AMD's EPYC™ Architecture Benefits Storage-Centric Solutions

Micron regularly evaluates new technologies and products to gain insight into their potential impact on storage in general and nonvolatile storage in particular. With trends like software-defined storage (SDS), NVM Express® (NVMe™) SSDs and the expanding bandwidth capabilities of Ethernet and Fibre Channel, customers have many options from which to meet their data center requirements.

With Micron 9200 NVMe SSDs, a server can support up to 264TB of capacity in a single rack unit design!

To best assess any technology, we need understand the impact on current and future solutions. These evaluations help us provide the knowledge you need to plan and the solutions you need to deploy within the solid-state data center.

AMD's recently announced EPYC™ architecture is one of the technologies that shows great promise to move server architecture forward. Whether you want to deploy an SDS solution, a data-hungry database application or a compute-heavy application server, effective and efficient data storage is always important. We think that the EPYC-based server architectures now being offered by various original equipment manufacturers (OEMs) and original device manufacturers (ODMs) are compelling options.

This whitepaper discusses EPYC architecture from the point of view of solid-state storage solutions and is based solely on AMD's documentation and claims. We have not included any specific testing performed by Micron due to the flexibility OEMs and ODMs have in server implementation and use of additional support components, which could ultimately affect solution performance. Because of this, we'll focus our discussion on the currently available EPYC products: the single- and dual-socket reference architecture demonstrated by AMD.

Architecture Overview

In June of 2017, AMD released its new EPYC line of enterprise-class server processors based on its Zen micro architecture. Using its Infinity Fabric to interconnect up to four system-on-chip (SoC) die, each with 8 cores within a single CPU package, AMD has built a highly capable solution that truly scales. The result is a 32-core, single-socket solution for enterprise platforms. AMD designed Infinity Fabric to minimize core-to-core communication and, with current 2 socket implementations, there are no more than 2 hops between any two processor cores.

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

- 8 memory channels with 2 DIMMs per channel
- 128 PCIe 3.0 lanes for both the single- and dual-socket solutions
- Native SATA, SATA Express controllers on-chip

Memory Architecture

Current EPYC implementations support up to 16 DIMMs (RDIMM, LRDIMM and NVDIMM) per CPU providing linear scalability as the socket count increases. Reference designs from AMD and products from several ODMs and OEMs support single- and dual-socket designs for a total of up to 32 DIMMs per server.

This enables immense memory capacity. Based on AMD's documented specifications, a common example server solution could use Micron's 32GB RDIMMs — supporting up to 512GB of RAM in a single-socket solution; AMD claims support for a maximum of 2TB per socket.

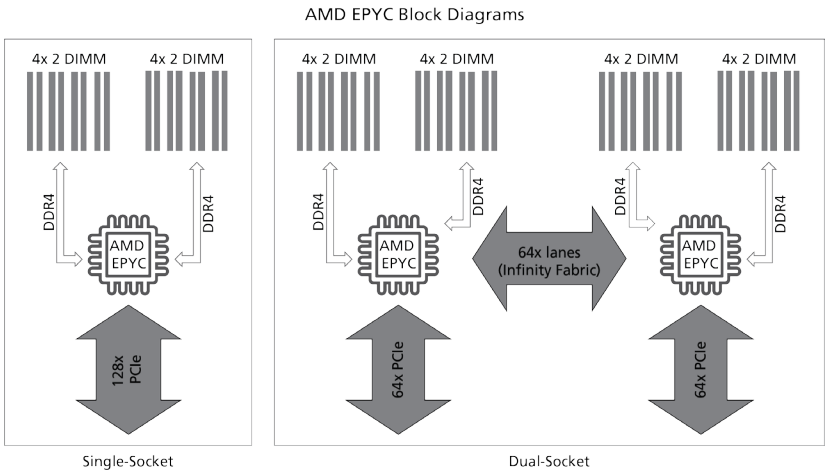

Figure 1: AMD EPYC High-level Architecture

Based on this design, the architecture may support over 21 GB/s of DRAM bandwidth per channel for a total of over 170 GB/s total memory bandwidth per socket. For dual-socket solutions, double these specifications.[1]

AMD indicates that EPYC supports NVDIMM-N technology, allowing you to use storage-class memory that is well-suited for deploying high-performance, nonvolatile OLTP solutions (NVDIMM-based storage is a great solution for transaction log acceleration). Micron's currently available 32GB NVDIMM-Ns enable up to 64GB per memory channel. Depending on actual OEM/ODM EPYC designs (your maximum NVDIMM capacity may vary), using only four of the available eight channels per socket for NVDIMM storage could achieve 256GB per socket of nonvolatile, memory-speed storage per socket.

