PowerEdge NUMA Configurations with AMD EPYC Processors

Tech Note by:
Jose Grande

SUMMARY
With the introduction of AMD's EPYC (Naples) x86 Server CPUs featuring four Zeppelin dies per package, there is a need to clarify how AMD's new silicon design establishes Non-Uniform Memory Access (NUMA) domains across dies and sockets.

This tech note explains how Dell EMC PowerEdge Servers leverage AMD's EPYC CPUs to configure NUMA domains for optimal performance by using Dell EMC BIOS Settings.

AMD EPYC is a Multi-Chip Module (MCM) processor and per silicon package there are four Zeppelin SOCs/dies leveraged from AMD Ryzen. Each of the four dies have direct Infinity Fabric connections to each of the other dies as well as a possible socket-to-socket interconnect. This design allows, at most, four NUMA nodes per socket or eight NUMA nodes in a dual-socket system.

AMD EPYC's four dies each have two Unified Memory Controllers (UMC), that each control one DDR channel with two DIMMs per channel, along with one controller for IO, as shown in Figure 1 below:

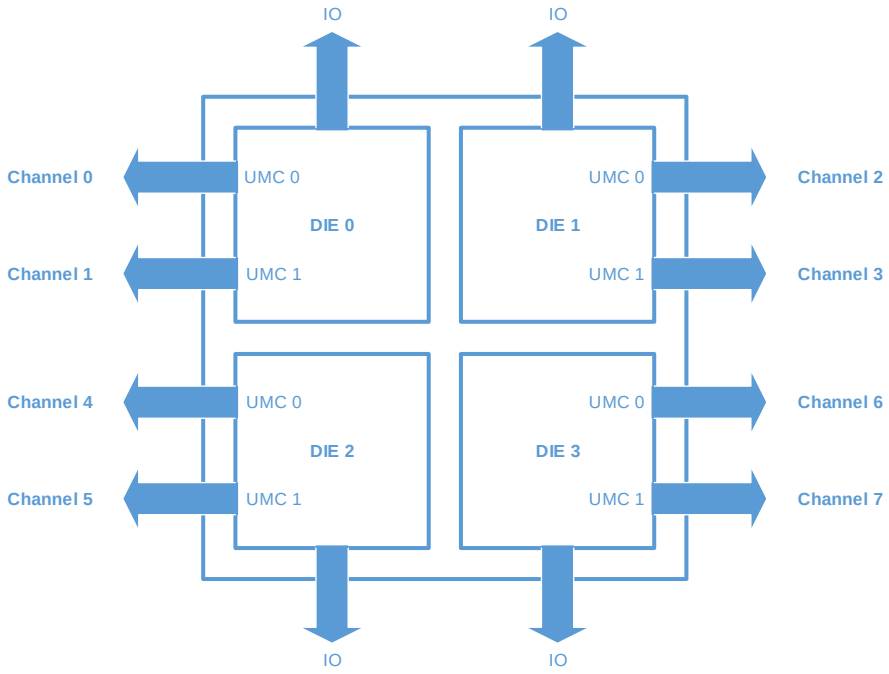

Figure 1: Die layout of the AMD EPYC processor

Memory Interleaving Options and Rules

AMDs EPYC processor supports four memory interleaving options:

- Socket Interleaving (2 processor configurations)
- Die Interleaving
- Channel Interleaving
- Memory Interleaving disabled

The following are the rules for each memory interleave options:

- The system can socket interleave, but only if all channels in the entire system have the same amount of memory. Die interleaving must be enabled as well.
- The system can die interleave, but only if all channels on the socket have the same amount of memory. Channel interleaving must be enabled as well.
- The system can channel interleave as long as both channels have at least one DIMM. The channels do not have to be symmetrical. This is the default configuration.
- No interleave at all, where each channel is stacked on top of the previous channel. However, it should be noted that probe filter performance may be affected if there is one UMC with less memory than the other UMC on the same die.

NUMA Domains per Memory Interleave Option

AMD's new silicon architecture adds nuances on how to configure platforms for NUMA. The focus of AMDs scheme to NUMA lies within its quad-die layout and its potential to have four NUMA domains.

Socket interleaving is memory interleave option meant only for inter-socket memory interleaving, and is only available with 2-processor configurations. In this configuration memory across both sockets will be seen as a single memory domain producing a non-NUMA configuration.

Die interleaving is available for all configurations. Die interleaving is the intra-socket memory interleave option that creates one NUMA domain for all the four dies on socket. In a 2-processor configuration, this will produce two NUMA domains, one domain pertaining to each socket, providing customers with the first option for NUMA configuration. In a one socket platform, die interleaving will be the maximum option of memory interleaving, and will produce one memory domain thus producing a non-NUMA configuration.

Channel interleaving is also available with all configurations. Channel interleaving is the intra-die memory interleave option and is the default setting for Dell EMC platforms. With channel interleaving the memory behind each UMC will be interleaved and seen as one NUMA domain per die. This will generated with four NUMA domains per socket.

