

# 4TH GEN AMD EPYC™ PROCESSOR ARCHITECTURE



together we advance\_data center computing

November 2022

---

# 4TH GEN AMD EPYC PROCESSOR ARCHITECTURE

## CONTENTS

---

<b>INTRODUCTION</b>	<b>3</b>
<b>HYBRID MULTI-DIE ARCHITECTURE</b>	<b>4</b>
Decoupled Innovation Paths	4
CPU Core Innovation	4
I/O Die Innovation	5
AMD Infinity Architecture	6
<b>‘ZEN 4’ CORE</b>	<b>7</b>
Double-Digit IPC Improvements	7
New Instructions	7
Security Enhancements	7
<b>SYSTEM-ON-CHIP DESIGN</b>	<b>8</b>
AMD Infinity Fabric™ Technology and the I/O Die SERDES	9
NUMA considerations	9
<b>MULTIPROCESSOR SERVER DESIGNS</b>	<b>11</b>
Single-Socket Server Configurations	11
2-Socket Server Configurations	11
<b>AMD INFINITY GUARD FEATURES</b>	<b>12</b>
Cutting-Edge Security Features	12
AMD Secure Processor	12
<b>CONCLUSION</b>	<b>14</b>

---

## INTRODUCTION

---



The information technology industry is changing rapidly, with many different workload facets demanding innovation that can satisfy their specific needs. High-performance computing and cloud applications need high-density CPUs that provide high core counts for highly parallelized workloads. Enterprise applications need a balance between CPU and I/O capability. Artificial intelligence, data analytics, high-performance computing, as well as structured and unstructured data applications are driven by the strength and speed of individual cores and accelerated mathematical functions. And network infrastructure, networking, security, and edge applications need cost-optimized systems that can be deployed widely and securely in locations around the globe.

The design decisions we have made in the 4th generation of AMD EPYC™ processors have evolved a platform that can support all of these needs. Three goals have driven the design of the AMD EPYC 9004 Series processors: performance, with a goal of double-digit instructions-per-clock (IPC) and frequency improvements; latency, with a goal of reducing average latency with higher cache sizes and effectiveness; and throughput, with a goal of reducing dynamic power to enable significantly higher core counts.

This white paper presents the processor architecture that supports the EPYC 9004 Series, and future enhancements that can enable a single-socket architecture to branch out and address a continuously widening universe of workload demands. Our hybrid, multi-chip architecture enables us to decouple innovation paths and deliver consistently innovative, high-performance products. The 'Zen 4' core represents a significant advancement from the last generation, with new support for highly complex machine learning and inferencing applications. Our system-on-chip approach helps server vendors to accelerate their designs and get innovative products into customers' hands quickly. AMD EPYC processors are the only x86 server CPUs with an integrated, embedded security processor that is "hardened at the core" to help secure customer data whether in a central data center or distributed across locations at the network edge. Finally, this paper will review some of the design choices that enable no-compromise single-socket servers as well as some of the most powerful two-socket servers on the planet.



# HYBRID MULTI-DIE ARCHITECTURE

The most important innovation in AMD EPYC processors is the hybrid multi-die architecture. We anticipated the fact that increasing core density in monolithic processor designs would become more difficult over time. One of the primary issues is the fact that the process technology that can create a CPU core is on a different innovation path than the technology that lays down the analog circuitry to drive external pathways to memory, I/O devices, and an optional second processor. These two technologies are linked together when creating monolithic processors and can impede the swift delivery of products to market.

## DECOUPLED INNOVATION PATHS

AMD EPYC processors have decoupled the two innovation paths for CPU cores and I/O functions into two different types of dies that can be developed on timelines appropriate for what they need to do. In today's 4th Gen processors, the 'Zen 4' CPU dies are produced with 5nm technology, while the I/O die is created using 6nm processes. The AMD EPYC 9004 Series processors are built with up to 12 CPU dies with up to eight cores each, a large L3 cache shared across all cores within each CPU die, and an I/O die. The result is an estimated 24% integer and 52% more floating-point top-of-stack performance

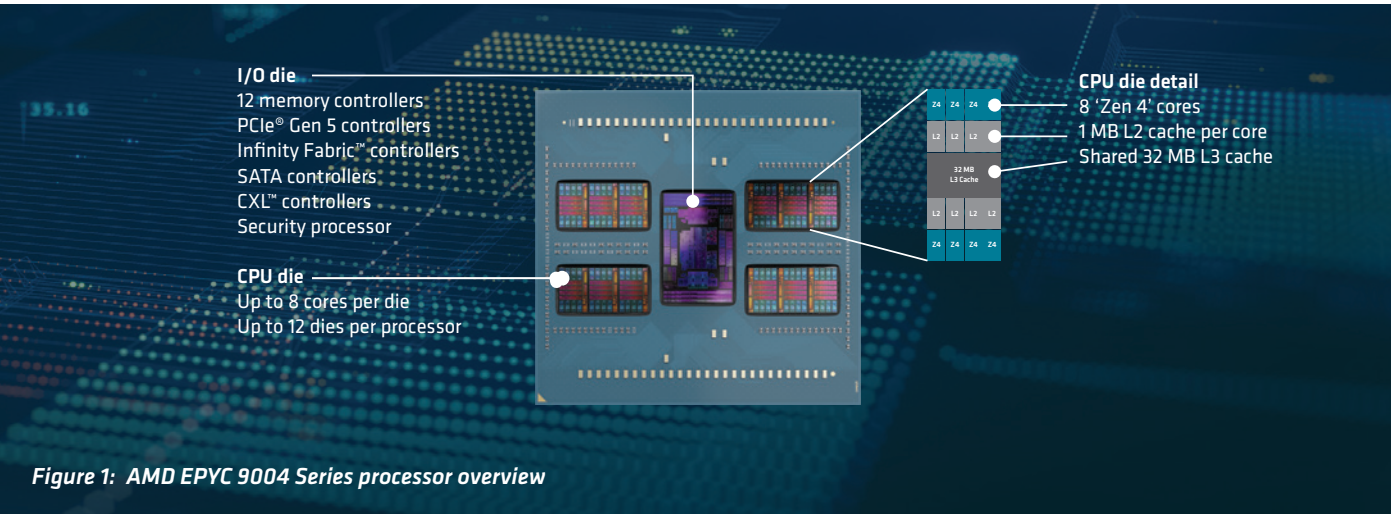
per watt over the prior generation, which frees thermal envelopes to deliver more computing power.<sup>SP5-003A, SP5-004A</sup>

