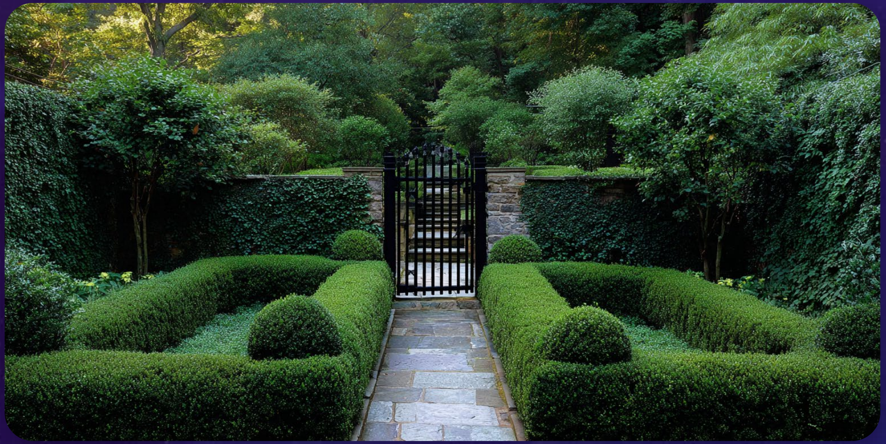
Leveraging Open Standards to Address the AI Data Crossroads

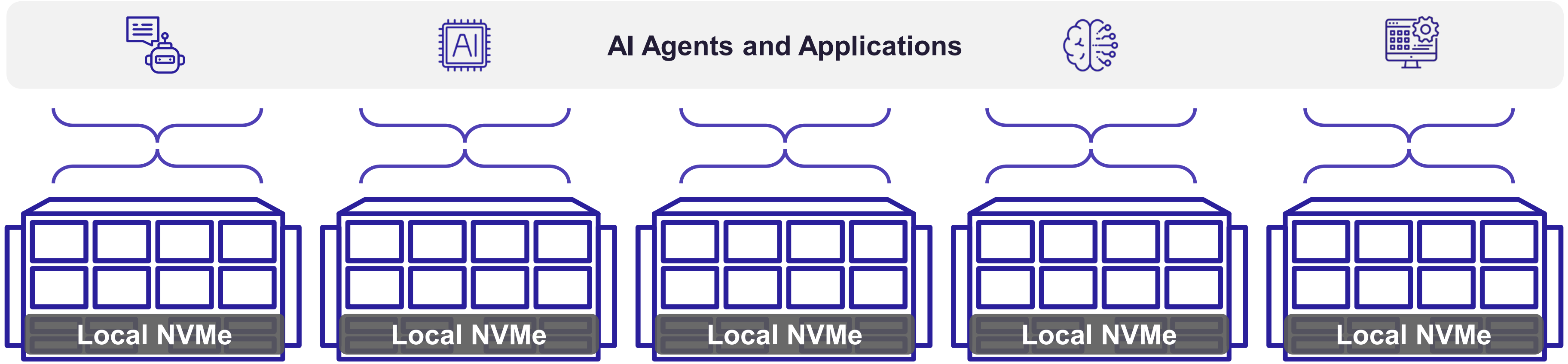
Building Open Highways vs Walled Gardens

- Performance: Tier 0 in GPU Server, no added software
- Capacity: extended operating life, power efficient Open Flash Platform
- Standards-based: your data, your AI, anywhere



- Performance: over provisioned, specialized infrastructure
- Capacity: limited operating life, power hungry
- Proprietary: copy proliferation, siloed, rigid environment

