# Intel and Micron Pathfinding Discussion

## 4-4-2023

# Attendees

- Intel attendees
  - Host - Dimitrios Ziakas
  - Bains, Kuljit S kuljit.s.bains@intel.com;
  - Fazio, Al al.fazio@intel.com;
  - Agarwal, Rajat rajat.agarwal@intel.com;
  - Kumar, Mohan J mohan.j.kumar@intel.com;
  - Benavides, Tony tony.benavides@intel.com;
  - Hinck, Todd A todd.a.hinck@intel.com;
  - Debendra Sharma, debendra.das.sharma@intel.com
  - Tomishima, Shigeki
  - Guliani, Sandeep
  - Pickett, Jay
- Micron Attendees
  - Steve Pawlowski – CVP, AMS
  - Mark Helm – Sr. Fellow, Pathfinding & Strategy
  - Brent Keeth -Sr. Fellow, Pathfinding
  - Balint Fleischer - Sr. Director, AMS
  - Ameen Akel – DMTS, Systems Research
  - Tony Veches – Director, Product Architecture
  - Mike Williams – Director, Tech and Products Strategy
  - Tony Brewer – Sr. Systems Architect
  - John Groves – Director, SW Engineering
  - Ilker Cebeli - Sr. Director, Tech and Products Strategy

# Meeting Agenda

- TCDRAM – Ameen Akel/Brent Keeth
    - Objective: Discuss potential L3 memory alternatives and identify collaboration
- LP in Servers – Steve Pawlowski
    - Objective: Share LP memory controller architecture ideas, and discuss potential collaboration areas
- CXL 3.0 Features – Balint Fleischer
    - Objective: Discuss potential CXL feature(s) and collaborations areas

# TCDRAM

DMTS, Ameen Akel
Sr. Fellow, Brent Keeth

# TCDRAM: An Inflection in Memory Systems

- TCDRAM: Closer integration of DRAM and leading-edge logic through hybrid bonding
  - Addresses bandwidth challenges of DSA design within silicon perimeter limits
  - TCO reduction: increased energy efficiency, faster compute and/or reduced silicon area

- Multiple stacks per ASIC possible → Linear increase in bandwidth per stack

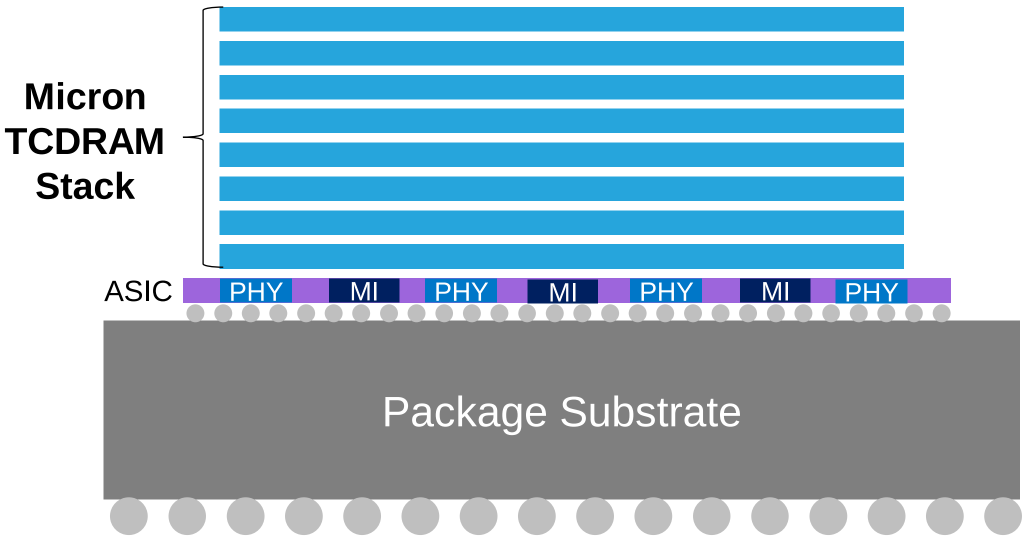
| | 3D TCDRAM v1 | 2.5D TCDRAM v1 | HBM2e | HBM3 |
|---|---|---|---|---|
| Bandwidth | 10TB/s | 3TB/s | 409GB/s | 1TB/s |
| 128B Energy/bit | 1.0pJ/b | 1.5pJ/b | 7pJ/b | 3.6pJ/b |
| 128B Power at Max Bandwidth | 80W | 36W | 22.9W | 28.8W |
| Page Size | 512B | 512B | 1KB | 2KB |
| Capacity | 24GB | 24GB | | |
| Timeline | 2026-2027 | 2026-2027 | | |

TCDRAM offers dramatic increases in bandwidth for the same system power.

# TCDRAM v1 Architecture Overview

- FROST Architecture: 10TB/s per stack at 1.0pJ/b (128B/page)
- Hybrid-bonded construction at 6u pitch: W2W stack construction, C2W to ASIC
- Optimized DRAM interface and timings
- Optimized, on-ASIC memory interface blocks (MIBs)
- Optimized, very low-energy communication circuits across the stack

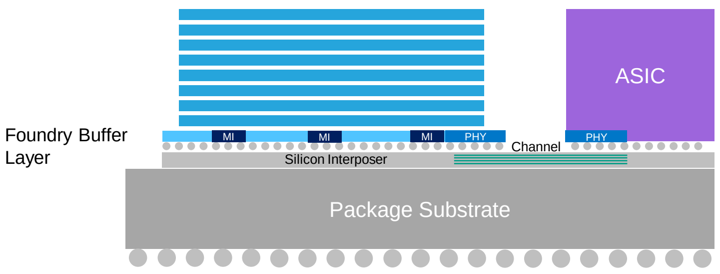
| | TCDRAM v1 Arch. |
|---|---|
| Channels | 512 physical channels<br>1024 pseudo-channels |
| Banks / Pseudo-channel | 8 |
| Page Size / Bank | 512B |
| Data IOs / Pseudo-channel | 64 |
| Die / Stack | 8 |
| Capacity / Stack | 24GB |
| Area / Stack | 107.7mm$^2$ |



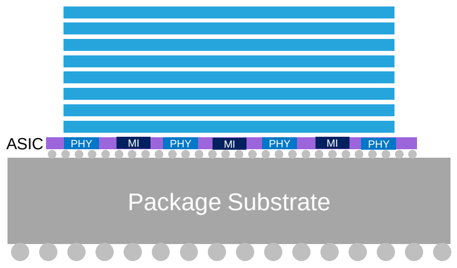
# TCDRAM Product Concepts

| | 2.5D TCDRAM v1 with Base | TCDRAM v1 |
|---|---|---|
| Bandwidth | 3TB/s | 10TB/s |
| 128B Memory Energy/bit | >1.5pJ/b | 1.0pJ/b |
| 128B/page Power | 36W | 80W |
| Estimated ASIC Energy/bit** (Memory + Data Movement) | ~1-2pJ/b | 0.25pJ/b |
| Estimated Total Energy/bit** | 2.5-3.5pJ/b | 1.25pJ/b |
| Pin Bandwidth | >24Gb/s | 1.28Gb/s |
| Physical Interface (# Data IOs) | HBM-like (1K) | Physical FROST (64K) |
| Logical Interface | Packet-based Protocol | Digital Read/Write Interface |
| Integration | 2.5D | 3D |
| ASIC Architecture Impact | Low | Medium+ |

