FPGA Based 400 GBit/s Data Recorder - Insight into different pitfalls and design choices

Andreas Schuler
David Epping

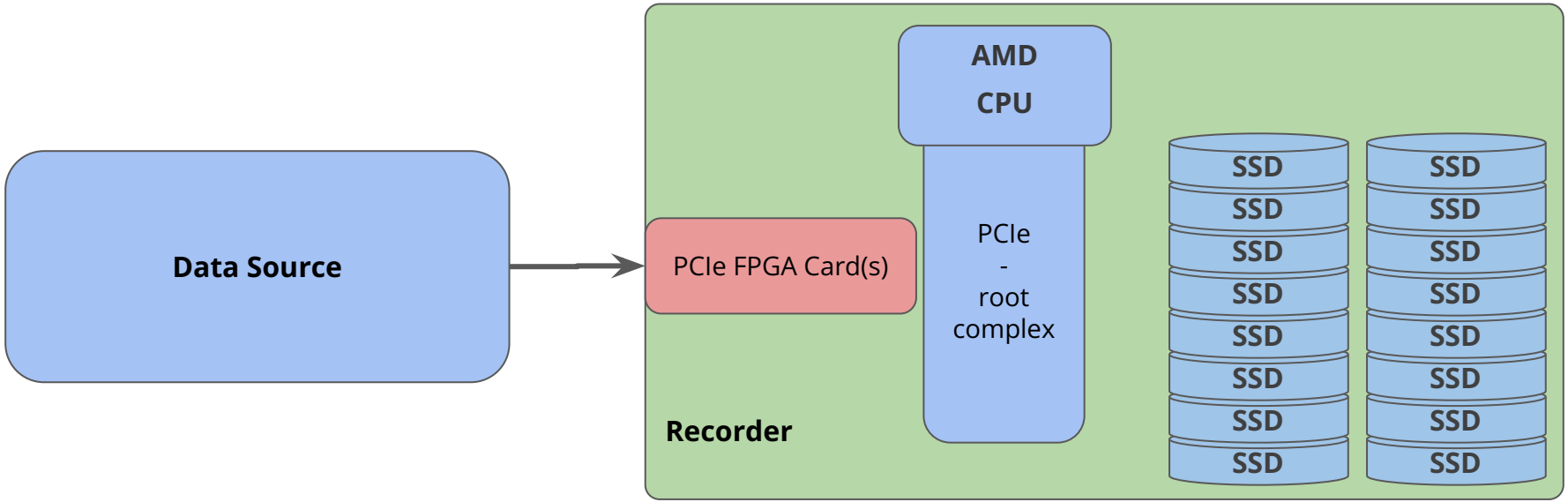
Agenda

- System Overview and requirements
  - Recording Modes
  - Hardware Architecture and Components
- SSD
  - Requirements
  - Intro NVMe
  - RAID Scaling Limitations
  - NVMe Streamer
  - Crossbar and PCIe Bridge
- Data transfer
  - Peer to peer communication
  - CPU Architecture

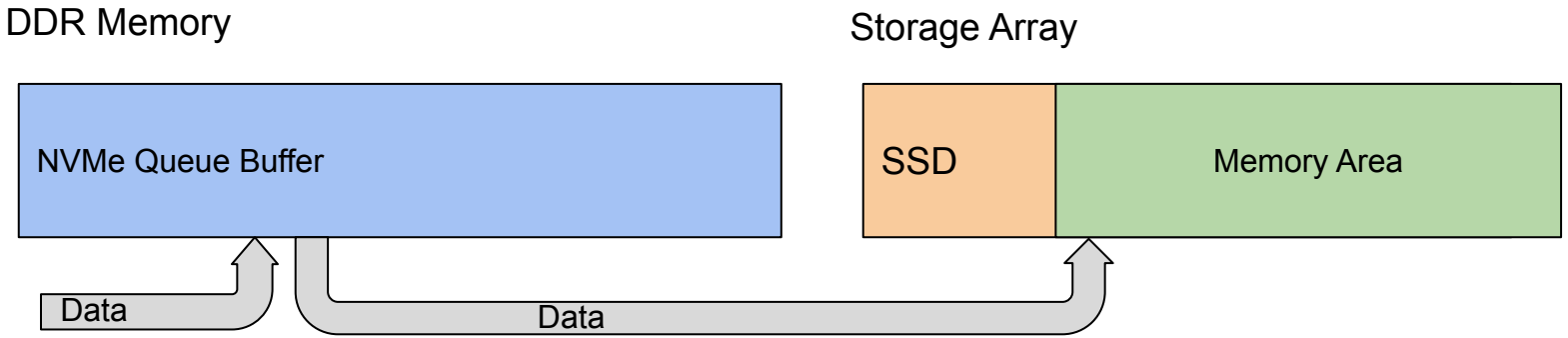
Agenda

- **System Overview and requirements**
  - Recording Modes
  - Hardware Architecture and Components
- SSD
  - Requirements
  - Intro NVMe
  - RAID Scaling Limitations
  - NVMe Streamer
  - Crossbar and PCIe Bridge
- Data transfer
  - Peer to peer communication
  - CPU Architecture
  -

