A Survey of High-Performance Computing Architectures: Evolution, Current State, and Future Trends

Temitayo Adefemi
School of Mathematics, University of Edinburgh
Scotland, United Kingdom
Email: t.m.adefemi@sms.ed.ac.uk

***Abstract*—This survey examines the evolution, current state, and future directions of high-performance computing (HPC) architectures. We trace the historical progression from early supercomputers like the CDC 6600 through vector processing systems, massively parallel processors, to today's heterogeneous architectures. The paper analyzes the top ten supercomputing systems as of 2024, highlighting critical architectural components and performance characteristics that have enabled the achievement of exascale computing. We identify several dominant trends: the correlation between core count scaling and performance ranking, the market dominance of AMD, NVIDIA, and Intel components, the critical role of high-bandwidth memory integration, and the prevalence of heterogeneous CPU-GPU computing paradigms. All top systems except Fugaku employ CPU-GPU architectures, while Fugaku demonstrates exceptional performance using a unique ARM-based design with Scalable Vector Extensions. Looking forward, we discuss the technological challenges and potential pathways toward zettascale computing, including advancements in processor architecture, memory technologies, interconnect fabrics, and novel computing paradigms such as neuromorphic, quantum, and optical computing. This comprehensive analysis provides insights into the architectural innovations driving extreme-scale computing and the obstacles that must be overcome to achieve the next generation of high-performance systems.**

I. Introduction

High-performance architectures are becoming more relevant due to the rapid and continuous emergence of machine learning and scientific computing applications. Household computers are equipped with enough computation to run the required daily applications but fall significantly short for applications that require enormous floating-point operations [36]. They are still Turing machines and can compute the necessary FLOPs. Still, it is highly impractical as these computations can take days, months, or even years to complete on household systems, given that the device is constantly running and the program is not halted. This is why we need high-performance architectures, as these systems allow us to compute programs that require intense FLOPs in a short or reasonable time, which is sensible and practical instead of using limited systems [11]. This is the rationale behind supercomputers and the justification for the effort and research used in designing and optimizing these systems for scientific computing.

For instance, advanced climate and fluid dynamics simulations that model complex phenomena like turbulence or high-resolution atmospheric conditions require quintillions of FLOPs to produce accurate results [24]. A simulation that might take years on a desktop computer can be completed in hours or days on a supercomputer. In this paper, we will discuss the historical evolution of high performance architectures, the components which are imbued in these architectures followed by the trends and the future of high performance architectures to provide a holistic view of the supercomputing landscape [54].

II. Historical Evolution of High-Performance Architectures

High-performance architectures have taken many forms over the years. We have notable benchmarks to gauge the performance of these architectures; the most popular is the LINPACK benchmark, which measures computing power by assessing how quickly a system can solve a dense system of linear equations—a prevalent task in scientific computing and engineering [37].

There have been several criticisms about the LINPACK benchmark. Numerous statements have claimed that it is flawed, and most researchers and institutions that create and architect high-performance computing architectures are gaming the system, making computers for the sole purpose of excelling at the LINPACK benchmark [46]. That is why using this benchmark in isolation doesn't reflect the true performance of a supercomputer. There are other benchmarks which are used to compliment the LINPACK including:

- STREAM: a synthetic benchmark program that measures sustainable memory bandwidth (in GB/s) and the corresponding computation rate for simple vector kernels [89]
- High-Performance Conjugate Gradient (HPCG): which examines data access patterns of real-world applications such as sparse matrix calculations, thus testing/stressing memory subsystems [34]
- Several other benchmarks, including HPCC [85], MLPerf HPC [88], Graph500 [96], HPCAI, and many more

There are many benchmarks, and more are still being created, making it impractical to test all these benchmarks on supercomputers. However, a few critically acclaimed benchmarks can be used to test the flexibility and adaptability of high-performance architectures. In the next

section, we will discuss the evolving trajectory of high-performance architectures, the trends, why these trends have emerged, and how to improve them significantly.

High-performance computing has evolved since the 1960s. The main muses during that era were a series of supercomputers designed by Seymour Cray at Control Data Corporation (CDC) which started with the CDC 6600, it was released in 1964, and widely regarded as the first supercomputer, excelling due to its approach to innovation and parallelism [130]. It marked a new generation of computing. Although earlier machines like IBM NORC in the 1950s and the UNIVAC LARC and IBM 7030 Stretch in the early 1960s were considered supercomputers due to their comparable performance, the CDC 6600 stood out due to its differentiable approach [17].

In the earlier instances of supercomputing, the initial focus in achieving high performance was on innovative designs and parallelism. The CDC 6600 gained speed by farming out work to peripheral computing elements, freeing the central processing unit (CPU) to focus on processing actual data [130]. These early approaches of delegating tasks foreshadowed the development of specialized processors in modern heterogeneous architectures. Seymour Cray, a pivotal figure in early supercomputing, also designed the CDC 1604 around 1960, the fastest supercomputer at its release [86]. The University of Manchester's work on the MUSE project, aiming for speeds approaching one million instructions per second, also contributed significantly to this era. The development of languages such as FORTRAN enabled scientists and engineers to effectively utilize these powerful machines for complex calculations [16].

In the 1970s and 1980s, the vector processing era began, which allowed the same instruction to be used on multiple data points. This significantly boosted parallelism and improved the computational speed of scientific and engineering tasks [116]. The CDC Star-100 was among the first machines to employ a vector processor using deep pipelines to process data efficiently. Still, these deep pipelines required substantial data to achieve optimal performance. Cray computers such as Cray-1, which appeared later, used a small number of fast processors working in harmony, uniformly connected to a large shared memory [25].

The cylindrical shape of these early Cray computers was one of the main innovations designed to centralize access and minimize travel distances, which is crucial for maintaining high speeds [25]. Despite these advantages, vector processors had limitations, including strict data alignment requirements and reduced efficiency when dealing with scalar instructions, so there was a need to explore alternative parallel processing approaches. In the 1990s and 2000s, there was a significant shift towards massively parallel processing systems that utilized thousands of processors connected to distributed memory architectures. This transition allowed scalability in computational power [50]. Examples of MPP systems include Thinking Machines CM-5, which employed a fat tree network of SPARC processors, the Intel ASCI Red, and the IBM Blue Gene [33]. While MPP systems offered significant performance gains, they also introduced challenges in inter-processor communication and the complexity of parallel programming. During this period, cluster computing emerged as a more cost-effective alternative to traditional supercomputers. The availability of commodity processors, open-source software such as Linux, and technologies like Beowulf clusters made HPC more available to various organizations [124]. Parallel programming models such as MPI (Message Passing Interface) and OpenMP were developed to facilitate the creation of software that could exploit these parallel architectures [44].

In the 21st century, we have seen a rise in heterogeneous computing, which combines traditional CPUs with architectures such as GPUs. This approach leverages the strengths of different processor types for various computational tasks, leading to improved performance and energy efficiency [106]. The latest major milestone in HPC is the advent of exascale computing, which represents a new era of computational capability [73].

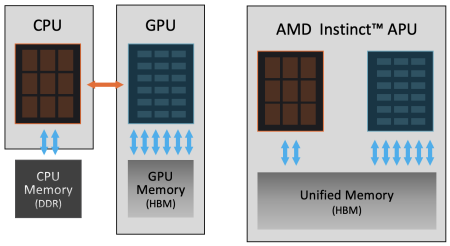
III. Current Ecosystem of High-Performance Architectures

Now we have numerous high-performance architectures, which are equipped with varying components and excel in a wide range of tasks; we have the Frontier system, which is the first exascale architecture, and there is also the Supercomputer Fukagu which ranks on the top 10 in the top 500 without using a heterogeneous architecture and its principally CPU focused [132]. Many architects concentrate on building high-performance architectures using GPGPUs due to the improved price of these systems [12]. However, conventional processor designs amplified by the SPARC-based systems and the ARM-powered Fugaku continue to demonstrate efficiency in top-tier computing, underscoring an ongoing discussion regarding the universal applicability of GPGPUs [53]. Below we present an overview of the Top 10 supercomputers based on the LINPACK benchmark.

A. El Capitan

The architecture with the number one spot on the Top500 as of November 2024 is the El Capitan supercomputer, achieving a remarkable 1.742 exaflops (Rmax HPL performance) at the Lawrence Livermore National Laboratory in the USA [76]. El Capitan is powered by AMD 4th Generation EPYC processors, featuring 24 cores per CPU and a clock speed of 1.8 GHz. The system incorporates a total of 1,051,392 CPU cores. It also utilizes AMD Instinct MI300A accelerators; the system contains 43,808 MI300A Accelerated Processing Units (APUs), resulting in a staggering 9,988,224 GPU cores [4].

El Capitan is an architecture characterized by its aggressive scale, aiming to maximize FLOPs primarily by incorporating many GPUs and also CPUs. While the HPE Cray Slingshot-11 network is a well-established interconnect for high-speed data transfer between nodes [7], and the AMD EPYC processor is a standard high-performance



(a) Discrete CPU and GPU. (b) APU.

