AMD EPYC™ 7002 Architecture Extends Benefits for Storage-Centric Solutions

Overview

With the release of the second generation of AMD EPYC™ family of processors, Micron believes that AMD has extended the benefits of EPYC as a foundation for storage-centric, all-flash solutions beyond the previous generation. As more enterprises are evaluating and deploying commodity server-based software-defined storage (SDS) solutions, platforms built using AMD EPYC 7002 processors continue to provide massive storage flexibility and throughput using the latest generation of PCI Express® (PCIe™) and NVM Express® (NVMe™) SSDs.

With this new offering, Micron revisits our previously released analysis of the advantages that AMD EPYC architecture-based servers provide storage-centric solutions. To best assess the AMD EPYC architecture, we discuss the EPYC 7002 for solid-state storage solutions, based on AMD EPYC product features, capabilities and server manufacturer recommendations. We have not included any specific testing performed by Micron. Each OEM/ODM will have differing server implementation and additional support components, which could ultimately affect solution performance.

Architecture Overview

The new EPYC 7002 series of enterprise-class server processors, AMD created a second generation of its "Zen" microarchitecture and a second-generation Infinity Fabric™ to interconnect up to eight processor core complex die (CCD) per socket. Each CCD can host up to eight cores together with a centralized I/O controller that handles all PCIe and memory traffic (Figure 1). AMD has doubled the performance of each system-on-a-chip (SoC) while reducing the overall power consumption per core through advanced 7nm process technology over the first generation's 14nm process, doubling memory DIMM size support to 256GB LRDIMMs while also providing a 2x peripheral throughput increase with the introduction of PCIe Generation 4.0 I/O controllers. The result is a CPU architecture that supports up to 64 cores per socket for enterprise platforms, 4TB of system memory, and 128x PCIe 4.0 lanes in single-socket and up to 160x PCIe 4.0 lanes in select dual-socket configurations, if the OEM/ODM chooses. (See "PCIe" section and Figure 2.)

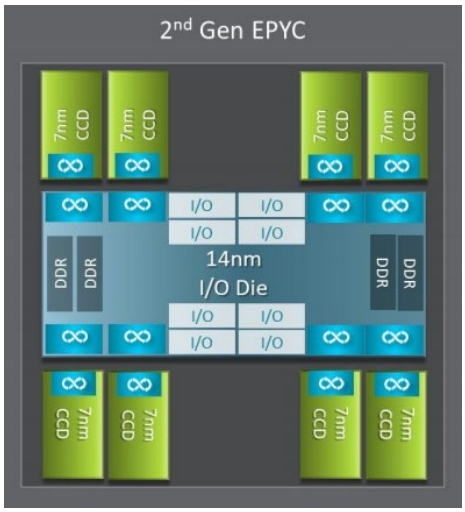

Figure 1**: **AMD EPYC 2 CPU Block Diagram

There are several features that we will discuss from a storage perspective:

- 8x memory channels per socket with 2 DIMMs per channel
- 128x or 160x PCIe 4.0 lanes

Memory Architecture

Each EPYC 7002 socket supports up to 16 DIMMs and can support various memory types (RDIMM, LRDIMM and NVDIMM) per CPU, providing 2x memory bandwidth with the second socket. Supplying eight dual-DIMM channels per CPU socket enables immense memory capacity. Based on AMD's documented specifications, a common example server solution could support up to 2TB per socket using Micron 128GB LRDIMMs, aligning with the claimed support for a maximum of 4TB in AMD collateral.

Using Micron 3200MT/s 64GB RDIMMs (MTA36ASF8G72PZ-3G2), servers may support over 51.2 GB/s of DRAM bandwidth per channel for a total of over 410 GB/s total memory bandwidth per socket.

At the time of writing, reference designs from ODMs and OEMs support EPYC single- and dual-socket designs for a total of up to 16 and 32 DIMMs per server, respectively.