GPU Server-Local Storage is Siloed



AI Agents and Applications

Local NVMe

Local NVMe

Local NVMe

Local NVMe

Local NVMe

Local Storage Attributes

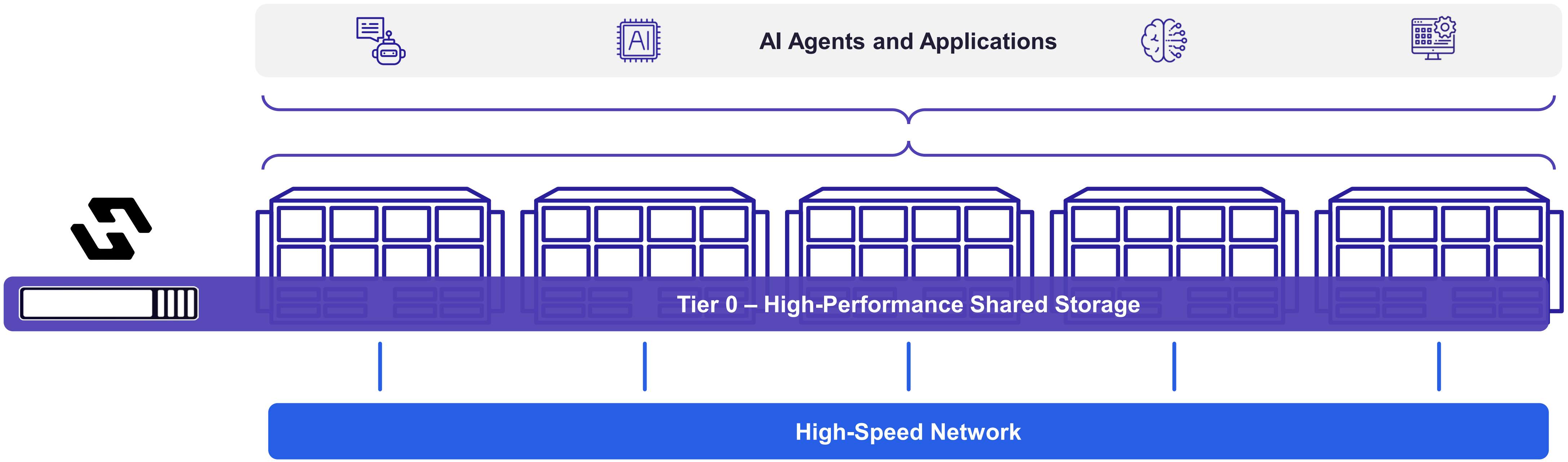
- Included with purchase (sunk cost)
- Physical servers and cloud VMs
- 60 TB to 492 TB per server today
- 1.97 PB per server in future
- Highest throughput, lowest-latency
- Avoids network bottlenecks

Challenges with Using

- Siloed: Lack of shared access
- Not protected
- Manual data management
- Operational complexity

