AI at the Intersection of Storage and Networking

A Perspective from the Chair of SNIA and UEC

REGIONAL SDC 24
BY Developers FOR Developers
APRIL 24, AUSTIN, TX

Dr. J Metz, AMD
Chair, Ultra Ethernet Consortium
Chair, SNIA

A SNIA Event

Agenda

- The Needs of AI
- Impacts on the Network
- Impacts on Storage
- Intersection of SNIA and UEC
- Conclusion

Special Thanks: Pratik Mishra (AMD), Mark Nowell (Cisco); Jason Molgaard (Solidigm), Shyam Iyer (Dell)

REGIONAL SDC 24

AI for Storage/Networking, or Storage/Networking for AI?

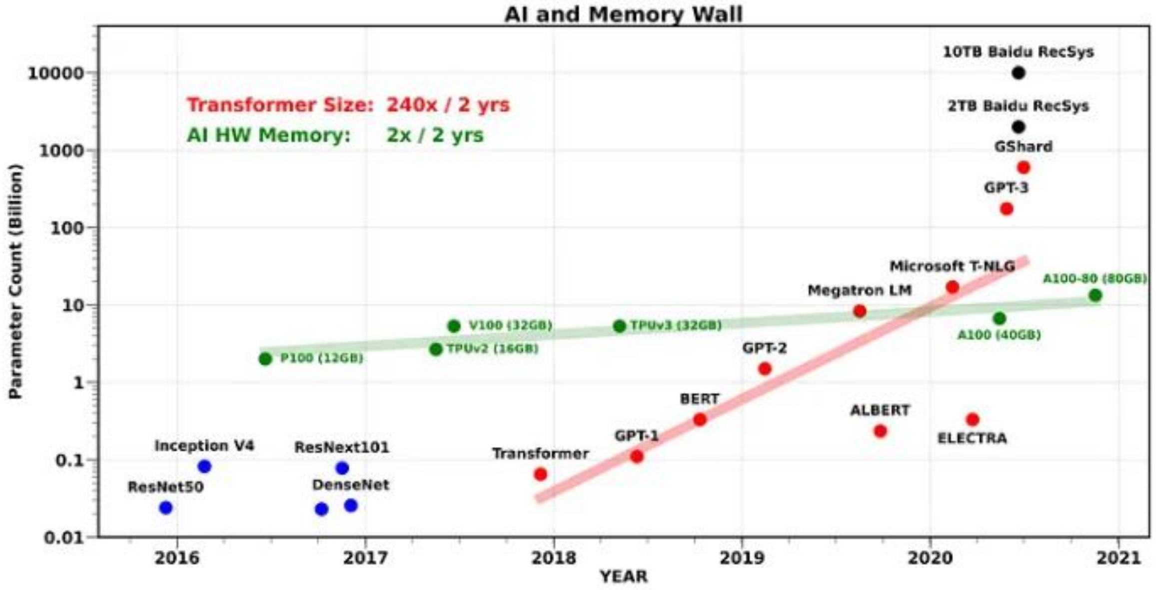
- Lots of talk about how AI will change networking infrastructure
  - … but what network infrastructure do you need to have, for enough AI to change the networking infrastructure?
  - Is it more than just superfast speeds and feeds?
- Updating data/storage infrastructure
  - Massive data sets, parallel processing requirements
  - Compression/Decompression techniques, offloading for migration, replication, and synchronization efficiency
  - Memory buffer data transfers and abstraction trade-offs
  - Where does the data need to be, and when?

The Needs of AI

REGIONAL SDC 24
BY Developers FOR Developers

The AI Monster

- AI workloads need
  - Ever-increasing Memory Bandwidth
  - Ever-increasing Memory Capacity
  - (Near) Instantaneous Data Access (Exabytes)
- Intermittent data surges
- ”Straggler” data (tail latency) significantly impacts completion time
- Extended operation duration (hours, days)



Parameter (Billions) count over the years*

* Gholami, Amir, et al. (2021). AI and Memory Wall. https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

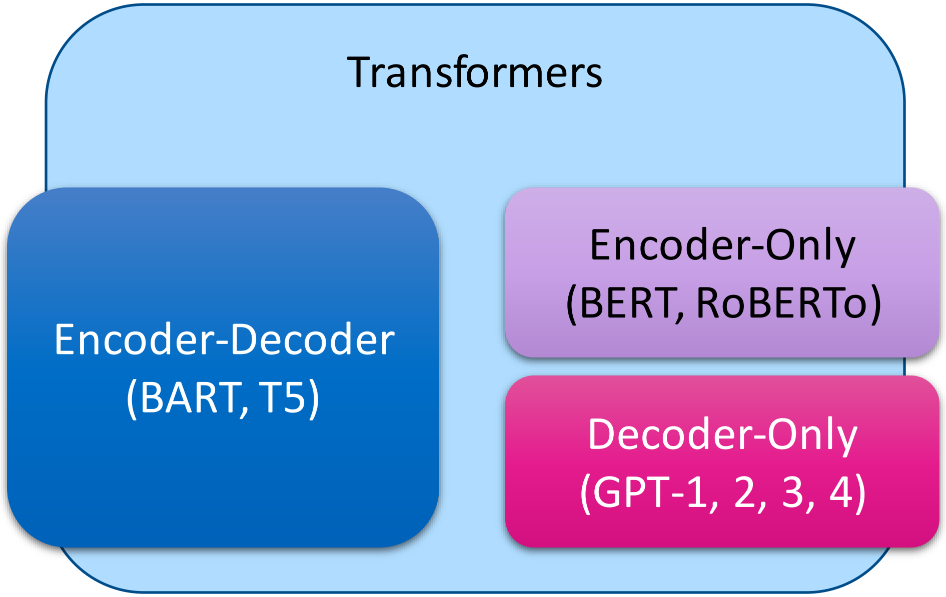
REGIONAL SDC 24

New Architectures

- Transformers
  - Model of text generation applications
  - Two building blocks:
    - Encoders
    - Decoders
- Encoders
  - Parallel processing of all input tokens into learned information
  - A.k.a. Understanding Context
- Decoders
  - Takes input tokens one-by-one to generate output (sequential)
  - A.k.a. Generating tasks (text)