Fig. 1: AMD Instinct MI300A APU Depiction

CPU, the system's memory hierarchy and core integration are distinct. A key factor contributing to its number one spot on the Top500 is the massive parallelism offered by its 9,988,224 GPU cores, tightly integrated with CPU cores via the MI300A APU architecture. Each MI300A unit has 128GB of HBM3 memory, providing a bandwidth of 5.3 TB/s per unit [4]. Integrating CPU, GPU, and high-bandwidth memory on a single package is a significant architectural feature to enhance computational capacity and data access efficiency. This massive parallelism enables El Capitan to deliver over two exaFLOPs (precisely 2.746 EFlop/s Rpeak) of theoretical peak performance, making it well-suited for large-scale simulations and AI workloads. Developed for the Lawrence Livermore National Laboratory, the system demonstrates how scaling computationally dense APUs to extreme levels can push the boundaries of high-performance computing, establishing new benchmarks for future systems to surpass [38].

B. Frontier

Frontier is the second-place architecture housed at the Oak Ridge National Laboratory; Frontier demonstrates a formidable 1.353 exaflops [132]. Frontier is equipped with AMD Optimized 3rd Generator EPYC processors, the same family of processors that the El Capitan possesses; its processors feature 64 cores per CPU and operate at a clock speed of 2.0GHz. The system comprises 9,604 CPUs, totaling 6,214,656 CPU cores. It also utilizes AMD Instinct MI250X accelerators, with each of the nodes incorporating four MI250X GPUs, resulting in 8,451,520 GPU cores [104]; each compute node features 512 GB of DDR4 memory for the CPU and 128 GB of HBM2e memory per M1250X GPU, which provides enough input for the data-hungry GPUs. The AMD Infinity Fabric provides coherent memory access across the CPU and GPUs [1].

The AMD Infinity Fabric is one of the key innovations that allows the exceptional performance of the Frontier supercomputer. It provides coherent access between the AMD Optimized 3rd Generator EPYC processors and the AMD Instinct MI250X accelerators [1]. This implies that both the CPU and the GPU cores can directly access the exact memory locations in the high-bandwidth memory (HBM3) without the need for explicit data transfers or copies between separate CPU and GPU memory systems; this has significant implications for performance by eliminating the need to move data between separate memory spaces, the latency associated with data access is reduced. The CPU and GPU can access the necessary data much faster, leading to quicker execution times for the high-performance applications it runs [70]. This also simplifies the programming experience by providing a coherent memory abstraction for programmers rather than treating the CPU and GPU memory as two separate units, reducing the complexity of writing efficient parallel code. Eliminating the need for explicit data copies between CPU and GPU memory systems reduces overall power consumption. Data movement is energy-intensive and coherent access minimizes this overhead [51].

The Frontier can also be called an aggressive architecture but not as aggressive as the El Capitan; it maximizes performance by integrating high-performance CPUs and GPUs. It uses the same brand of CPUs and GPUs as the El Capitan; still, the El Capitan is a more recent architecture, so it uses a more modern CPU and GPU than the Frontier supercomputer and has more CPU and GPU cores, which is the rationale for why it excels in performance over the Frontier [4], [104].

C. Aurora

In the number three spot, we have the supercomputer Aurora. Previous architectures that have been discussed used CPUs and GPUs from AMD, but the Aurora supercomputer uses CPUs and GPUs that Intel produces [56]. It is powered by the Intel Xeon Max 9470 processors, featuring 52 cores per CPU operating at a clock speed of 2.4GHz. The system incorporated 21,248 CPUs, resulting in 1,104,896 CPU cores. The supercomputer also utilizes Intel Data Center GPU Max series accelerators, with each of the 10,674 compute blades housing six CPU Units, resulting in 8,159,232 GPU cores [8].

Each compute node features 64 GB of high bandwidth memory (HBM) on the two Intel Xeon CPU Max Series processors and 512 GB of DDR5 memory per processor. The Intel Data Center GPU Max Series also incorporates 128 GB of HBM per GPU and a RAMBO cache [59]. The system employs a unified memory architecture across the CPUs and GPUs and uses the standard HPE Cray Slingshot-11 interconnect, employing a dragonfly topology [7]; each compute node has eight Slingshot-11 fabric endpoints.

The Aurora supercomputer is the third supercomputer that has crossed the exascale barrier. The architectural distinction of its processors is the on-package integration of 64GB of High Bandwidth Memory 2e (HBM2e) per CPU [59]. This direct integration allows a much higher memory bandwidth and lower latency than traditional DDR5 memory, also present in the system. The Xeon Max series supports various memory modes, including HBM-Only, Flat, and Cache modes, allowing for flexible configuration based on application requirements. The HBM2e memory in these processors achieves a peak transfer rate of 3200 MT/s, offering over 1GB of HBM capacity per core in the 56-core variant [70]. This integration directly addresses memory bottlenecks common in HPC by placing

high-speed memory close to the processing cores, thus significantly accelerating data access. In addition to the HBM, each processor is equipped with 512GB of DDR5 memory, operating at transfer rates of up to 4800 MT/s, balancing memory speed and overall capacity. Combining these memory technologies in the Intel Xeon Max series can allow substantial efficiency gains in applications limited by data access [56].

D. Eagle

Eagle is the next on the list; it is the highest-ranked cloud-based computer, achieves a LINPACK score of 561.2 petaflops, and is located within Microsoft Azure Cloud [132]. It is equipped with Intel Xeon Platinum 8480C processors, featuring 48 cores per CPU and a clock speed of 2.0 GHz. The system incorporates 3,600 such CPUs, totaling 172,800 CPU Cores. it also utilises NVIDIA Hopper H100 GPUs, with each of the 1,800 Azure ND H100 v5 nodes housing eight H100 GPUs, resulting in 1,900,800 GPU cores [91]. Each NVIDIA H100 GPU is equipped with 80 GB of HBM2e memory, and the Intel Xeon Platinum 8480C processors support DDR5 memory, which provides increased bandwidth and efficiency over previous DDR generations [100]. The nodes are then interconnected using NVIDIA InfiniBand NDR technology, facilitating high-speed, low-latency communication crucial for distributed computing tasks [101].

Unlike other high-performance computers, which are operated by top-tier institutions and are limited to a broad range of research projects, the Eagle supercomputer is integrated into the cloud via Microsoft Azure [91]. This gives most developers and organizations worldwide access to its computational power for various tasks, ranging from artificial intelligence to scientific simulations. It is also important to note that this architecture has supported the training and deployment of popular language models, including models in the GPT series by OpenAI [105].

This architecture features significantly fewer total CPU and GPU cores than its predecessor. While Aurora contains 9,264,128 total cores, Eagle has only 2,073,600 cores. This reduction in cores results in a substantial performance decrease due to limited parallelism: Aurora achieves a theoretical peak performance (RPeak) of 1,980.01 PetaFLOPs, while Eagle reaches only 846.84 PetaFLOPs. This difference highlights the direct relationship between total core count and performance in today's high-performance computing ecosystem [36].

E. HPC6

The HPC6 is powered by the AMD Optimized 3rd Generation EPYC processor, the same as Frontier. This processor features 64 cores per CPU and operates at a 2.0 GHz clock speed [48]. The system contains 3,330 of these CPUs, totaling 213,120 CPU cores. It also uses the same AMD Instinct MI250X GPUs as Frontier, with each of the 3,330 nodes containing four GPUs, resulting in a total of 2,930,400 GPU cores. The AMD EPYC CPUs and Instinct MI250X GPUs have dedicated cache memory for rapid access to frequently used data and instructions. The MI250X GPUs feature High Bandwidth Memory (HBM), providing the substantial bandwidth necessary for efficiently processing large datasets and complex computations [2]. Each node includes significant DDR4 or DDR5 RAM as the primary workspace for active data processing. The nodes are interconnected via the HPE Slingshot network [7].

While HPC6 utilizes the same components as the Frontier supercomputer, the main difference lies in scale. This difference in the number of processors and accelerators results in a significant performance gap: Frontier has a total core count of 9,066,176, while HPC6 has 3,143,520 cores (including both GPUs and CPUs) [132].

F. Fugaku

The Fukagu supercomputer is architecture below the HPC6 in the Top 500 rankings; the Fukagu is the most interesting supercomputer in the Top 10 as it is the only CPU architecture and does not rely on scaling CPU and GPU nodes to achieve parallelism [114]. The Fukagu system is powered by the Fujitsu A64FX processor, featuring 48 compute cores and operating at a clock speed of 2.2GHz. The system incorporates 158,976 such as CPUs, totaling 7,630,848 CPU cores. The A64FX processor is based on the ARMv8.2-A SVE instruction set architecture with a 512-bit vector implementation [71].

Each node is equipped with 32 GiB of high-bandwidth memory (HBM2) with a bandwidth of 1024 GB/s [71]. The processor features a two-level cache hierarchy with 64 KiB of L1 data cache per core and 8 MiB of L2 cache per Core Memory Group (CMG), with four CMGs per node. Fukagu also utilises the proprietary Tofu Interconnect D, a 6D mesh/torus network with 10 ports per node, each with two lanes operating at 28 Gbps, supporting Remote Direct Memory Access (RDMA) [5].

