MSST
The International Conference on
Massive Storage Systems and Technology

# The Perfect DPU

## for Storage Target Markets in the Agentic Era

MSST

June 2nd 2026

Ted Weatherford

VP Business Development

Xsight Labs

Sales@xsightlabs.com

# Outline

- Xsight Labs Introduction
  - 800G DPU overview
- Agentic Era Memory Wall
- Highest Density KV Cache Rack System
  - Open Flash Platform Organization: Just a Bunch of Flash (JBOF) system

# Xsight Labs

Delivering the Right Products at the Right Time

## Xsight Labs Company Overview

- Fabless semiconductor company founded in 2017
- 260 Employees (60 contractors)
- Only start up driving end-to-end connectivity for AI factories, Edge AI and Hyperscale CSPs with two Ethernet chip products shipping today
- Exceeded 2025 revenue targets and on track to do the same 2026

## Strong ecosystem of investors & strategic partners:

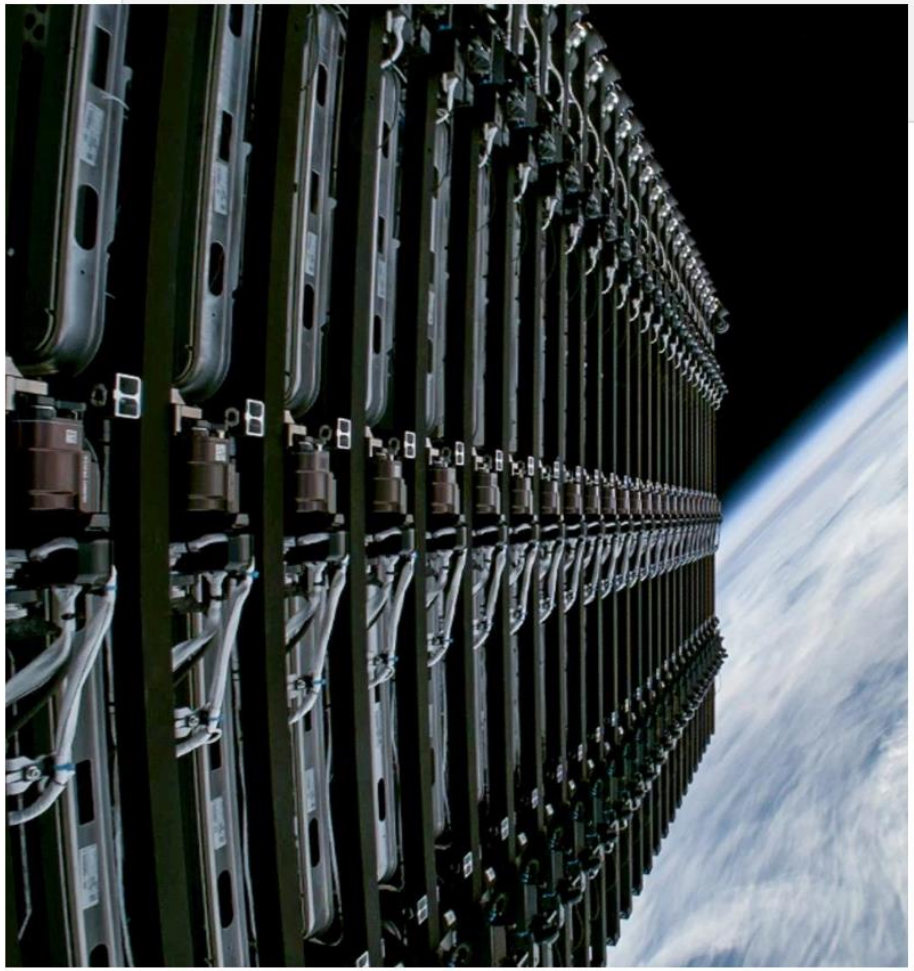
- Raised $440M up until today

ATREIDES MANAGEMENT · Fidelity · Battery · intel capital · M12 · AMD · MARVELL

## Multidisciplinary team with a proven track record:

- Previously **Galileo (Marvell Prestera) Technology, Mellanox, Annapurna Labs (AWS), Habana Labs (Intel) and Leaba (Cisco SiliconOne), Astera Labs (ALAB)…**
- Avigdor Willenz is the Chairman of the board and founding investor

# Xsight's X2 Switch Enables Starlink V3 Satellite



- X2 routes traffic within satellite and between satellites; terabits down and gigabits up.
- Delivers the performance required for rapid data processing, stable telemetry, and real-time operational responsiveness for Starlink
- X2's flexible, programmable architecture adapts seamlessly to diverse orbital and ground-based networking workloads
  - Developed a custom data plane application
- Performance Testing:
  - Extensive throughput and latency tests
  - Full suite of environmental qualification tests in extreme environments confirming its readiness for deployment in orbital missions, including
    - Vibration
    - Radiation
    - Temperature
    - Thermal stress

# Philosophy

- Open
- Programable
- No performance compromises

E-Series DPU

- Cloud on a Chip.

