire Rapids, followed by **Granite Rapids** and **Sierra Forest**.

Following their switch towards "chiplet" designs (with Sapphire Rapids), from Granite Rapids and Sierra Forest onwards Intel CPUs will be based around either P-cores or E-cores (performance or efficiency, respectively). Granite Rapids will comprise of P-cores, targeted at High-Performance workloads, while Sierra

Forest will comprise of E-cores. Both of these architectures will have support for **Advanced Performance Extensions (APX)** – an extension of the x86 instruction set with more general purpose registers, meaning fewer load instructions (∼10% less) and fewer store instructions (∼20% less). Both will also have support for the new **AVX10** vector instruction set, bringing capabilities from AVX-512 to 256-bit registers (on E-cores).

Table 2: Key Features of Granite Rapids and Sierra Forest

	Granite Rapids	Sierra Forest
Launch	Expected 2024	Expected 2024
Node Technology	Intel 3 (5 nm)	Intel 3 (5 nm)
Configurations	up to 56 P-cores, up to 8 sockets	up to 144 E-cores, up to 2 sockets
Instruction Sets	AVX2, AVX-512, AVX10, AMX, APX	AVX2, AVX10, APX
Memory Support	12-channel DDR5	12-channel DDR5
Connectivity	PCIe gen 5, CXL 2.0	PCIe gen 5, CXL 2.0

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 is now supported by Intel's new **Xe Ponte Vecchio** discrete GPU. The successor to Ponte Vecchio was set to be **Rialto Bridge**, but following its cancellation in March 2023, it will now be followed by the **Falcon Shores** GPU in 2025. Falcon Shores was originally intended as an "XPU" with a CPU and GPU "on-package", but it will now be a discrete GPU[1].

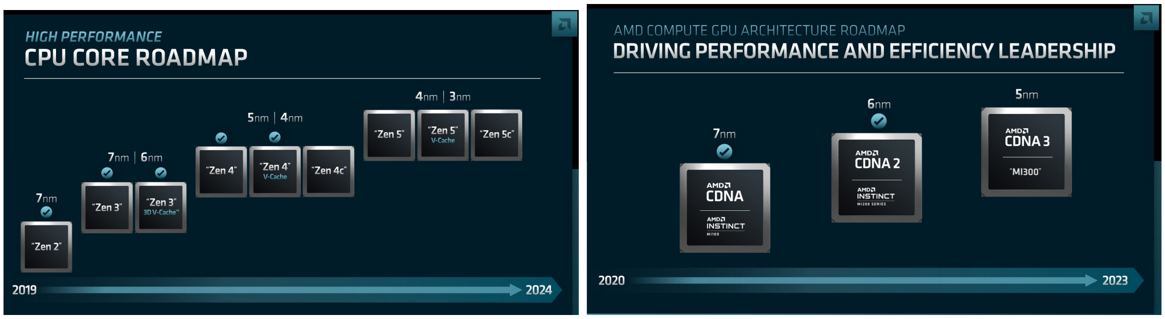
Table 3: Key Features of Intel Xe Product Line

	Ponte Vecchio	Falcon Shores
Launch	Released 2023	Expected 2025
Node Technology	7-10 nm	–
Configurations	56-128 Xe-cores	–
DP Performance	22.2-52.4 TFLOP/s	–
Memory Support	48-128 GB HBM2e	HBM3
Connectivity	PCIe gen 5, CXL	–

[1] https://www.anandtech.com/show/18756/intel-scraps-rialto-bridge-gpu-next-server-gpu-will-be-falcon-shores-in-2025

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.


HIGH PERFORMANCE
CPU CORE ROADMAP

AMD COMPUTE GPU ARCHITECTURE ROADMAP
DRIVING PERFORMANCE AND EFFICIENCY LEADERSHIP

Figure 2: AMD's CPU and GPU Process Roadmap

2.2.1 CPUs

AMD re-entered the server market 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**. Frontier is powered by a Milan-variant, codenamed **Trento**. The fourth iteration of the Zen architecture is used in the latest **Genoa** and **Bergamo** lines.