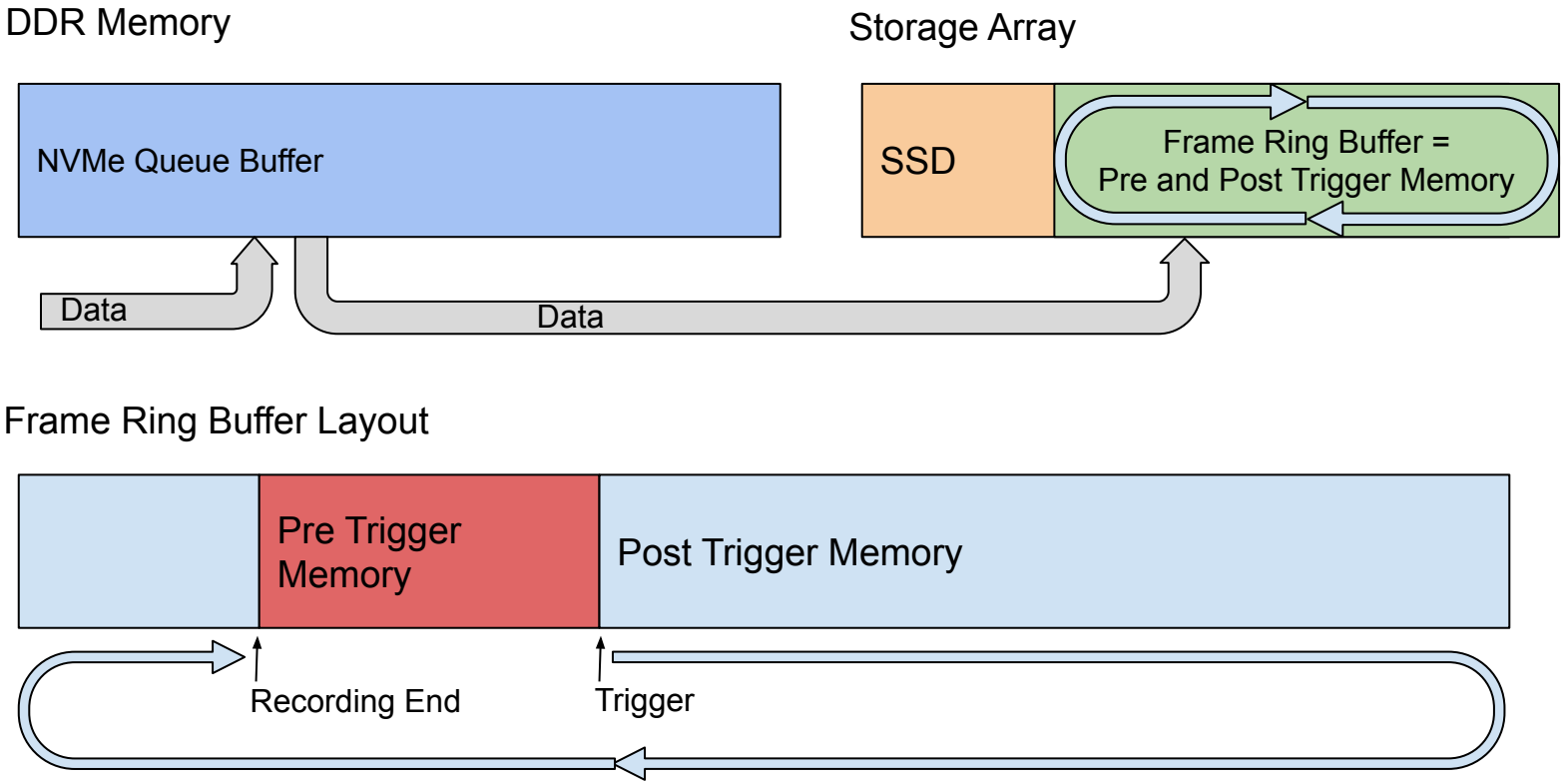
System Overview



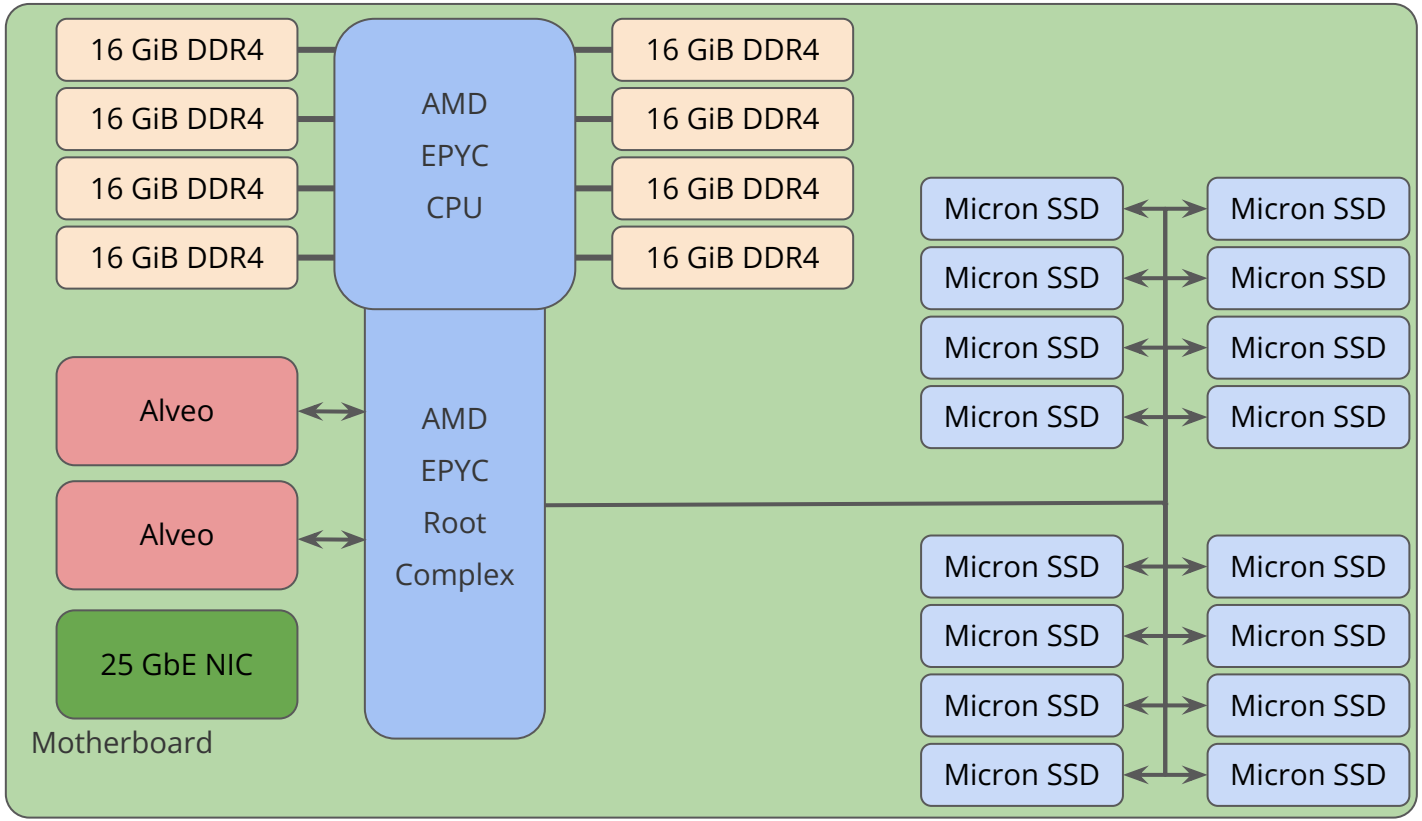
Recording Modes - Normal Recording Mode



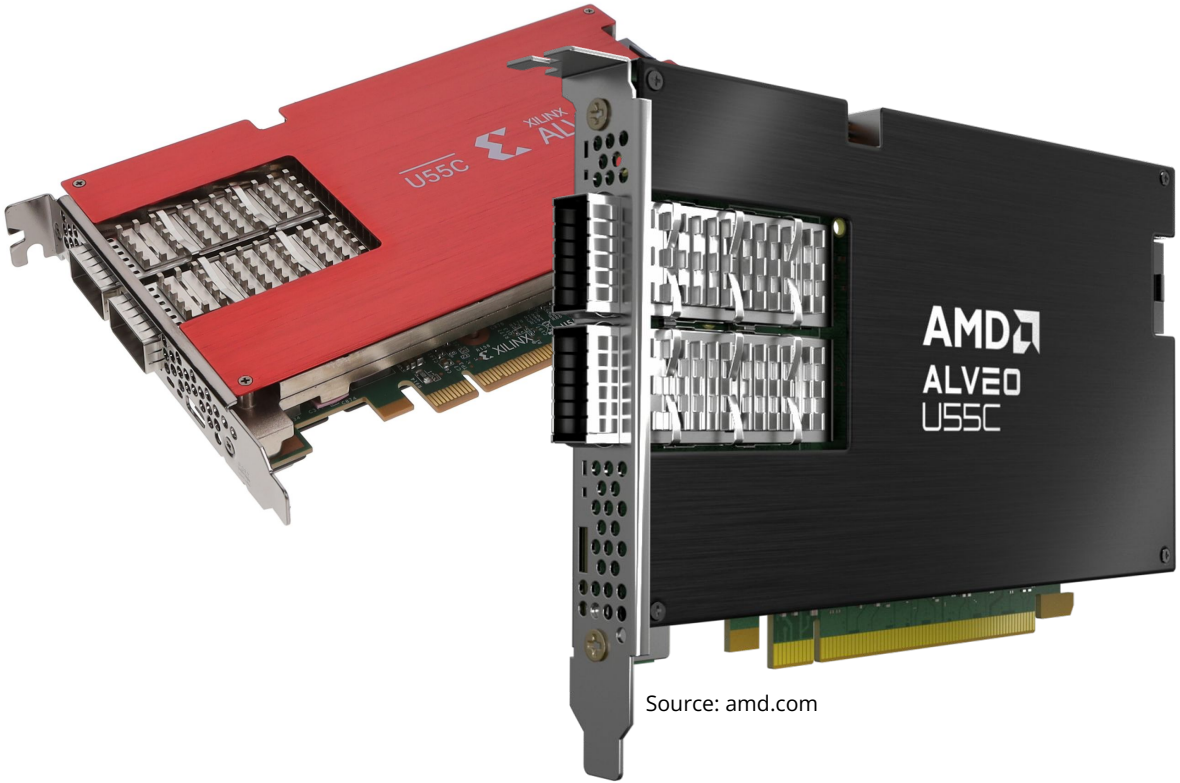
Recording Modes - PreTrigger Recording Mode



Hardware Architecture



AMD Alveo U55C FPGA Card



Source: amd.com

- Dual QSFP28
  - Eight 25 GbE links
- Dual PCIe 16 GT/s x8
  - Bifurcation of x16 PCIe edge connector
- Virtex UltraScale+ HBM
  - 2.6 M FF, 1.3 M LUTs
  - 3 SLRs (dies)
  - 16 GiB HBM2 Memory @ 460GB/s
- Compact FHHL, 1 Slot
- Relatively cheap < 5 k$

Agenda

- System Overview and requirements
  - Recording Modes
  - Hardware Architecture and Components
- **SSD**
  - Requirements
  - Intro NVMe
  - RAID Scaling Limitations
  - NVMe Streamer
  - Crossbar and PCIe Bridge
- Data transfer
  - Peer to peer communication
  - CPU Architecture
  -

NVMe SSDs

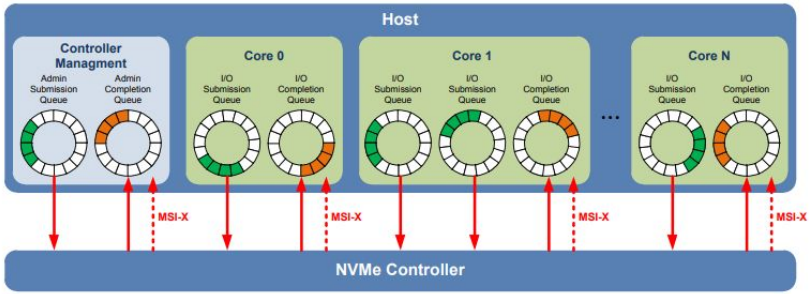
Requirements:

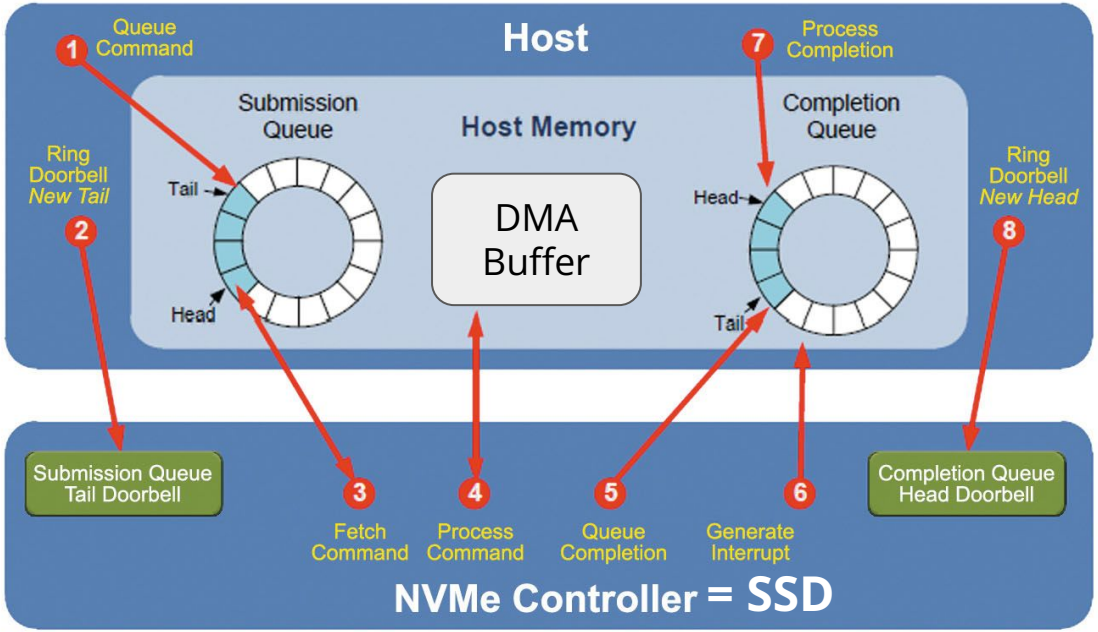
- High performance ~ 5 to 6 GiB/s **sustained** write speed per SSD
- **Thermal stability** with cooling options
- High endurance ~60 DWPD
- Capacity per SSD: 1 to 6 TiB
- RAID0 configuration
- Minimum RAID0 speed 200 Gbit/s => 23.3 GiB/s

Models:

- Micron XTR
- Kioxia FL6

NVMe Protocol Overview





Raid Setup

What is RAID 0?

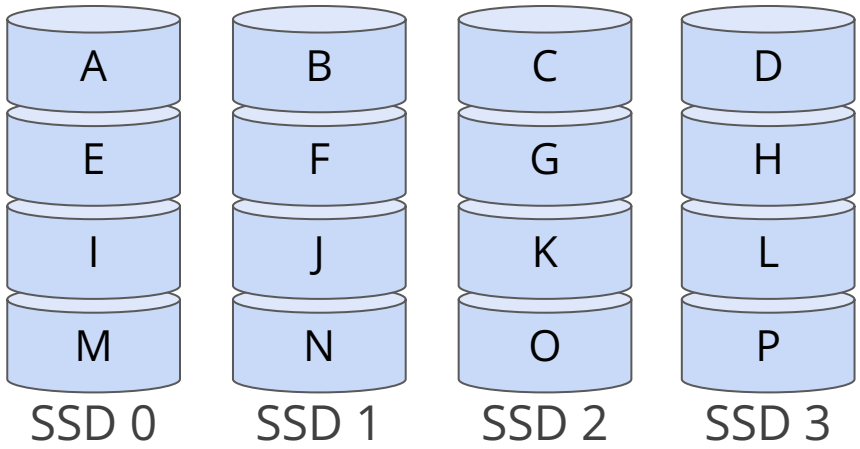
- Data is striped across all SSDs

Benefit of using RAID 0?

- write /read speed increase due to parallel access

Risk using RAID 0?

- No parity information
- Raid Scaling Limitations



Raid Scaling Limitations

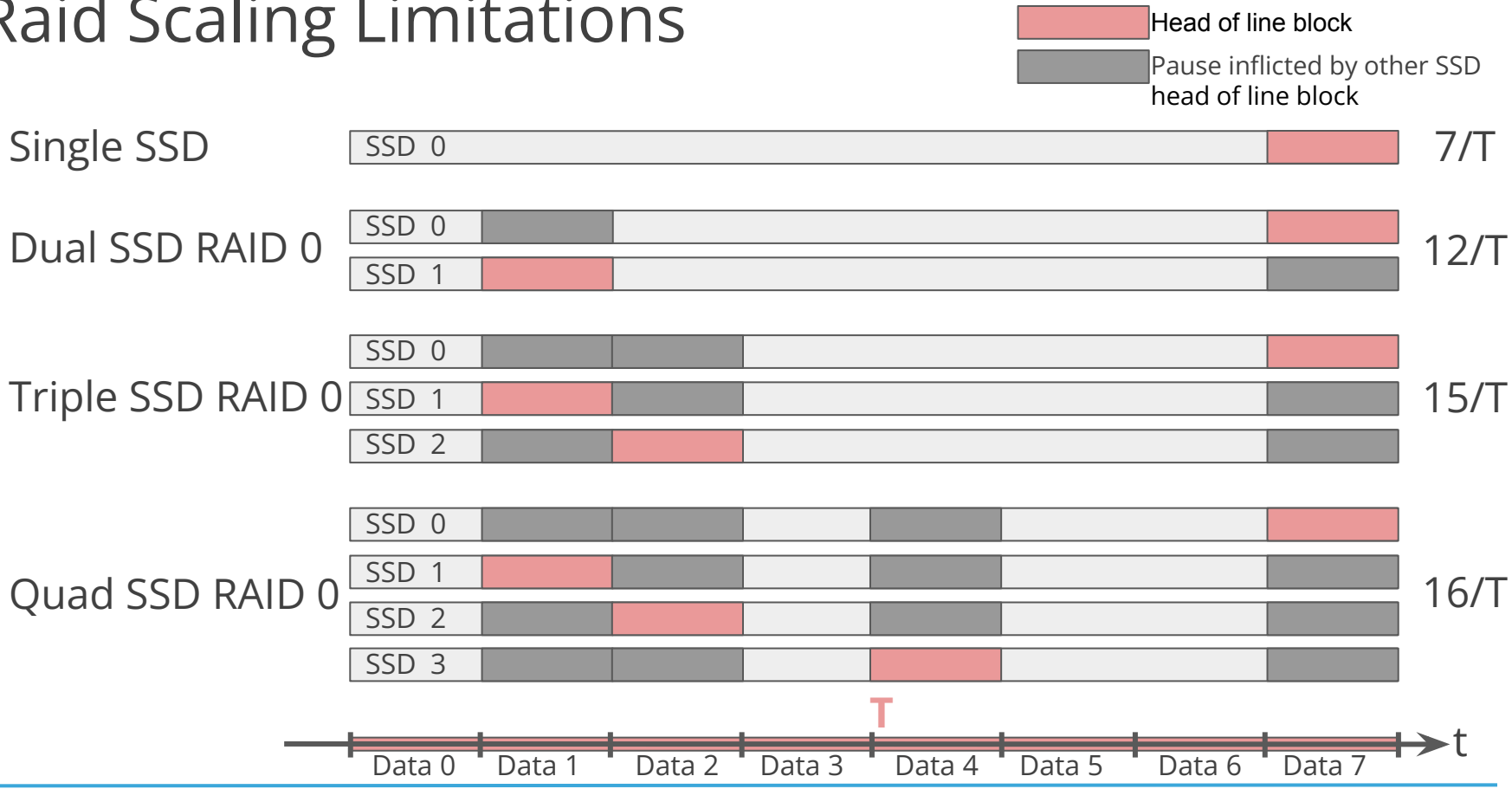
The efficiency of an RAID 0 can be limited when the number of SSDs reach a critical number.

The sequential datastream and RAID 0 striping, data / command distribution is strictly Round Robin