This decoupling has enabled us to leap ahead of the market and stay there. The approach we have taken is more flexible and dynamic than trying to force all aspects of a processor into one fabrication technology. We believe that it is faster to deliver new and high-performance products to market by assembling modules into a processor than to create large, monolithic CPUs.

## CPU CORE INNOVATION

We can innovate with our CPU cores, and innovate we have, with smaller process sizes leading to more cores within a given thermal envelope. This, plus our continuous improvement in instructions per cycle, has resulted in double-digit performance gains with every new generation (Table 1). But that's not all. Taking a modular approach enables us to look at the CPU core complex as a unit of innovation where we can make variants to better address specific workloads. It's a flexible unit that we can swap in and out with alternatives that benefit workloads such as the following:

- **ACCELERATING PER-CORE-LICENSED SOFTWARE:** You want to get the most out of your license, so we offer a set of processors with



fewer cores and higher clock speed. Instead of using a single CPU die with eight cores, for example, we can spread those eight cores across eight CPU dies, each with one core. This spreads the thermal load across the processor and enables us to increase the clock frequency. With this eight-core example, each core would attach to its own memory channel. This spreads memory references across memory channels, helping to reduce memory latency.

- **DRIVING COMPUTER-AIDED ENGINEERING PERFORMANCE:** An innovation that you've seen in 3rd Gen AMD EPYC processors is AMD 3D V-Cache™ technology. We literally stack additional cache memory on top of each CPU die. This uses a direct copper-to-copper hybrid bonding process that enables more than 200 times the interconnect densities of current 2D technology and more than 15 times the interconnect technologies that use solder bumps for the connection.<sup>EPYC-026</sup> This innovation delivers 768 MB of L3 cache in our 3rd Gen processors with 3D V-Cache. And the future that we envision doesn't restrict 3D silicon to just memory expansion.
- **ACCELERATING CLOUD AND EDGE WORKLOADS:** CPU dies with more than eight cores hold the promise for increasing overall density for cloud computing environments, helping host more virtual

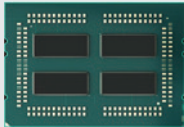
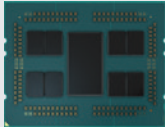
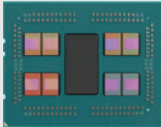
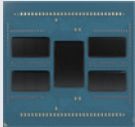
machines per server. Dies with low-power cores can be targeted to support the needs of edge locations such as the immense amount of processing associated with 5G installations.

- **SPEEDING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING:** Our latest core design includes support for the AVX-512 instructions (256-bit data path) to help speed AI and ML workloads, including BFLOAT16, and Vectorized Neural Network Instruction (VNNI).
- **SIMPLIFIED PRODUCTION:** The multi-die architecture can help reduce waste in the fabrication process. When we place many (relatively) small CPU dies on a silicon wafer, the inevitable production flaws affect a small number of dies that fail testing and are not integrated into any processors. In comparison, if the wafer contains fewer, larger, monolithic processors, a single flaw can cause the entire processor to be rejected, reducing the overall yield in terms of average number of processors produced per wafer. This can contribute to higher costs.

## I/O DIE INNOVATION

The I/O die is a place for parallel innovation. In the EPYC 9004 Series we have doubled the I/O bandwidth of the CPU from the past generation by incorporating PCIe® Gen 5 capabilities onto the I/O die. Not being satisfied with just doubling the I/O bandwidth,

*Table 1: The multi-die architecture has enabled significant improvements for each processor generation since the beginning*

	AMD EPYC 7001 'NAPLES'	AMD EPYC 7002 'ROME'	AMD EPYC 7003 'MILAN'	AMD EPYC 9004 'GENOA'
				
Core Architecture	'Zen'	'Zen 2'	'Zen 3'	'Zen 4'
Cores	8 to 32	8 to 64	8 to 64	16 to 96
IPC Improvement Over Prior Generation	N/A	24% <sup>ROM-236</sup>	19% <sup>MLN-003</sup>	14% <sup>EPYC-038</sup>
Max L3 Cache	Up to 64 MB	Up to 256 MB	Up to 256 MB*	Up to 384 MB
PCIe® Lanes	Up to 128 Gen 3	Up to 128 Gen 3	Up to 128 Gen 4	Up to 128 Gen 5 8 bonus lanes Gen 3
CPU Process Technology	14nm	7nm	7nm	5nm
I/O Die Process Technology	N/A	14 nm	14 nm	16 nm
Power (Configurable TDP [cTDP])	120-200W	120-280W	155-280W	200-400W
Max Memory Capacity	2 TB DDR3-2400/2666	4 TB DDR4-3200	4 TB DDR4-3200	6 TB DDR5-4800
* Up to 768 MB for processors with AMD 3D V-Cache™ technology				

---

the I/O subsystem on the I/O die supports AMD Infinity Fabric™ interconnects, SATA disk controllers, and Compute Express Link (CXL™) 1.1+ memory controllers that can be flexibly assigned to specific functions at server design time. The I/O die is where the dedicated security processor resides, close to the memory controllers that manage the range of memory encryption mechanisms that are part of our AMD Infinity Guard<sup>GD-183</sup> feature set.

