



Flash Memory Summit

Getting started with OMI: *Memory controller implementation example*

Bruno Mesnet, IBM Client Engineering
OpenCAPI - OMI enablement



Just Simple and Open

Overview

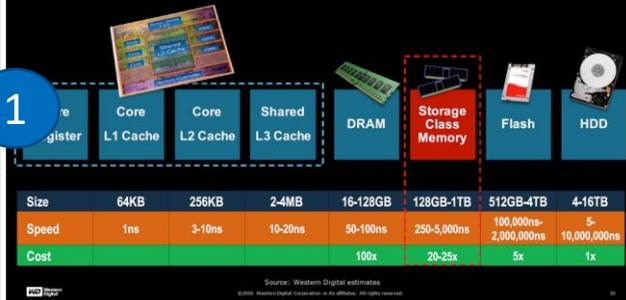


Flash Memory Summit

Memory in a server



Moving Mountains of Data



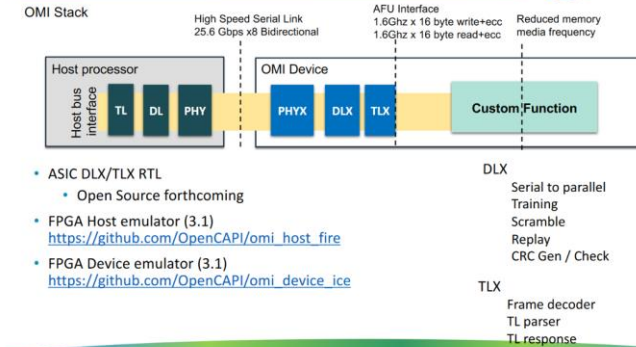
The further away from the core, the longer it takes for data to be exchanged

3 | ©2022 Flash Memory Summit. All Rights Reserved.

OMI Reference material (pdf)



Current Open Memory Interface (OMI) Reference



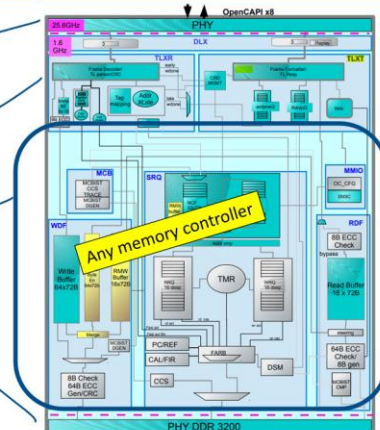
- ASIC DLX/TLX RTL
 - Open Source forthcoming
- FPGA Host emulator (3.1)
 - https://github.com/OpenCAPI/omi_host_fire
- FPGA Device emulator (3.1)
 - https://github.com/OpenCAPI/omi_device_ice

7 | ©2022 Flash Memory Summit. All Rights Reserved.

ASIC Block Diagram



- PHY
- OMI ASIC Device Reference Design
 - https://github.com/OpenCAPI/omi_asic_device_reference_design
- User logic + Memory Controller
 - MMIO
 - Write path (WDF)
 - Read path (RDF)
 - Scheduling (SRQ)
- Memory PHY

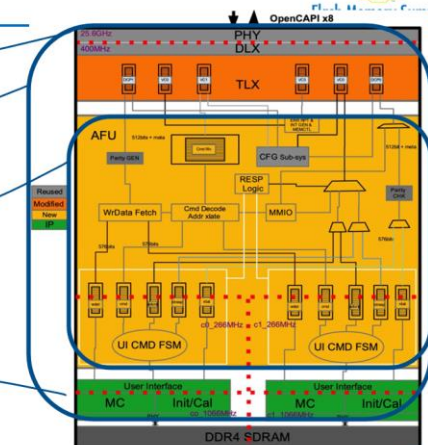


10 | ©2022 Flash Memory Summit. All Rights Reserved.