* Still under definition
** Based on reticle-limited GPU architectures

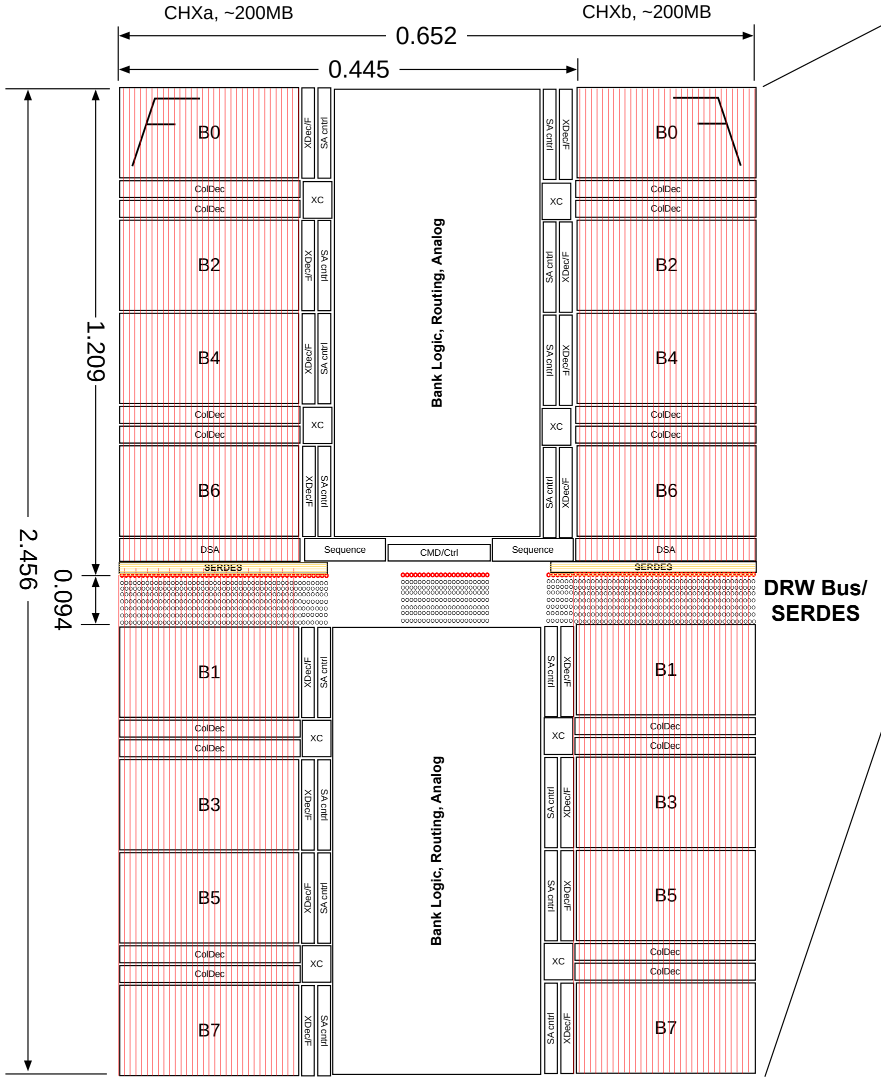



**TCDRAM v1 + Buffer**

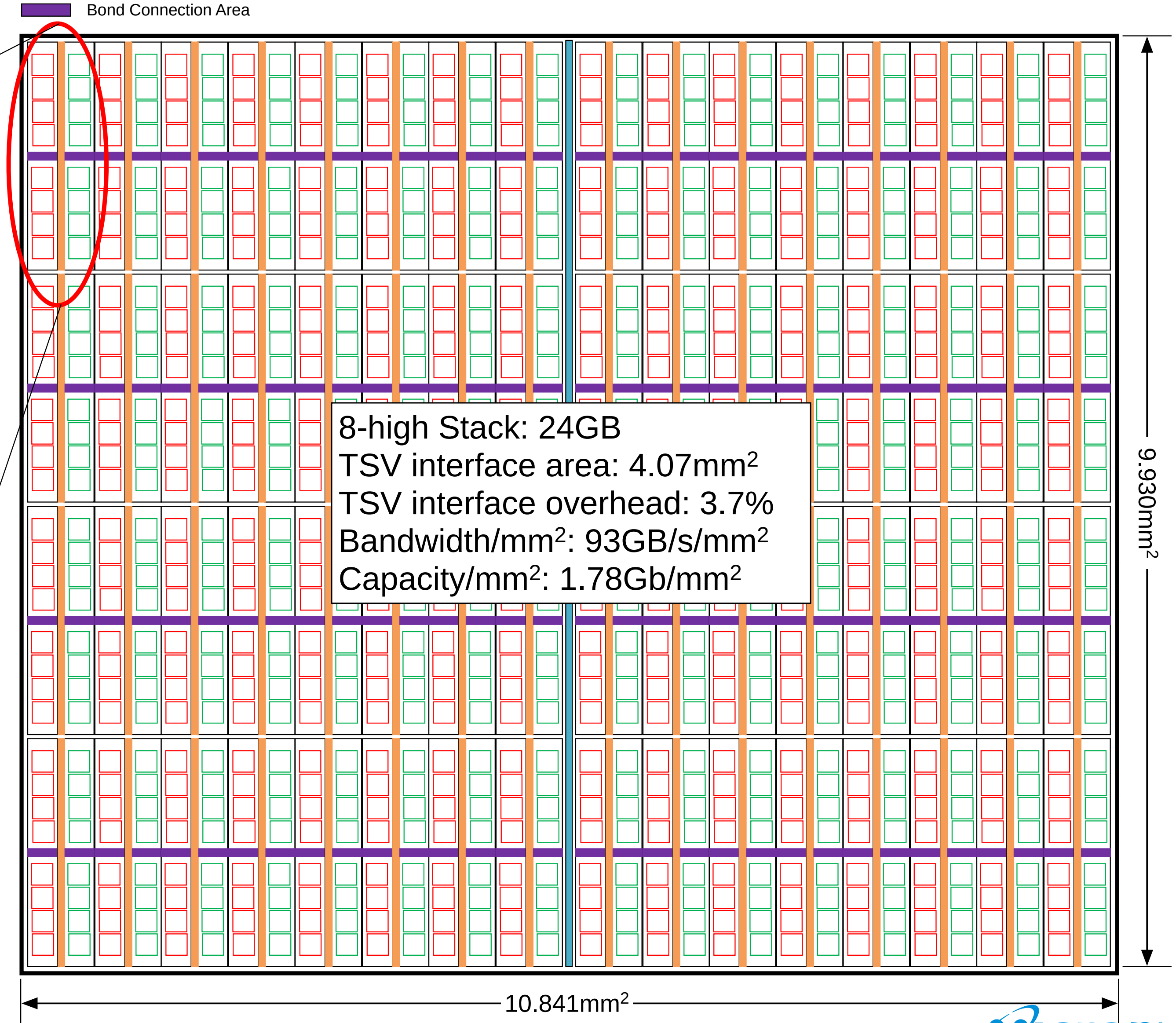



**TCDRAM** v1

# TCDRAM v1 Component Overview



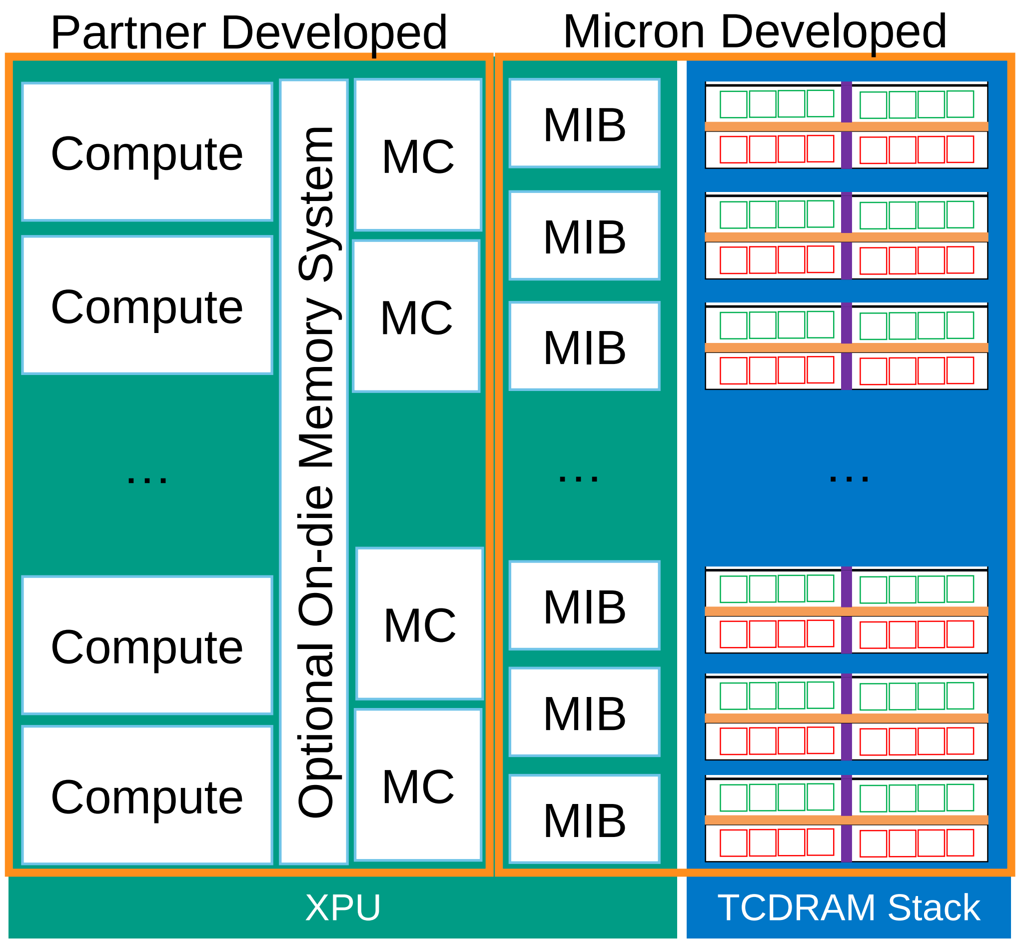

Channel pair area ~ 1.60 $mm^2$




Total Area: 107.7$mm^2$

# TCDRAM Memory Interface Block (MIB)

- On-logic-die extension of memory-centric logic
  - Enables more efficient and performant system solution
  - Reduces waste by unifying often duplicated memory logic
- Opportunity to reimagine architecture / memory interface
  - Combined RAS implementation → no RAS / ECC overlap
  - Unified ECC and media management
  - Included low-level scheduling simplifies memory system
  - Memory built-in self test (BIST) engines
  - Optimized logic-memory interface and IOs
- MIB evolves with newer generations of TCDRAM
  - Evolve around in-logic digital interface
  - Allow physical TCDRAM interface to improve over time without disrupting accelerator architecture and design

