A STUDY ON THE LATEST INTEL, AMD AND ARM SERVER PROCESSORS

Utsav Jaiswal
Dept. of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur
uj8866@srmist.edu.in

Sandesh Pandey
Dept. of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur
sp2044@srmist.edu.in

Sneha Puri
Dept. of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur
sp8526@srmist.edu.in

***Abstract*—This electronic document presents a comprehensive study of the latest server processors from industry leaders Intel, AMD, and ARM. Analyzing the architectural designs, performance metrics, and technological innovations of Intel Xeon Phi, AMD EPYC 9000 series, and ARM armv9 Neoverse processors, we provide insights into their respective strengths and weaknesses. Our findings reveal distinct architectural features, performance characteristics, and market impacts unique to each processor family. We highlight the increasing importance of energy efficiency, scalability, and versatility in server processor design, reflecting evolving demands in data center and cloud computing environments. By comparing and contrasting these leading solutions, we contribute to a deeper understanding of the current state and future trends in server processor technology. The conclusions drawn from this research offer valuable guidance for industry stakeholders, informing strategic decisions and investments in next-generation server infrastructure.**

***Keywords*—Cache, Cores, Threads, Clock Speed, Memory Bandwidth, Transistors, Multithreading, Microarchitectures, Virtualization, Prefetchers**

I. Introduction

The advancement of server processors stands at the forefront of modern computing, underpinning the infrastructure that powers the digital world. Over the years, major players such as Intel, AMD, and ARM have continuously pushed the boundaries of performance, efficiency, and innovation in their server processor architectures.

As data volumes continue to skyrocket and demand for computational resources intensifies, the need for high-performance, energy-efficient server processors has never been greater. The latest advancements from Intel, AMD, and ARM promise to redefine the landscape of server computing, offering unprecedented levels of performance, scalability, and versatility. Birthed from decades of innovation and hard work, the current state of server processors is remarkable. Intel's Xeon Phi series, AMD's EPYC 9000 series and ARM's armv9 Neoverse series have created a high benchmark for the upcoming server processors to beat. This study covers the current state of the server processors manufactured by the aforementioned big players of current industry.

Our analysis will encompass architectural design differences, performance metrics evaluation, technological innovations, scalability, versatility, and possible future trends. By addressing these key questions and topics, this paper aims to provide a comprehensive understanding of the evolving landscape of server processor technology and its implications for future computing environments.

II. Study on Latest Intel Server Processors

A. Overview

Intel's newly released server processors, belonging to the Xeon series, can be considered the embodiment of the company's commitment to performance, reliability, and ingenuity as far as data centers are concerned. The Xeon family, considered by many Industry trailblazers for reliability, quality, and performance in the enterprise, Cloud and HPC workloads [1], has a longstanding history of living up to expectations.

The current Xeon systems, which encompass the microarchitectures of Skylake, Cascade Lake, Ice Lake or their later editions, are the key to the combination of the most advanced technologies and close design work. These processors have been developed by the manufacturers to address the different needs of an advanced data center hardware appliance with their edge in the areas of performance, scalability, and security.

Skylake Xeon processors, adopting the 2017 introduction, marked a new era of server computing with increased Core Counts, Memory Bandwidth enhancement and top-level (fine-grained) security features [5]. The following versions, such as Cascade Lake and Ice Lake, have been extensively tested and of course, further pushed the performance and efficiency boundaries, implementing technology as well as architectural optimizations [2][9].

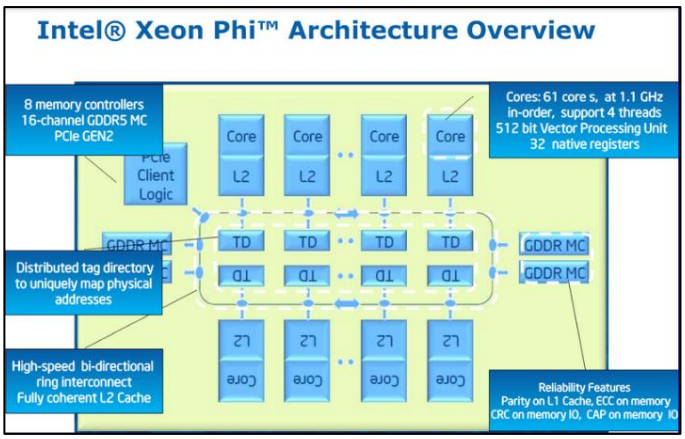

Fig. 1. Architecture of Intel Xeon Server Processor

One such example is the Cascade Lake Xeon Scalable processors which were equipped with the Intel Deep Learning Boost (DL Boost) technology and performed faster workloads while also improving deep learning for AI applications [2]. Ice Lake Xeon processors constructed on 10-nm process nodes boast higher-level SMT (Simultaneous Multithreading), faster memory channel, and support PCIe Gen4, providing enhanced capabilities for cloud and HPC computing [1].

This is proven true by Intel's never giving up on the best performance through the improvements and enhancements in each building generation of Xeon processors. The company keeps spending cash on research and new technologies to react to the market and shifting demands of the users, so that Xeon processors remain the game's main players in server computing..

Through its range of server processors aimed at different types of workloads, Intel equips companies with tools to achieve optimum performance, scalability, and efficiency in their data center investments. Xeon family's heritage of reliability, adaptability, and leadership in the industry shows Intel's role as a leading force in server computing development.

B. Specifications and Architecture

The Intel Xeon processors hold a progressive performance in servers and workstations that are verified by its leading technology innovations. The new generation of Xeon processors, with its state-of-the-art features and architecture, tries to solve the hard problems and the increasing needs of the modern Data Center environment.

