SCSI Trade Association
A SNIA Community

# Storage Trends in AI


**Patrick Kennedy**
Principal Analyst, ServeTheHome

**J Metz**
Chair, SNIA
AMD, Technical Director

**Jeremiah Tussey**
At-Large Board Member, STA
Microchip Technology Inc. Alliances Manager, Product Marketing, Data Center Solutions

# SNIA Legal Notice

- The material contained in this presentation is copyrighted by SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

# Housekeeping

- Ask questions in the dialogue box on your screen
- Questions will be answered at the end of the webinar
  - If we don't get to your question today, we will post a follow up blog to answer it
- Slides are available for download via the SNIA Educational Library
- Polls will be presented and participation is encouraged

**Polling question:** What is the biggest challenge when it comes to AI storage?

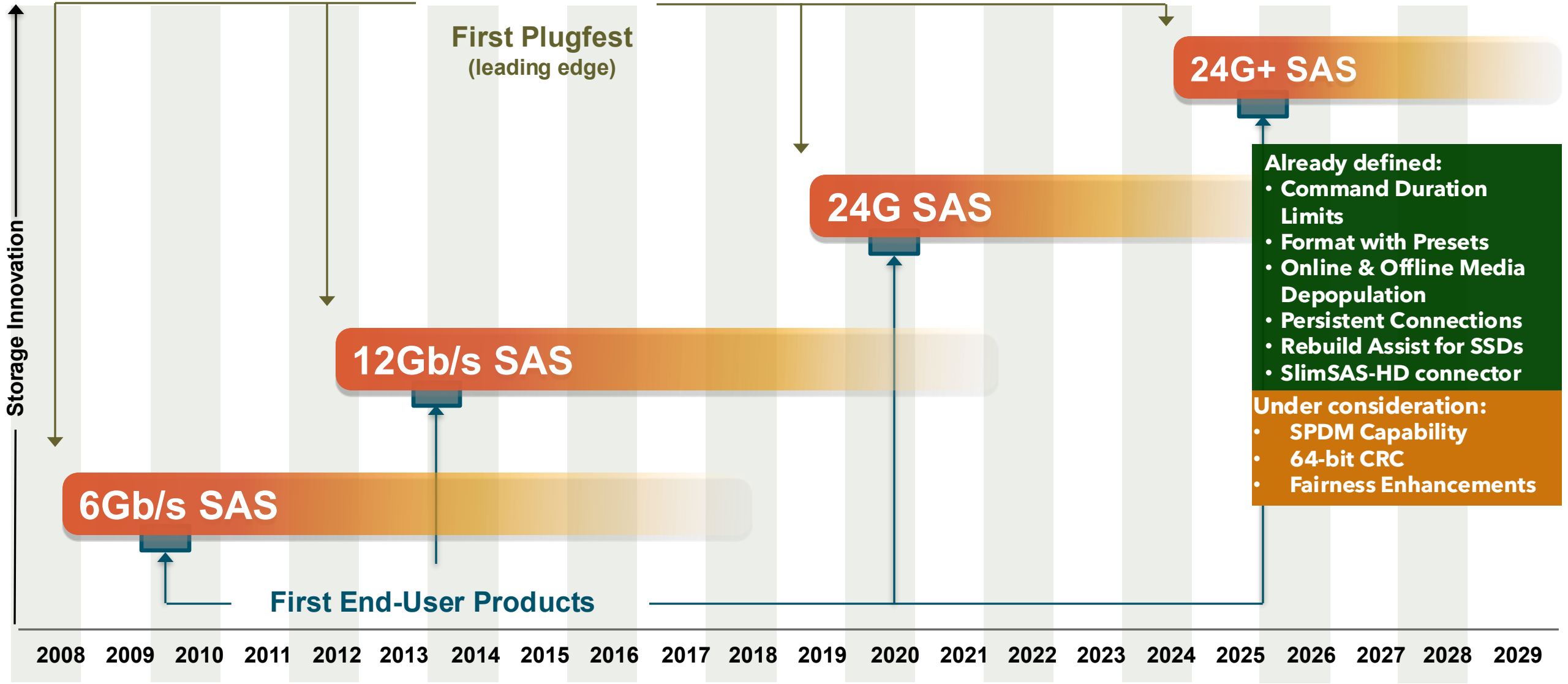
- Managing data volume
- Data security
- Impact on workload performance
- Cost & ROI

(answer in the polling window)