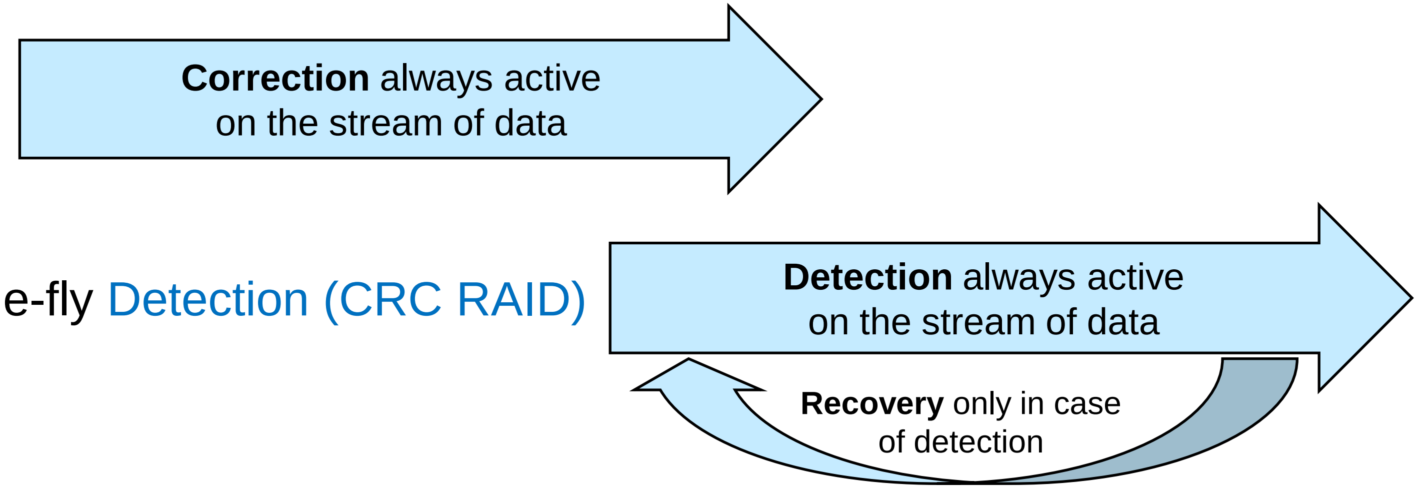


# LP5 Considerations in Servers

## CVP AMS, Steve Pawlowski

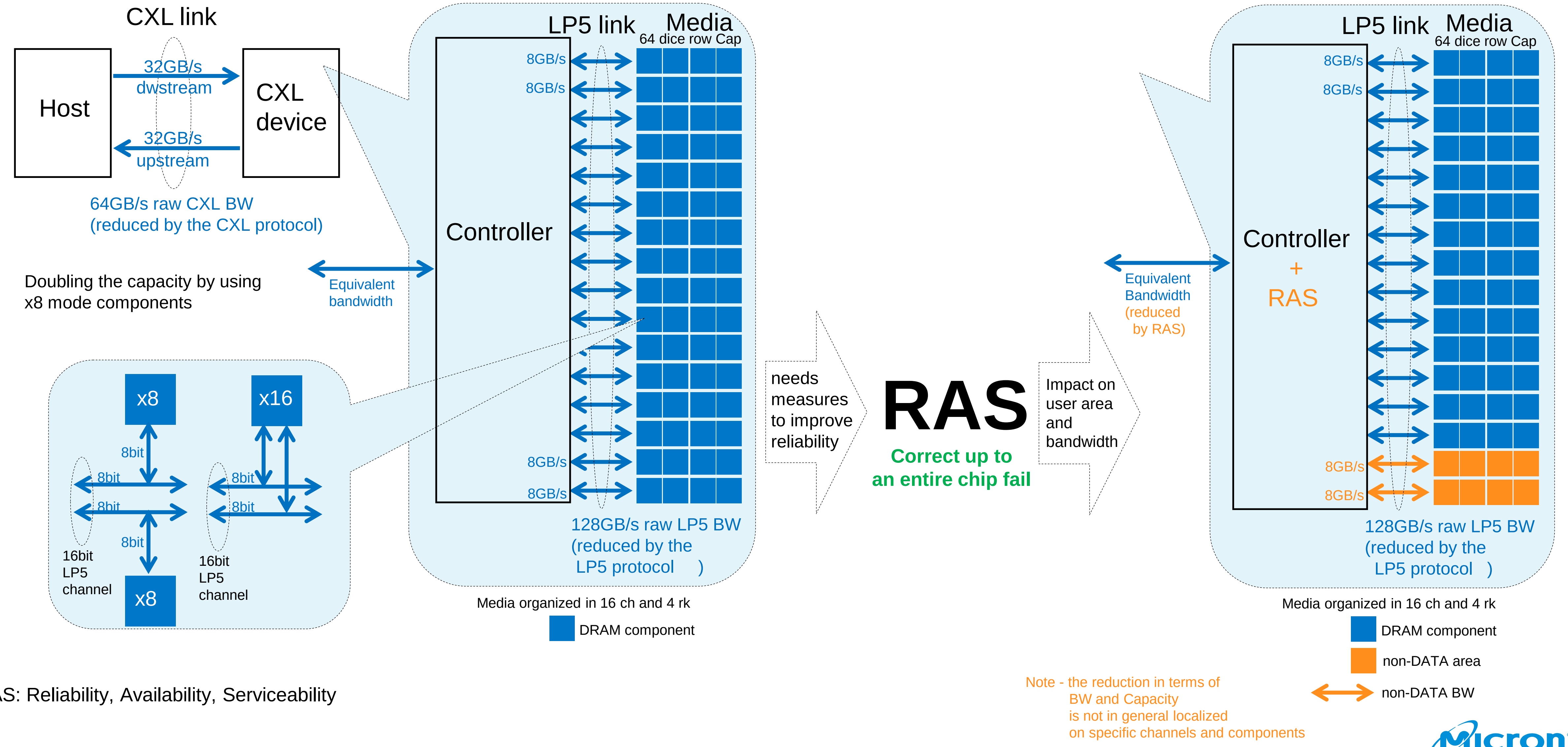
# LP5 RAS

- Objective : Correct up to a failure involving an entire chip (ChipKill-like protection)
- LP5 characteristics (like x16 mode) require the development of a dedicated RAS solution
- Two approaches
  - On-the-fly Correction (LPCK)
  - Recovery assisted by the on-the-fly Detection (CRC RAID)




- In both cases data stored on different components are bound together (→ BW reduction)

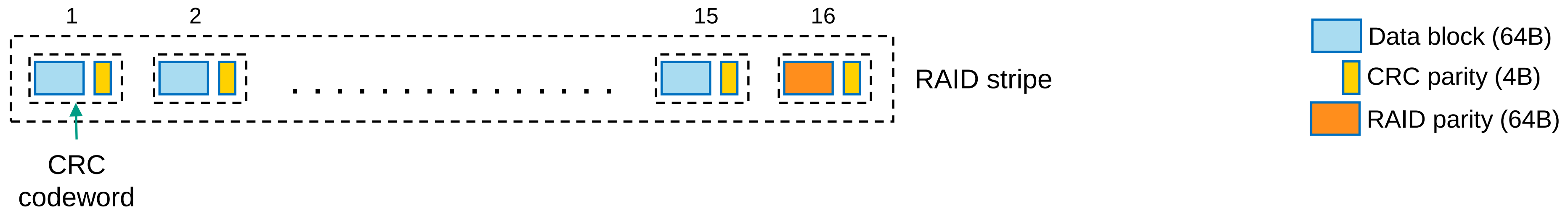
# RAS in the CXL device: impact on BW and Capacity



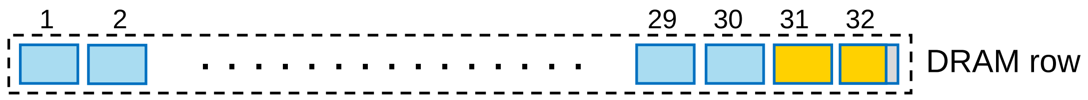

- RAS: Reliability, Availability, Serviceability

# The CRC RAID solution with LP5 components

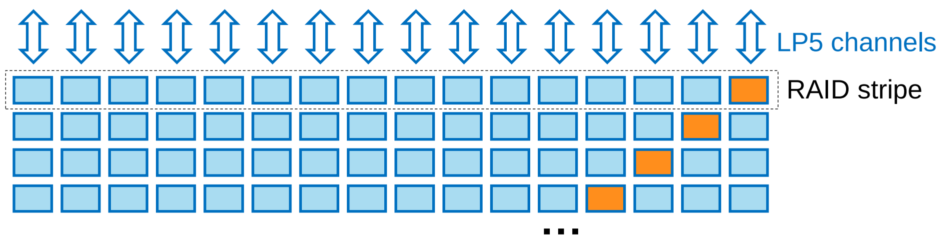
CRC RAID logical representation




CRC bits are placed inside the DRAM row, reducing the user capacity

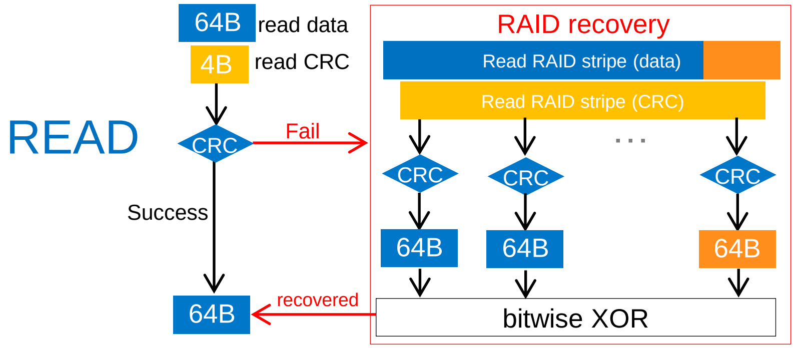



RAID parity rotation for uniform channels utilization




Amplification on READ

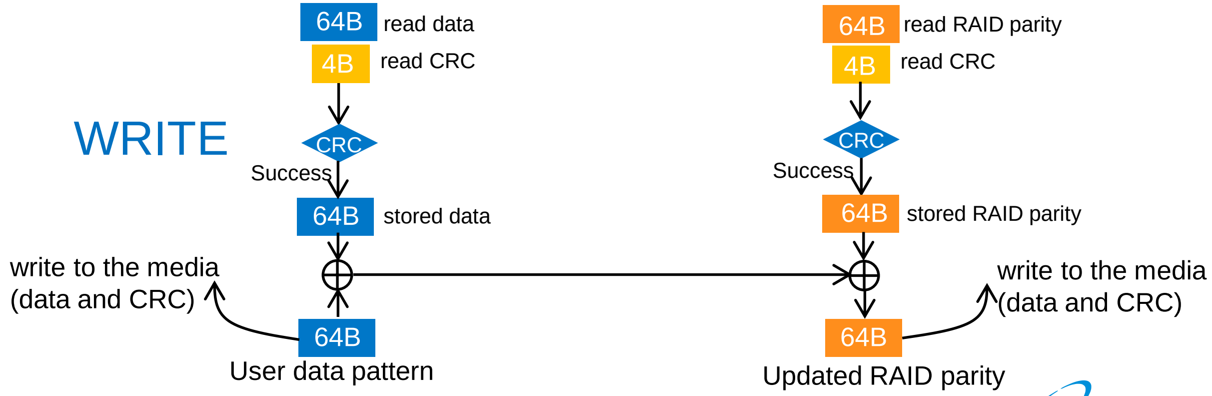
1 host read → 1.5 LP5 read (BL32 + BL16) $\boxed{\times\ 1.5}$