1.) Performance

Leverages microarchitectures like Skylake, Cascade Lake, and Ice Lake from Intel’s advanced technologies which delivers better results. Ranges from 64 cores (Nvidia GeForce RTX 3090) to 56 cores (Nvidia GeForce RTX 3080) and with the Titan RTX, now featuring 144 cores. This offers tremendous capacity for multi-threaded workloads.

2.) Cache

Although the size of L3 caches depends on the chips, some multi-chip models can go up to 60MB of cache per socket. Employs Smart Cache technology of Intel for procuring data in an efficient way and performance boost.

3.) I/O

Provides a broad platform for the PCIe Gen4 and Gen5 connections with the external components, solutions and accelerators.

4.) Memory Support

Its memory support spans from DDR4 to DDR5 with eight channels for DDR4 and twelve channels for DDR5. Ram provides high memory bandwidth which helps in quick processing of data and facilitating faster response from the system.

5.) Other Specifications

Once the transistors are manufactured on enhanced process nodes such as 10nm and 14nm processes to accomplish high performance and low power consumption [1]. Comprehending Intel's security features, including Intel Hardware Shield and Intel Trusted Execution Technology (TXT) for execution of software, and Intel Software Guard Extensions (SGX) for system protection against security vulnerabilities [2].

Meanwhile, the Intel Xeon processors have proven to outperform competition and are designed to provide extreme performance, scalability, and reliability for a wide range of server and workstation requirements in business environments [1].

Intel Xeon processors make use of sophisticated microarchitectures which were developed internally by Intel and designed to meet the fast application and ever-increasing demands on the productivity of the modern server and workstation. While Intel does not typically disclose detailed architectural information publicly, I can provide an overview of some key features and advancements found in recent Intel Xeon processor microarchitectures [3].

While Intel does not typically disclose detailed architectural information publicly, I can provide an overview of some key features and advancements found in recent Intel Xeon processor microarchitectures:

1.) *Skylake Microarchitecture:*

Launched in 2015, Skylake is known among previous microarchitectures as the most incredible improvement, with

the features of marvelous performance, power efficiency, and workability. These updates included increasing the CPU counts, improved cache hierarchy, and support for enhanced AVX-512 instructions. So, better parallel processing capabilities were made available for computer-intensive tasks leading to the delivery of higher performance.

2.) *Cascade Lake Microarchitecture:*

Cascade Lake, the product produced in 2019 and onwards, utilizes Skylake as the baseline to offer further performance improvements in all dimensions such as security, reliability, and memory performance. It begins with Intel Deep Learning Boost (DL Boost) technology for the quick AI workloads and covers hardware possessing the Meltdown and Specter vulnerabilities for security purposes. Cascade Lake also adds more capacities for RAM and faster ERAM encryption features.

3.) Ice Lake Microarchitecture:

Ice was built with the 10 nm process node and with big changes in the architecture design. It has more cores, greater IPC (Instructions Per Clock ratio), and additional features such as PCIe Gen4 and AVX-512 support. The Ice Lake family of processors provides much better performance and power management compared to the previous generation, making them perfect for all types of servers and workstations.

4.) Future Microarchitectures

Intel is still an inventor of microarchitecture designs to stay on top of changing market needs and technology typhoons. Usually, the specific details of forthcoming microarchitectures remain confidential until the moment which is closer to the release of their new generation. Intel's roadmap generally highlights three major competence factors, namely performance, efficiency, and security, to meet data center customers' requirements.

C. Strengths

1.) Innovation and Evolution

The technology of the Xeon processors is evolving and it is upgraded with the latest stuff in the field of architectural engineering. With intellectual thoughtlessness, the high quality that Xeon processor keeps up the level of performance, efficiency and security is considered to be on the top of today's computing systems that change all the time.

2.) Customization and Optimization

Xeon chips facilitate the users to come up with various customization options and modules specific for workload and optimization with the tools for specific tasks. Companies can make use of the abilities which, for example, Intel Performance Maximisers or Intel DCT can bring to any of their systems, and improve it towards the goals to make it as efficient, reliable, and powerful as possible.

3.) Industry Leadership

On the other hand, the strong points of Xeon CPUs from Intel include the company's huge number of years developing and researching, as well as a lot of experience in this field. They are underpinned by a successful community of hardware and software suppliers, supplying full the variety of solutions and broad compatibility with various server and workstation applications.

4.) Industry Leadership

The Xeon series of processors is based on Intel's knowledge and benefits from Intel's long-term investment in research and development in CPU design. They are heavily supported by an industry environment of hardware as well as software vendors that will cover all needs and compatibility of the system in a complex server environment.

5.) Advanced Virtualization Support

Xeon processors are concerned with virtualization aid which is provided through the technologies such as the Intel Virtualization Technology (VT-x) and Intel VT-d that help in virtualized environment utilisation. Hence, businesses can use the resource fully through utilising the virtual servers, expand easily, and create the virtualized server deployment more efficiently with it.

D. Weaknesses

1.) Cost:

The main difference between Xeon processors and consumer-grade CPUs is the expense. This makes it a challenge for small firms and price-conscious users. The high upfront price, with the need for high-end components and infrastructure to supplement them, may be a hurdle for some businesses, financially.

2.) Power Consumption:

Xeon processors, designed with a focus on performance, could demand more power in general than their less powerful cousins. It might result in an above average energy usage rate which will consequently increase the overall operational costs especially in large-scale data centre systems where generally there are power efficiency issues.