## AMD INFINITY ARCHITECTURE

When creating a processor based on a hybrid, multi-chip architecture, the performance of the interconnect is of paramount importance. The heart of the AMD Infinity Architecture is a leadership interconnect that supports extraordinary levels of scale at every layer. Components communicate using AMD Infinity Fabric technology—a connection that is used between CPUs, between components in the multi-chip architecture, and to connect ‘Zen 4’ processor cores, memory, PCIe® Gen 5 I/O, and security mechanisms. As a result, the architecture delivers breakthrough performance and efficiency to deliver on the promise of next-generation computing.



# 'ZEN 4' CORE

At AMD, our core design is an undertaking of continuous optimization. The 'Zen 4' core integrated into 4th Gen AMD EPYC processors is the first and only x86 server CPU built with 5nm fabrication technology. Because we build our server processors as part of a multi-chip architecture, the core complex is a component that can be innovated and enhanced independently of the I/O die. For example, we enhanced the 'Zen 3' core with AMD 3D V-Cache technology to dramatically increase the amount of L3 cache on enabled processors, and enhancements such as these can be expected with the 'Zen 4' core as well. The core complex used in EPYC 9004 Series processors consists of up to eight cores, dedicated 1 MB L2 cache per core, and a 32 MB cache shared between the eight cores (Figure 2).

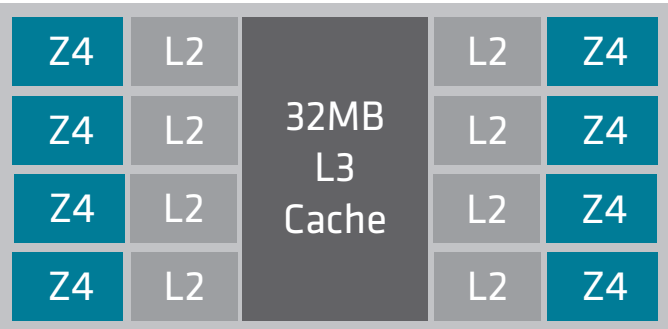


Figure 2: Layout of the 'Zen 4' core complex illustrating an 8-core die

## DOUBLE-DIGIT IPC IMPROVEMENTS

For each generation, we strive for double-digit percentage improvements in instructions per cycle, which we have been able to deliver with each new EPYC processor series (see Table 1). Improvements over the 'Zen 3' core include 1 MB L2 private cache per core, branch-prediction improvements, larger operation cache, and deeper internal buffers.

## NEW INSTRUCTIONS

The 'Zen 4' core introduces new instructions designed to advance artificial intelligence, machine learning, and high-performance computing workloads. The full set of AVX-512 instructions are implemented to match industry standards. These include BFLOAT16 and Vectorized Neural Network Instruction (VNNI). Our implementation of these data-heavy instructions enables applications that are hard coded for AVX-512 to work without modification. Our approach uses the same 256-bit data paths that exist through the CPU and enable the two parts to execute on sequential clock cycles. This means no throttling of the CPU clock is necessary to manage thermal envelopes.

With the potential to expand memory through CXL controllers, virtual memory is now addressable through 57 bits, and a fifth level of nested page tables has been implemented to support this.

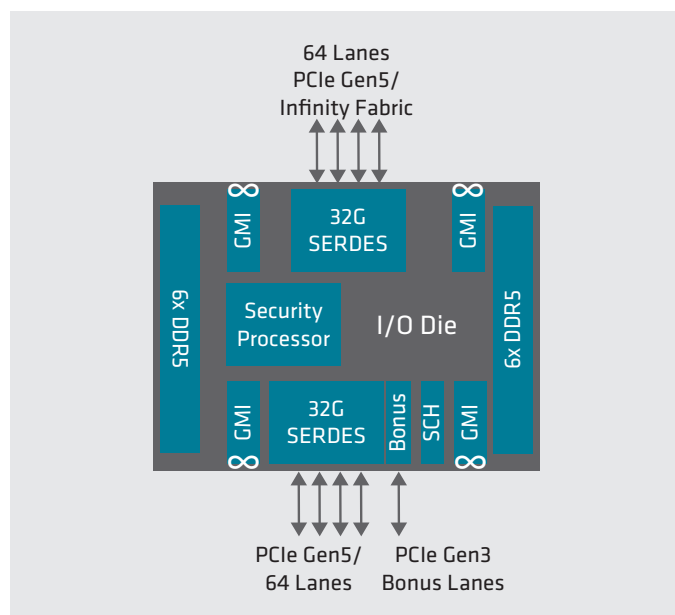
## SECURITY ENHANCEMENTS

Each 'Zen' core generation builds upon the security features of the previous one, and they incorporate mitigations for known vulnerabilities with no modifications necessary to application software. The original 'Zen' core has resisted side-channel attacks in part because of the tagging of memory to threads once read into the processor caches. This helps reduce the possibility of one thread being able to view another thread's data when in use in the processor. For the 'Zen 4' core we introduced the capability for guest operating systems in virtualized environments to run exclusively on one core—thus introducing further solutions that can help protect against side-channel attacks targeted at cached memory.