READ operations access a single channel

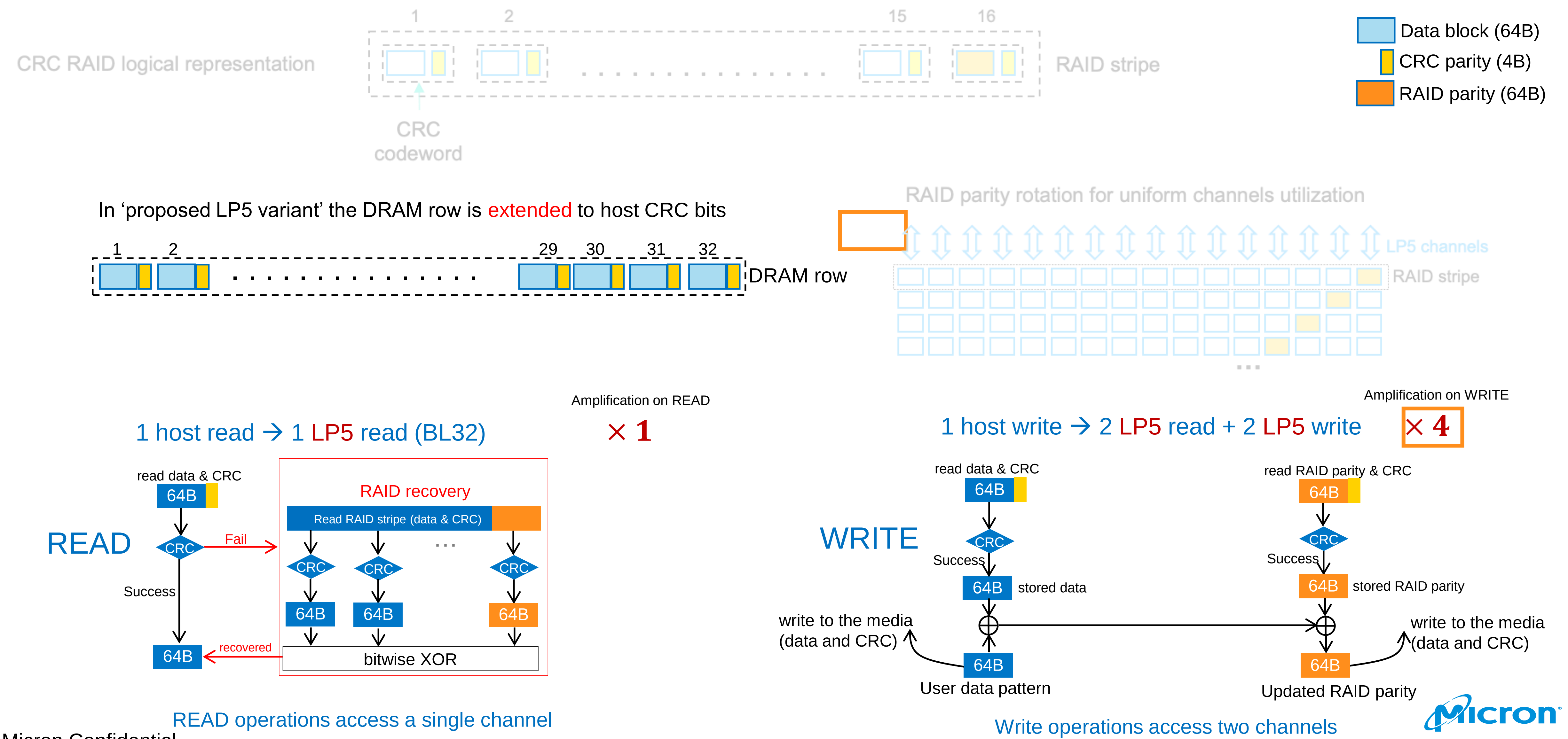
Amplification on WRITE

1 host write → 3 LP5 read + 3 LP5 write $\boxed{\times\ 6}$




Write operations access two channels

# The CRC RAID solution with LP5 components

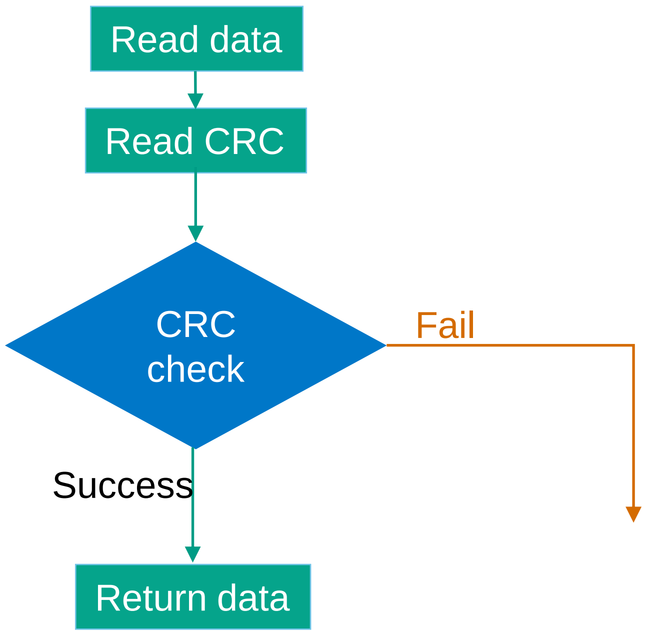


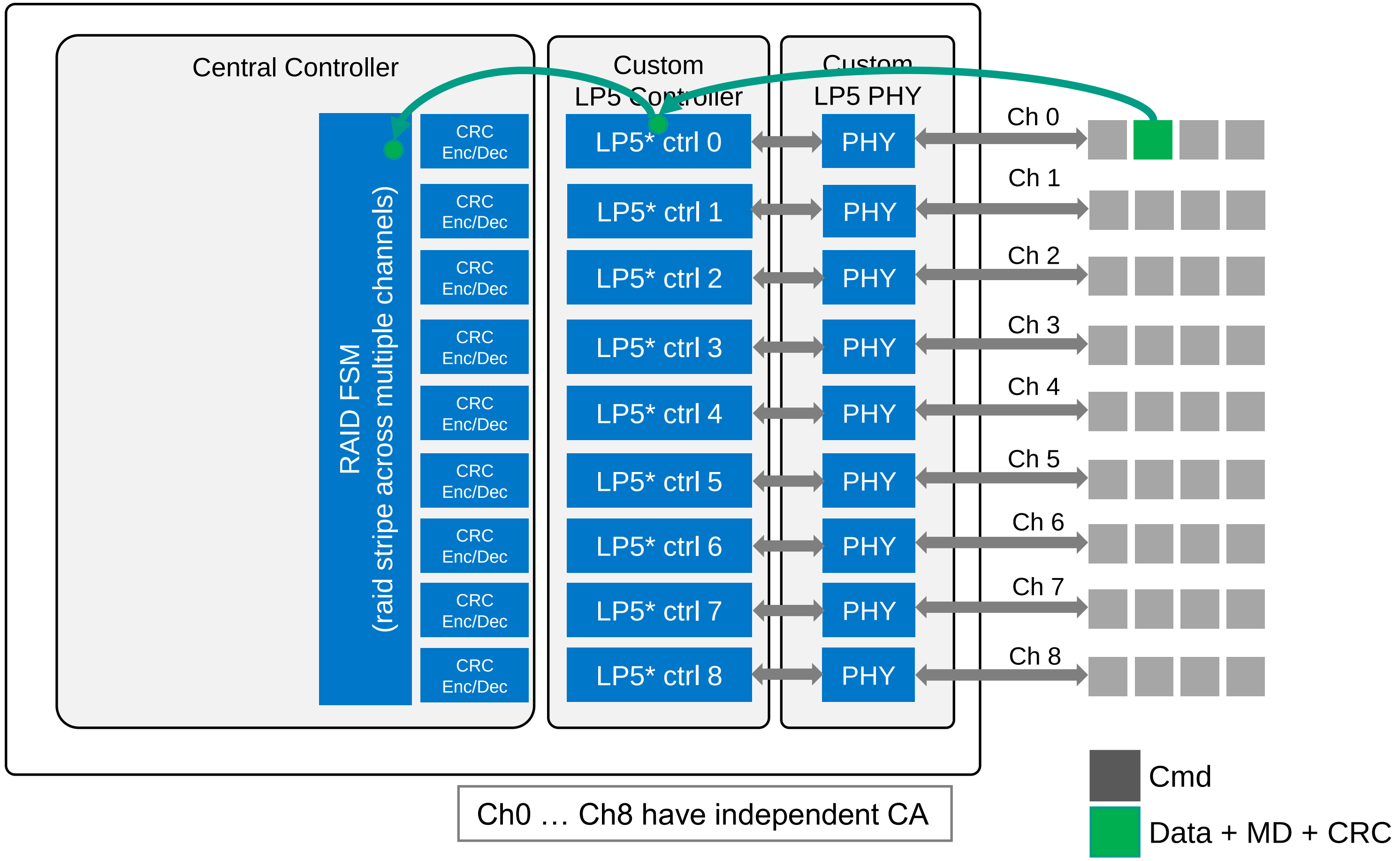
# CRC+RAID (8+1): 64B read command flow (No Errors)