REGIONAL SDC 24

Digging Deeper



- GPT: Generative Pre-Trained Transformers
  - Popular in cloud-services (particularly text-generation)
  - Decoder-only
    - Uses pre-trained matrices
- Two Key Stages:
  - Summarization (SUM)
    - Processes large input context simultaneously (parallel)
    - Computation-bound, higher weights reusability
    - Well-suited to GPUs
  - Generation (GEN)
    - Produces single word at a time (iterates)
    - Memory-bound, lower weight reusability
    - Performs poorly on GPUs
    - Sequential computation: maximum contribution to latency
    - Capacity- and bandwidth-limited

REGIONAL SDC 24

Impact on Networking and Storage/Data

- Compute, Memory, and Bandwidth constraints
  - What's the impact on data movement (Network/Storage)?
  - What happens when you hit 1 Million endpoints?
- Things that break (or, at least, hurt):
  - Congestion signaling, notification, spreading and mitigation (e.g., reaction time)
  - Data ordering and sequencing
  - Timely telemetry
  - Multipath flow-hashing and load-balancing
  - Dataset-specific best practices that require manual tuning
  - Recovery methods
  - Management techniques
  - I/O Amplification
  - Security

AI Problems To Solve

- Memory Bandwidth vs. Capacity vs. Latency
- Computation-Bound Workloads
  - E.g., Summarization: processes large input context simultaneously (parallel)
- Memory-Bound Workloads
  - E.g., Generation: produces single word at a time (iteration)
- Recommendation Workloads spend almost 60% of time in Network I/O*
- I/O Tax: 70% of AI model training is spent on data movement
- I/O Blender: Multiple AI phases occurring at the same time
- Impact of checkpointing, (de-) Compression, Encryption, Replication, etc.

