Orchestrating the Future

Seamless Workload & Data Management in Distributed HPC

David Flynn
CEO & Founder, Hammerspace

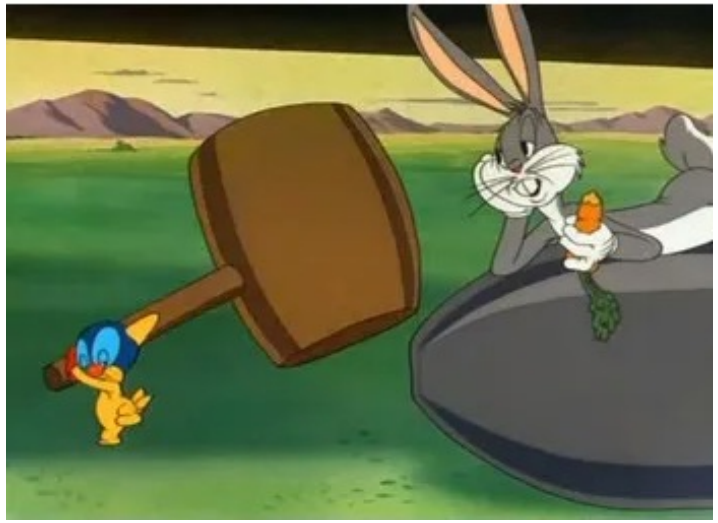
Matthew Shaxted
CEO & Founder, Parallel Works

HPC | QUANTUM | DATA | AI

“Hammerspace is Like Magic”

Principal Engineer, Meta

Hammerspace is an extra-dimensional space that is instantly accessible and infinite in size.



The Dawn of a New Era

Evolution of Data Permanence and Access

Evolution of Data Permanence and Access

Evolution of Data Permanence and Access

Evolution of Data Permanence and Access

Evolution of Data Permanence and Access

android

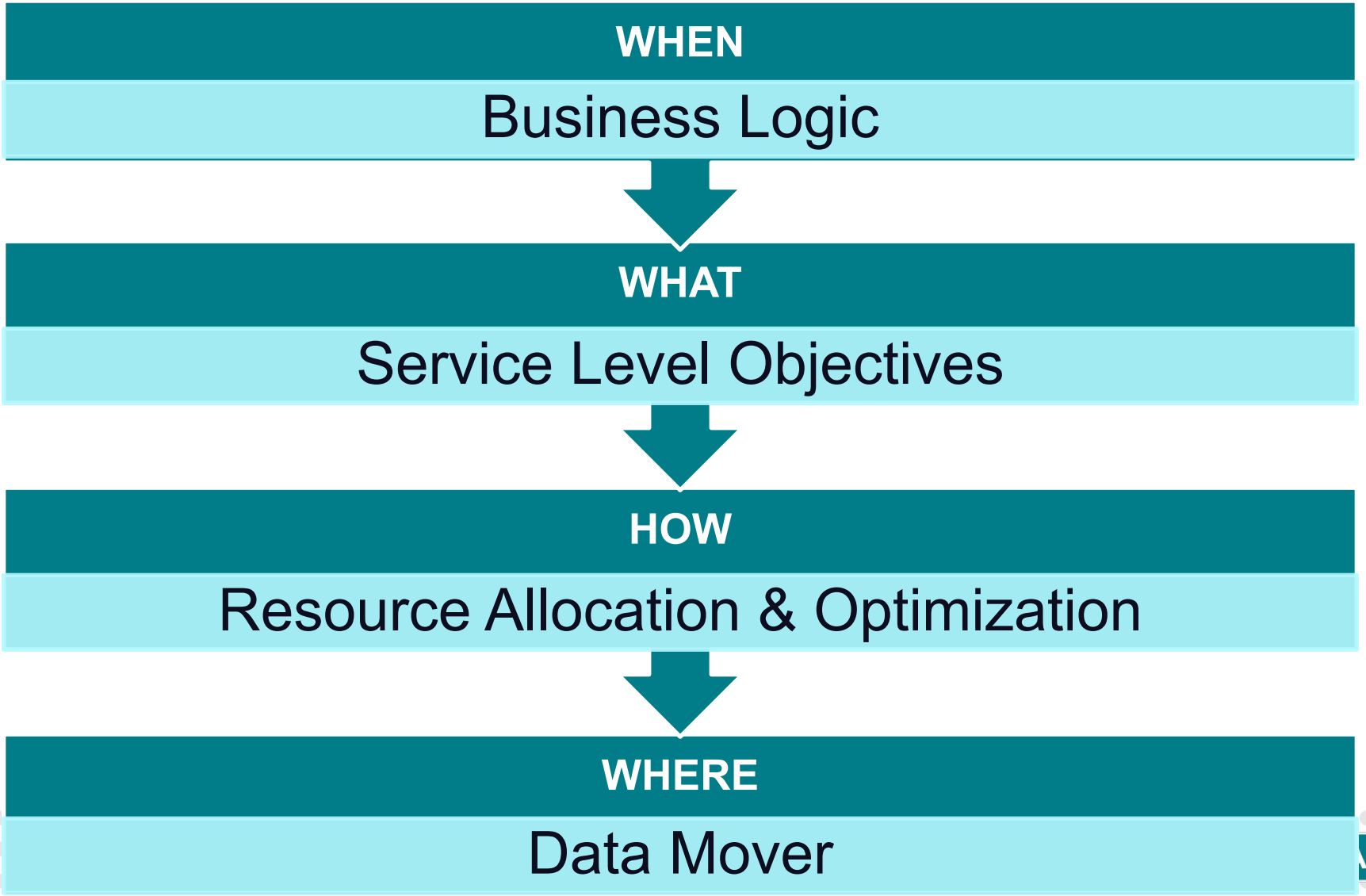
iOS

Enterprise Data Architectures Are Evolving Fast to Meet Demands of AI

FROM:
Data at Rest
Siloed in Storage

TO:
Data in Motion
to Compute or Storage

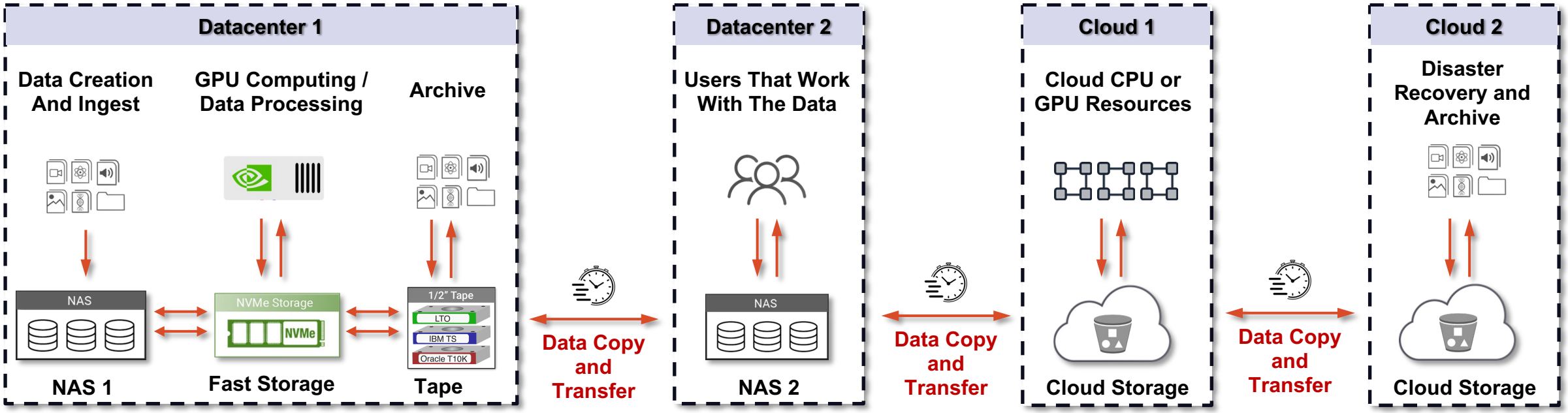
Business Objectives Dictate Data Placement