64B Data | 32B: CRC+MD

1x host read ➔ 1x mem read + 1 CRC read

Check (CRC) and Recovery (RAID)



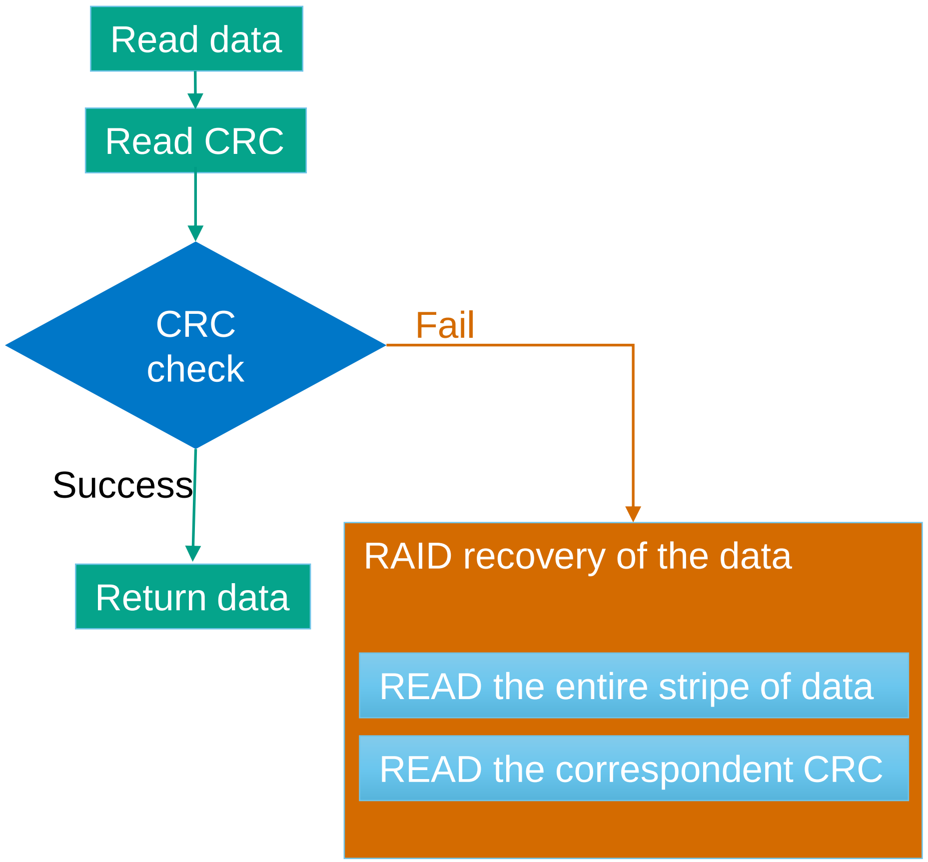

1x host Read ➔ 1x mem read + 0.5x mem read

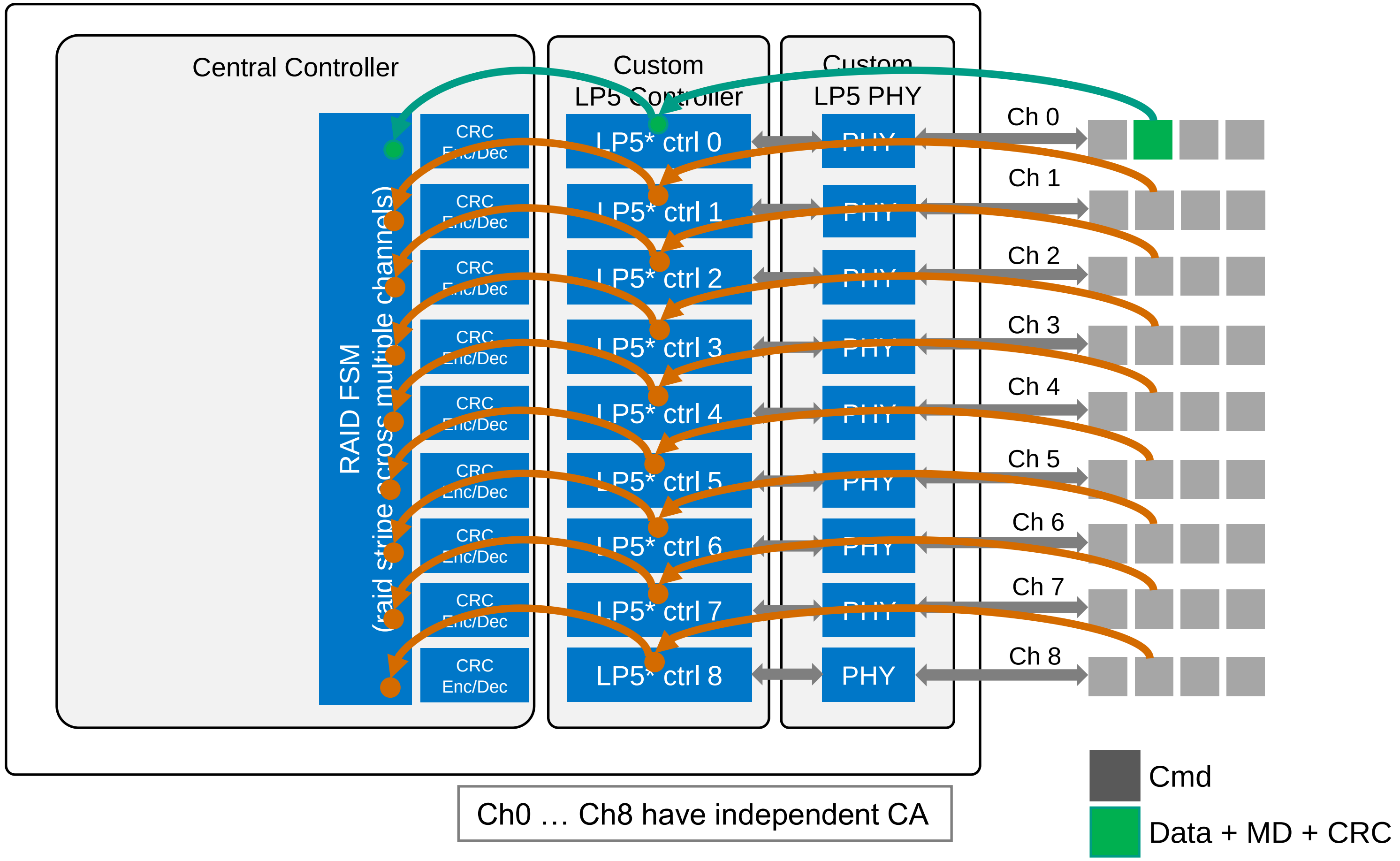
# CRC+RAID (8+1): 64B read command flow (No Errors)

64B Data | 32B: CRC+MD

1x host read ➔ 1x mem read + 1 CRC read

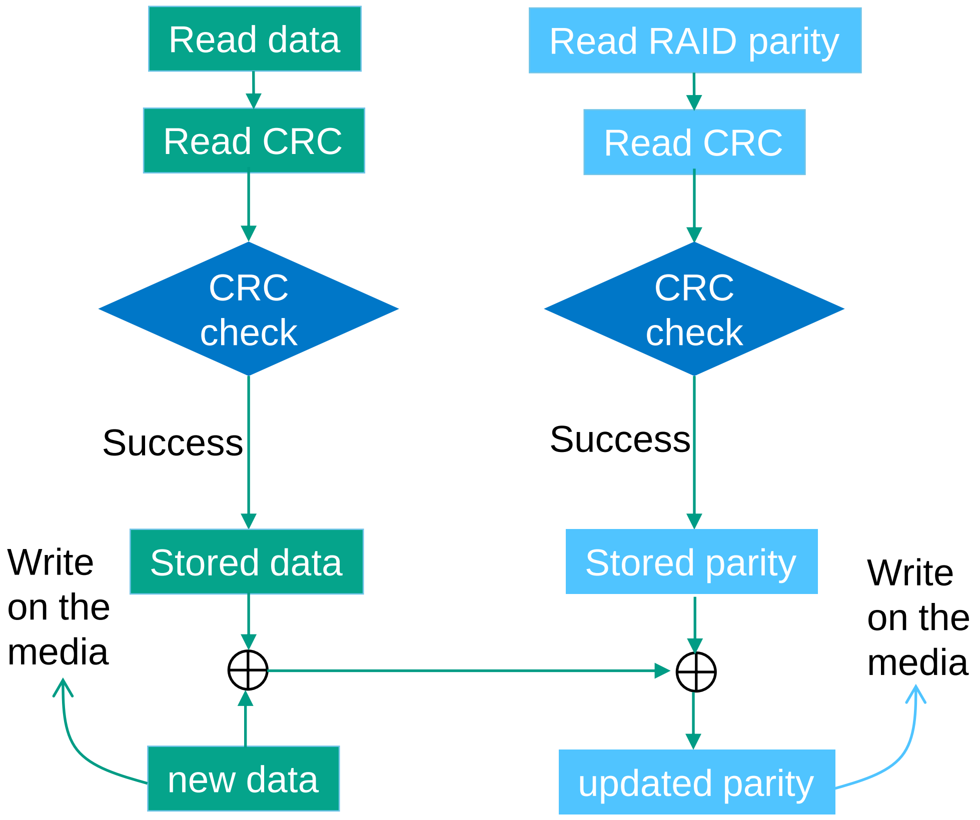
Check (CRC) and Recovery (RAID)