Table 4: Key Features of Rome, Milan, and Trento (Milan-variant)

	Rome	Milan (Trento)
Launch	Released 2019	Released 2021
Node Technology	7 nm, 14 nm I/O	7 nm, 14 nm I/O
Architecture	Zen 2	Zen 3
Configurations	up to 64 cores	up to 64 cores
Instruction Sets	AVX, AVX2	AVX, AVX2
Memory Support	8-channel DDR4	8-channel DDR4
Connectivity	PCIe gen 4	PCIe gen 4, InfinityFabric 3.0 (Trento)

Table 5: Key Features of Genoa and Bergamo

	Genoa	Bergamo
Launch	Released 2022	Released 2022
Node Technology	5 nm	5 nm
Architecture	Zen 4	Zen 4c
Configurations	up to 96 cores	up to 128 cores
Instruction Sets	AVX2, AVX-512	AVX2, AVX-512
Memory Support	12-channel DDR5 1.1 GB L3 (Genoa-X)	12-channel DDR5
Connectivity	PCIe gen 5, InfinityFabric 3.0	PCIe gen 5, InfinityFabric 3.0

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, and LUMI, installed at CSC in Finland. The will also be a key component of the El Capitan supercomputer, to be installed at the Lawrence Livermore National Laboratory.

The most recent MI300 series will also include AMDs first "APU" (accelerated processing unit), named the MI300A. In addition to its GPU compute units, it will integrate 24 Zen 4 EPYC cores on package, providing up to 122 TFLOP/s of double-precision performance[2].

Table 6: Key Features of Vega and Arcturus Product Lines

	MI50/MI60 (Vega)	MI100 (Arcturus)
Launch	Released 2018	Released 2020
Node Technology	7 nm	7 nm
Architecture	5th gen GCN	CDNA
Configurations	60/64 compute units	120 compute units
DP Performance	6.6/7.3 TFLOP/s	11.5 TFLOP/s
Memory Support	16/32 GB HBM2	64 GB HBM2
Connectivity	PCIe gen 4, InfinityFabric	PCIe gen 4, InfinityFabric

Table 7: Key Features of Aldebaran and Aqua Vanjaram Product Lines

	MI210/MI250/MI250X (Aldebaran)	MI300/MI300X (Aqua Vanjaram)
Launch	Released 2021/22	Released 2023
Node Technology	6 nm	5 nm
Architecture	CDNA 2.0	CDNA 3.0
Configurations	104/208/220 compute units	220/304 compute units
DP Performance	22.6-47.87 TFLOP/s	47.87-81.72 TFLOP/s
Memory Support	64/128 GB HBM2e	128/192 GB HBM3
Connectivity	PCIe gen 4, InfinityFabric	PCIe gen 5, InfinityFabric

[2] https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300a-data-sheet.pdf

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).

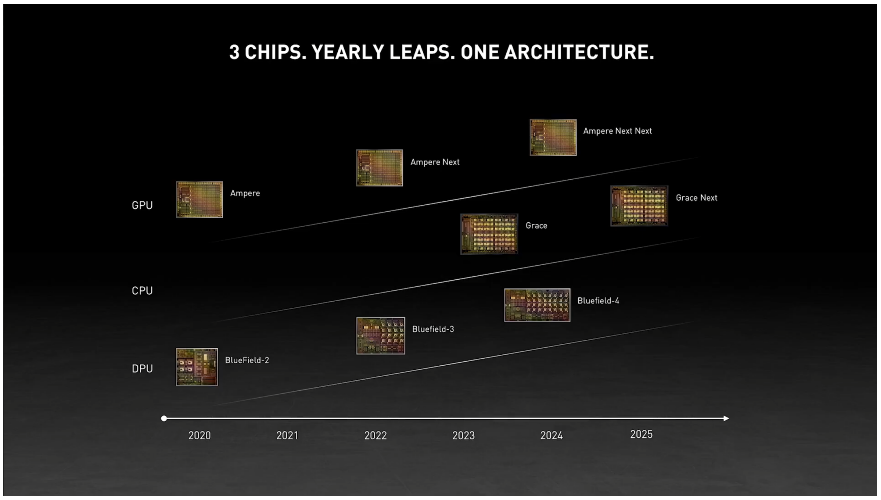

3 CHIPS. YEARLY LEAPS. ONE ARCHITECTURE.

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 **Ampere** and **Hopper**. Hopper H100 GPUs are in use in Microsoft's Eagle system (#3 in Top500), and the Flatiron Institute's Henri system (#1 in the Green500).

Table 8: Key Features of NVIDIA's Ampere and Hopper Product Lines

	A100 (Ampere)	H100 (Hopper)
Launch	Released 2021/22	Released 2023
Node Technology	7 nm	4 nm
Architecture	Compute Capability 8.0	Compute Capability 9.0
Configurations	108 SMs	132 SMs
DP Performance	9.75 TFLOP/s	25.6-31.04 TFLOP/s
Memory Support	40/80 GB HBM2e	80/96 GB HBM2e
Connectivity	PCIe gen 4, NVLink 3.0	PCIe gen 5, NVLink 4.0

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". The two components of each package are interconnected with NVLink-C2C, which NVIDIA claims is 7× faster than PCIe gen 5.