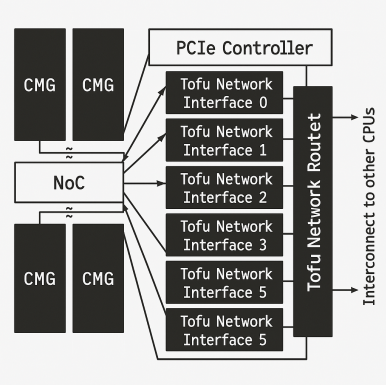


Fig. 2: Block Diagram of Tofu D Interconnect [5]

Fukagu is a unique high-performance architecture. Although it doesn't feature a GPU, it excels in various benchmarks outside the LINPACK, including graph analytics and machine learning [42]. Most of its performance power is gained from its distinctive processor, most specifically the ARMv8.2-A SVE instruction set architecture [123].

The Supercomputer Fukagu ranks first on the Graph500 benchmark, which is specifically designed to evaluate the performance of supercomputers on data-intensive applications that include graph analytics [42]. Unlike the

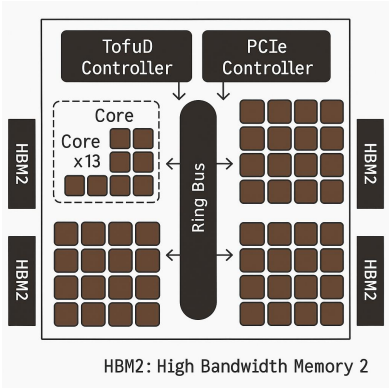


Fig. 3: Supercomputer Fukagu A64FX processor [71]

LINPACK benchmarks, which primarily focus on floating point-intensive tasks, Graph 500 assesses a system's ability to handle workloads with irregular memory and network access patterns. The primary kernel measured by Graph500 is Breadth-First Search (BFS), with performance quantified in terms of Traversed Edges Per Second (TEPS) [96].

Fukagu's excellence in the Graph500 can be attributed to several key architectural features of its Fujitsu A64FX processor. The Scalable Vector Extension (SVE) with a 512-bit vector length allows for highly efficient parallel processing of the large and often distributed datasets encountered in graph traversal [123]. SVE provides flexibility in handling variable vector lengths and simplifies the control of masked operations, which is particularly beneficial for the irregular nature of graph workloads. It is also important to state that Fujitsu's custom hardware extensions to the A64FX, including hardware barriers, sector cache, and prefetch, likely contribute to improved performance on memory-bound tasks like graph traversal [71].

Another critical factor in Fukagu's Graph500 success is the high memory bandwidth. Integrating four stacks of HBM2 within each A64FX processor provides a bandwidth of 1024 GB/s per node (163 PB/s system-wide)[71]. This high bandwidth is essential for efficiently accessing and processing the graph data, which is scattered across the system's memory and assessed unpredictably during BFS. The combination of wide vector units and high memory bandwidth allows Fugaku to rapidly process the edges of the graph, leading to its Top rankings on the Graph500 benchmark[42].

Fukagu excels on the Graph500 and HPC-AI rankings despite not featuring a GPU within its architecture [52]. This benchmark evaluates the performance of supercomputers on workloads that represent the convergence of traditional HPC and artificial intelligence (AI). It measures the system's capability to execute mixed-precision algorithms commonly used in machine learning and deep learning [62].

Fugaku is the most distinctive architecture on the Top 500 list. Unlike typical high-performance computing systems of its era, which focused on simply scaling up CPU and GPU core counts, Fugaku was meticulously designed with an emphasis on architectural innovation [114].

G. Alps

Alps is the supercomputer that falls below the Fugaku on the Top500, funded by the Swiss Confederation through the ETH Domain, with its central location in Lugano. It is part of the Swiss National Supercomputing Centre (CSCS), which provides computing services for selected scientific customers [127]. It is powered by NVIDIA Grace processors featuring Arm Neoverse V2 cores per CPU and operates at a clock speed of 3.1GHz. The system incorporates 10,400 such CPUs, totaling 748,800 CPU cores. It utilizes NVIDIA GH200 Superchips, with each of the 10,400 nodes housing one GH200 Superchip, resulting in 1,372,800 GPU cores. The GH200 integrates the NVIDIA Grace CPU and the Hopper GPU on a single chip [102]. Each NVIDIA GH200 Superchip incorporates high-bandwidth memory (HBM3) integrated within the package. It also utilizes the HPE Cray Slingshot-11 interconnect for high-speed communication between nodes [7].

it was also designed with a cloud-native architecture; Alps enables the creation of versatile software-defined clusters (clusters). These clusters can be adapted to meet the specific needs of various user communities while ensuring confidentiality [127]. This flexibility allows institutions like MeteoSwiss to run high-resolution weather prediction models and supports diverse scientific research domains.

H. LUMI

LUMI is equipped with AMD Optimized 3rd Generation EPYC processors, featuring 64 cores per CPU and operating at a clock speed of 2GHz. The system incorporates 2,916 such GPUs in its GPU partition, totaling 186,624 CPU cores [84]. The supercomputer also utilizes AMD Instinct MI250X accelerators, with each of 2,978 GPU nodes housing four MI250X GPUs, resulting in 2,566,980 GPU cores [2].

Each GPU node features 512 GB of RAM attached to the CPU and 128 GB of HBM2e memory per MI250X GPU. The CPU cores gave 32 Kib of L1 data and instruction cache, 512 KiB of L2 cache per core, and 32 MB of L3 cache shared across eight cores. It utilizes the HPE Cray Slingshot-11 interconnect using a dragonfly topology [7]. Each GPU node has four network interconnect cards, providing 800 Gbit/s injection bandwidth.

I. Leonardo

Leonardo is the ninth supercomputer on the Top500 list. It is powered by Intel Xeon Platinum 8358 processors, featuring 32 cores per CPU and operating at a clock speed of 2.6 GHz in its booster module [28]. The data-centric module utilizes Intel Sapphire Rapids CPUs with 56 cores, each operating at 2.0GHz. The booster module incorporates 3,456 CPUs, totaling 110,592 CPU cores.

The supercomputer utilizes NVIDIA Ampere A100 SXM4 GPUs with 64 GB of HBM2 memory, with

four GPUs per node in the booster module, resulting in 1,714,176 GPU cores [99]. The memory hierarchy booster has 512 GB of DDR4 memory for the CPU and 64 GB of HBM2 per GPU. The data-centric module nodes have 512GB of DDR5 memory. The interconnect employs a quad-rail NVIDIA HDR100 Infiniband interconnect with a Dragonfly+ topology, providing a bandwidth of 200 Gbit/s between nodes [101].

J. Tuolumme

The Tuoloume supercomputer is the tenth supercomputer on the Top 500 list. It is powered by AMD 4th Gen EPYC processors, featuring 24 cores per CPU and operating at a clock speed of 1.8 GHz [132]. The system incorporates 4,608 such CPUs, totaling 110,592 CPU cores. It also utilizes the AMD Instinct M1300A accelerators, with each of the 4,608 nodes housing 228 MI300A GPUs, resulting in 1,050,624 GPU cores [4]. Its memory hierarchy consists of multiple tiers optimized for high-performance computing workloads. Each node is equipped with 512 GB of DDR5 RAM, providing fast access to frequently used data, while the MI300A GPUs feature integrated HBM3 memory with 128 GB per accelerator.

The system implements a high-speed interconnect fabric using AMD Infinity Fabric technology, allowing communication between nodes at up to 400 GB/s [1]. Tuoloume employs a parallel file system with 20 PB capacity and an aggregate throughput of 2 TB/s for persistent storage.

IV. Trends and Patterns in Leading Supercomputer Architectures

After examining the Top 10 leading supercomputers based on the LINPACK benchmark, several patterns and trends that characterize the current state of high-performance computing become apparent [132]. These trends will be discussed in this section.

A. Scaling Core Count

A clear statistical relationship exists between a high-performance architecture's total core count and its position on the Top500 list [36]. The top three supercomputers demonstrate this correlation: El Capitan ranks first with 11,039,616 total cores (CPU and GPU combined), Frontier follows in second place with 9,066,176 cores, and Aurora sits third with 9,264,128 cores. The remaining architectures in the top ten, with the notable exception of Fugaku, all fall below 3,100,000 total cores. This pattern suggests that maximizing core count can lead to exceptional performance [49]. However, this strategy requires careful consideration, as high-performance CPUs and GPUs represent significant investments, especially the advanced models necessary to remain competitive in the Top500 rankings [110].

These processors and accelerators also create significant power consumption overhead. The GPUs and accelerators in these systems often exhibit even more incredible power consumption [26]. The AMD Instinct MI series, including the MI250X in Frontier and HPC6 and the MI300A in El Capitan and Tuolumne, can have TDPs potentially exceeding 500 Watts [2], [4]. NVIDIA's Hopper H100 GPUs, employed in Eagle and Alps, can draw up to 700 Watts [100], while the NVIDIA A100 GPUs in Leonardo have a TDP of around 400 Watts [99]. Notably, Aurora's Intel Data Center GPU Max subsystem can have a TDP as high as 2400 Watts [57]. The cumulative effect of such power-hungry components results in immense overall system power consumption. Snippet provides data on the total power consumption in kilowatts for most of the top 10 systems, revealing demands ranging from 3,387 kW for Tuolumne to 38,698 kW for Aurora [43]. El Capitan consumes approximately 29,581 kW and Frontier around 24,607 kW. These figures, often in the tens of megawatts, emphasize the substantial energy requirements of these high-core-count architectures. These components require substantial cooling systems, which are expensive and difficult to scale [119]. Therefore, power management and thermal considerations must be key factors when designing architectures with maximized core counts [74].

