SNIA COMPUTE + MEMORY + STORAGE SUMMIT

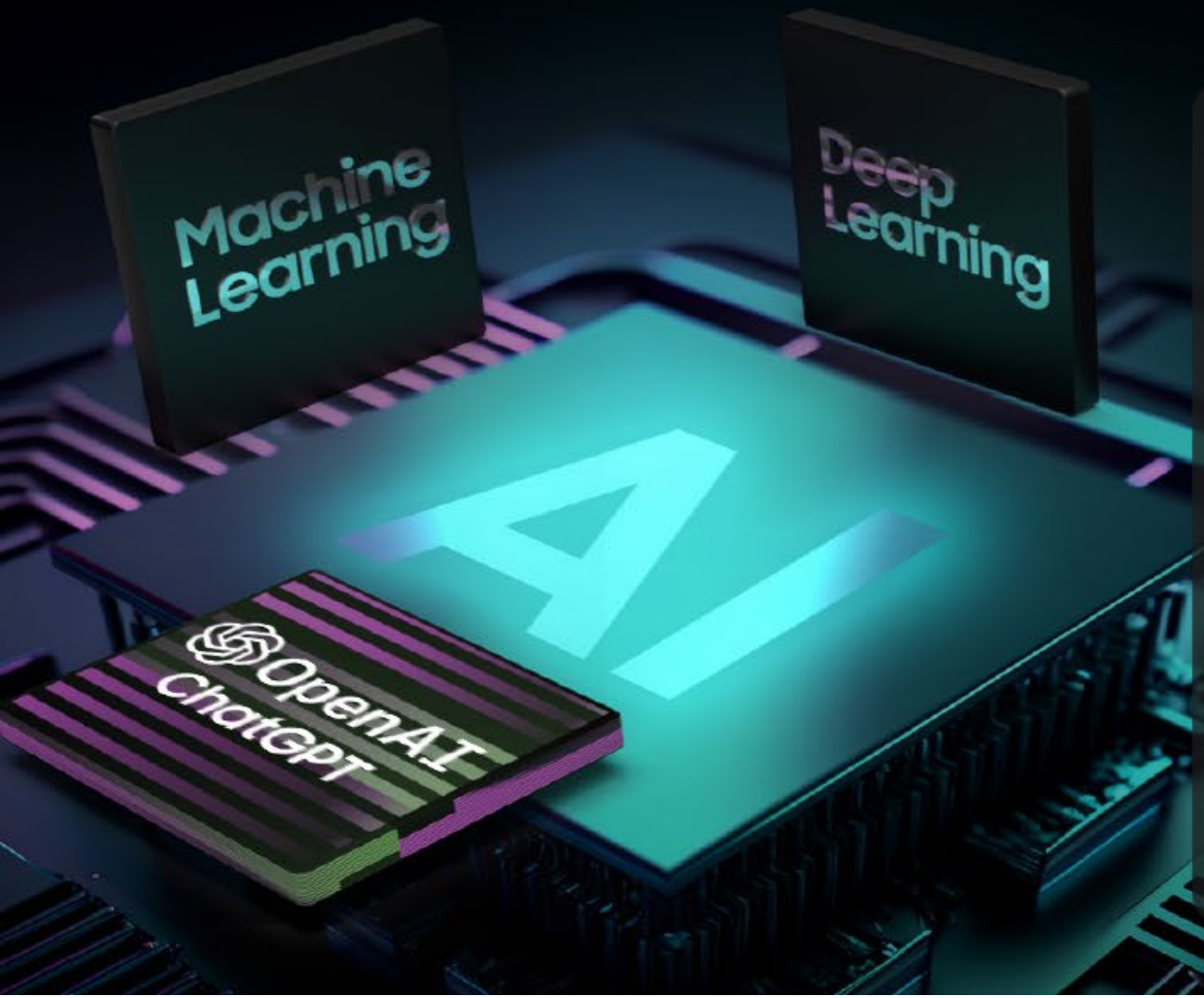
Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023

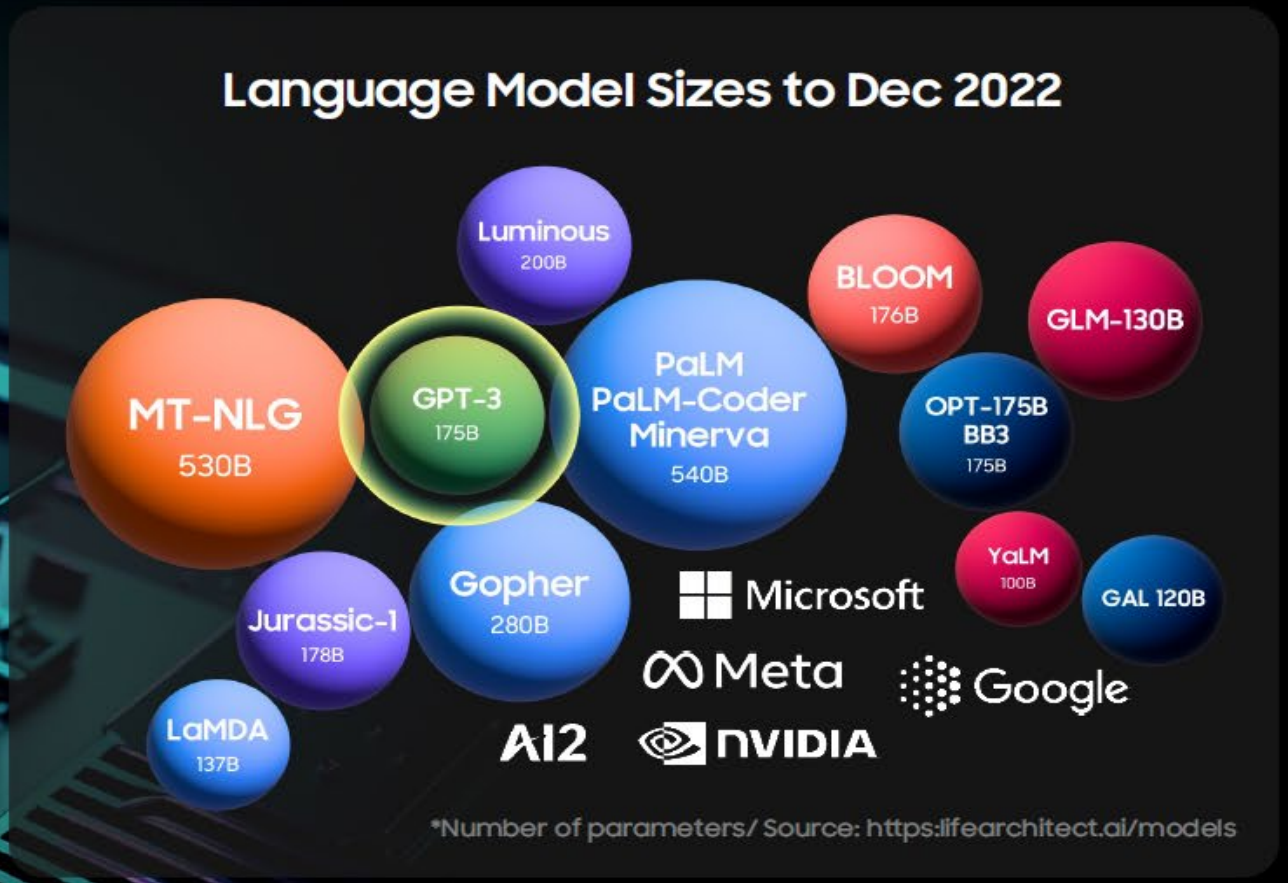
Compute, Memory and Storage: Optimized Configurations for a New Era of Workloads

