



SAMSUNG



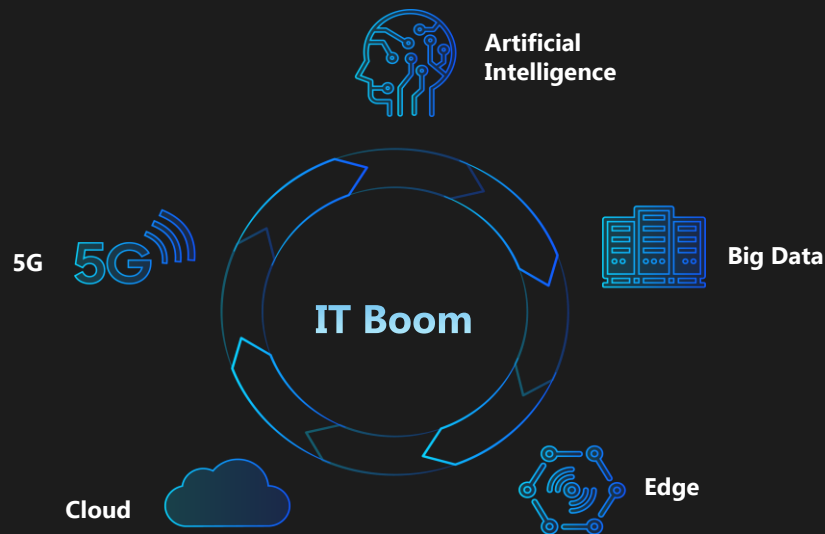
Flash Memory Summit

# Data Centric Compute and Data Tiering with CXL

David McIntyre, Product Planning  
Samsung Semiconductor Device Solutions America



# Industry Trends and Challenges



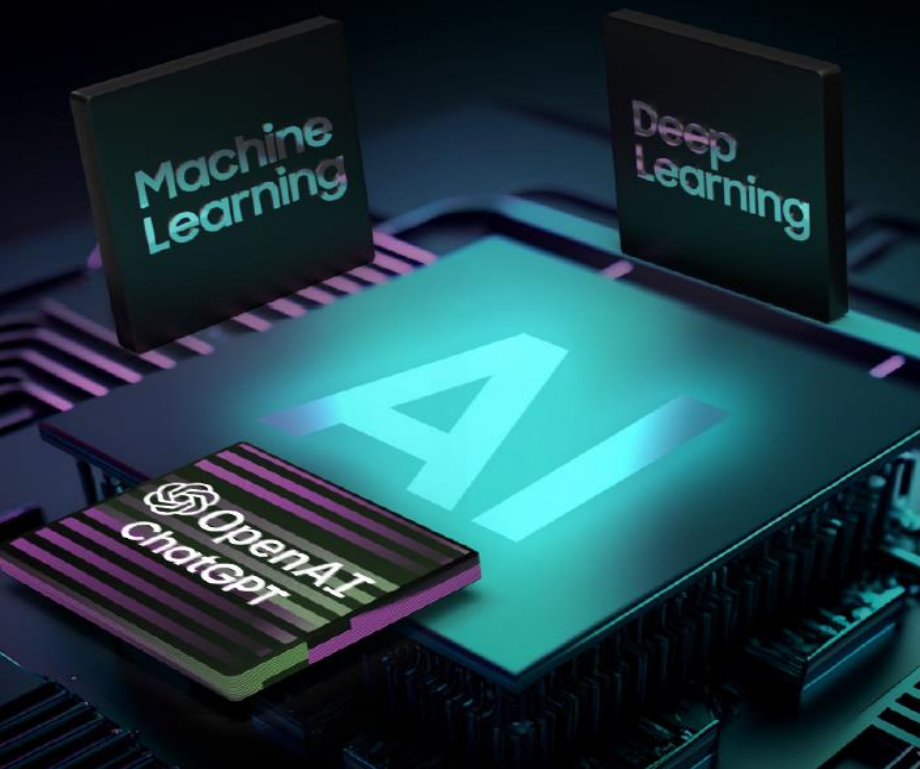
- Massive growth of data-centric technologies and applications
- Memory bandwidth and density not keeping up with increasing CPU core count
- Gap between CPU and memory performance steadily increasing
- Need a next gen interconnect for memory bandwidth/density expansion, heterogeneous computing and memory disaggregation