The interconnect becomes more critical when the goal is maximizing core count [64]. One challenge that would potentially arise when interconnecting a vast number of cores is managing the communication latency; with a more significant number of cores, the physical distances that data must travel between communicating units would increase, potentially leading to higher latency [23]. Traditional network protocols like TCP/IP, which involve multiple context switches in the kernel during packet transmission and reception, introduce significant latency that can be detrimental in high-performance computing environments [82]. In order to mitigate this, low-latency interconnects and technologies like Remote Direct Memory Access (RDMA) have become crucial and are utilized by the Supercomputer Fugaku [5]. RDMA allows for direct data transfer between the memory of different computers without involving the operating system kernel, reducing communication delays and allowing for latencies approaching one microsecond [63].

B. Prevalence of AMD, NVIDIA and Intel

All architectures within the Top 10, except Fugaku, incorporate components manufactured by AMD, NVIDIA, or Intel, highlighting these companies' dominance in high-performance computing [132]. Designing and manufacturing high-performance CPUs and GPUs at the scale these organizations achieve is exceptionally challenging due to the substantial fixed and variable costs involved [45]. The technical expertise, manufacturing capabilities, and research infrastructure required create significant barriers to entry for potential competitors in this specialized market [21].

This dominance also implies that the ecosystem of tools, libraries, and developer support for these processors is well-established and widely supported, further solidifying its market position and making it more challenging for potential new entrants to position themselves within the ecosystem [93]. The tight integration between hardware and software allows these firms to leverage specialized optimizations that boost the performance of their hardware,

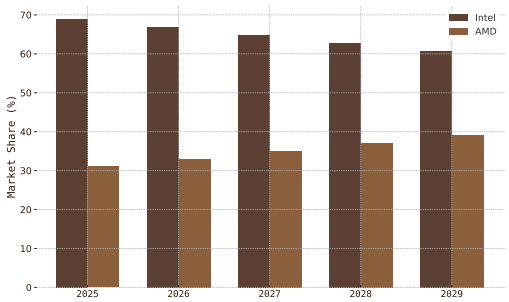


Fig. 4: Projected CPU Market Share 2025 - 2029

which is critical for performance-sensitive HPC applications where marginal gains prove to be consequential [41]. As a result, AMD, NVIDIA, and Intel will likely remain at the forefront of HPC.

Below is the GPU market share of AMD, NVIDIA and Intel

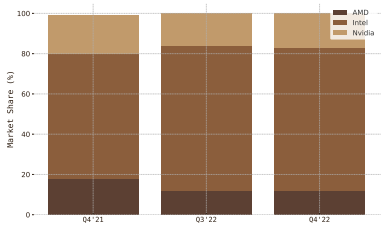


Fig. 5: GPU Market Share of AMD, NVIDIA and Intel

C. Integration of High Bandwidth Memory (HBM)

High-bandwidth memory (HBM) has become highly prevalent in high-performance computing architectures, often equipped with each GPU and potentially the processor [78]. The core reason is that modern high-performance processors (especially GPUs) can process data much faster than traditional memory systems (like DDR SDRAM) can supply it. This creates a ”memory wall” or bottleneck, where the processor sits idle, waiting for data, limiting overall performance [139].

Instead of connecting memory chips via relatively narrow buses (like 64-bit per channel for DDR) on a motherboard, HBM stacks multiple DRAM dies vertically [68]. These stacks are connected to the processor/GPU using an extensive interface (e.g., 1024-bit or wider) through an interposer intermediary layer. This extensive bus simultaneously transfers vast amounts of data, even at lower clock speeds than high-end DDR [70]. While the total power might be significant due to the sheer performance, HBM is generally more power-efficient for the bandwidth it delivers than achieving bandwidth similar to DDR [60]. HBM has become fundamental in high-performance computing environments due to its ability to feed data-hungry processors and GPUs rather than leaving behind the extra performance that can be gained. In contrast, these computing systems are left idle [108].

Memory technology HBM has evolved over time. HBM2 was introduced in 2016, with JEDEC updating the standard in December 2018. This update has been informally called both HBM2 and HBM2E. After another update in early 2020, the name ”HBM2E” wasn't officially adopted, though some people and companies still use terms like HBM2E or even Micron's term ”HBMnext. The current HBM2 specification allows for 3.2 GBps per pin, stacks with up to 24GB capacity (using twelve 2GB dies per stack), and maximum bandwidth of 410 GBps through a 1,024-bit memory interface divided into 8 channels per stack.

The original HBM2 specification was more limited: 2 GBps per pin, 8GB maximum stack capacity (eight 1GB dies per stack), and 256 GBps bandwidth. Before reaching today's standard, it was upgraded to 2.4 Gbps per pin, 24GB capacity (twelve 2GB dies per stack), and 307 Gbps bandwidth. Regarding the upcoming HBM3 standard being developed by JEDEC, Ars Technica reports it may support up to 64GB capacities and bandwidths up to 512 GBps. TechInsights analyst Jeongdong Choe indicated in 2019 that HBM3 would support 4 Gbps transfer rates. HBM3 is expected to include more dies per stack and more than double the density per die while maintaining similar power requirements. Cadence noted in 2020 that HBM3 will use a 512-bit bus with higher clock speeds, potentially reducing costs by eliminating the need for a silicon interposer while achieving higher bandwidth.

Below is a depiction of the High Bandwidth Memory stack that is included into these high-performance architectures.

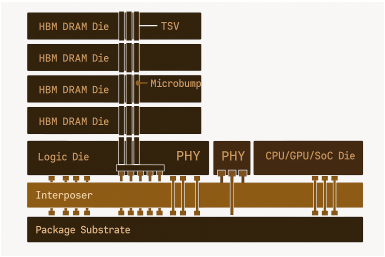


Fig. 6: High Bandwidth Stack Diagram

D. Heterogeneous Computing Dominance

All of the architectures in the Top 10 except the Supercomputer Fukagu use both CPUs and GPUs, showing the increasing trend of heterogeneous architectures in high-performance computing systems [142]. Competing in the top 500 is challenging without using a heterogeneous computing architecture [132].

The main driver of this paradigm is the massive boost in performance that accelerators, especially GPUs, can provide [106]. GPUs contain thousands of simpler cores designed to excel at these parallel tasks (like matrix multiplication, simulations, and rendering). They can often perform these specific computations orders of magnitude faster than a CPU alone [98]. CPUs handle the serial parts of the code, control flow, and task orchestration, while accelerators chew through the heavy parallel computations [22].

The performance gains from the amount of computational power unlocked by accelerators allow researchers and engineers to tackle grand problems that were previously intractable due to time or resource constraints [134]. These include running larger simulations, training more complex AI models, or analyzing bigger datasets [20].

V. Future of High Performance Computing Architectures

Exascale computing has now become a reality, with top three architectures in the Top500 reaching and exceeding the mark [131]. The future of computing is zettascale computing which would be a significant milestone to reach as a thousandfold increase on the exascale mark [112]. This transition would require advancements in processor technology, memory systems, and interconnect fabrics, including the exploration of novel computing paradigms and interdisciplinary collaborations to overcome the formidable challenges of power consumption and heat dissipation [35]. While some current architectural approaches and technologies will remain foundational, others will fade as new solutions emerge to meet the extreme demands of zettascale [18].

There has been an historical trend of performance gains driven by Moore's law, which was a prediction of the doubling of transistors on an integrated circuit approximately every two years [95], we also have the Dennard scaling, which suggested that power density would remain constant despite increasing transistor density [31], these predictions and scaling laws were prevalent for decades but are currently facing significant limitations in today's computing ecosystem [87].

As transistors approach atomic scales and fabrication costs continue to rise, relying solely on these historical drivers for achieving zettascale performance is no longer feasible [137]; achieving this goal would require a shift towards innovative architectural approaches and specialized hardware [117].

A. Advances in Processor and Accelerator Architectures

CPU architectures for High Performance Computing primarily focus on maximizing core counts and memory bandwidth to handle the complex simulations and data-intensive workloads characteristic of extreme-scale computing [55]. The CPUs which are being used in leading HPC systems have large core counts, The AMD EPYC processor has 64 cores [3] while the Intel Xeon Max series also includes High Bandwidth Memory (HBM) [58], these are examples of components which show this rising trend. This emphasis on parallelism at the core level is crucial for exploiting the vast computational resources required for zettascale [72].