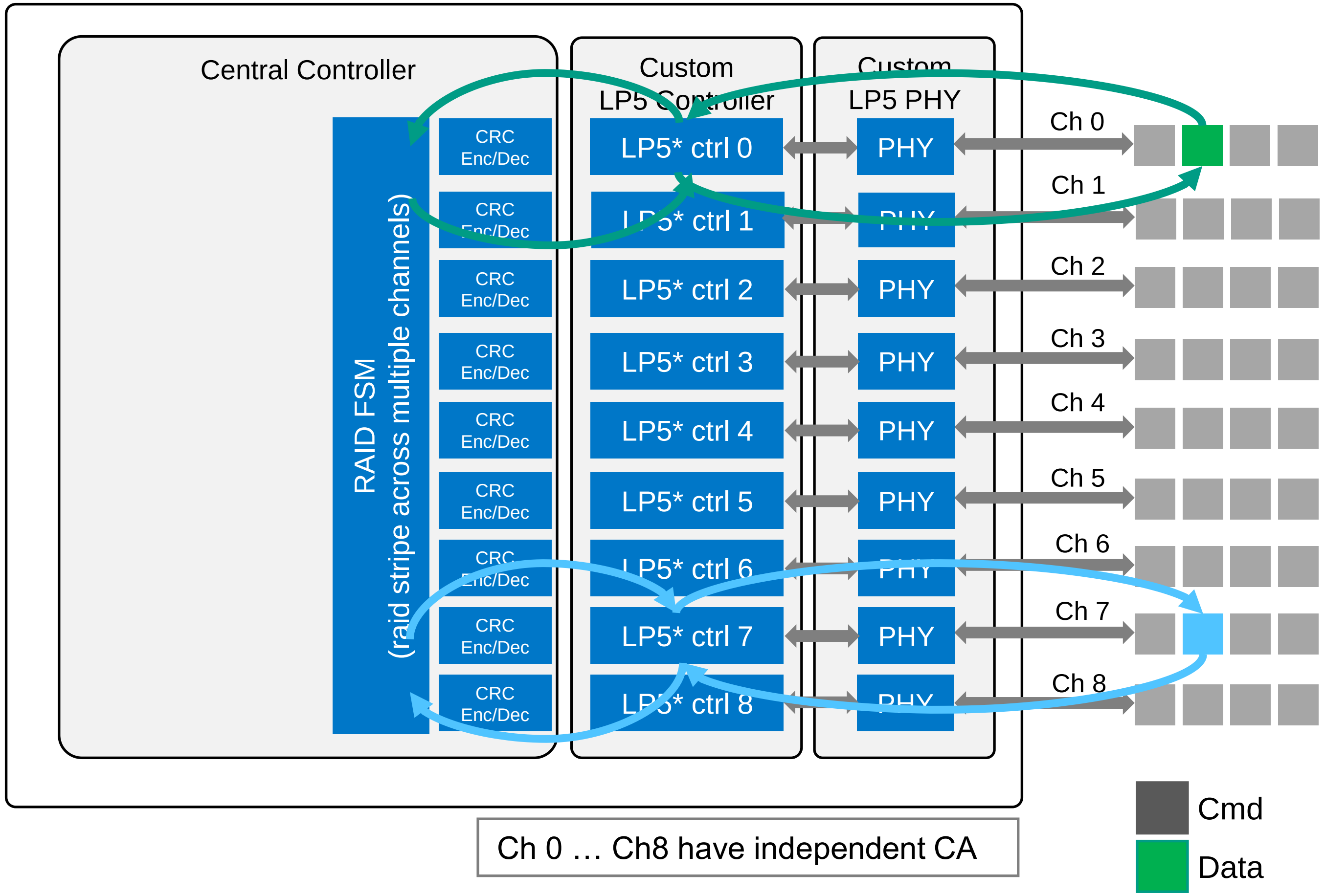


# CRC+RAID (8+1): 64B write command flow (No Errors)

Read (Data and CRC), Write Data and Update (CRC)