Presented by David McIntyre
Director, Product Planning
Samsung

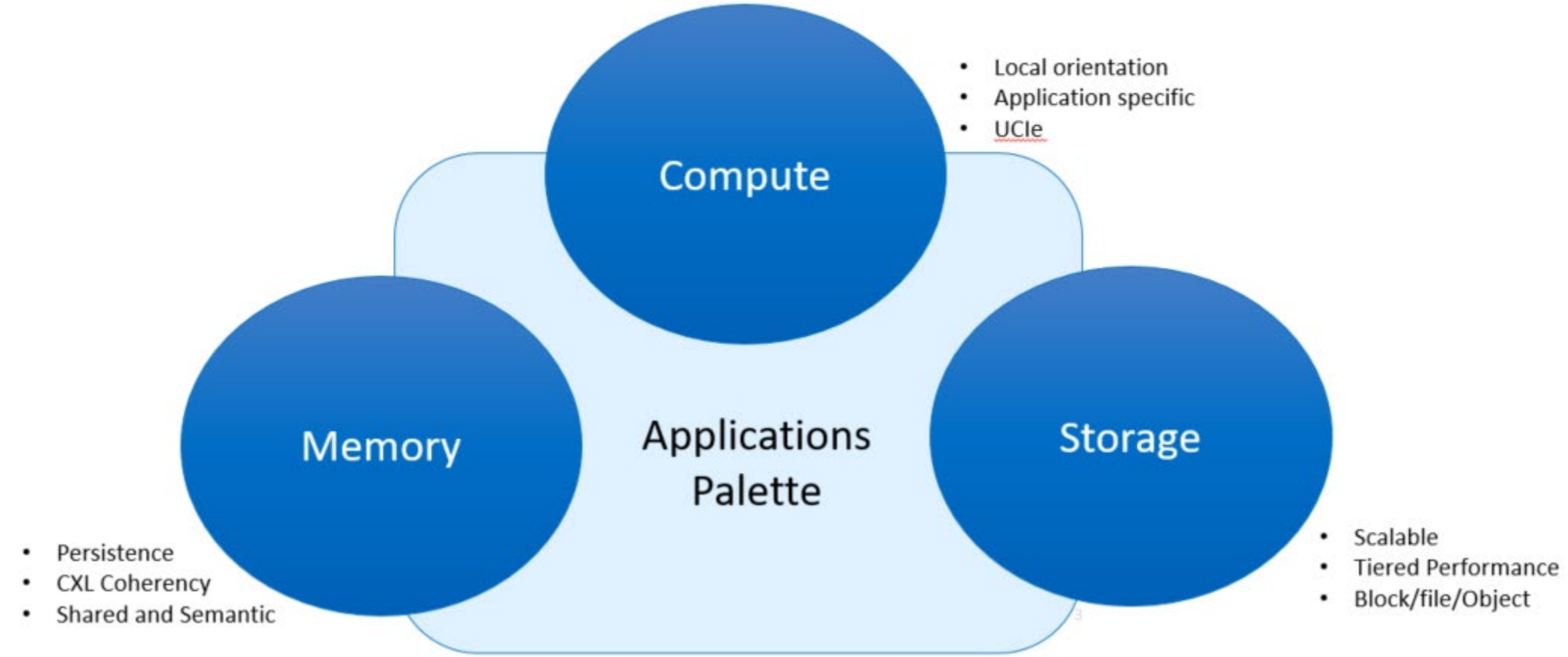
In the Era of AI & ML

Swift increase in demand for memory capacity and performance



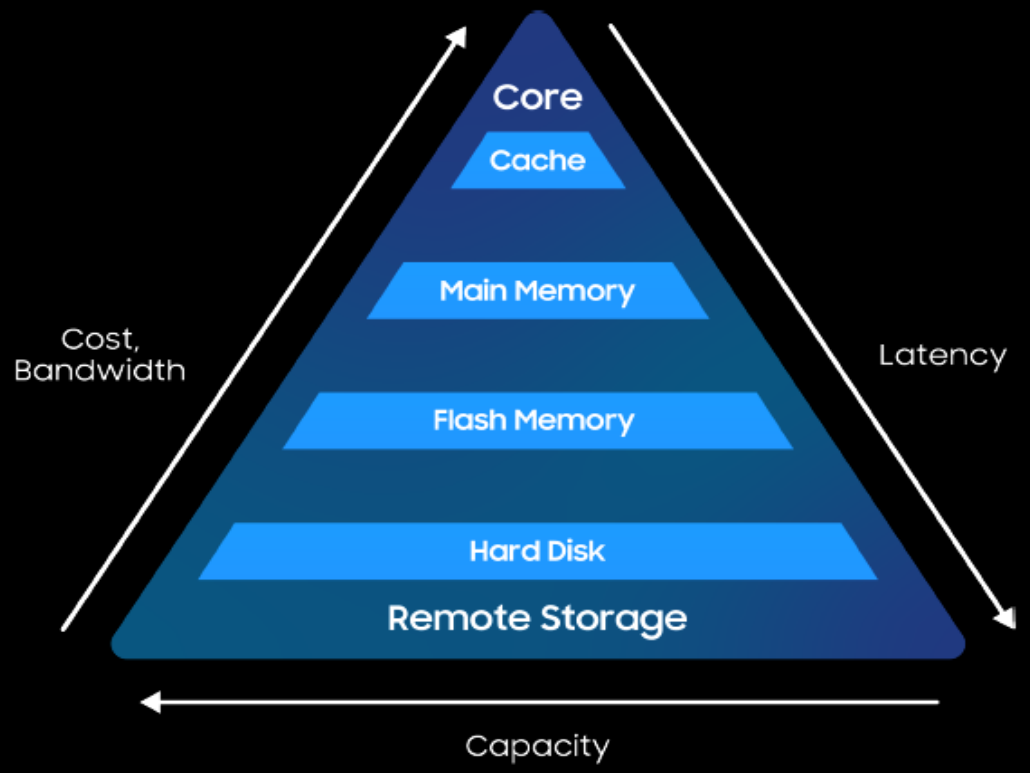


(1) Balancing Application-Driven Resources

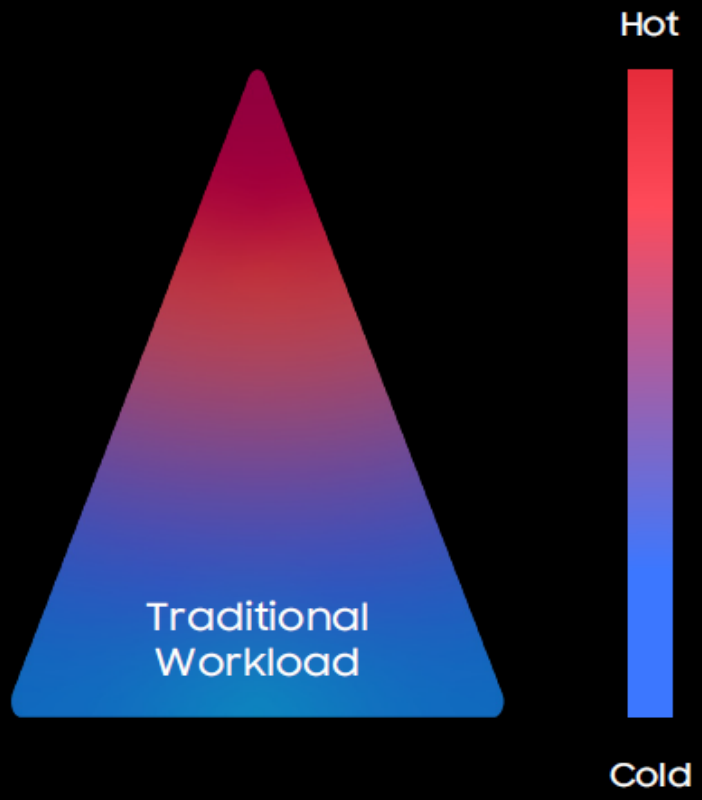


Memory Hierarchy