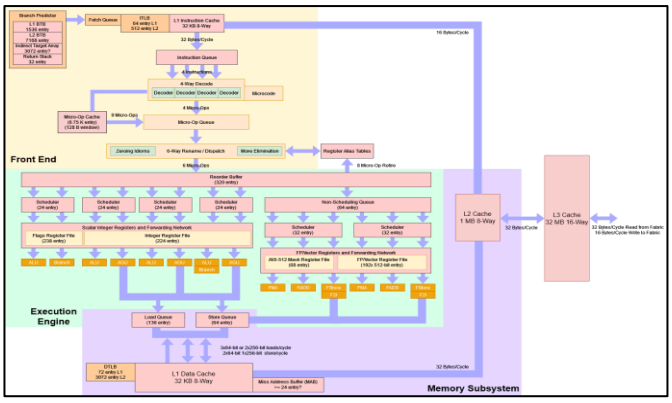

Fig. 2. Block Diagram of AMD's Zen 4 architecture

3.) Complexity:

The deployment and administration of the Xeon-based servers are not an easy task and call for privileged human resources with experience in server hardware, software, and system administration. Case of small businesses or even organisations having no dedicated personnel in the IT sector can be a bit challenging in setting up and maintaining Xeon-powered infrastructure since such a sector requires some skills.

4.) Limited Compatibility

The advantage of Xeon is that it has wide industry compatibility with standard server locations and programs. However, there may be some application or legacy systems issues and compatibility problems. Those LLCs that own personalised or niche software often face problems implementing their programs in Xeon-based systems. Moreover, purchasing the latest if 2nd and 3rd generation Xeon processors involves software revisions or driver updates, often resulting in incompatibility issues with the older systems.

5.) Heat Dissipation

Xeon processors generally increase heat production during operation, especially the ones that have many cores designed specifically for high performance. Timely heat dissipation is a key to the proper functioning of the component and to deliver a great performance. Such may entail investment in better cooling techniques, like more advanced air or liquid cooling systems which could lead to more spending and system complexity of Xeon processor-based server establishment.

III. AMD

A. Overview

AMD's latest server processor, part of the EPYC lineup, represents a culmination of the company's commitment to performance, energy efficiency, and innovation in the data center market [1]. The EPYC series has a rich history of challenging industry norms and delivering exceptional value to cloud, enterprise, and HPC workloads [2].

The current generation of EPYC processors, known as the 4th Gen EPYC, introduces significant advancements built upon the Zen 4 microarchitecture and fabricated using a cutting-edge 7nm process [3]. Codenamed Genoa, Bergamo, and Siena, these processors signify a new era in server computing [1].

Genoa, unveiled on November 10, 2022, boasts a range of 16 to 96 Zen 4 cores, setting a new standard for core count and performance [1]. With support for PCIe 5.0 and DDR5 memory, Genoa optimizes data throughput and system responsiveness while emphasizing energy efficiency [4]. This focus on efficiency translates into a lower total cost of ownership for enterprise and cloud data center clients, a key selling point highlighted by AMD CEO Lisa Su [5].

Furthermore, AMD's continuous innovation is evident in the introduction of Genoa-X, featuring 3D V-Cache technology for enhanced technical computing performance, and Bergamo for cloud-native computing applications [1]. These variants expand the EPYC portfolio to cater to diverse workload requirements, ensuring flexibility and scalability for a wide range of use cases [6].

In September 2023, AMD introduced the Siena lineup, targeting low-power computing scenarios with the Zen 4c microarchitecture [7]. With support for up to 64 cores and utilizing the SP6 socket, Siena offers a balance of performance and efficiency tailored to specific deployment needs [8].

The EPYC Genoa processor, with its up to 96 Zen 4 cores, support for DDR5 memory, PCIe Gen 5 connectivity, and innovative features such as 3D V-Cache technology, represents a significant leap forward in server processor capabilities [1]. AMD's relentless pursuit of performance, efficiency, and versatility underscores its commitment to driving the future of data center computing [9].

B. Specifications

The AMD EPYC 9000 series is known for its Exceptional Performance. It utilizes a Zen 4 architecture for enhanced performance while offering up to 128 cores, setting a new standard for processing power. It also iincorporates innovative 3D V-Cache Technology, boosting cache capacity and improving performance. Similarly, it also supports high-bandwidth DDR5 memory, ensuring efficient data access and throughput.

In the security aspect, the processor is protected by AMD Infinity Guard, providing comprehensive security features to safeguard critical data and infrastructure.

AMD EPYC™ processors are recognized as the most energy-efficient x86 servers in the industry. Balances exceptional performance with reduced energy consumption, leading to cost savings for data center operators.

1.) Cores and Threads:

The processor boasts a powerful Zen 4 core architecture, offering up to 128 cores and 256 threads depending on the specific model [1]. This means that the processor can handle a very large number of tasks simultaneously, making it ideal for demanding workloads such as scientific computing, video editing, and large-scale simulations [2].

2.) Cache:

The processor offers a substantial L3 cache, reaching up to 1152 MB per socket depending on the specific model. This large cache size helps store frequently accessed data readily available for the processor, reducing the need to fetch data from slower main memory. Additionally, some models come equipped with 3D V-Cache technology, further expanding the cache size and enhancing performance in tasks that heavily rely on cache access, such as gaming and professional software applications.

3.) I/O:

This motherboard supports the latest PCIe Gen 5 standard [1]. This translates to high-speed communication between the motherboard and other components like graphics cards and storage devices [2]. Simply put, data can move much faster between these parts, resulting in improved performance for tasks that rely on heavy data transfer, such as demanding video games and professional content creation applications [3].

4.) Memory Support:

This system boasts up to 12 channels of DDR5 memory [1]. DDR5 is the newest generation of memory technology, offering significantly faster data transfer speeds compared to previous generations [2]. Combined with the high memory channel count, this allows for exceptional memory bandwidth, meaning data can flow much quicker between the processor and memory [3]. This translates to improved performance in applications that rely heavily on data movement, such as video editing, scientific computing, and multitasking with demanding programs [4].

