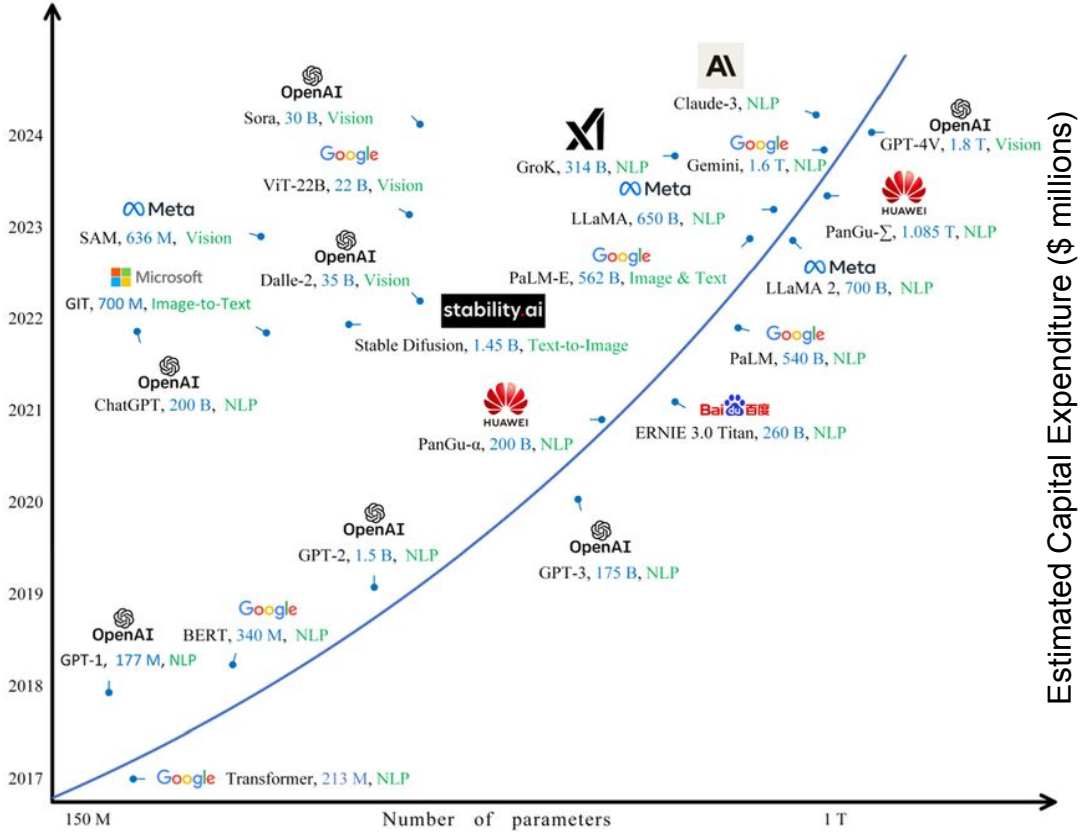
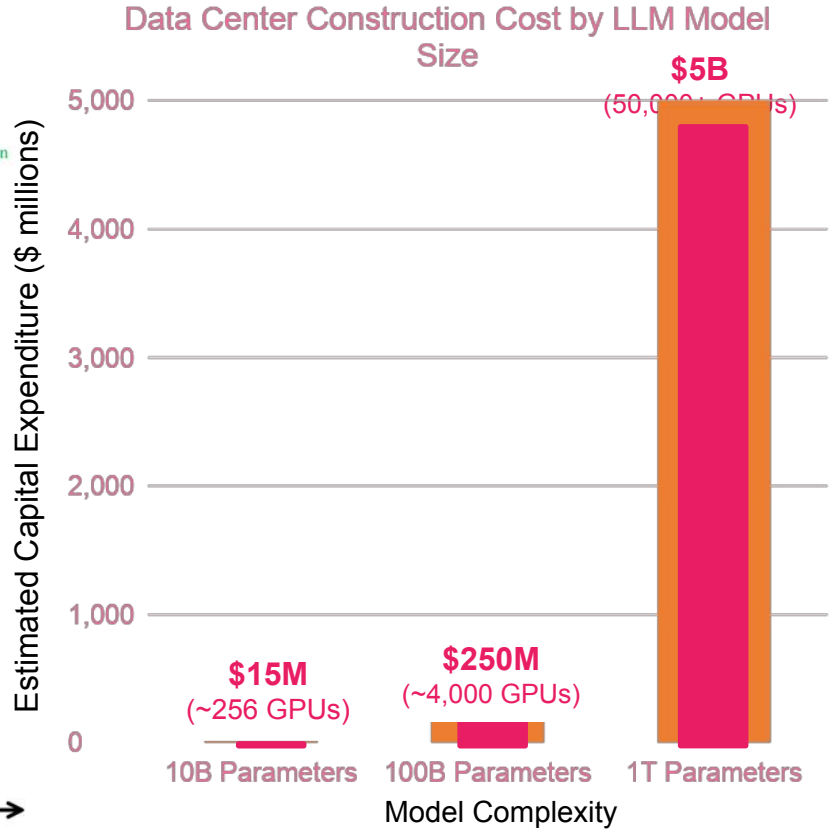
KubeFlash