  *Compute adds acceleration, not acceleration adds compute*

  *Its just a Linux ARM server capable of 800G!*

X-Series Ethernet Switch

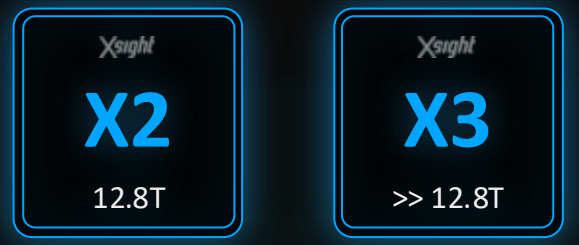
- Started with a blank sheet of paper, building an architecture to last for decades

  *Data Center Switch Performance but with a true SDN programming model!*

# Differentiation: Open, Software Defined, Performance/Watt


## E-Series – SDN Infrastructure Processor

- **Full control plane and data path programmability** – using standard programming models, enabling customers to innovate at the pace of software not hardware.
- **Standard-based approach** avoids vendor lock-in and enables an open ecosystem
- High compute density based on ARM®
- **OPEX and CAPEX** savings
- **Scalable** from 8 to 64 ARM cores to cover broad application space from control plane to storage to AI infrastructure




## X-Series – Open ISA Ethernet Switch

- **OPEX and CAPEX savings**, driven by industry leading power and cost for its class
- **Scalable new switch architecture** enables easy migration from current 12.8T solution to next generations, and beyond (eg 409.6T)
- The industries only programmable switch, along with **Industry's first open ISA** ($X^{ISA}$), enabling customers and partners innovation, flexibility, feature velocity and Time-To-Market

# The Main AI Building Blocks are 6 Chips

## The 6 Main Chips Required for AI Token Factories 2026-2030…


- Heart of the mathematics behind AI Learning and Inference. Memory and I/O bandwidth extremely high (HBM).
- Orchestrating the system. DRAM centric memory; Many DRAM controllers. Multi-threading. High Power per DMIPs.
- Off-loading CPU with reliable secure virtualization: networking and storage. Front-end networks & Storage Targets
- Highest performance, fixed pipeline networking interfaces. Typically dominated by one vendor for RDMA and IB.
- Low Latency for XPU/GPU to XPU/GPU in system or in rack clustering. Going to multi-rack soon. Copper today for reliability and lowest cost.
- Ethernet embedded memory switches provide the performance, economics and scale (L3 and Radix). Allows 1M plus nodes.

# Xsight provides 2 of the main 6 now

## The Main 6 of the AI chip building blocks

2026

1. XPU / GPU 
2. CPU 
3. DPU  E-Series
4. NIC 
5. Scale-Up Switch 
6. Scale-out Switch  X-Series

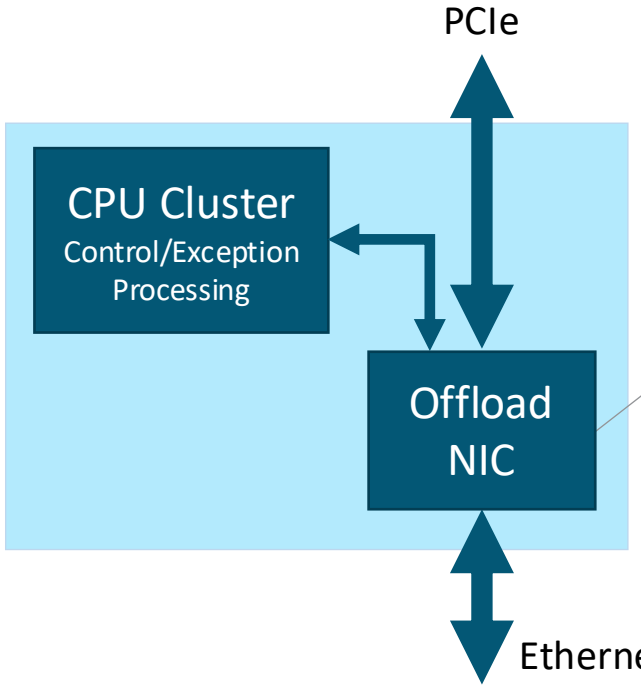
- DPU provide intelligent front end Ethernet terminations
- DPU integrates Network, Compute, Storage and Security
- DPU replaces x86 in flat Warm FLASH Context Cache Tiers
- Teams building DPU can by extension build CPUs & NIC

- A Scale-out Switch allows large connection space efficiently
- Ethernet continues to dominate, destine to expand to backend

# DPU Architectural Differentiation

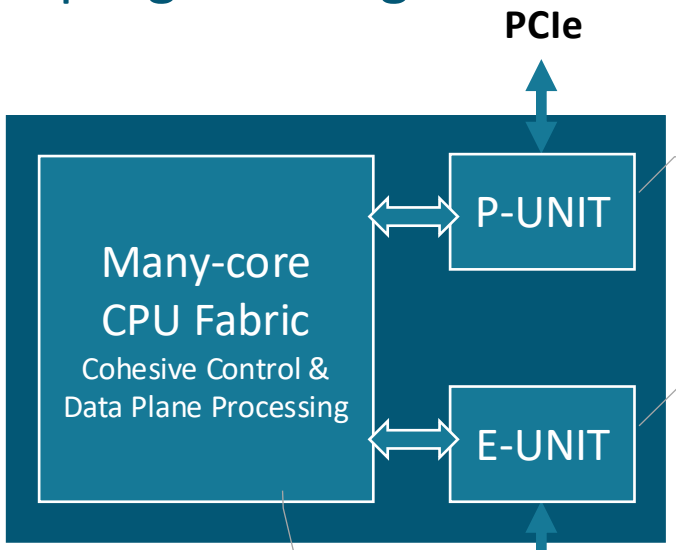
## Fundamentally different approach to dataplane programming

### Traditional Smart-NIC/DPU




- **Inline Encryption**
- **Packet Processing**
  - *Stateless offloads*
  - *Parse-match-action pipeline*
- NPU-like SW development
- Constrained Buffering

Nvidia Bluefield | AMD Pensando | Intel IPU




- **Device Emulation Services**
- **Inline Storage Offloads**
- **Packet Processing**
  - *Stateless offloads*
  - *Parse-match pipeline*

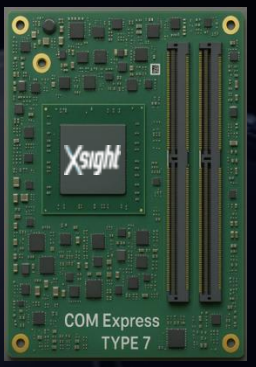
- **Inline Encryption**
- **Inline Packet Processing**
  - *Stateless offloads*
  - *Parse-match pipeline*