FPGA device block diagram



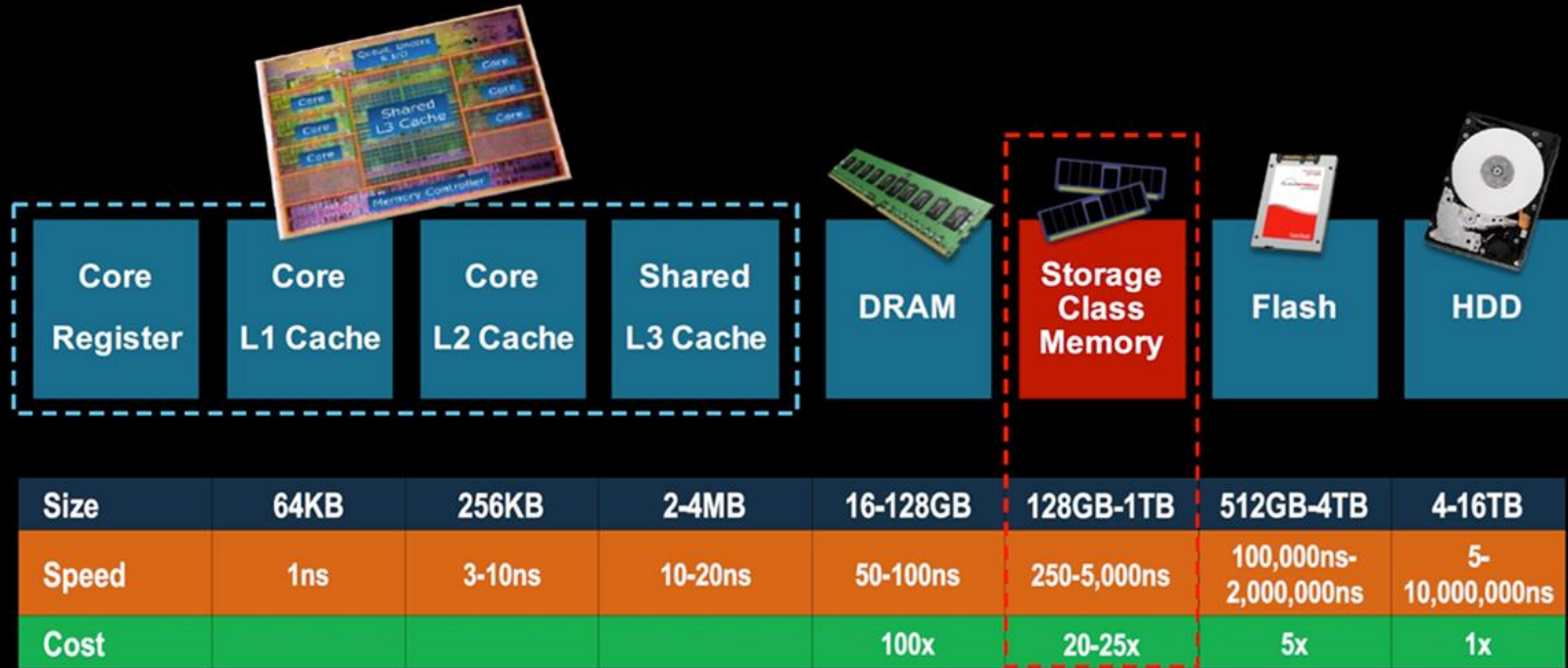
- PHY: Xilinx tranceiver
- OMI FPGA Device Reference Design
 - https://github.com/OpenCAPI/omi_device_ice
- User logic
 - MMIO
 - Write path (WDF)
 - Read path (RDF)
 - Scheduling (SRQ)
- 2x Xilinx DDR4 memory controller
- 2x Xilinx DDR4 PHY



11 | ©2022 Flash Memory Summit. All Rights Reserved.

Memory in a server

Moving Mountains of Data



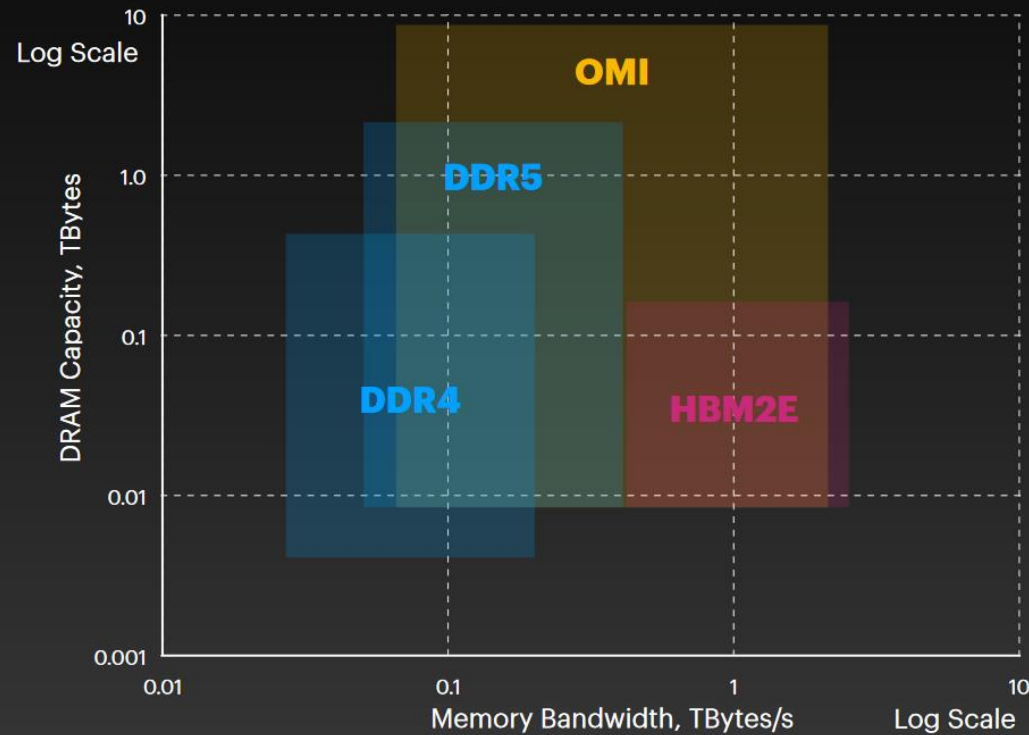
The further away from the core, the longer it takes for data to be exchanged