Keep hot data close to CPU using data locality



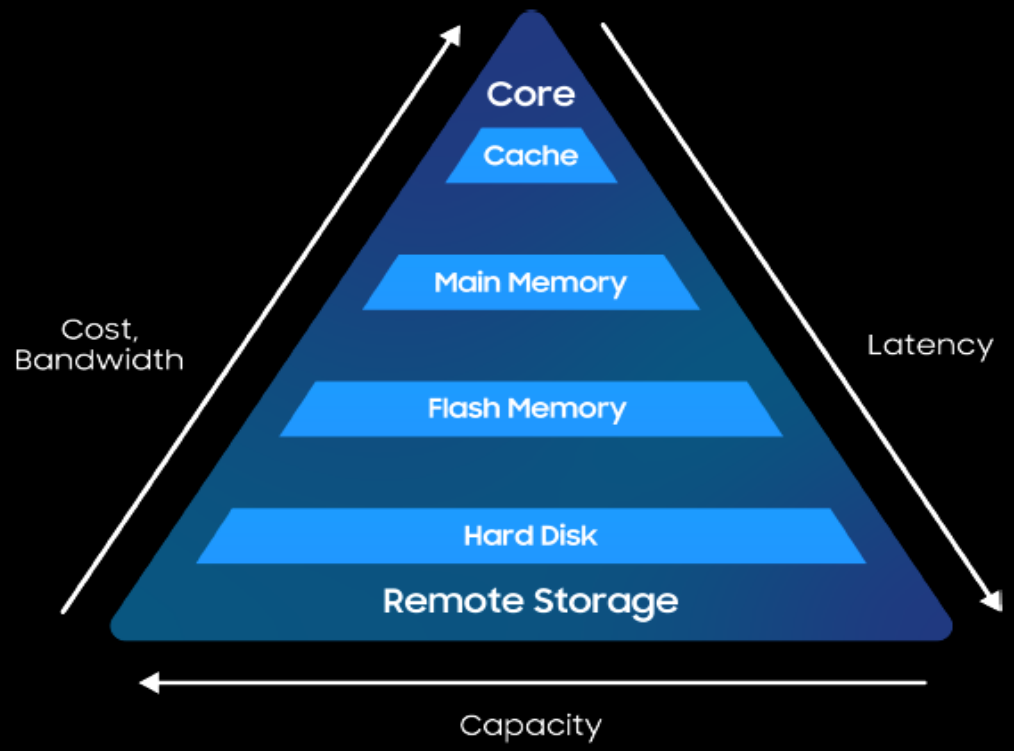
Memory Hierarchy



Traditional Workload

Memory Hierarchy Disparity for Modern Workloads

Not all workloads exhibit the conventional pattern of data locality

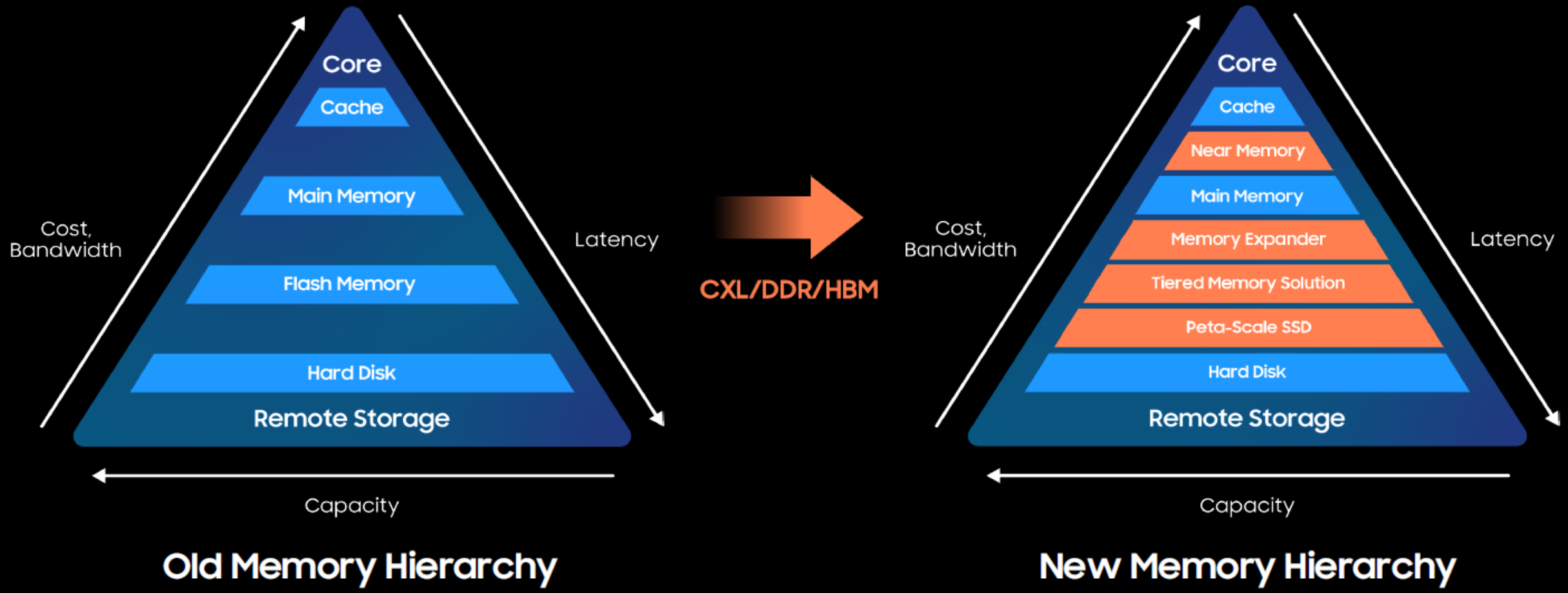


Memory Hierarchy

Modern Workload

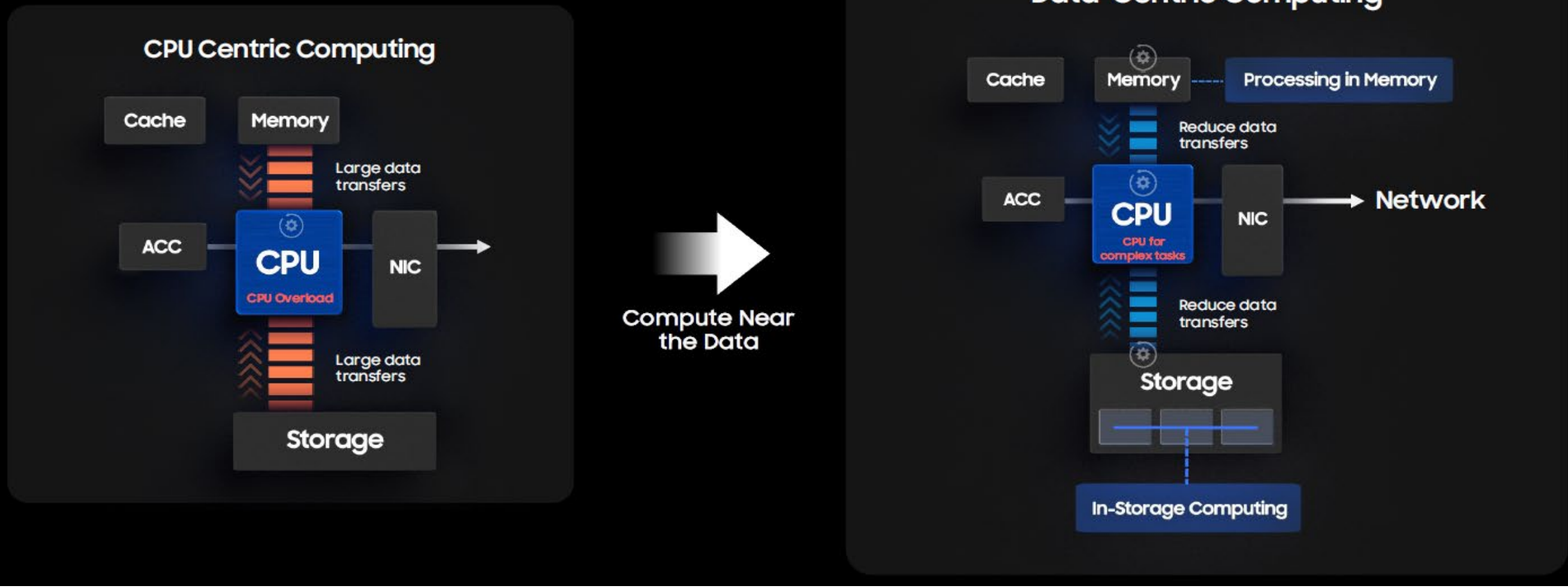