# Storage Architecture

# SAS Technology Roadmap




Updated March 2025

# 24G+ SAS Feature Overview

| | Feature | Benefit |
|---|---|---|
| **Already Defined:** | | |
| Command Duration Limits (CDLs) | Gives user control of HDD latencies | Better supports SLAs |
| Format with presets | Ability to format as SMR or CMR HDD technology. | Reduction of SKUs |
| Online & Offline Media Depopulation (Depop) | Ability to isolate media components and logically remove them from data set. | Less service calls in the data center; allows drives to degrade in place, extending the life of the device, thus improving sustainability. |
| Rebuild Assist for SSDs | Device reports failed LBAs to reduce read time of stripes; | Improves the rebuild time for RAID systems |
| Persistent Connections | Ability to sustain a connection | Less overhead & higher performance (overhead is associated with opens & closes) |
| SlimSAS-HD connector (SFF-TA-1016) | Adding a higher density connector | More connectivity out of a low profile board and adding compatibility with CopperLink internal cables. |
| **Under Consideration:** | | |
| SPDM Capability | Attestation and authentication of devices | Advanced systems security |
| Grow the PI (Protection Information) field | Greater capacity, support 64-bit CRC | Increase data reliability |
| Fairness enhancements for Large Topologies | Expanders assess quality of service across the topology | Improved quality of service for all devices. |

**Designed to enhance SAS architectures that include 24G SAS, 12Gb/s SAS, and 6Gb/s SATA products**

# Phases of AI and Storage Workload Requirements

**Ingest**
Importing vast amounts of raw data

**Transform**
Preparing data into Proper formats for analysis, cleaning and validating

**Training/Tuning**
Processing through AI algorithm, evaluate results and refine for accuracy

**Inference**
Model recognizes new patterns and extrapolates conclusions


## Ingest
- High capacity
- Low power
- High sequential read throughput

## Transform
- High capacity
- Low power
- Random reads
- High sequential write throughput

## Training/Tuning
- High burst throughput
- Random reads
- Occasional large sequential writes during checkpoints
- Low latency
- Higher GPU utilization