5.) Other Specifications:

This processor is built on a cutting-edge 5nm manufacturing process. This translates to smaller transistors, which can improve both performance and efficiency. In simpler terms, the processor can pack more power while using less energy. Additionally, it supports AMD Infinity Guard security features, offering a comprehensive suite of technologies to safeguard your system from potential security threats. For more information, Table 1 can be referenced.

C. Architecture

The latest generation of AMD Server processor is seen to use the newest Zen 4 architecture under the hood [1]. Zen 4 is a microarchitecture developed by AMD as a successor to Zen 3 [1]. Zen 4 was first mentioned by Forrest Norrod during AMD's EPYC One Year Anniversary webinar [2]. During the next horizon event which was held on November 6, 2018, AMD stated that Zen 4 was at the design completion phase [3].

Apart from the EPYC series, even the newest Ryzen 7000 series implement this architecture [4].The Zen 4 architecture is a significant advancement in microarchitecture, featuring a 64-bit superscalar design with out-of-order execution and 2-way simultaneous multithreading (SMT) [5]. It incorporates advanced dynamic branch prediction and supports 4-way decoding of x86 instructions with a stack optimizer [6].

Key features include multiple caches, such as an Op cache for decoded instructions, and prefetchers for code and data [7]. It includes four integer/address and two floating-point instruction schedulers, along with 3-way address generation and 5-way integer execution [7].

MODEL	CORES	THREADS	MAX. BOOST CLOCK	ALL CORE BOOST SPEED	BASE CLOCK	L3 CACHE	TDP
9654P	96	192	Up to 3.7GHz	3.55GHz	2.4GHz	384MB	360W
9654	96	192	Up to 3.7GHz	3.55GHz	2.4GHz	384MB	360W
9634	84	168	Up to 3.7GHz	3.1GHz	2.25GHz	384MB	290W
9554P	64	128	Up to 3.75GHz	3.75GHz	3.1GHz	256MB	360W
9554	64	128	Up to 3.75GHz	3.75GHz	3.1GHz	256MB	360W
9534	64	128	Up to 3.7GHz	3.55GHz	2.45GHz	256MB	280W
9474F	48	96	Up to 4.1GHz	3.95GHz	3.6GHz	256MB	360W
9454P	48	96	Up to 3.8GHz	3.65GHz	2.75GHz	256MB	290W
9454	48	96	Up to 3.8GHz	3.65GHz	2.75GHz	256MB	290W
9374F	32	64	Up to 4.3GHz	4.1GHz	3.85GHz	256MB	320W
9354P	32	64	Up to 3.8GHz	3.75GHz	3.25GHz	256MB	280W
9354	32	64	Up to 3.8GHz	3.75GHz	3.25GHz	256MB	280W
9274F	24	48	Up to 4.3GHz	4.1GHz	4.05GHz	256MB	320W
9254	24	48	Up to 4.15GHz	3.9GHz	2.9GHz	128MB	200W
9224	24	48	Up to 3.7GHz	3.65GHz	2.5GHz	64MB	200W
9684X	96	192	Up to 3.7GHz	3.42GHz	2.55GHz	1152MB	400W
9384X	32	64	Up to 3.9GHz	3.5GHz	3.1GHz	768MB	320W
AMD EPYC™ 9184X	16	32	Up to 4.2GHz	3.85GHz	3.55GHz	768MB	320W

Table 1. AMD Official datasheet for EPYC 9000 series with detailed specifications

The speculative, out-of-order load/store unit can handle up to three loads or two stores per cycle with a sizable load and store queue. Zen 4 also introduces AVX-512 instruction support, increased cache sizes, improved paging capabilities, higher transistor density due to the 5nm process, and enhancements to register files and reorder buffers.

Additionally, Zen 4 is capable of higher all-core clock speeds, with some models reaching 5GHz or higher. The architecture offers improved performance for various instructions and operations, making it a significant upgrade over its predecessor, Zen 3. Specifically in the EPYC 9004 "Genoa" package we are introduced to a maximum core/thread count of 96/192, further enhancing performance and scalability for server application.

D. Strength

The strengths of AMD EPYC processors lie in their exceptional performance, energy efficiency, and versatile capabilities, making them a preferred choice for modern data center environments. Here are some distinguished strengths:

1.) Unmatched Core Count:

The 9000 series offers up to 128 cores, currently the highest core count available in server processors. This allows for exceptional parallel processing capabilities, making them ideal for workloads that can be broken down into many smaller tasks, such as large-scale rendering, scientific simulations, and high-density virtualization.

2.) 3D V-Cache Technology (on select models)

This innovative feature significantly increases the amount of readily available cache. This translates to dramatic performance improvements in applications that rely heavily on cached data access, like in-memory databases, weather forecasting, and complex engineering simulations.

3.) Increased Memory Bandwidth:

With support for 12 channels of DDR5 memory, the 9000 series boasts significantly faster data transfer rates between the processor and memory. This is beneficial for applications that require frequent data movement, such as real-time analytics, large memory databases, and in-memory computing.

4.) *Enhanced I/O Capabilities:*

The 9000 series offers PCIe Gen 5 support, doubling the data transfer rate compared to the previous generation. This enables faster communication with other high-performance components like GPUs, accelerators, and high-speed storage devices, leading to an overall performance boost for data-intensive workloads.

5.) Improved Manufacturing Process:

Manufactured on a 5nm process, the 9000 series benefits from increased transistor density, potentially leading to better performance and improved power efficiency compared to the previous generation built on a 7nm process.

E. Weaknesses

Despite its impressive performance and features, the AMD EPYC 9000 series is not without its drawbacks. These weaknesses may pose challenges for certain deployment scenarios and require careful consideration by prospective users. Some of the weaknesses are highlighted below:

1.) High Cost:

As top-of-the-line server processors, the 9000 series likely carries a hefty price tag. This can be a significant barrier for adoption, especially for cost-conscious businesses.