- **Scalable Performance**
- **Standard SW Development Model**
  - *Continuous Innovation*
  - *Flexibility & Feature Velocity*
  - *Release Schedule Determinism*
  - *Full Visibility, Debuggability*
- **Flexible/Deep Buffering**


Xsight Labs EPU | AWS Nitro

# E-Series




SOC | Edge Server | AIC | COMEX

# E-Series' High Lights

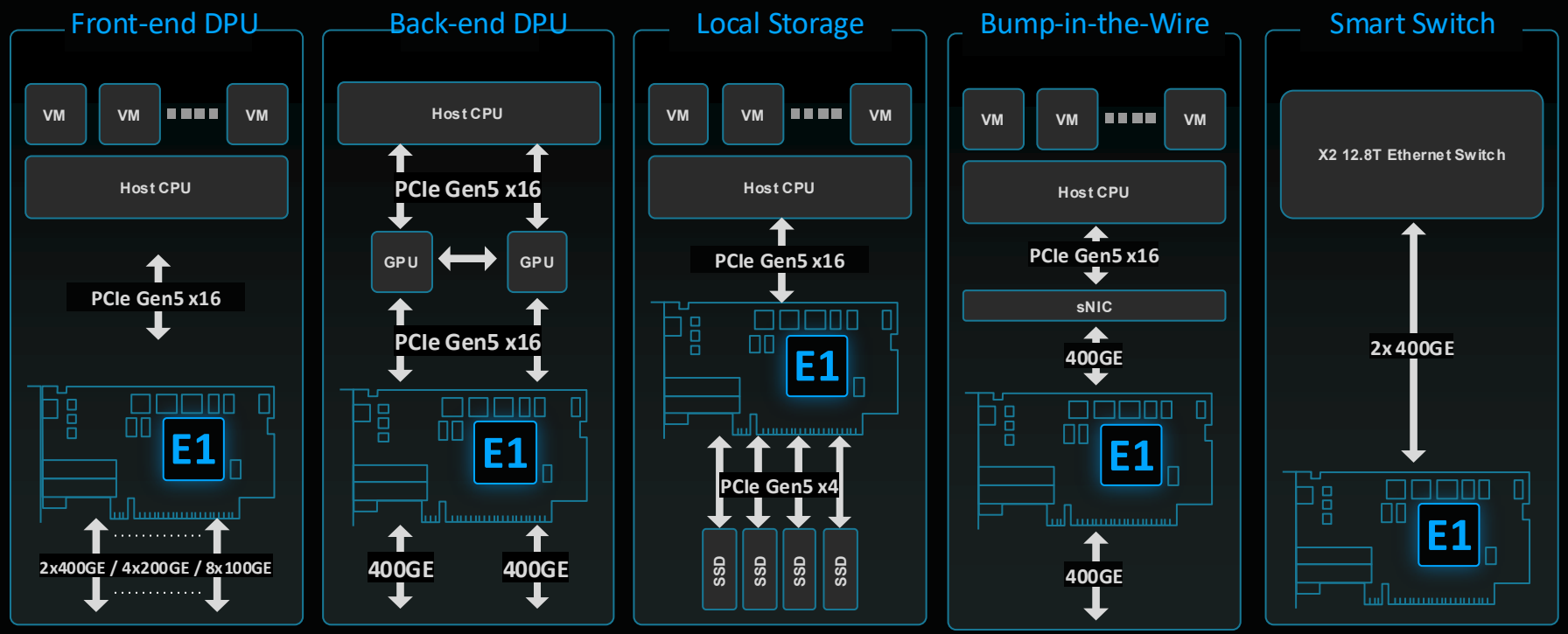

- ARM System Ready Level6 Server Ready; Linux Stack
- High Memory BW w/four DRAM Banks
- First to 800G networking over all DPUs
- 800G OF NVMe over Fabric I/O
- Soft RDMA RoCE for required agility / interoperability
- No slow path behavior
- Open true line rate SDN Datapath on Linux DPDK
- Lower Power than BlueField4; save 70% energy OPEX
- One year ahead of the pack!

# E1 DPU Example Applications

## Standards-based Ecosystem Driven

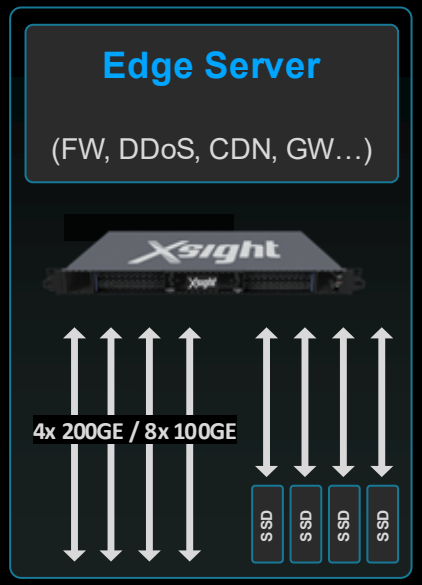
debian · eBPF · Linux · ubuntu · SONiC · SPDK · DPDK

Open-Source Libs., Apps, Linux Dataplane, OS, Distros, ARM® v9, ARM® SystemReady™, L4-L7, Crypto, ML, RAN, CCA, Trusted FW