New Memory Hierarchy

Deeper and more efficient memory hierarchy to fill the performance gap



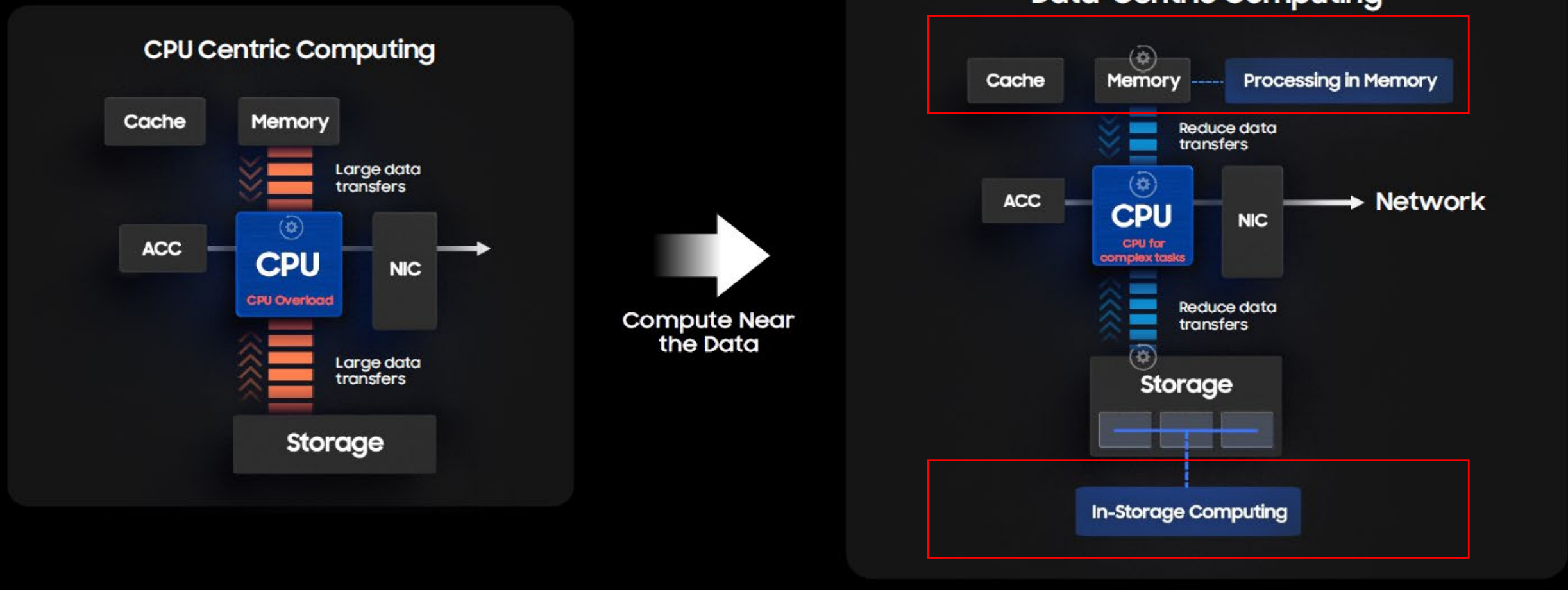
Data-Centric Computing Concept

Move the computation to the data for large datasets



Data-Centric Computing Concept

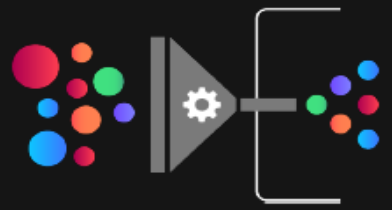
Move the computation to the data for large datasets