Data Silos Inhibit Productivity

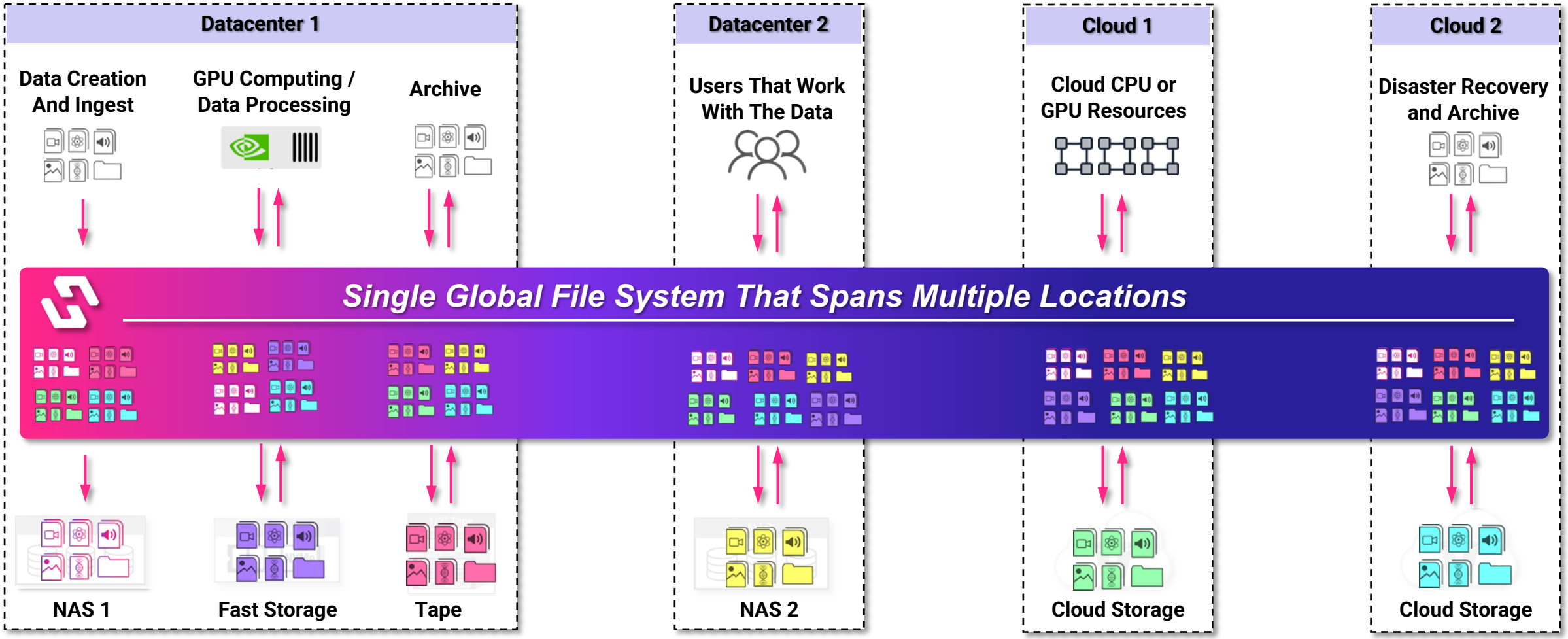
HPC + AI WALL STREET

- Valuable data trapped in silos
- Getting data to global users takes too long
- Infrastructure is not ready for AI
- Data copy sprawl impacts governance and security
- Lack performance to keep GPUs utilized
- Lack agility to use elastic cloud resources