# In the Era of AI & ML

Swift increase in demand for memory capacity and performance



Flash Memory Summit



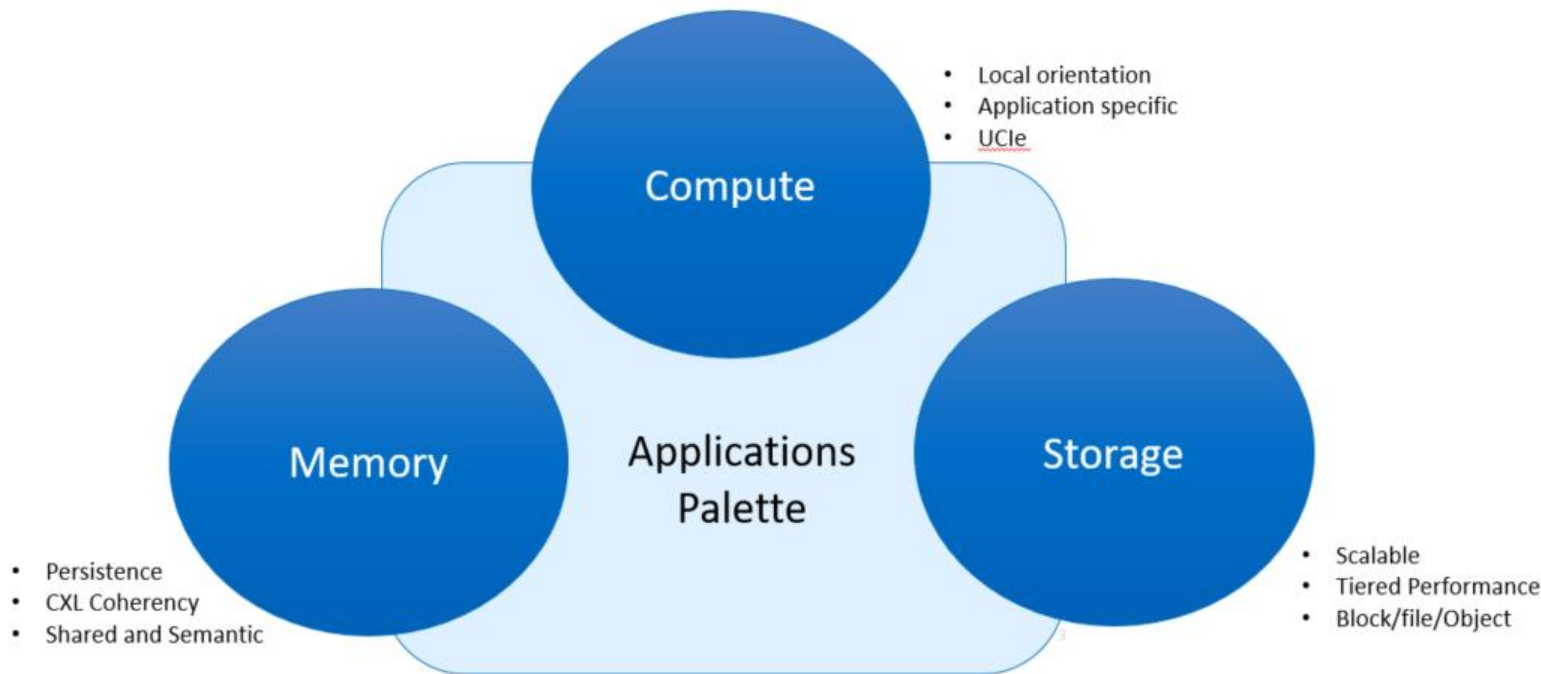
## Language Model Sizes to Dec 2022



\*Number of parameters/ Source: <https://lfearchitected.ai/models>



# Balancing Application-Driven Resources

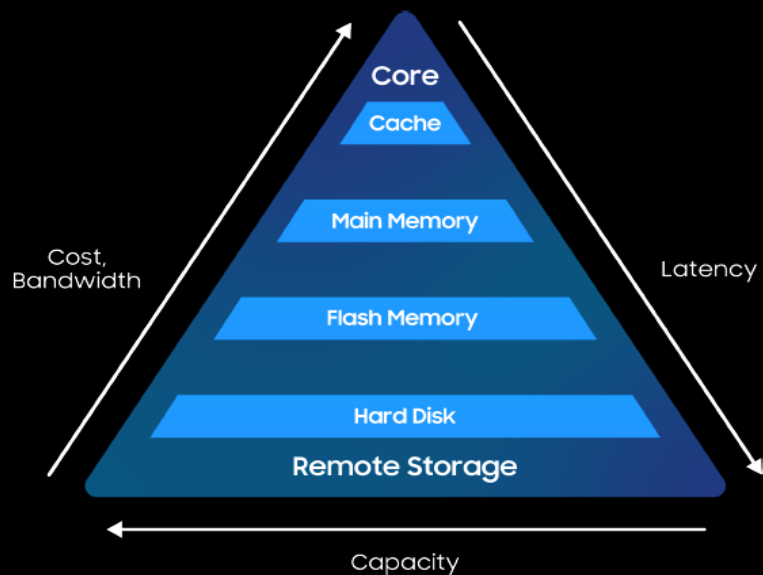


# Memory Hierarchy

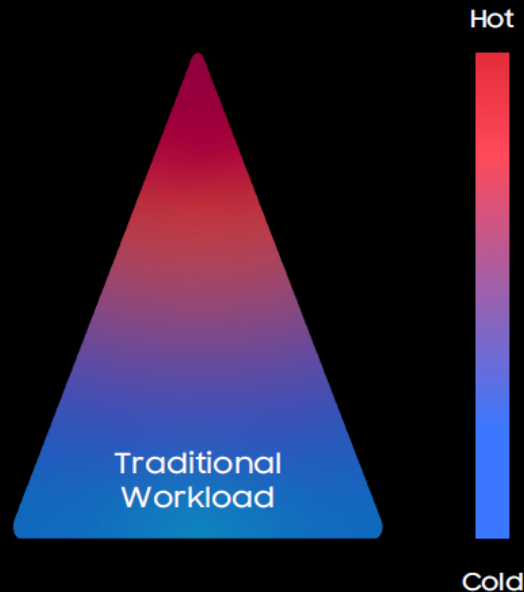
Keep hot data close to CPU using data locality