Memory interleaving disabled - When memory interleaving is disabled, four NUMA nodes will be seen as in the case for channel interleaving. In this case however, the memory will not be interleaved but rather stacked next to one another.

Number of Processors	Socket Interleave	Die Interleave	Channel Interleave
2	1	2	8
1	NA	1	4

Figure 2: NUMA Domain count per Memory Interleave Option

Performance Tuning

For best performance from AMD EPYC processors, it is recommended that each die have one DIMM populated in each channel. This allows all IO behind each die to access memory, with optimal latency.

Memory DIMM Population Guidelines

- Populate empty channels, with the same type/capacity of DIMMs, before populating two DIMMs on a given channel
- Recommendations for best performance:
  - 1 DIMM per channel dedicates full memory bandwidth
  - Populating 2 DIMMs per channel will increase capacity but will lower the clock speed, resulting in lower memory bandwidth. There is a dependency between memory speed and the bandwidth of the Infinity Fabric.

Memory Bus Speed	Infinity Fabric Speed (Die to Die in same CPU socket)	Infinity Fabric Speed (Socket-to-socket)
2666 MT/s	5.3 GT/s	10.6 GT/s
2400 MT/s	4.8 GT/s	9.6 GT/s
2133 MT/s	4.2 GT/s	8.5 GT/s
1866 MT/s	3.7 GT/s	7.4 GT/s

Figure 3: Memory Bus Speed to Infinity Fabric Bus Speed

- Minimum recommended:
  - At least 1 DIMM is per die in the system for a total of 4 DIMM per CPU
- On Dell EMC platforms populate DIMM 1 first (white slots in Figure 4, below)
- A 2 socket system (2 CPUs are populated) will need equivalent memory configurations on both CPUs for optimal performance.

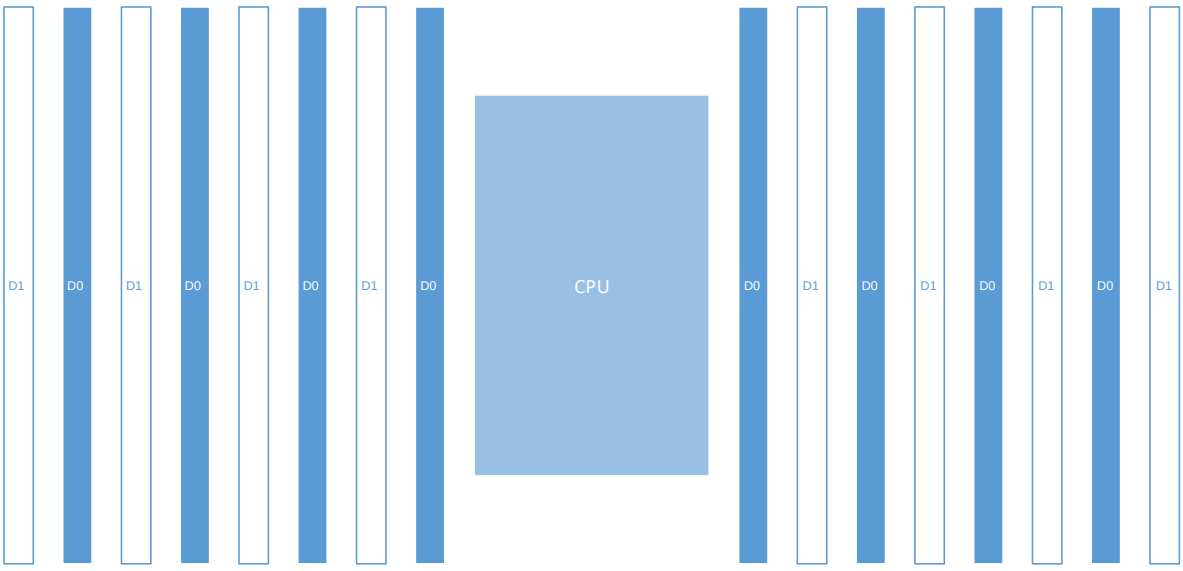

Figure 4: DIMM layout of AMD EPYC processors

PCIe Configuration Guidelines

- When PCIe cards are populated into particular slots with NUMA-unaware application/software, make sure to have memory DIMMs populated in the corresponding NUMA-node mapping as local memory. Mappings can be found in Section 4 of Platform Specific NUMA/Die Domain Details
- Consider also pinning the interrupts to local CPUs to get maximum performance. For instructions on how to tune network cards for better performance on AMD EPYC processors, access the following links and download provided documentation:
  - https://support.amd.com/TechDocs/56224.pdf
  - https://developer.amd.com/resources/epyc-resources/epyc-white-papers/

BIOS Setup

The "Memory Interleaving" setting controls whether the system is configured for Socket, Die, or Channel interleaving. In System Setup (F2 prompt during system boot), enter System BIOS > Memory Settings, and navigate to "Memory Interleaving" to choose the memory interleave for desired configuration. This option is also available in system management consoles such as RACADM.

