Dell EMC NUMA Configuration for AMD EPYC (Naples) Processors

Dell Engineering
February 2018

A Dell EMC Deployment and Configuration Guide

Revisions

Date	Description
February 2018	Initial release

Table of contents

Executive 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.

The goal of this Dell EMC Deployment and Configuration Guide is demonstrate how Dell EMC Servers leverages AMD's EPYC CPUs to configured NUMA domains for optimal performance by using Dell EMC BIOS Settings.

1 AMD EPYC Architecture

1.1 Zeppelin Die Layout

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 sockets system

AMD EPYC processor'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:

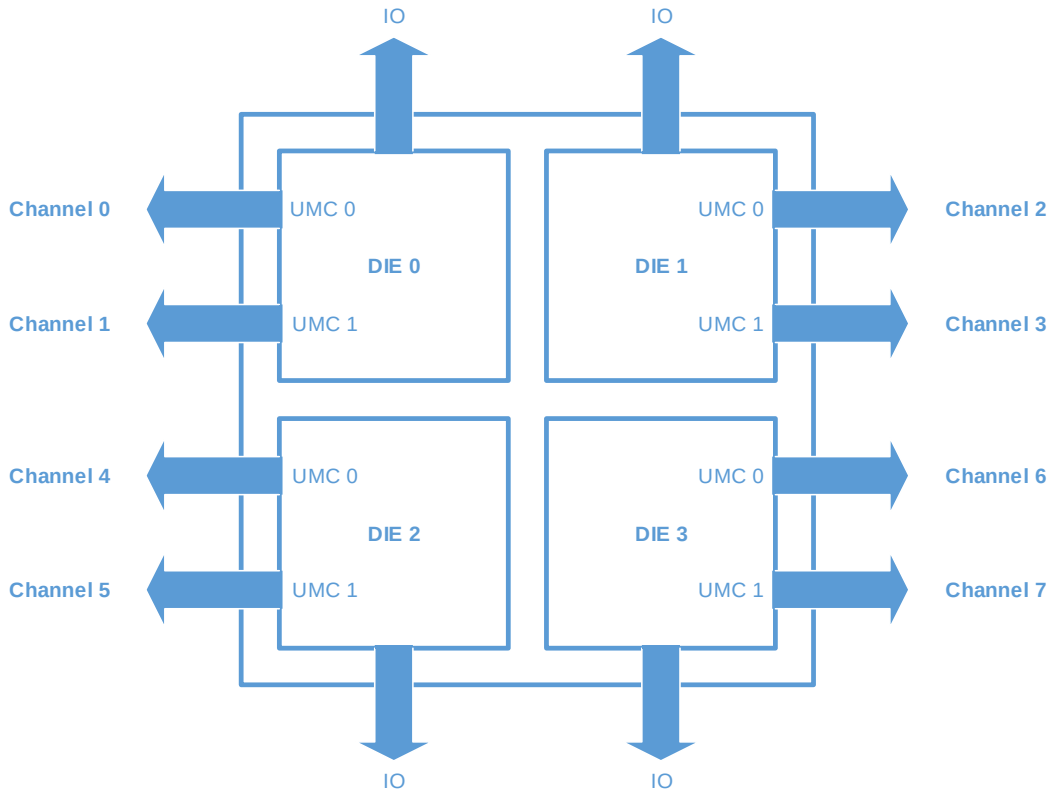


Figure 1 Figure 1 Zeppelin Die Layout

1.2 Memory Interleaving

The Memory Interleave feature for AMD EPYC processors is what controls how may NUMA domains are generated. AMD EPYC processors support 4 memory interleaving options. Each option becomes available based on system configuration.

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

1.2.1 Memory Interleaving Rules

The following are the rules for each memory interleave option:

- 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.

1.2.2 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 the only memory interleave option meant 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 the intra-socket memory interleave option that creates one NUMA domain for all the four dies on a 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 1-processor configuration die interleaving will be the maximum option for memory interleaving, and will produce one memory domain thus producing a non-NUMA configuration.

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 1 NUMA domain per die. This will generate four NUMA domains per socket.

Memory Interleave disabled - When memory interleave is disable 4 NUMA nodes will be seen as in the case for channel interleaving but the memory will not be interleaved yet stacked next to one another.

Figure 2 NUMA Domain Count per Memory Interleave Option

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

2 Performance Tuning

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

2.1 Memory DIMM Population Guidelines

- Populate empty channels, with the same type/capacity of DIMMs, before populating 2 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

Figure 3 Memory Bus Speed to Infinity Fabric Bud Speed

Memory Bus Speed	Infinity Fabric Speed (Die to Die within 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

- 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.