## Inference/Tuning
- Mix of random reads and writes
- Quickly feed compute resources
- Reduce idleness

# AI Storage Problems To Be Solved

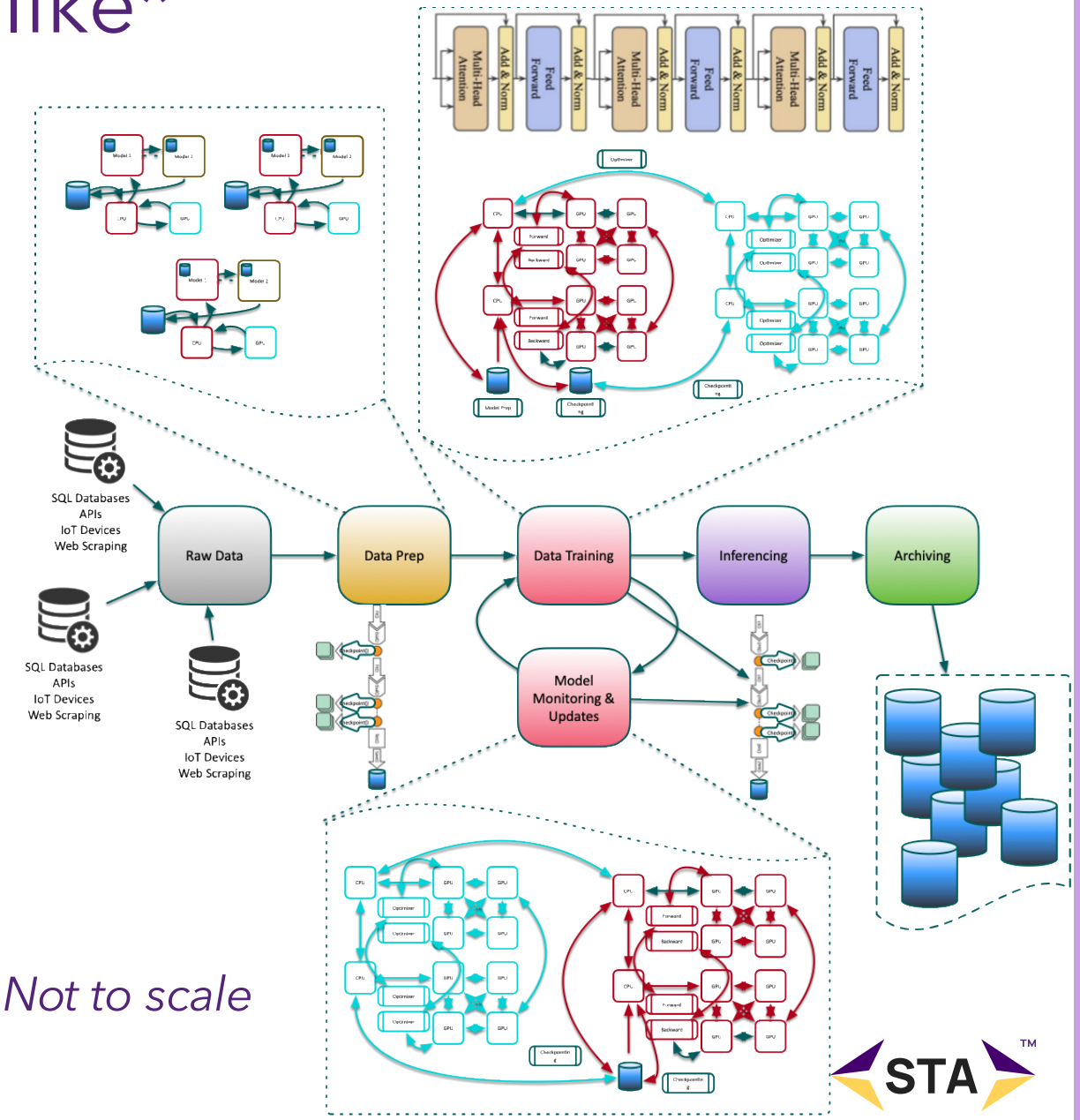
# The AI Monster

- 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)

# What workloads are really like*

- Storage needs vary at each stage
  - Ingest –
    - High capacity and sequential write performance
    - High capacity, heavy read, moderate write
  - Data prep –
    - Sequential read and write performance
    - Moderate capacity, heavy read/write
  - Training –
    - Random read performance
    - Moderate capacity, high read/write
  - Checkpointing –
    - Sequential write performance
    - Moderate-to-high capacity, high read, moderate-to-high write
  - Inference –
    - Random read/write performance
  - Archive –
    - Very High Capacity
    - Heavy capacity, heavy read, moderate write