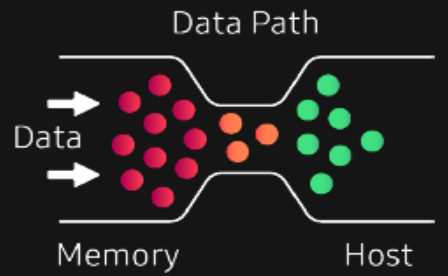
Data-Centric Computing Benefits

Power-optimized scalable processing for large data

Low Power Computing



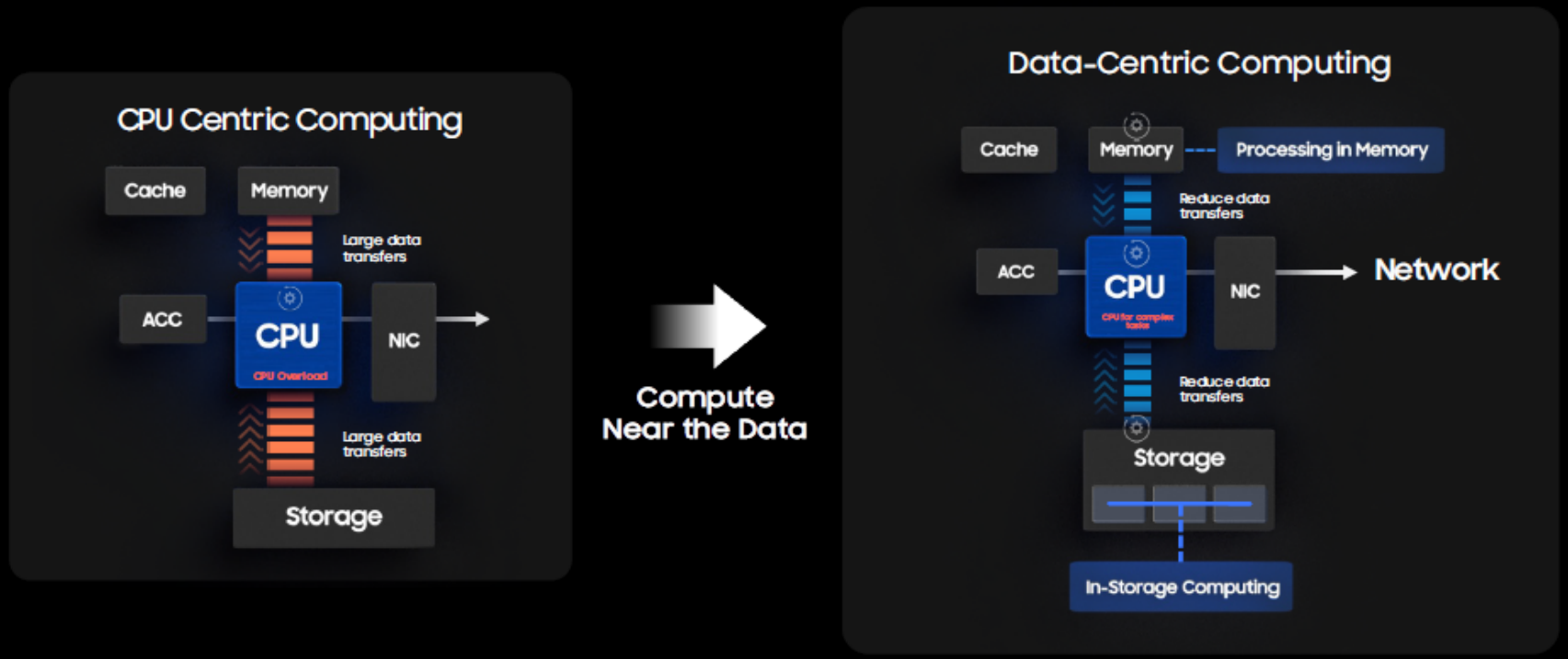
Data Reduction



High Effective Bandwidth

Scalable Computing

Challenges in Data-Centric Computing



Interface Interference

Killer Application

Ease of Use

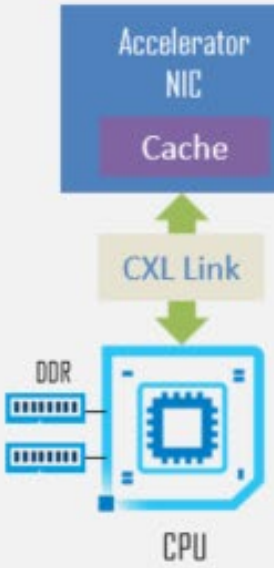
CXL™ 1.0/CXL 1.1 Usage Models

Type 1 Device