Size versus Bandwidth



Flash Memory Summit

Memory Interface Comparison OMI - Bandwidth of HBM at DDR Latency, Capacity & Cost

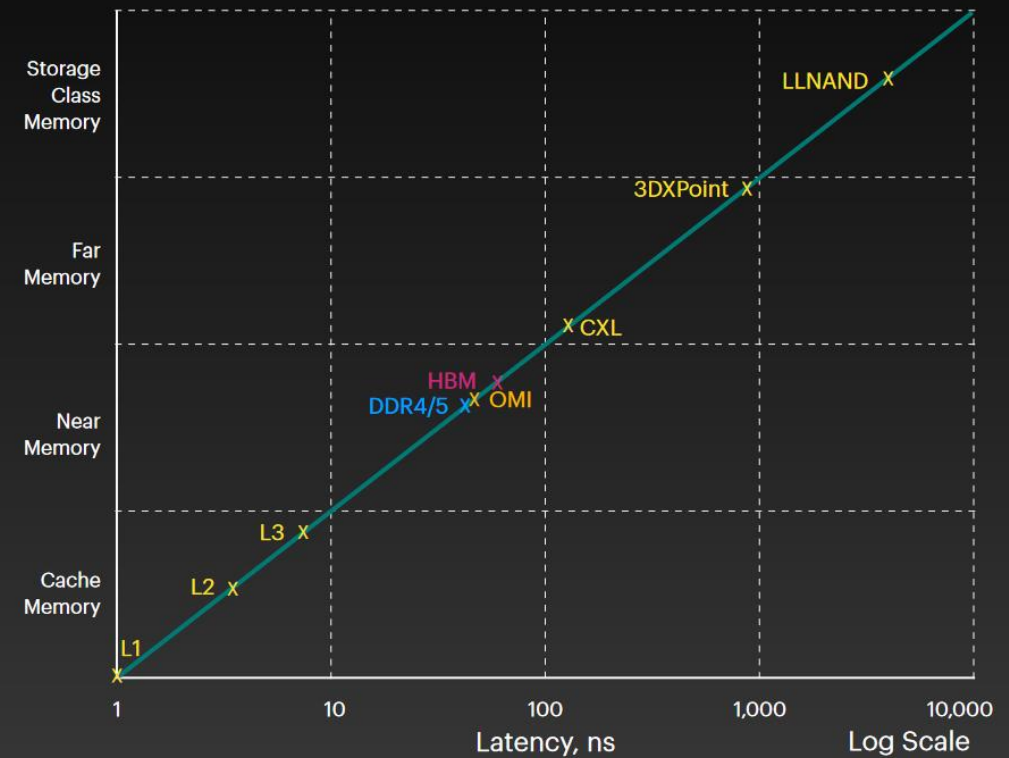


Graph depicts Practical Minimum : Maximum Channels per Processor

DDR = 1 : 8

OMI = 1 : 32

HBM = 1 : 6



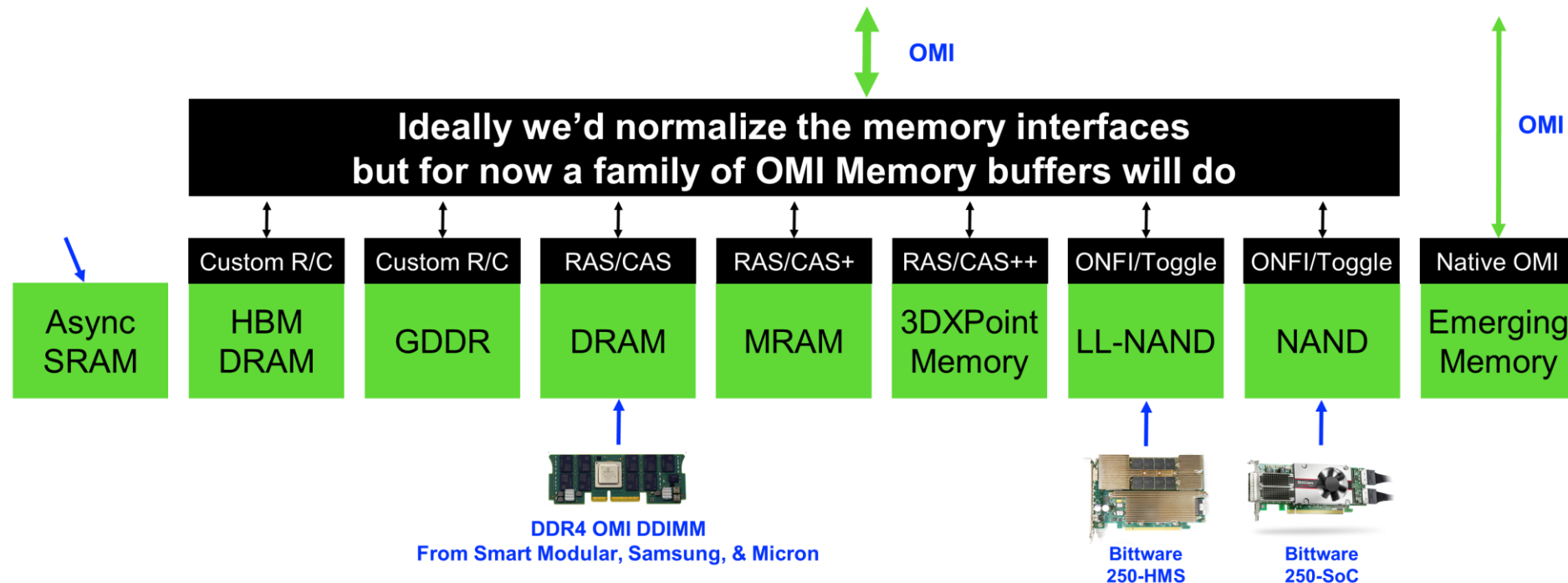
NALLASWAY



OpenCAPI Memory Interface, OMI

Gathering the Divergent Memory Choices

- Today's Tiered Memory Choices - Valid Members of a Data Centric Memory Fabric