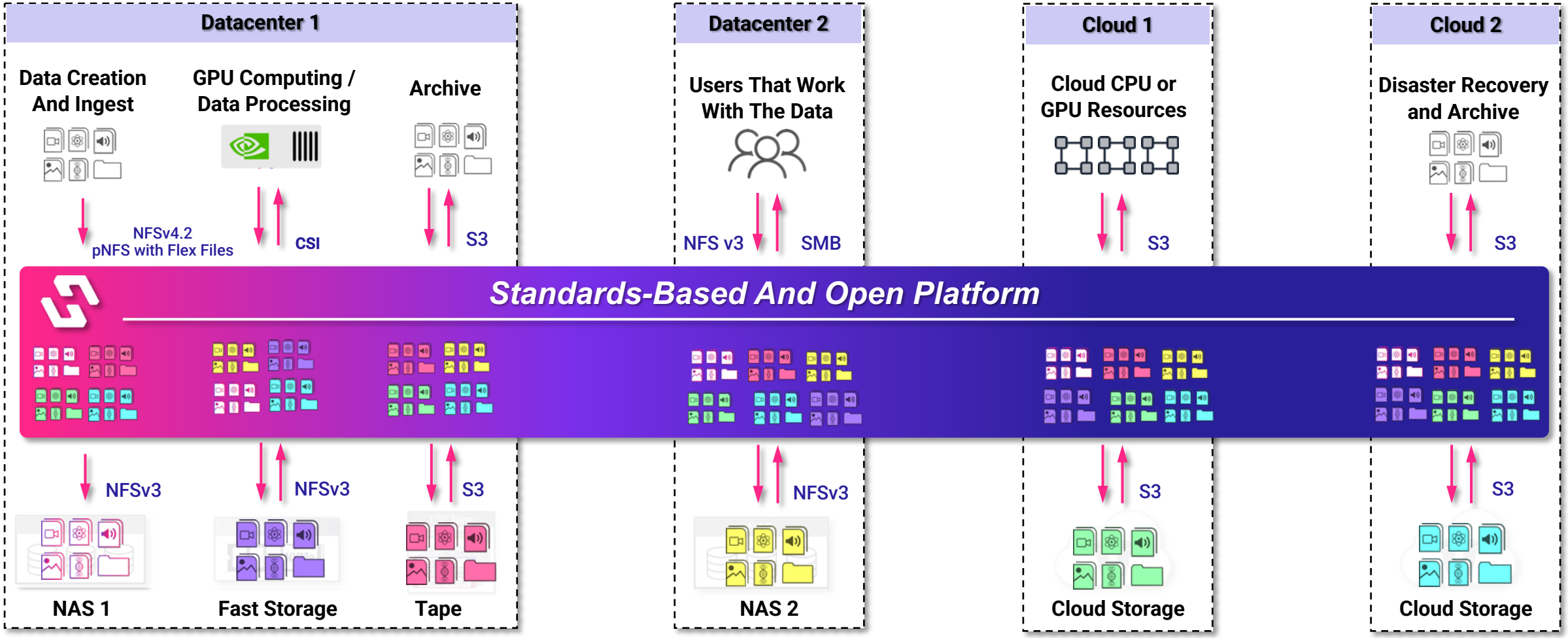
Hammerspace

Breaks Down Data Silos



Hammerspace

Eliminates Infrastructure Lock-in

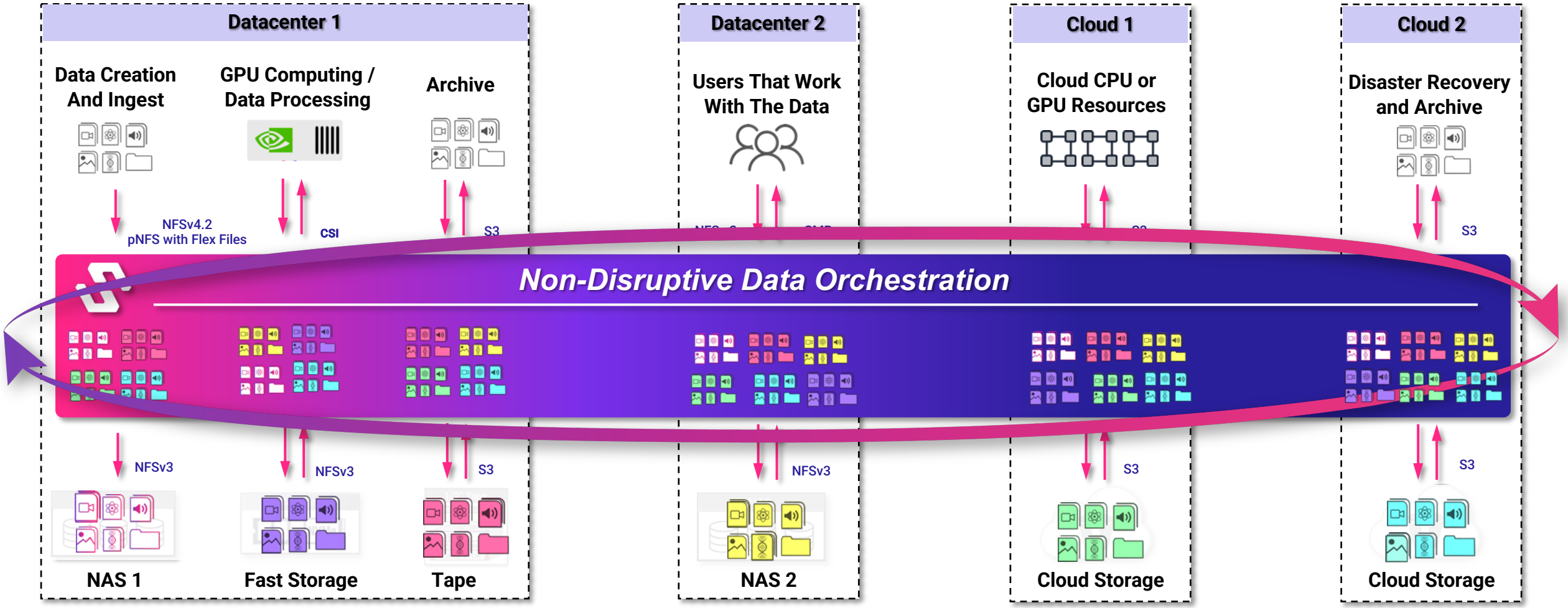


HPC + AI WALL STREET

HPC QUANTUM DATA AI

Hammerspace

Orchestrates Dynamic Data Movement

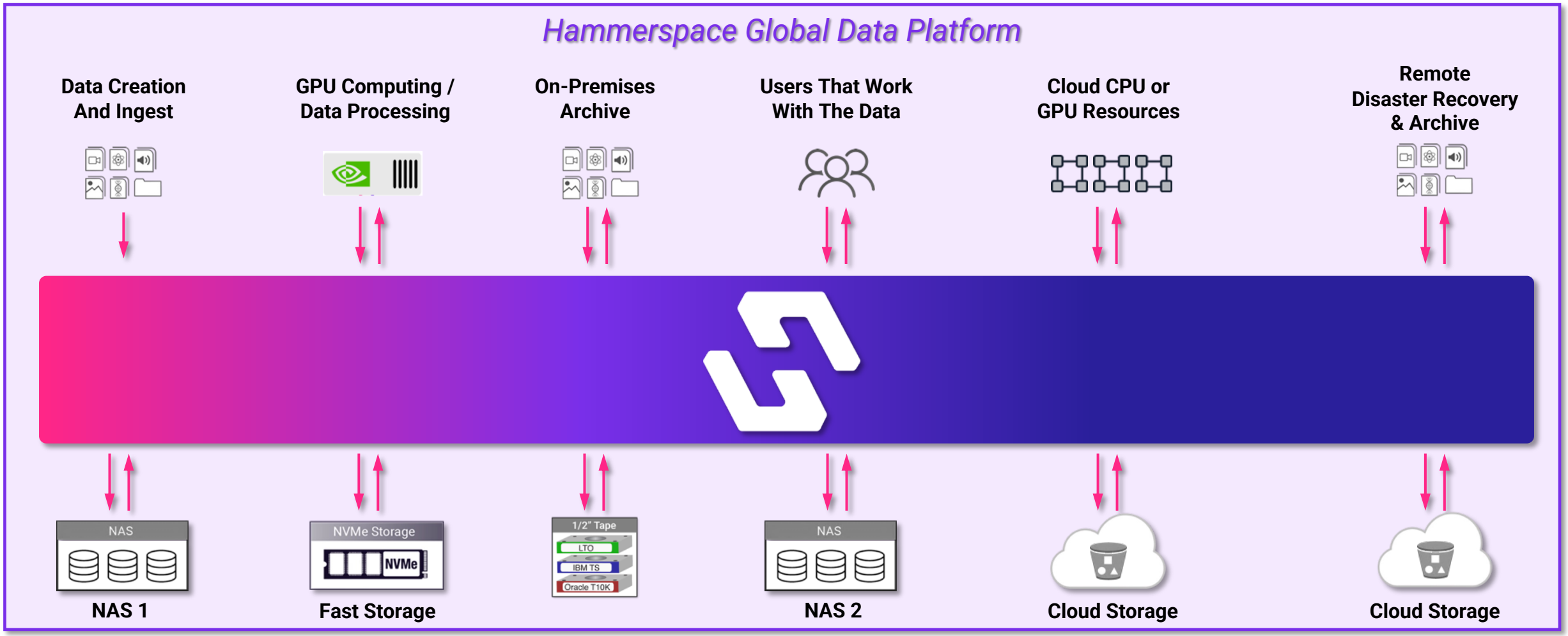


HPC + AI WALL STREET

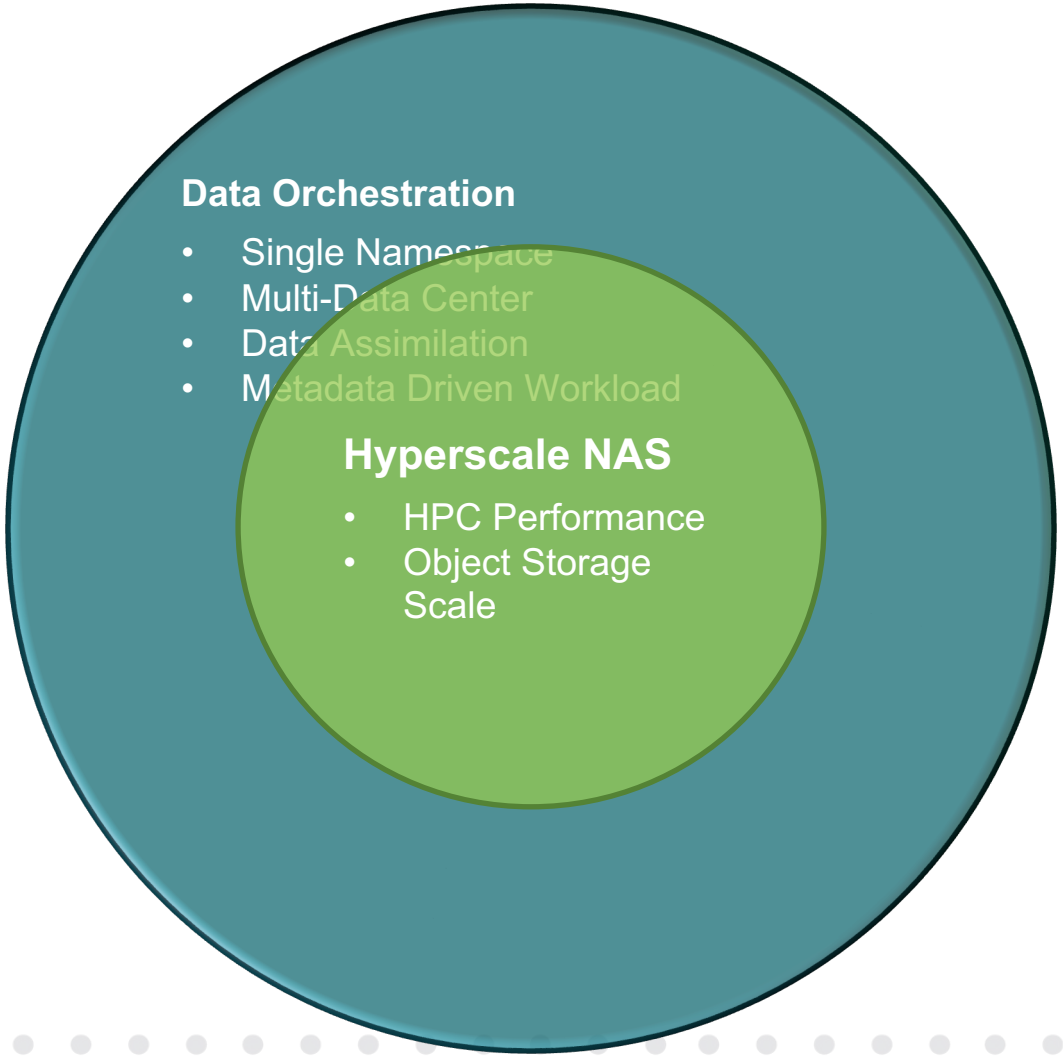
HPC QUANTUM DATA AI

Hammerspace

Global Data Platform for File and Object Data



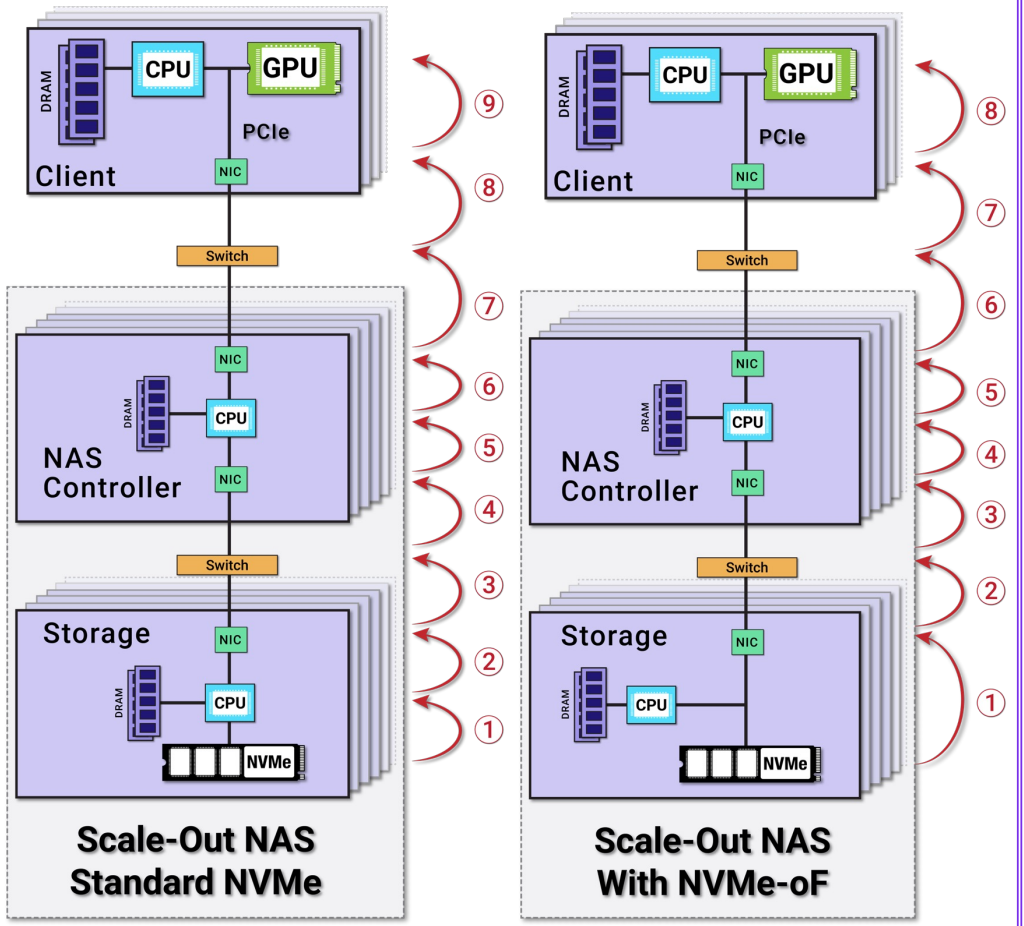
Hammerspace Superpowers



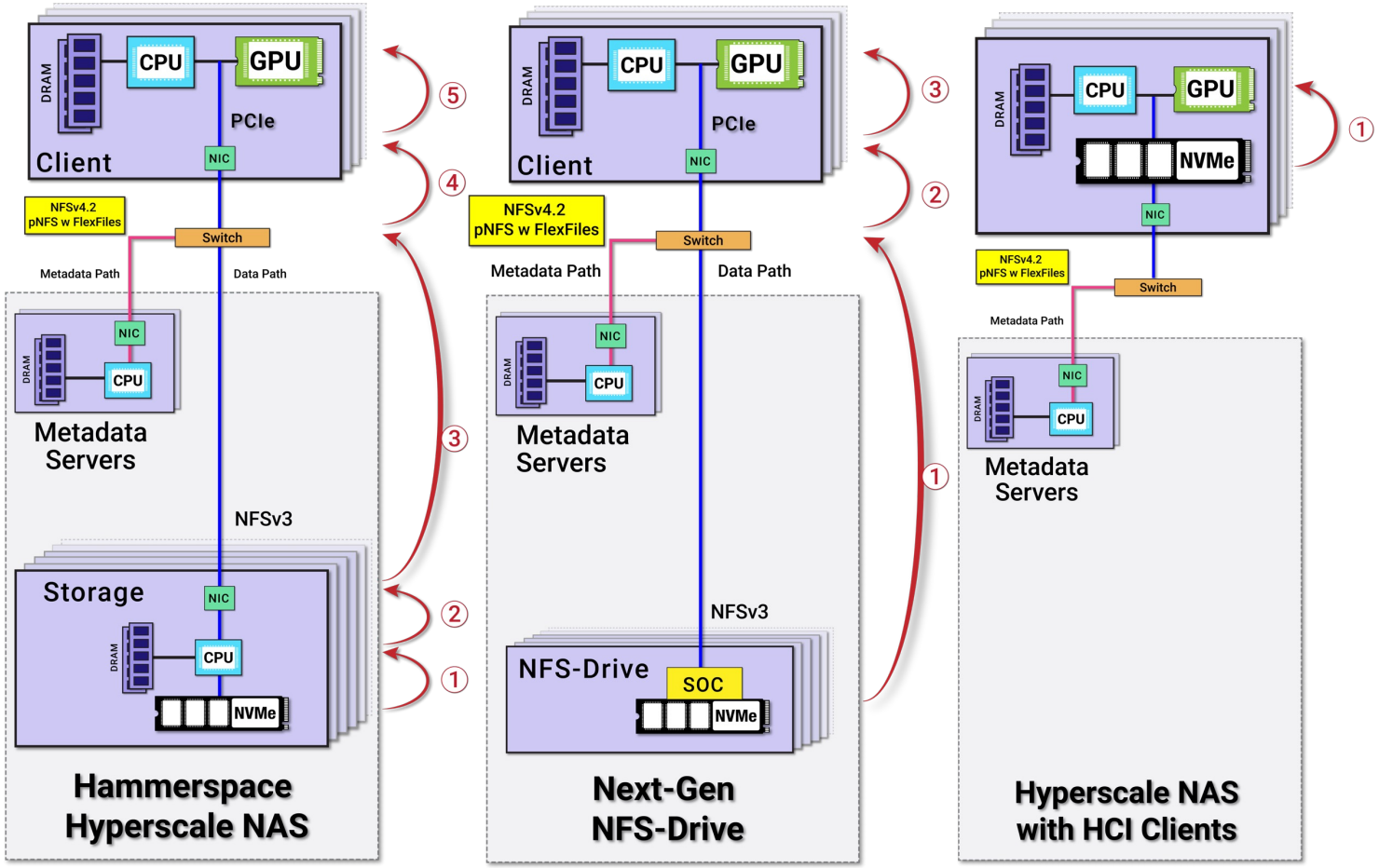
Future Advancements Will Continue to Drive File and Object Storage Infrastructure Efficiency and Performance

HPC + AI WALL STREET

Traditional NAS Architectures



Hyperscale NAS Architectures



2x