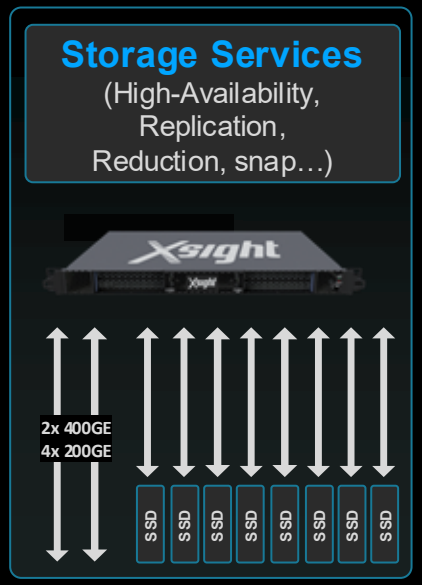



- DPU Smart Switch
- Software Load Balance
- TLS/kTLS Offload
- Stateful Dist. NGFW/IPS
- Multi-host optimized Smart NIC
- NVMe-oF/TCP Initiator & Target
- Bare Metal Provisioning
- TLS/DTLS Termination

# E1 Edge Server and Storage Applications




- Linux Arm System Level 6 Ready Servers
- CDN WebServer
- 5G & 6G Applications
- Stateful Distributed Firewall IP IP Services
- SDWAN, VPN Routing/Internet Gateway
- Software Load Balancers
- >10,000 Arm N2 cores per rack; <20 kilowatts




- Distributed AI Storage Systems
- Warm AI Storage Server
- NVMe-oF/TCP Initiator & Target
- Open Flash Platform Organization
- Open Compute Platform Organization

# Xsight Labs' E1 800G DPU
# Used as a Flash Warm Storage Server

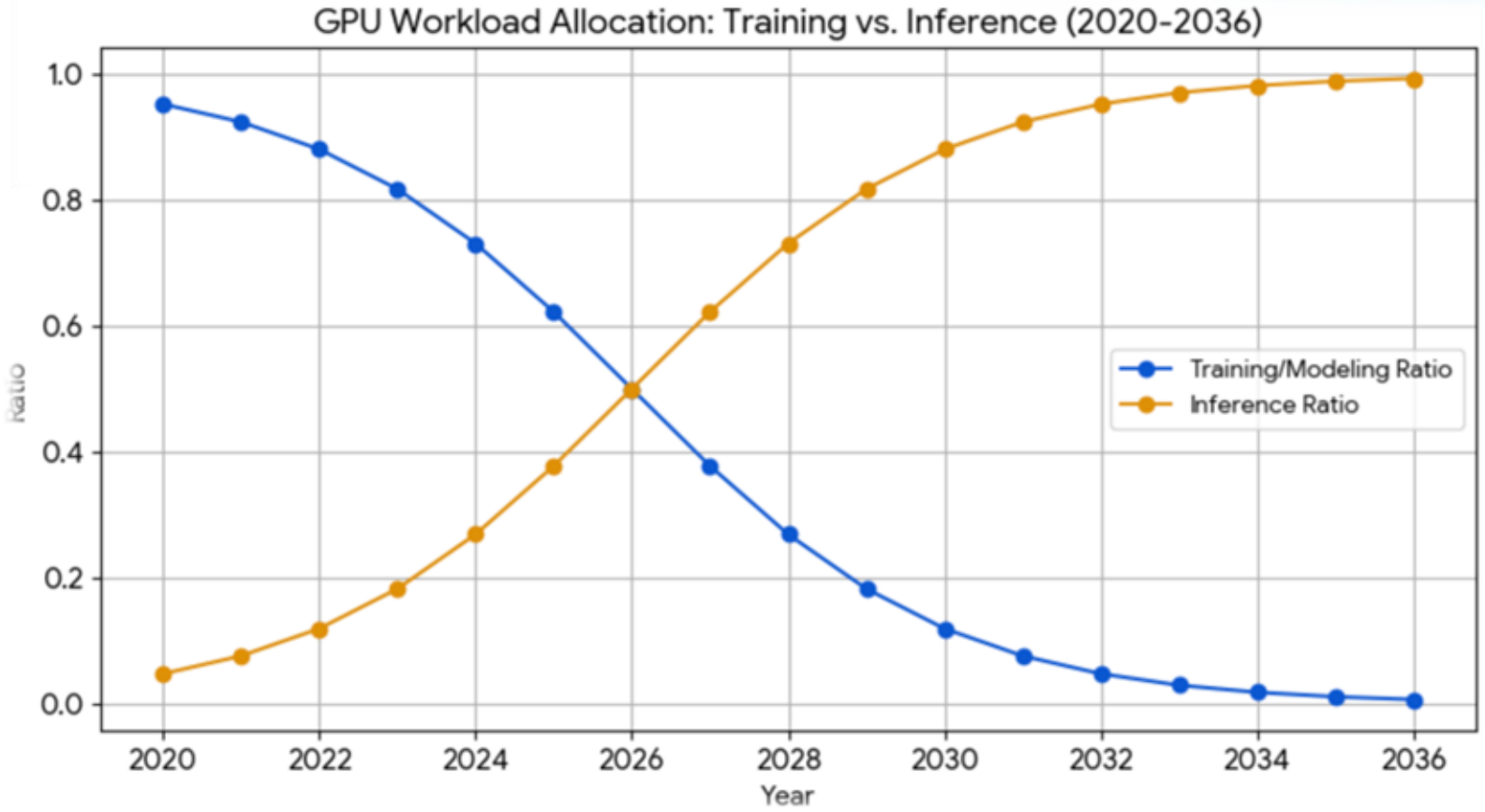
800G
DPU


E-Series
"Edge Server SOC"

- SDx Linux Programming Model
- 64 x N2 ARM Cores
- 8x112G (2x400G, 4x200G, 8x110G/50G/25G
- n5 TSMC
- 40 lanes PCIe Gen5
- 4 x Channels DDR
- 75W Typical


# Agentic Era Memory Challenges

# Inference Dominates over Training Now




*A huge transition in AI workloads and memory architectures begins in 2026. This is a shift from compute-centric training to Context Centric Agentic Inference.*

**2020–2025**
The period was characterized by heavy training loads as organizations raced to build frontier foundation models

**2026**
This year marks a pivotal inflection point where inference capacity requirement begins to eclipse training.

**2030–2036**
The workload is dominated by persistent, stateful inference, where models are rarely "newly trained" but are constantly performing complex, multi-turn reasoning tasks that require deep access to long-term memory.

