Storage Disaggregation: Infrastructure Fungibility & Efficiency at Hyperscale

Madhavan Ravi | HW Systems Engineer, Meta

MSST 2026

Agenda

1. The Problem Statement
2. The Solution
3. Storage Use Cases
4. Industry Landscape & Open Ecosystem
5. Key Takeaways & Future Directions

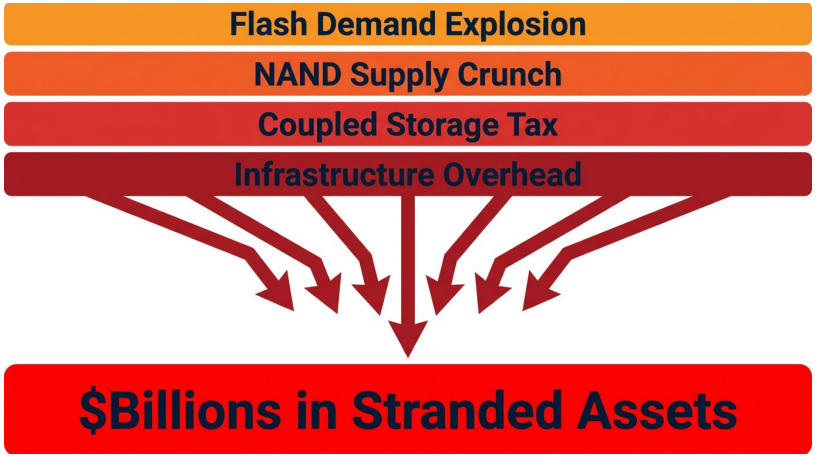
Why Coupled Storage Fails at Hyperscale

Flash Demand Exploding
30%+ CAGR growth; sub-optimal utilization leads to billions in stranded flash assets industry-wide

NAND Supply Shifting
Small-capacity SSDs harder to source; price volatility makes per-server SSD model unsustainable

Coupled Storage Inefficiencies
SKU proliferation, NPI overhead per server type, lifecycle lock-in between storage and compute

Infrastructure Tax
Single digit % host CPU consumed by infra daemons; encryption, virtualization, and packet filtering compete for cycles. Some new usecases make ir worse.



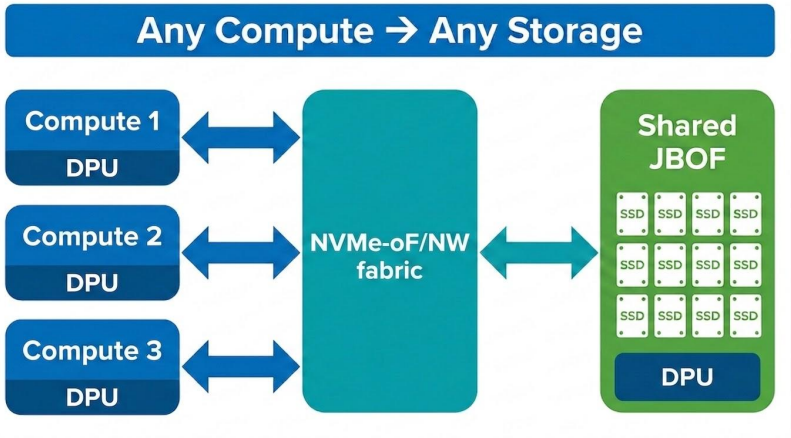
The Solution: DPU-Based Infrastructure/Servers

Architecture

- Replace per-server SSDs with shared DPU-managed JBOF
- NVMe-oF block interface (TCP, RoCEv2)
- Thin provisioning and overcommit

Infrastructure Offloading

- Storage + Network virtualization on DPU
- HW accel: compression, crypto, serialization
- 5%+ host CPU reclaimed for applications



Converged Building Blocks

2 Compute (x86, ARM) + 3 Storage (HDD, TLC, QLC)
= Any Server Shape

Storage Use Cases: From Boot to AI Inference

Boot Storage Disaggregation

- A JBOF provides boot volumes over NVMe-oF
- 2-4x overcommit (boot partitions mostly static)
- Eliminates per-server boot drive procurement
- Enables truly diskless compute

Impact: CapEx reduction, fewer HW SKUs, faster deployment

Data Storage for GP Compute

- Fabric-attached flash pool as block devices
- Services consume storage elastically
- DPU enforces QoS, encryption, isolation
- No more oversized servers

Impact: Collapse 6-10 server types into 2-3 building blocks

KV Cache for LLM Inference (the buzz these days)

- DPU-managed flash as KV cache offload tier
- NVMe-oF low-latency access from GPU nodes
- Flexible flash-per-accelerator sizing

Impact: Better tokens/s, longer context, flash BW parallelism, lower cost

Industry Landscape: DPU Ecosystem

Hyperscaler Deployments

- **AWS** - Nitro (pioneered DPU offload)
- **Google** - Intel partnership for custom IPU
- **Microsoft** - Acquired Fungible, MANA
- **Oracle** - AMD, Nvidia, Marvell + custom ‘Acceleron’
- **ByteDance / Alibaba** - Custom DPUs

Merchant Silicon Options

- **Xsight** - ARM cores + DPDK
- **Intel IPU** - ARM + DPDK + P4 MPU
- **AMD Pensando** - P4 MPU-heavy
- **NVIDIA BlueField** - Many ARM cores + P4
- **Chelsio T7** - Fixed-function HW offload
- **Marvell Octeon** - ARM + DPDK
- **Many FPGA based DPU** companies…..

Build vs Buy Trade-offs: Time-to-market | Power + Cost | Multi-sourcing | Long-term Optimization

Key Challenge

Different architectures, speeds/feeds, different management paths, differing server implementation feature support across vendors. Lack of open collaboration has resulted in industry fragmentation. GPU platform vendors (Nvidia, AMD) are also enabling DPUs for their own unique ecosystems.

Open Ecosystem Vision: Vendor-Neutral Smart Storage

Goal: Open, vendor-neutral smart storage architecture with flash as a shared, elastic resource

Standard Protocols

- NVMe-oF interfaces
- Common abstraction layers (host SW + DPU)
- Open management and orchestration APIs

Multi-Vendor Supply

- Multiple DPU vendors via open competition
- Better optionality and supply chain resilience
- Independent storage + compute scaling

Simplified Fleet

- Converge to 2-3 building block types
- Any rack serves any workload
- Fewer SKUs, faster NPI execution

Let's collaborate on open standards!

Key Takeaways & Future Directions

- **Infrastructure Fungibility** - Any compute can access any storage
- **NPI Simplification** - Fewer SKUs, faster development cycles
- **Fleet Efficiency** - Billions in CapEx savings via storage pooling
- **AI Readiness** - Elastic flash pools, KV cache tiering for LLMs ← can't present anything these days without saying 'AI' :)
- **DPUs as Enablers** - NVMe-oF, HW root-of-trust, encryption, isolation, more

Open collaboration is needed to end industry fragmentation in DPU offerings and Server Integration mechanisms

Thank You

Questions?