AMD's 2nd Gen EPYC supports NVDIMM-N technology as well, allowing use of storage-class memory that is well-suited for deploying high-performance SDS solutions. Micron's currently available 32GB NVDIMM-N modules enable up to 64GB per memory channel. Depending on actual OEM/ODM designs, maximum NVDIMM capacity may vary. For our discussion, we assume that support for four of the available eight channels per socket for NVDIMM storage could achieve 256GB per socket of nonvolatile, memory-speed storage per socket. These robust memory capabilities provide a strong set of options to best match the requirements of SDS solutions for years to come.

PCIe

EPYC 7002 now provides support for 128x PCIe 4.0 lanes per socket, doubling the available bandwidth (2 GB/s each direction) per lane from the previous PCIe generation. PCIe provides I/O for all physical slot connections and various peripherals within the server. For dual-socket designs, AMD allows OEMs to convert either 48 or 64 PCIe lanes into 3x or 4x xGMI processor socket interconnects as shown in Figure 2.

In this brief, we focus solely on the dual-socket configuration that uses the 4x xGMI Infinity Fabric, called "Dual-Socket 64 Lane Infinity Fabric" in Figure 2.

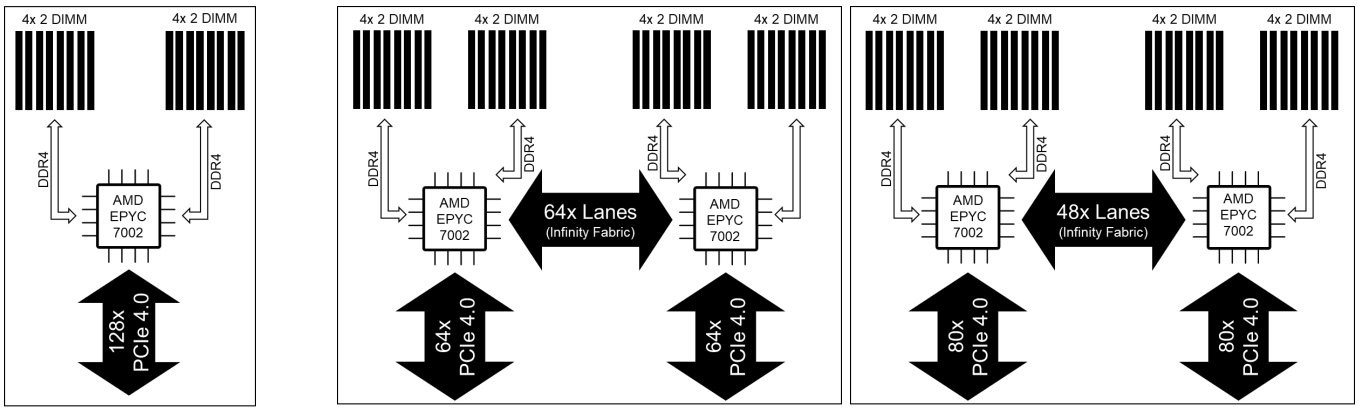

Figure 2: AMD EPYC High-Level Architectures

What Does This Mean for Storage?

It's All About PCIe Lanes

PCIe lanes mean options. It means we have the expandability we need without using cumbersome, expensive PCIe switches or expanders. Both single- and dual-socket designs have a total of 128x PCIe lanes per socket available for peripherals such as option slots, Infinity Fabric xGMI socket interconnect lanes and integrated PCIe and SATA device connections, combined enabling far greater platform flexibility than previously available. Additionally, those lanes now provide twice the throughput of the previous-generation EPYC designs that used PCIe 3.0. The impact for storage-centric solutions of this many super-fast PCIe lanes is clear: OEMs now have more options when designing a compute- or storage-centric platform, resulting in more options for businesses that are looking for high-performance solutions for their workloads.

Compute-centric solutions, such as artificial intelligence and machine learning (AI/ML), and real-time analytics require an architecture where we can deploy more DRAM and high-performance NVMe SSDs to feed applications with the data they need. The benefit is more efficient operations and deployments that could enable deployment of fewer compute nodes while providing similar capability.