Improvement in Latency, BW and IOPS per $, Watt and Rack U

5x
Improvement in Latency, BW and IOPS per $, Watt and Rack U

10x
Improvement in Latency, BW and IOPS per $, Watt and Rack U

Parallel Works

Contact

Matthew Shaxted, CEO
shaxted@parallelworks.com

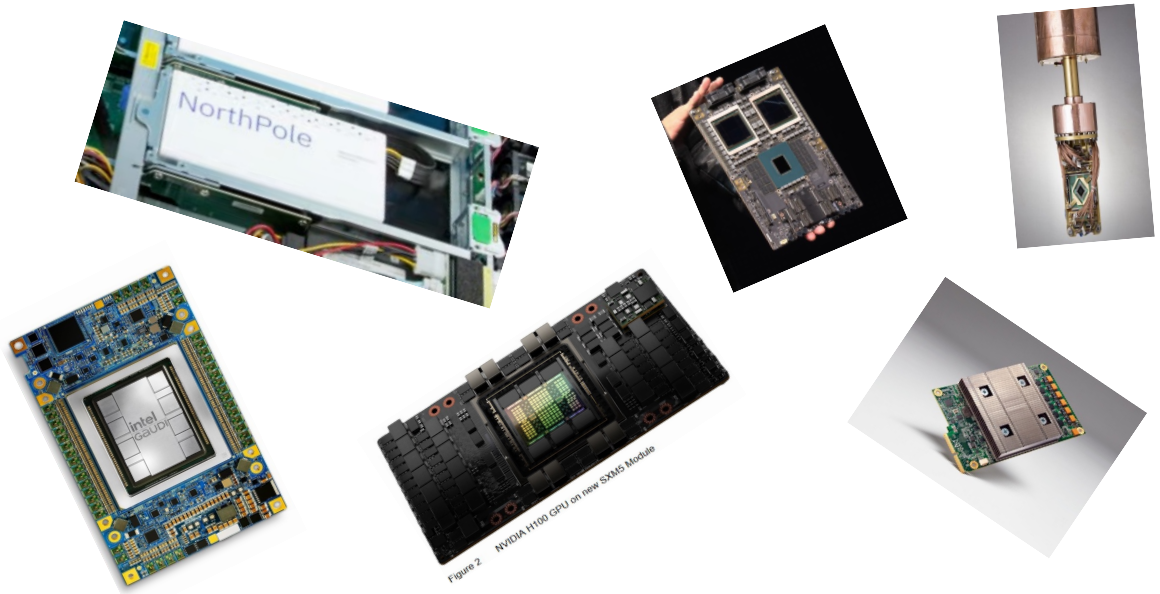
Challenges for the Modern Computing Organization

Technology diversity & capacity constraints necessitate agile task placement

HPC + AI WALL STREET

Rapidly evolving technology (cpu, gpu, fpga, neural, quantum) requires agility to get first mover advantage.

When on-premises resources are constrained or in maintenance, seamless **access to diverse computing sites is essential**.



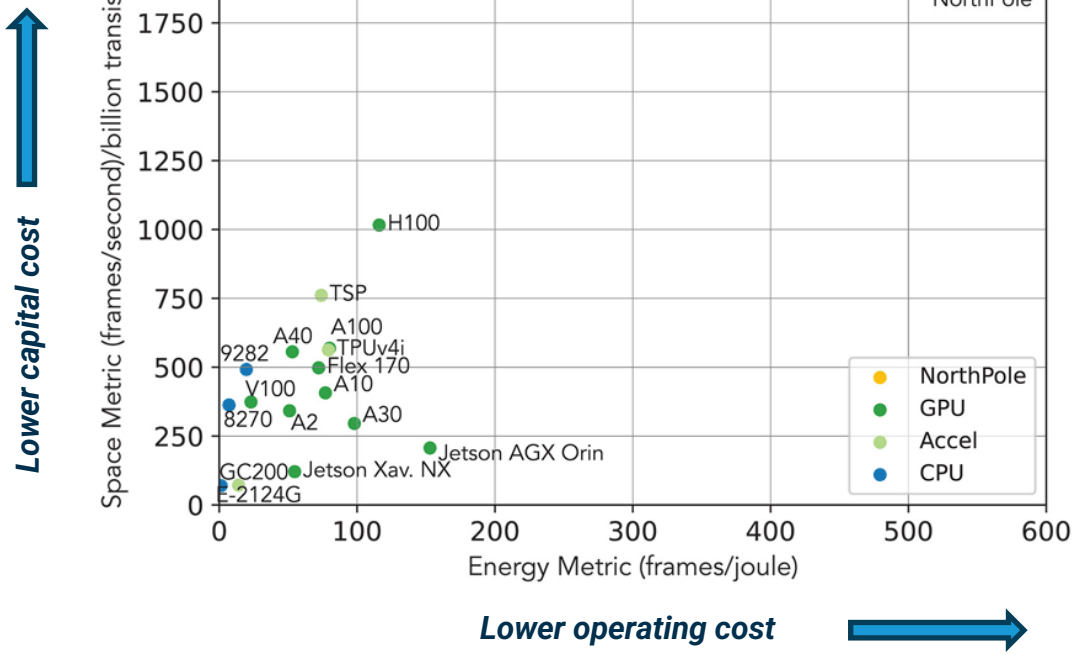


Chart from Carlos Ortega-Otero IBM NorthPole:
https://www.youtube.com/watch?v=7s1M09z_qI8