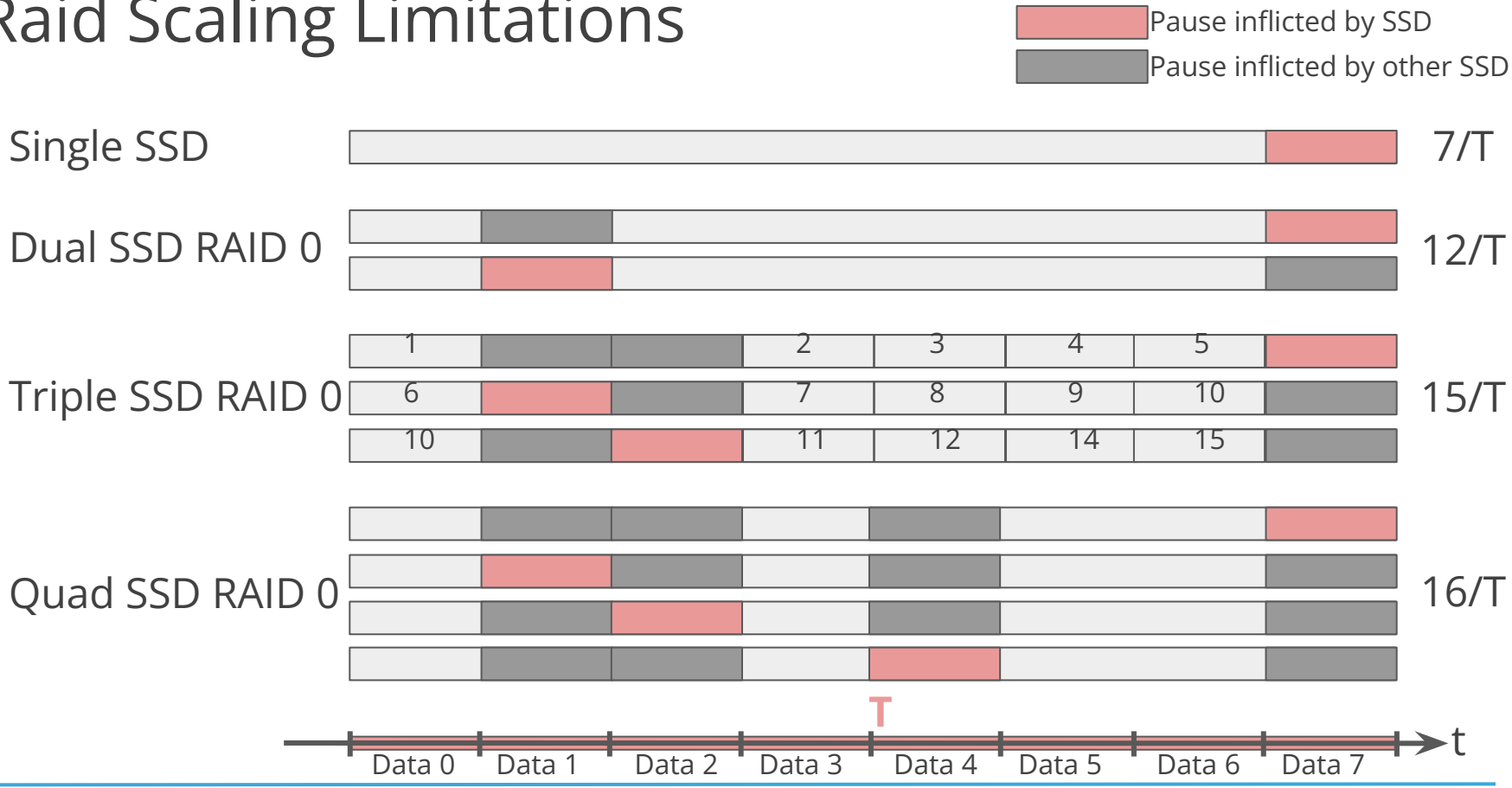
If one SSD has full queue and takes longer to process a pending command than others take to empty their queue, all SSDs have to wait

If SSDs don't synchronize their “pauses”, bandwidth increase using more raid members is limited

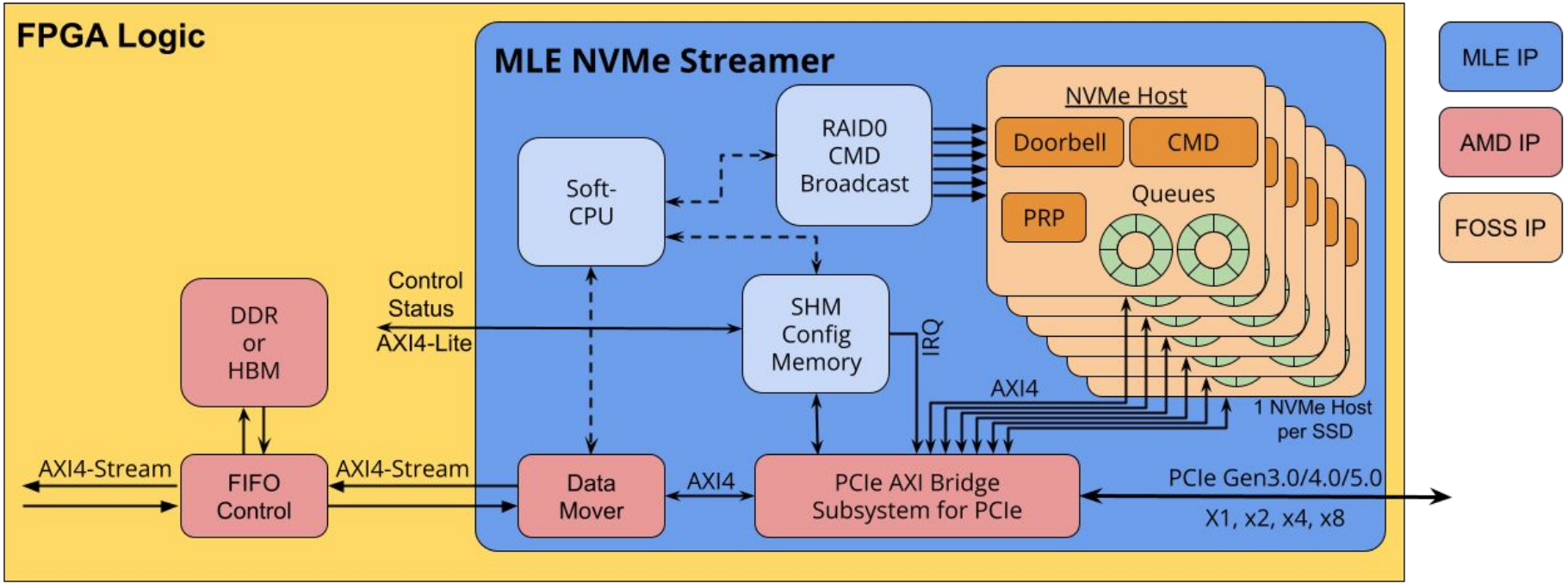
Raid Scaling Limitations



Raid Scaling Limitations



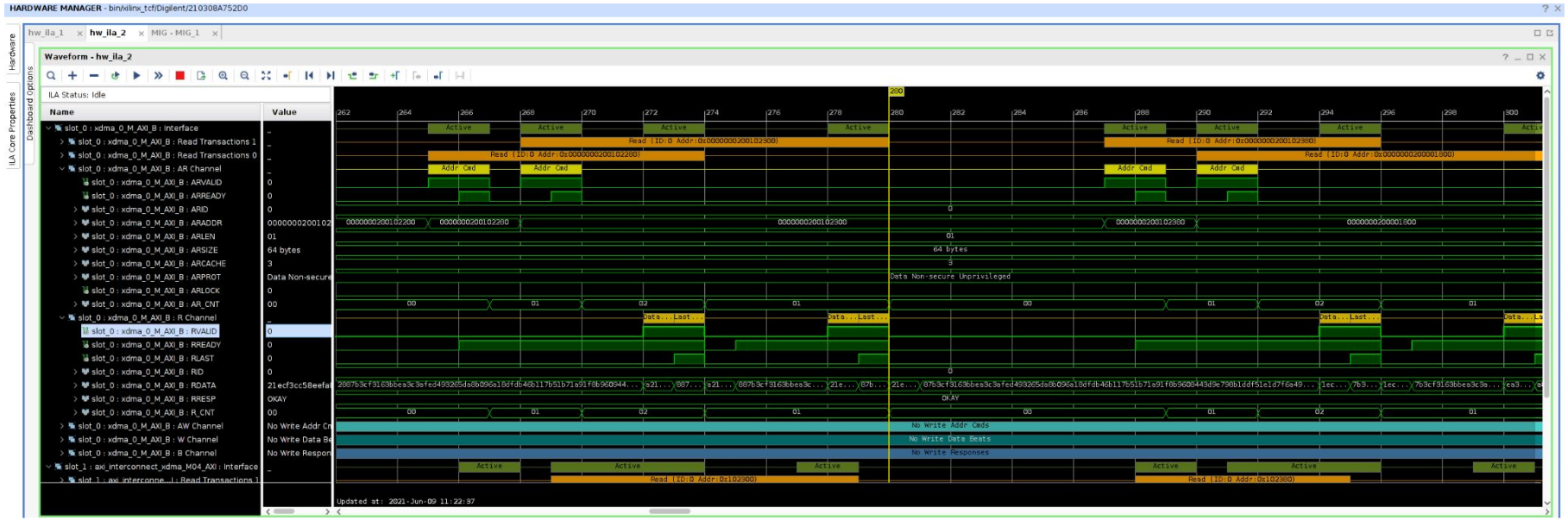
NVMe Streamer



High Level Data Organization

- Sixteen NVMe Streamer instances, one per 25 Gbit/s data link, to allow for sixteen independent sources (data rate, triggering, recording start and length)
  - Instances are distributed across multiple FPGAs and across the two PCIe links per FPGA
  - (Linux) Software can configure any combination of NVMe Streamers to operate in concert for a single data source with multiple 25 Gbit/s links
- Each NVMe Streamer comprises sixteen NVMe host controllers, one per SSD in two RAID0 configurations, operating in sync.
- Thus, each SSD is configured with sixteen IO Submission and Completion Queue pairs, one per associated NVMe host controller.
- (Linux) Software is responsible for managing the SSD / RAID storage areas and assigning them to individual NVMe Streamer instances.