New support for virtualized environments includes secure multi-key encryption (SMKE) that enables hypervisors to selectively encrypt address space ranges on CXL-attached memory. Memory encrypted with SMKE can be accessed by the CPU across reboots, and the existing software encryption framework works seamlessly with CXL-attached memory as well independent of device implementation.

## SYSTEM-ON-CHIP DESIGN

The I/O die (Figure 3) implements many of the functions that would normally be implemented with external chip sets, thus qualifying AMD EPYC processors as systems on chip (SOCs). This approach helps reduce server design complexity and power consumption due to fewer chips. Our all-in philosophy means that every offering in our product line has the same built-in features that are listed below. This takes the mystery out of CPU selection. Just choose the core count, frequency, and L3 cache size your workload requires, and the rest are included at no extra cost.



*Figure 3: The I/O die implements many functions that would otherwise require external chip sets*

- **12 DDR5 MEMORY CONTROLLERS**—50% more memory controllers than any other x86 processor.<sup>EPYC-033</sup> Having more, and more powerful CPU cores creates a higher demand for memory, and additional memory channels and higher bandwidth keeps this equation in balance. Memory interleaving on 2, 4, 6, 8, 10, and 12 channels helps optimize for both small- and large-memory configurations. The memory controllers include inline encryption

engines for implementing AMD Infinity Guard features discussed below.

- **UP TO 128 PCIe GEN 5 LANES IN A 1P CONFIGURATION; UP TO 160 LANES IN A 2P CONFIGURATION.** The PCIe Gen 5 lanes can be dedicated to support higher-level functions including up to 32 PCIe lanes configurable as on-chip SATA controllers for massive disk capacity and up to 64 lanes configurable as CXL 1.1+ memory controllers for cache-coherent memory expansion and support for persistent memory. In server designs, the bonus lanes are often used for access to performance-insensitive I/O such as to M.2 drives used for system boot.
- **UP TO 12 PCIe GEN 3 'BONUS' LANES** in a 2-socket configuration, or 8 lanes in a single-socket configuration.
- **2X FASTER AMD INFINITY FABRIC CONNECTIVITY** over the prior generation for CPU-to-CPU connectivity. Rather than invent new connectivity mechanisms that can delay time to market, we use the same physical interfaces for Infinity Fabric connections as for the PCIe Gen 5 I/O, with different protocols layered on the physical (PHY) layer. This affords server designers the freedom to trade off more PCIe I/O lanes in exchange for fewer interprocessor communication links. AMD supports use of 3 or 4 links each of which correspond to x16 PCIe physical connections. With Infinity Fabric protocols running on these interfaces, four links can support a maximum theoretical bandwidth of 512 GB/s between servers, which more than matches maximum theoretical memory speeds of 460.8 GB/s. What this means is that remote memory access from one CPU to another can flow nearly at memory speeds.
- **UPDATED INFINITY FABRIC INTERFACE** offers up to 36 Gb/s for communication between the 'Zen 4' core complex and I/O die. The new 'Zen4' core complex can use one or two Infinity Fabric interfaces, allowing for double the CPU-core-to-I/O die bandwidth (up to 72 Gb/s) based on the number of cores in the complex. The 4th Gen EPYC I/O die offers great flexibility with twelve Infinity Fabric interfaces, enabling 4, 8, or 12 core complexes depending on the performance and power requirements per customer use case. (This is known internally as the Global Memory Interface (GMI) and is labeled this way on many figures.)



- **INTEGRATED SECURITY PROCESSOR** that supports confidential computing with features including secure root of trust, secure memory encryption (SME), and secure encrypted virtualization (SEV).<sup>GO-183</sup> This is discussed in a separate section below
- **A SERVER CONTROLLER HUB** helps minimize the required chip set for basic server control functions. It includes direct USB connectivity, 1 Gb/s LAN-on-motherboard, and various UART and I2C and I3C bus connectivity.

### AMD INFINITY FABRIC™ TECHNOLOGY AND THE I/O DIE SERDES

The use of the same physical layer to support I/O functions including AMD Infinity Fabric technology reflects our philosophy of using industry-standard, well-understood technologies that offer server designers flexibility to design innovative servers, and simplifies our CPU designs over inventing proprietary interconnects.

The PCIe Gen 5 I/O is supported in the I/O die by serializer-deserializer (SERDES) silicon with one independent set of traces to support each port of 16 PCIe lanes. The I/O die contains eight SERDES devices, and typically four are used to connect to a second processor and four connect to I/O devices. Each of these devices can be customized so that the underlying PCIe Gen 5 PHY circuitry can be used for:

- Up to 4 links of Gen3 AMD Infinity Fabric connectivity
- 128 lanes of PCIe Gen 5 connectivity to peripherals (up to 160 lanes in 2-socket designs)
- Up to 64 lanes that can be dedicated to CXL 1.1+ connectivity to extended memory
- Up to 32 I/O lanes that can be configured as SATA disk controllers

The lanes in each SERDES can be bifurcated given constraints described in server design documentation. Each SERDES has specific constraints, for example some are restricted to PCIe and Infinity Fabric connectivity, while others enable the richer set of functions.