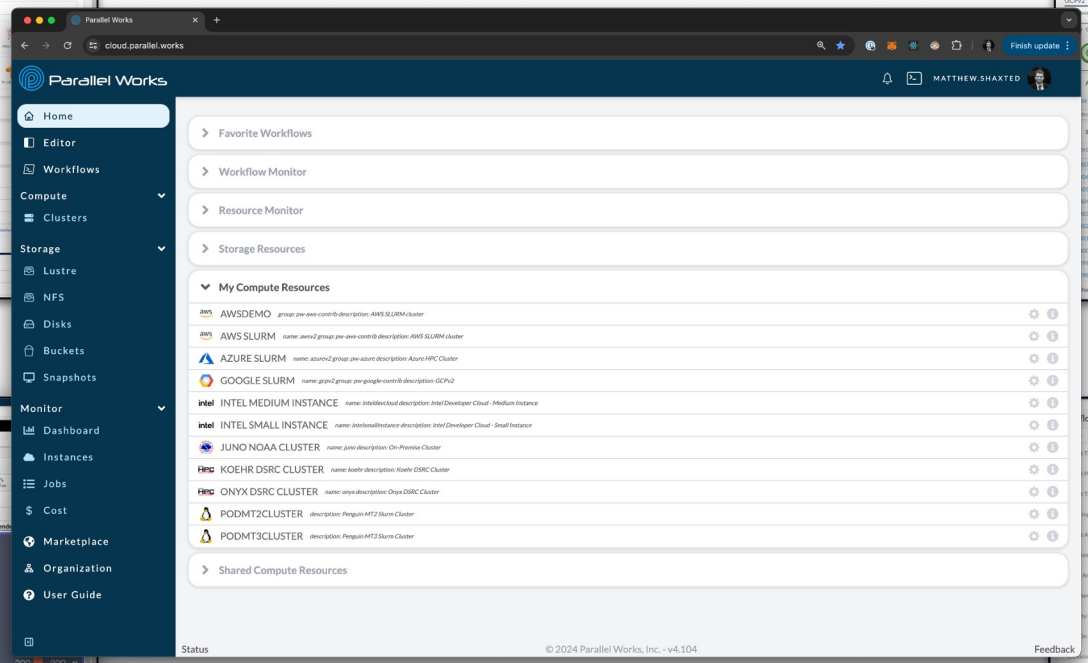
HPC QUANTUM DATA AI

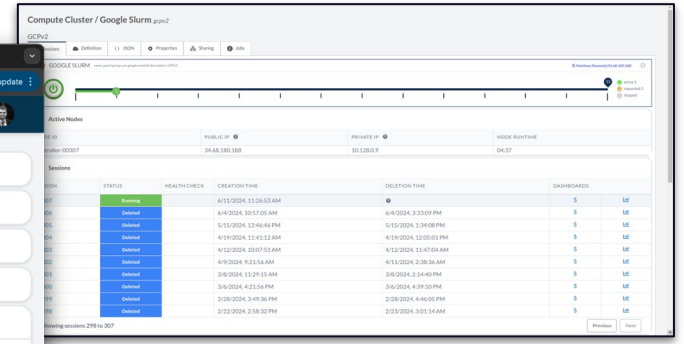
Parallel Works ACTIVATE

HPC + AI WALL STREET

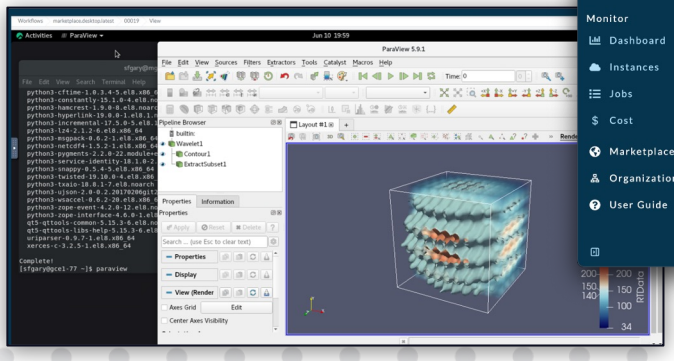
HPC & AI control plane for R&D teams who rely on compute-intensive applications and AI models. Easily and consistently provision, manage, and share on-premises, cloud and hybrid high-performance compute resources at scale for simulation, analytics and AI.