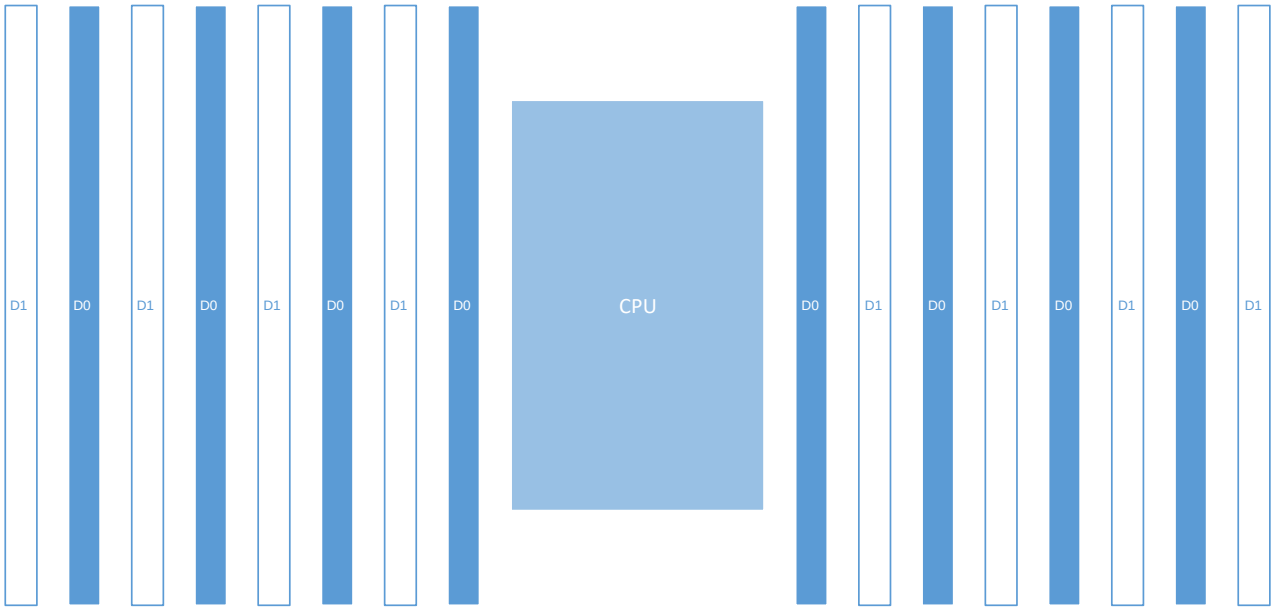


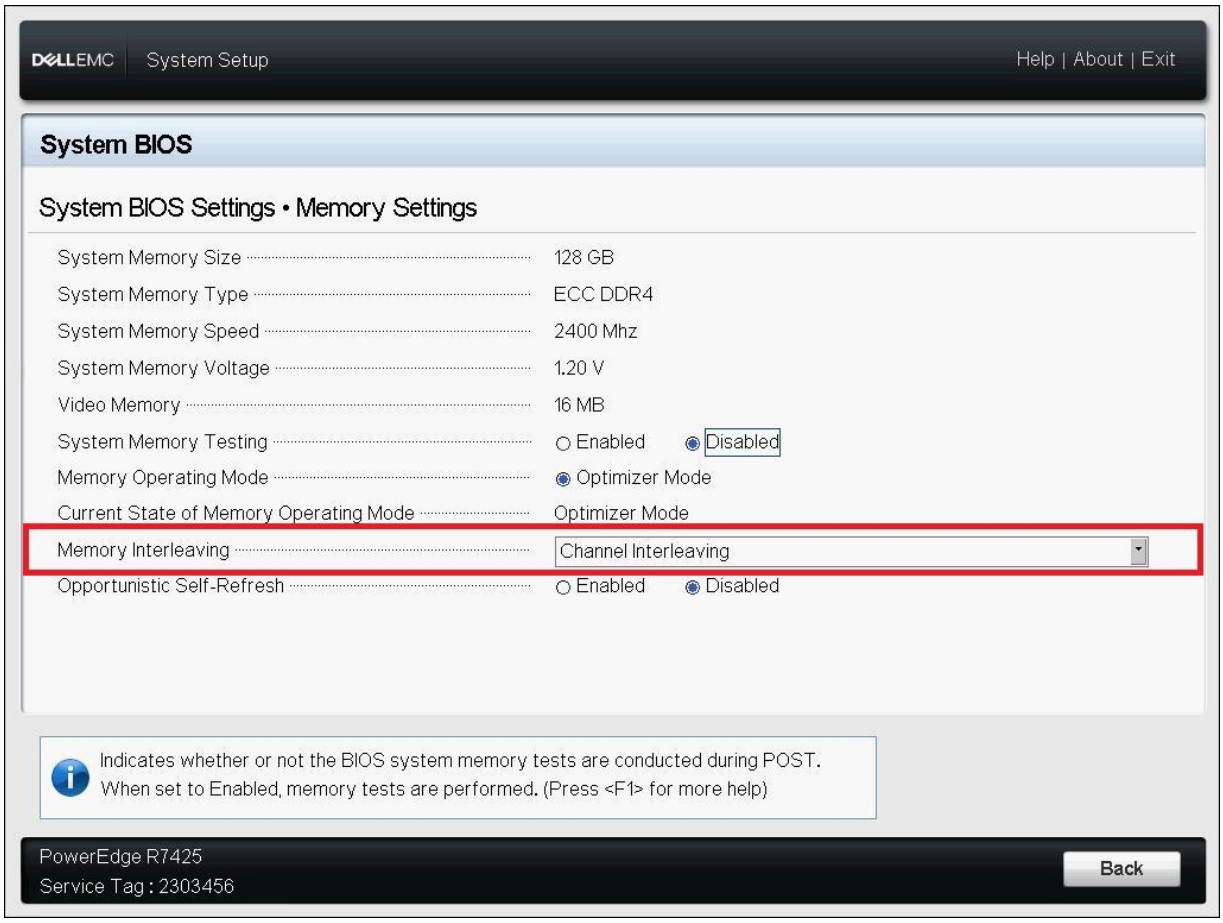
Figure 4 DIMM Layout

2.2 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 Platform Specific NUMA/Die Domain Details
- Considering 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, go to the following links and download provided documentation:
  - https://support.amd.com/TechDocs/56224.pdf
  - https://developer.amd.com/resources/epyc-resources/epyc-white-papers/

3 BIOS Setup

The “Memory Interleaving” setting controls whether the system is configured for Socket, Die, 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.



4 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 EPYC based platforms, PowerEdge R6415, R7415, and R7425.

PowerEdge R6415			
PCIe Slot / Device	**CPU Die**	**Memory Slots**	**NUMA Domain**
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

PowerEdge R7415 (Sys config without rear side disk)			
PCIe Slot / Device	**CPU Die**	**Memory Slots**	**NUMA Domain**
Slot1	2	A7, A8, A15, A16	2
Slot2	1	A1, A2, A9, A10	1
Slot3	0	A3, A4, A11, A12	0
Slot4	2	A7, A8, A15, A16	2
Slot5	3	A5, A6, A13, A14	3
Embedded LOM	2	A7, A8, A15, A16	2
Mini PERC	3	A5, A6, A13, A14	3

PowerEdge R7415 (Sys config. with rear side disk)			