Caching Devices/Accelerators

Usages:
- PGAS NIC
- NIC atomics

Protocols:
- CXL.io
- CXL.cache

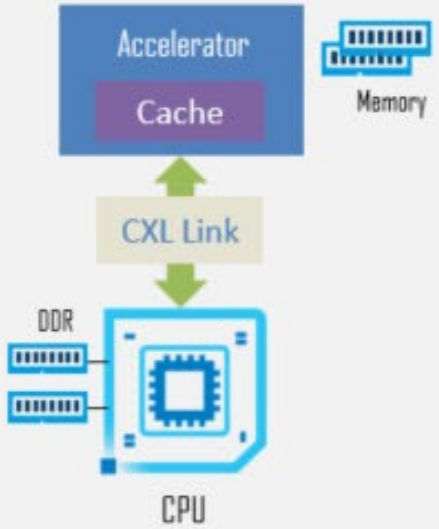


Type 2 Device

Accelerators with Memory

Usages:
- GPU
- FPGA
- Dense
- Computation

Protocols:
- CXL.io
- CXL.cache
- CXL.memory

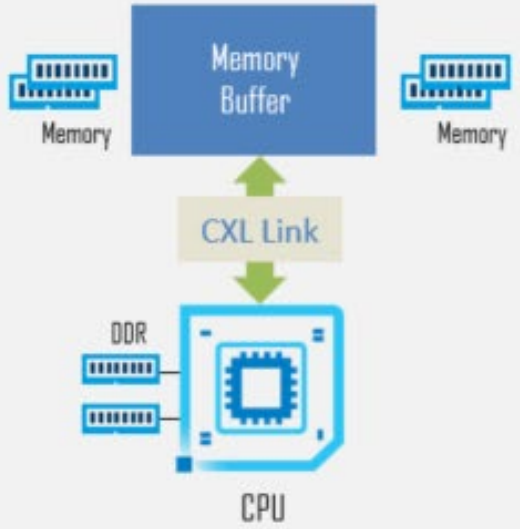


Type 3 Device

Memory Buffers

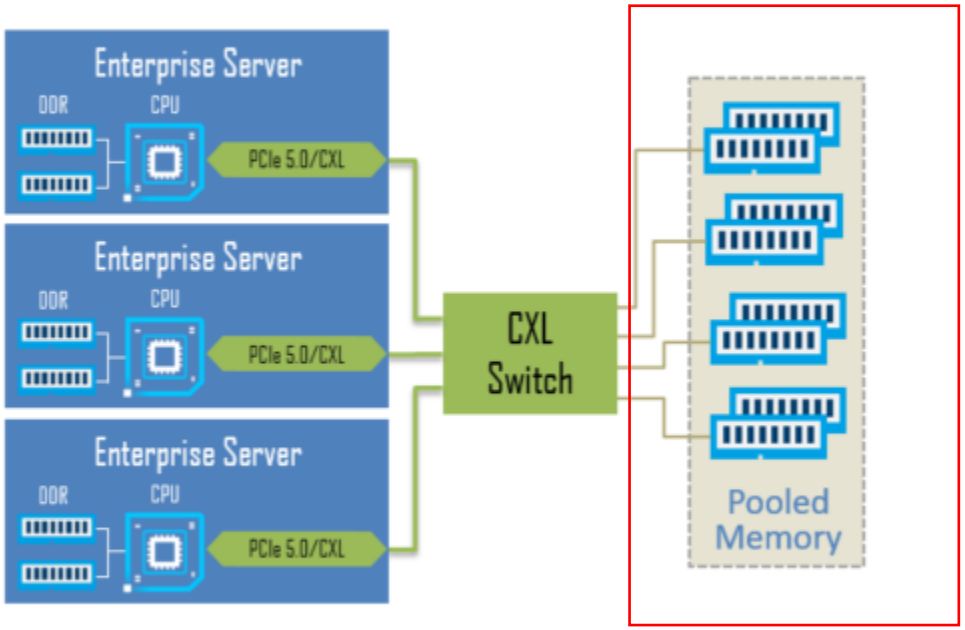
Usages:
- Memory BW expansion
- Memory capacity expansion
- 2LM