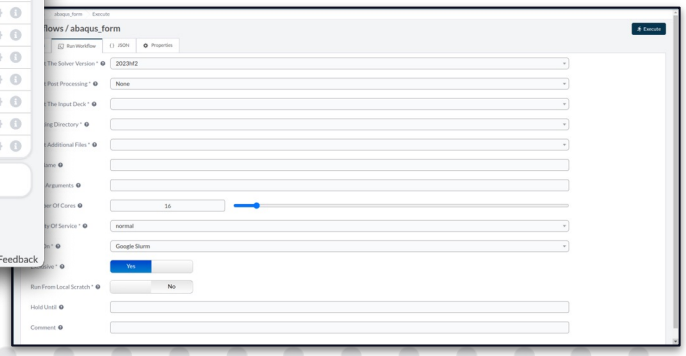
Access





Workflow

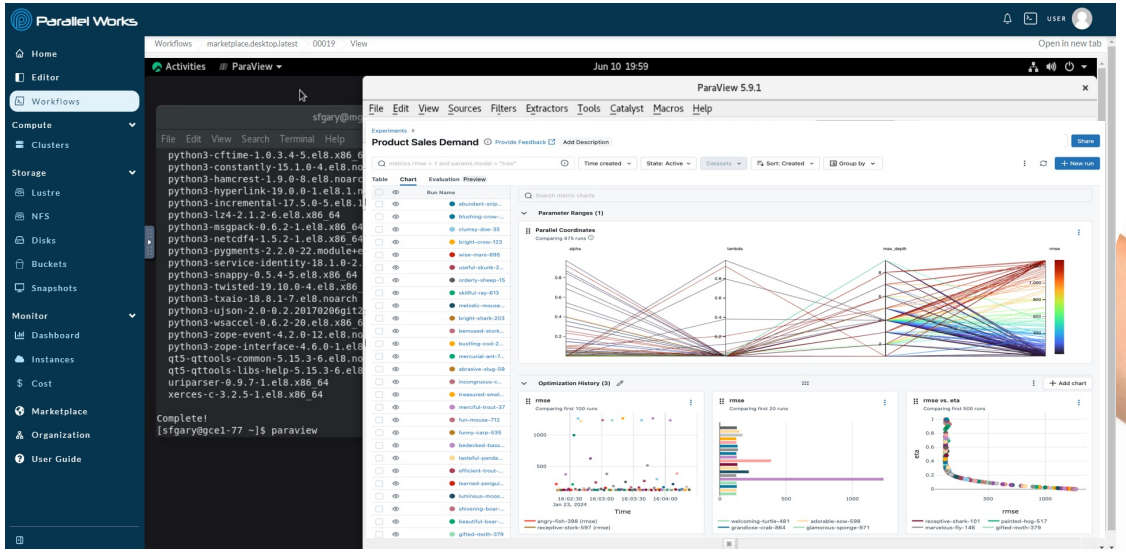


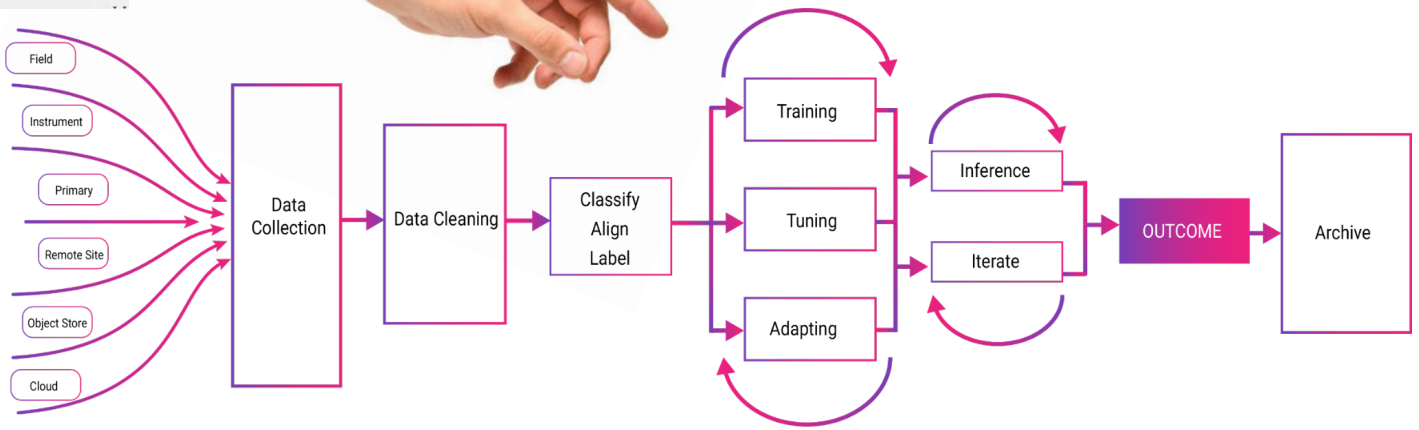


HPC QUANTUM DATA AI

R&D, Task Steering & Control to Optimal (or Available) Environments

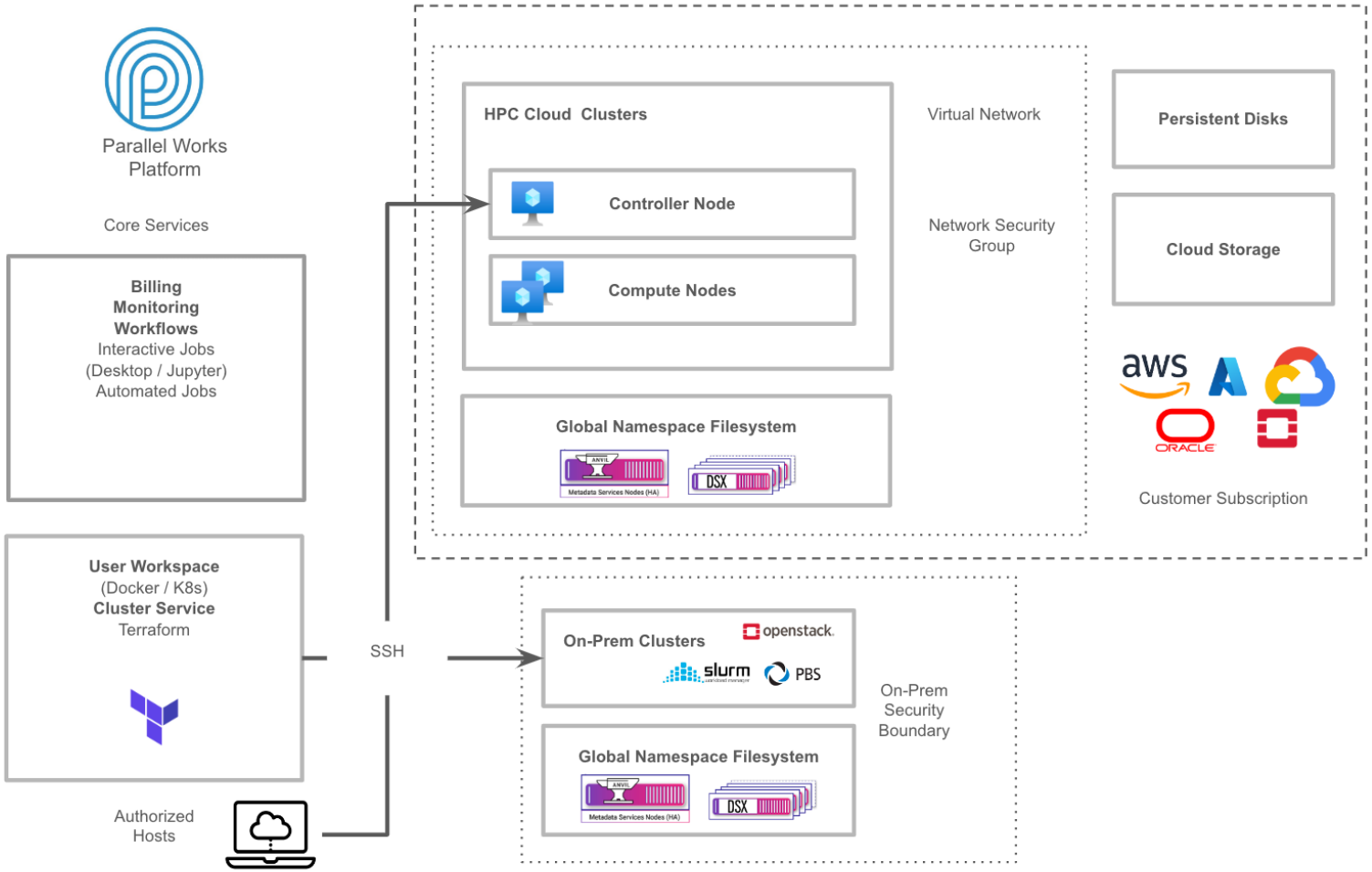
HPC + AI WALL STREET



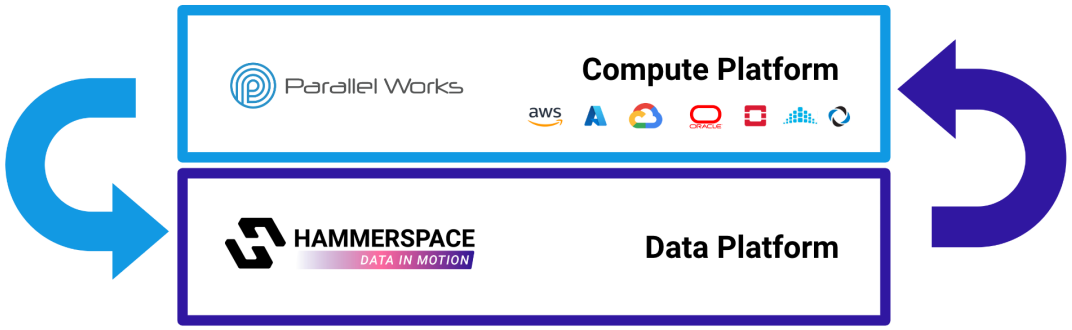


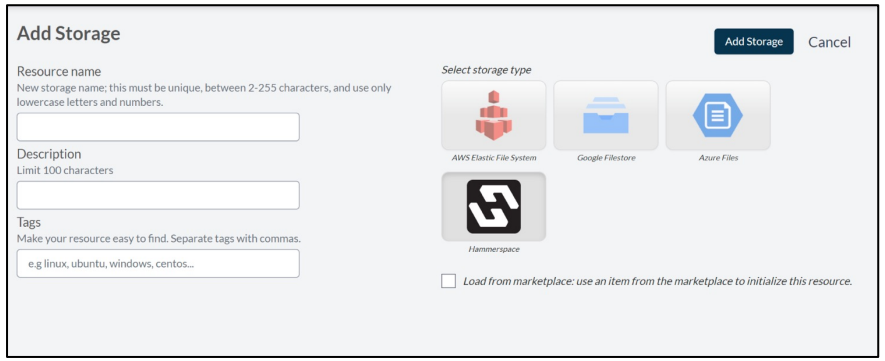
HPC QUANTUM DATA AI

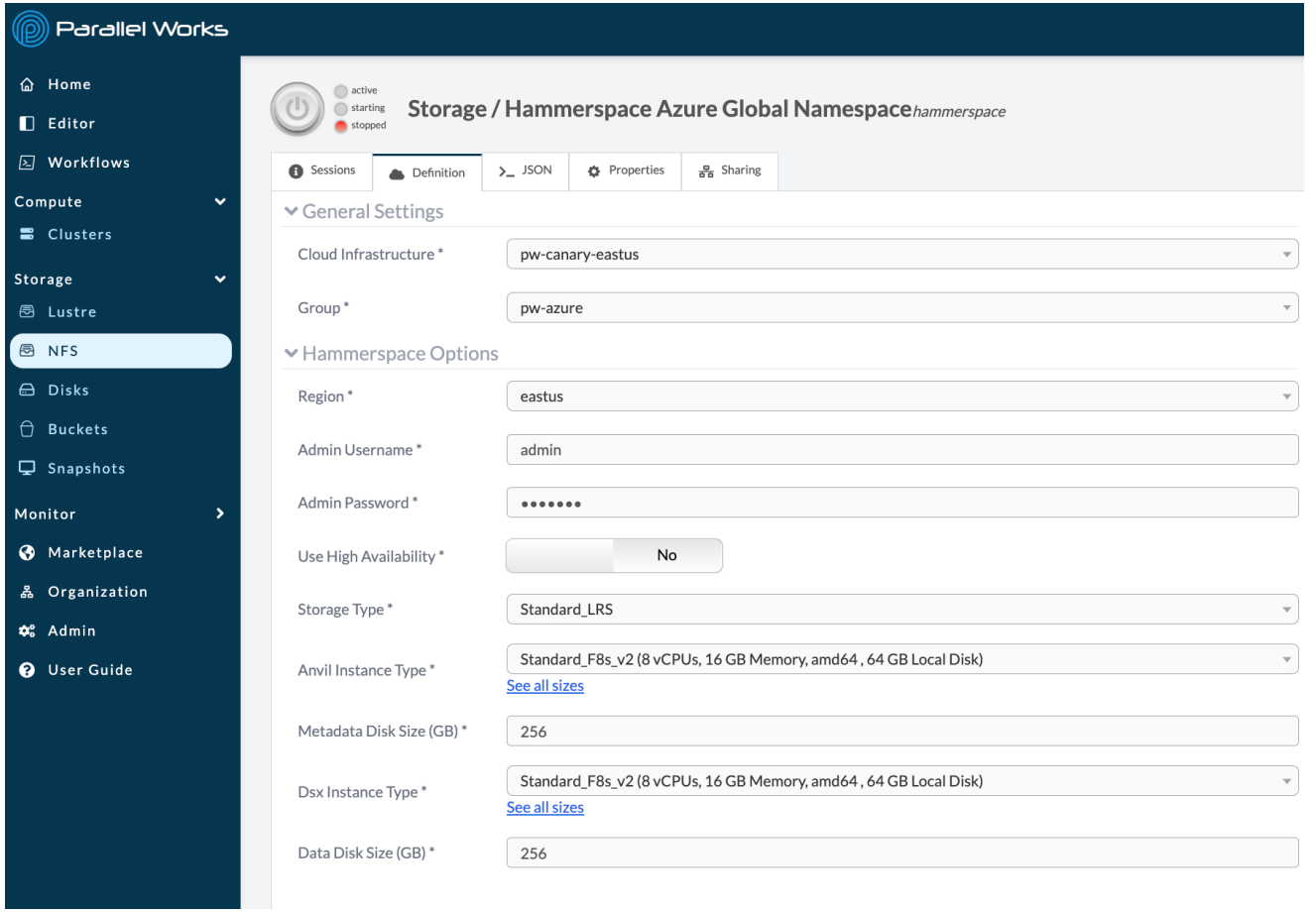
Connection to / Assembly of Uniform Compute Clusters for Application Portability