128 PCIe Lanes

EPYC provides support for 128 PCIe lanes per socket. That is 3X the number of PCIe lanes from previous generation architectures. These lanes will primarily be used for PCIe connections, but depending on the OEM/ODM requirements, up to 32 lanes could be reconfigured to support native SATA interfaces, removing the requirement for an external SATA controller. A single-socket configuration has all 128 lanes available for these peripheral devices/interfaces; a dual-socket architecture requires that 64 lanes of each socket be designated for CPU-to-CPU interconnect running AMD's Infinity Fabric; the remaining 64 PCIe lanes per socket can be used for peripheral devices/interfaces (refer to Figure 1).

What Does This Mean for Storage?

It's All About PCIe Lanes

PCIe lanes mean options. It means we finally have the expandability we need without using cumbersome, expensive PCIe switches or expanders. Both single- and dual-socket designs have a total of 128 PCIe lanes, enabling far greater platform flexibility than previously available. The storage impact of this many lanes is clear: based on the EPYC architecture specifications, OEMs now have more options when designing a compute or storage platform.

Compute-centric solutions require an architecture where we can deploy more high-performance NVMe SSDs to feed applications with the data they need. The benefit is more efficient operations and deployments that could potentially use fewer nodes while providing similar capability.

[1] Based on AMD publicly available specification: http://www.amd.com/en/products/epyc-7000-series.

Storage-centric solutions require a balance between network I/O and fast SSDs, ensuring that data can be moved from the storage array to the application servers. The expanded lane count of these solutions allows potentially more flexibility in OEMs designing solutions that balance network access bandwidth with storage access bandwidth and do so sufficiently to enable high-performance solutions.

AMD EPYC provides 128 PCIe lanes, giving you more solution design options.

Server architectures prior to EPYC limited us to four to six NVMe SSDs mainly due to the limited number of free PCIe lanes available without resorting to the use of PCIe switches or expanders. With EPYC that is less of a concern. Based on the documented architecture specifications, and assuming a few PCIe lanes need to be reserved for embedded peripheral interfaces (USB, SATA and the like — all varying by ODM/OEM), we can expect as many as 110 available PCIe lanes for use for PCIe-native NVMe and/or SATA SSDs and high bandwidth x8 and x16 PCIe I/O slots. AMD's own dual-socket reference design motherboard has 112 PCIe lanes available.[2]

Assuming a hypothetical 60%/40% split between storage and PCIe slot usage of those lanes, there could be around 64 lanes available for NVMe devices. With each NVMe device using 4 PCIe lanes, the solution would have 16 NVMe SSDs per single- or dual-socket server without the use of slow PCIe switches or expanders getting in the way. Using higher storage-to-PCIe slot ratios, some OEM/ODM designs support up to 24 NVMe SSDs connected directly to the CPU complex, though these designs offer fewer physical PCIe slots. Alternatively, 32 of these lanes could be converted to dedicated SATA ports using EPYC's on-chip SATA controller, resulting in 32 SATA drives without any extra controllers — saving costs and eliminating PCIe controller-based bottlenecks.

Network Bandwidth to Support Maximum I/O

More PCIe lanes also means potentially more networking interconnects and more available bandwidth.

While network bandwidth is important to both compute-centric and storage-centric solutions, it is doubly important for storage-centric SDS solutions because they must be able to move large amounts of data to external application servers via high-bandwidth networks. The more network interfaces we can use or the higher bandwidth those networks can transmit, the more data we can move between the storage solution and the application servers.

The EPYC architecture has so many available PCIe lanes it allows the OEM/ODM to create solutions that provide more network bandwidth to move data to and from all those SSDs as PCIe lanes directly translate to higher data throughput. A single 100Gb Ethernet port consumes 16 PCIe lanes (the same number of lanes required for four NVMe SSDs) and many current EPYC designs provide multiple x16 PCIe slots, each supporting 100 Gb/s of throughput.

Storage Design Options

Capacity-Focused Solutions

Capacity-focused solutions are designed around a large number of high-capacity SSDs. (Micron's latest 5100 ECO SSDs can reach up to 8TB and 9200 ECO NVMe SSDs can reach 11TB).