Flash Memory Summit



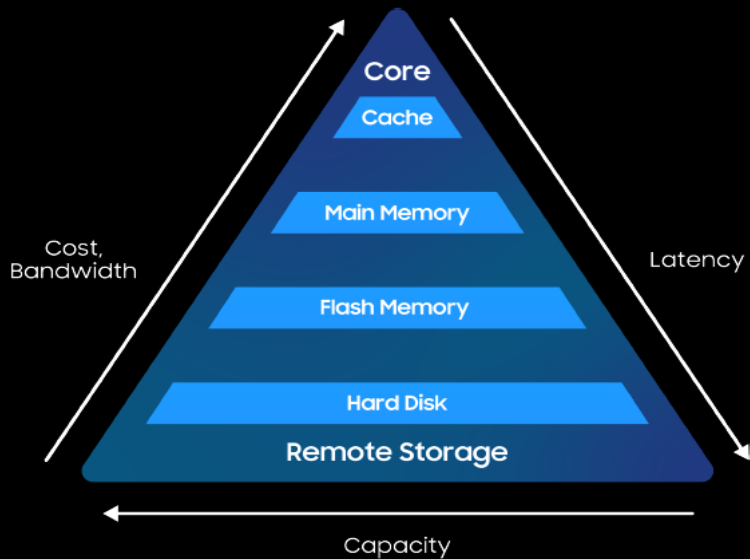
Memory Hierarchy



Traditional Workload

# Memory Hierarchy Disparity for Modern Workloads

Not all workloads exhibit the conventional pattern of data locality



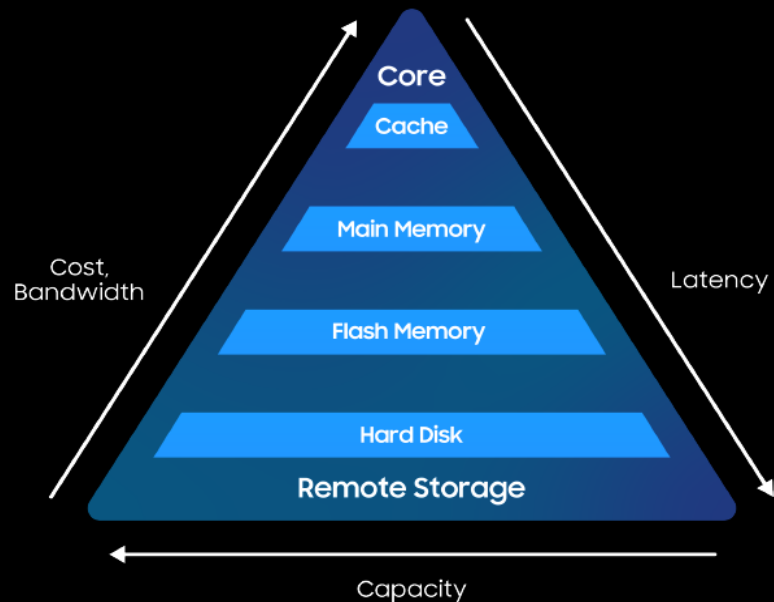
Memory Hierarchy



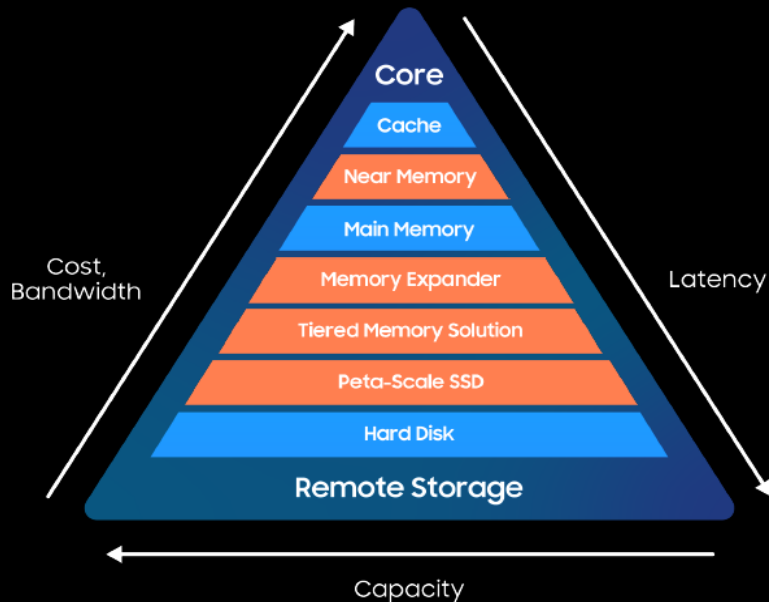
Modern Workload

# New Memory Hierarchy

Deeper and more efficient memory hierarchy to fill the performance gap



Old Memory Hierarchy



New Memory Hierarchy



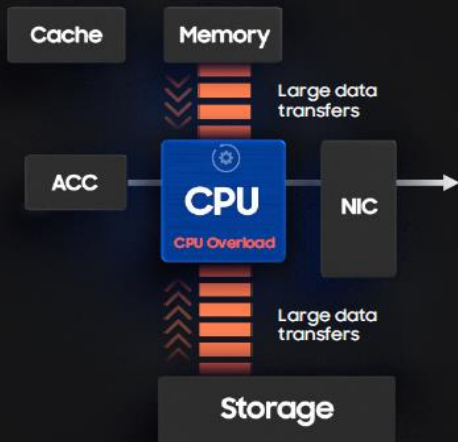
# Data-Centric Computing Concept

Move the computation to the data for large datasets



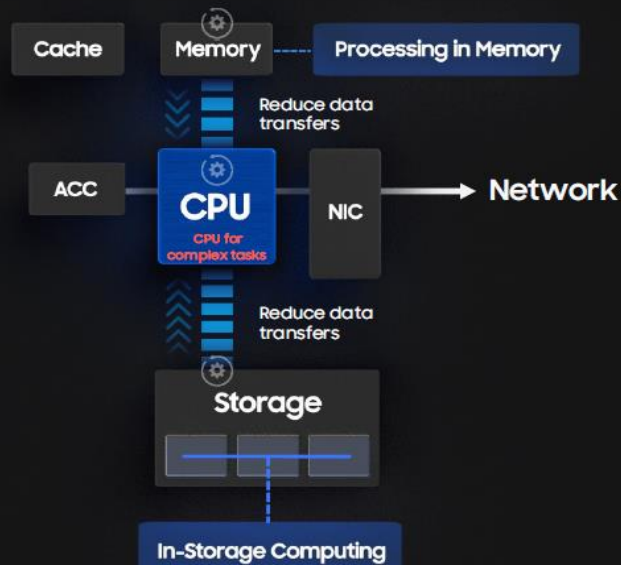
Flash Memory Summit

## CPU Centric Computing



Compute Near  
the Data

## Data-Centric Computing





## Move the computation to the data for large datasets



# Data-Centric Computing Benefits

Power-optimized scalable processing for large data



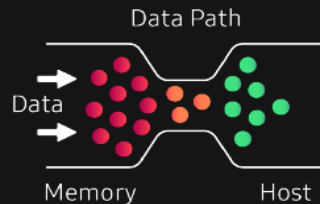
Flash Memory Summit



**Low Power  
Computing**



**Data  
Reduction**

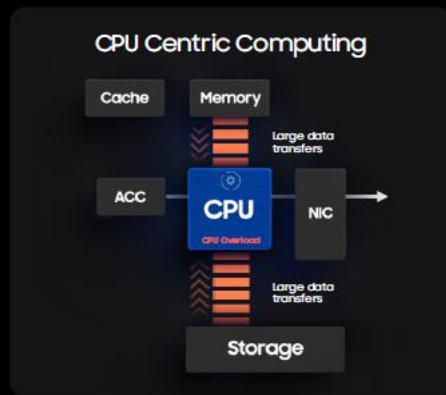


**High Effective  
Bandwidth**

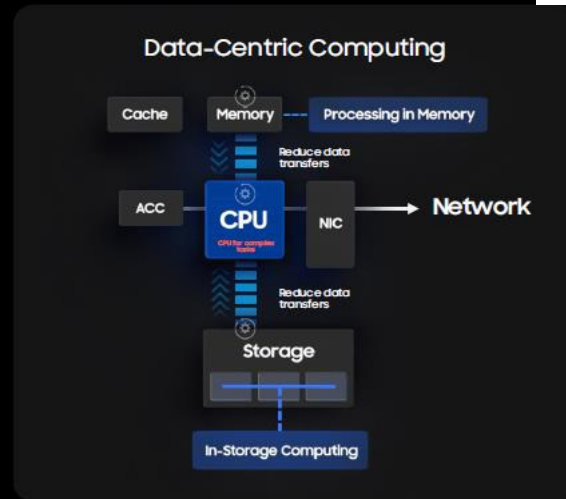


**Scalable  
Computing**

# Challenges in Data-Centric Computing



Compute  
Near the Data



Interface  
Interference



Killer  
Application



Ease of Use



# CXL™ Features - Summary

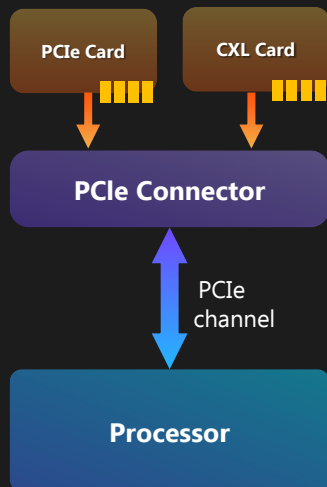
- CXL consortium - open standard (non-proprietary)
- Regular specification updates (CXL 1.1, CXL 2.0, CXL 3.0)
- High speed, low latency interconnect
- PCIe Physical layer (PCIe 5.0, PCIe 6.0)
- Supports various types of memories (volatile, non-volatile)
- CPU and CXL device memory coherency
- Supports switching (multi-level), memory pooling & sharing
- Direct peer-to-peer (P2P) communication



**CXL Momentum**

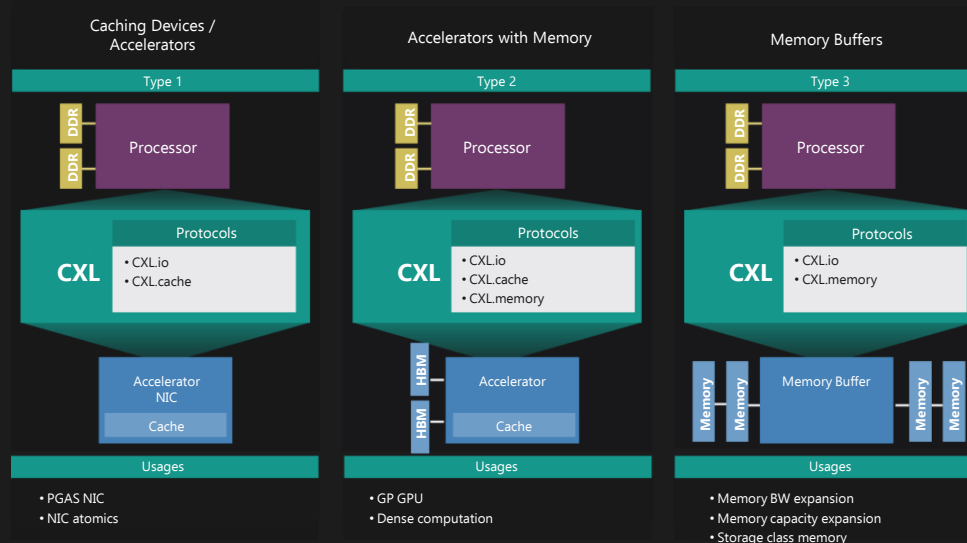


# CXL™ Introduction



Source: CXL™ Consortium

COLLABORATE. INNOVATE. GROW.