# THE HBM MEMORY WALL is here…

To support this shift toward agentic inference, the memory hierarchy per POD is undergoing an exponential divergence…

We hit the ”Context Memory Wall”-
the limitation where model context exceeds available GPU HBM

This necessitated the adoption of a tiered storage approach where Flash (JBOF) acts as an AI-native memory extension

# Why is Flash Exploding in the Agentic Era?

1. **The "Context Memory Wall":** HBM is limited by physical constraints on the GPU package (die area, power, cost). You cannot simply increase HBM to infinite capacity. Conversely, Flash storage is highly scalable and cost-effective.
2. **Stateful Agentic Workflows:** Modern AI is shifting from "stateless" (where every query starts from scratch) to "stateful" (where agents retain memories of past interactions). Storing years of agent history requires Petabyte-scale capacity that only a Flash JBOF can provide.
3. **The Role of the DPU:** As we discussed, the DPU is the critical enabler here. Without the DPU to manage the "Flash-to-GPU" data pipeline via RDMA, a 114:1 memory ratio would be unusable because the CPU bottleneck would stall the inference process. With the DPU, this huge Flash capacity acts as an "extended memory pool" that the GPU can pull from as needed, effectively creating a near-infinite context window for your models.

# FLASH Storage Dominates AI Mem. Tiers

- The most critical trend is the **massive scaling of the Flash JBOF tier**.
- Network attached FLASH (JBOF) is the only way to practically scale AI Agents

| AI Memory Types | 2020 Capacity | 2036 Est. | Driving Factor |
|---|---|---|---|
| SRAM | 0.003 TB | ~0.015 TB | Physical die area constraints (on die GPU & XPU) |
| DRAM | 0.3 TB | ~6 TB | System-wide overhead and staging CPU mainly |
| HBM | 0.4 TB | ~50 TB | Dense tensor computation requirements |
| Flash (JBOF) | 1.0 TB | ~2000 TB | Persistent, long-term agentic context |

- By 2036, inference PODs will require Petabyte-scale context storage to maintain the historical state, RAG-indexed libraries, and reasoning chains necessary for advanced autonomous agents.
- The integration of **RDMA (Remote Direct Memory Access)** is what allows this vast flash pool to behave like a high-speed memory layer, ensuring that the GPU can pull "context blocks" from the JBOF with latency that is orders of magnitude lower than traditional disk storage.

# AI Inference HBM vs. FLASH JBOF per POD

- By 2036, inference PODs will require Petabyte-scale context storage to maintain the historical state, RAG-indexed libraries, and reasoning chains necessary for advanced autonomous agents.

- The integration of **RDMA (Remote Direct Memory Access)** is what allows this vast flash pool to behave like a high-speed memory layer, ensuring that the GPU can pull "context blocks" from the JBOF with latency that is orders of magnitude lower than traditional disk storage.

### The 2032 Ratio: Flash vs. HBM

Based on the scaling projections for high-performance AI inference PODs, the disparity between high-bandwidth memory (HBM) and the flash-based storage tier (Flash JBOF) is widening significantly.

In the year 2032, our projections for an enterprise inference POD indicate:

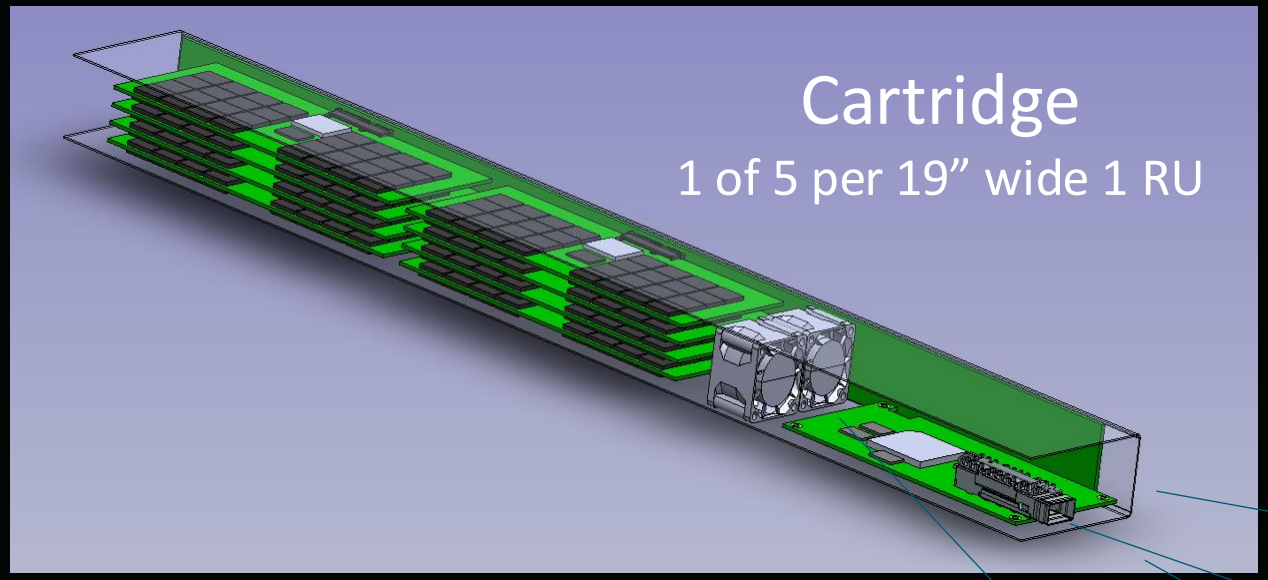
- **HBM (GPU-Local Memory):** $\approx 35\ \text{TB}$
- **Flash JBOF (Capacity Tier):** $\approx 4000\ \text{TB}$