Hammerspace Tier 0

Turn GPU Server-Local NVMe Into High-Performance Shared Storage

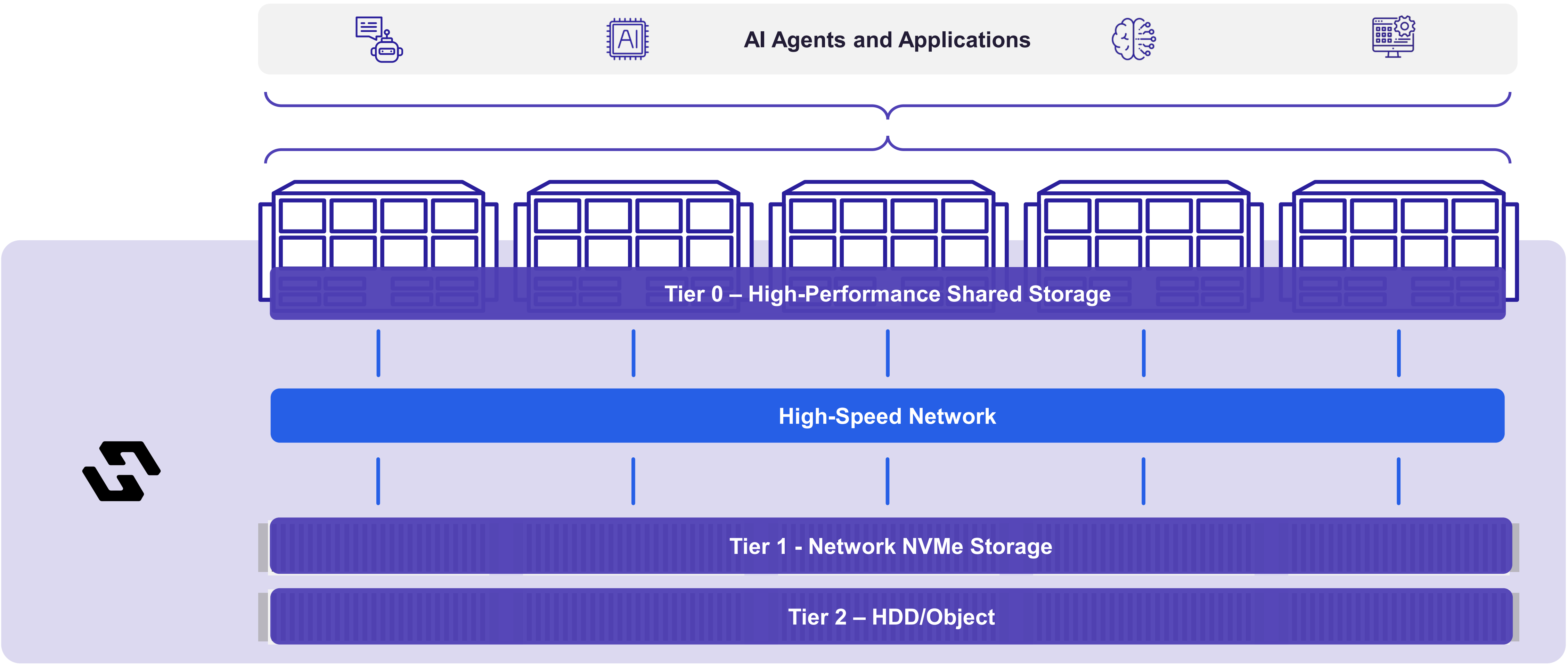

AI Agents and Applications
AI
Tier 0 – High-Performance Shared Storage
High-Speed Network

Activate in Hours, No New Storage

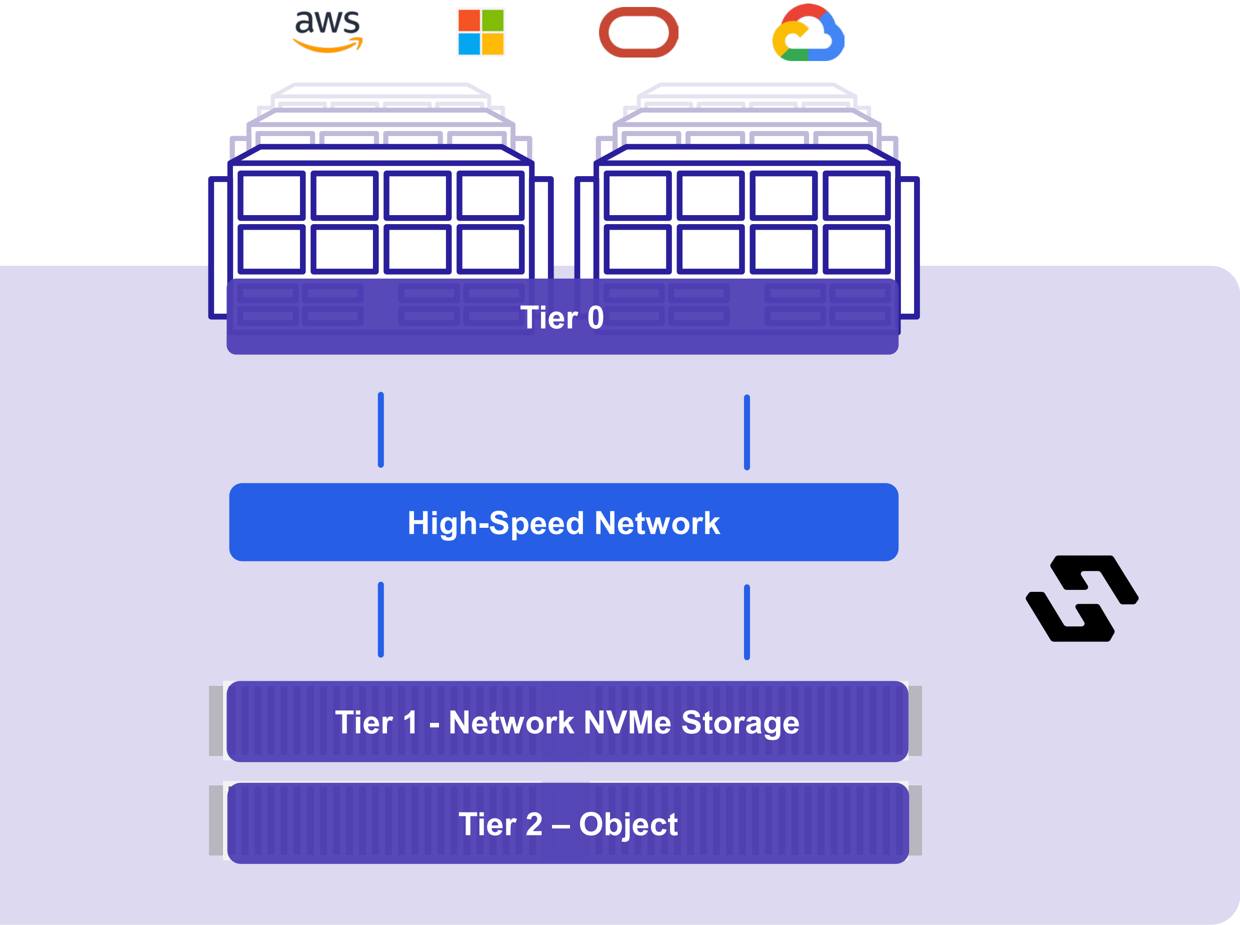
Feed On-Prem and Cloud GPUs up to 10x Faster