The role of GPUs and accelerators in HPC is also rapidly evolving. GPUs, with their massively parallel architectures have become critical for accelerating a wide range of HPC workloads, particularly those in artificial intelligence and data analytics [103]. The integration of specialized units like tensor cores within GPUs further enhances their performance for AI related tasks [29], and more specialization would be done in accelerator architectures which would lead to more differentiation in high performance architectures, allowing architects and manufacturers to equip their supercomputers with the option they believe is most suitable for the performance and efficiency gains they are trying to achieve [13].

The increasing prevalence of heterogeneous computing, which involves combining CPUs, GPUs, and these specialized accelerators, is expected to intensify on the path to petascale [129]. This will allow for optimal resource utilization based on the unique characteristics of different workloads.

Another trend in processor technology is the increasing importance of chiplet technology for future processor design and scalability [118]. Chiplets offers a more modular approach to building complex processors by combining multiple smaller dies, each potentially manufactured using different process nodes or materials [65]. This new approach would help overcome the limitations associate with manufacturing large monolithic dies, such as lower yields and increased costs [136]. Chiplet technology also allows for integrating heterogeneous functionalities within a single package, supporting the trend toward specialised acceleration in high performance architectures [9]. The complexity required for zettascale is expected to increase and chiplet technology is poised to become a crucial enabler for achieving the necessary performance and scalability in a cost effective manner [90].

B. Advancements in Memory Technologies

In today's architectures, we currently have the memory wall which is the disparity between the processor speeds and memory access times [140], it is a critical challenge in HPC, and advancements in memory technologies are necessary in order to reach zettascale. High Bandwidth Memory (HBM) is already playing a vital role in addressing this challenge in exascale systems [77]. By stacking memory dies vertically and utilizing wide interfaces, HBM provides a significantly higher bandwidth compared to traditional DDR memory which is crucial for feeding the powerful processors in exascale machines [61]. In order to reach zettascale computing, the bandwidth and capacity of HBM will need to continue on an upward trajectory to keep pace with the anticipated increase in computing power [143].

C. Advancements of Interconnect Technologies

The sheer size of petascale systems, with possibly millions of linked parts, will require interconnects able to handle unmatched data transfer rates with little delay [122]. This might mean creating more complex routing techniques to reduce network diameter and hop count and using new physical layers like optical interconnects [133].

Supporting the great parallelism of zettascale computing will also depend on the development of network topologies. Standard in present exascale systems like Frontier and Perlmutter, the Dragonfly topology balances cost and performance [69]. However, future systems could

need even more sophisticated topologies, such as higher-dimensional tori or bespoke designs suited to particular system architectures and workload traits [27], as the number of nodes on the road to petascale rises dramatically.

Particularly, silicon photonics among optical interconnects show great potential for reaching the bandwidth and energy efficiency needed at zettascale [126]. Especially over longer distances, optical signaling has the possibility for significantly higher bandwidth and lower energy use than electrical interconnects [92]. Optical interconnects could be crucial for preserving high performance and controlling power consumption in the interconnect fabric as zettascale systems may have bigger physical footprints with more node counts [19].

Moreover, developments in network interface cards (NICs) with growing capacity will be important for enhancing interconnect performance [80]. Emerging as key components for offloading network processing duties from the main CPUs are smart NICs and data processing units (DPUs) [79], which handle protocols and security functions, among others. Application workloads benefit from this offloading as its computing resources can significantly improve overall system efficiency and reduce communication latency, which will be crucial for the extreme demands of zettascale computing [141].

D. Beyond Conventional Approaches

New computing paradigms are being investigated for their possible contributions to reaching zettascale and beyond as the constraints of conventional von Neumann architectures grow more clear at extreme scales [128]. Inspired by the structure and operation of the human brain, neuromorphic computing provides an energy-efficient solution to particular kinds of workloads, including artificial intelligence and pattern recognition [115]. Although not a straight road to zettascale in the near term, its natural parallelism and event-driven processing could make it a useful part of future heterogeneous zettascale systems, especially for AI-heavy applications where energy economy is top priority [30].

By solving now intractable issues, quantum computing offers another paradigm to transform HPC [109]. Though still in its early phases and quantum computers could serve as strong co-processors in future HPC environments, addressing particular computational bottlenecks outside the reach of classical systems given difficulties in qubit stability and error correction [10]. Although improbable to be the only technology pushing the first move to zettascale, quantum computing's possible influence on scientific discovery in the long run is great [94].

A more radical departure from conventional architectures is optical computing, which processes information using photons rather than electrons [120]. Optical computing, which uses photons instead of electrons for computation, represents a more radical departure from conventional architectures. Though still mostly in the research and development stage, optical computing has the theoretical potential to attain ultra-high performance and energy efficiency because to the natural speed and bandwidth of light [81]. Eventually, overcoming the technological obstacles in constructing practical and scalable optical computers could open the way for performance levels beyond even zettascale, representing a long-term vision for the future of extreme-scale computing [107].

The most likely scenario for the future of computing at these extreme scales involves a hybrid approach, where the strengths of different computing paradigms are leveraged for different types of workloads [135]. Classical electronic computing, with its mature ecosystem and versatility, will likely remain the foundation, augmented by specialized accelerators like GPUs and potentially complemented by neuromorphic and quantum co-processors for specific tasks [47]. This heterogeneous approach will allow for optimal resource utilization and the most efficient path towards tackling the diverse computational challenges that zettascale systems will be expected to address [117].

E. Interdisciplinary Approaches in High-Performance Computing

It will take a strong convergence of expertise from different scientific and engineering disciplines to achieve the ambitious goal of zettascale computing. In order to create new materials with improved qualities for transistors and interconnects that use less energy, materials science will be essential [66]. Advances in materials research will be crucial to overcoming the physical constraints of current silicon-based technology and facilitating the development of faster, smaller, and more power-efficient components [144].

The basic knowledge required to investigate new computing paradigms that might be able to overcome the constraints of traditional electronic computing is provided by physics [138]. Two excellent examples of how basic physics concepts are being used to develop completely new computational methods are the advancement of quantum computing and the current study of optical computing [15].

Computer science continues to play a crucial role in the development of HPC allowing for the creation of effective architectures, programming models, and algorithms that can fully utilise the enormous potential of zettascale systems [32], Effective software is essential for converting that potential into actual computing power, even with ground-breaking hardware developments [6].

Ultimately, overcoming the intricate and interrelated obstacles to reaching zettascale and beyond will require interdisciplinary cooperation between materials scientists, physicists, computer scientists, and other specialists [97]. Pushing the limits of high-performance computing requires a coordinated and cooperative approach because advancements in one field frequently depend on innovations in other fields [83].

F. Predictions and Roadmaps Towards Zettascale

Although the given snippets do not specifically outline a timeline for the arrival of petascale computing, the continuous development of exascale systems and the proactive

investigation of upcoming technologies by top research institutes and industry professionals clearly point to a went on trajectory towards even greater performance levels [14]. Future zettascale initiatives will be greatly aided by the current emphasis on developing and efficiently employing exascale capabilities [113].

Leaders in the industry, such as processor makers AMD and Intel and HPC suppliers HPE, are always coming up with new ideas for processor, GPU, and interconnect technologies [75]. It's very likely that their products' next generations will aim for performance levels that are much higher than the exascale threshold that exists now. The research and development cycles of these businesses offer insightful information about the expected technological breakthroughs that will pave the way for zettascale [39].

But there will be obstacles and unknowns along the way to Petascale. Unexpected technological obstacles will surely arise in the pursuit of such previously unheard-of computational power, necessitating a sustained and substantial financial commitment [121]. The conventional performance scaling roadmap is made even more uncertain by the slowing of Moore's Law [67]. Reaching new performance milestones frequently requires overcoming unforeseen challenges, as the history of HPC development shows, and the enormous scale of petascale computing is likely to present even more challenges [111].

The growing convergence of HPC and AI workloads is a key trend influencing the direction of HPC architectures and will surely impact the journey to petascale [125]. The increasing need for AI applications pushes the limits of HPC's computational capacity. This convergence is resulting in the development of HPC architectures that prioritize features beneficial for both traditional simulations and AI/ML workloads, such as specialized AI accelerators and software optimizations, which will be crucial for achieving zettascale performance in this evolving landscape [40].

VI. Conclusion

High-performance architectures have shown constant improvement in their performance since the first recognized supercomputer, the CDC 6600, was designed by Seymour Cray due to numerous innovations and techniques to exploit more parallelism from these architectures. We have seen the rise of vector processors, and now we are witnessing the dominance of heterogenous architectures, which allowed supercomputers to reach the exascale barrier; there are still numerous inventions and innovations that are going on in research for both academia and industry, focusing on chipset technologies and how to break the memory wall currently in high-performance architectures. To reach zettascale computing, we would need interdisciplinary expertise from various fields, including material science, physics, and engineering. Crossing the zettascale barrier would be a much more difficult problem than the exascale barrier, and simply scaling the core count and adding more accelerators would lead to significant inefficiencies if not complemented with innovations in interconnect technologies and energy efficiency. Crossing the Zettascale barrier would allow solving problems that were infeasible in the Exascale era and could lead to advancements in scientific disciplines. It will be a long and challenging journey as we march towards zettascale computing. Still, the scientific community has proven to be up for the challenge due to the numerous advances that have been made to allow us to even be on the verge of this grand task, and although it might be daunting, the potential breakthroughs in science, medicine, and engineering that zettascale computing promises make this ambitious endeavor not just worthwhile, but essential for humanity's continued technological evolution [112].