2.) Potentially High Power Consumption:

While AMD claims improved efficiency, the high core count and processing power likely result in significant power draw. This translates to higher electricity costs and the need for robust cooling solutions to maintain optimal performance.

3.) Heat Generation:

The high power consumption inevitably leads to substantial heat generation. Improper thermal management can throttle performance and potentially lead to hardware issues in the long run.

4.) Software Compatibility (Potential Issue):

Being a new architecture, there's a possibility of encountering software compatibility issues, especially with older applications. Thorough testing is recommended before widespread deployment to ensure all essential software functions as expected.

5.) Limited Availability (Potential Issue):

New processor launches can sometimes experience supply chain constraints or high demand, leading to limited availability at launch.

IV. ARM

A. Overview

Arm's Neoverse processor family has established itself in data center performance and efficiency [1]. The latest addition, the Neoverse V3, builds on this legacy by delivering significant advancements for cloud computing, High-Performance Computing (HPC), and Artificial Intelligence (AI) applications [2].

The Neoverse family began in 2017 with Arm's vision for a range of server-grade CPU cores, with the high-performance V series, the balanced E series, and the power-efficient A series [3].

The Neoverse V1, codenamed Zeus, was the first iteration, followed by the V2 which further refined performance. Following its predecessor, Arm's latest iteration, the Neoverse

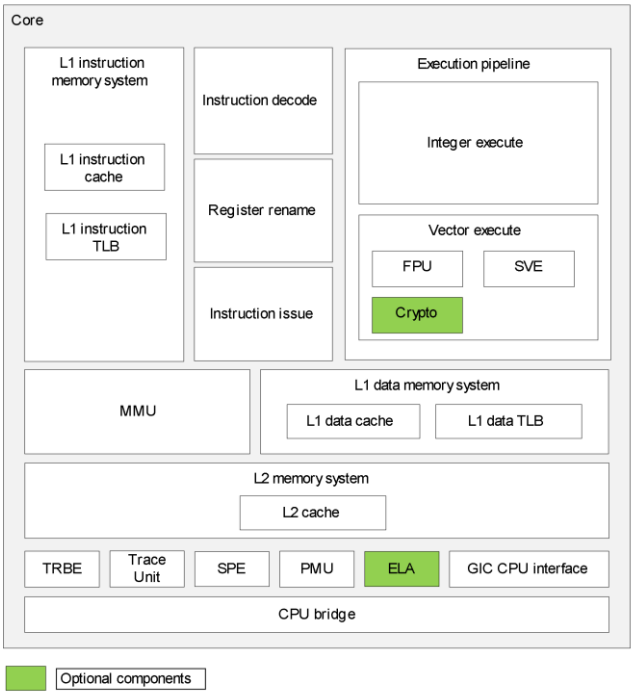

Fig. 1.1. Neoverse V3 core components

V3 processor boasts significant advancements built upon the Armv9.2-A architecture and manufactured using a cutting-edge fabrication process. Nicknamed "Poseidon", the Neoverse V3 reigns supreme as Arm's fastest CPU core to date. It delivers double-digit performance improvements over its predecessor, the V2, particularly in workloads like databases, caching, and machine learning. This translates to faster processing times and improved responsiveness for demanding applications.

The V3 isn't just about raw power. It boasts a robust memory subsystem, high-speed connections between processing cores, and the capability to integrate seamlessly with AI accelerators [9]. This combination makes it the ideal foundation for developing next-generation AI solutions [10].

The Neoverse V3 improves security by introducing support for Arm Confidential Compute Architecture [11]. This innovative feature enables the creation of highly secure, memory-encrypted cloud virtual machines, safeguarding sensitive data in the cloud environment [12].

The V3 caters to diverse workloads with its support for configurations ranging from a single die with 64 cores to a two-die configuration offering 128 cores per socket [13]

B. Specifications

The Neoverse V3 utilizes a compelling set of specifications designed to deliver exceptional performance and efficiency within the data center. At its core, the V3 leverages the Armv9.2-A architecture, ensuring compatibility with the latest software and tools[1]. In terms of memory, the V3 offers

support for up to 12 channels of DDR5/LPDDR5, enabling high bandwidth and rapid data transfer rates. This translates to faster processing of large datasets and improved responsiveness for applications that rely heavily on memory access. Additionally, the V3 has up to 64 lanes of PCIe Gen 5 or CXL (Cache-coherent Link) I/O, facilitating high-speed communication with accelerators and storage devices. This combination of advanced memory and I/O technologies significantly reduces bottlenecks, leading to an overall boost in system performance and responsiveness. The V3 also provides flexibility by offering a choice between 2MB and 3MB of L2 cache per core, allowing for customization based on specific workload requirements. These specifications solidify the Neoverse V3's position as a powerful and versatile processor, well-equipped to handle the growing demands of modern data centers.

1.) Cores and Threads:

The Neoverse V3 Compute Subsystem (CSS v3) is offered in various configurations with up to 128 Neoverse V3 cores in a socket, and 2-socket configurations supported. It can also be scaled down to support smaller configurations (ex, 32-cores).

2.) Cache:

The L1 instruction memory system includes a 64KB, 4-way set associative L1 instruction cache with 64-byte cache lines add a fully associative L1 instruction Translation Lookaside Buffer (TLB) with native support for 4KB, 16KB, 64KB, and 2MB page sizes.

The L2 cache is private to the core and is 8-way (2MB) or 12-way (3MB) set associative. Configurations can be chosen between with 2MB or 3MB of L2 cache per core.

3.) I/O:

Up to 64 lanes of PCIe Gen 5: The V3 supports the latest PCIe Gen 5 standard, hence significantly faster communication with accelerators, storage devices, and network cards compared to previous generations.