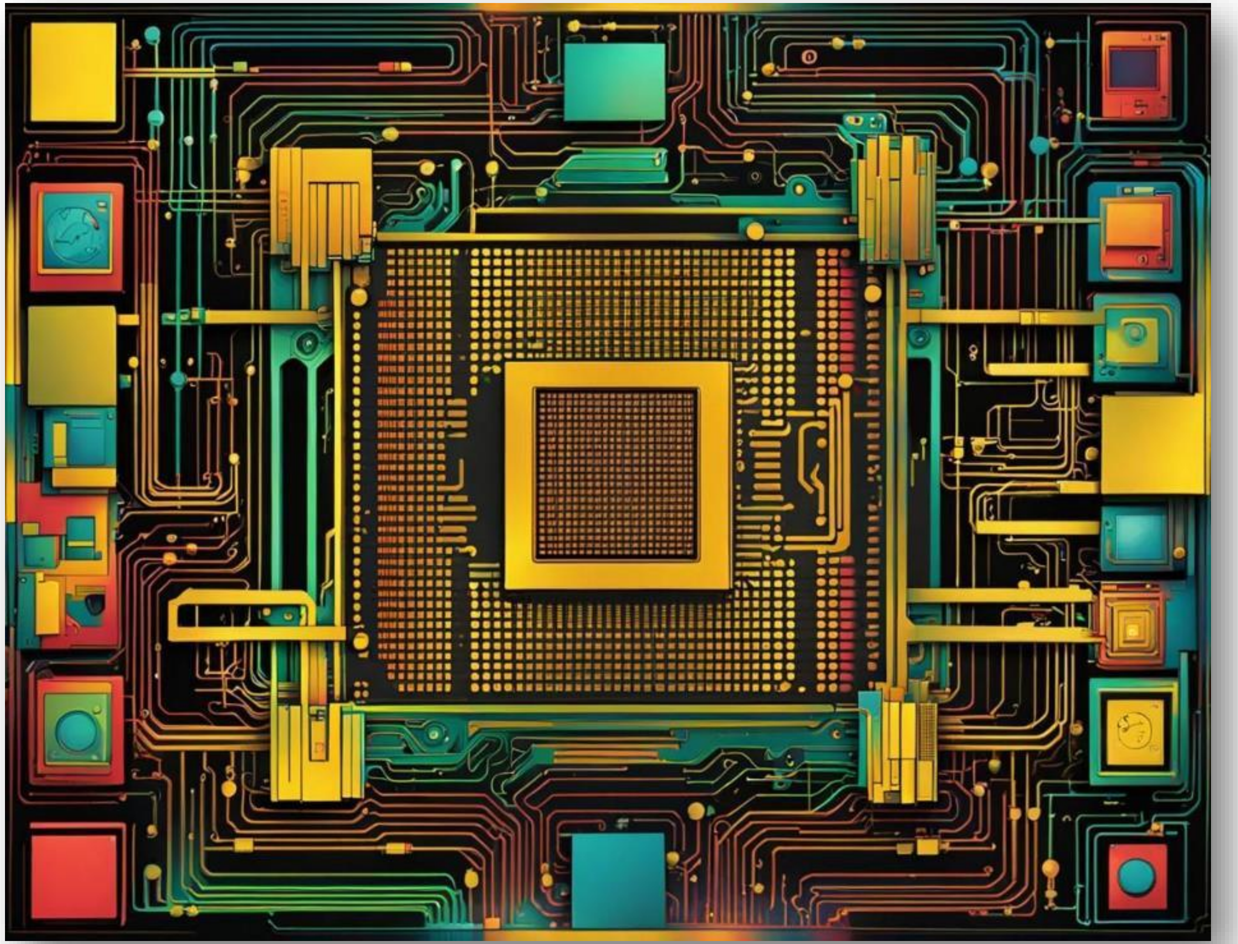
*Meta, OCP 2022 Global Summit

REGIONAL SDC 24

Impacts on the Network

REGIONAL SDC 24
BY Developers FOR Developers

Remote Access to Memory



- Issues
  - Verbs API limits efficiency by preventing OOO packet data from being delivered straight through the network to the application buffer (final destination)
  - Go-Back-*N* recovery methods retransmit *N* packets for any single packet loss
- Impact
  - Ties up network bandwidth for recovery
  - Causes under-utilization of available links
  - Increases tail latencies
- Ideal Solution
  - All links are used; order is only enforced when the AI workload requires it

REGIONAL SDC 24

Bandwidth and Latency

- Training is highly *latency*-bound, where tail latency negatively impacts the frequent computation and communications phases
  - Generation stage is maximum contribution to latency; 60-80% of total
  - Latency increases with # of output tokens
- Large models (e.g., from 175B parameters in GPT-3 to 1T in GPT-4) drive larger messages on the network
- Underperforming networks therefore underutilize expensive resources

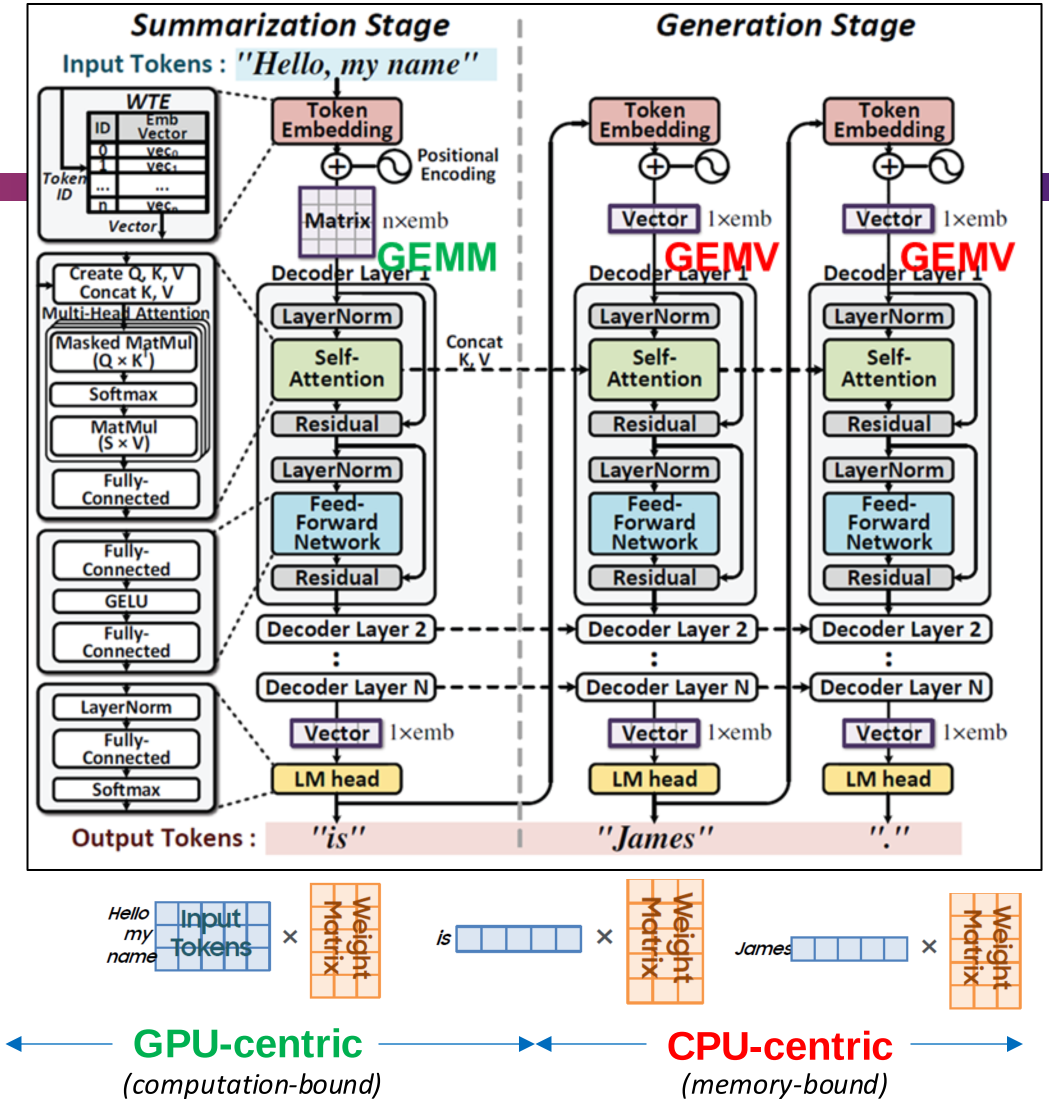


Image credit: Hong, Seongmin, et al. "DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation." *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2022.