Table 9: Key Features of NVIDIA's Grace Superchips[3]

	Grace CPU Superchip	Grace Hopper Superchip
Launch	Released 2023	Released 2023
Node Technology	4 nm	4 nm
Architecture	Arm v9	Arm v9, Compute Capability 9.0
Configurations	144 cores per socket	72 cores per socket, 132 SMs
DP Performance	7.1 TFLOP/s	3.55 TFLOP/s (Grace) 34 TFLOP/s (Hopper)
Memory Support	up to 480 GB LPDDR5X	up to 480 GB LPDDR5X (Grace) 96-144 GB HBM3e (Hopper)
Connectivity	NVLink-C2C	NVLink-C2C

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

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 (previously Cavium), NVIDIA (discussed above), Fujitsu and Amazon (AWS) have been producing Arm-based CPUs for use in HPC systems and data centres.

Table 10: Key Features of Marvell and Fujitsu Arm CPUs

	Marvell ThunderX2	Fujitsu A64FX
Launch	Released 2018	Released 2019
Node Technology	16 nm	7 nm
Architecture	ARMv8.1-A	ARMv8.2-A 512-bit Scalable Vector Extensions
Configurations	32 cores per socket, 4-way SMT	48 cores per socket, plus assistant cores
Memory Support	up to 480 GB LPDDR5X	32 GB HBM2

Table 11: Key Features of AWS Graviton CPUs

	Graviton 2	Graviton 3
Launch	Released 2020	Released 2021
Node Technology	7 nm	7 nm
Architecture	Neoverse N1, ARMv8.2-A	Neoverse V1, ARMv8.4-A 256-bit SVE
Configurations	64 cores per socket	64 cores per socket
Memory Support	DDR4	DDR5

In addition to these architectures, the new JUPITER EU Exascale supercomputer will make use of **SiPearl Rhea1** CPUs. These CPUs will debut in 2024, using Neoverse V1 cores, providing 256-bit wide SVE, High-bandwidth memory in the form of HBM2e, and support for DDR5 and PCIe.

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.

2.5.1 IBM Power Architectures

All of the **IBM BlueGene** systems were driven by PowerPC architectures. In the case of Sierra and Summit, they were both powered by **Power9** CPUs, with NVIDIA V100 GPUs providing the majority of the computational performance. The Power9 CPU was manufactured on a 14 nm process, with up to 24 cores and 4-way SMT. Additionally, it included the NVLink interconnect, allowing direct GPU-GPU communication.

The **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, and it supports PCIe Gen 5, but has dropped support for NVLink.

2.5.2 Architectures Found Primarily in China

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.

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 contains 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 [6, 7].

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 [6, 7] 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.

Table 12: Key Features of AMD/Xilinx FPGAs

	Alveo U280	Versal VCK5000[4]
Launch	Released 2019	Released 2021
Node Technology	16 nm	7 nm
Memory Support	8 GB HBM2	16 GB DDR4
Connectivity	PCIe gen 4	PCI gen 4
Other Features	–	Two dual-core Cortex CPUs

Table 13: Key Features of Intel (formerly Altera) FPGAs

	Stratix 10	Agilex M-Series[5]
Launch	Released 2013	Expected 2024
Node Technology	13 nm	Intel 7 (10 nm)
Memory Support	8/16 GB HBM2	HBM2e
Connectivity	PCIe gen 4	CXL
Other Features	–	400 G Ethernet Network-on-Chip

2.7 Comparison and Summary

The end of the "free lunch" [8] and the breakdown of Dennard scaling [9] 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 in excess of 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.

[4] https://www.hpcwire.com/2022/03/08/amd-xilinx-takes-aim-at-nvidia-with-improved-vck5000-inferencing-card/

[5] https://www.hpcwire.com/off-the-wire/intel-introduces-agilex-m-series-fpgas/

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. 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.

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 these systems.

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 commercial systems in the UK are those found at research laboratories such as the Met Office, ECMWF, and AWE. Each of these systems are homogeneous clusters using Intel Xeon 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. Table 14 provides a summary of these systems.