Reduce Costs, Power, and Rackspace

Expand to Multiple Tiers in a Single Global Namespace


AI Agents and Applications
Tier 0 – High-Performance Shared Storage
High-Speed Network
Tier 1 - Network NVMe Storage
Tier 2 – HDD/Object

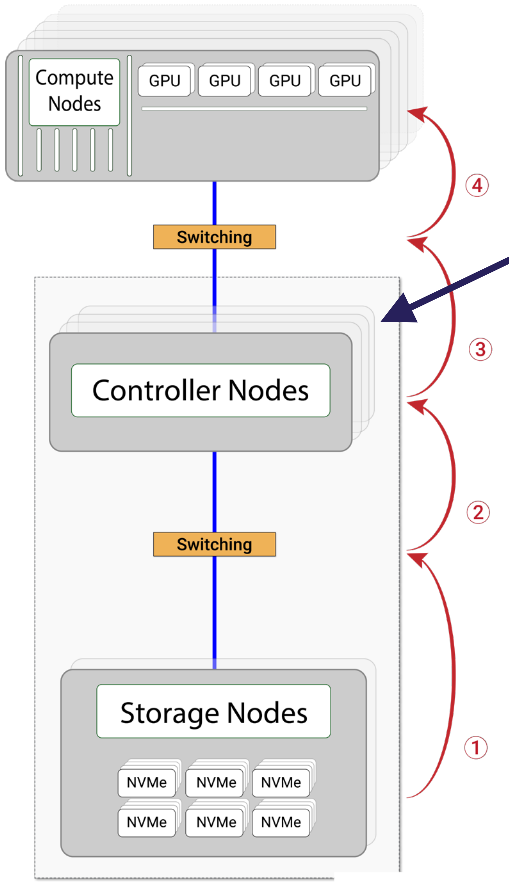
Hammerspace Tier 0 in the Cloud



- Use NVMe drives included in many GPU instance types
- Use fast East-West inter-node networks
- 2x to 10x faster performance
- 50% less than Managed Lustre
- Multi-protocol access, enterprise data services, global namespace

Standards-base, High-Efficiency Storage for AI

Closed Scale-Out NAS Architecture



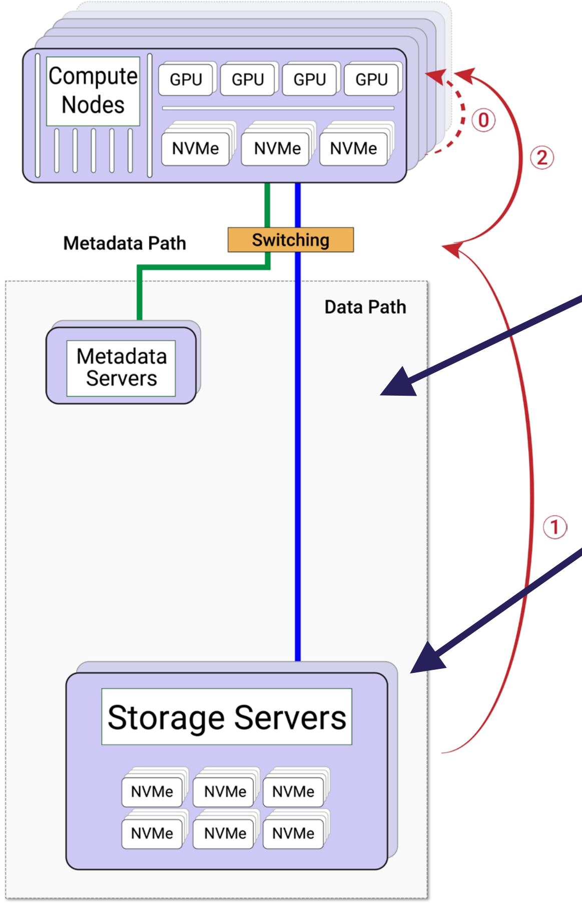
Closed Proprietary Systems

- Data funnels through Controller Nodes.
- Puts data within a proprietary silo
  - Adds latency & degrades performance
  - Limits scalability
  - Adds power, cost and complexity.