For the following example, we'll use our 8TB 5100 SATA SSD and a two rack unit (2RU) server. AMD's EPYC-based solutions can support up to 32 dedicated SATA ports, though the actual number of SSDs implemented in

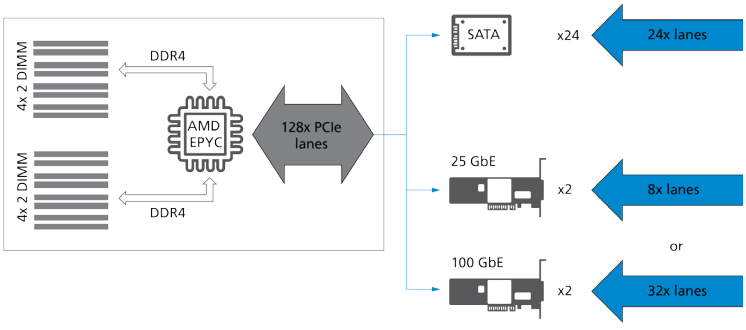

Figure 2: Capacity-Focused Design

[2] Slide 18 of Open Compute Project "Microsoft Project Olympus Server" presentation.

an OEM/ODM solution may be fewer due to design choices or the physical constraints of the chassis.

Current 2RU server designs limit the maximum number of front-panel accessible SSDs to around 24 U.2/2.5-inch drives, but custom-designed solutions can support up to 32 SATA SSDs (with a total capacity of 256TB).

Based on AMD's claims for EPYC and our hypothetical 2RU server platform, we can create a solution that supports 24x 2.5-inch Micron 5100 ECO SATA SSDs at 8TB per drive. The resulting solution provides up to 192TB of storage capacity. Each of these drives has direct access to the EPYC CPU complex and does not have to traverse an external PCIe SATA controller. Alternatively, we can use NVMe SSDs, such as the Micron 9200 ECO at 11TB, that will yield a potential total of 264TB in that same 2RU design. Using solutions from some OEMs,[3] this capacity can be attained using a 1RU server!

IOPS-Focused Solutions

When maximizing input/output operations per second (IOPS) is your goal, NVMe SSDs and storage-class memory (SCM), such as Micron's NVDIMM-Ns, should be the solution considered. Two-tiered storage solutions using NVDIMM-N as a cache tier and NVMe SSDs for a capacity tier can provide extremely high IOPS performance.

Dual-socket designs offer an additional advantage when using NVDIMM-N by providing access to twice the memory slots of the single-socket EPYC solutions.

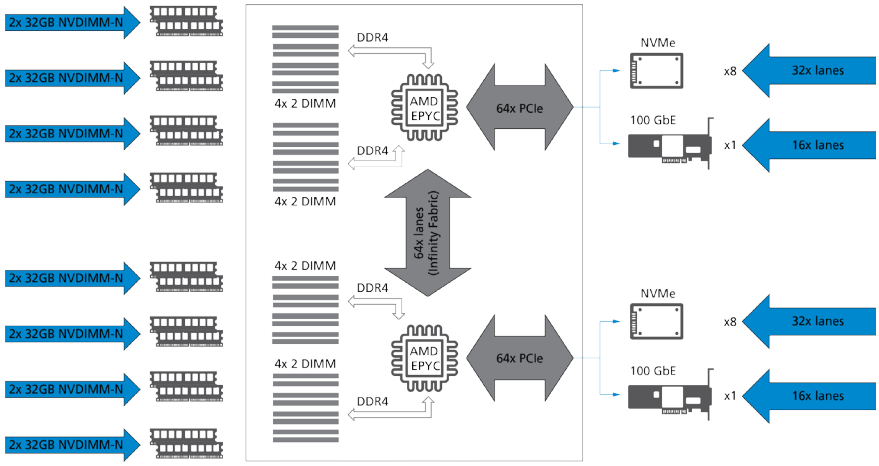

Figure 3: Hypothetical High IOPS-Focused Server Deployment

EPYC's design benefits are clear: greater numbers of high-performance devices in each server means more IOPS. 128 PCIe lanes and 8 memory channels per socket enable you to create large scale IOPS-focused storage solutions in compact enclosures.

OEM/ODM 2RU single- or dual-socket products using EPYC currently support from 10-24 NVMe devices without using PCIe switches. While the maximum number of NVDIMMs in OEM/ODM products will vary by OEM/ODM, one can imagine products that support NVDIMMs in half the available memory channels. To support high-IOPS solutions, sufficient network bandwidth must also be available.

For example, using a hypothetical dual-socket design (refer to Figure) and assigning half of the memory channels to NVDIMMs would result in 512GB of SCM when using Micron's 32GB NVDIMM-N. To supply sufficient IOPS, our solution would use dual 100 GbE network ports, which will require 32 PCIe lanes. Using our 60%/40% drive-to-PCIe slot ratio from our previous example, we could support up to 16 NVMe devices.