Table 14: Summary of UK Commercial Systems

System	Architecture	Rmax (PFLOP/s)	Ranking
Met Office	Cray XC40, Intel Xeon CPU	7.04	#109
ECMWF	Cray XC40, Intel Xeon CPU	3.94	#193
ECMWF	Cray XC40, Intel Xeon CPU	3.94	#194
UK AWE (Damson)	Bull Sequana, Intel Xeon CPU	3.24	#239
Met Office	Cray XC40, Intel Xeon CPU	2.80	#307
Met Office	Cray XC40, Intel Xeon CPU	2.80	#308

In 2021, the Met Office announced that its next system will also be a homogeneous cluster, but will be based on AMD 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. Five of these systems appear on the Top500. The regional (Tier-2) centres typically host smaller systems that contain a wealth of architectural diversity. In particular, the **Isambard** system, installed at the University of Bristol, contains at least 9 different architectures for evaluation. Table 15 summarises some of these systems.

The upcoming **Isambard-3** system will be based on the NVIDIA Grace CPU Superchip, providing at least 55,000 cores, while the **Isambard-AI** system will be based around NVIDIA Grace Hopper Superchips.

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.

Table 15: Summary of some of the UK University Systems

System	Architecture	Rmax (PFLOP/s)	Ranking
EPCC (ARCHER2)	Cray XC40, AMD EPYC CPU	19.54	#39
Cambridge (Dawn)	Intel Xeon CPU, Intel Xe Max GPU	19.46	#41
STFC (DiRAC, Tursa)	AMD EPYC CPU, NVIDIA A100 GPU	5.23	#152
Cambridge (Wilkes-3)	AMD EPYC CPU, NVIDIA A100 GPU	2.29	#419
Cambridge (Cumulus)	Intel Xeon CPU, Intel Xeon Phi MIC	2.27	#431
Bristol (Isambard)	Marvel ThunderX2, Fujitsu A64FX, Intel Xeon CPU, Intel Xeon Phi MIC, NVIDIA P100 GPU, NVIDIA V100 GPU, NVIDIA A100 GPU, AMD MI100 GPU, IBM Power9	–	–
Durham (Bede)	IBM Power9, NVIDIA V100 GPU, NVIDIA Grace Hopper	–	–
York (Viking 2)	AMD EPYC CPU, NVIDIA A40 GPU, NVIDIA H100 GPU	–	–
Warwick (Avon)	Intel Xeon CPU	–	–

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 Tier-0 systems are listed in Table 16

Table 16: Summary of EU Tier-0 Systems

System	Architecture	Rmax (PFLOP/s)	Ranking
FZJ (JUWELS)	AMD EPYC CPU, NVIDIA A100 GPU	44.12	#18
Cineca (Marconi)	IBM Power9 CPU, NVIDIA V100 GPU	21.64	#35
LRZ (SuperMUC)	Intel Xeon CPU	19.48	#40
HLRS (Hawk)	AMD EPYC CPU	19.33	#42
BSC (MareNostrum 4)	Intel Xeon CPU	6.47	#121
	IBM Power9 CPU, NVIDIA V100 GPU	1.15	–
	AMD EPYC CPU, AMD MI50 GPU	–	–
	Fujitsu A64FX CPU	–	–
CEA (Joliot-Curie)	Cray XC40, Intel Xeon CPU	2.80	#307

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. Table 17 summarises these systems.

Table 17: Summary of EuroHPC JU Systems

System	Architecture	Rmax (PFLOP/s)	Ranking
Finland (LUMI)	AMD EPYC CPU, AMD MI250X GPU	379.70	#5
	AMD EPYC CPU	6.30	#125
Cineca (LEONARDO)	Intel Xeon CPU, NVIDIA A100 GPU	238.70	#6
	Intel Xeon CPU	7.84	#101
BSC (MareNostrum 5)[6]	Intel Xeon CPU, NVIDIA H100 GPU	138.20	#8
	Intel Xeon CPU	40.10	#19
Luxembourg (MeluXina)	AMD EPYC CPU, NVIDIA A100 GPU	10.52	#71
	AMD EPYC CPU	2.29	#421
Czechia (Karolina)	AMD EPYC CPU, NVIDIA A100 GPU	6.75	#113
	AMD EPYC CPU	2.84	#302
Bulgaria (Discoverer)	AMD EPYC CPU	4.52	#166
Slovenia (Vega)	AMD EPYC CPU	3.10	#198
	AMD EPYC CPU, NVIDIA A100 GPU	3.10	#268
Portugal (Deucalion)	Fujitsu A64FX	–	–

3.1.3 United States

In the United States, there is a long history of supercomputing within the Department of Energy. Currently their largest systems (excluding the ExaFLOP-capable Frontier system) are Aurora (discussed in Section 3.2), Summit, Sierra, and Perlmutter. Table 18 details the DoE systems within the Top 100, but excludes Frontier and Aurora.