CXL (Cache-coherent Link) Support: The V3 offers compatibility with CXL. This allows for cache coherency between the processor and attached devices like accelerators.

4.) Memory Support:

There are up to 12 channels of DDR5/LPDDR5 memory along with high Bandwidth Memory (HBM) Support. Memory Management Unit (MMU) includes several Translation Lookaside Buffers (TLBs), an MMU Translation Cache (MMUTC), and a translation table prefetcher.

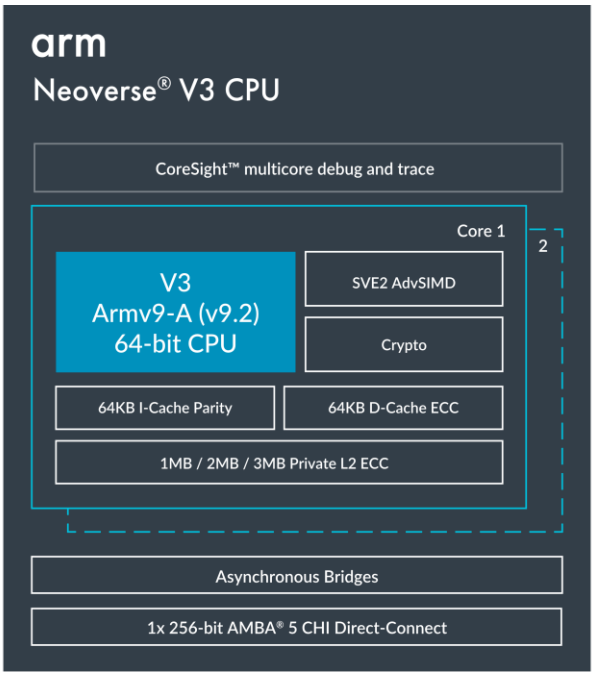

Fig. 1.2. Neoverse V3 CPU

5.) Other Specifications:

Armv9.2-A A64 instruction set, ensuring compatibility with the latest software and development tools optimized for Arm processors. The V3 incorporates a Memory Management Unit (MMU) with support for 48-bit Physical Addresses (PA) and 48-bit Virtual Addresses (VA).

Generic Interrupt Controller (GIC) CPU interface that facilitates communication with an external interrupt distributor, ensuring timely handling of interrupts and improved system responsiveness. Implementation of the Scalable Vector Extension (SVE) with a 128-bit vector length and Scalable Vector Extension 2 (SVE2).

The V3 boasts an Activity Monitoring Unit (AMU). This unit provides valuable insights into system activity, allowing for performance optimization and resource management. Support for the optional Cryptographic Extension is also ensured in this processor.

C. Architecture

As seen in Fig 1.2, The V3 is a true 64-bit processor, capable of addressing massive amounts of memory and handling complex data structures with ease. It employs out-of-order execution, a technique that enables it to dynamically rearrange instructions and executing them independently for optimal performance. The V3 incorporates sophisticated branch prediction mechanisms that allow it to anticipate the most likely execution path. It also utilizes speculative execution. This technique involves predicting future instructions and executing them speculatively, assuming the prediction is

correct. Futhermore, V3 employs a superscalar architecture, meaning it can decode and execute multiple instructions per clock cycle. Fig 1.1 shows the different core components that are present in this CPU. The components mentioned in the figure are explained in simple terms below.

1.) Instruction Decode Unit: Decodes complex AArch64 instructions into a format understandable by the processor for further execution.

2.) Register Rename Unit: Facilitates out-of-order execution.

3.) Instruction Issue Unit: Controls the dispatching of decoded instructions to various execution pipelines based on their type and dependencies.

4.) Integer Execute Pipeline: Handles arithmetic and logical operations on integer data.

5.) Vector Execute Pipeline: Processes Advanced SIMD (NEON technology) instructions for media and signal processing, floating-point operations, and supports Scalable Vector Extensions (SVE/SVE2) for efficient vector data manipulation The optional Cryptographic Extension can further accelerate cryptographic algorithms when included.

6.) Trace Unit and Trace Buffer Extension (TRBE): Enable advanced debugging and performance analysis capabilities. The Trace Unit captures execution details, while the TRBE writes this trace data directly to memory for later analysis.

7.) Statistical Profiling Extension (SPE): Offers software developers insights into instruction performance to optimize code.

8.) Performance Monitoring Unit (PMU): Provides performance counters for monitoring core and memory system activity for debugging and profiling purposes.

9.) Activity Monitoring Unit (AMU): Provides information valuable for system power management.

10.) Generic Interrupt Controller (GIC) CPU Interface: Facilitates communication with an external interrupt controller for efficient interrupt handling.

*11.) CPU Bridge:*Acts as an asynchronous interface between each core and the external system, enabling independent frequency, power, and area scaling for individual cores.

D. *Strengths*

1.) Advanced Architecture: The V3's 64-bit architecture, out-of-order execution, and sophisticated branch prediction enable efficient handling of large datasets and complex workloads. This translates to faster processing times for critical tasks in cloud computing, HPC simulations, and AI training.

2.) Boosted Vector Processing: Support for Scalable Vector Extensions (SVE/SVE2) and a dedicated Vector Execute Pipeline significantly accelerate vector operations. This is crucial for HPC workloads involving dense linear algebra and AI tasks that heavily rely on vector processing for tasks like image recognition and natural language processing.

3.) Configurable Cores and L2 Cache and High-Bandwidth Memory Support: The V3 offers flexibility in core count (32-128) and L2 cache size (2MB or 3MB per core). This allows data centers to tailor their infrastructure to specific workloads, optimizing performance and cost-efficiency for cloud deployments and HPC clusters. Additionally, up to 12 channels of DDR5/LPDDR5 memory provide exceptional data transfer rates, minimizing bottlenecks and ensuring smooth data flow for memory-intensive HPC simulations and large-scale AI training processes.