Storage-centric solutions such as SDS need a balance between high-bandwidth network I/O and fast, high-capacity SSDs, ensuring the rapid movement of mission-critical data from the storage array cluster to the application servers. The expanded lane count of these solutions may allow more flexibility for OEMs designing solutions that balance network bandwidth with storage access bandwidth and do so sufficiently to enable high-performance solutions. By supporting multiple SSDs along with high-speed 100 Gb/s (and even 200 Gb/s) Ethernet ports using the same CPU complex mean less intersocket data movement — a traditional bottleneck in legacy CPU architectures. When intersocket transfers occur, the AMD Infinity Fabric architecture offers a nonblocking architecture in its standard, default configuration using 64x PCIe lanes for xGMI socket interconnect, which provides an equal amount of bandwidth to that socket's available PCIe lanes (64x PCIe 4 lanes + 64x Infinity Fabric lanes).

Pre-EPYC server architectures restricted us to four to six NVMe SSDs, attached directly to the CPU, due to the limited number of free PCIe lanes available without resorting to the use of PCIe switches or expanders. EPYC 7002 addresses that limitation. Based on the documented architecture specifications and currently available OEM/ODM designs, we can see more than 120x available PCIe lanes for use for NVMe and/or SATA SSDs and high-bandwidth x8 and x16 PCIe I/O slots. In addition, servers based on EPYC can support dual-mode storage interfaces where a mix of SATA and NVMe can directly connect to CPU interfaces.

In this brief, we assume the use of a dual-socket 2RU server offering[1] with the configuration listed in Table 1 for all-NVMe or all-SATA configurations, as appropriate.

Table 1: Slot Configuration for NVMe and SATA Servers Used in This Brief

I/O Slot Type	24x NVMe (front) + 2x SATA (rear)	24x SATA (front) + 2x SATA (rear)
x16 PCIe 4.0	3	2
x16 PCIe 3.0	1(OCP)	1(OCP)
x8 PCIe 4.0	4	5
x8 PCIe 3.0	1(OCP)	1(OCP)
M.2 PCIe 3.0	1	1
Board-to-Board (x4 PCIe 3.0)	4	4

[1] https://www.gigabyte.com/us/Rack-Server/R282-Z92-rev-100#ov(NVMe) for all-NVMe PCIe configuration example
https://www.gigabyte.com/us/Rack-Server/R282-Z91-rev-100#ov(SATA) for all-SATA PCIe configuration example

Based on these configurations, many drive + network configurations can be deployed. The following sections describe sample designs for a storage-centric solution that focuses on various technical requirements such as maximizing capacity, IOPs or throughput, and creating a balanced solution. We will make an additional assumption that we are using PCIe 4.0-compliant network adapters such as the Mellanox ConnectX®-6 SmartNIC or the Broadcom® NetExtreme® P2100 series.

Storage Design Options

Capacity-Optimized Storage Solutions

Servers designed to focus on maximum capacity per node will require many high-capacity SSDs for a given budget. This requirement means capacity solutions will prioritize lower-cost SATA SSDs or mainstream NVMe SSDs. Micron 5300 SATA or Micron 7300 NVMe SSDs can reach 7.68TB at a low cost per GB price.

For the following example, we use our 7.68TB Micron 5300 or Micron 7300 NVMe SSD and the 2RU server described in the last section. This server can host 24x Micron 7300 NVMe SSDs in front-facing U.2 drive bays. The all-SATA configuration will use a SATA HBA in one x8 PCIe 3.0/4.0 slot for drive management.