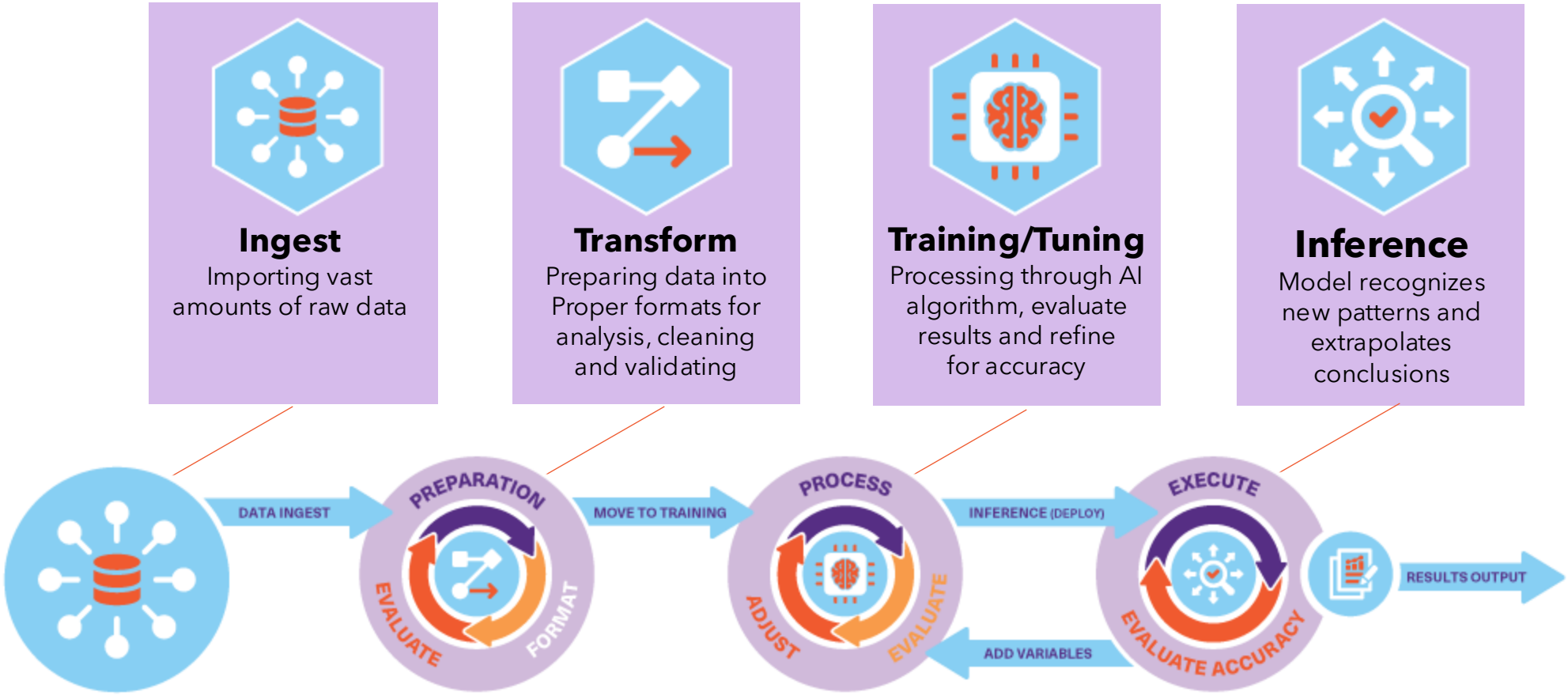



** Not to scale*

# Phases of AI and Storage Workload Requirements

## *SNIA Initiatives*




### Ingest

- Storage Data Placement (SDP)
- SDXI
- SFF – Connectors, Transceivers, Form Factors
- STA
- NVMe-oF Drive Spec.
- DNA (future)

### Transform

- Computational Storage
- SDXI

### Training/Tuning

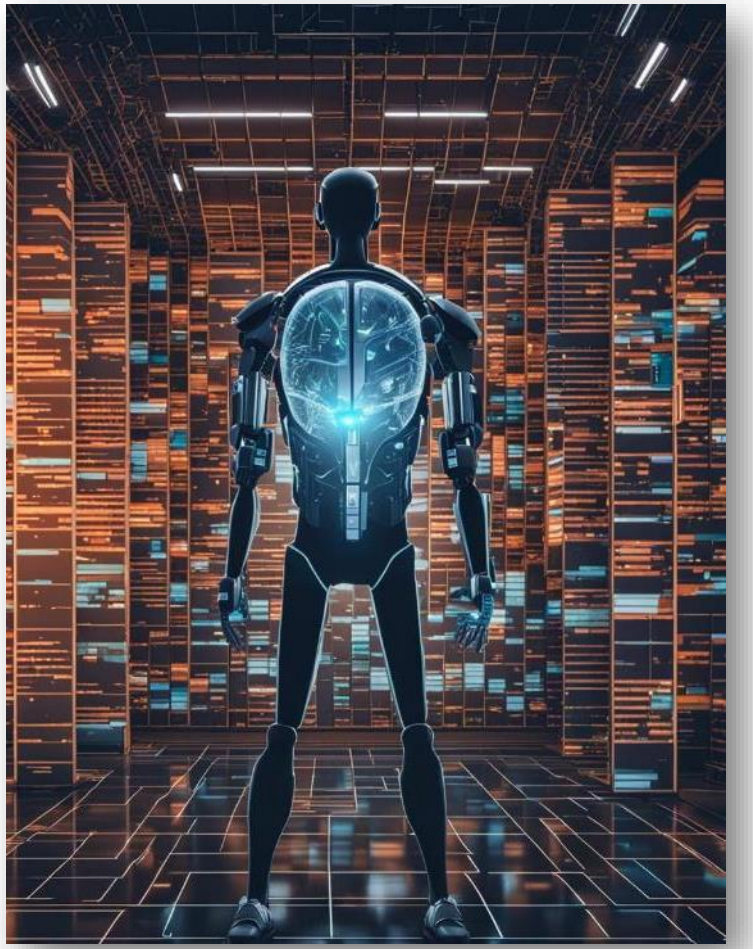
- Computational Storage
- DPU
- Green Storage/Energy Efficiency