4.) Optional Cryptographic Extension: The Cryptographic Extension introduces new instructions specifically designed for cryptographic operations. These instructions leverage the processing power of the V3's vector units, significantly accelerating encryption and decryption compared to relying solely on software implementations. This translates to faster processing of secure communication protocols, improved data protection performance in cloud deployments, and faster training of secure AI models.

5.) Arm Confidential Compute Architecture Support: This innovative technology enables the creation of highly secure, memory-encrypted virtual machines in the cloud.

E. *Weaknesses*

1.) Higher Cost: As a processor with advanced features, the V3 might carry a higher price which could be a deciding factor for cost-sensitive data centers.

2.) Limited Public Availability: While the V3 has been announced, specific details about pricing and widespread

availability is limited initially. This can make it challenging to accurately plan infrastructure upgrades or budget for these processors.

3.) Software Ecosystem Maturity: New processors often require time for a robust software ecosystem to develop. While the V3 leverages the Armv9 architecture ensuring compatibility with existing tools, optimized software specifically designed to take full advantage of the V3's unique capabilities might take time to mature. This could potentially affect performance until fully optimized software becomes widely available.

4.) Power Consumption: Achieving high performance often comes at the expense of increased power draw. Data centers with strict power efficiency requirements might need to carefully evaluate the V3's power consumption compared to their current infrastructure or alternative processor options.

5.) Limited Benchmarks: Since the V3 is a relatively new product, independent benchmark results comparing its performance to established processors might be scarce initially. This can make it challenging for potential users to get a clear picture of real-world performance gains compared to existing solutions.

V. Conclusions

For the performance comparison and ranking of the aforementioned state of the art server processors, our team has taken reference from PassMark software's benchmarking as well as GeekBench. They are open-source companies that perform parallel processing and heavy tasking tests on a CPU along with standard specifications by first isolating a single core and also performing the same activating all the cores of the CPU. This gives us insights on the overall performance of the CPU. There are a few key terms to be familiar about before studying the benchmark.

The socket type, which dictates motherboard compatibility, and the CPU class, such as consumer, high-performance, or server, establish the fundamental positioning and intended use case for a given processor. The base clock-speed and maximum turbo speed provide an initial gauge of raw processing power, while the number of physical cores and amount of cache memory indicate the CPU's potential for multitasking and handling heavily threaded workloads. Power consumption metrics like Thermal Design Power (TDP) and estimated yearly running costs are important considerations for heat, cooling, and electricity expenses. To holistically evaluate and compare CPUs, benchmark scores like single-thread performance, overall CPU Mark, and price-to-performance "CPU value" offer a robust set of quantitative and qualitative measures. By examining this comprehensive set of specifications and benchmarks, users can make informed decisions in selecting the optimal CPU for their specific computing requirements and budget. The comparisons are shown in Table 9.1,9.2 and 9.3.

Intel Xeon Phi 7210 @ 1.30GHz	7,306
ARM Neoverse-N1 128 Core 3000 MHz	43,229
AMD EPYC 9654	122,091

Table 9.1 CPU PassMark Rating As of 16th of April 2024 - Higher results represent better performance

	X	**N**	**E**
Socket Type	SVLCLGA3647	NA	SP5
CPU Class	SERVER	NA	SERVER
Clock Speed	1.3 GHz	3.0 GHz	2.4 GHz
Turbo Speed	UP TO 1.5 GHZ	UP TO 3.5 GHZ	UP TO 3.7 GHZ
# of Physical Cores	64	128	96
# of Threads	64	128	128
TDP	215W	NA 2	360W
Yearly Running Cost	$39.24	$45.90	$65.70
CPU Value	3.9	0	18.2
Single Thread Rating	460 (-84.3%)	1323 (-54.7%)	2922 (0.0%)
CPU Mark	7306 (-94.0%)	43229 (-64.6%)	122091 (0.0%)

Table 9.2 Benchmarking data results X –Intel Xeon Phi 7210 @ 1.30GHz, N - ARM Neoverse-N1 128 Core 3000 MHz, E - AMD EPYC 9654

Intel Xeon Phi 7210 @ 1.30GHz	3.9
ARM Neoverse-N1 128 Core 3000 MHz	2.2
AMD EPYC 9654	18.2

Table 9.3 CPU Value (CPU Mark / $Price) As of 16th of April 2024 - Higher results represent better value

The table 9.2 showcases benchmarking results for three processors: Intel Xeon Phi 7210, ARM Neoverse-N1 128 Core, and AMD EPYC 9654. Let's delve into their key specifications for comparison.

In terms of raw processing speed, ARM Neoverse-N1 128 Core stands out with its 3.0 GHz clock speed, followed by AMD EPYC 9654 at 2.4 GHz and Intel Xeon Phi 7210 at 1.30 GHz. However, the AMD EPYC 9654 boasts the most cores (96) for tackling heavily threaded workloads, while ARM Neoverse-N1 128 Core comes in a close second with 128 cores. All three processors offer the same number of threads, matching their core count.

When it comes to power efficiency, observing table 9.2, the Intel Xeon Phi 7210 shines with the lowest thermal design power (TDP) of 215W, indicating less heat generation and potentially lower cooling requirements. This translates to a lower estimated yearly running cost of $39.24 compared to AMD EPYC 9654's $65.70. ARM Neoverse-N1 128 Core falls in between at $45.90.

To gauge overall performance for your dollar, CPU value is a helpful metric. Here, AMD EPYC 9654 takes the lead with a value of 18.2, followed by Intel Xeon Phi 7210 at 3.9 and ARM

Neoverse-N1 128 Core at 0. This suggests AMD EPYC 9654 offers the best performance for its price.