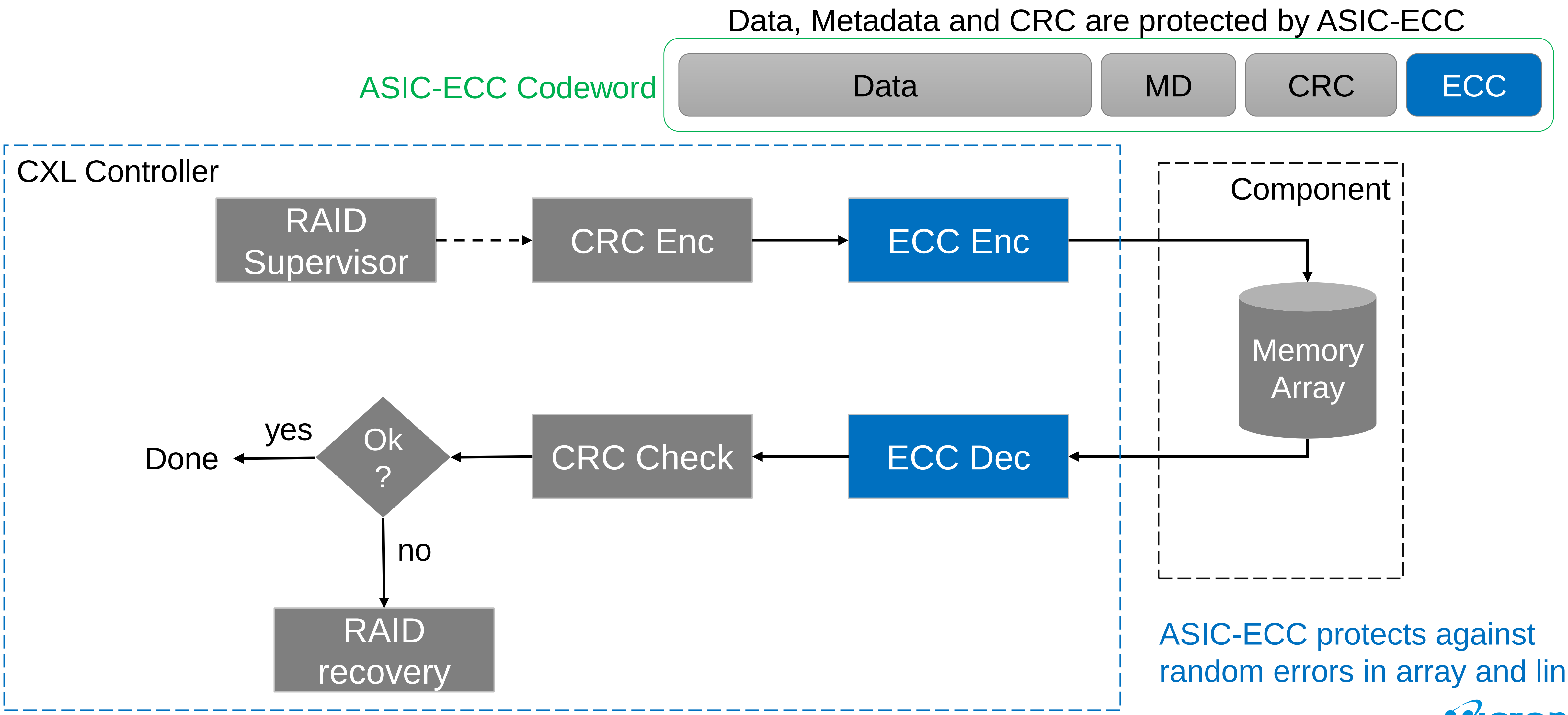

Cmd

Data

RAID parity

1x host write ➔ 3x mem read + 3x mem write

# RAS Solution for On ASIC ECC (LP5) no on-die ECC




ASIC-ECC protects against random errors in array and link

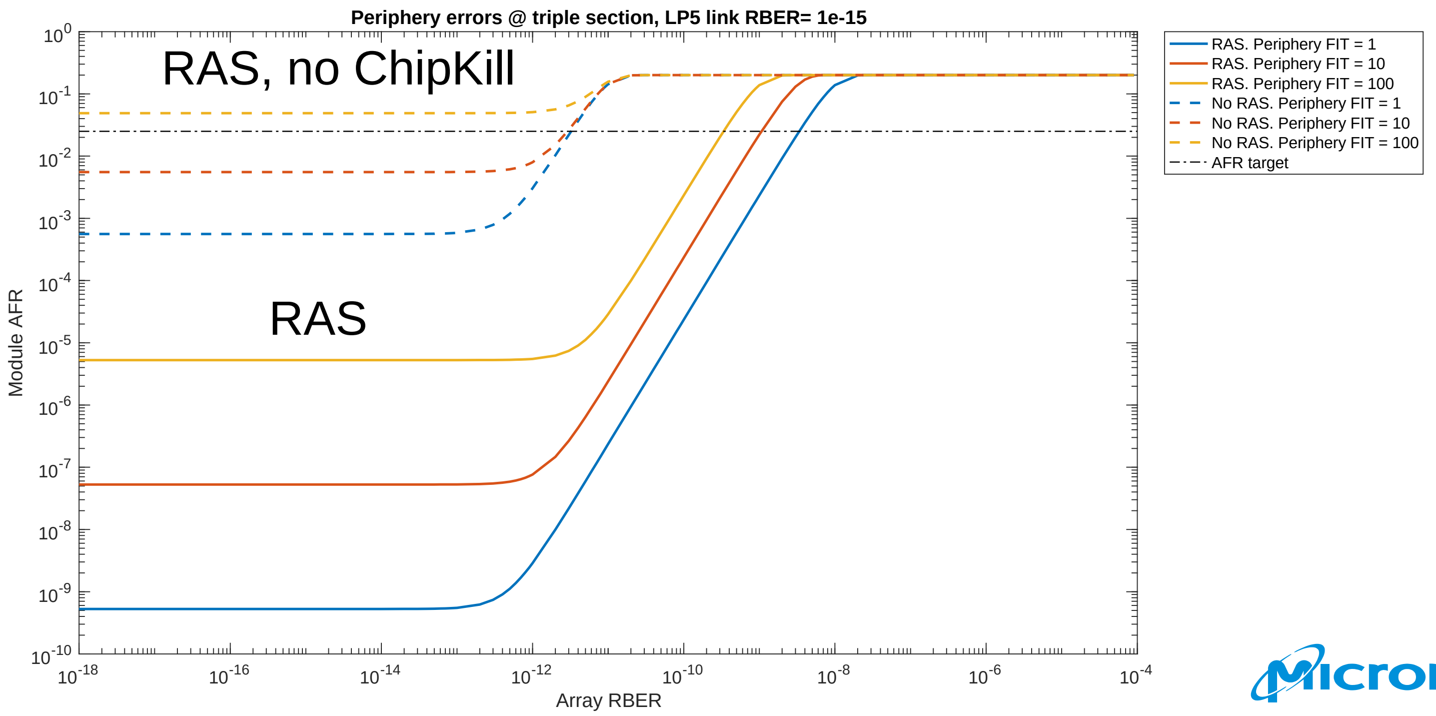
# Some Considerations w/RAS Solution

# Reliability

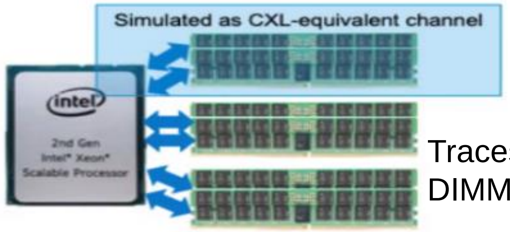
- **Module AFR** (with 64 components) with and without RAS
  - Hypothesis: A periphery failure is a hard failure and impacts a DRAM triple-section (~3000 rows)

- LPCK and RAID are in the same ball-park with respect to reliability performances
  - ASIC reliability not considered

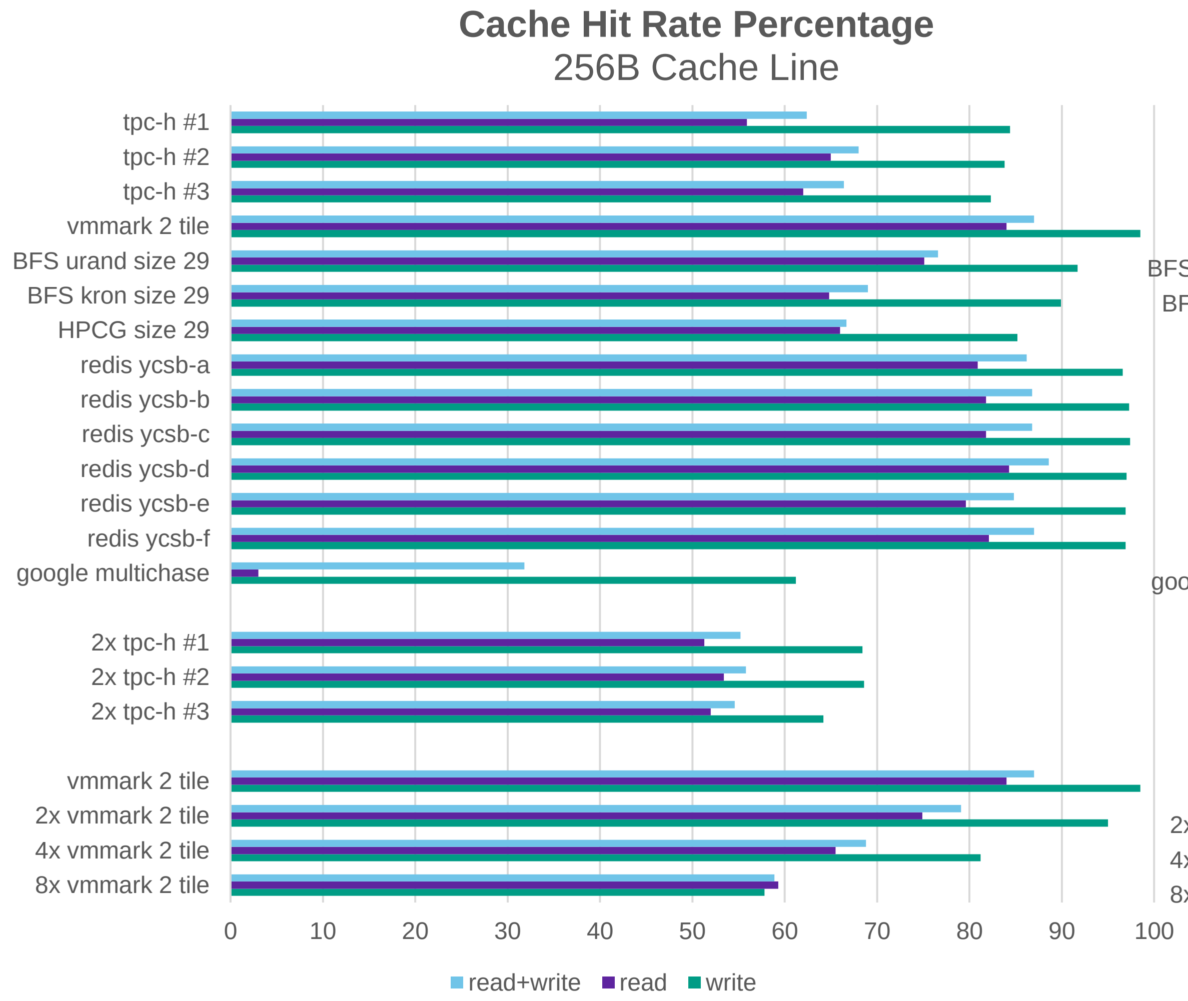
| Periphery FIT | AFR No CK | AFR RAID/LPCK | SDC RAID/No CK (CRC28) | SDC LPCK (CRC14) |
|---|---|---|---|---|
| 1 | 5e-4 | 5e-10 | 2e-12 | 3e-14 |
| **10** | **5e-3** | **5e-8** | **2e-11** | **3e-12** |
| 100 | 5e-2 | 5e-6 | 2e-10 | 3e-10 |