Revolutionary All Flash Technology For HCI and Kubernetes

Cost of AI data centers grows even faster than LLM mode



Source: https://www.researchgate.net/

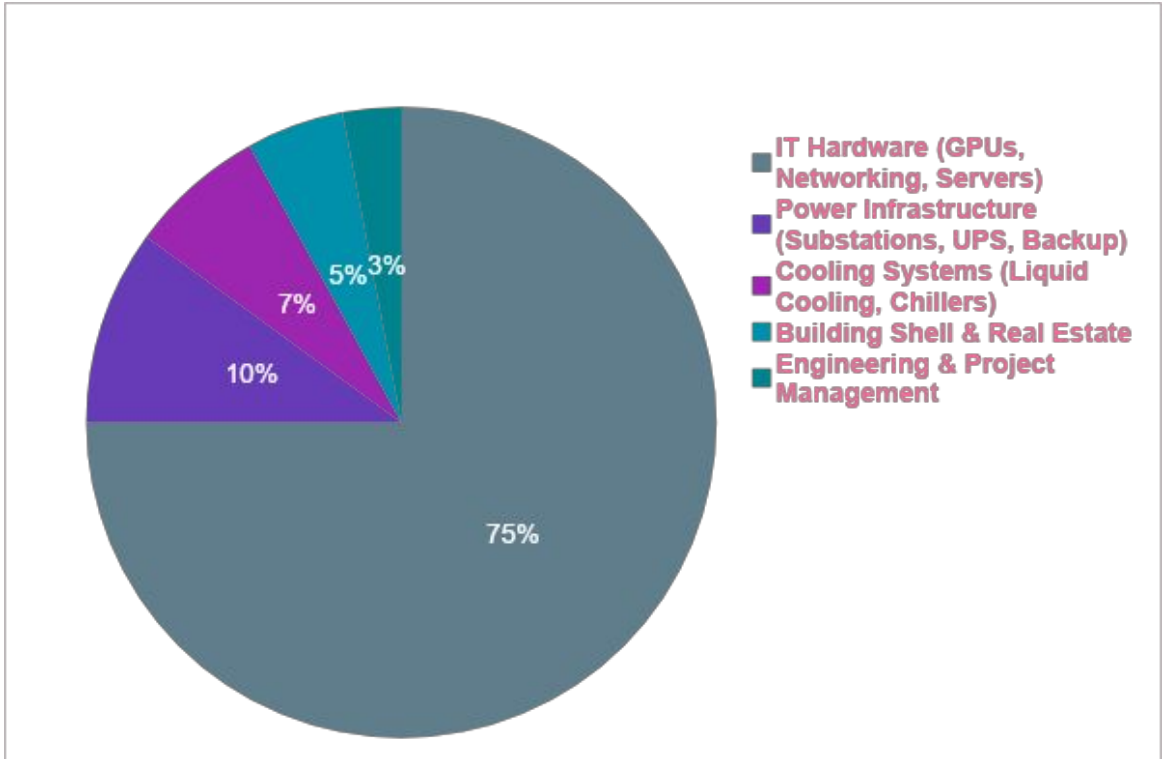


The Need to Democratize AI Data Centers

	Public AI	Private AI
Data Privacy	Shared environment; risk of data leakage.	Air-gapped security; zero external exposure.
IP Protection	IP may inadvertently "train" a competitor's model.	100% ownership of model weights and logic.
Cost Profile	Variable OpEx; high scaling "tax" at volume.	Fixed CapEx; lower TCO for high-utilization.
Customization	General-purpose; limited hardware tuning.	Bespoke architecture for specific domain tasks.
Governance	Subject to vendor Terms of Service (ToS).	Full control over compliance and ethics.