PCIe Bridge Queue Depth vs Memory Latency



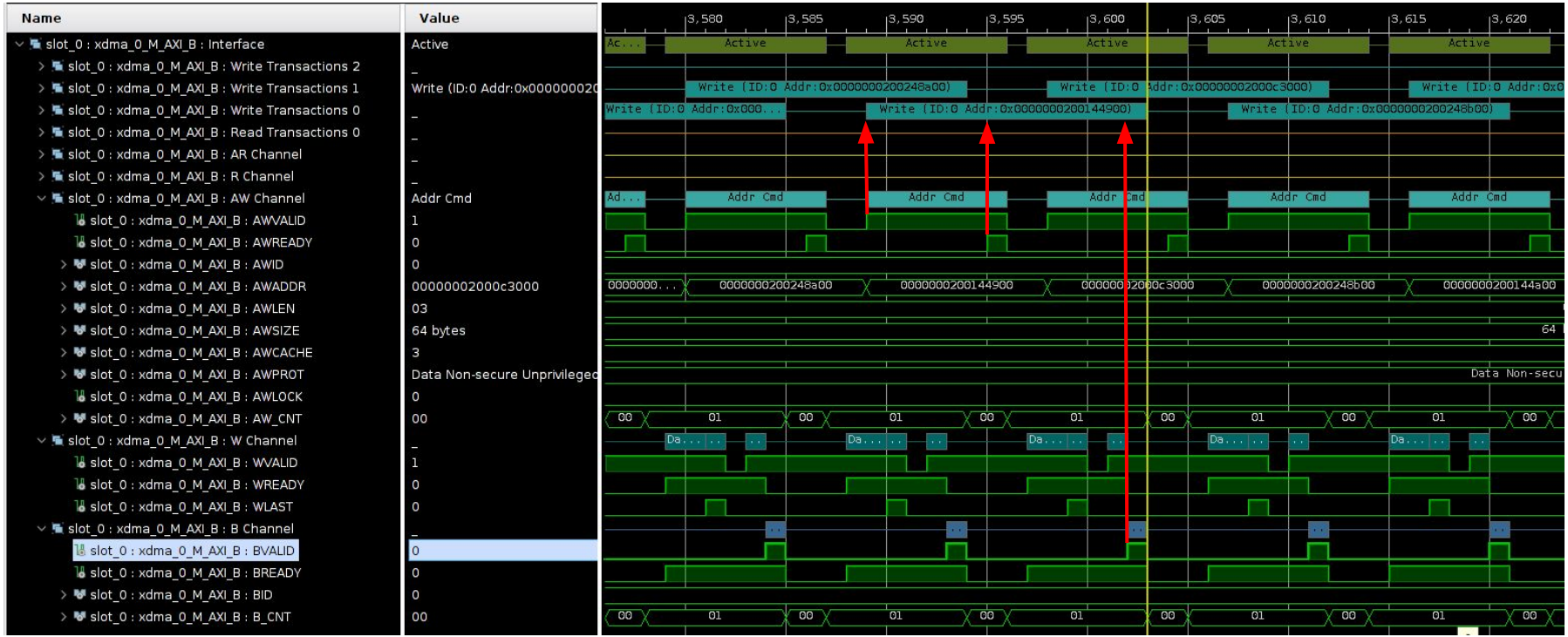
Xilinx Crossbar Single Slave per ID

- In certain architectures the Xilinx Vivado AXI Interconnect IP can perform suboptimal
- One case is a setup of one single threaded AXI Master accessing multiple AXI Slaves in parallel
- This is not directly possible with the Xilinx Vivado AXI Interconnect IP and accesses get serialized, making latency a major factor for bandwidth

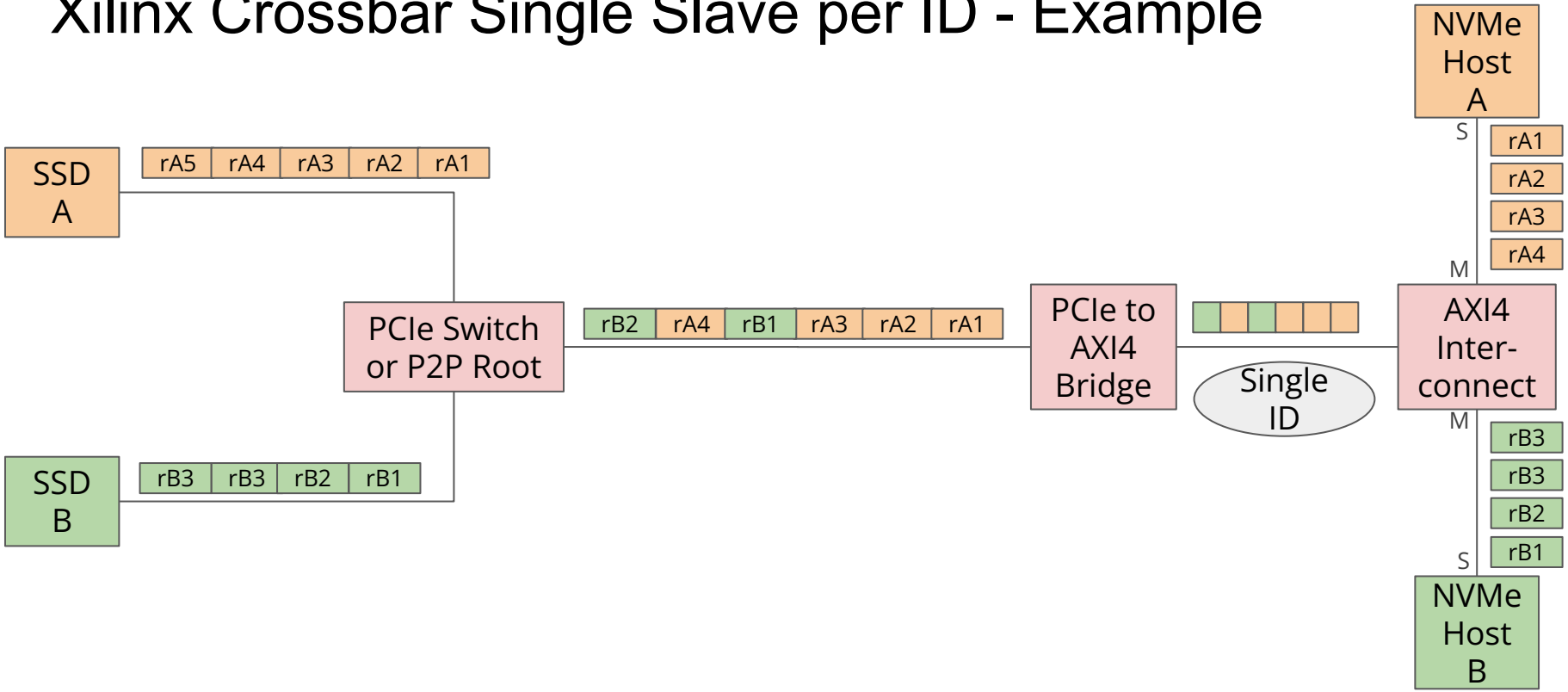
Explain loss of Performance in certain architectures

- If a single AXI Master needs to access multiple AXI Slaves in parallel, the Xilinx AXI interconnect imposes a performance bottleneck.
- In this application 16 NVMe SSDs access the FPGA internal AXI infrastructure via a single Xilinx Bridge for PCI Express operating as the AXI Master. Accesses are highly parallel because of the many independent SSDs and the multitude of FPGA memory channels used.
- However, as can be seen in this ILA capture at the AXI Bridge for PCI Express AXI Master interface, the Xilinx AXI interconnect accepts only one write transfer at a time (awvalid assertion is waiting for bvalid assertion), completely removing