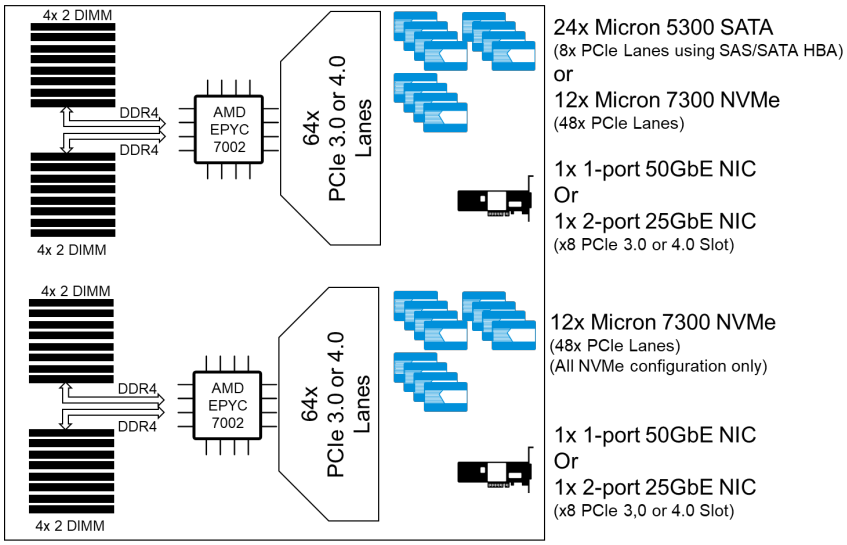

Figure 3: Capacity-Optimized Storage Node

Table 2 highlights the slot assignments for either an all-SATA or all-NVMe configuration, as illustrated in Figure 3:

Table 2: Server PCIe Slot Assignments for Capacity-Optimized Storage Server

I/O Slot Type	24x NVMe (front) + 2x SATA (rear)			24x SATA (front) + 2X SATA (rear)	
	Available Slots/Ports	Configured Slots		Available Slots/Ports	Configured Slots
x16 PCIe 4.0	3	3x slots 12x NVMe SSDs		2	1x slot 24x SATA SSDs
x16 PCIe 3.0	1(OCP)	1x OCP 3.0 4x NVMe SSDs		1(OCP)	-
x8 PCIe 4.0	4	2x slots 4x NVMe SSDs	2x slots 2x 25/50GbE NICs	5	2x slots 2x 25/50GbE NICs
x8 PCIe 3.0	1(OCP)	1x OCP 2.0 4x NVMe SSDs		1(OCP)	-
M.2 PCIe 3.0	1	1x M.2 Micron 5300 SATA boot drive		1	1x M.2 Micron 5300 boot drive
Board-to-Board (x4 PCIe 3.0)	4	2x sockets 2x SATA SSDs		4	1x socket 2x SATA SSDs

Storage totals for an all-NVMe configuration is 24x 7.68TB for a total raw capacity over 184TB in a single 2RU server. The all-SATA configuration can support 26x 7.68TB drives for a total raw capacity of 200TB in a single 2RU server.

Networking consists of either two single-port 50GbE NICs each installed in a x8 PCIe 4.0 slot or a dual-port 25GbE NIC. Either configuration provides a total of 100Gbps per node of network bandwidth.

I/O-Optimized Storage Solutions

Servers designed to focus on maximum I/O operations per node will require many high-IOP SSDs. I/O-optimized server configurations focus on maximizing IOPS per dollar spent. Therefore, I/O optimized solutions will prioritize high-performance NVMe SSDs. Micron 9300 NMVe SSDs can support up to 835,000 read and 310,000 write IOPs for small-block (4KB) block sizes.

For this example, we use the 6.4TB Micron 9300 MAX NVMe SSD and the 2RU all-NVMe server described earlier and illustrated in Figure 4.