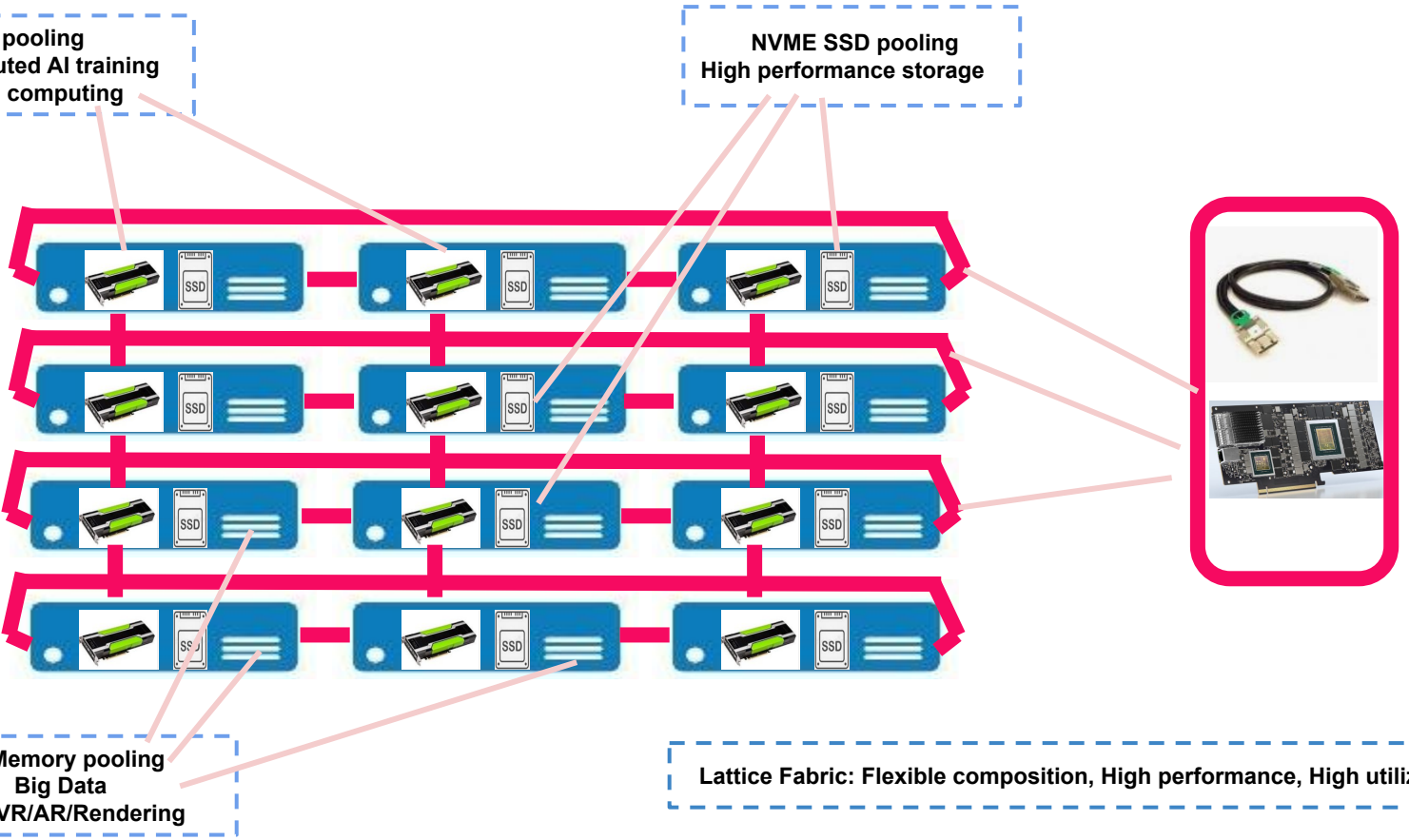
Despite the necessity of private AI, >99.9% of companies **cannot** afford owning proprietary AI data centers, limiting AI's proliferation in enterprise applications.