- Trade = Depth vs Latency vs Bandwidth vs Persistence vs Power vs Cost

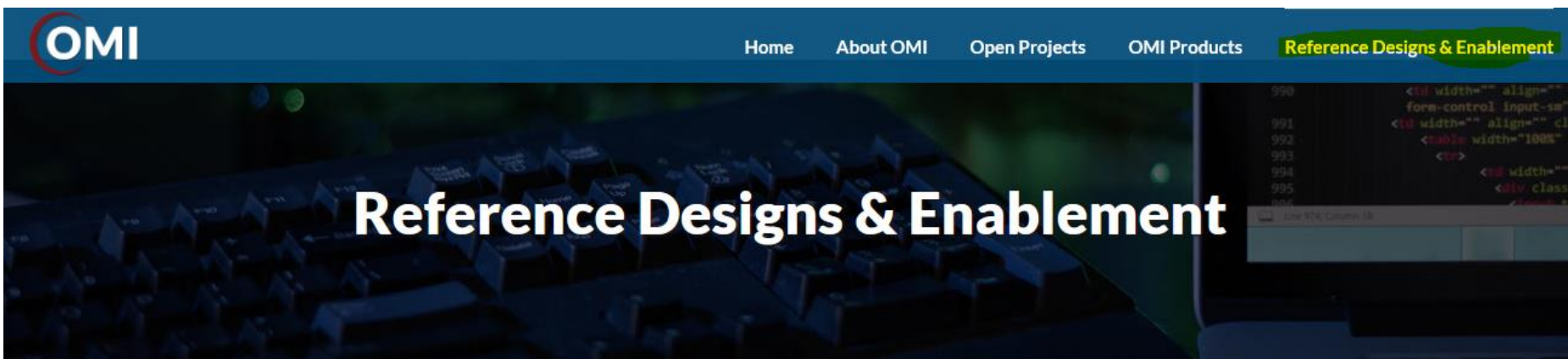


Go to...

<https://openmemoryinterface.org/>



Flash Memory Summit



A critical component of growing any ecosystem is to have enablement available for respective developers. This includes specifications, tools, and reference designs. Below is a list of available enablement today. It is suggested to periodically visit as more enablement will be added over time.