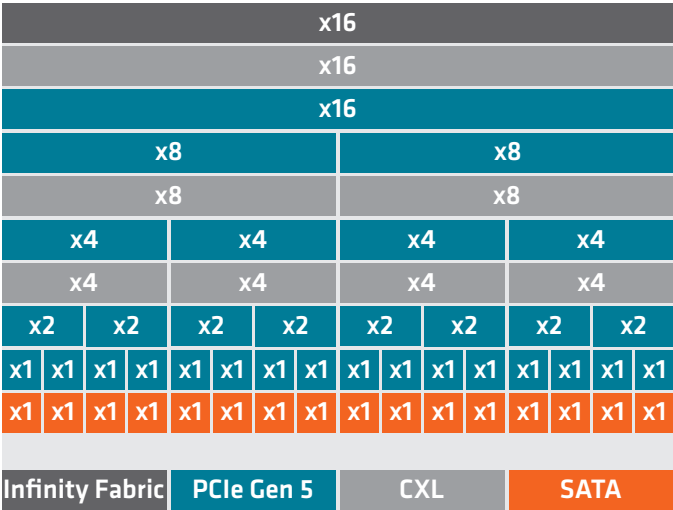


Figure 4: Idealized example of SERDES lane bifurcation options

An idealized bifurcation diagram—no single SERDES provides all of them—is illustrated in Figure 4, indicating that the entire port can be dedicated to 16 lanes of Infinity Fabric, PCIe, or CXL connectivity. These can be broken down to various combinations of x8, x4, x2, and x1 bandwidth. For example, if SATA controllers share connectivity with PCIe on a SERDES, a maximum of eight x1 SATA controllers can be allocated. Or, as the diagram illustrates, CXL connections must use a minimum of four lanes.

### NUMA CONSIDERATIONS

In a multi-chip architecture, there can be varying amounts of memory latency depending on the connectivity between memory controllers and CPU dies. This is known as non-uniform memory access, or NUMA. For applications needing to extract every last percent of latency out of memory accesses, they can take advantage of these varying latencies to create an affinity between specific address ranges and the CPU cores closest to that memory.



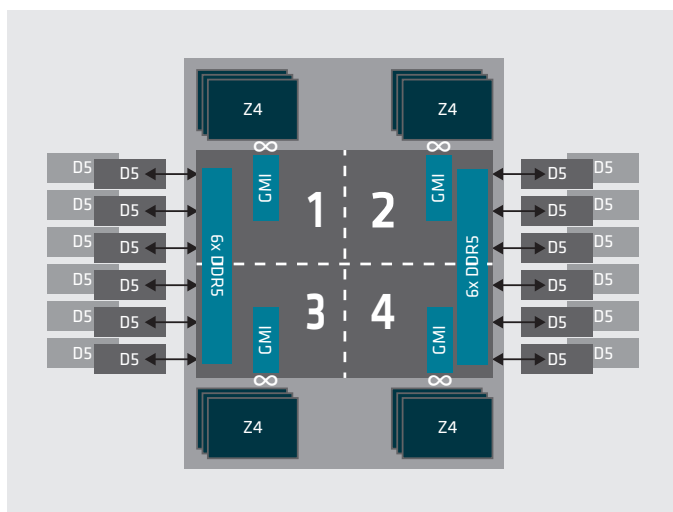
In AMD EPYC 7001 Series processors, memory controllers were located on the same die with up to eight CPU cores, creating a tight affinity between the memory controlled by the die and the CPU cores

on the die. When a memory controller had to request data destined for a different set of cores, the data had to pass from one die to another over an internal Infinity Fabric connection.

Beginning with AMD EPYC 7002 Series processors, non-uniform latency was reduced dramatically by locating memory controllers onto the I/O die. In AMD EPYC 9004 Series processors, optimizations to the Infinity Fabric interconnects reduced latency differences even further.

Still, for applications that need to squeeze the last one or two percent of latency out of memory references, creating an affinity between memory ranges and CPU cores can improve performance. Figure 5 illustrates how this works. If you divide the I/O die into four quadrants for an 'NPS=4' configuration, you will see that six DIMMs feed into three memory controllers, which are closely connected via Infinity Fabric (GMI) to a set of up to three 'Zen 4' CPU dies, or up to 24 CPU cores.

Most applications don't need to be concerned about using NUMA domains, and using the AMD EPYC processor as a single domain (NPS=1) gives excellent performance. The [AMD EPYC 9004 Architecture Overview](#) provides more details on NUMA configurations and tuning suggestions for specific applications.



*Figure 5: Dividing the AMD EPYC processor into four NUMA domains can give small performance improvements for some applications*

# MULTIPROCESSOR SERVER DESIGNS

The flexibility of the SERDES enables the Infinity Fabric interconnects to share the same physical infrastructure of chip's PCIe I/O. In Figure 6, these are labeled as 'G' and 'P' links, each of which support 16 lanes of PCIe Gen 5 connectivity. In a single-socket configuration, all Infinity Fabric links are dedicated to PCIe I/O, affording 128 lanes of Gen 5 bandwidth on AMD EPYC 9004 Series processors.

## SINGLE-SOCKET SERVER CONFIGURATIONS

AMD EPYC processors with no 'P' suffix can be used in single-socket and 2-socket configurations. Processor part numbers with a 'P' suffix are optimized for single-socket servers by dedicating the 'P' links for PCIe I/O connections only.