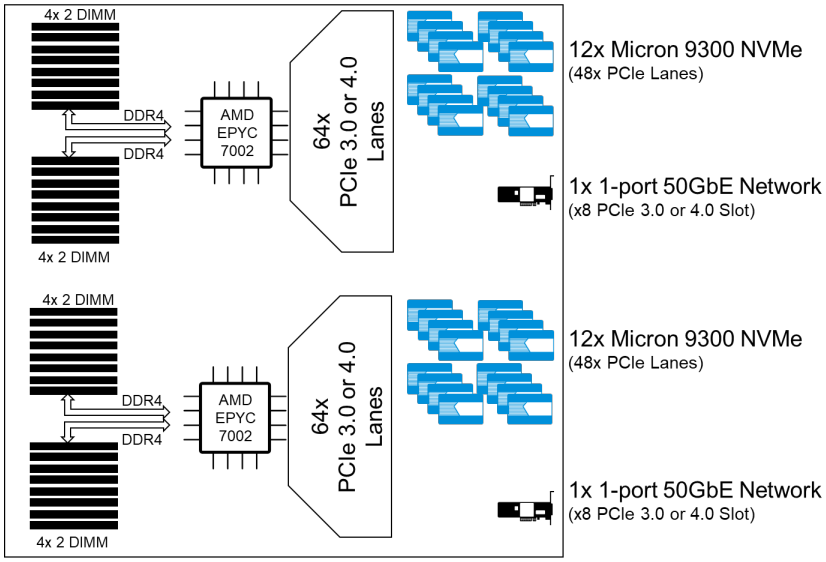

Figure 4: I/O-Optimized Storage Node

Storage totals for this configuration are 24 x 6.4TB for a total raw capacity of over 153TB in a single 2RU server. This server configuration has a theoretical performance potential of over 20 million read and over 3.6 million write IOPS per node.[2] Note that the actual application performance may be lower due to the operating system and data protection configuration used. Based on these theoretical values, this configuration should show impressive I/O performance for a single server.

Networking bandwidth consists of a single-port 50GbE NIC on each CPU complex, installed in a x8 PCIe 3.0/4.0 slot and enabling a total of 100 Gb/s per node of network bandwidth. This configuration is balanced such that each CPU complex hosts 12x SSDs and one 50GbE network interface. Network bandwidth is less critical for I/O-focused solutions configured to manage small block sizes and has proven not to be a restricting factor in many application workloads.

Table 3 describes how the PCIe slots and ports used to support 24x Micron 9300 NVMe SSD:

[2] Performance claim is based on individual drive-documented IOPS performance multiplied by the number of drives installed. Actual performance may vary due to software and/or hardware restrictions not foreseen in this discussion.

Table 3: Server PCIe Slot Assignments for Performance-Optimized Storage Server

I/O Slot Type	Available Slots/Ports	As Configured	
x16 PCIe 4.0	3	3x slots 12x NVMe SSDs	
x16 PCIe 3.0	1 (OCP 3.0)	1x slot 4x NVMe SSDs	
x8 PCIe 4.0	4	2x slots 2x 1-port 50GbE NICs	2x slots 4x NVMe SSDs
x8 PCIe 3.0	1 (OCP 2.0)	1x 2x NVMe SSDs	
M.2 PCIe 3.0	1	1x M.2 Micron 5300 SATA boot drive	
Board-to-Board (x4 PCIe 3.0)	2	2x sockets 2x NVMe SSDs	

Throughput-Optimized Solutions

Throughput-optimized solutions balance storage and network bandwidth.

PCIe 4.0 provides around 2 GB/s of throughput in each direction per lane (double that of PCIe 3.0). Using these numbers and understanding that each 100GbE network port can handle 12.5 GB/s into or out of the NIC, we can calculate the number of Gen 3 or Gen 4 PCIe lanes required to support each 100 GbE port (Table 4a).

Since server slots typically come in either 8-lane or 16-lane configurations, there will be some inefficiency in the required number of lanes and the slot configuration (1x 100GbE port requiring 13 PCIe 3.0 lanes or 7 PCIe 4.0 lanes). There may be some underutilized PCIe bandwidth on some occupied slots.

Each PCIe 3.0-compliant NVMe SSD can support up to 3.5 GB/s into or out of the drive. The Micron 9300 MAX SSDs can provide approximately 3.5 GB/s based on the available firmware as of this writing. Since the 9300 is a single-port 4-lane NVMe Gen 3 SSD, each SSD will use four PCIe 3.0 or 4.0 lanes (Table 4b). Therefore, each drive will require approximately 25Gb of network bandwidth to move data into or out of the server.