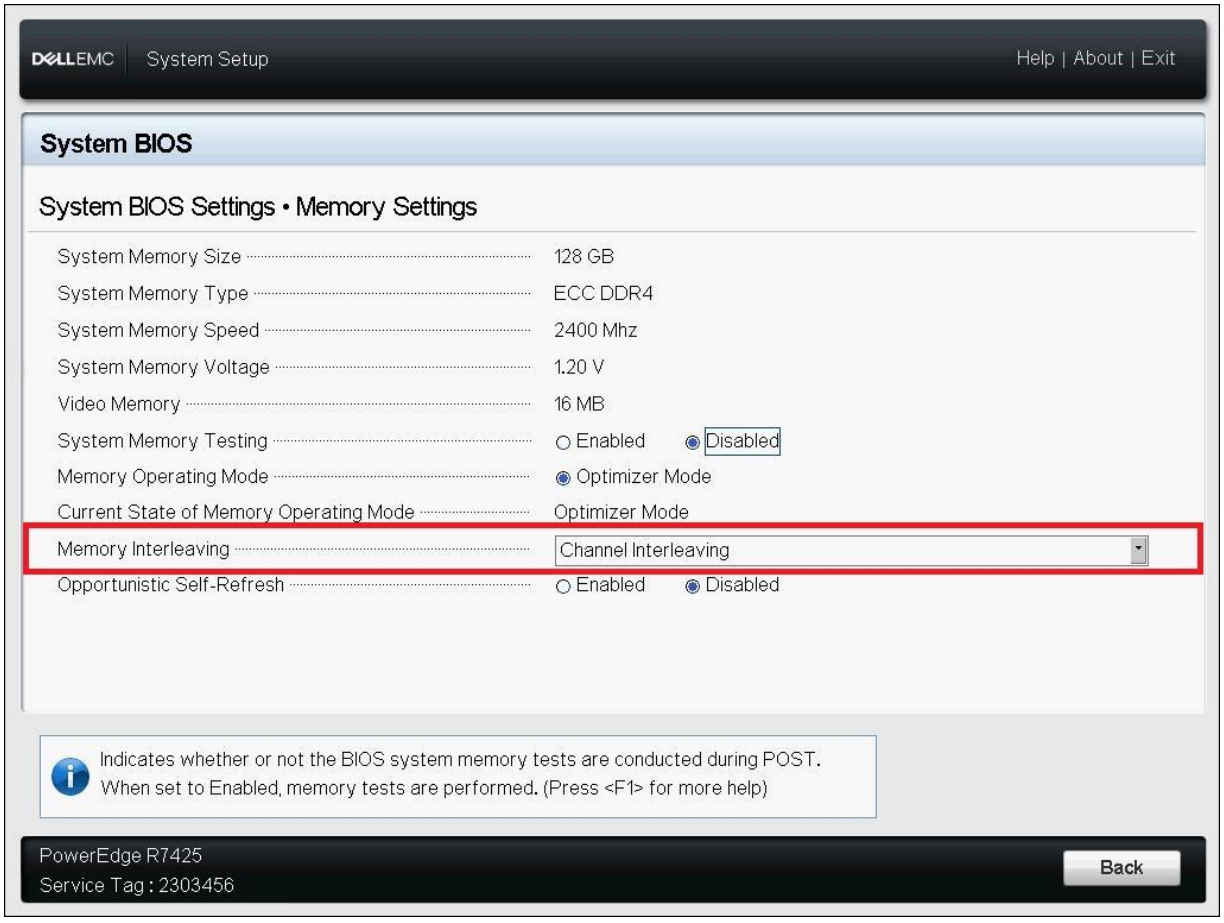

Figure 5: Using the BIOS Setup screen to select Memory Interleaving option

Platform-specific NUMA/Die Domain Details

The following matrices shows how CPU die, memory and PCIe slots are physically grouped to each NUMA domain for Dell EMC PowerEdge EPYC-based platforms, PowerEdge R6415, R7415, and R7425.

PowerEdge R7415 (Sys configuration *with* rear side disks)			
PCIe Slot / Device	**CPU Die**	**Memory Slots**	**NUMA Node**
Slot1	2	A7, A8, A15, A16	2
Slot2	0	A3, A4, A11, A12	0
Slot3	2	A7, A8, A15, A16	2
Embedded LOM	2	A7, A8, A15, A16	2
Mini PERC	3	A5, A6, A13, A14	3

CONCLUSIONS

Dell EMC PowerEdge Servers with the new AMD EPYC processors can deliver significant performance improvements for key applications and workloads. By following the guidelines above and implementing the desired NUMA configuration with the PowerEdge system BIOS, optimal performance can be attained.

For further information, please see:

Dell TechCenter, at https://www.dell.com/community/Dell-Community/ct-p/English is an online technical community where IT professionals have access to numerous resources for Dell EMC software, hardware and services.

https://community.amd.com/community/server-gurus EPYC Server Community Forum

https://developer.amd.com/resources/epyc-resources/epyc-white-papers/ Linux Network Tuning Guide for AMD EPYC Processor Based Servers