### Inference/Tuning

- Computational Storage
- DPU
- Green Storage

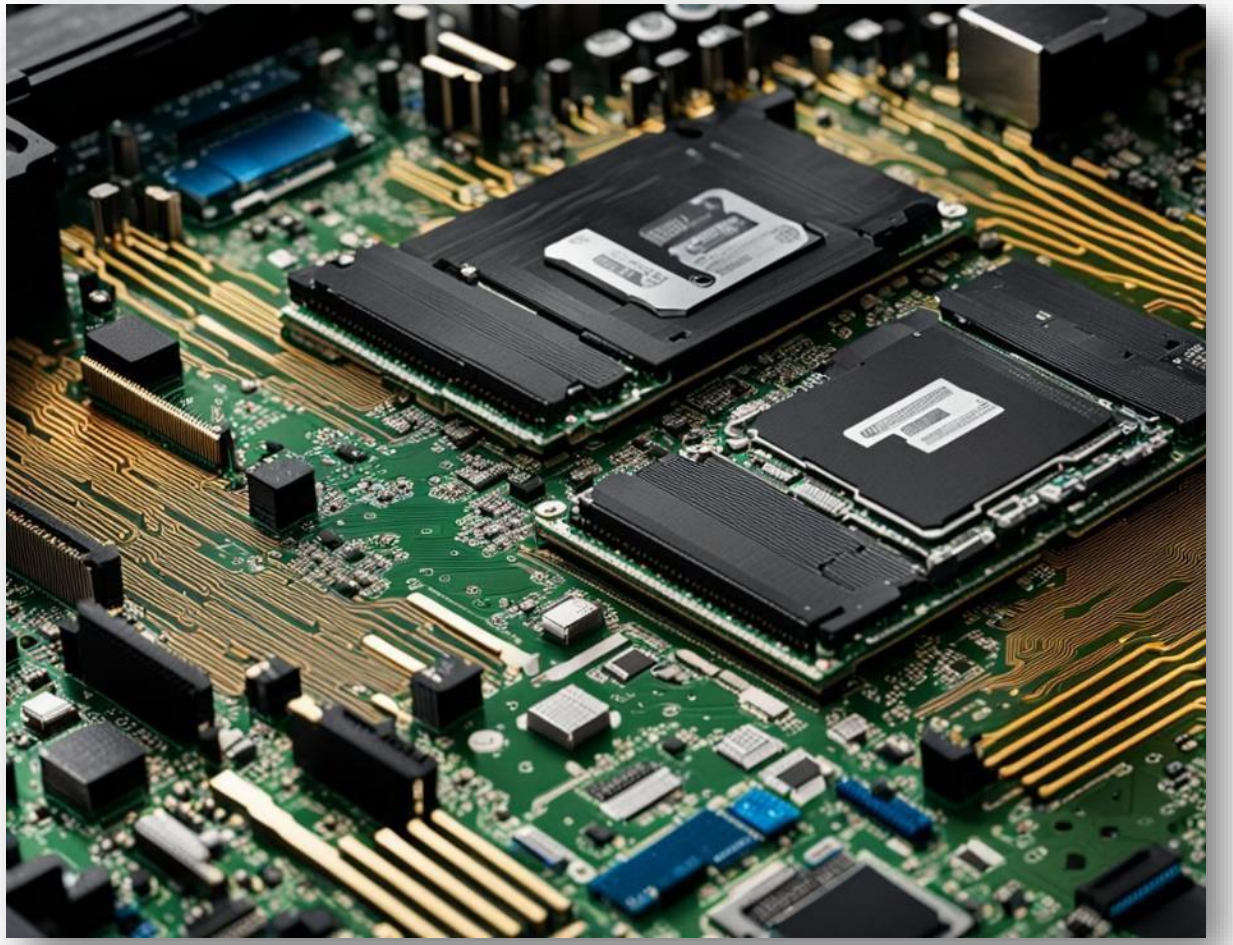
# Storage AI Needs

- Scalability and Performance
- Data Diversity and Edge Computing
- Cloud Integration
- Cost-Effectiveness
- AI-Specific Features



# 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

# SNIA Addresses AI Storage Needs


- Standards for AI-Driven Data Storage
  - Computational Storage Architecture 1.0 & 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
- R&D Initiatives
- Advocacy for AI-Friendly Policies

# Industry Ecosystem – It's a Big Problem!

- Everyone has a stake in the game:








… and more!