Looking at benchmark results, AMD EPYC 9654 excels in both single-threaded performance (2922 rating) and overall CPU Mark (122091), making it ideal for tasks that leverage a single core or require significant processing power. ARM Neoverse-N1 128 Core comes in second for CPU Mark (43229) but trails behind in single-threaded performance (1323 rating). The Intel Xeon Phi 7210 falls short in both categories (460 single-threaded rating and 7306 CPU Mark).

Ultimately, the best processor depends on your specific needs. If single-threaded performance is crucial, AMD EPYC 9654 is the way to go. For heavily threaded workloads, consider the high core count of AMD EPYC 9654 or ARM Neoverse-N1 128 Core. Keep power consumption and budget in mind as well. The Intel Xeon Phi 7210 excels in these areas but may not offer the raw processing power of the other two options.

ACKNOWLEDGMENT

We would like to extend a warm gratitude of thanks to Dr. Aswathy K Cherian a distinguished faculty in SRM Institute of Science and Technology for guiding us through every step in the assignment and completion of this research. Furthermore, we would like to thank Mr. Mohan Das for educating us about the Industry Essentials of Cloud Computing which led to the selection of this topic for exploration and research.

REFERENCES

[1] S. Naffziger et al., "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product," 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 2021, pp. 57-70, doi: 10.1109/ISCA52012.2021.00014.

[2] Markus Velten, Robert Schöne, Thomas Ilsche, and Daniel Hackenberg. 2022. Memory Performance of AMD EPYC Rome and Intel Cascade Lake SP Server Processors. In Proceedings of the 2022 ACM/SPEC on International Conference on Performance Engineering (ICPE '22). Association for Computing Machinery, New York, NY, USA, 165–175. https://doi.org/10.1145/3489525.3511689

[3] R. Bhargava and K. Troester, "AMD Next Generation "Zen 4" Core and 4th Gen AMD EPYC™ Server CPUs," in IEEE Micro, doi: 10.1109/MM.2024.3375070. keywords: {Registers;Microarchitecture;Servers;Throughput;Vectors;Computer architecture;Charge coupled devices},

[4] The AMD 5nm Area-Optimized ×86-64 Microprocessor Core," 2024 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2024, pp. 38-40, doi: 10.1109/ISSCC49657.2024.10454507. keywords: {Charge coupled devices;Cloud computing;Microprocessors;Sockets;FinFETs;Energy efficiency;Transistors},

[5] Y. S. Shao and D. Brooks, "Energy characterization and instruction-level energy model of Intel's Xeon Phi processor," International Symposium on Low Power Electronics and Design (ISLPED), Beijing, China, 2013, pp. 389-394, doi: 10.1109/ISLPED.2013.6629328.

[6] S. J. Pennycook, C. J. Hughes, M. Smelyanskiy and S. A. Jarvis, "Exploring SIMD for Molecular Dynamics, Using Intel® Xeon® Processors and Intel® Xeon Phi Coprocessors," 2013 IEEE 27th International Symposium on Parallel and Distributed Processing, Cambridge, MA, USA, 2013, pp. 1085-1097, doi: 10.1109/IPDPS.2013.44.

[7] Saule, Erik, Kamer Kaya, and Ümit V. Çatalyürek. 2014. "Performance Evaluation of Sparse Matrix Multiplication Kernels on Intel Xeon Phi."

[8] In Parallel Processing and Applied Mathematics, edited by Roman Wyrzykowski, Jack Dongarra, Konrad Karczewski, and Jerzy Waśniewski, 559–70. Berlin, Heidelberg: Springer Berlin Heidelberg.

[9] S. M. Tam et al., "SkyLake-SP: A 14nm 28-Core xeon® processor," 2018 IEEE International Solid-State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2018, pp. 34-36, doi: 10.1109/ISSCC.2018.8310170.

[10] H. P B, S. R. Anireddy, J. F T and V. R, "Introduction to ARM processors & its types and Overview to Cortex M series with deep explanation of each of the processors in this Family," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-8, doi: 10.1109/ICCCI54379.2022.9740768.

[11] L. Padoin, D. A. G. d. Oliveira, P. Velho and P. O. A. Navaux, "Time-to-Solution and Energy-to-Solution: A Comparison between ARM and Xeon," 2012 Third Workshop on Applications for Multi-Core Architecture, New York, NY, USA, 2012, pp. 48-53, doi: 10.1109/WAMCA.2012.10.

[12] R. Christy et al., "8.3 A 3GHz ARM Neoverse N1 CPU in 7nm FinFET for Infrastructure Applications," 2020 IEEE International Solid-State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2020, pp. 148-150, doi: 10.1109/ISSCC19947.2020.9062889.

[13] Pellegrini and C. Abernathy, "Arm Neoverse N1 Cloud-to-Edge Infrastructure SoCs," 2019 IEEE Hot Chips 31 Symposium (HCS), Cupertino, CA, USA, 2019, pp. 1-21, doi: 10.1109/HOTCHIPS.2019.8875640.

[14] M. Bruce, "Arm Neoverse V2 platform: Leadership Performance and Power Efficiency for Next-Generation Cloud Computing, ML and HPC Workloads," 2023 IEEE Hot Chips 35 Symposium (HCS), Palo Alto, CA, USA, 2023, pp. 1-25, doi: 10.1109/HCS59251.2023.10254718.

[15] A. Pellegrini, "Arm Neoverse N2: Arm's 2nd generation high performance infrastructure CPUs and system IPs," 2021 IEEE Hot Chips 33 Symposium (HCS), Palo Alto, CA, USA, 2021, pp. 1-27, doi: 10.1109/HCS52781.2021.9567483.