Protocols:
- CXL.io
- CXL.mem



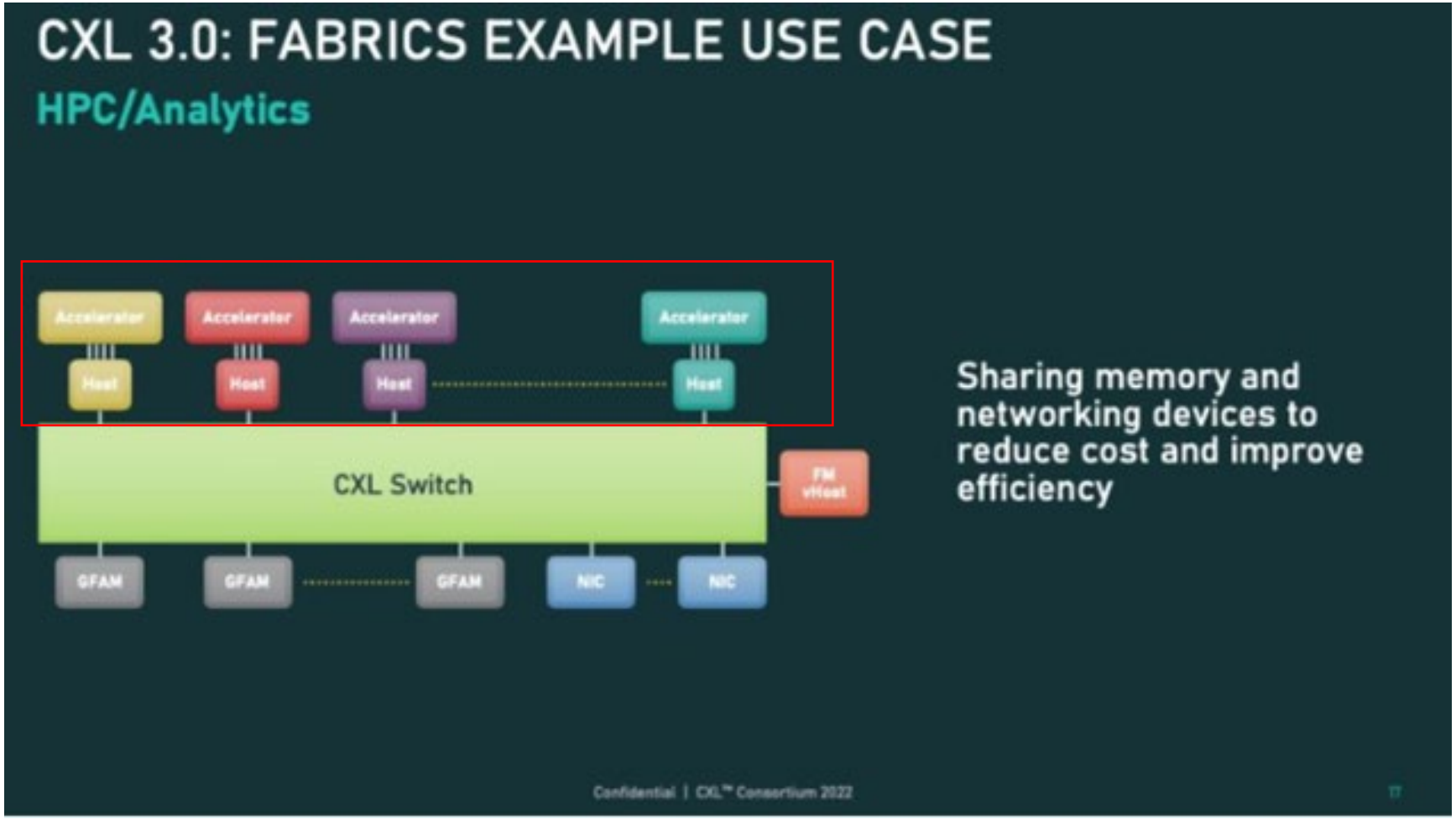
CXL™ 2.0: Resource Pooling at Rack Level, Persistent Memory

- Resource pooling/disaggregation
  - Managed hot-plug flows to move resources
  - Type-1/Type-2 device assigned to one host
  - Type-3 device (memory) pooling at rack level
  - Direct load-store, low-latency access – similar to memory attached in a neighboring CPU socket (vs. RDMA over network)
- Persistence flows for persistent memory
- Fabric Manager/API for managing resources
- Security: authentication, encryption
- Beyond node to rack-level connectivity!