AI Data Center Cost Breakdown

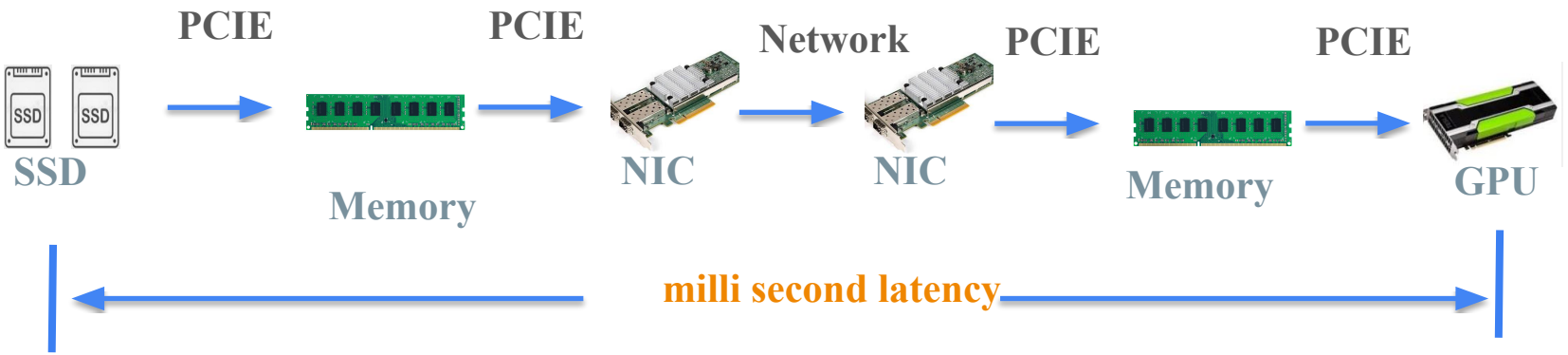


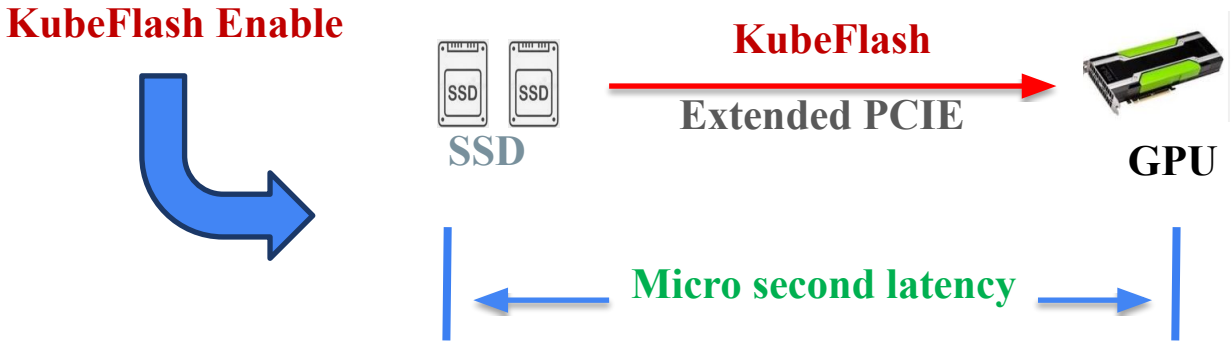
All cost factors are based on **the number of GPUs** provisioned.