Source: CXL™ Consortium



# CXL™ : Targeting Usage Models

## Type 1 Device

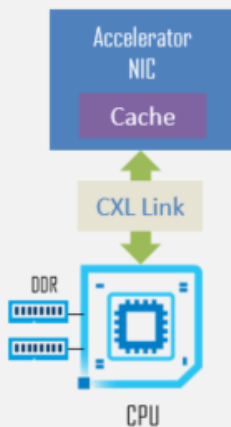
Caching Devices/Accelerators

Usages:

- PGAS NIC
- NIC atomics

Protocols:

- CXL.io
- CXL.cache



## Type 2 Device

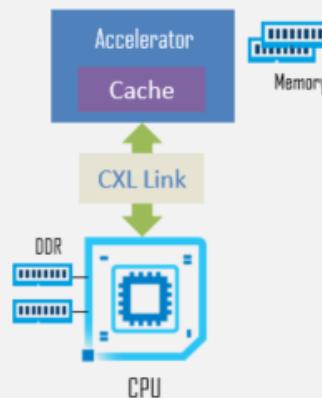
Accelerators with Memory

Usages:

- GPU
- FPGA
- Dense
- Computation

Protocols:

- CXL.io
- CXL.cache
- CXL.memory



## Type 3 Device

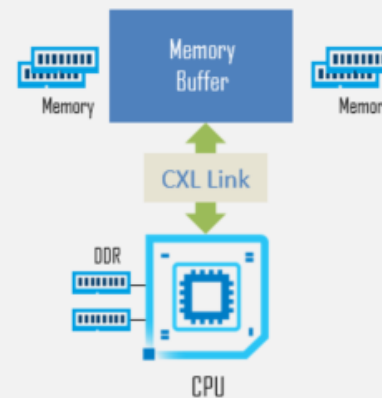
Memory Buffers

Usages:

- Memory BW expansion
- Memory capacity expansion
- ZLM

Protocols:

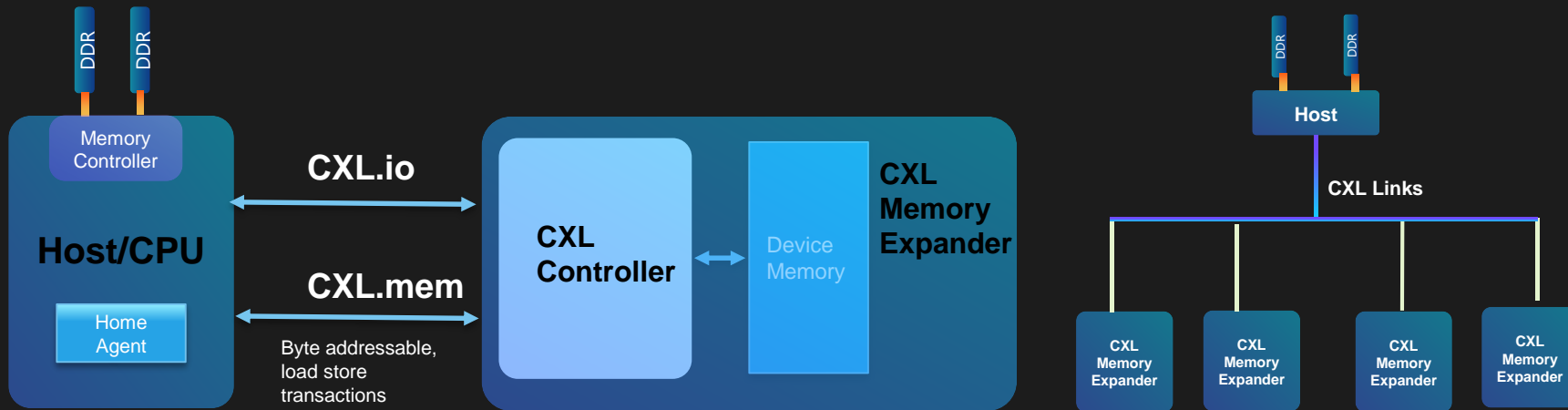
- CXL.io
- CXL.mem







# Typical CXL Type 3 Device



Memory attached to a CXL device is mapped coherently to the system address space



# CXL Type 3 Device - Memory Expansion

Max. 8TB for 1CPU

CPU

Primary DDR channels  
8x 2DPC (DIMM/channels)



Max. 12TB for 1CPU

CPU

8x 2DPC  
(DIMM/channels)

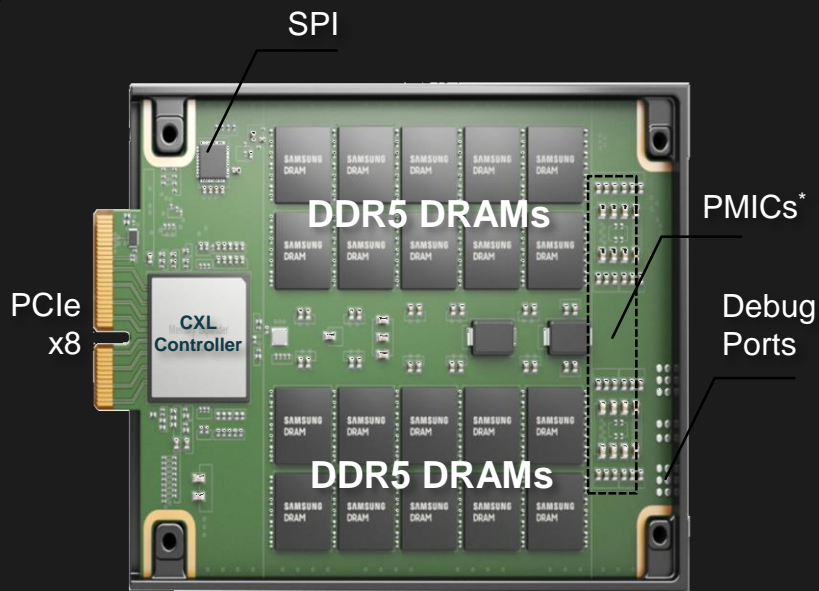


4x CXL links



CXL enables systems to significantly scale memory capacity and bandwidth

# Samsung Memory Expander



E3.S Form Factor

\* Bottom-side

- Form Factor – EDSFF (E3.S-2T)
- CXL 2.0 Support
- CXL Link Width - x8
- Media - DDR5
- Module Capacity – 128GB, 512 GB
- Maximum CXL Bandwidth – 32GB/s (unidirectional)
- Viral and Data Poisoning
- Memory Error Injection
- Multi-symbol ECC, Media scrubbing
- Availability – Currently available for evaluation/testing