Disaggregated system with CXL optimizes resource utilization delivering lower TCO and power efficiency

CXL 3.0 supports Heterogeneous Compute

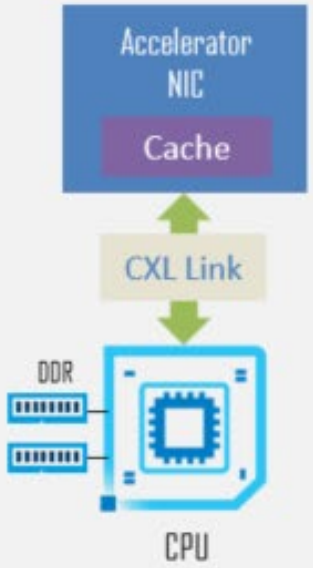


CXL™ : *Targeting* Usage Models

Type 1 Device
Caching Devices/Accelerators

Usages:
- PGAS NIC
- NIC atomics

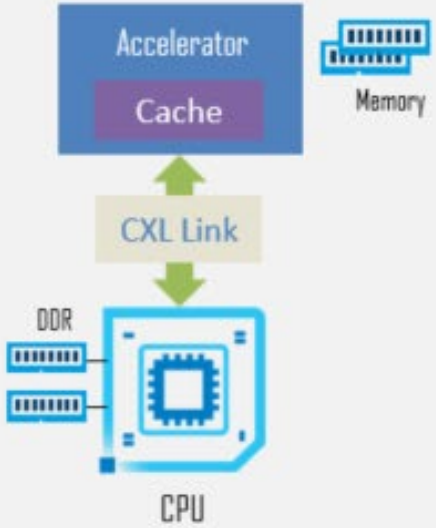
Protocols:
- CXL.io
- CXL.cache



Type 2 Device
Accelerators with Memory

Usages:
- GPU
- FPGA
- Dense
- Computation

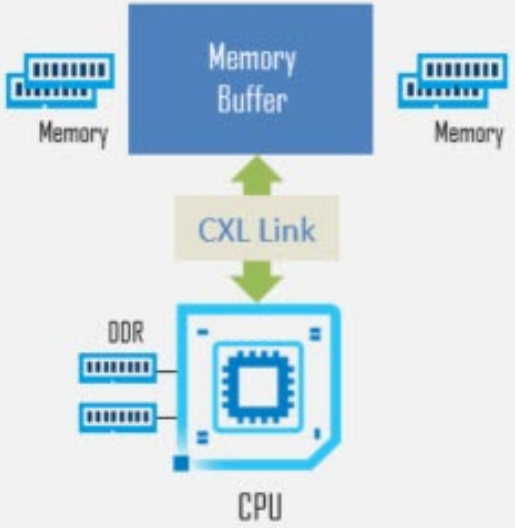
Protocols:
- CXL.io
- CXL.cache
- CXL.memory



Type 3 Device
Memory Buffers

Usages:
- Memory BW expansion
- Memory capacity expansion
- 2LM

Protocols:
- CXL.io
- CXL.mem



CXL Memory Device Types

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

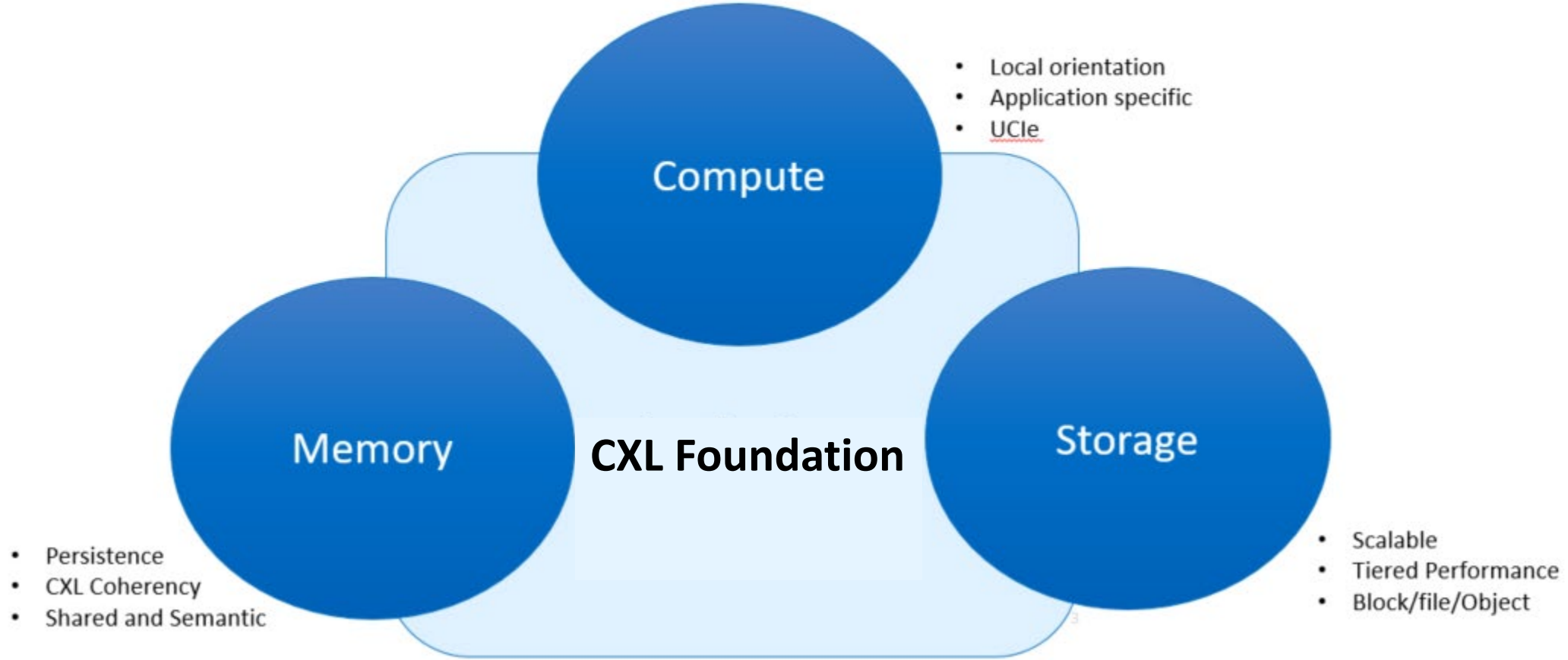
CXL device with .mem and .io as active data path

Accelerator Attached Solution

CXL Type 2/3 device

Accelerator with CXL interface

(2) Blending 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

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023

Please take a moment to rate this session.

Your feedback is important to us.