REGIONAL SDC 24

Collective Communications

- More than just point-to-point connectivity; inter-accelerator communication in AI is part of “collective” communication operations
- Proper network architecture enables benefits of packet-spraying in bandwidth-intensive operations by eliminating the need to reorder packets before delivery

All-Reduce:

- Imagine you have a group of friends, and each friend has a number written on a piece of paper.
- You want to find the total sum of all those numbers. Here’s how All-Reduce works:
  - Each friend shares their number with everyone else.
  - Everyone adds up all the numbers they receive.
  - The final result is the sum of all the original numbers, and everyone gets that same total.
- In parallel computing, All-Reduce is used to combine data from different processors or nodes to compute a global result (like the total sum in our example).

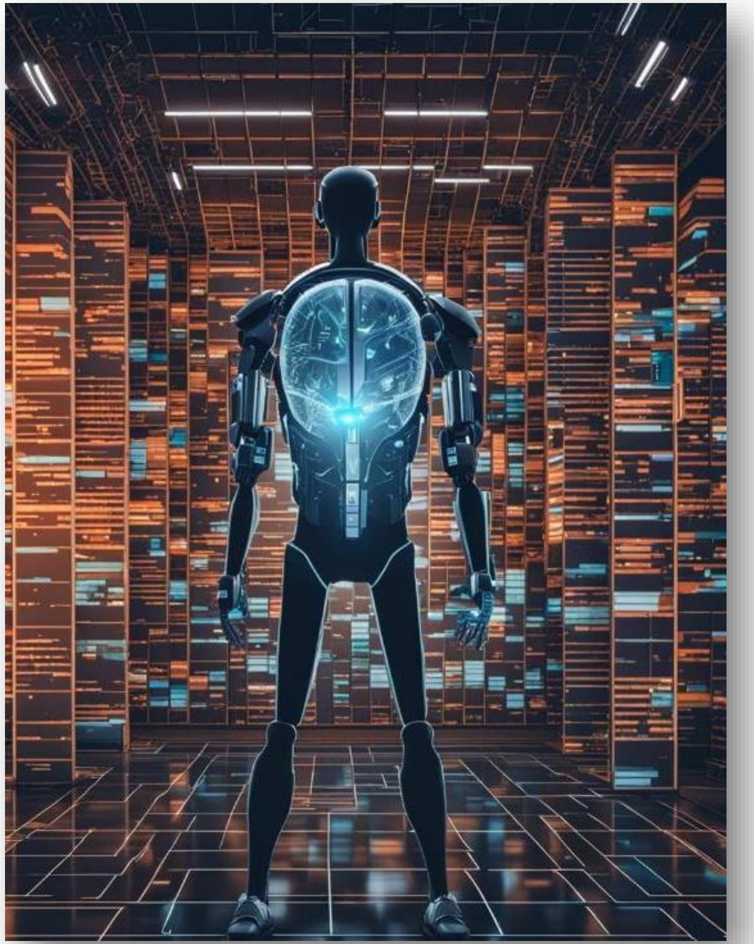
All-to-All:

- Imagine you’re hosting a potluck dinner, and each guest brings a different dish.
- You want everyone to taste every dish. Here’s how All-to-All works:
  - Each guest shares their dish with every other guest.
  - Everyone gets a taste of every dish.
- In parallel computing, All-to-All is used to exchange data between all processors or nodes. Each processor communicates with every other processor, ensuring that everyone has the necessary information.

Impacts on Storage

Storage AI Needs

- **Scalability and Performance**:
  - Scale-Out Architectures: Direct accelerator access to storage via networking Fabric (e.g., NVMe-oF)
  - High I/O Rates and Low Latency
  - Power Restrictions
- **Data Diversity and Edge Computing**:
  - Data Sources (such as DPU Computing; support for offloads, programmability, control + data path optimization)
  - Edge-to-Core Processing
- **Cloud Integration**:
  - Hybrid Cloud
  - Flexibility
- **Multi-modal GenAI jobs – images, text, video**
  - True for both AI Training and Inference
  - Data + Metadata cannot fit in GPU (memory-hierarchy)
  - Accelerator (e.g., GPU) and remote network data paths is a bottleneck
- **AI-Specific Features**:
  - AI-Aware Algorithms
  - GPU Integration
  - Integrated and cooperative host/driver software stack with storage devices