# CXL Memory Device Types



Flash Memory Summit

## Memory Expander

CXL Type 3 device

CXL device with high bandwidth and low latency without a long tail



## Tiered Memory Solution

CXL Type 3 device

CXL device with .mem and .io as active data path



## Accelerator Attached Solution

CXL Type 2/3 device

Accelerator with CXL interface

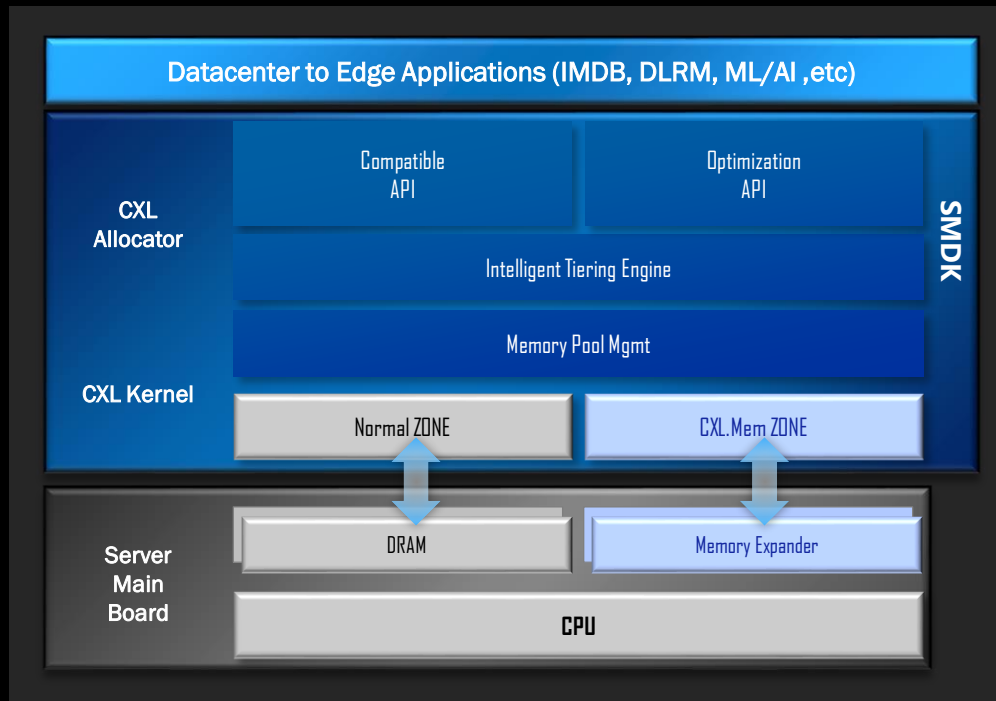


COLLABORATE. INNOVATE. GROW.

SAMSUNG



# SMDK- Scalable Memory Development Kit



- To enable main memory and CXL memory to work together.
- Memory Zone – Distinguishes between native attached DDR and CXL memory.
- Memory Pool Management – Two pools of memory appear as one to applications.
- Intelligent Tiering – Tier based on latency, capacity
- Compatible APIs and Optimization APIs.
- Revision 1.4 of SMDK available on GitHub



# CXL™ Specification Overview

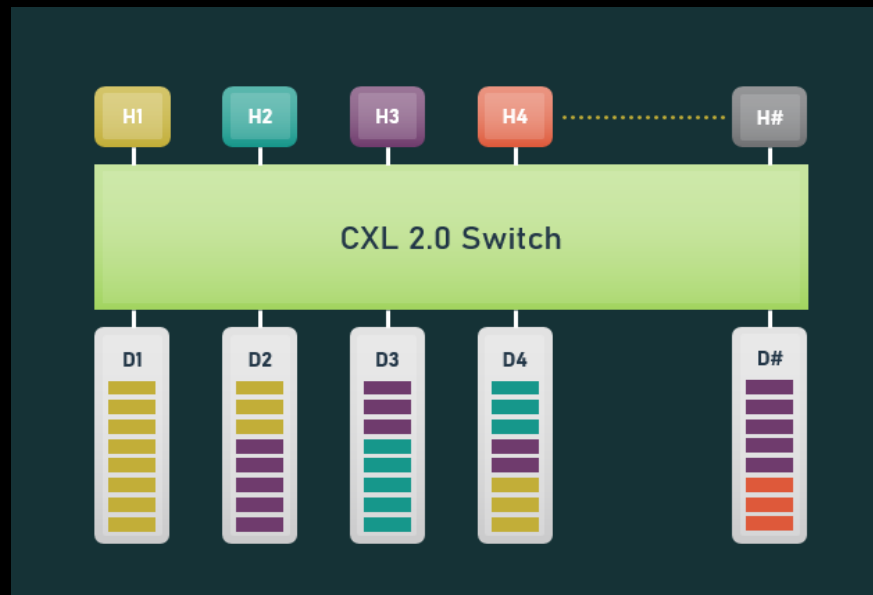
CXL 3.0 builds on CXL 2.0 features to further enhance composable disaggregated infrastructure support

Features	CXL 1.0 / 1.1	CXL 2.0	CXL 3.0	
Release date	2019	2020	2022	
Max link rate	32GTs (PCIe 5.0)	32GTs (PCIe 5.0)	64GTs (PCIe 6.0)	
Flit 68 byte (up to 32 GTs)	✓	✓	✓	
Flit 256 byte (up to 64 GTs)			✓	1
Type 1, Type 2 and Type 3 Devices	✓	✓	✓	CXL 3.0 Additions
Memory Pooling w/ MLDs	CXL 2.0 Additions	✓	✓	
Global Persistent Flush		✓	✓	
CXL IDE		✓	✓	
Switching (Single-level)		✓	✓	
Switching (Multi-level)			✓	2
Multiple Type 1/Type 2 devices per root port			✓	3
Memory sharing (256 byte flit)			✓	4
Symmetric coherency (256 byte flit)			✓	5
Direct memory access for peer-to-peer			✓	6
Fabric capabilities (256 byte flit)			✓	7