PCIe Slot / Device	**CPU Die**	**Memory Slots**	**NUMA Domain**
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

PowerEdge R7425					
PCIe Riser	**PCIe Slot / Device**	**CPU SKT**	**CPU Die**	**Memory Slots**	**NUMA Domain**
1A	1	1	2	A7, A8, A15, A16	2
	3	1	3	A5, A6, A13, A14	3
1D	1	1	2	A7, A8, A15, A16	2
	2	1	3	A5, A6, A13, A14	3
	3	1	3	A5, A6, A13, A14	3
1E	1	1	2	A7, A8, A15, A16	2
	2	1	3	A5, A6, A13, A14	3
2A	4	2	2	B7, B8, B15, B16	6
	6	2	3	B5, B6, B13, B14	7
2C	4	2	2	B7, B8, B15, B16	6
2D	4	1	1	A1, A2, A9, A10	1
	5	2	2	B7, B8, B15, B16	6
	6	2	3	B5, B6, B13, B14	7
3A	7	2	3	B5, B6, B13, B14	7
	8	2	0	B3, B4, B11, B12	4
3B	7	2	1	B1, B2, B9, B10	5
	8	2	0	B3, B4, B11, B12	4

5 Technical support and resources

5.1 Dell

Dell.com/support is focused on meeting customer needs with proven services and support.

Dell TechCenter is an online technical community where IT professionals have access to numerous resources for Dell EMC software, hardware and services.

Storage Solutions Technical Documents on Dell TechCenter provide expertise that helps to ensure customer success on Dell EMC Storage platforms.

5.2 AMD

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

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