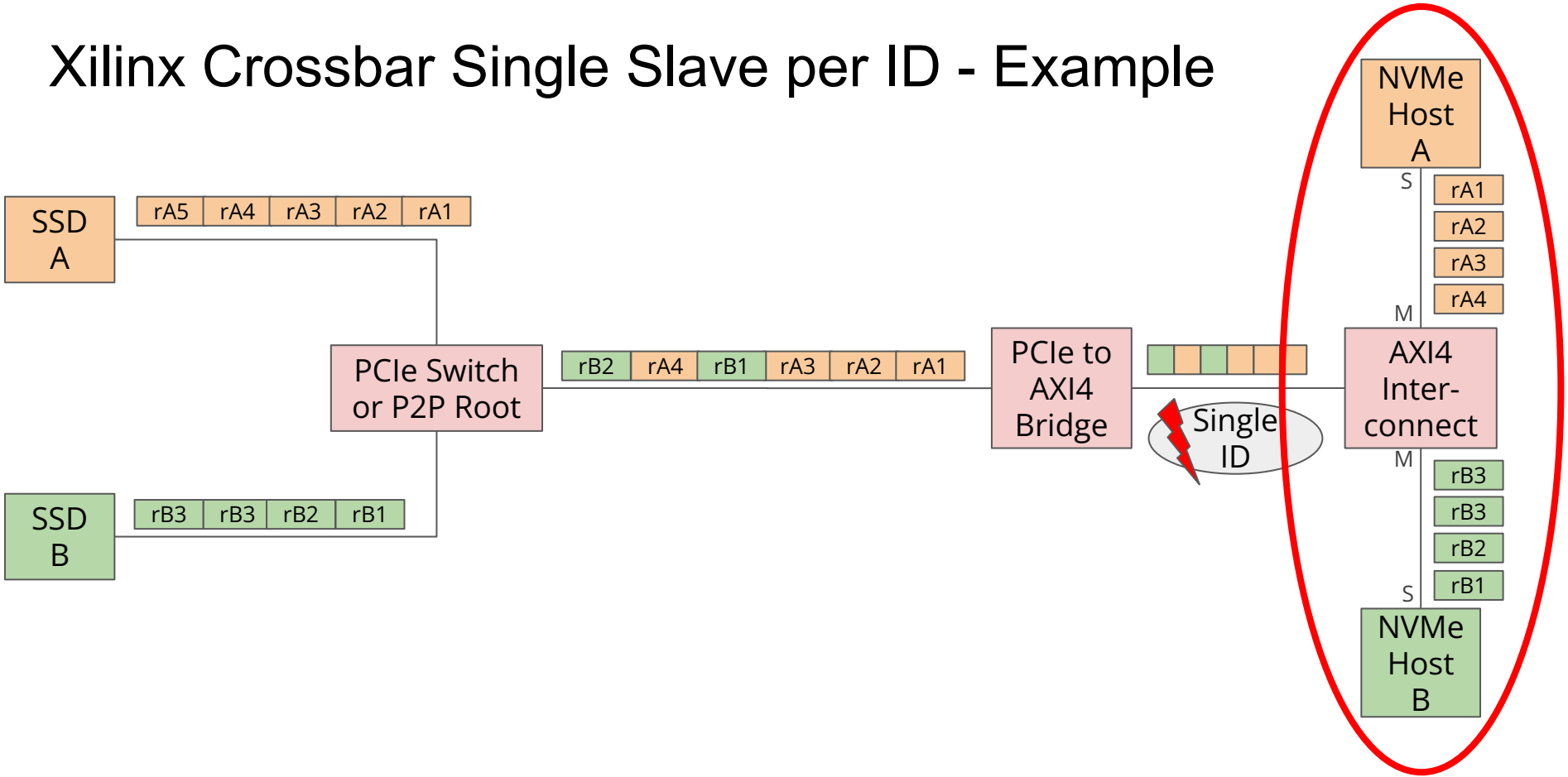
Xilinx Crossbar Single Slave per ID



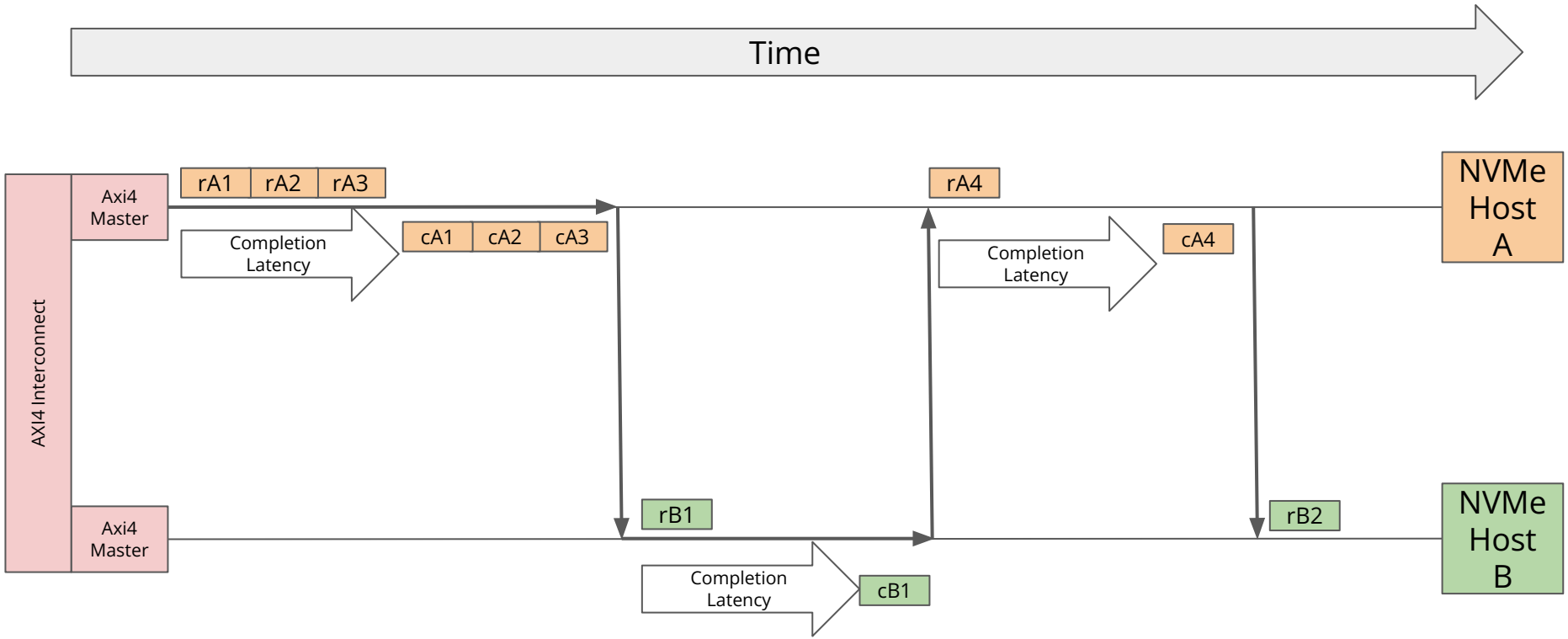
Xilinx Crossbar Single Slave per ID - Example



Xilinx Crossbar Single Slave per ID - Example



Xilinx Crossbar Single Slave per ID - Timing



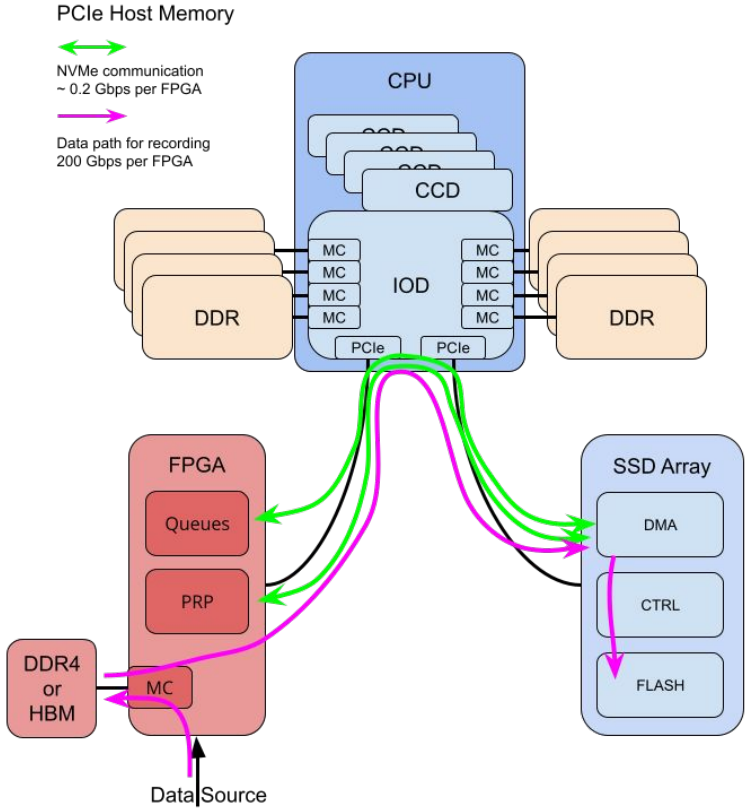
Agenda

- System Overview and requirements
  - Recording Modes
  - Hardware Architecture and Components
- SSD
  - Requirements
  - Intro NVMe
  - RAID Scaling Limitations
  - NVMe Streamer
  - Crossbar and PCIe Bridge
- **Data transfer**
  - Peer to peer communication
  - CPU Architecture
  - Peer to Peer