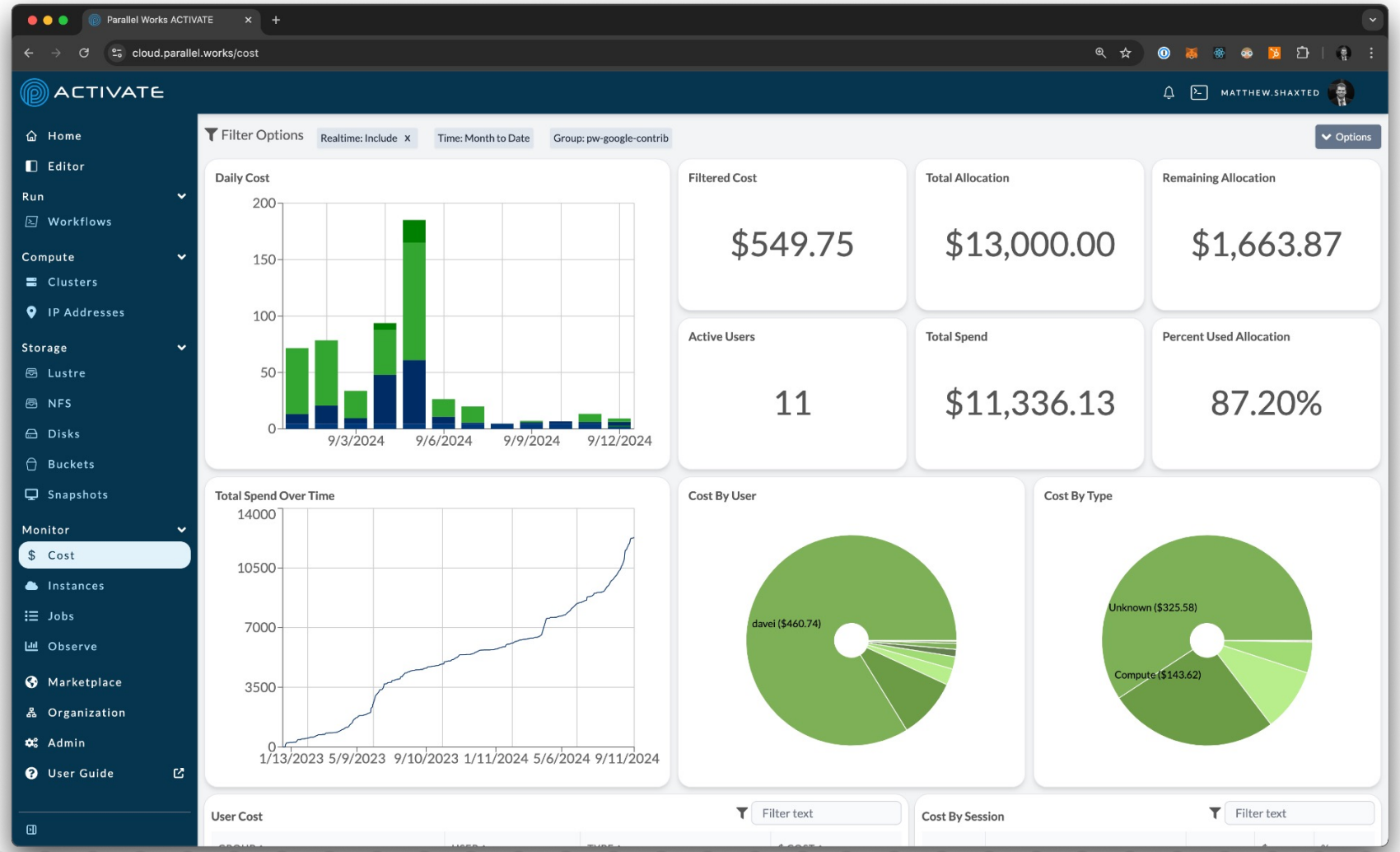
Fit Storage Requirements Precisely to Workloads







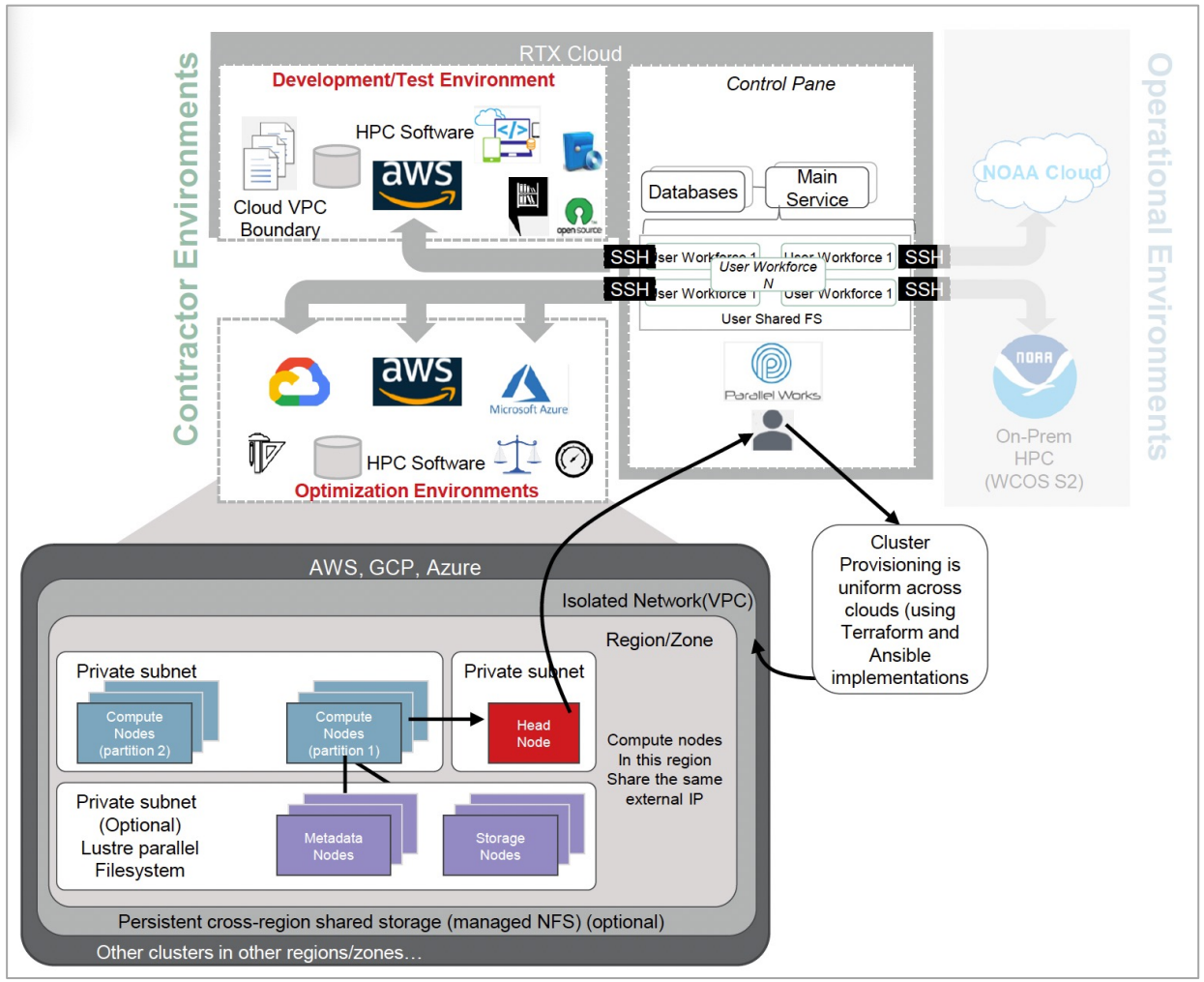
~3 min Cost Control in Cloud Environments: FinOps for HPC



HPC + AI WALL STREET

HPC QUANTUM DATA AI

Streamlining Research to Operations

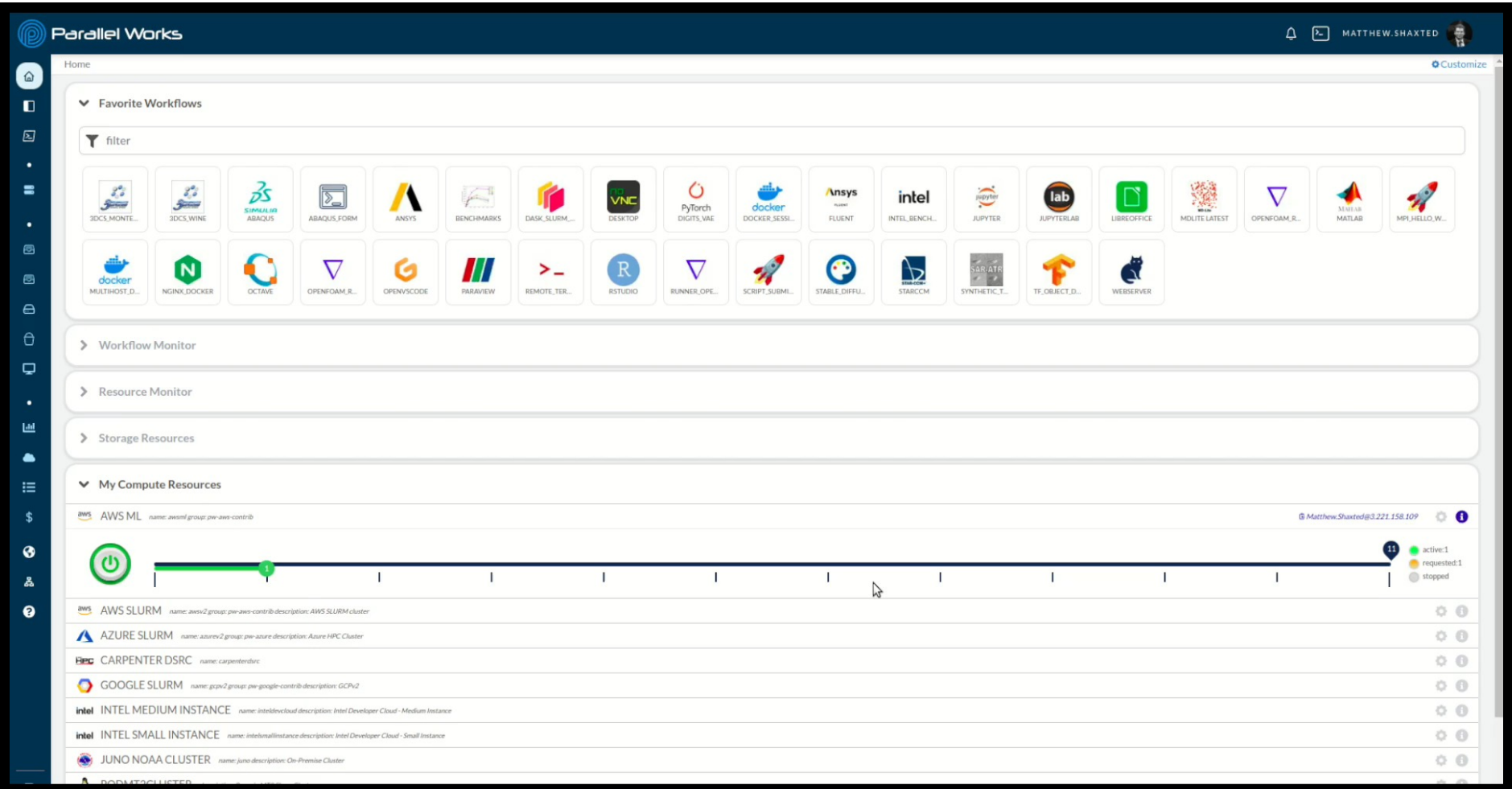


Problem: Customer experienced a high barrier when moving production workloads from local dev environments to operational systems.