STA™

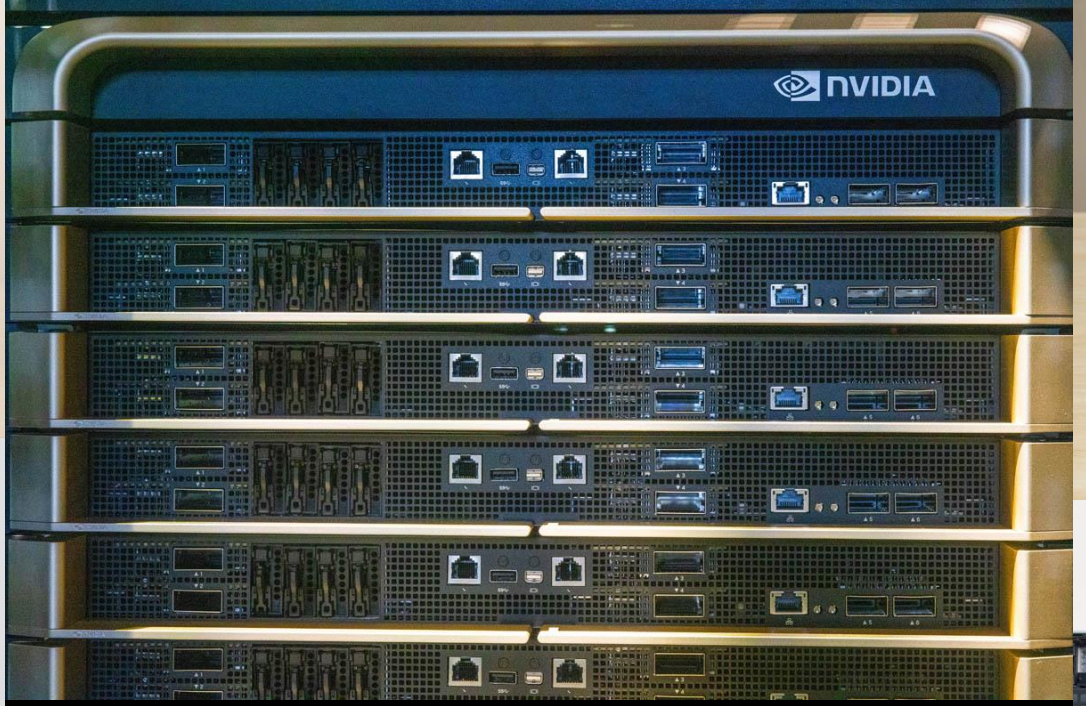
**Polling question:** Which aspect of storage for AI workloads is most critical?

- Speed
- Capacity
- Reliability
- Cost

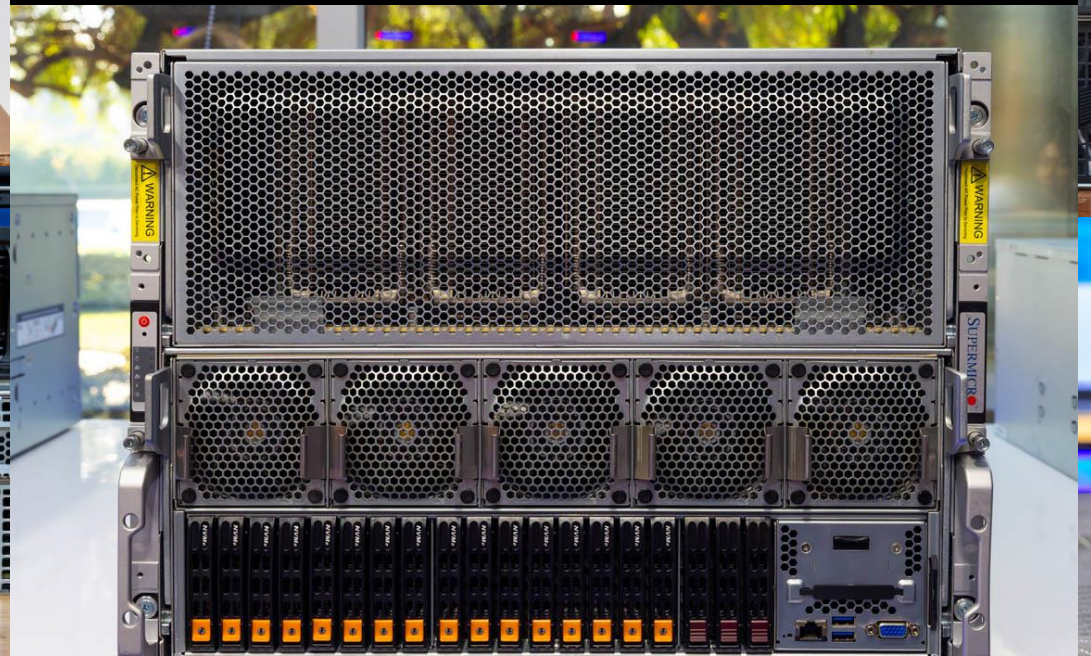
(answer in the polling window)