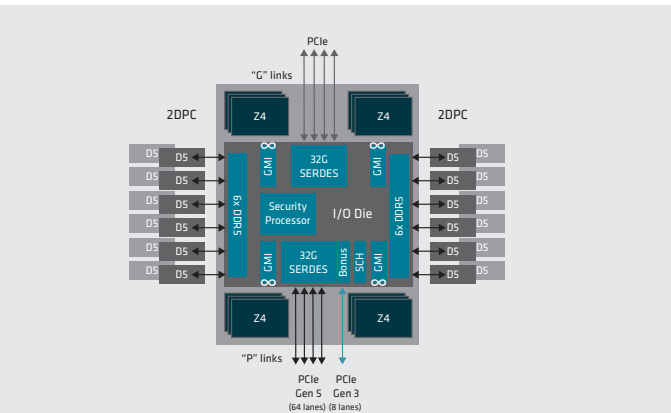


Figure 6: 4th Gen AMD EPYC processor in a single-socket server configuration with all links dedicated to PCIe connectivity

## 2-SOCKET SERVER CONFIGURATIONS

In these configurations, three or four 16-lane 'G' links are used to connect to the second processor. For I/O-intensive server designs, three links can be used as Infinity Fabric interconnects and one additional link from each CPU can be dedicated to PCIe Gen 5 I/O, bringing the server I/O capacity to 160 lanes (Figure 7).

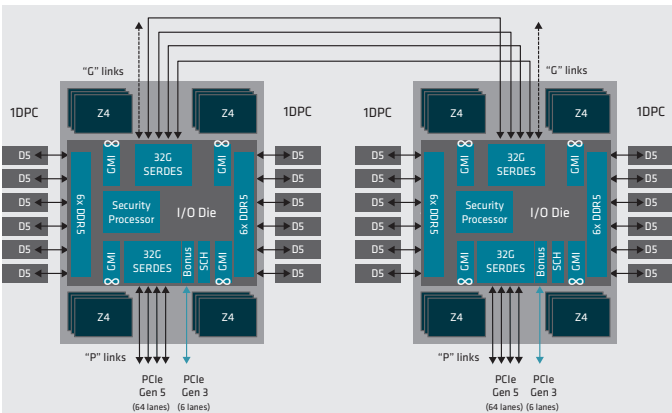


Figure 7: AMD EPYC processors in a 2-socket configurations

# AMD INFINITY GUARD FEATURES

Data is every organization's most precious asset, and AMD Infinity Guard security features are designed to help protect your data from malicious users, hypervisors, and even administrators. This approach can help mitigate the risks of attacks against physical DIMMs or attacks against guests in virtualized and hyperconverged environments.

## CUTTING-EDGE SECURITY FEATURES

Cutting-edge security features are built into our processors, and, like our core designs, they are the outcome of continuous improvement. Figure 8 illustrates the generation-over-generation improvements we have made to help hypervisors increase the isolation of virtual machines. We are proud to report that select EPYC 9004 processors are on track for United States Federal Information Processing Standard (FIPS) 140-3 certification in 2023.

## AMD SECURE PROCESSOR

Security features are managed by the AMD Secure Processor, a 32-bit microcontroller that runs a hardened operating system. The hardening process removes unnecessary components and applies previous security patches in the microcontroller to help reduce attack

surfaces. It provides cryptographic functionality for key generation and key management, and it supervises hardware-validated boot, where the foundation for platform security starts. AMD Infinity Guard security features must be enabled by server OEMs and/or cloud service providers to operate. Check with your OEM or provider to confirm support of these features. These include:

- **HARDWARE-VALIDATED BOOT** helps verify that the operating system or hypervisor software that you intended to load is what is actually loaded. The AMD Secure Processor loads the on-chip boot ROM that loads and authenticates the off-chip boot loader. The boot loader, in turn, authenticates the BIOS before any of the 'Zen' cores can execute the code. Once the BIOS is authenticated, the OS boot loader loads the operating system or hypervisor.
- **AMD SECURE MEMORY ENCRYPTION (SME)** can be used to encrypt all of main memory with no changes required to the operating system or application software. SME helps protect against attacks on the integrity of main memory (such as certain cold-boot attacks) because it encrypts the data. 256-bit AES-XTS encryption engines are built into the EPYC 9004 Series memory controllers to help reduce performance impact during reading and writing of encrypted memory. These engines can be used to encrypt memory with either 128 or 256-bit keys. The new, 256-bit

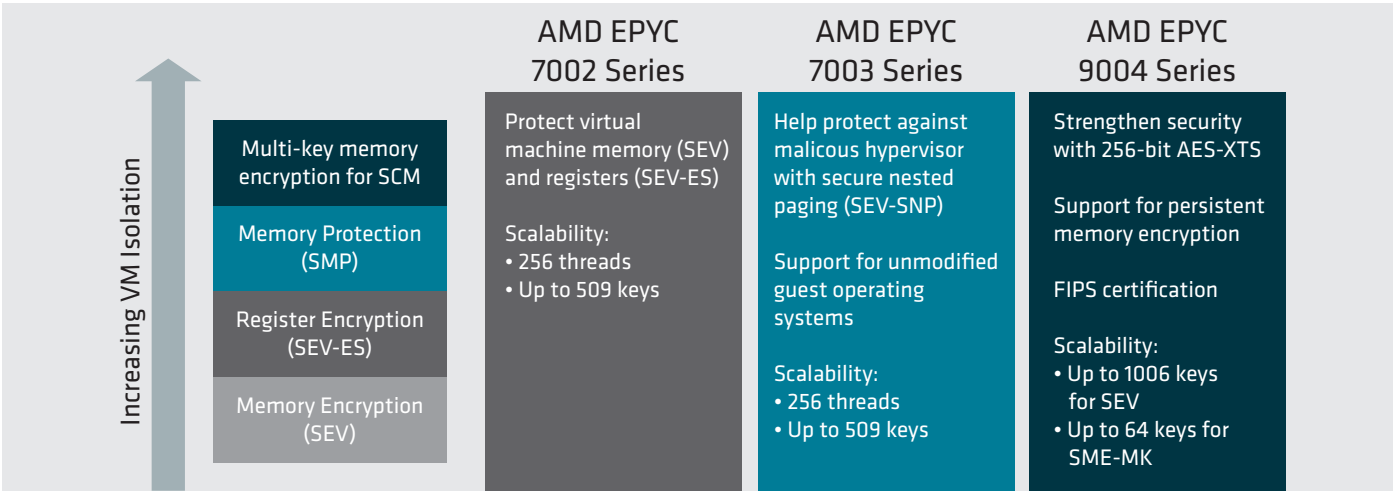


Figure 8: Each new AMD EPYC processor generation delivers more features to help isolate virtual machines