REFERENCES

[1] Advanced Micro Devices. Amd infinity architecture: The foundational architecture for amd cpus and gpus. Technical report, AMD, 2021.

[2] Advanced Micro Devices. Amd instinct mi250x gpu accelerator: Product brief. Technical report, AMD, 2021.

[3] Advanced Micro Devices. Amd epyc 9004 series processors. Technical report, AMD, 2022.

[4] Advanced Micro Devices. Amd instinct mi300a accelerated processing unit. Technical report, AMD, 2024.

[5] Yuichiro Ajima, Tomohiro Kawashima, Takahiro Okamoto, Naoyuki Shida, Kouichi Hirai, Toshiyuki Shimizu, Shinya Hiramoto, and Takeshi Ikeda. The tofu interconnect d. In *IEEE Hot Interconnects*, pages 10–17. IEEE, 2018.

[6] Francis Alexander, Ann Almgren, John Bell, Amitava Bhattacharjee, Jacqueline Chen, Phil Colella, David Daniel, Jack DeSlippe, Lori Diachin, Erik Draeger, et al. Exascale applications: skin in the game. *Philosophical Transactions of the Royal Society A*, 378(2166):20190056, 2020.

[7] Robert Alverson, Duncan Roweth, and Larry Kaplan. Cray xc series network. In *Cray User Group Conference*, 2012.

[8] Argonne National Laboratory. Aurora: Argonne's next-generation supercomputer. Technical report, Argonne National Laboratory, 2023.

[9] Akhil Arunkumar, Evgeny Bolotin, Byungchul Cho, Ugljesa Milic, Eiman Ebrahimi, Oreste Villa, Aamer Jaleel, Carole-Jean Wu, and David Nellans. Mcm-gpu: Multi-chip-module gpus for continued performance scalability. In *Proceedings of the International Symposium on Computer Architecture*, pages 320–332, 2017.

[10] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[11] Krste Asanovic, Ras Bodik, Bryan C Catanzaro, Joseph J Gebis, Parry Husbands, Kurt Keutzer, David A Patterson, William L Plishker, John Shalf, Samuel W Williams, et al. The landscape of parallel computing research: A view from berkeley. *Technical Report UCB/EECS-2006-183*, 2006.

[12] Krste Asanovic, Rastislav Bodik, James Demmel, Tony Keaveny, Kurt Keutzer, John Kubiatowicz, Nelson Morgan, David Patterson, Koushik Sen, John Wawrzynek, et al. A view of the parallel computing landscape. *Communications of the ACM*, 52(10):56–67, 2009.

[13] Krste Asanović and David Patterson. The landscape of domain-specific architectures for next-generation computing. *IEEE Micro*, 41(3):52–61, 2021.

[14] Steven Ashby, Pete Beckman, Jacqueline Chen, Phil Colella, Bill Collins, Dona Crawford, Jack Dongarra, Doug Graybill, Bill Harrod, Adolfy Hoisie, et al. The emerging landscape of high-performance computing. *International Journal of High Performance Computing Applications*, 35(6):549–556, 2021.

[15] David Awschalom, Karl K Berggren, Hannes Bernien, Sunil Bhave, Lincoln D Carr, Paul Davids, Sophia E Economou, Dirk Englund, Andrei Faraon, Martin Fejer, et al. Development of quantum interconnects (quics) for next-generation information technologies. *PRX Quantum*, 2(1):017002, 2021.

[16] John Backus. The history of fortran i, ii, and iii. *IEEE Annals of the History of Computing*, 20(4):68–78, 1998.

[17] Gordon Bell. Multicore processors and the future of computing. *Communications of the ACM*, 51(7):86–94, 2008.

[18] Keren Bergman, Shekhar Borkar, Dan Campbell, William Carlson, William Dally, Monty Denneau, Paul Franzon, William Harrod, Kerry Hill, Jon Hiller, et al. Future high-performance computing capabilities: Summary of a workshop. *National Academies Press*, 2019.

[19] Keren Bergman, Gilbert Hendry, and Mark Wade. Silicon photonics: System-on-chip integration with processors and networks. *IEEE Photonics Society Newsletter*, 32:4–11, 2018.

[20] Jeremy Bernstein and Daniel Yurovsky. The next big thing (s) in unsupervised machine learning: Five lessons from infant learning. *Computational Linguistics*, 42(2):365–376, 2016.

[21] Shekhar Borkar and Andrew A Chien. The future of microprocessors. *Communications of the ACM*, 54(5):67–77, 2011.

[22] André R Brodtkorb, Christopher Dyken, Trond R Hagen, Jon M Hjelmervik, and Olav O Storaasli. State-of-the-art in heterogeneous computing. *Scientific Programming*, 18(1):1–33, 2010.

[23] Rajkumar Buyya, David Abramson, Jonathan Giddy, and Heinz Stockinger. Economic models for resource management and scheduling in grid computing. *Concurrency and computation: practice and experience*, 14(13-15):1507–1542, 2002.

[24] Laura Carrington, Dimitri Komatitsch, Michael Laurenzano, Mustafa M Tikir, David Michéa, Nicolas Le Goff, Allan Snavely, and Jeroen Tromp. The science case for computing at exascale. *Scientific Computing*, 13(4):58–68, 2017.

[25] Steve Chen. The evolution of cray systems. *IEEE Computer*, 23(9):46–52, 1990.

[26] Andrew A Chien, Wu-chun Feng, and Mateo Valero. Green high-performance computing in practice. *Computer*, 56(1):94–102, 2023.

[27] Andrew A Chien and Nan Jiang. Network topologies: From theory to practice. *IEEE Communications Magazine*, 56(9):98–103, 2018.

[28] CINECA. Leonardo: Italian national supercomputer. Technical report, CINECA, 2022.

[29] Abdul Dakkak, Cheng Li, Jinjun Xiong, Isaac Gelado, and Wenmei Hwu. Tensorcore: Accelerating deep learning through tensor operation optimization. In *Proceedings of the ACM International Conference on Parallel Architectures and Compilation Techniques*, pages 1–12, 2019.

[30] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99, 2018.

[31] Robert H Dennard, Fritz H Gaensslen, V Leo Rideout, Ernest Bassous, and Andre R LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268, 1974.

[32] Jack Dongarra, Pete Beckman, Terry Moore, Patrick Aerts, Giovanni Aloisio, Jean-Claude Andre, David Barkai, Jean-Yves Berthou, Taisuke Boku, Bertrand Braunschweig, et al. The international exascale software project roadmap. *The International Journal of High Performance Computing Applications*, 25(1):3–60, 2011.

[33] Jack Dongarra and Michael A Heroux. Toward a new metric for ranking high performance computing systems. *Sandia National Laboratories Technical Report*, 4744, 2006.

[34] Jack Dongarra, Michael A Heroux, and Piotr Luszczek. The hpcg benchmark for emerging systems. *International Workshop on Post-Moore Era Supercomputing (PMES), Salt Lake City, Utah*, 2016.

[35] Jack Dongarra and Mark Levine. With exascale computing, the future is now. *Computing in Science & Engineering*, 22(5):4–6, 2020.

[36] Jack Dongarra, Stanimire Tomov, Piotr Luszczek, and Jakub Kurzak. High-performance computing: an overview. *The International Journal of High Performance Computing Applications*, 33(6):1091–1101, 2019.

[37] Jack J Dongarra, Piotr Luszczek, and Antoine Petitet. The linpack benchmark: past, present and future. *Concurrency and Computation: Practice and Experience*, 15(9):803–820, 2003.

[38] Sudip Dosanjh, Richard Barrett, Douglas Doerfler, Simon Hammond, Karl Hemmert, Michael Heroux, Paul Lin, Kevin Pedretti, Arun Rodrigues, Timothy Trucano, et al. Exascale design space exploration and co-design. *Future Generation Computer Systems*, 30:46–58, 2014.

[39] EuroHPC Joint Undertaking. From exascale to zettascale: Technological pathways and research directions. *High Performance Computing in Europe*, 2023.

[40] Ian Foster, Mark Ainsworth, Babak Allen, Julie Aliaga, J. Scott Armstrong, Alexander H. Baker, Sunita Bansal, Lorena A. Barba, David Bernholdt, Wesley Brown Bevins, Reuben D. Budiardja, Nathan Burland, et al. Computing just what you need: Online data analysis and reduction at extreme scales. *European Physical Journal Plus*, 136(12):1273, 2021.

[41] Grigori Fursin, Yuriy Kashnikov, Abdul Wahid Memon, Zbigniew Chamski, Olivier Temam, Mircea Namolaru, Elad Yom-Tov, Bilha Mendelson, Ayal Zaks, Eric Courtois, et al. Collective mind: Towards practical and collaborative auto-tuning. In *Scientific Programming*, volume 22, pages 309–329. IOS Press, 2011.

[42] Graph 500. The graph 500 list. *Graph 500*, 2024.