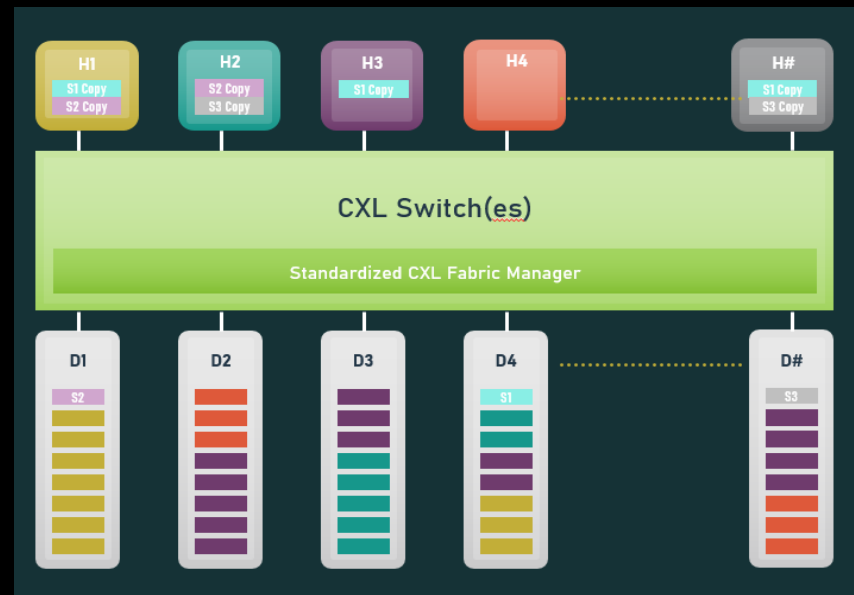




# CXL 2.0 vs 3.0 – Memory Pooling/Sharing



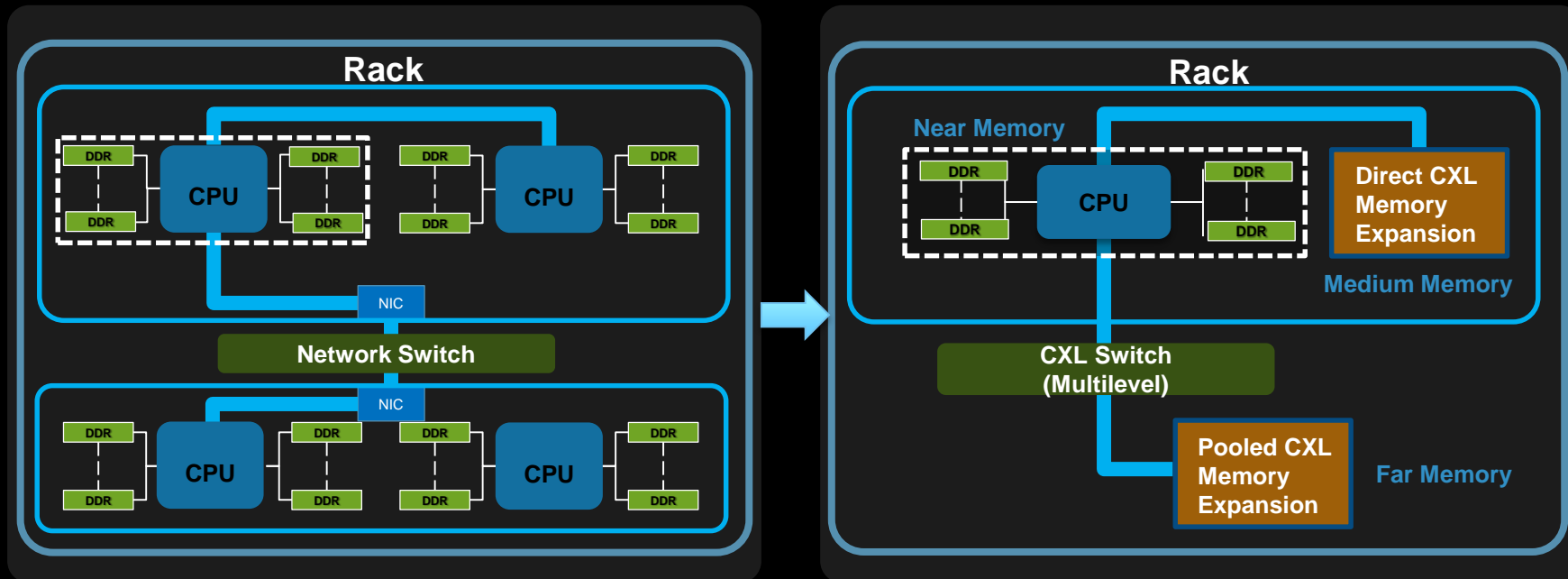
CXL 2.0 - Memory Pooling  
Multi Logical Devices (MLDs) - Finer Grain Memory allocation



CXL 3.0 – Memory Pooling and Sharing  
Memory sharing to improve memory utilization



# CXL Memory Architecture

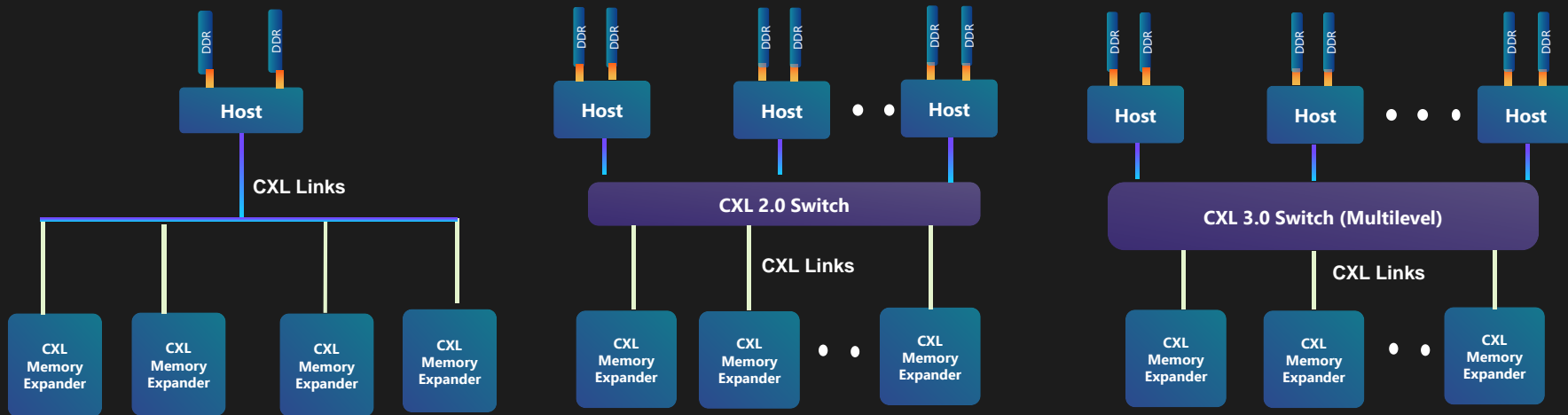


CXL Memory Usage: Capacity Expansion, BW Expansion, Tiered Memory, Memory Pooling/Sharing

CXL Memory Applications: In-Memory Databases, HPC, AI Training/Inference, General Purpose Virtualization etc.



# CXL Memory Expansion, Switching, Pooling & Sharing



CXL 1.1 – Direct Memory Expansion

CXL 2.0 – Memory Pooling

CXL 3.0 – Memory Pooling and Sharing



# Memory-Semantic SSD™ for the Memory-Centric Computing Era

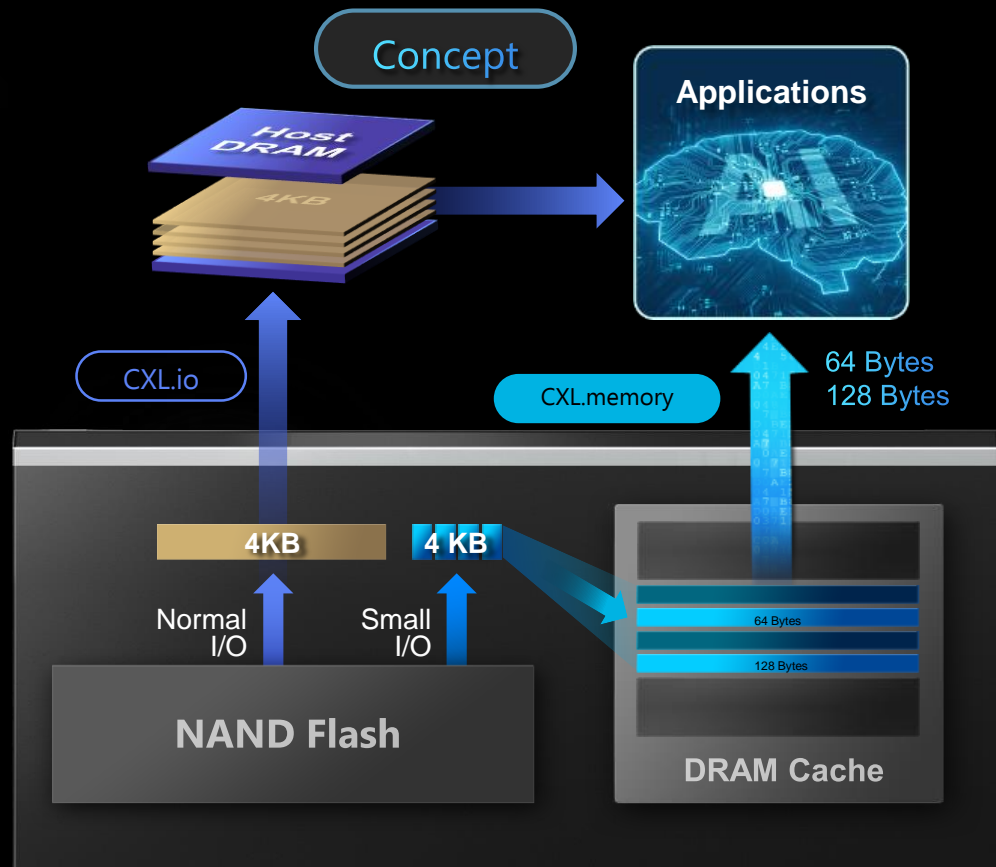
## CXL-based SSD with built-in DRAM

### Built-in DRAM

- Processing AI and ML applications, usually need relatively small-sized data chunks
- Applications can write data to the DRAM cache at DRAM speed