The resulting ratio of Flash to HBM capacity is:

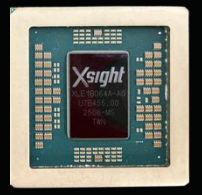
$$\text{Ratio} = \frac{4000\ \text{TB}}{35\ \text{TB}} \approx 114.3$$

This means that by 2032, the **Flash storage tier will be approximately 114 times larger than the HBM tier.**

# DPU as Storage Target-Edge Server with 2 x 400G



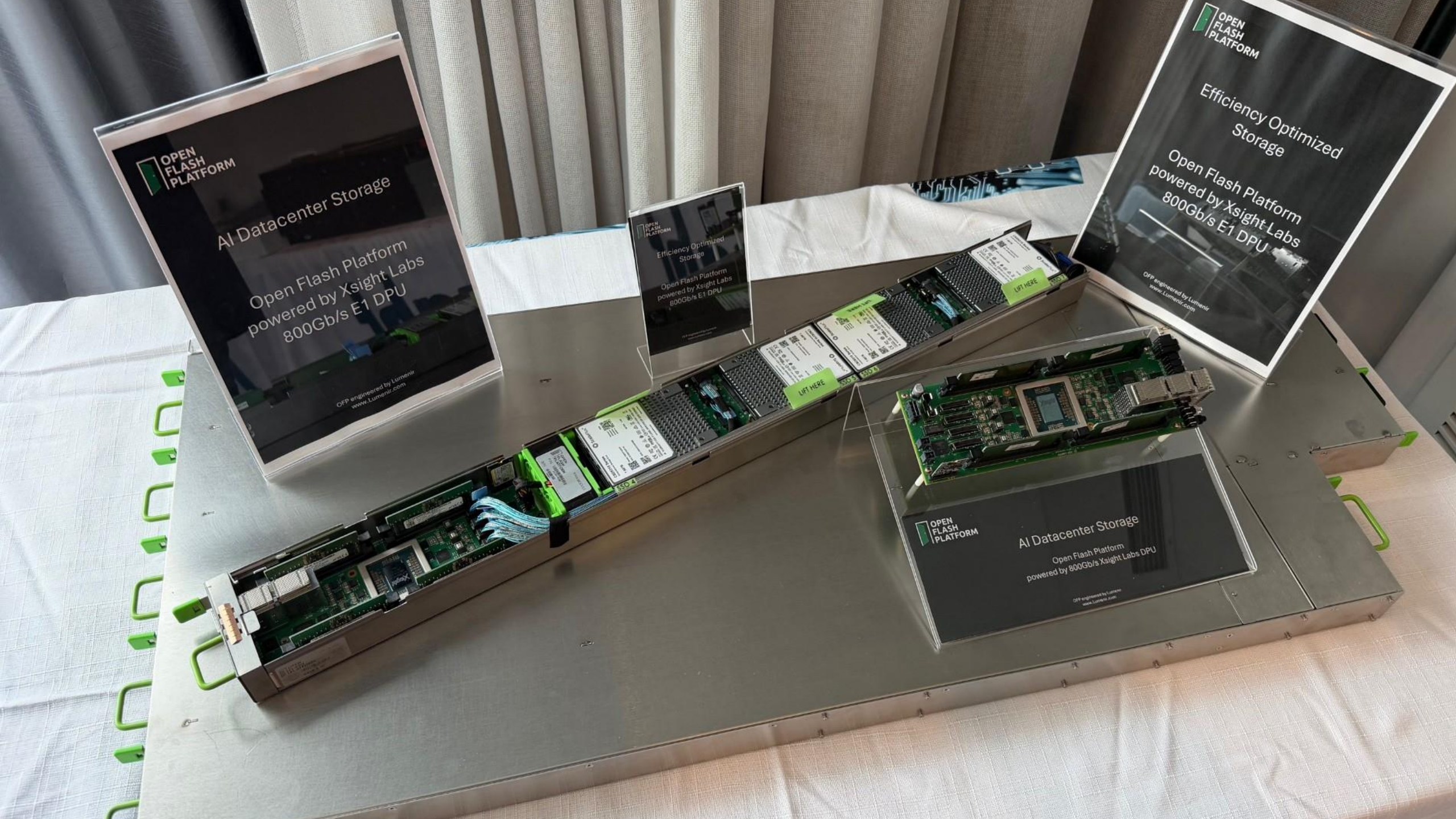

800G DPU



DPU Controller PCB

2x 400G QSFP-112

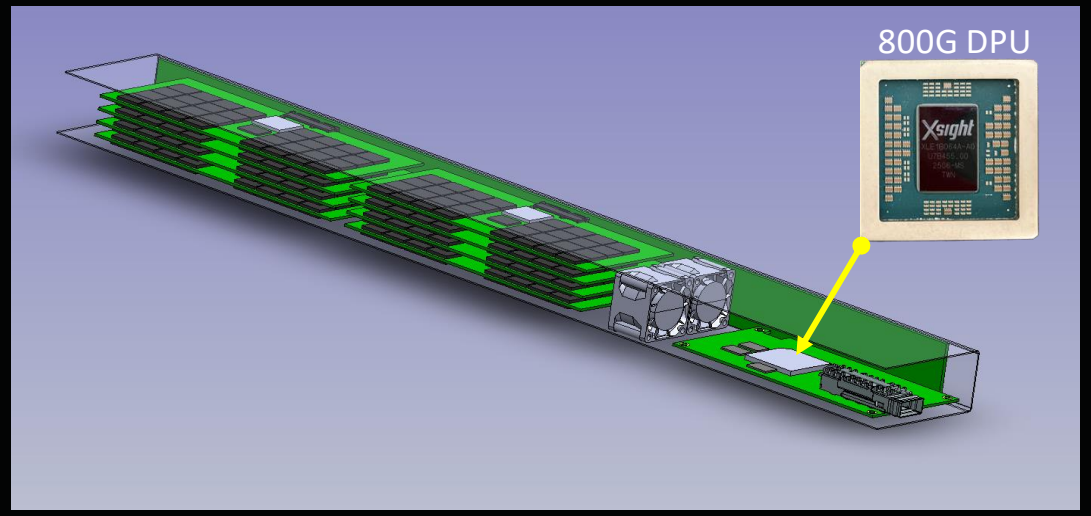
Xsight + HAMMERSPACE


OPEN FLASH PLATFORM
AI Datacenter Storage
Open Flash Platform powered by Xsight Labs 800Gb/s E1 DPU
OFP engineered by Lumenir www.Lumenir.com
OPEN FLASH PLATFORM