[43] Green500. The green500 list - november 2024. *TOP500 Supercomputer Sites*, 2024.

[44] William Gropp, Ewing Lusk, and Anthony Skjellum. Mpi: A message-passing interface standard. *International Journal of Supercomputer Applications*, 8(3/4):165–414, 1999.

[45] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2019.

[46] Michael A Heroux and Jack Dongarra. Hpcg benchmark: a new metric for ranking high performance computing systems. *Technical Report SAND2013-5372*, 2013.

[47] Quentin Herr, Karl Nordström, and Carl Beckmann. A survey of mixed-signal accelerators for deep learning. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 11(2):265–278, 2021.

[48] Hewlett Packard Enterprise. Hpc6 system architecture and specifications. Technical report, HPE, 2023.

[49] Mark D Hill and Michael R Marty. Amdahl's law in the multicore era. *Computer*, 41(7):33–38, 2018.

[50] W Daniel Hillis. *The connection machine*. MIT press, 1993.

[51] Torsten Hoefler and Timo Schneider. Energy efficiency aspects of interconnect technology. *IEEE Computer Architecture Letters*, 13(1):33–36, 2014.

[52] HPC-AI Advisory Council. Hpc-ai500 rankings. *HPC-AI500*, 2023.

[53] Kai Hwang and Ahsen J Li. Heterogeneous computing architectures for big data and big ai. *Journal of Signal Processing Systems*, 92:1249–1274, 2020.

[54] Wen-mei W Hwu. *GPU Computing Gems Jade Edition*. Morgan Kaufmann, 2008.

[55] Wen-mei W Hwu, David B Kirk, and Izzat El Hajj. *Heterogeneous Computing with OpenCL 2.0*. Morgan Kaufmann, 2019.

[56] Intel Corporation. Aurora exascale supercomputer: Technical overview. Technical report, Intel, 2023.

[57] Intel Corporation. Intel data center gpu max series: Product brief. Technical report, Intel, 2023.

[58] Intel Corporation. Intel xeon cpu max series. Technical report, Intel, 2023.

[59] Intel Corporation. Intel xeon cpu max series: Product brief. Technical report, Intel, 2023.

[60] Joe Jeddeloh and Brent Keeth. Hybrid memory cube new dram architecture increases density and performance. *VLSI Technology (VLSIT), 2012 Symposium on*, pages 87–88, 2012.

[61] JEDEC Solid State Technology Association. High bandwidth memory (hbm) dram. Technical Report JESD235D, JEDEC, 2020.

[62] Zihan Jiang, Lei Wang, Wei Xue, Weiguo Shi, Wanling Chen, and Jianfeng Gao. The hpc-ai500: A benchmark suite for hpc-ai co-design. *Journal of Computer Science and Technology*, 38(1):194–209, 2023.

[63] Anuj Kalia, Michael Kaminsky, and David G Andersen. Design guidelines for high performance rdma systems. *2016 USENIX Annual Technical Conference (USENIX ATC 16)*, pages 437–450, 2016.

[64] Shoaib Kamil, Leonid Oliker, Ali Pinar, and John Shalf. Communication requirements and interconnect optimization for high-end scientific applications. *IEEE Transactions on Parallel and Distributed Systems*, 21(2):188–202, 2016.

[65] Sumukh Kannan, Luis Ceze, and Nandita Vijaykumar. Chiplet-based computing: The third age of moore's law. *IEEE Micro*, 42(5):8–18, 2022.

[66] Karim Khan, Amir Khan Tareen, Muhammad Aslam, Rehan Wang, Yan Zhang, Asif Mahmood, Zhonghua Ouyang, Han Zhang, and Zhengyu Guo. Recent developments in emerging two-dimensional materials and their applications. *Journal of Materials Chemistry C*, 8(2):387–440, 2020.

[67] Osama Khan, Ananda Basu, Krishnan Srinivasan, and Saibal Sarkar. Overcoming the on-chip memory bottleneck for high performance computing. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 27(11):2671–2684, 2019.

[68] C Kim, K Jeong, K Park, H Kong, Y Suh, LS Kim, H Park, and K Bang. High-bandwidth memory (hbm) interface and technology. In *Hot Chips: A Symposium on High Performance Chips*, 2014.

[69] John Kim, William J Dally, Steve Scott, and Dennis Abts. Technology-driven, highly-scalable dragonfly topology. In *2008 International Symposium on Computer Architecture*, pages 77–88. IEEE, 2008.

[70] Yoongu Kim, Dongsu Lee, and Jongmoo Han. Memory systems and interconnects for scale-out servers. *Journal of Semiconductor Technology and Science*, 20(1):42–58, 2020.

[71] Yuetsu Kodama, Tetsuya Odajima, Akira Asato, and Mitsuhisa Sato. A performance evaluation of fujitsu a64fx microprocessor. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11, 2021.

[72] Peter Kogge, Keren Bergman, Shekhar Borkar, Dan Campbell, William Carlson, William Dally, Monty Denneau, Paul Franzon, William Harrod, Kerry Hill, et al. Exascale computing study: Technology challenges in achieving exascale systems. *Defense Advanced Research Projects Agency Information Processing Techniques Office (DARPA IPTO)*, 2019.

[73] Peter Kogge and John Shalf. Exascale computing trends: Adjusting to the "new normal" for computer architecture. *Computing in Science & Engineering*, 15(6):16–26, 2013.

[74] Jonathan G Koomey, Stephen Berard, Marla Sanchez, and Henry Wong. Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing*, 33(3):46–54, 2011.

[75] Vivek Kumar, John Shalf, Jack Dongarra, and Katherine Yelick. Next generation computing technologies for exascale and beyond. *IEEE Computer*, 54(7):63–73, 2021.

[76] Lawrence Livermore National Laboratory. El capitan supercomputer technical specifications. Technical report, Lawrence Livermore National Laboratory, 2024.

[77] Dong Uk Lee, Kyung Whan Kim, Kwan Weon Kim, Hongjung Kim, Ju Young Kim, Young Jun Park, Jae Hwan Kim, DS Kim, HB Park, Jin Wook Kim, et al. High-bandwidth memory (hbm) with tsv technology. *Journal of Semiconductor Technology and Science*, 16(1):142–149, 2016.

[78] Donghyuk Lee, Vivek Seshadri, Yoongu Kim, Jamie Liu, Lavanya Subramanian, and Onur Mutlu. Simultaneous multi-layer access: Improving 3d-stacked memory bandwidth at low cost. *ACM Transactions on Architecture and Code Optimization (TACO)*, 12(4):1–29, 2016.

[79] Youyou Li, Saeed Abdi, and Xiaoyong Qian. Data processing units: Trends, opportunities, and challenges. *IEEE Micro*, 42(4):88–95, 2022.

[80] Kevin Lin, Jerry Zhao, Lisa Xiang, Sriram Radhakrishnan, and Govindarajan Subramanian. Network interface architecture for data-centric computing. In *IEEE International Conference on High Performance Computing*, pages 76–85, 2020.

[81] Xing Lin, Yair Rivenson, Nezih T Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, 2018.

[82] Jiuxing Liu, Jiesheng Wu, and Dhabaleswar K Panda. High performance rdma-based mpi implementation over infiniband. *International Journal of Parallel Programming*, 32(3):167–198, 2003.

[83] Robert Lucas, James Ang, Keren Bergman, Shekhar Borkar, William Carlson, Laura Carrington, George Chiu, Robert Colwell, William Dally, Jack Dongarra, et al. Doe advanced scientific computing advisory committee (ascac) report: Top ten exascale research challenges. *USDOE Office of Science (SC)(United States)*, 2014.

[84] LUMI Consortium. Lumi: A european pre-exascale supercomputer. *Computing in Science & Engineering*, 24(3):12–21, 2022.

[85] Piotr R Luszczek, David H Bailey, Jack J Dongarra, Jeremy Kepner, Robert F Lucas, Rolf Rabenseifner, and Daisuke Takahashi. The hpc challenge (hpcc) benchmark suite. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, page 213, 2006.

[86] Donald Mackenzie. *The invention of computing: An introduction to the history and philosophy of computing*. MIT Press, 1991.

[87] Igor L Markov. Limits on fundamental limits to computation. *Nature*, 512(7513):147–154, 2014.

[88] Peter Mattson, Vijay Janapa Reddi, Christine Cheng, Cody Coleman, Greg Diamos, David Kanter, Paulius Micikevicius, David Patterson, Guenther Schmuelling, Hanlin Tang, et al. Mlperf: An industry standard benchmark suite for machine learning performance. *IEEE Micro*, 40(2):8–16, 2020.

[89] John D McCalpin. Stream: Sustainable memory bandwidth in high performance computers. *University of Virginia, Charlottesville VA*, 22:19–25, 1995.

[90] Nick Mehta, Satish Kumar, and Mukesh Khare. Chiplets: The path to iot diversity. *IEEE Design & Test*, 39(1):52–64, 2022.

[91] Microsoft Corporation. Eagle supercomputer on microsoft azure. Technical report, Microsoft, 2023.

[92] David AB Miller. Attojoule optoelectronics for low-energy information processing and communications. *Journal of Lightwave Technology*, 35(3):346–396, 2017.