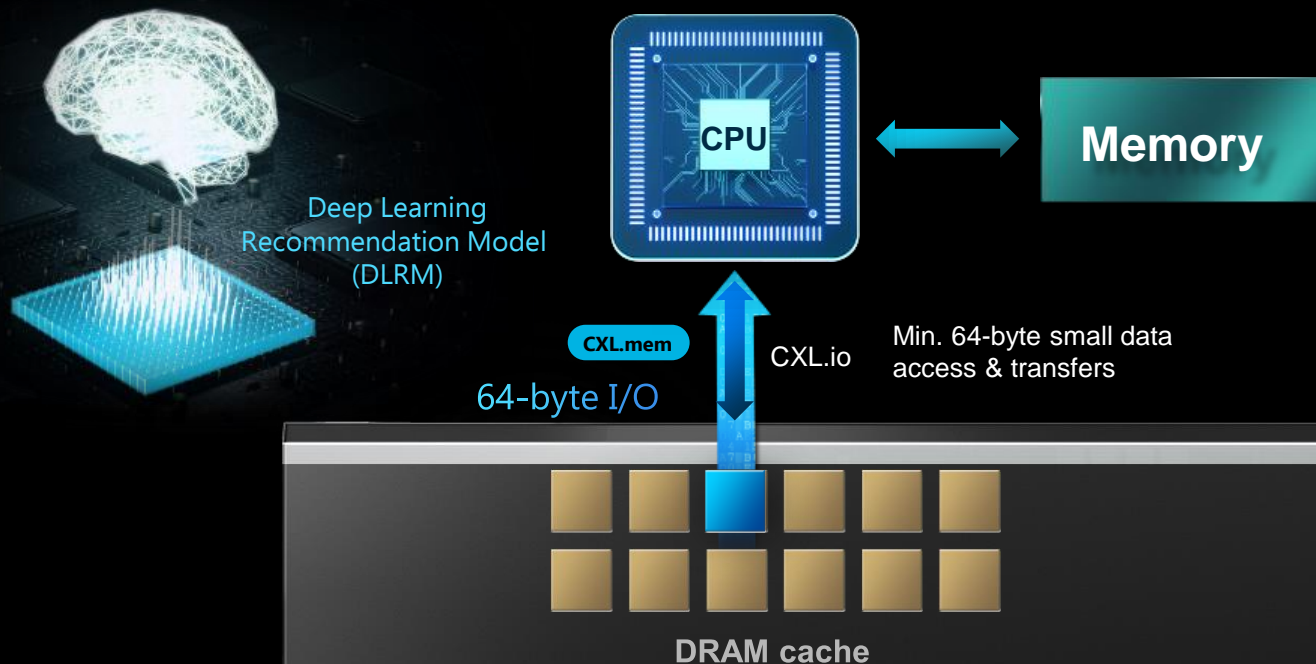
### CXL Technology

- Low latency enabled by CXL.memory protocol

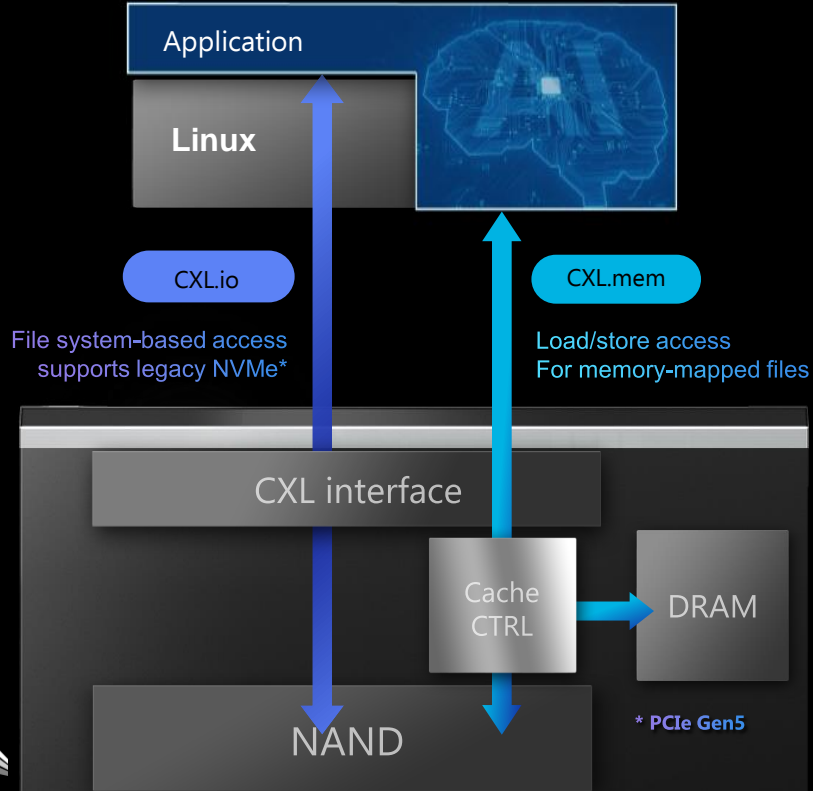




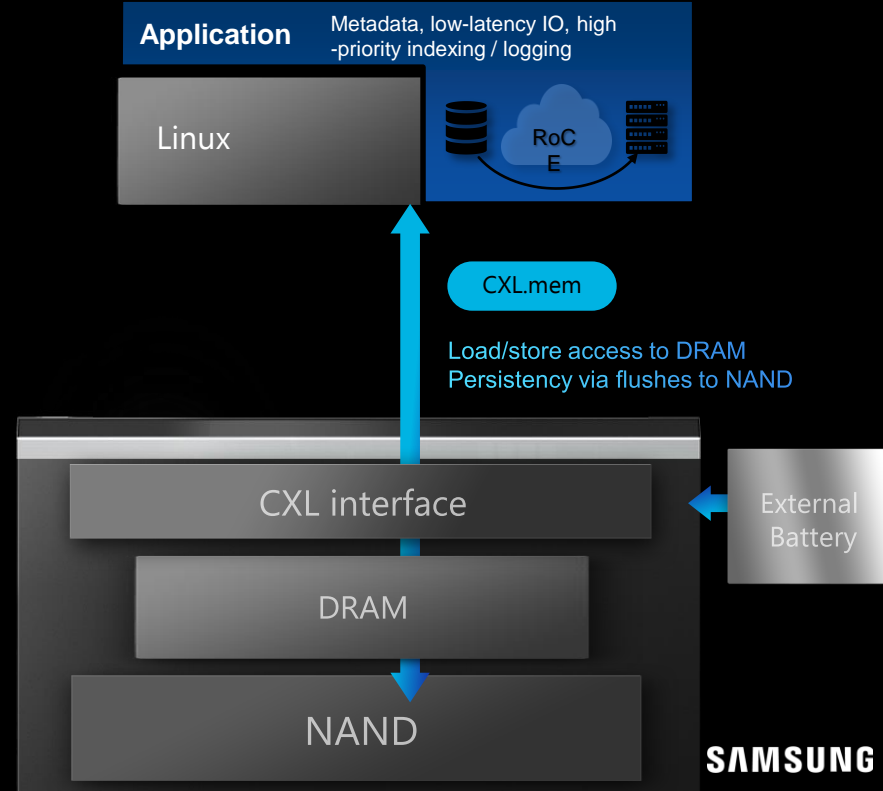
# Small Granularity Access



# Dual Mode Support

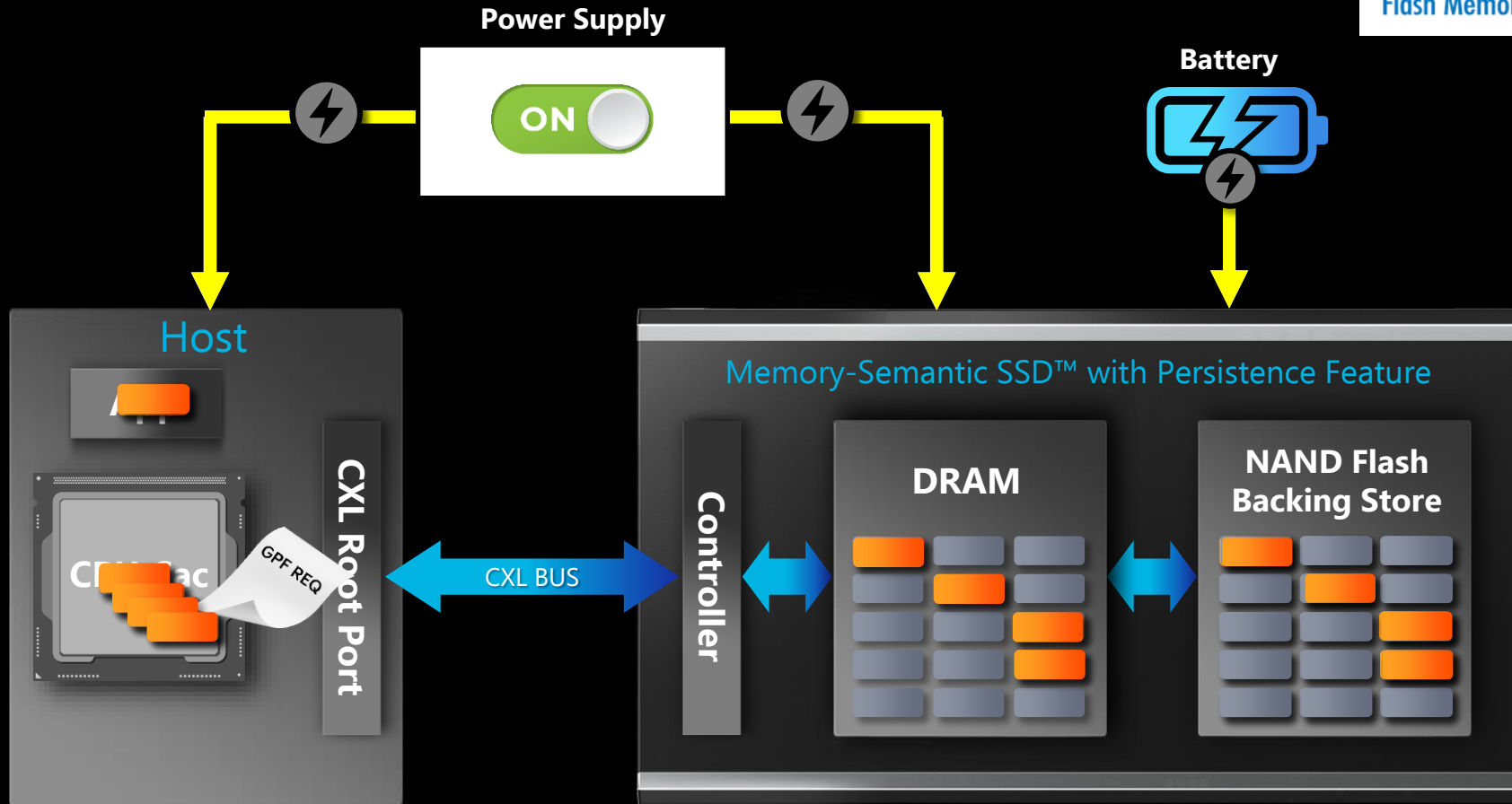


# Persistent Memory Mode



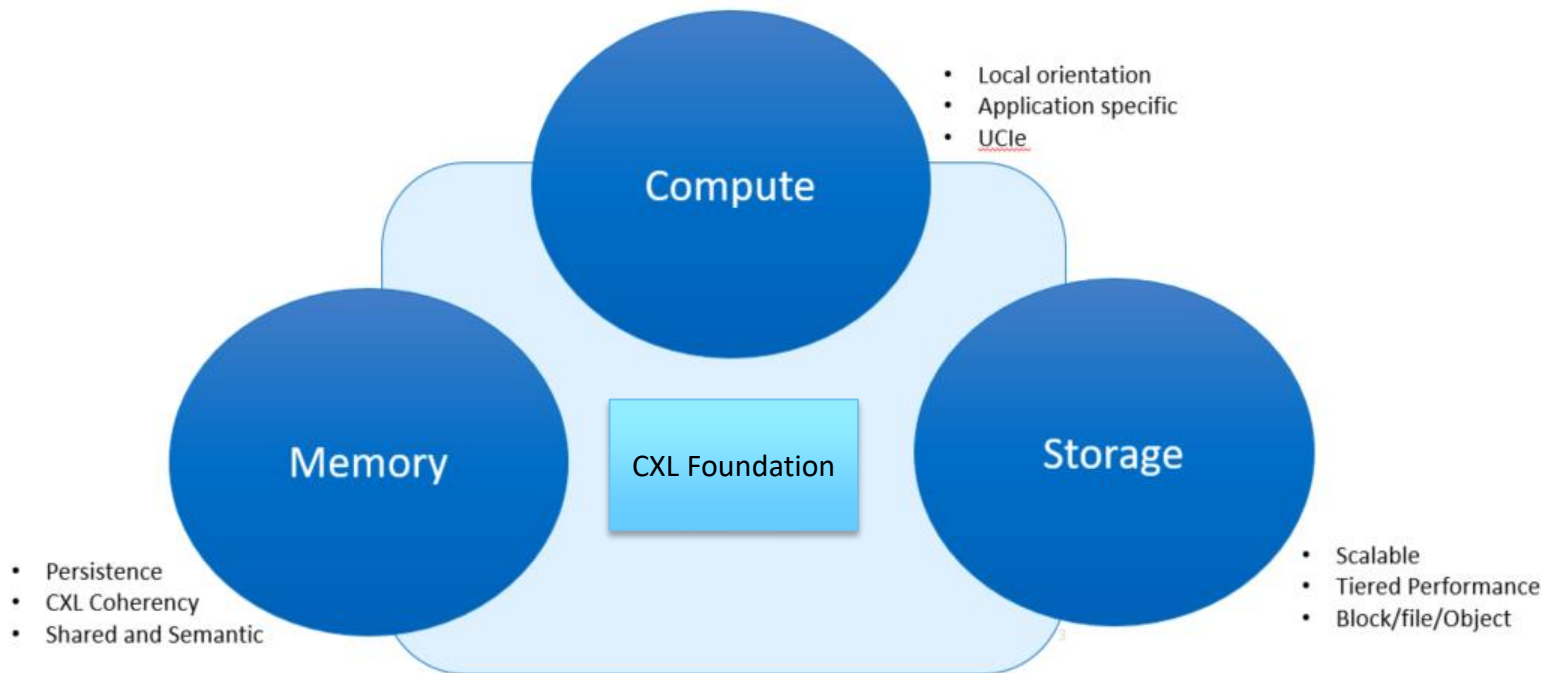


# Persistent Memory Mode Illustration





# Balancing Application-Driven Resources



# Summary

- CXL is the enabling foundation for:
  - Application-oriented memory topologies
  - Data-centric Computing
  - Heterogeneous Compute
- Challenges to exploit CXL-based architectures
  - Architectures that address CXL latencies by coupling to the application layer
  - Open source accelerator programming frameworks
  - Data-centric and heterogeneous computing adoption
  - Workload validation and support

***Support the End Market: Become One With Our Application Developers***



Flash Memory Summit

# Thank You