KubeFlash extends PCI Express for server interconnect

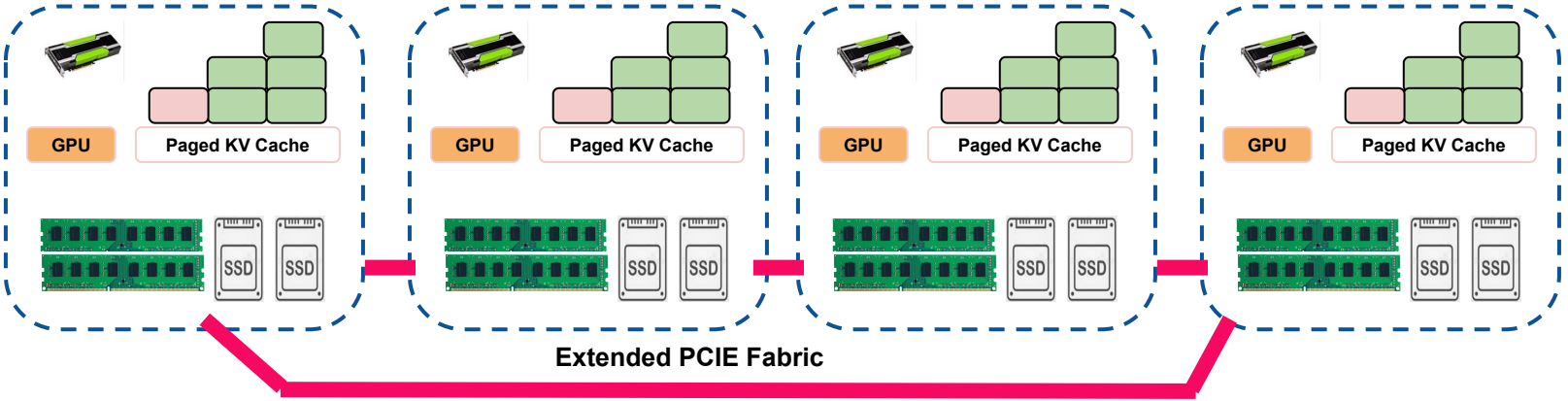


GPU/SSD Data Exchange



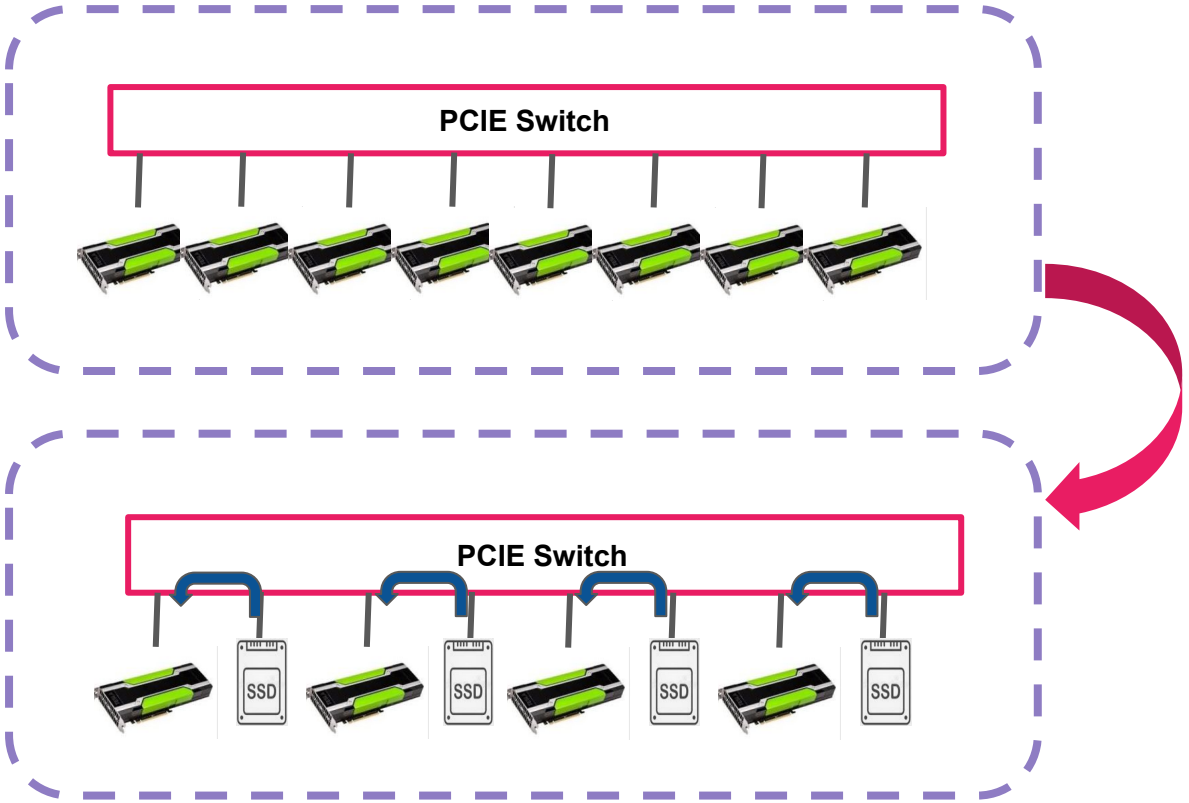


KubeFlash: Tightly coupled GPU/SSD KV Cache for LLM



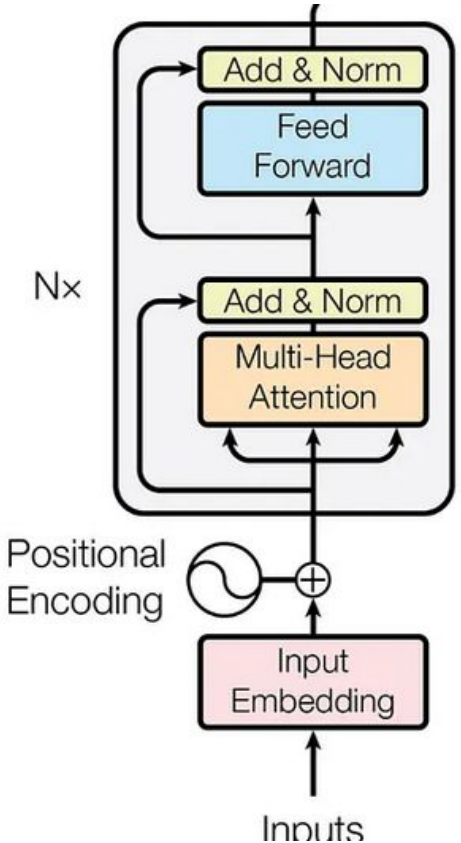
- Extended PCIE fabric provide pooling for memory and SSD
- Data movement all in hardware, minimal driver/FW involvement
- Big improvement in bandwidth and latency compared with traditional RDMA

KubeFlash's Revolutionary KFNative™ Technology



- Each GPU paired with SSD
- GPUs have dedicated computing scratch pad
- SSD hold all model parameters, gradients, optimizer states
- GPU exchange data through neighboring SSD