OMI ARCHITECTURE SPECIFICATIONS

Design Overview

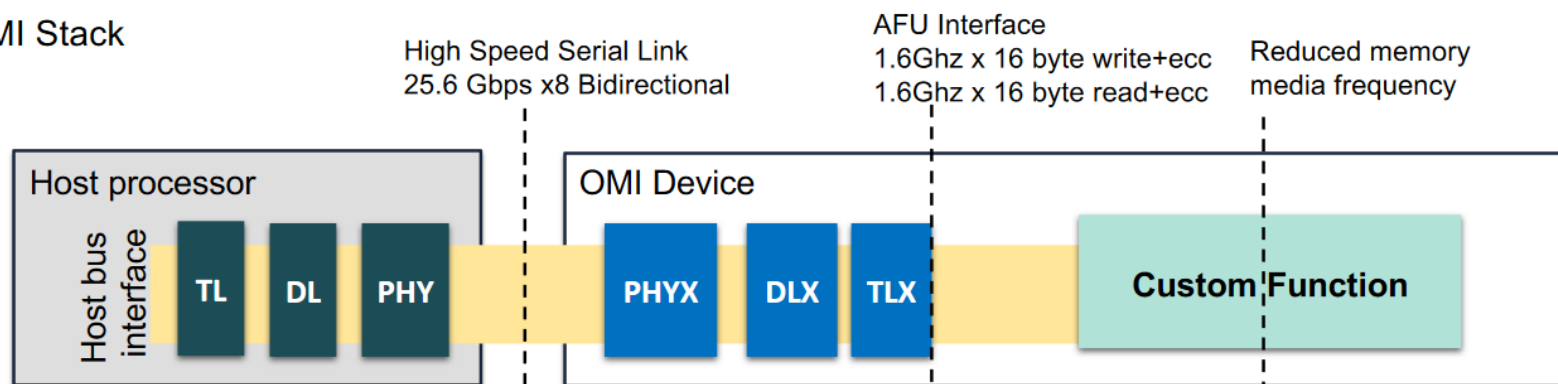
A general understanding of OMI is helpful before consuming architecture and design details. The following [OMI reference material \(PDF\)](#) shows block diagrams of the OMI stack, PCIe interoperability, and the relationship of OMI to the Data Link layer and Transaction Link layer.

OMI Reference material (pdf)

Current Open Memory Interface (OMI) Reference



OMI Stack



- ASIC DLX/TLX RTL
https://github.com/OpenCAPI/omi_asic_device_reference_design
- FPGA Host emulator (3.1)
https://github.com/OpenCAPI/omi_host_fire
- FPGA Device emulator (3.1)
https://github.com/OpenCAPI/omi_device_ice

DLX

Serial to parallel
Training
Scramble
Replay
CRC Gen / Check

TLX

Frame decoder
TL parser
TL response

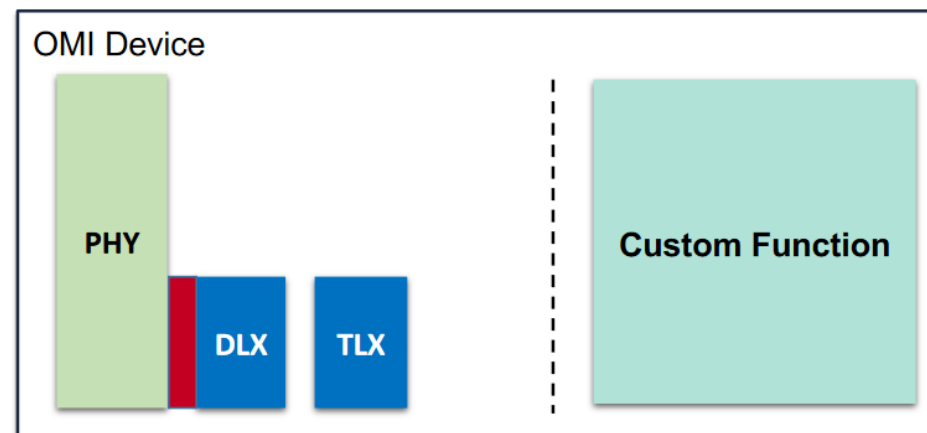
OMI Reference material (pdf)



- OMI and PCIe PHY are compatible/interoperable
 - OMI PHY layer is Based on the OIF CEI 28G SR specification
 - PCIe 5 SerDes PHY x16
 - OMI 32 Gbps PHY x8
 - OMI reference clock is 133.33 MHz
 - PCIe reference clock is 100 MHz

- OMI 3.1 available with P10

The DLX of OMI
Is directly connected to the serdes pins of the PHY
No connection to PCS/PIPE as in the PCI-e stack