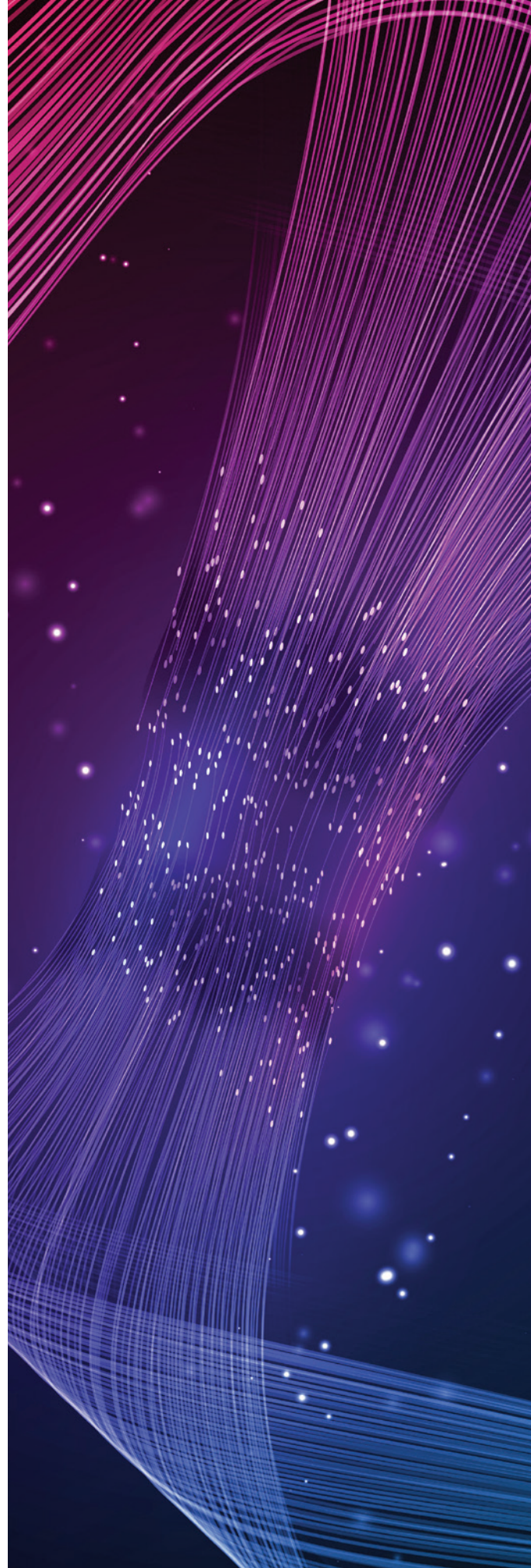


---

encryption option is integrated into the I/O die in order to support United States Federal Information Processing Standards (FIPS) 140-3 compliance. All of this is done without the encryption key being visible outside of the AMD Secure Processor.

- **AMD SECURE ENCRYPTED VIRTUALIZATION (SEV)** enables hypervisors and guest virtual machines to be cryptographically isolated from one another. Thus, if malicious software is successful in evading the isolation provided by the hypervisor, or if the hypervisor itself is compromised, reading memory from another virtual machine will expose only encrypted data for which the key is stored inside of the AMD Secure Processor and memory controllers. In 4th Gen AMD EPYC processors, up to 1006 keys can be used for virtual machine encryption.
- **AMD SECURE ENCRYPTED STATE (SEV-ES)**, introduced in 2nd Gen AMD EPYC processors, encrypts virtual machine state when interrupts cause it to be stored in the hypervisor. With this information encrypted with the virtual machine's encryption key, a compromised hypervisor is unable to view a virtual machine's registers.
- **AMD SECURE NESTED PAGING (SEV-SNP)** introduced in 3rd Gen AMD EPYC processors, builds on SEV and SEV-ES by adding strong encryption to virtual machine nested page tables to help prevent attacks such as data replay, memory remapping, and more—all with the goal to create confidential, isolated execution environments for virtual machines. With the 57-bit physical memory enabled by 4th Gen AMD EPYC processors, we have increased the page table depth that can be encrypted to five levels.
- **AMD SECURE MULTI-KEY ENCRYPTION (SMKE)**, introduced in 4th Gen AMD EPYC processors, enables fast encryption for storage-class memory, which helps data stored on CXL-attached memory to remain encrypted across a system reboot, helping protect even persistent memory from prying eyes.

This powerful set of security features, is enabled in turn by a multi-layered set of technologies accessible by all of the major hypervisor vendors. It is an innovative set of modern security features that help decrease potential attack surfaces as software is booted, executed, and processes your data. Built-in at the silicon level, AMD Infinity Guard features offer state-of-the-art capabilities to help defend against internal and external threats. Whether yours is a small- or medium-size business or an enterprise organization, implementing robust security features on premises or in the cloud is streamlined with AMD Infinity Guard technology.