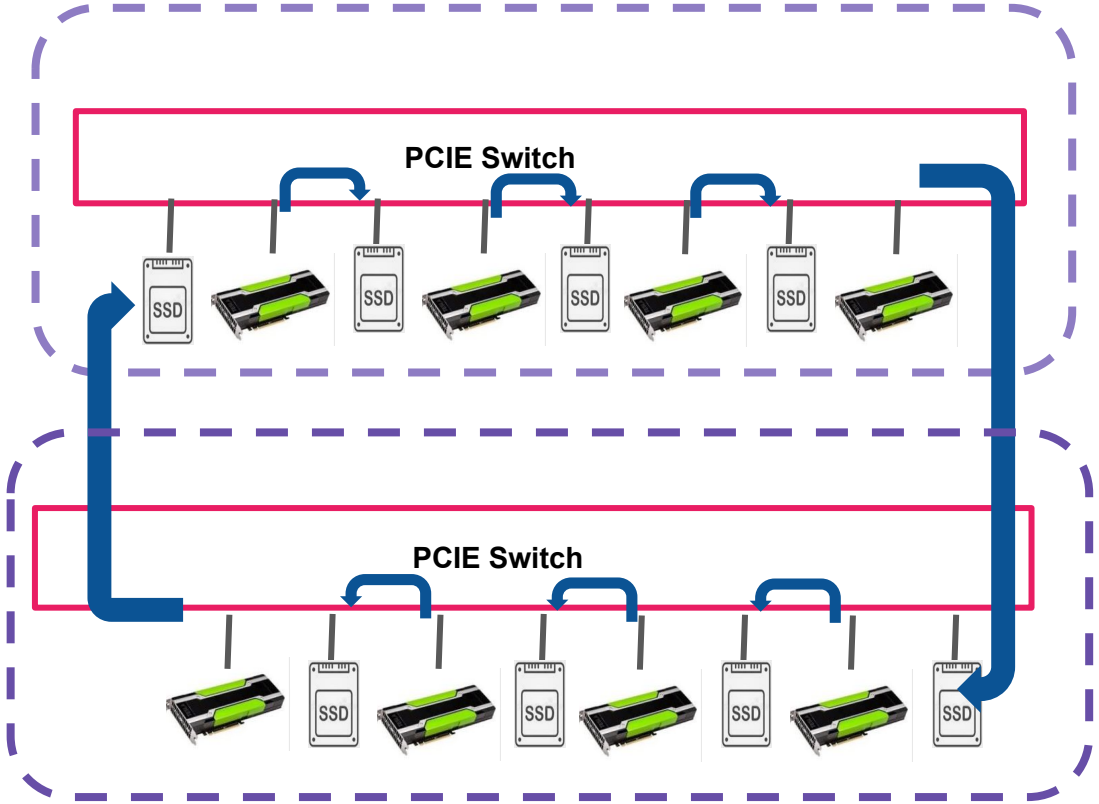
Comparable Performance with Far Fewer GPUs



Case Study of Streaming Computing during LLM Inference

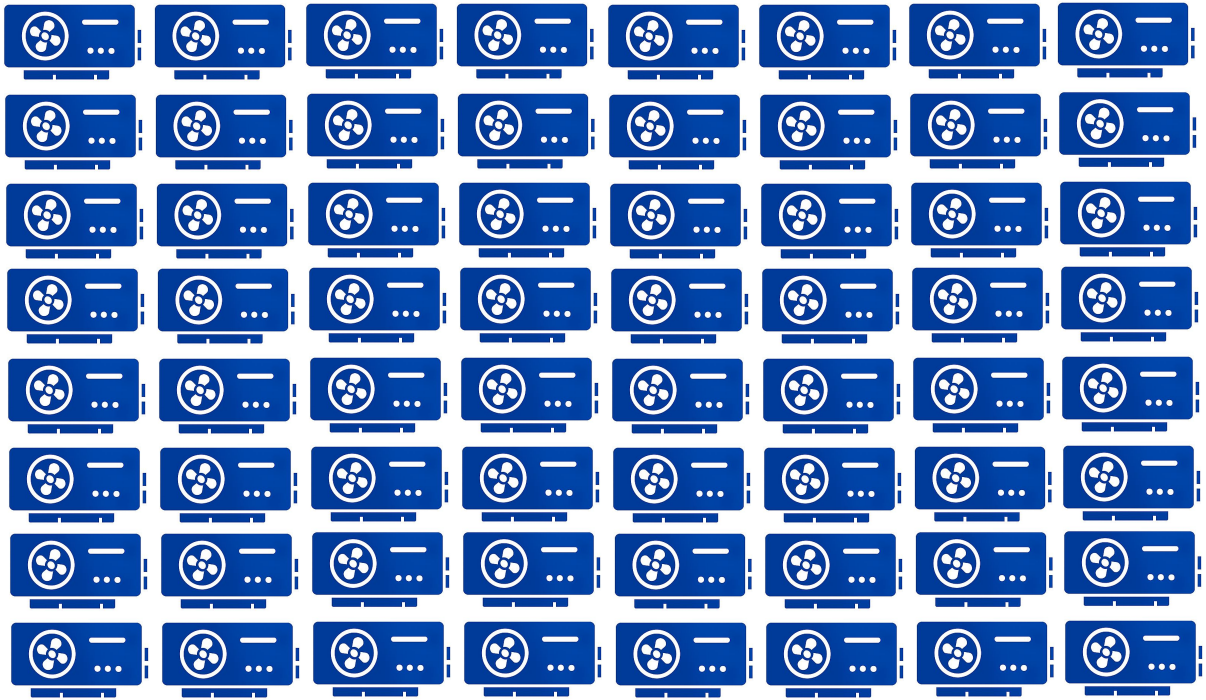
- GPT-3 175B, 96 layer
- Each layer ~1.8B, 3.6G Bytes
- Gen6x16 SSD bandwidth ~ 100GB/s
- Model loading time for one layer: 3.6/100 ~ 35ms
- Computing: 100K x 1.8B / 1000 T ~ 180ms