The resulting IOPS performance of this example is impressive. If each 32GB NVDIMM-N is capable of around 1.1 million IOPS.[4] Using 16x 32GB NVDIMM-Ns in our example, we have the potential to reach 17 million IOPS in the NVDIMM-N layer alone. When added to the 16 NVMe SSD using 11TB Micron 9200 SSDs (800,000 read IOPS per SSD),[5] we can produce an additional 9.6 million IOPs at the NVMe layer. The combined NVDIMM-N and NVMe IOPS exceed 26 million in 2RU! Real-world performance is, of course, application, workload and file system dependent.

[3] For example, HPE CL3150-Gen4 can support 24x NVMe devices in a 1RU design. HPE has not qualified the Micron 9200 NVMe device for use in their specific products.

[4] Micron Solution Brief "HPE, Micron, and Microsoft Windows Server 2016/NVDIM-N Solution."

[5] Micron 9200 data sheet

Throughput-Maximized Solutions

Throughput-maximized solutions add a dependency on network bandwidth and are typically focused on disaggregated, storage-centric solutions rather than scale-out, compute-centric application solutions. For this reason, storage-centric SDS solutions should be designed with an equal amount of network and SSD bandwidth.

Each 100Gb network port requires 16x PCIe lanes for full bandwidth, which is equivalent to 4 NVMe SSDs or 16 SATA SSDs.

Based on current OEM/ODM EPYC designs, most motherboards will have at least two x16 PCIe slots and up to 24 NVMe drive slots, but by only supporting two 100Gb Ethernet ports, each requiring a x16 slot, we can only support 8 NVMe drives before we create a bottleneck at the network interface. Figure 4 shows this design example. This solution leaves up to 64x PCIe lanes unused. Custom motherboard designs could easily improve this to supporting up to 12x NVMe SSDs and 3x 100 GbE network interfaces and still have available PCIe lanes for ancillary peripherals.

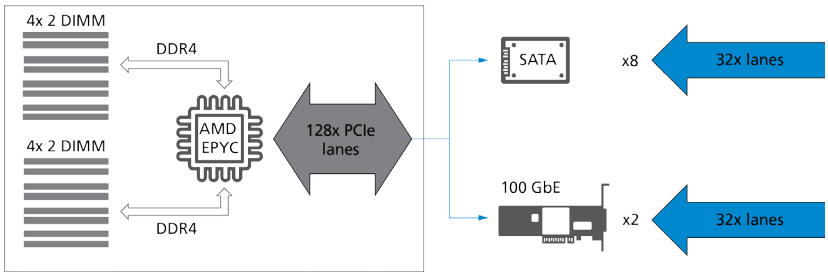

Figure 4: Throughput-Maximized Solution

Balanced Designs for Multi-Socket Server Solutions

Regardless of CPU vendor or model, when considering a multiple socket server for your storage solution, it is important to consider how the board design is balanced across all CPU complexes. In many designs, OEM/ODM products will distribute PCIe physical slots, NVMe ports and USB ports, each of which will unevenly consume a CPU's PCIe lane capacity. For example, putting a x16 slot on one CPU complex, but only putting two x8 slots on the other. This will prevent you from supporting multiple 100Gb Ethernet ports because each single-port 100Gb network interface card (NIC) will require a x16 slot. In other designs, NVMe devices may be attached to one CPU, but those same lanes on a different CPU will be dedicated to USB or SATA ports.

For storage-centric solutions such as SDS, it is very important that the allocation of NVMe and x8 and x16 PCIe slots be evenly distributed across the multiple CPU sockets in the system. This will minimize the possibility of data having to travel across CPU socket interconnects. While the AMD EPYC SoC provides the ability to create balanced two-socket solutions, it is important to evaluate each OEM/ODM product for how balanced the PCIe lane usage is presented. A balanced EPYC solution will offer a x16 PCIe port on each CPU as well as the same number of NVMe connections on each CPU.

Conclusions

AMD's EPYC System on Chip solution has raised the stakes for x86 architecture. This is especially impactful for storage centric solutions. Based on the AMD reference architecture and early products from OEMs and ODMs, the ability to support up to 20+ NVMe or up to 32 SATA devices using an on-CPU controller, along with support for 128 PCIe lanes – 3x the number of lanes available in past x86 architecture products – is a potential game changer. Support for 8 channels of memory per socket, that can support traditional DRAM or NVDIMM-N modules supports high-performance, low-latency storage capabilities as well as hybrid SCM/SSD two-tiered storage implementations provides a lot of flexibility for system designers. Based on these specifications, we believe that further, in-depth analysis is warranted. Whether you are designing your storage solutions to focus on throughput, IOPS, or simply capacity, EPYC should be part of your solution considerations. Micron is committed to providing the best storage solutions that meet a wide variety of needs across as many platforms as possible and we will be investigating AMD's EPYC architecture potential further.

micron.com