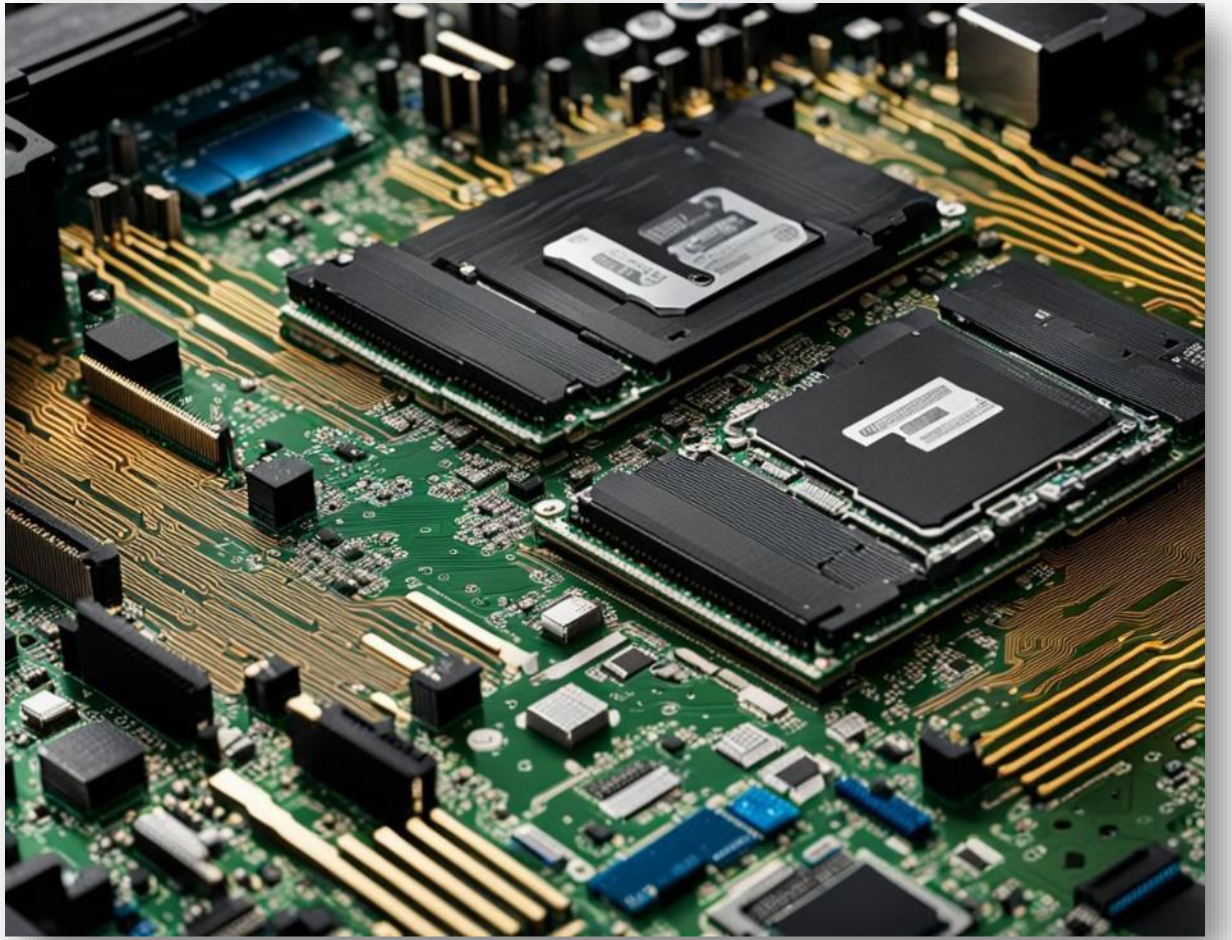


SNIA and AI

- Persistent Memory Programming Model
- Green Storage (Emerald Program)
- Security Standards
- Vendor-Neutral Object Storage (CDMI)
- Automotive Storage
- Near-Data Compute (Computational Storage)
- Smart Data Accelerator Interface (SDXI)
  - Example to follow!

REGIONAL SDC 24

Memory Infrastructures

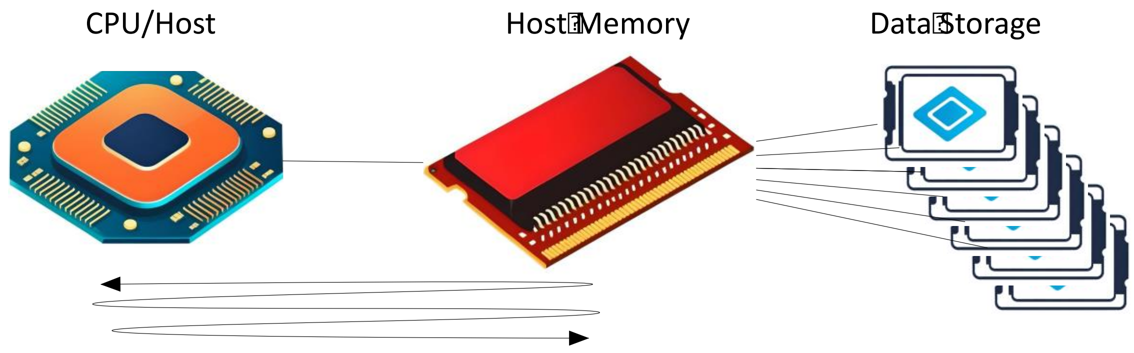


- High-Concept Futures
  - Computational Fabric-Attached Memory
  - Hierarchical memory pooling
  - Intra- and Inter-processor network fabric end-points
  - Disaggregated multi-access Ethernet-based storage/data
- Low-Level Efficiency Improvements
  - Kernel-Bypass for memory access
  - In-process data mutation
  - Processing-near-data