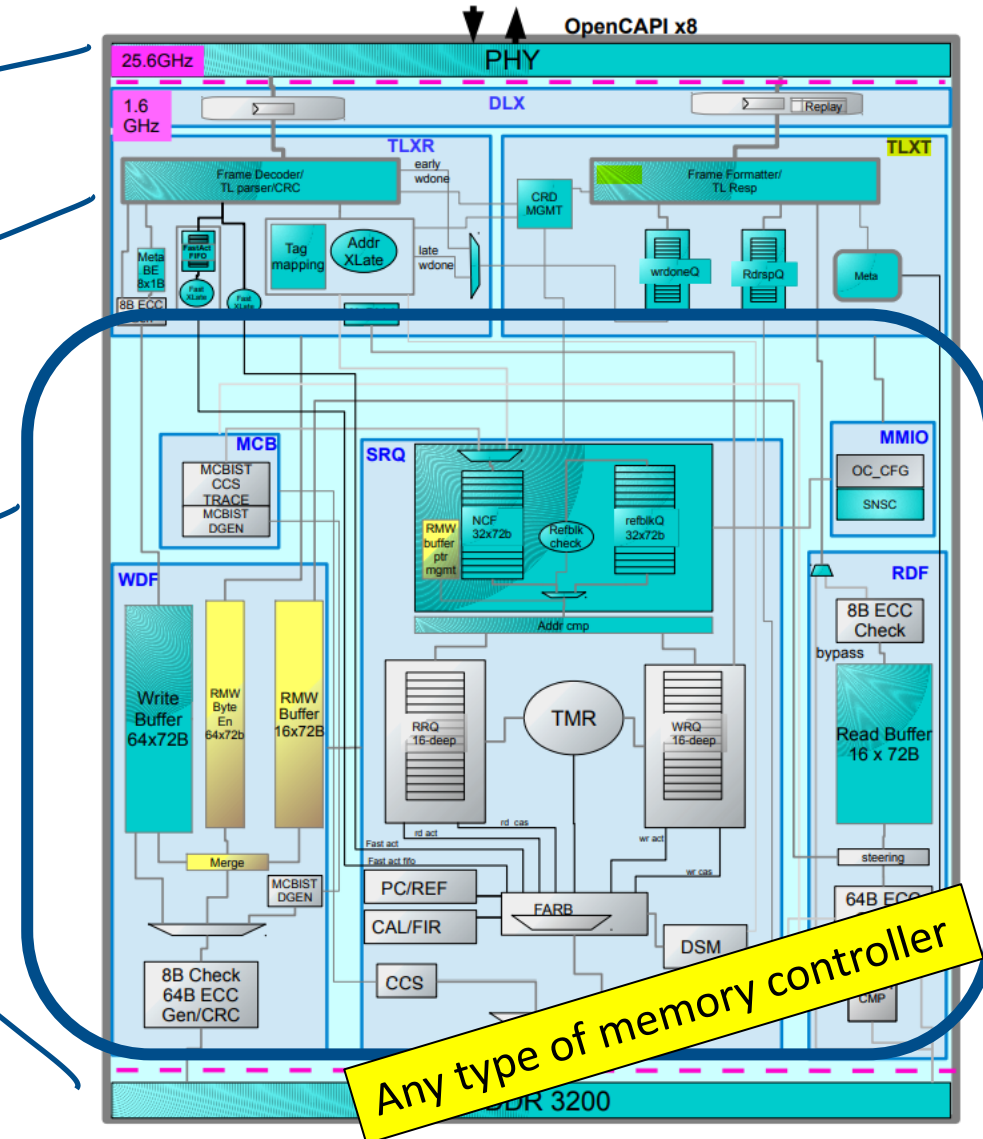


ASIC Device Block Diagram



Flash Memory Summit

- PHY: Alphawave
- OMI ASIC Device Reference Design
 - https://github.com/OpenCAPI/omi_asic_device_reference_design
- User logic + Memory Controller
 - MMIO
 - Write path (WDF)
 - Read path (RDF)
 - Scheduling (SRQ)
- Memory PHY: IP vendor