Table 18: Summary of Department of Energy Systems in Top 100

System	Architecture	Rmax (PFLOP/s)	Ranking
ORNL (Summit)	IBM Power9 CPU, NVIDIA V100 GPU	148.60	#7
LLNL (Sierra)	IBM Power9 CPU, NVIDIA V100 GPU	94.64	#10
LBNL/NERSC (Perlmutter)	AMD EPYC CPY NVIDIA A100 GPU	79.23	#12
LANL/SNL (Crossroads)	Intel Xeon CPU Max	30.03	#24
ANL (Polaris)	AMD EPYC CPU, NVIDIA A100 GPU	25.81	#27
LANL/SNL (Trinity)	Intel Xeon CPU, Intel Xeon Phi MIC	20.16	#38
ORNL (Frontier TDS)	AMD EPYC CPU AMD MI250X GPU	19.20	#44
LLNL (Lassen)	IBM Power9 CPU, NVIDIA V100 GPU	18.20	#46
LBNL/NERSC (Cori)	Intel Xeon Phi MIC	14.01	#60

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*;

[6] Two further partitions will be added to the system, including an NVIDIA Grace CPU partition.

each supernode is connected via an infiniband interconnect [10].

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.

3.2.1 The United Kingdom

The UK's Exascale strategy is currently focused 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 £1.2 billion in new supercomputing infrastructure for the Met Office[7].

In the March 2023 Budget, the Chancellor of the Exchequer announced £3.5 billion to "make the UK a scientific and technological superpower"[8]. This included £900 million to invest in a new Exascale supercomputer.

In October 2023, it was announced that the University of Edinburgh's Parallel Computing Centre (EPCC) will host a new system Exascale system in a new wing of their Advanced Computing Facility[9].

3.2.2 Europe

The first Exascale system in Europe will be **JUPITER**, hosted at the Jülich Supercomputing Centre. It will be completed in 2024, and will consist of two parts. A Booster module, with its performance delivered by close to 24,000 NVIDIA Grace Hopper GH200 Superchips, and a Cluster module, with its performance delivered by new European ARM CPUs from SiPearl[10].

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

[8] https://www.gov.uk/government/news/government-commits-up-to-35-billion-to-future-of-tech-and-science

[9] https://www.epcc.ed.ac.uk/whats-happening/articles/edinburgh-lead-new-era-uk-supercomputing

[10] https://www.fz-juelich.de/en/news/archive/press-release/2023/with-jupiter-we-will-have-perhaps-the-most-powerful-ai-supercomputer-in-the-world

3.2.3 United States

The DoE have currently installed two of three planned Exascale systems, with the final system to be delivered within 2024. These machines are Aurora, Frontier, and El Capitan. Each of them are heterogeneous systems, consisting of mixture of CPUs and GPUs. Table 19 summarises these systems.

Table 19: Summary of Department of Energy Exascale Systems

System	Architecture	Rmax (PFLOP/s)	Ranking
ORNL (Frontier)	AMD EPYC CPU, AMD MI250X GPU	1,194.00	#1
ANL (Aurora)	Intel Xeon CPU, Intel Xe Max GPU	585.34	#2
LLNL (El Capitan)	*AMD MI300A APU*[11]	>2,000.00	–

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 have not been submitted to the Top500, but details of the machines were revealed in a number of paper submissions that were presented at the Supercomputing 2021 conference[12][13].

OceanLight is the successor to the TaihuLight system, installed in Qingdao, China. The system is reportedly capable of 1.2 ExaFLOP/s LINPACK performance, from a theoretical peak of approximately 1.5 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) [11].

The **Tianhe-3** system is reportedly capable of 1.5 ExaFLOP/s on LINPACK, out of an estimated 2.0 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.

[11] https://www.tomshardware.com/news/amd-instinct-mi300-apu-with-zen-4-and-cdna-3-up-and-running-in-the-lab

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

[13] As of Feb 2024, the performance of these machines has still not been publicly announced, but they are unofficially listed here: https://www.nextplatform.com/2023/11/13/top500-supercomputers-who-gets-the-most-out-of-peak-performance/.

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.

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] 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.

[7] 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.

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

[9] 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.

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

[11] 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.