Qumulo DELL VAST
NetApp HITACHI

HAMMERSPACE

Open Architecture



OPEN FLASH PLATFORM

pNFS v4.2

- **Separates Data Path from Control / Metadata**
- **Standard Linux – No proprietary client software or storage servers needed in the data path**
- **Ensures Extreme Linear Scalability**

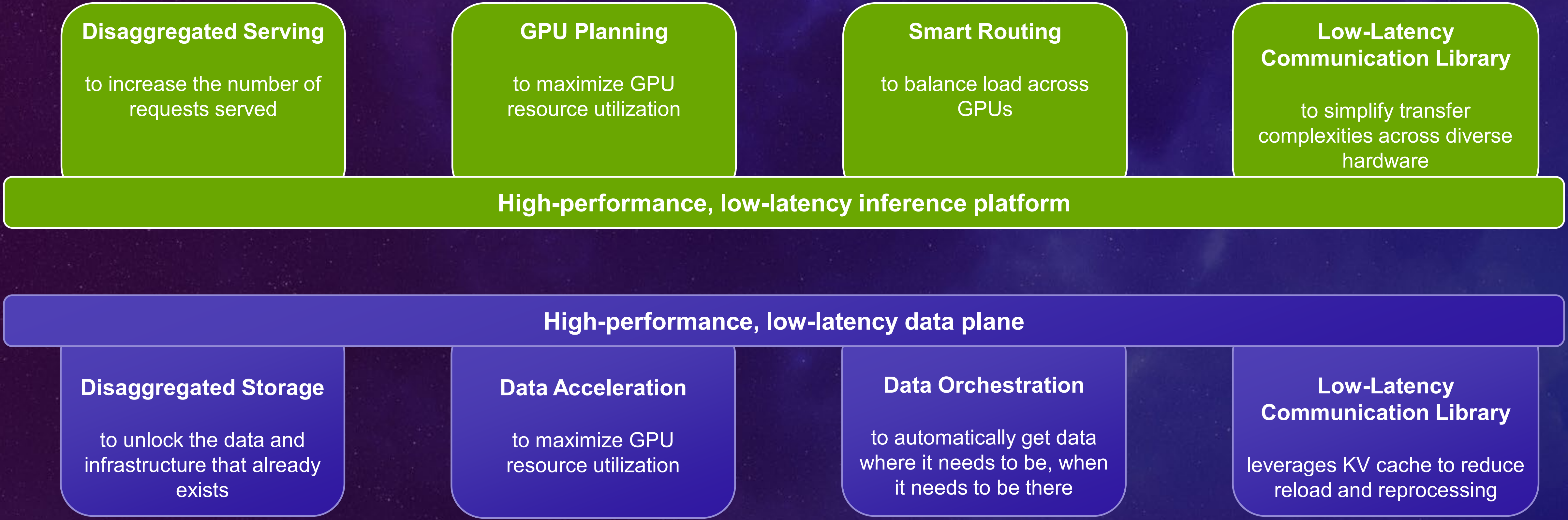
OFP

>10x Capacity Increase

60% Cost/Capacity Reduction

90% Power Savings

Open, Unified Data for Accelerated Inference & Training



Disaggregated Serving

to increase the number of requests served

GPU Planning

to maximize GPU resource utilization

Smart Routing

to balance load across GPUs

Low-Latency Communication Library

to simplify transfer complexities across diverse hardware

High-performance, low-latency inference platform

High-performance, low-latency data plane

Disaggregated Storage

to unlock the data and infrastructure that already exists

Data Acceleration

to maximize GPU resource utilization

Data Orchestration

to automatically get data where it needs to be, when it needs to be there

Low-Latency Communication Library

leverages KV cache to reduce reload and reprocessing

Thank You

david.flynn@hammerspace.com