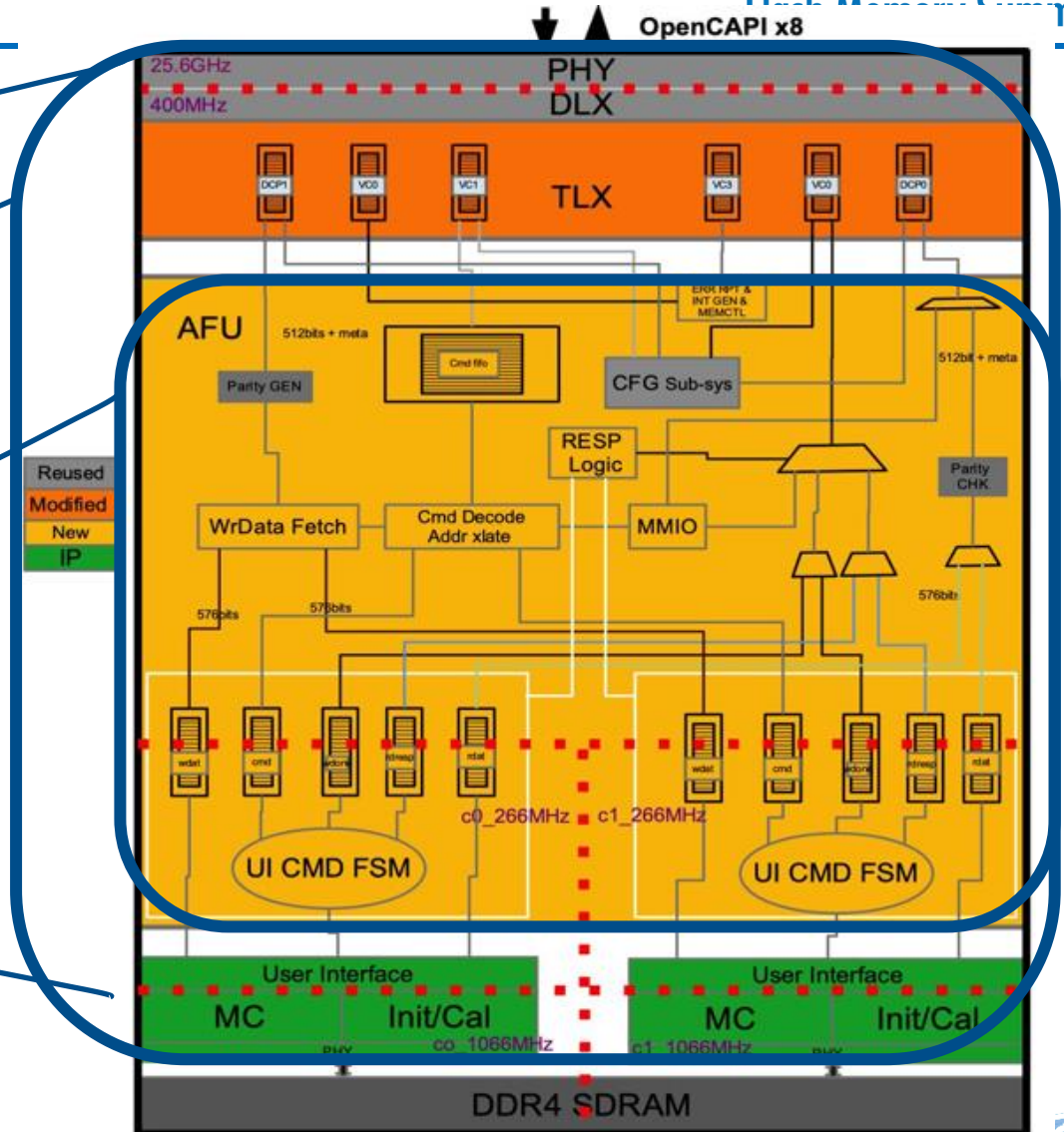


FPGA Device block diagram



Flash Memory Summit

- PHY: Xilinx tranceiver
- OMI FPGA Device Reference Design
 - https://github.com/OpenCAPI/omi_device_ice
- User logic
 - MMIO
 - Write path
 - Read path
 - Scheduling
- 2x Xilinx DDR4 memory controller
- 2x Xilinx DDR4 PHY



FPGA Device (ICE) design features

- User logic interface TLX
- Supports *mem_rd*, *mem_wr*, *partial mem_rd* and *partial mem_wr*.
 - *Error reported on unsupported command.*
- I2C and jtag control
- 133MHz clocking received from DDIMM connector
- Flow control:
 - Each memory port has 256 credits for Rd + 256 credits for Wr
 - 4 credits for MMIO commands
- Address remapping to allow maximum performance at DDR4 interface
- 2x Xilinx DDR4 memory controller (x8 16Gb 1066MHz)
 - *576bits per cycle at 1066MHz = 19GB/sec per port*
- 2x Xilinx DDR4 PHY: *frequency ratio between PHY and MC: 4:1*
 - *Xilinx PHY performs memory programming and internal calibration – re-calibration on failure*
 - 32GB of memory space can be initialize to 0 or data equal to address

ASIC FPGA differences



Flash Memory Summit

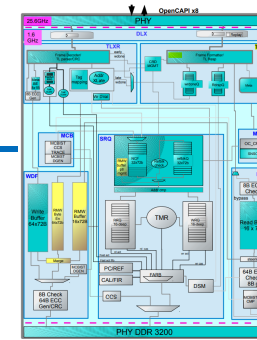
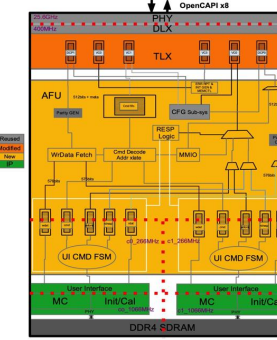


Figure 1: High-Level Memory Data Flow



- OMI PHY
- Frequency ratio between PHY and DL:
- Internal logic clocking:
- Round trip latency over DDR4 RDIMM:

- *FPGA logic example is functional but not latency optimized*

- Memory controller calibration:
- Memory PHY
- Memory

Alphawave true PHY

16:1

1.6GHz

< 10ns

FW

IP vendor

4U 512GB 409GB/s per skt

Xilinx GTY transceiver

64:1

400MHz

not representative

Internal Microblaze

Xilinx DDR4 PHY

32GB @ 19GB/s per port



Just Simple and Open

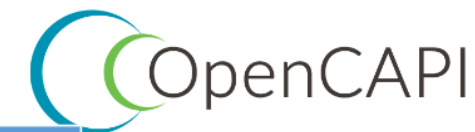
Questions



Flash Memory Summit

Backup

DL <-> PHY Interface



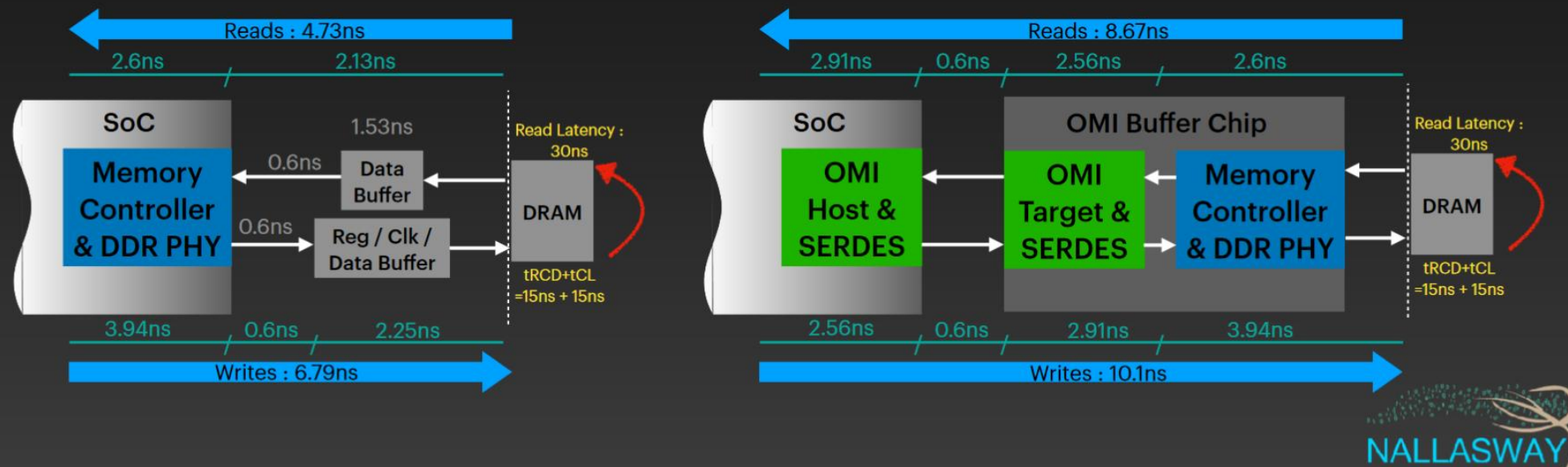
| DL <-> PHY Signal | Comment |
|-------------------------------|--|
| PHY_DL_CLOCK_<7:0> | Recovered captured clock for this lane |
| PHY_DL_LANE_<7:0>(15:0) | 16 bits of Rx Data for this lane |
| PHY_DL_INIT_DONE_<7:0> | Indication from the PHY that it is trained and has good eyes |
| PHY_DL_RECAL_DONE_<7:0> | Indication from the PHY that calibration is complete |
| PHY_DL_IOBIST_RESET | Reset to the DL driven from the PHY to kick off IOBIST |
| PHY_DL_RX_PSAVE_STS_<7:0> | Indicates if the Rx Lane has responded to a Power Saving Request and is in Low Power State |
| PHY_DL_TX_PSAVE_STS_<7:0> | Indicates if the Tx Lane has responded to a Power Saving Request and is in Low Power State |
| DL_PHY_IOBIST_PRBS_ERROR(7:0) | DL to PHY to indicate a PRBS error |
| DL_PHY_LANE_<7:0>(15:0) | 16 bits of Tx Data for this lane |
| DL_PHY_RUN_LANE_<7:0> | Indication to the PHY to run in high speed mode |
| DL_PHY_TX_PSAVE_REQ_<7:0> | Indication to the PHY to turn off the driver logic to save power |
| DL_PHY_RX_PSAVE_REQ_<7:0> | Indication to the PHY to turn off the receiver logic to save power |
| DL_PHY_RECAL_REQ_<7:0> | Indication to the PHY to run calibration on this lane |



LRDIMM & OMI DDIMM Latency Comparison

Based on 25.6G OMI SERDES Speeds

- OMI Buffer adds 7.16ns to unloaded LRDIMM memory access of 41.5ns
- In Loaded case OMI has 0ns Write and 3.94ns Read Latency Adder
- **Commands are in Memory Controller Queue**





OMI Maximum Bandwidth Performance Summary

From Simulation Data

| Read/Write Mix | OMI Bandwidth | % of Max DDR4 BW |
|-----------------------|---------------|------------------|
| 100% Read | 20.7 GBytes/s | 80.7% |
| 67% Read / 33% Writes | 20.5 GBytes/s | 80.2% |
| 100% Writes | 21.2 GBytes/s | 82.8% |

Setup

- 128 byte data
- Templates 0,1,5,9 for Tx and 0,1,4,7 for Rx
- Refresh cycles were not enabled
- MCBIST was not enabled
- Single rank was used
- CL-nRCD-nRP → 20/20/20
- DDR4 16Gb x4 3200