# Real World Insights

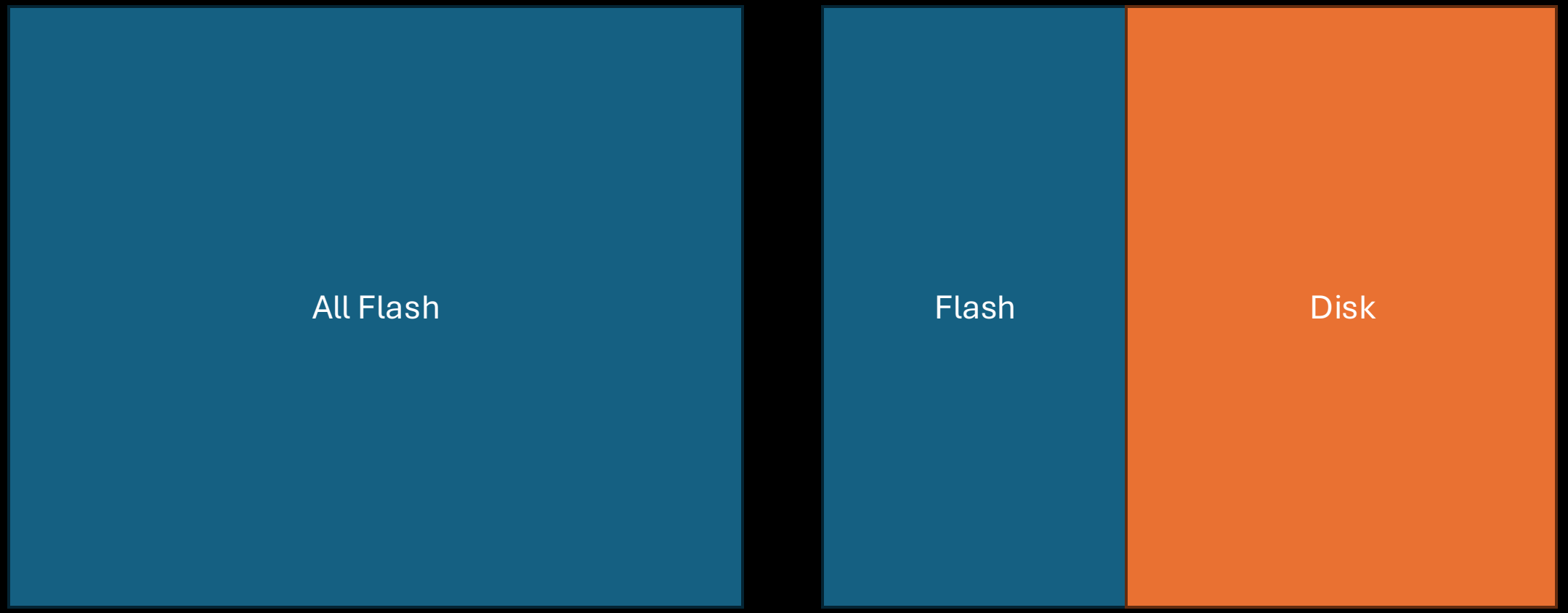



# AI Servers and Local Storage




# Shared Cluster Storage

## Two Camps in AI Clusters



# Where is Data Today?

Enterprise Arrays

Cloud Storage

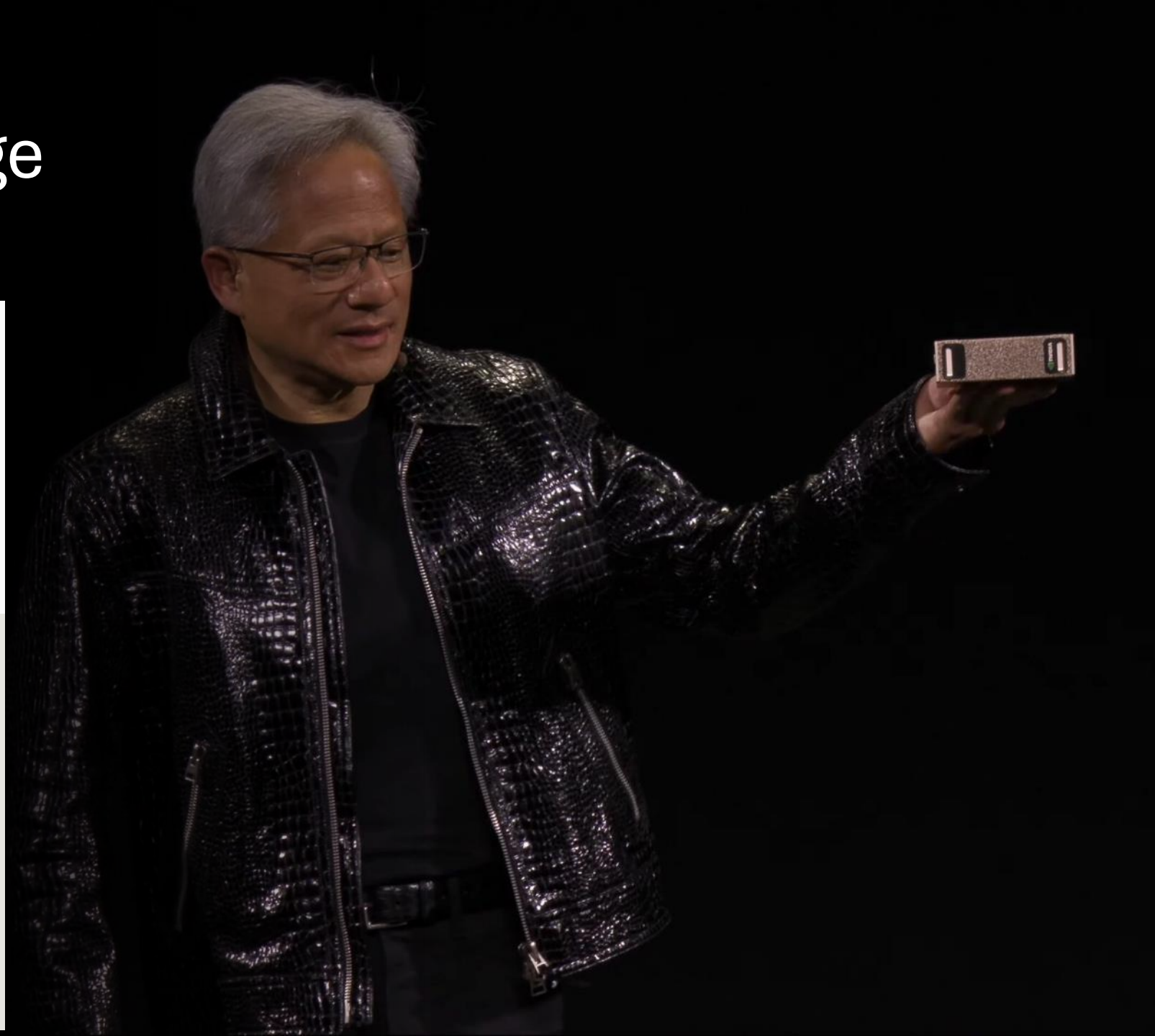
Local Storage

Back-up Storage

**Key Questions**

- Are the arrays fast enough to feed AI systems?
- Does the data proximity to AI accelerators meet the enormous speed and latency demands of an AI cluster?
- Folks focus on the GPUs and AI accelerators, but how do you solve for interconnects?

# What is the storage for AI Agents?



# Where Does AI Generated Data Land in 2027?

Enterprise Arrays

Cloud Storage

Local Storage

Back-up Storage

**Key Questions**

- Compliance needs
- Cost
- Performance
- When to tier

# Opportunities & Challenges

# Thank you!

**Patrick Kennedy**
Principal Analyst

https://www.servethehome.com/

https://www.youtube.com/ServeTheHomeVideo

**Dr. J Metz**
Chair, SNIA

https://www.snia.org/

https://www.linkedin.com/in/jmetz/

# Follow STA

https://www.snia.org/sta-forum

https://www.linkedin.com/company/snia/

**Serial Attached SCSI Playlist** on SNIAVideo
https://www.youtube.com/@SNIAVideo

https://x.com/SNIA