---

## CONCLUSION

---



AMD EPYC 9004 Series processors demonstrate how our hybrid, multi-die architecture delivers strong innovation, helping continue to deliver customer value, with every new generation. Decoupling our core and I/O innovation processes enabled us to shrink the CPU core complexes, which in turn makes room for more cores and provides for more energy-efficient performance. Innovation in the 'Zen 4' cores unleashes a voracious appetite for memory access and I/O capacity, and we set the table with a new I/O die that supports an industry-leading 12 DDR5 memory channels, 50% more than the prior generation. We doubled our I/O and AMD Infinity Fabric™ throughput by basing them on PCIe Gen 5 interfaces and also added 'bonus' lanes for less performance-sensitive devices. Support for domain-specific instructions such as AVX-512, and connectivity to next-gen GPU accelerators prepares AMD EPYC to excel in an increasingly important world of artificial intelligence and machine learning. And if that weren't enough, support for CXL 1.1+ technology enables new Infinity Guard features to help protect even your persistent memory pools from prying eyes. We have raised the bar for data center computing once again, and more enhancements to this 4th generation of AMD EPYC processors are in the development process.

---

## END NOTES

For details on the footnotes used in this document, visit [amd.com/en/claims/epyc](https://amd.com/en/claims/epyc), [amd.com/en/claims/epyc3x](https://amd.com/en/claims/epyc3x) and [amd.com/en/claims/epyc4](https://amd.com/en/claims/epyc4).

- EPYC-026 Based on calculated areal density and based on bump pitch between AMD hybrid bond AMD 3D V-Cache stacked technology compared to AMD 2D chiplet technology and Intel 3D stacked micro-bump technology.
- EPYC-033 AMD EPYC 9004 CPUs support 12 memory channels. Intel Scalable Ice Lake CPUs support 8 memory channels.  $12 \div 8 = 1.5x$  the memory channels or 50% more memory channels per <https://ark.intel.com/>.
- EPYC-038 Based on AMD internal testing as of 09/19/2022, geomean performance improvement at the same fixed-frequency on a 4th Gen AMD EPYC™ 9554 CPU compared to a 3rd Gen AMD EPYC™ 7763 CPU using a select set of workloads (33) including est. SPECrate®2017\_int\_base, est. SPECrate®2017\_fp\_base, and representative server workloads.
- GD-183 AMD Infinity Guard features vary by EPYC™ Processor generations. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at <https://www.amd.com/en/technologies/infinity-guard>.
- MLN-003 Based on AMD internal testing as of 02/1/2021, average performance improvement at ISO-frequency on an AMD EPYC™ 72F3 (8C/8T, 3.7GHz) compared to an AMD EPYC™ 7F32 (8C/8T, 3.7GHz), per-core, single thread, using a select set of workloads including SPECrate®2017\_int\_base, SPECrate®2017\_fp\_base, and representative server workloads.
- ROM-236 Based on AMD internal testing, average per thread performance improvement at ISO-frequency on a 32-core, 64-thread, 2nd generation AMD EPYC™ platform as compared to 32-core 64-thread 1st generation AMD EPYC™ platform measured on a selected set of workloads including sub-components of SPEC CPU® 2017\_int and representative server workloads.
- SP5-003A SPECrate®2017\_int\_base estimate based on internal AMD reference platform measurements and published score from [www.spec.org](http://www.spec.org) as of 09/27/2022. Comparison of estimated 2P AMD EPYC 9534 (1070 SPECrate®2017\_int\_base, 560 Total TDP W, 128 Total Cores, \$17606 Total CPU \$, AMD Est) is 1.24x the performance of published 2P AMD EPYC 7763 (861 SPECrate®2017\_int\_base, 560 Total TDP W, 128 Total Cores, \$15780 Total CPU \$, <http://spec.org/cpu2017/results/res2021q4/cpu2017-20211121-30148.html>) [at 1.24x the performance/W] [at 1.11x the performance/CPU\$], AMD 1Ku pricing and Intel ARK.intel.com specifications and pricing as of 8/22/22. OEM published scores will vary based on system configuration and determinism mode used (default cTDP performance profile)
- S05-004A SPECrate®2017\_fp\_base estimate based on internal AMD reference platform measurements and published score from [www.spec.org](http://www.spec.org) as of 09/27/2022. Comparison of estimated 2P AMD EPYC 9534 (1010 SPECrate®2017\_fp\_base, 560 Total TDP W, 128 Total Cores, \$17606 Total CPU \$, AMD Est) is 1.52x the performance of published 2P AMD EPYC 7763 (663 SPECrate®2017\_fp\_base, 560 Total TDP W, 128 Total Cores, \$15780 Total CPU \$, <http://spec.org/cpu2017/results/res2021q4/cpu2017-20211121-30146.html>) [at 1.52x the performance/W] [at 1.37x the performance/CPU\$], AMD 1Ku pricing and Intel ARK.intel.com specifications and pricing as of 8/22/22. OEM published scores will vary based on system configuration and determinism mode used (default cTDP performance profile)

© 2022 Advanced Micro Devices, Inc. All rights reserved. All rights reserved. AMD, AMD 3D V-Cache, the AMD Arrow logo, EPYC, Infinity Fabric, and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. CXL is a trademark of Compute Express Link Consortium, Inc. Intel and Xeon are trademarks of Intel Corporation or its subsidiaries. PCIe® is a registered trademark of PCI-SIG Corporation. SPEC, SPEC CPU, and SPECrate are trademarks of the Standard Performance Evaluation Corporation. See [www.spec.org](http://www.spec.org) for more information. Other names are for informational purposes only and may be trademarks of their respective owners. LE-85001-00 11/22