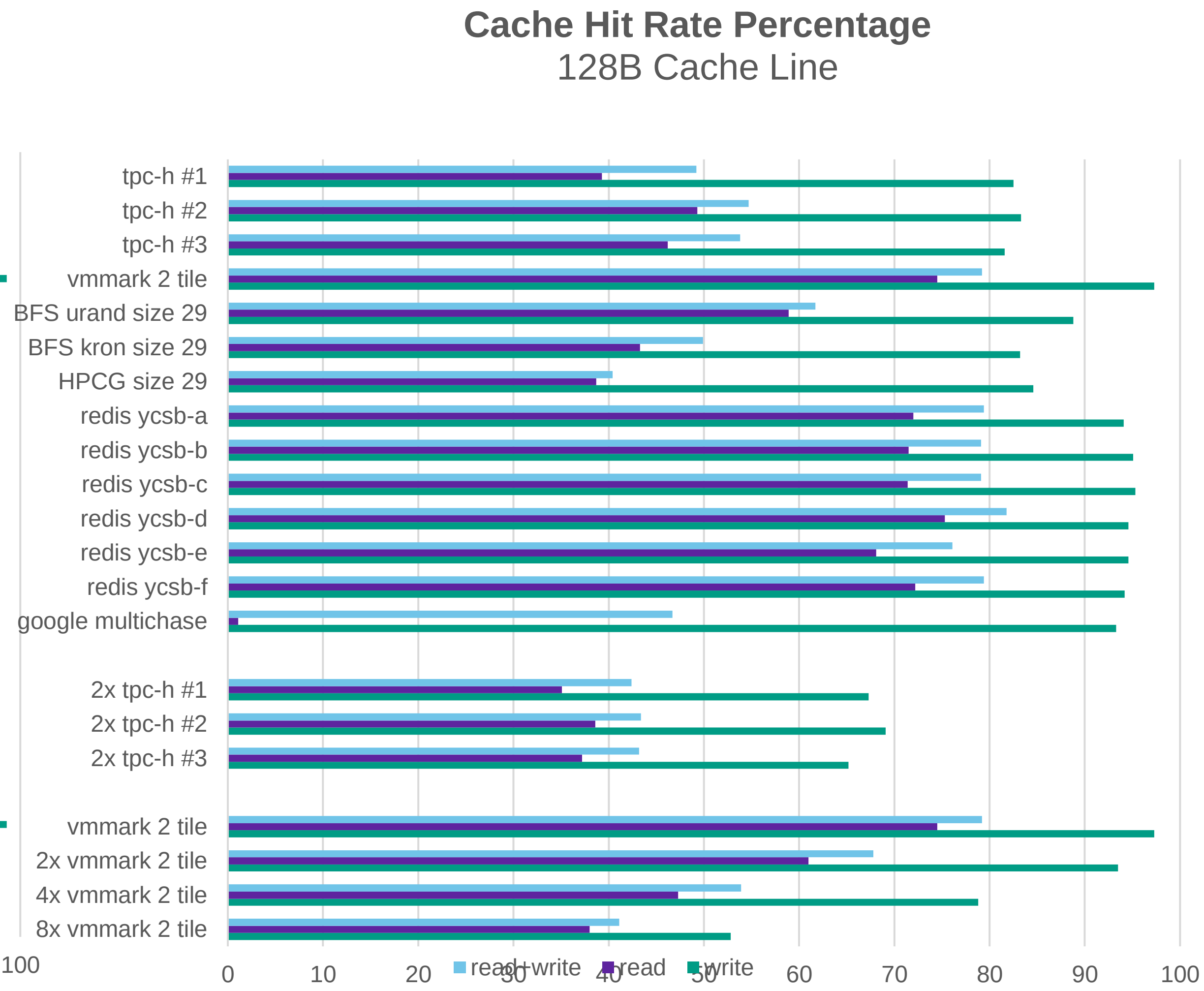


# Memory side Cache hit rate for different benchmarks




Traces captured from 2-channel DDR4 DIMM on Cascade Lake

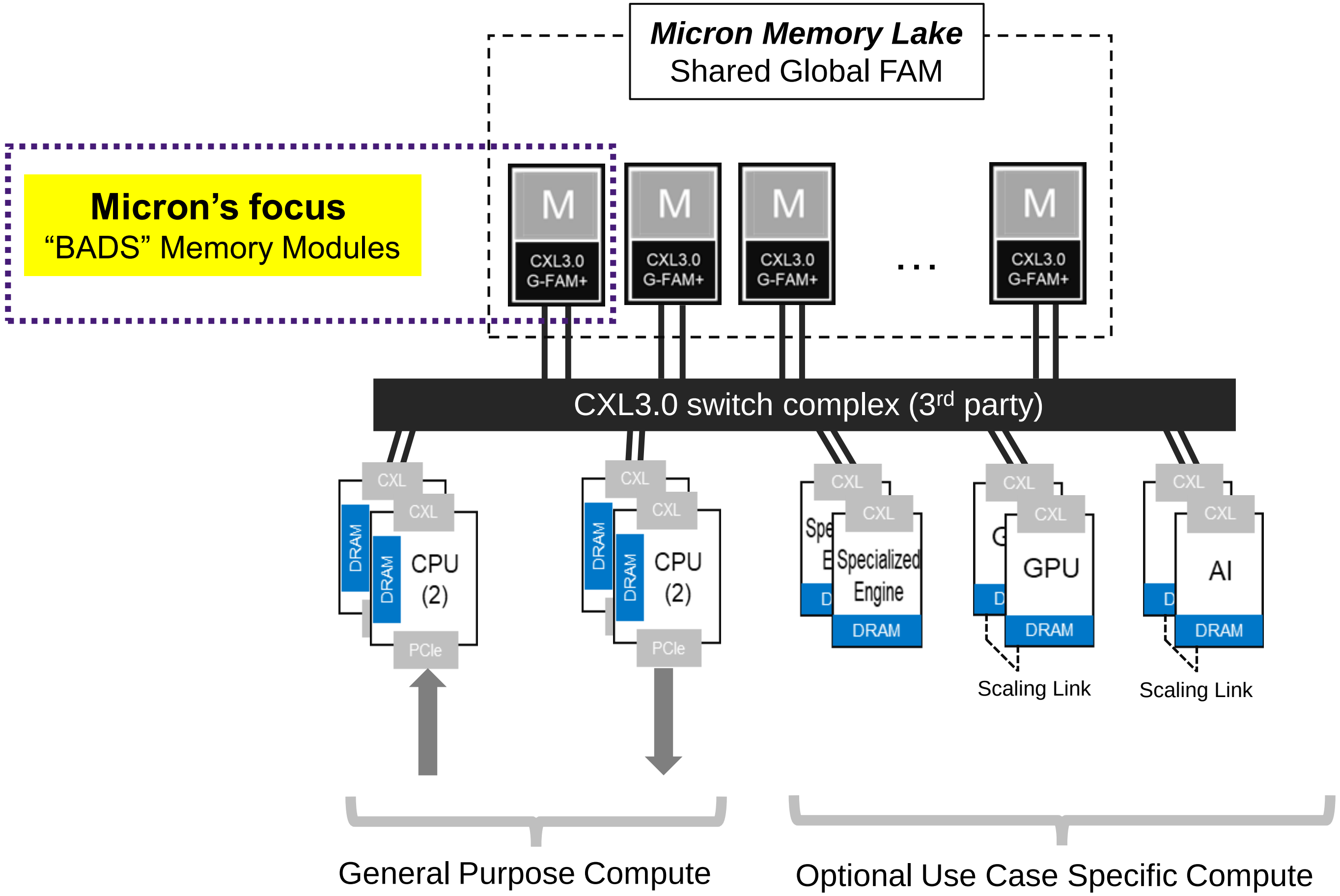


# Memory Centric Computing and CXL3.0

## Sr Director, Balint Fleischer, AMS/SMS

# Micron’s Memory Centric System Architecture vision




**Value Prop: accelerating “Time to Insight”**

**Initial target use cases:**

- AI preprocessing pipeline
- Graph Analytics
- Graph Inference
- Similarity Search
- Spark ML pipeline
- Multi-criteria Recommender System
- Hyperdimensional Data Analytics
- LLM Systems

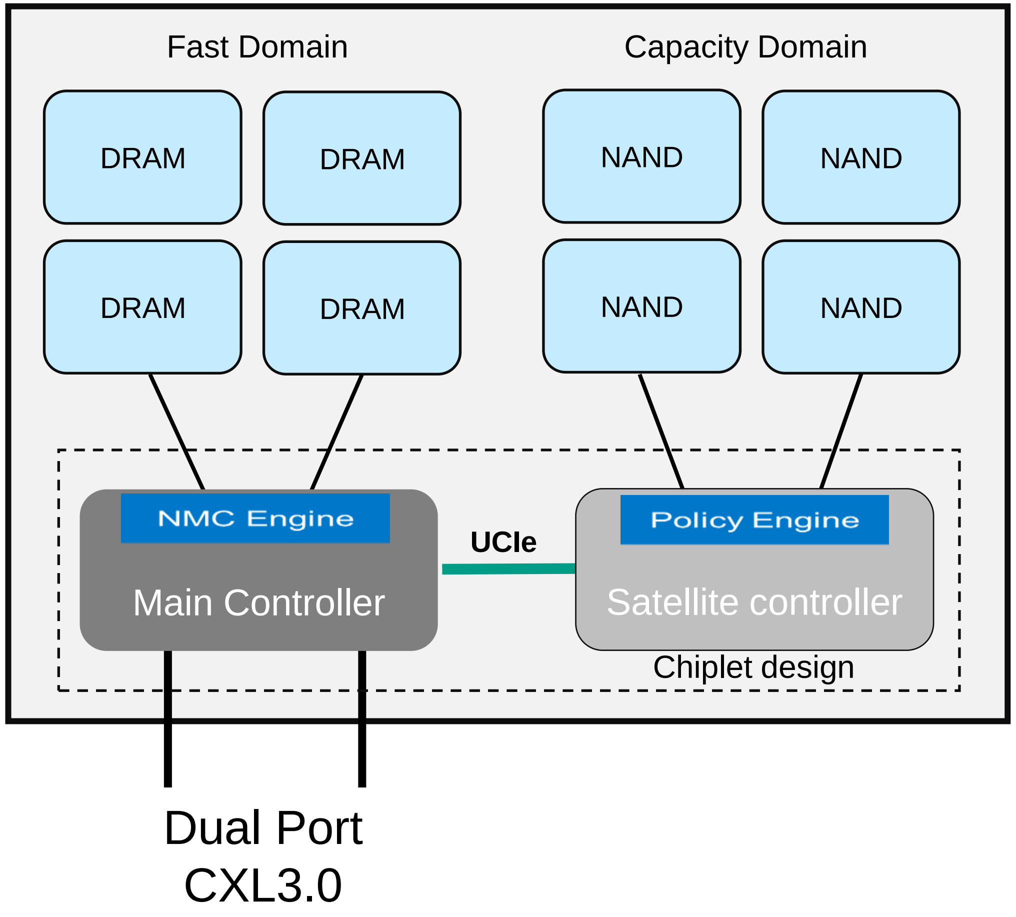
**Collaborating with:**

- SW ecosystem (Applications, Middlewares)
- HW ecosystem (CXL fabric switches, SiPH)

Private Memory

# BADS Module High Level overview

E3.S FF



# Proposed HW discussion topics

- CXL features to support GFD integrated Near Memory Computing (NMC)
    - SW Coherency with multiple cache hierarchies
    - Scalable ATS for GFD with integrated NMC
    - GFD to hosts interrupt mechanism
- CXL features to support redundant fabric failover
    - Reissue a timed out request on redundant fabric
- Enhancing DSA for GFD use cases
  – GFD devices will present different latency / bandwidth characteristics for host-based data movers
    - Supporting GFD devices with >= 1 uSec of latency
    - Mapping DSA to CXL.  CXL.mem and/or CXL.io (UIO), 64B and/or 256B transfers?

# Proposed SW Discussion topics

- Global Virtual Address, usage and implementation
- A Scalable, low latency NMC API definition
- Global FAM memory allocator and name server
- “Tagged DAX capacity” support
- FAM operation failure modes


Micron Confidential

# Proposed Collaboration Goals

- Strive toward “Better together” – enabling scalability of emerging workloads
- Support  Intel’s Heterogeneous Computing with Micron’s FAM
- Coordinate use case and ecosystem enabling
- Collaborate on CXL and on Linux to grow their support of Fabric use cases