REGIONAL SDC 24

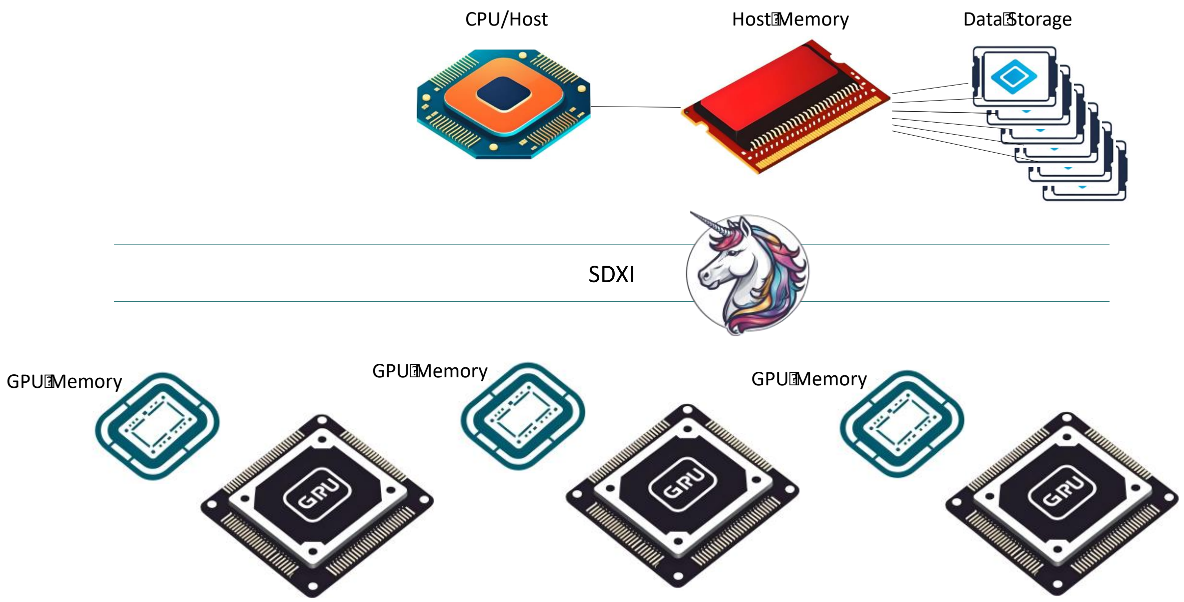
Memory Normalization – Example

- AI processing data is created/prepped in host memory
  - Cannot simply ingest the data from host memory – it's usually in storage
  - Must bring data from storage to host memory
- Data must be cleaned and prepped
  - Data structure/formats are changed
  - This happens in **host memory**
- Prime use case for Computational Storage (CS)
  - CS contributes to data prep and cleaning



Memory Normalization – Example (cont.)

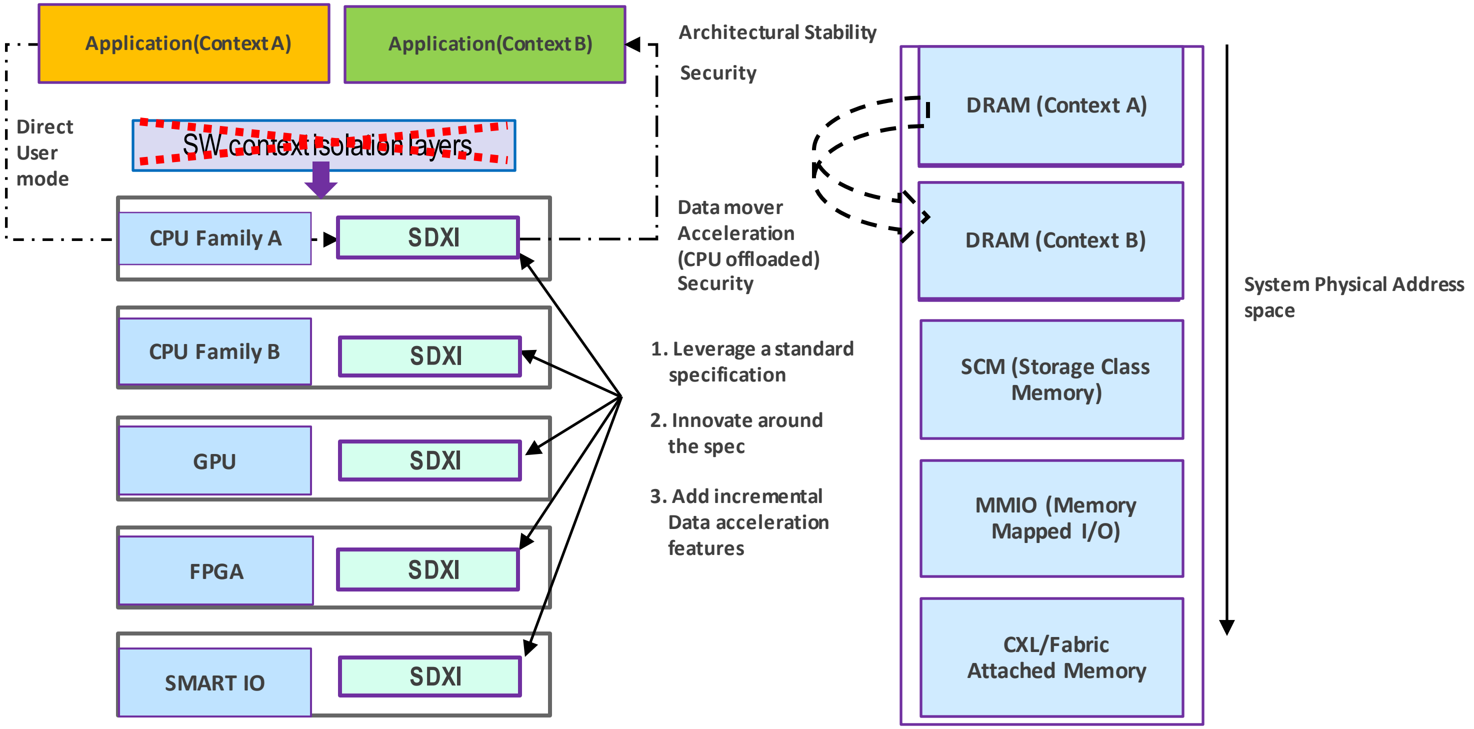
- Unicorn formatting
  - Varying data formats and intermediate data representations used in AI/ML data pipelines
    - E.g., file, Columnar, Binary, Text, Tabular, Nested, Array-based, Hierarchical
  - Need to build accelerator operations to be able to get to a format that AI models can be used to share weights (e.g.)
    - Example: sharing tensor vectors, lists of memory pointers
  - Tensors may be in different address spaces like Host Memory, GPU Memory, etc.
- Need operations to be able to perform:
  - Format Conversions
  - In-memory Vector/Tensor transformations like quantization, scaling, matrix operations, etc.
- Vendor-specific accelerator operations weaken TCO
  - Possible Solution: SDXI