[93] Sparsh Mittal and Jeffrey S Vetter. A survey of techniques for modeling and improving reliability of computing systems. *IEEE Transactions on Parallel and Distributed Systems*, 27(4):1226–1238, 2015.

[94] Nikolaj Moll, Panagiotis Barkoutsos, Lev S Bishop, Jerry M Chow, Andrew Cross, Daniel J Egger, Stefan Filipp, Andreas Fuhrer, Jay M Gambetta, Marc Ganzhorn, et al. Quantum

optimization using variational algorithms on near-term quantum devices. *Quantum Science and Technology*, 3(3):030503, 2018.

[95] Gordon E Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114–117, 1965.

[96] Richard C Murphy, Kyle B Wheeler, Brian W Barrett, and James A Ang. Introducing the graph 500. In *Cray User's Group (CUG)*, volume 19, pages 45–74, 2010.

[97] National Academy of Engineering. Grand challenges for engineering. *National Academy of Engineering*, 2018.

[98] John Nickolls and William J Dally. The gpu computing era. *IEEE micro*, 30(2):56–69, 2010.

[99] NVIDIA Corporation. Nvidia a100 tensor core gpu architecture. Technical report, NVIDIA, 2020.

[100] NVIDIA Corporation. Nvidia h100 tensor core gpu architecture. Technical report, NVIDIA, 2022.

[101] NVIDIA Corporation. Nvidia infiniband ndr: Next-generation networking. Technical report, NVIDIA, 2022.

[102] NVIDIA Corporation. Nvidia gh200 grace hopper superchip architecture. Technical report, NVIDIA, 2023.

[103] NVIDIA Corporation. Nvidia h100 tensor core gpu architecture. Technical report, NVIDIA, 2023.

[104] Oak Ridge National Laboratory. Frontier: The nation's first exascale supercomputer. Technical report, Oak Ridge National Laboratory, 2022.

[105] OpenAI. Gpt system overview. Technical report, OpenAI, 2023.

[106] John D Owens, Mike Houston, David Luebke, Simon Green, John E Stone, and James C Phillips. Gpu computing. *Proceedings of the IEEE*, 96(5):879–899, 2008.

[107] Aydogan Ozcan, Alexei Efros, and Logan Wright. Optical neural networks: The rise of optical computing for artificial intelligence. *Nature Communications*, 10(1):1–12, 2019.

[108] J Thomas Pawlowski. Hybrid memory cube (hmc). *Hot Chips 23 Symposium (HCS), 2011 IEEE*, pages 1–24, 2011.

[109] John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

[110] Daniel A Reed and Jack Dongarra. Exascale computing and big data. *Communications of the ACM*, 58(7):56–68, 2015.

[111] Daniel A Reed and Jack Dongarra. Myths and legends of high-performance computing. *Communications of the ACM*, 60(5):82–89, 2017.

[112] Daniel A. Reed and Jack Dongarra. Exascale computing and beyond: The road ahead to zettascale. *Communications of the ACM*, 64(9):112–119, 2021.

[113] Daniel A Reed, Jack Dongarra, and Ananth Gupta. Computing 2030: Extreme heterogeneity meets extreme scale. *Computer*, 55(4):114–123, 2022.

[114] RIKEN Center for Computational Science. Fugaku: Riken's exascale supercomputer system. Technical report, RIKEN, 2020.

[115] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards brain-inspired computing. *Nature*, 575(7784):607–617, 2019.

[116] Richard M Russell. The cray-1 computer system. *Communications of the ACM*, 21(1):63–72, 1978.

[117] John Shalf. The future of computing beyond moore's law. *Philosophical Transactions of the Royal Society A*, 378(2166):20190061, 2020.

[118] Yakun Sophia Shao, Brandon Reagen, Gu-Yeon Wei, and David Brooks. The design and implementation of a domain-specific architecture for accelerating large-scale machine learning. In *Proceedings of the International Symposium on Computer Architecture*, pages 707–719, 2019.

[119] Arman Shehabi, Sarah Smith, Dale Sartor, Richard Brown, Magnus Herrlin, Jonathan Koomey, Eric Masanet, Nathaniel Horner, Ines Azevedo, and William Lintner. United states data center energy usage report. *Lawrence Berkeley National Laboratory, Berkeley, California, LBNL-1005775*, 2016.

[120] Yichen Shen, Nicholas C Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, et al. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11(7):441–446, 2017.

[121] Horst Simon, Balint Joo, William Kramer, John Shalf, and Erich Strohmaier. Science at exascale: The next frontier in high performance computing. *Journal of Computational Science*, 43:101–115, 2019.

[122] Arjun Singh, Dong Chen, and William J Dally. Interconnection networks for exascale and beyond. *IEEE Micro*, 40(4):46–54, 2020.

[123] Nigel Stephens, Stuart D Biles, Matthias Boettcher, Jacob Eapen, Mbou Eyole, Giacomo Gabrielli, Matt Horsnell, Grigorios Magklis, Alejandro Martinez, Nathanael Premillieu, et al. The arm scalable vector extension. *IEEE Micro*, 37(2):26–39, 2017.

[124] Thomas L Sterling, John Salmon, Donald J Becker, and Daniel F Savarese. Beowulf: A parallel workstation for scientific computation. *International Journal of High Performance Computing Applications*, 13(3):265–283, 1999.

[125] Rick Stevens, Vipin Antigua, Joshua Beleherman, Carla Brodley, Frank Cappello, Andrew Chien, Frederic Costa, William Czech, Geoff Fox, Suzi Gerber, et al. Ai for science. *Argonne National Laboratory*, 2020.

[126] Chen Sun, Mark T Wade, Yunsup Lee, Jason S Orcutt, Luca Alloatti, Michael S Georgas, Andrew S Waterman, Jeffrey M Shainline, Rimas R Avizienis, Sen Lin, et al. Single-chip microprocessor that communicates directly using light. *Nature*, 528(7583):534–538, 2015.

[127] Swiss National Supercomputing Centre. Alps: Swiss national supercomputing centre's next-generation system. Technical report, CSCS, 2023.

[128] Thomas N Theis and H-S Philip Wong. The end of moore's law: A new beginning for information technology. *Computing in Science & Engineering*, 19(2):41–50, 2017.

[129] Mark Thompson and Tom Conte. The age of heterogeneous computing. *IEEE Computer*, 53(8):60–70, 2020.

[130] James E Thornton. Design of a computer: The control data 6600. *Scott, Foresman & Company*, 1970.

[131] TOP500. Top500 list - november 2023. *TOP500 Supercomputer Sites*, 2023.

[132] TOP500. Top500 list - november 2024. *TOP500 Supercomputer Sites*, 2024.

[133] Dana Vantrease and Nathan Binkert. Optical interconnects for high-performance computing. *IEEE Micro*, 38(2):15–24, 2018.

[134] Jeffrey S Vetter, Ron Brightwell, Maya Gokhale, Pat McCormick, Rob Ross, John Shalf, Katie Antypas, David Donofrio, Travis Humble, Catherine Schuman, et al. Computing at the exascale. *International Journal of High Performance Computing Applications*, 25(3):262–278, 2011.

[135] Jeffrey S Vetter, Ron Brightwell, Maya Gokhale, Pat McCormick, Rob Ross, John Shalf, Katie Antypas, David Donofrio, Travis Humble, Catherine Schuman, et al. Extreme heterogeneity 2018: Productive computational science in the era of extreme heterogeneity. *US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Washington, DC, USA, Tech. Rep.*, 2018.

[136] Vikram Viswanathan and Vijay Narayanan. Charting the chiplet frontier: Design considerations for multi-chip packages. In *Proceedings of the Design Automation Conference*, pages 1–6, 2021.

[137] M Mitchell Waldrop. The chips are down for moore's law. *Nature News*, 530(7589):144, 2016.

[138] Göran Wendin. Quantum information processing with superconducting circuits: a review. *Reports on Progress in Physics*, 80(10):106001, 2017.

[139] Wm A Wulf and Sally A McKee. Hitting the memory wall: implications of the obvious. *ACM SIGARCH computer architecture news*, 23(1):20–24, 1995.

[140] Wm A Wulf and Sally A McKee. Hitting the memory wall: implications of the obvious. *ACM SIGARCH computer architecture news*, 23(1):20–24, 1995.

[141] Jin Yang, Jerry Zhao, and Tao Li. High-performance data-centric networking with smartnics. In *ACM International Conference on Supercomputing*, pages 322–333, 2020.

[142] Mohamed Zahran. Heterogeneous computing: Here to stay. *Communications of the ACM*, 60(3):42–45, 2017.

[143] Xuehai Zhang, Iman Sadooghi, and Ioan Raicu. Memory systems and interconnect for future high-performance computing. *Journal of Parallel and Distributed Computing*, 134:32–49, 2019.

[144] Yazhou Zhao, Lyudmila V. Goncharova, Qian Zhang, Payam Kaghazchi, Qian Sun, Andrew Lushington, Biqiong Wang, Ruying Li, and Xueliang Sun. Two-dimensional nanostructures for sodium-ion battery anodes. *Nano Energy*, 79:105339, 2020.