Beyond KV-Cache

Topics for Today

- **KV-Cache: Bluefield DPU vs KubeFlash**
- **Deepseek Engram, implication for flash storage**
- **Beyond KV-Cache**

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

Bluefield/DPU essentially is All Flash Array

All Flash Array

Market Size ~ 25B

Major Player

KubeFlash Advantage

- Off-the-Shelf components
- Utilizing server from any vendor
- High performance and low CPU utilization

Bluefield

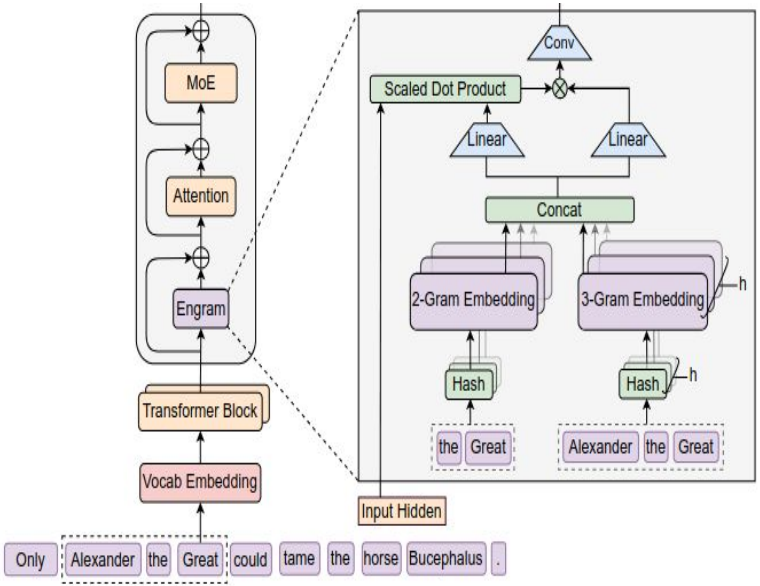
- Eliminate x86 CPU
- Redevelop full storage stack
- Proprietary Nvidia specific

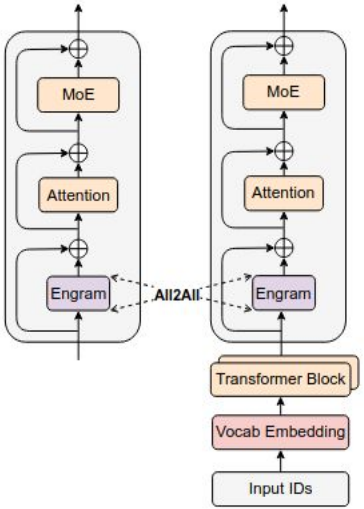
KubeFlash HCI mode best utilizing local native SSD