Tables 4a and 4b provides the ratio of 100GbE network ports to Micron 9300 NVMe SSDs:

Table 4a-b: Required PCIe Lanes for Each 100Gbe Network Port and Matching NVMe SSD Throughput

100GbE Ports	PCIe 3.0 Lanes for Networking	PCIe 4.0 Lanes for Networking
1	13	7
2	25	13
3	38	19
4	50	25

Number of Micron 9300 MAX NVME SSDs	PCIe 3.0/4.0 Lanes for Storage
4	12
8	24
12	36
16	68

Using the information shown in Table 4, we can now build a throughput-optimized configuration for varying amounts of available network bandwidth. For this paper, we focus on 100GbE as the incremental unit. The result is we can scale a host to 400 Gb/s of network I/O and support 16x NVMe SSDs in a well-balanced design where our storage and network I/O throughput is matched. The sample configurations are listed in Table 5.

Table 5: Server PCIe Slot Assignments for Throughput-Optimized Storage Server

I/O Slot Type	Available Slots/Ports	1x100GbE		2x100GbE		3x100GbE		4x100GbE	
x16 PCIe 4.0	3	1x slot 4x Micron 9300	1x slot 100GbE	2x slots 8x Micron 9300	1x slot 2x 100GbE Or 1x 200GbE	1x slot 4x Micron 9300	2x slots 3x 100GbE Or 1x200 + 1x100GbE	1x slot 4x Micron 9300	2x slots 4x 100GbE Or 2x 200GbE
x16 PCIe 3.0	1(OCP)	-		-		1x slot 4x Micron 9300		1x slot 4x Micron 9300	
x8 PCIe 4.0	4	-		-		2x slot 4x Micron 9300		4x slot 8x Micron 9300	
x8 PCIe 3.0	1(OCP)	-		-		-		-	
M.2 PCIe 3.0	1	1x M.2 Micron 5300 SATA boot drive							
BtB Sockets	4	-		-		-		-	

Consider Balanced Designs for Servers

Regardless of the CPU vendor or model, when considering a multiple-socket server for your storage solution, it is important to understand whether the board design is balanced across all CPU complexes. In some designs, products may distribute PCIe physical slots, NVMe ports and USB ports unevenly across multiple CPUs' complexes. For example, in some designs, vendors are putting a x16 PCIe slot on one CPU complex and two x8 PCIe slots on the other complex. In other designs, NVMe devices may directly attach to lanes on one CPU, but those same lanes on a different CPU attach to USB or SATA ports. These unbalanced designs could prevent you from supporting the desired configuration.

The market is starting to see the adoption of multiport 200GbE network adapters and PCIe 4.0-compliant NVMe interfaces, so continuing to focus on balanced server designs is still important.

For storage-centric solutions such as SDS and hyperconverged infrastructure (HCI), it is important that NVMe ports and x8 and x16 PCIe slots are evenly allocated across the multiple CPU sockets in the system. This layout will minimize the possibility of data travelling across CPU socket interconnects. It is essential to evaluate each OEM/ODM server design to ensure that it can support your desired solution requirements.

Conclusions

The AMD EPYC 7002 SoC solution has once again raised the stakes for x86 architecture. This is especially significant for storage-centric solutions such as SDS, HCI and AI. Supporting PCIe Generation 4.0, eight channels of 3200MT/s memory, and up to 64 cores in a single CPU provide unique opportunities for server OEMs and ODMs as well as storage vendors. Whether you are designing your storage solutions to focus on throughput, IOPS or capacity, EPYC 7002 should be part of your solution considerations. Micron is committed to providing the best storage solutions with high-performance Micron 9300 NVMe, mainstream Micron 7300 NVMe SSDs and Micron 5210 and 5300 SATA SSDs. These drives can meet a wide variety of needs across as many platforms as possible, and Micron has recently released a series of reference architectures built using EPYC 7002.

For More Information

To learn more about Micron products, visit micron.com/ssd. To learn more about AMD EPYC CPUs, visit amd.com.