GPU Server Ring Extension



- Commodity GPU, less HBM dependency
- No NVlink dependency
- Better parallelism All/All Ring reduce
- Train largest LLM with **64X** fewer GPUs

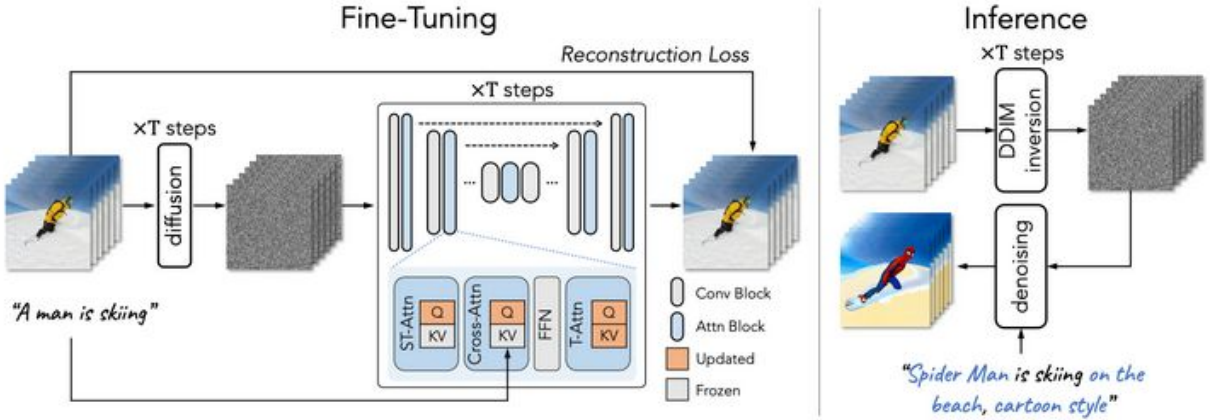
KubeFlash enable large scale training with low cost GPU or Alternative



- Nvidia 5090 ~ $2K
- 256 5090 ~$500K
- Aggregated ~ 50 Pflops
- Train any of the largest LLM
- Multi-Agent Reinforcement Learning

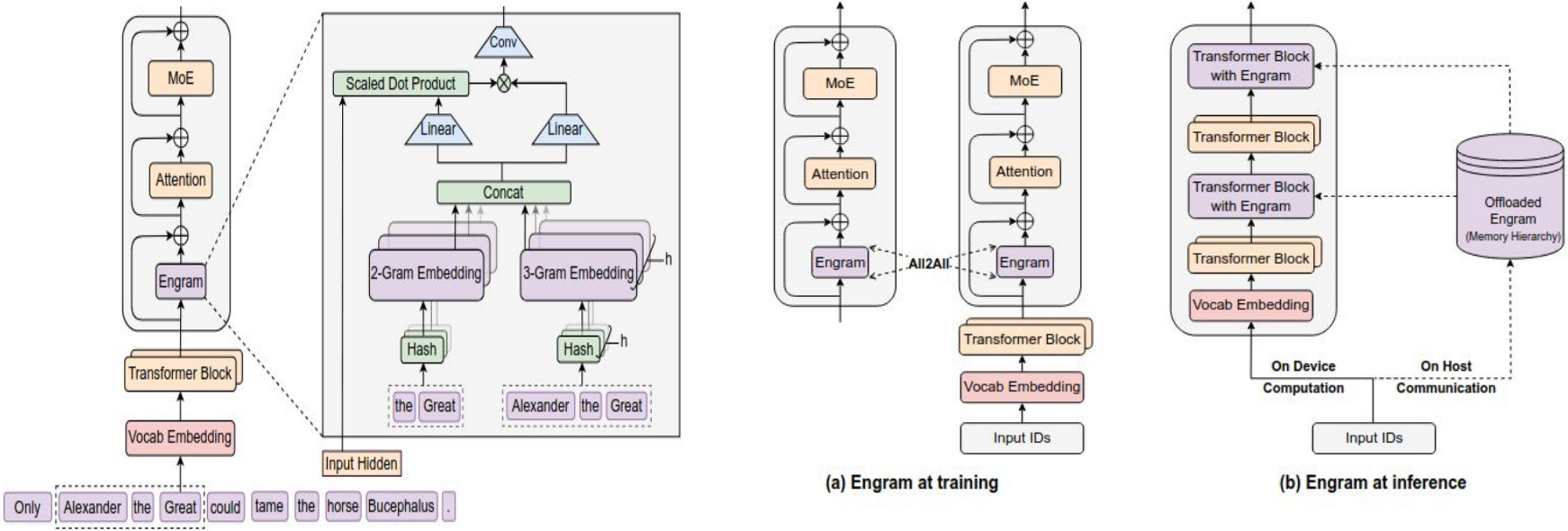
Data center cost is reduced **>100 times**.