Efficiency Optimized Storage
Open Flash Platform powered by Xsight Labs 800Gb/s E1 DPU
OPEN FLASH PLATFORM
Efficiency Optimized Storage
Open Flash Platform powered by Xsight Labs 800Gb/s E1 DPU
OFP engineered by Lumenir www.Lumenir.com
LIFT HERE
OPEN FLASH PLATFORM
AI Datacenter Storage
Open Flash Platform powered by 800Gb/s Xsight Labs DPU
OFP engineered by Lumenir www.Lumenir.com

# OCP Version

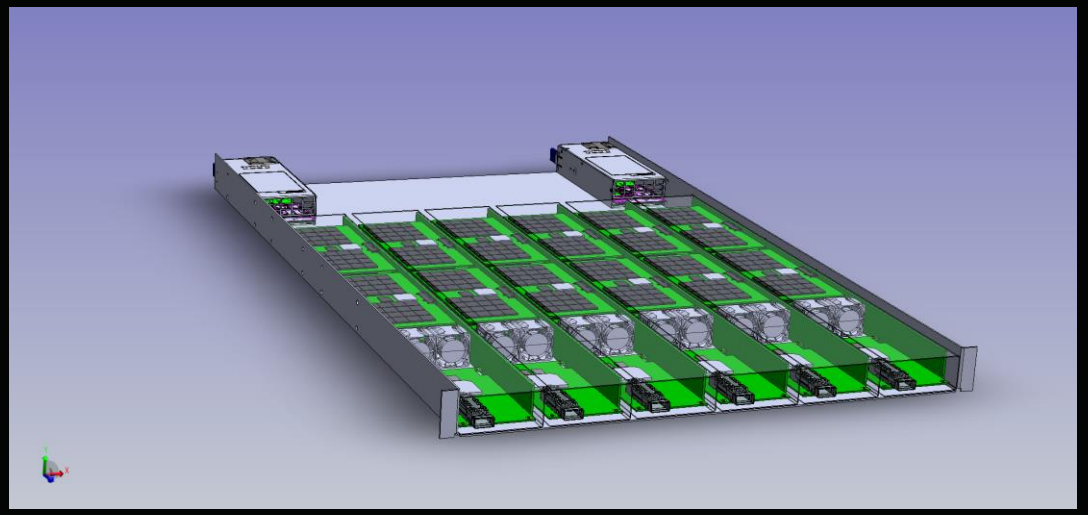
7

Cartridge



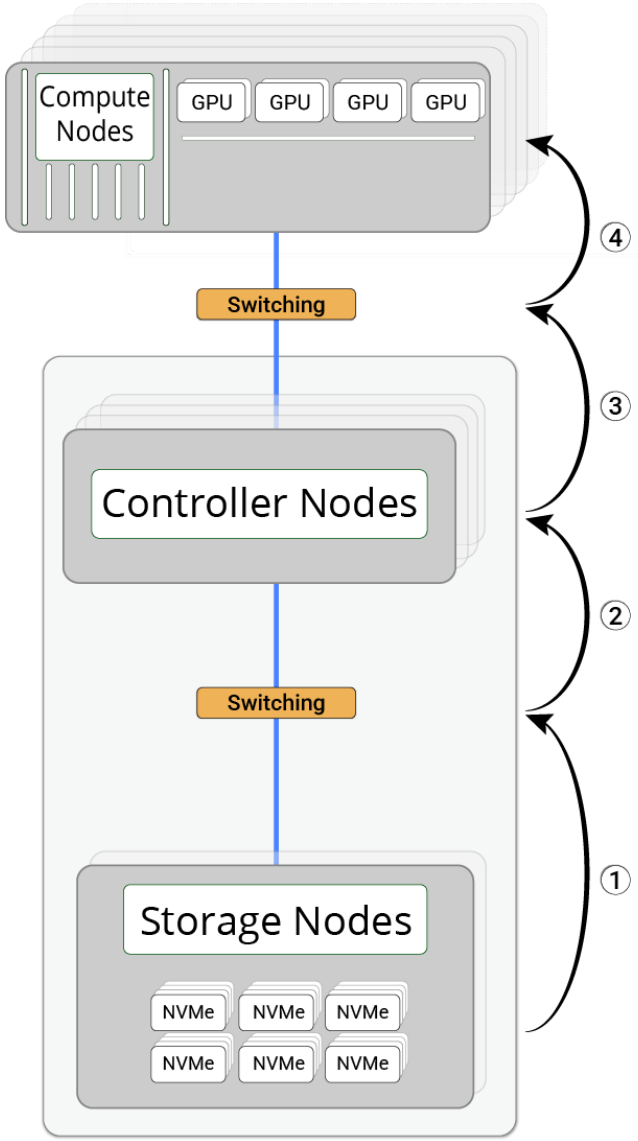

- One single 800G DPU Edge Server SOC
- 75W / SOC
- 480 GB/s Read, 240 GB/s Write
- 0.6 kW Read, 1.6 kW Write
- 100W Stand By Power