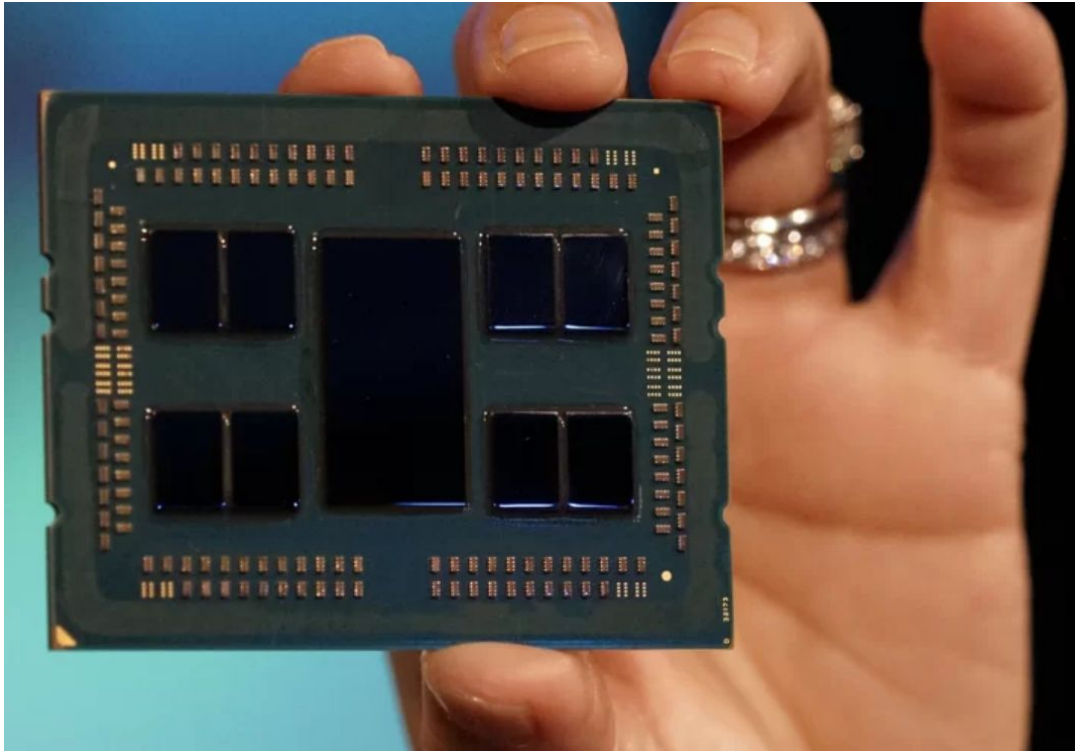
PCIe Peer to Peer communication

Initial idea was to utilize the AMD Root Complex for peer to peer communication, but...

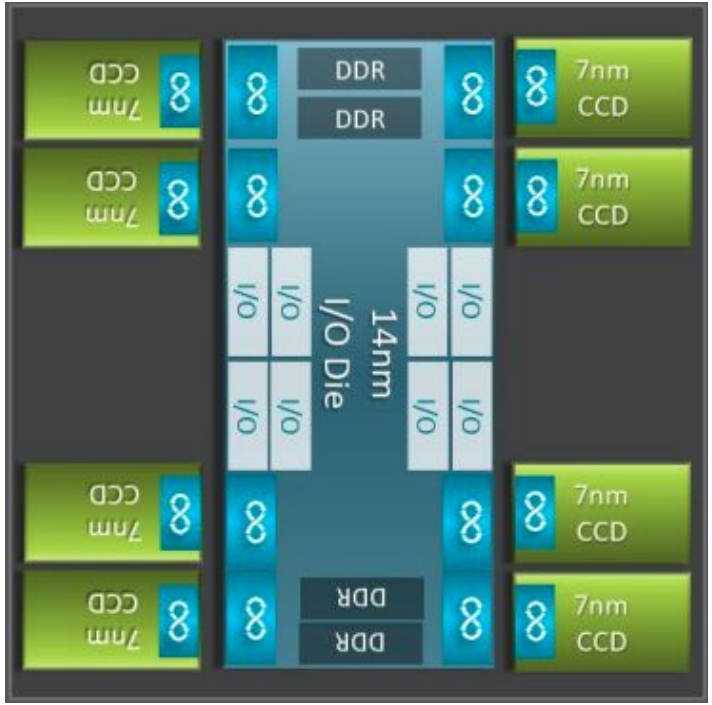
PCIe Root Complex **!=** PCIe Switch



AMD EPYC 2nd/3rd Generation



Source: https://www.tomshardware.com/picturestory/865-amd-epyc-supercomputer-slideshow-server.html

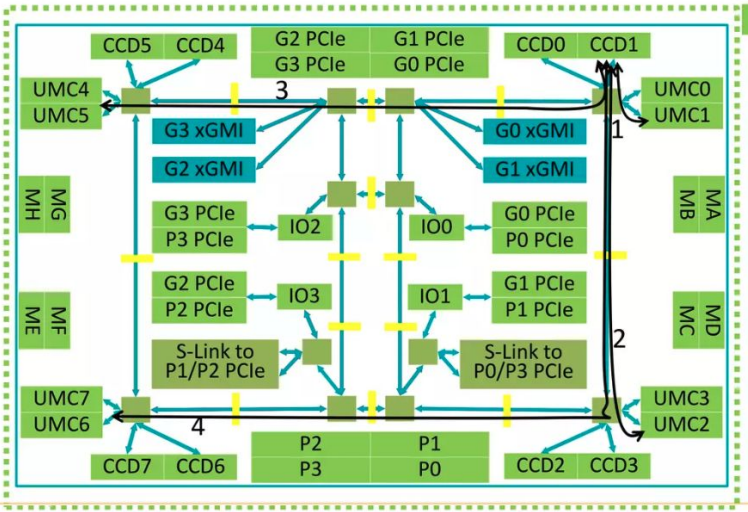


Source: https://www.amd.com/system/files/documents/TIRIAS-White-Paper-AMD-Infinity-Architecture.pdf

IOD Findings

2nd Gen AMD EPYC™ Improved Memory Latency

- Central IOD enables a single NUMA domain per socket
- Improved average memory latency[1] by 24ns (19%)[2]
- Minimum (local) latency only increases 4ns with chiplet architecture



Single Domain

CCD0, CCD1, IO0, CCD2, CCD3, IO1, CCD4, CCD5, IO2, CCD6, CCD7, IO3, MA/MB/MC/MD/ME/MF/MG/MH interleaved

1.46GHz / DDR2933 (coupled)[1]

1: Local 94ns
2: ~97ns
3: ~104ns
4: ~114ns

Measured Avg: ~104ns

Repeater: 1 FCLK (1.46GHz)

Switch: 2 FCLK (1.46GHz) (low-load bypass, best-case)

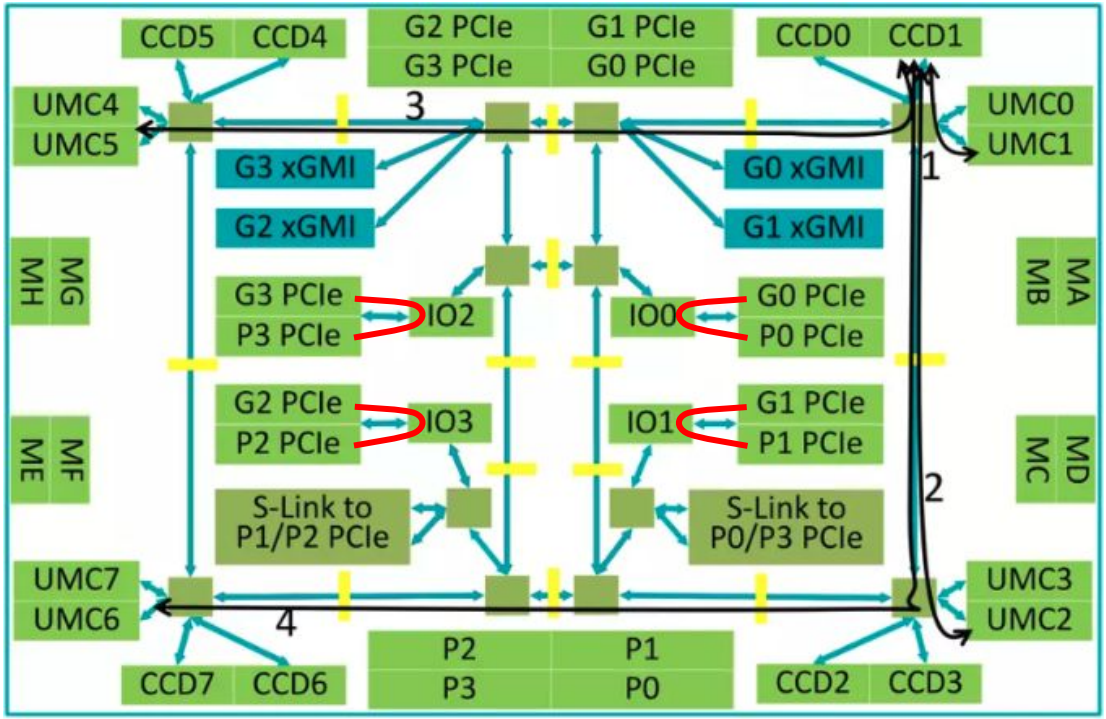
1: AMD internal testing with DRAM page miss
2: EPYC 7002 Series NUMA 1 vs EPYC 7001 Series Avg. Local; EPYC 7002 Series NUMA2 vs EPYC 7001 Series NUMA 3

Source: https://www.slideshare.net/AMD/amd-chiplet-architecture-for-highperformance-server-and-desktop-products