REGIONAL SDC 24

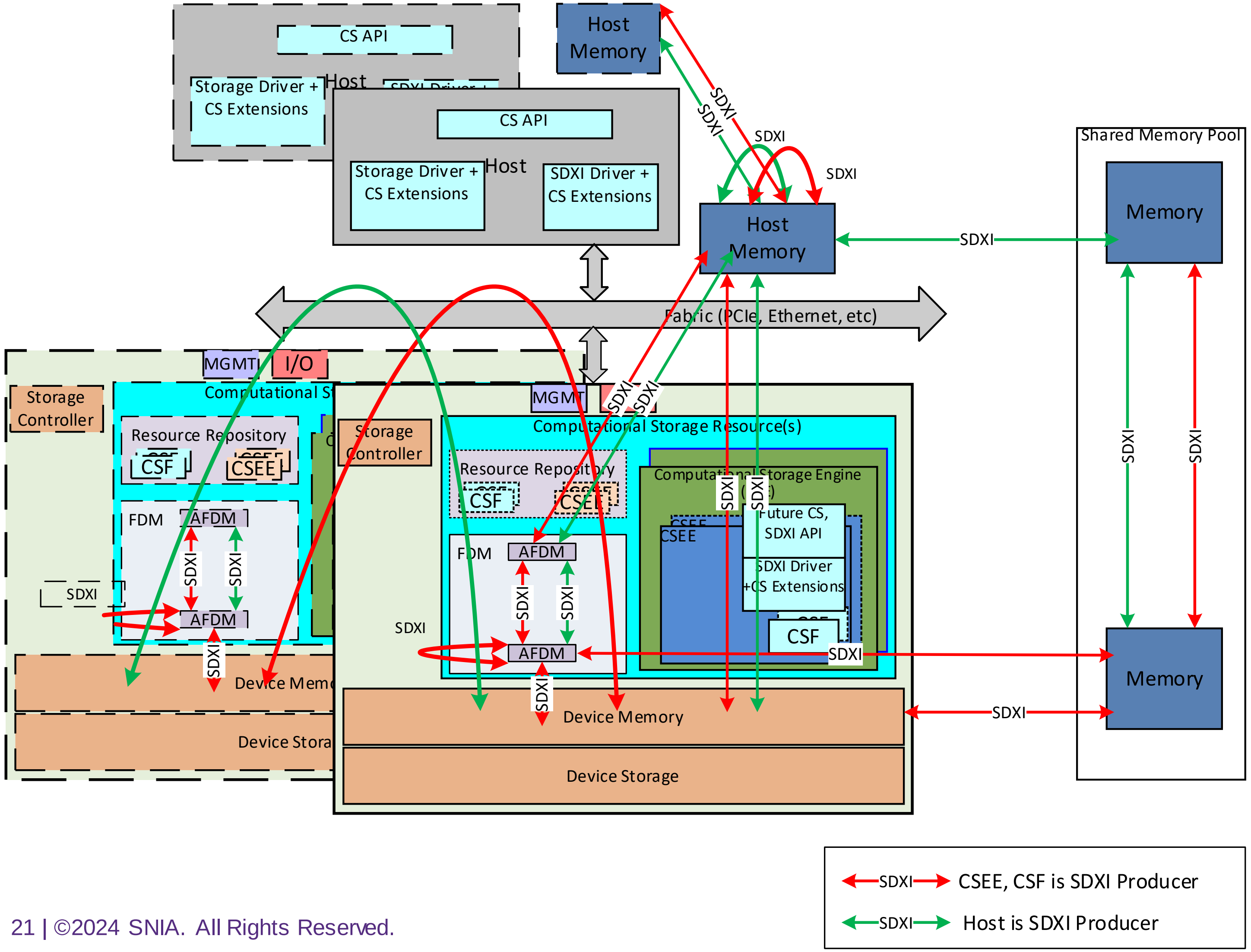
SDXI for Memory Normalization

- SNIA standard for a memory-to-memory data movement and acceleration interface
  - Low-level raw memory data movement
  - Data restructuring and transformation completed in-memory
- Extensible
- Forward-compatible
- Independent of I/O interconnect technology
  - Data movement between different address spaces
  - Standard extends to in-memory Offloads/transformations leveraging the architectural interface



Source: SNIA. SDXI Memory-To-Memory Data Movement.