OCP Shelf; 6 Cartridges/RU



- 1 EB Byte per Rack (4 PTB / Cartridge)
- 16,128 Arm N2 Core per Rack
- 504 x 800G per Rack
- 42 Shelfs per Rack

# Open Flash Platform Simplifies Storage



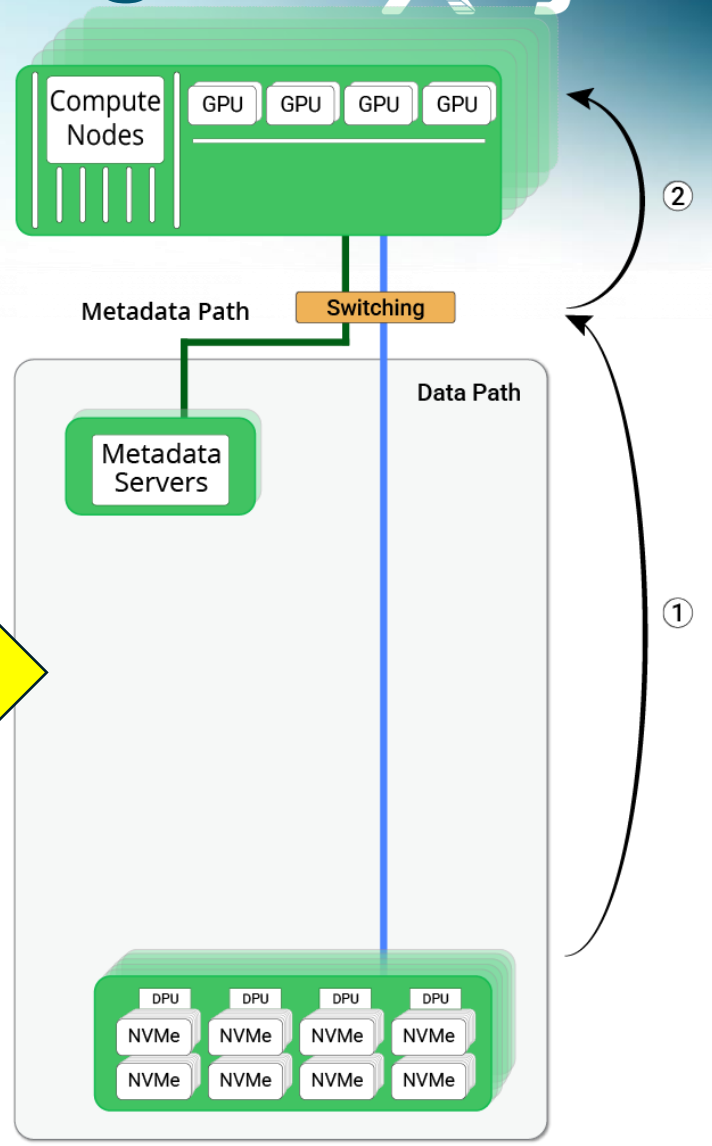

**THE OLD**

- **Proprietary data silo**
- **Limits performance**
- **Limits scalability**
- **Higher cost**
- **Higher complexity**
- **Inefficient**

**4 networking hops replaced w/only 2 steps**

**NEW, Simpler!**

- **Linux NFS on DPU**
- **No expensive x86**
- **Direct Storage Connection**
- **Distributed file system**
- **Collapse for performance and scale**



# World's First 800G DPU Powers OFP.org for NG AI Flash

## Why will CSPs and Enterprises Rapidly Adopt Open Flash Platform Systems?

**>10x**

**Increase in Density**

By optimizing the layout and integration of NVMe, OFP unlocks the true value of high-density flash.

**90%**

**Decrease in Power**

Replacing CPU-enabled storage servers with low-power IPU/DPU chips, OFP consumes far less power and generates less heat.

**60%**

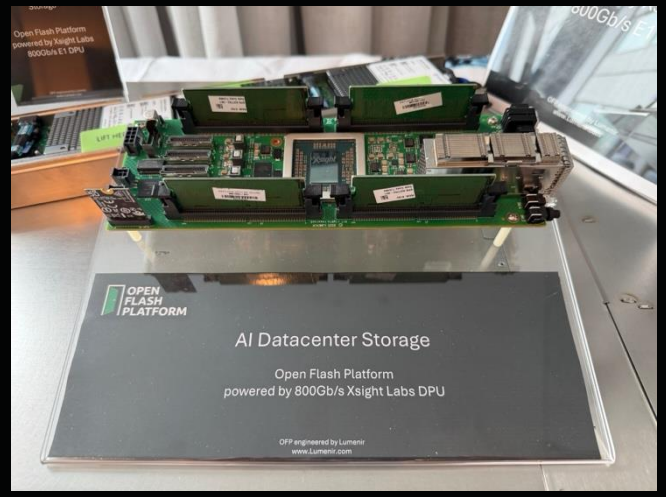
**Longer Operational Life**

Tie refresh cycles to the life of flash, rather than the life of a processor, resulting in a 3-year longer service life than the typical 5 years.

**50%**

**Lower TCO**

OFP reduces the amount of infrastructure, simplifies scaling, and extends the service life of storage investments.




Why did Hammerspace Select E-Series by Xsight Labs?

- 800G required
- Power & Area small
- More Processing  per Watt (leading efficiency)
- Simpler to use over other DPUs (AMD, INTEL, NVDA)
- Memory BW
- Feature Velocity of a Standard Linux model
- Great technical support

# Thank You