Cost effective, best investment/performance ratio

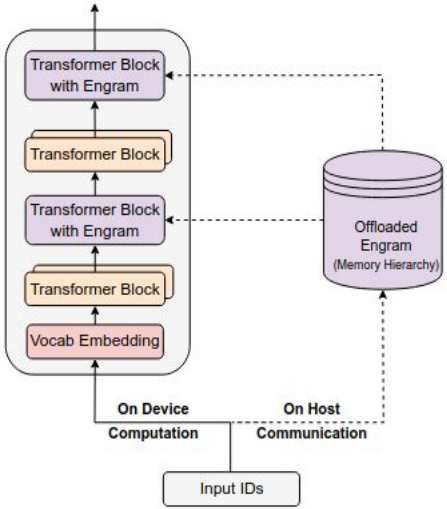
PCIE switch port limitation

Engram & Implication



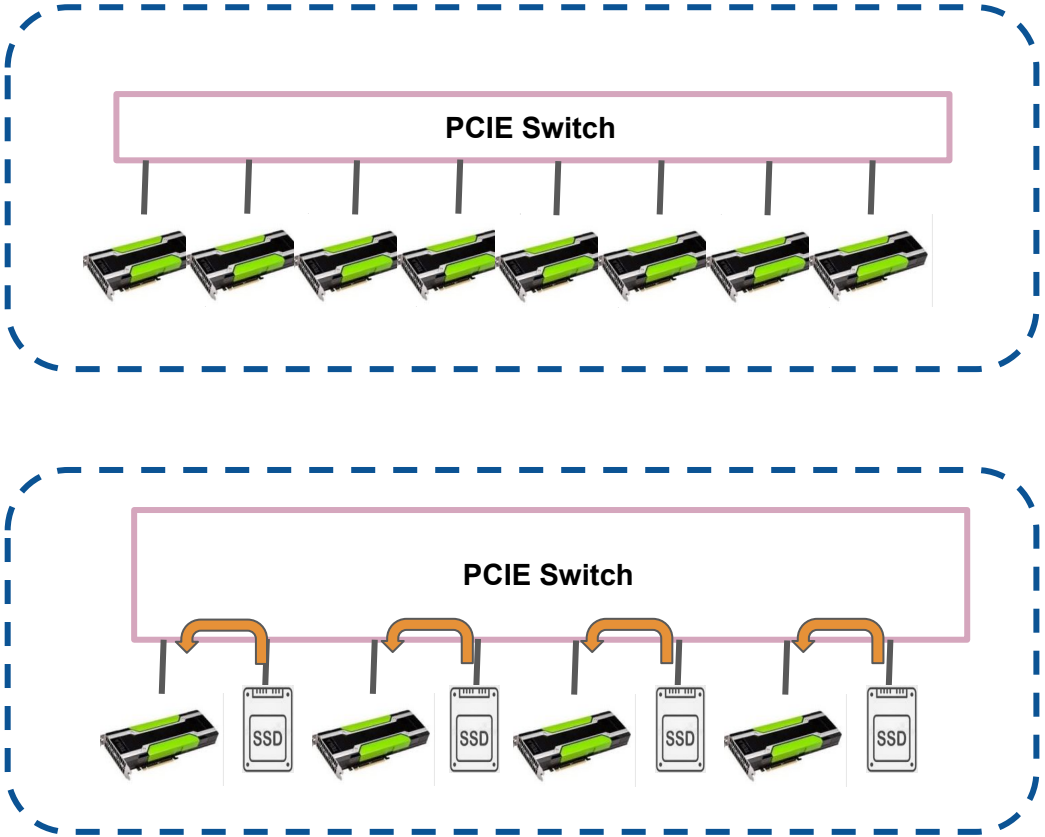


(a) Engram at training



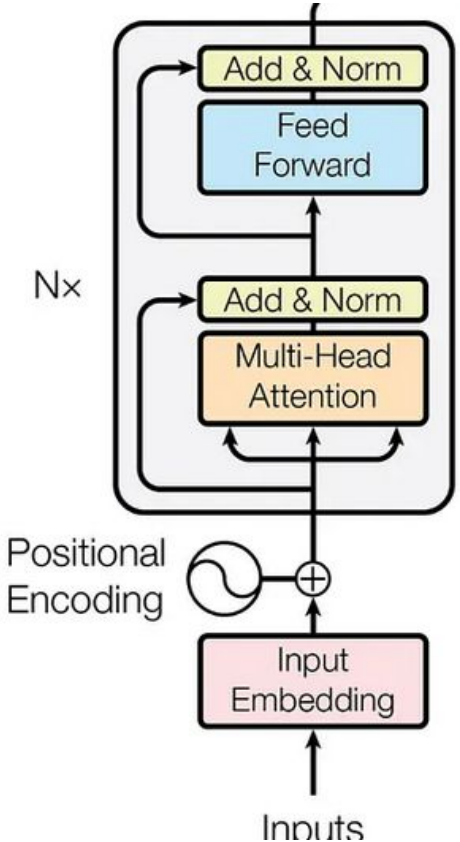
(b) Engram at inference

GPU Server Restructured



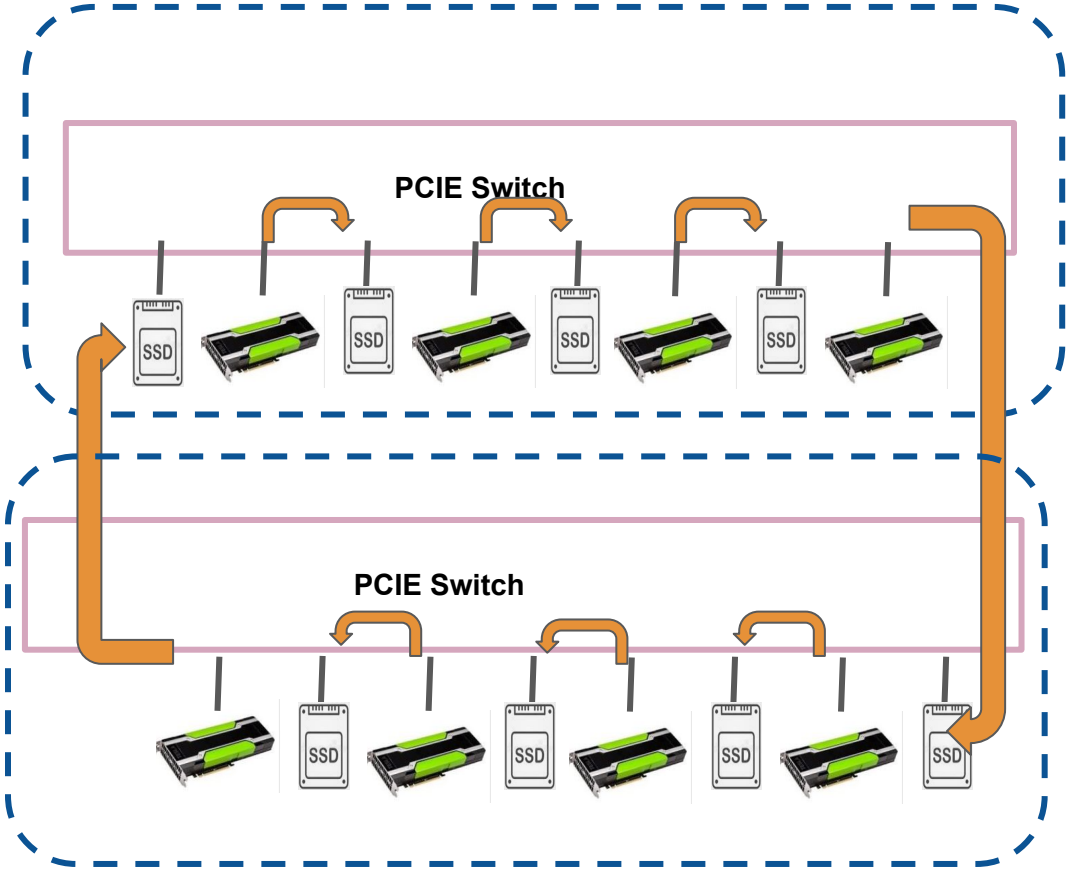
- GPU paired with SSD
- GPU now have dedicated computing scratch pad
- SSD will hold all model parameters, Gradients, optimizer states
- GPU exchange data through neighbour SSD

SSD LLM inference Streaming Computing



- GPT-3 175B, 96 layer
- Each layer ~1.8B, 3.6G Bytes
- Gen5x16 SSD bandwidth ~ 100GB/s
- Model loading time for one layer: 3.6/100 ~ **35ms**
- Computing: 100K x 1.8B / 1000 T ~ **180 ms**

GPU Server Ring Extension



- Commodity GPU, less HBM dependency
- No NVlink dependency
- Better parallelism
- Training largest LLM with a few tens of GPU