Solution: Parallel Works helped customer to spin up a cloud environment that closely mirrored the current on-premise operational environment (WCOSS2) to act as a testbed for operational workloads.

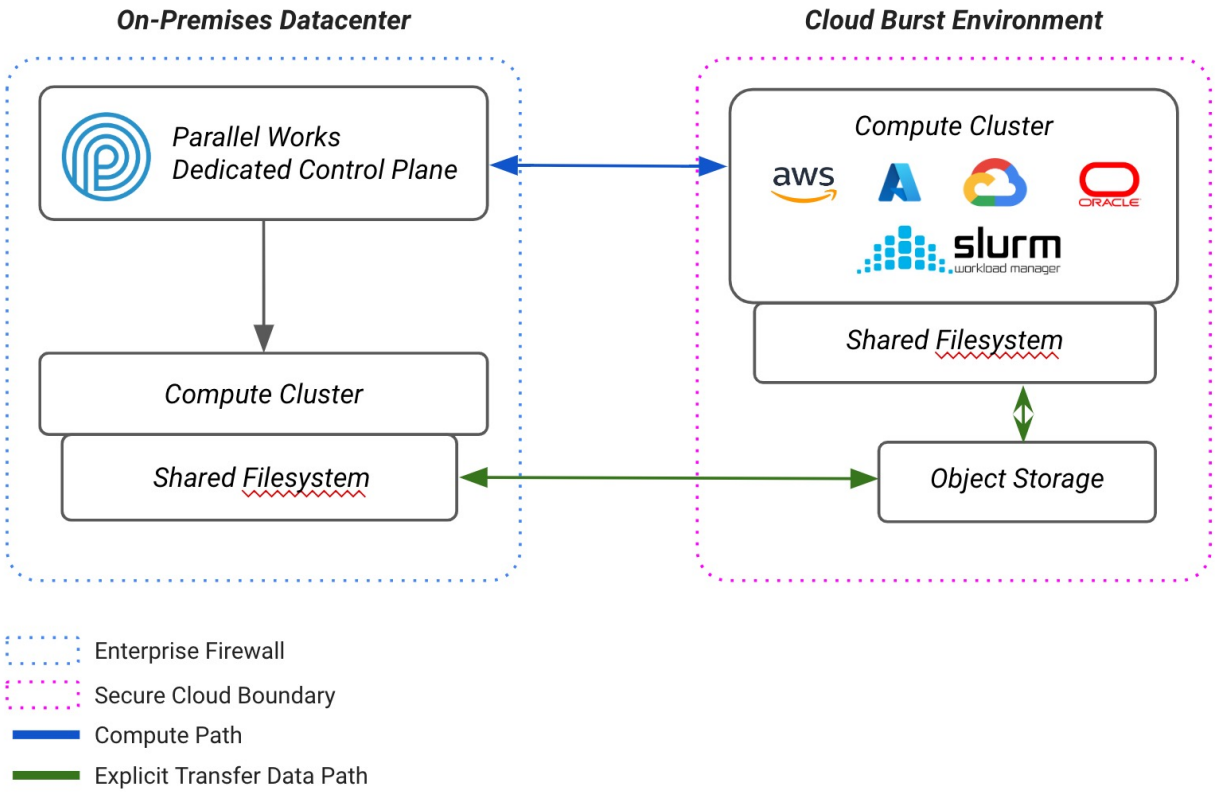
Key Results: The cloned cloud environment empowered practitioners to migrate production workloads to operational systems 3x faster than a local dev method, streamlining research to operations.

Lowering Friction for End Users To Do Their Work



Execute complex ML/GenAI stacks via turnkey webforms on cloud or on-premise GPU resources (e.g., Stable Diffusion).

Business-as-Usual for Burst Computing
Explicit Multi-Hop Transfers to Object Storage

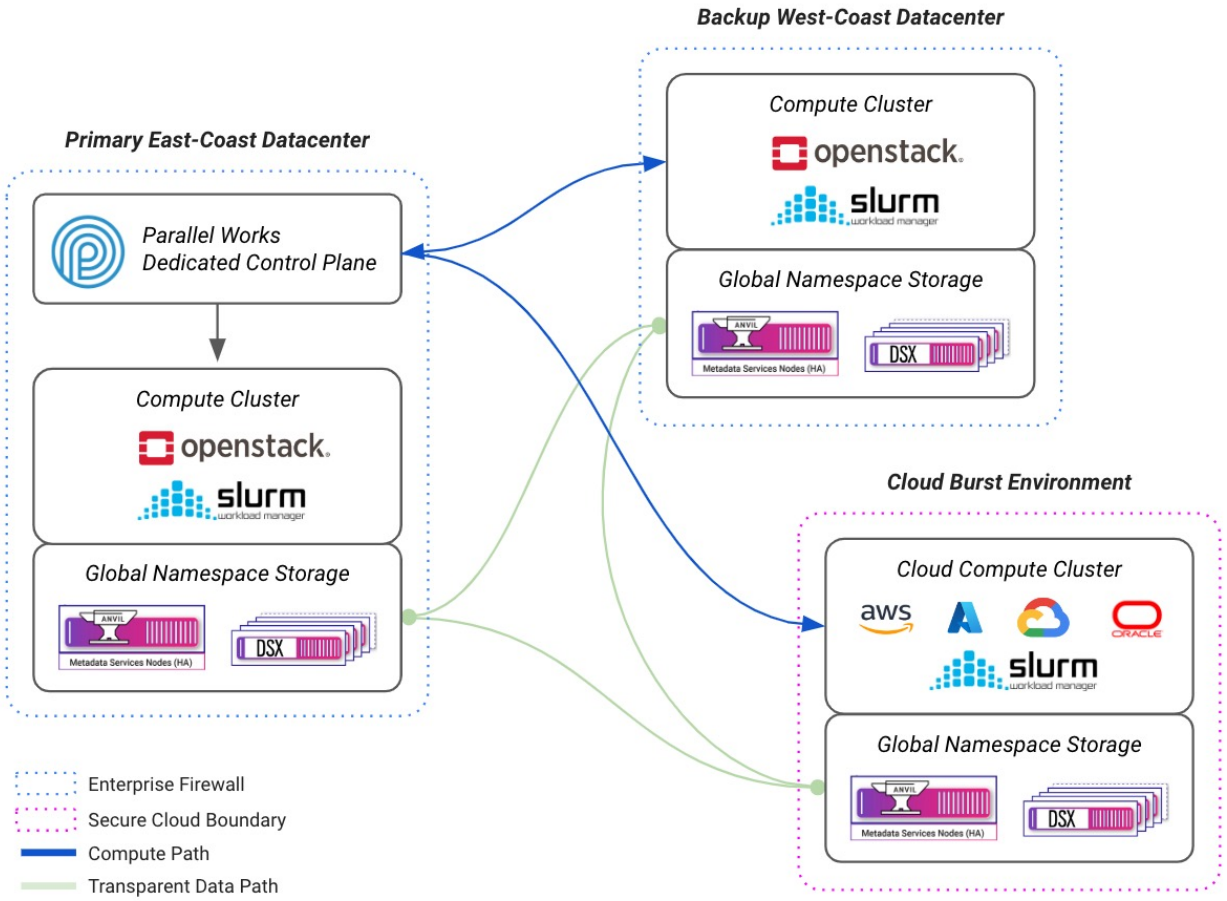


Different: running in cloud takes a different file availability approach as on-prem systems, adding to multi-site friction.

Difficult: requires knowledge and understanding of CSP-provided CLI tools.

Slow: multiple hops required to/from on-prem filesystem, object storage and cloud shared filesystem.

Converged Burst Computing
Transparent Data Availability at Any Site



Familiar: data presented exact same way as on-prem filesystems, lowering cognitive load for end users.

Portable: allows workloads to easily scale between all sites (cloud or on-prem).

Scalable: adding new sites or cloud keeps the same topology without having to learn new data movers.

Right-Sizing of HPC & AI Infrastructure

- Right-sizing is the process of matching workloads to optimal resource shapes to reduce TCO.
- Recent customer study looked at a $30m cloud program budget w/ multiple FS performance tiers in comparison to traditional full on-prem hardware refresh.
- Showed that using a cloud mix could deliver competitive TCO for a portion of workloads, but not a full 1:1 mapping.
- Cloud benefits such as supply chain agility for new large system and tech updates throughout the program lifecycle factor into the “more is better” tradeoff.
- Additionally global namespace filesystem as opposed to Managed Services resulted in >25% additional compute and storage.

Thank you!

david.flynn@hammerspace.com
shaxted@parallelworks.com

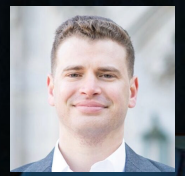
HPC + AI WALL STREET

2024

Orchestrating the Future: Seamless Workload and Data Management in Distributed HPC Environments

Moderator:
Alex Woodie
Managing Editor, Datanami

David Flynn
Co-founder and CEO,
Hammerspace



Matthew Shaxted
CEO,
Parallel Works

HPC | QUANTUM | DATA | AI