Stacking Technologies – SDXI and Computational Storage



- Multiple SDXI producers in Computational Storage architecture
  - Enables data movement across multiple active functional memory regions
- Reduce tromboning (round-tripping) with host environment for chained data processing
  - Data cleaning, structure alignment, encryption/decryption, data mutation, etc.

REGIONAL SDC 24

The Road Ahead

REGIONAL SDC 24
BY Developers FOR Developers

UEC Addresses AI Network Needs

Traditional RDMA-Based Networking	Ultra Ethernet Consortium
Required In-Order Delivery, Go-Back-*N* recovery	Out-of-Order packet delivery with In-Order Message Completion
Security external to specification	Built-in high-scale, modern security
Flow-level multi-pathing	Packet Spraying (packet-level multipathing)
DC-QCN, Timely, DCTCP, Swift	Sender- (and/or receiver-) based congestion control across multiple paths
Rigid networking architecture for network tuning	Semantic-level configuration of workload tuning
Scale to low tens of thousands of simultaneous endpoints	Targeting scale of 1M simultaneous endpoints

ultraethernet.org

REGIONAL SDC 24

SNIA Addresses AI Storage Needs

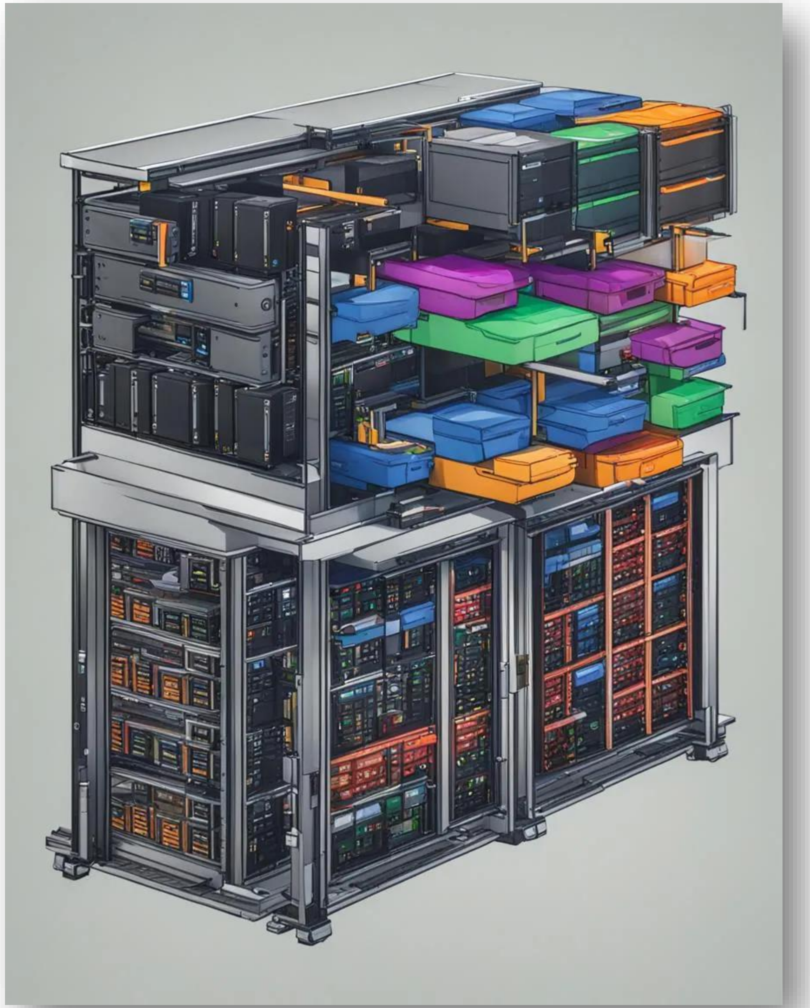
- Standards for AI-Driven Data Storage
  - Computational Storage Architecture 1.0
  - Computational Storage API 1.0
  - SNIA Emerald™ Power Efficiency Measurement Specification v4
  - Native NVMe-oF Drive Specification v1.0.1
  - Persistent Memory (PM) Performance Test Specification (PTS) v1.0
- Best Practices for AI Data Management
  - Swordfish™ Scalable Storage Management API Specification v1.2.6
  - Flexible Data Placement; Zoned Storage Models v1.0
- Collaboration with AI and Data Science Communities and Technologies
  - Current: CXL, DMTF, Open Fabrics Alliance, NVM Express, SODA Foundation, The Green Grid, among others
  - In process; Ultra Ethernet Consortium, Open Compute Project, Linux Foundation Projects, among others
- R&D Initiatives
- I/O Traces, Tools, and Analysis (IOTTA) suite
- Advocacy for AI-Friendly Policies

Conclusion

REGIONAL SDC 24
BY Developers FOR Developers

Summary and Key Takeaways

- AI Workloads are capitalizing on solid foundations in networking and data storage, but also requiring new ways of thinking
- Processing, Memory, Networking and Data are intersecting in new and non-traditional ways, and at scale much larger than ever before
- Boundary limitations (memory, bandwidth, processing, latency) are shifting both physically and logically
- The problem requires broad, open support for both networking and data storage services
- UEC and SNIA are working towards standardized, open, industry ecosystems to solve these problems



THANK YOU

Please take a moment to rate this session.

REGIONAL SDC 24
BY Developers FOR Developers