IOD Findings

Fastest Peer to Peer connections

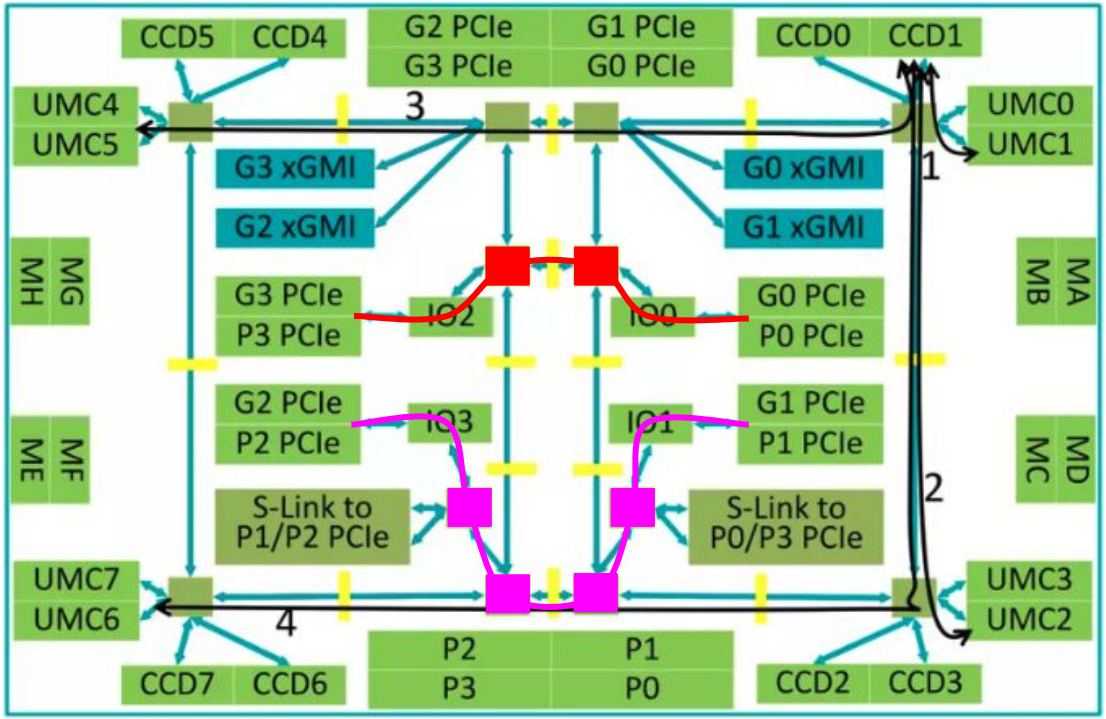
The recorder needs more than two PCIe Slots



IOD Findings

Second fastest connection (two switches)

Attention: This route is slower than the upper side (four switches)

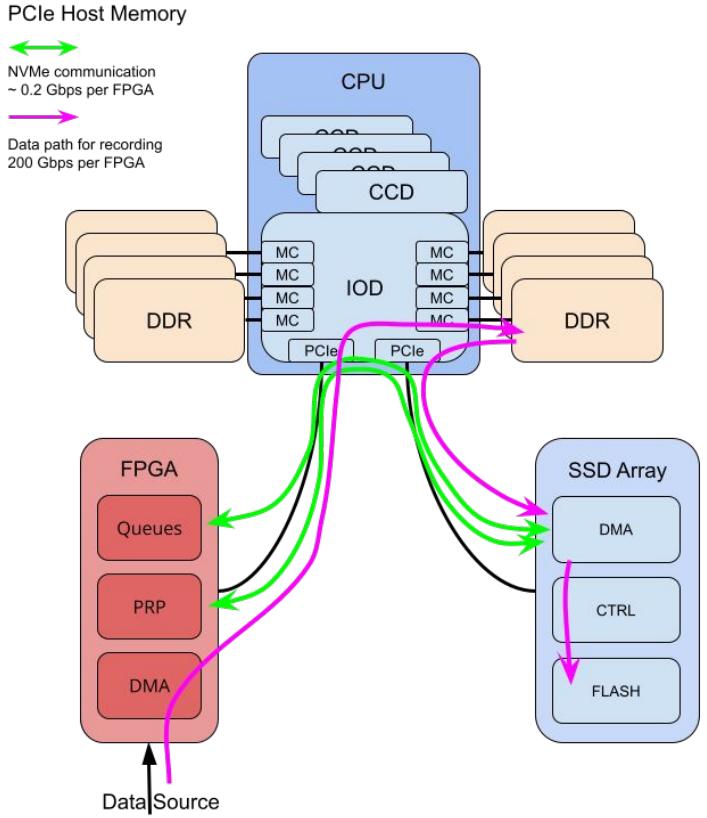


Peer to Peer with Host Memory Utilisation

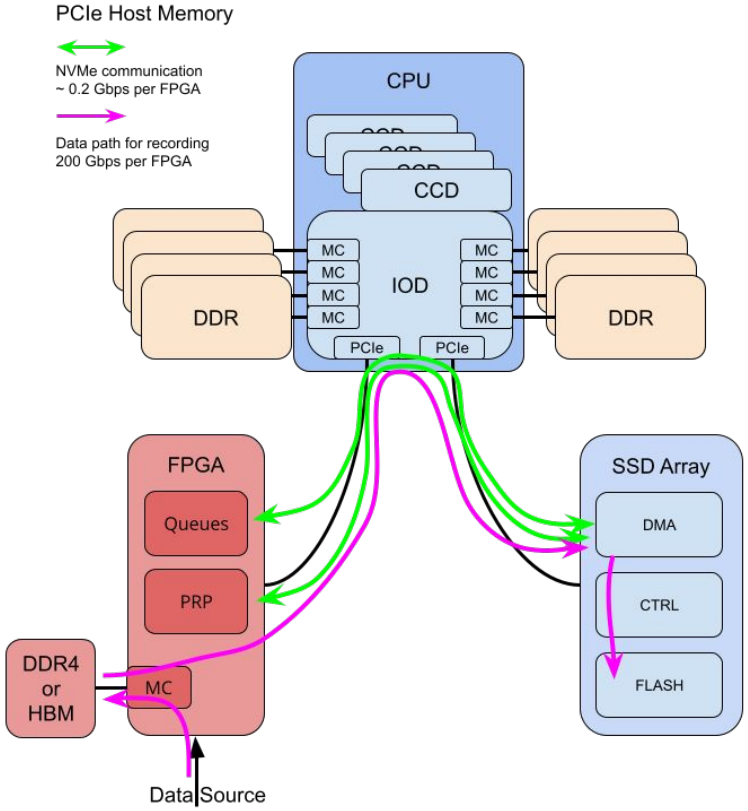
With the knowledge of the Root Complex Architecture and limitations a new approach was testet:

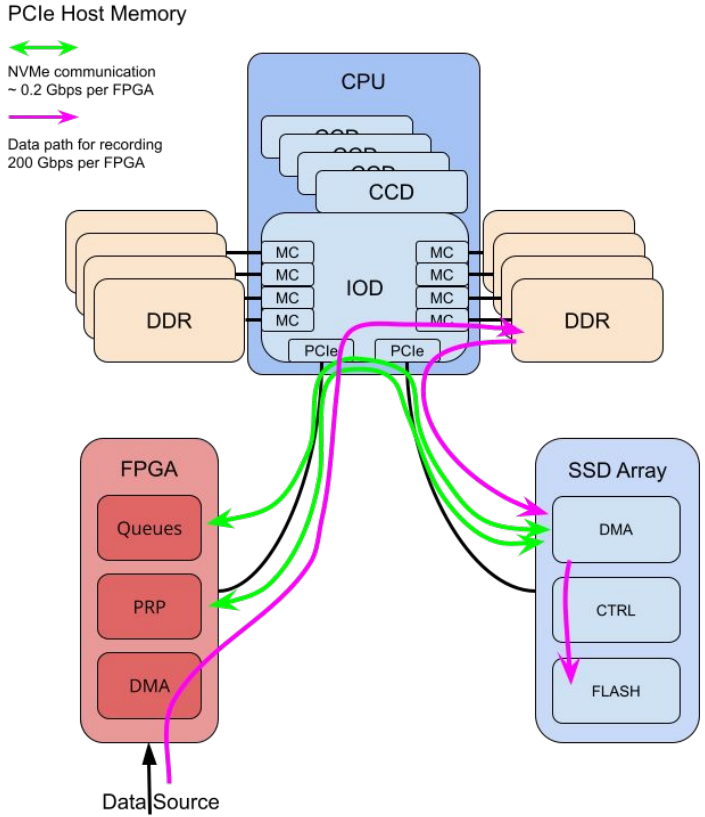
Peer to Peer with Host Memory as bounce buffer

- Highly optimized paths for small data granularity of SSDs



Comparison of Different Peer to Peer Options





Contact:

Andreas Schuler

Andreas.schuler@missinglinkelectronics.com

David Epping

David.Epping@missinglinkelectronics.com

Missing Link Electronics

Industriestraße 10
89231 Neu-Ulm

www.missinglinkelectronics.com