Diffusion Transformer (DiT) - Video Generation



- **Inverse and Denoise process both have long computing to overlap with IO loading**
- **High definition video gen requires large HBM in current approach**
- **Kubeflash provide high quality video Gen solution with single GPU**

Greater Advantages with Engram LLM

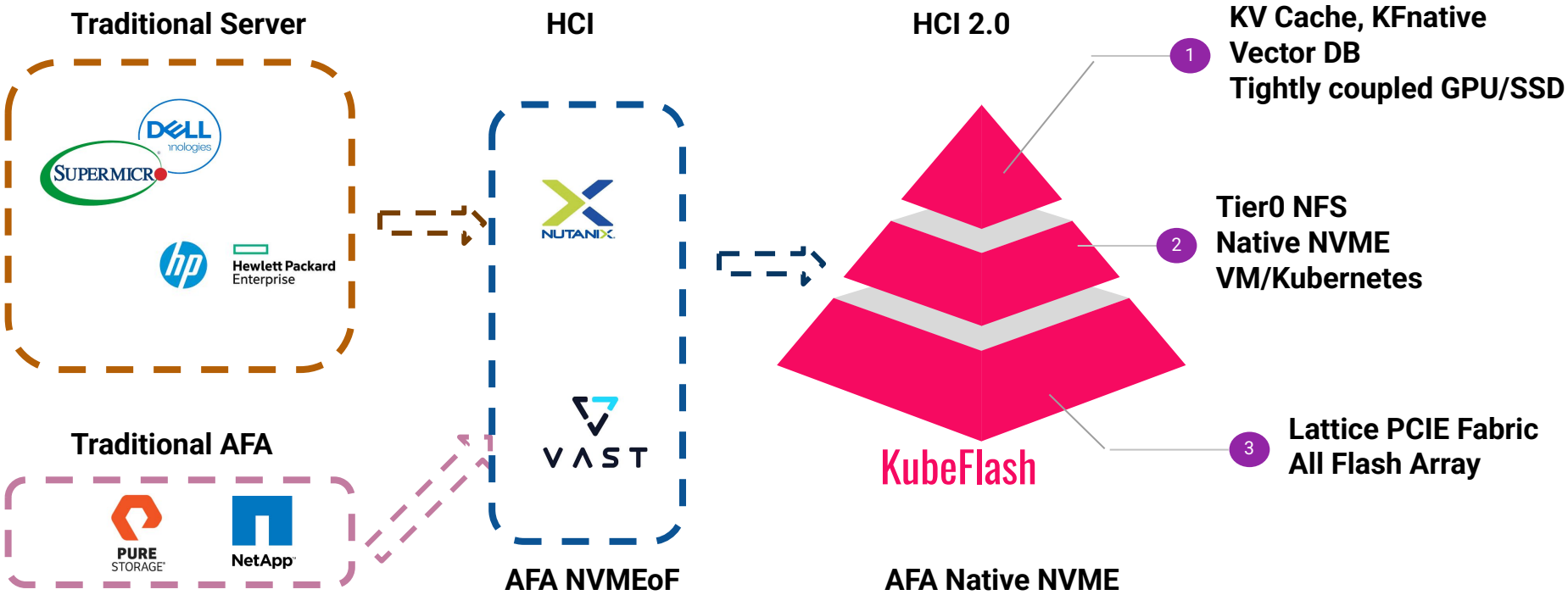


- **Both training and inference could leverage KFnative flash IO technology**
- **Provide large size engram vocabulary from SSD**
- **Natively combined engram lookup in the computing data path**

KubeFlash integrating with any platform with ease

- **Extended PCIE fabric could turn any server platform into high performance cluster**
- **All Flash Array built from standard off shelf components**
- **Converged and native SSD performance across cluster**

KubeFlash Computing infrastructure evolving



Market Segments for KubeFlash

	All Flash Array	HCI	AI & Data Analytics
Market Size	~ 25B	~ 15B	Billions
Major Player	PURE STORAGE, NetApp, VAST	NUTANIX, VMware vSAN™	Many Players
KubeFlash Advantage	- Off-the-Shelf components - Utilizing server from any vendor - High performance and low CPU utilization	- Native local NVME performance - Power efficient - Revolutionary all flash solution for Kubernetes -	- Converged SSD/Memory Caching - Native NVME throughput for KV cach - Boosting data analytics with native flash storage

Founding Team

- Wei Zhou, worked in Storage industry for over 20 years, have held senior management position at Marvell and SK hynix, leading teams in storage HW/FW/system engineering. Master degree of Applied Physics from Stanford.
- K. Wong, have worked in various fields in storage, have broad experience in SSD FW engineering and linux kernel engineering, have held principle engineer positions at Marvell and Sk Hynix
- X. Chen, 20 years plus experience in storage industry, strong kernel and OS engineering expertise, have held various engineering and management positions at Marvell and Startups