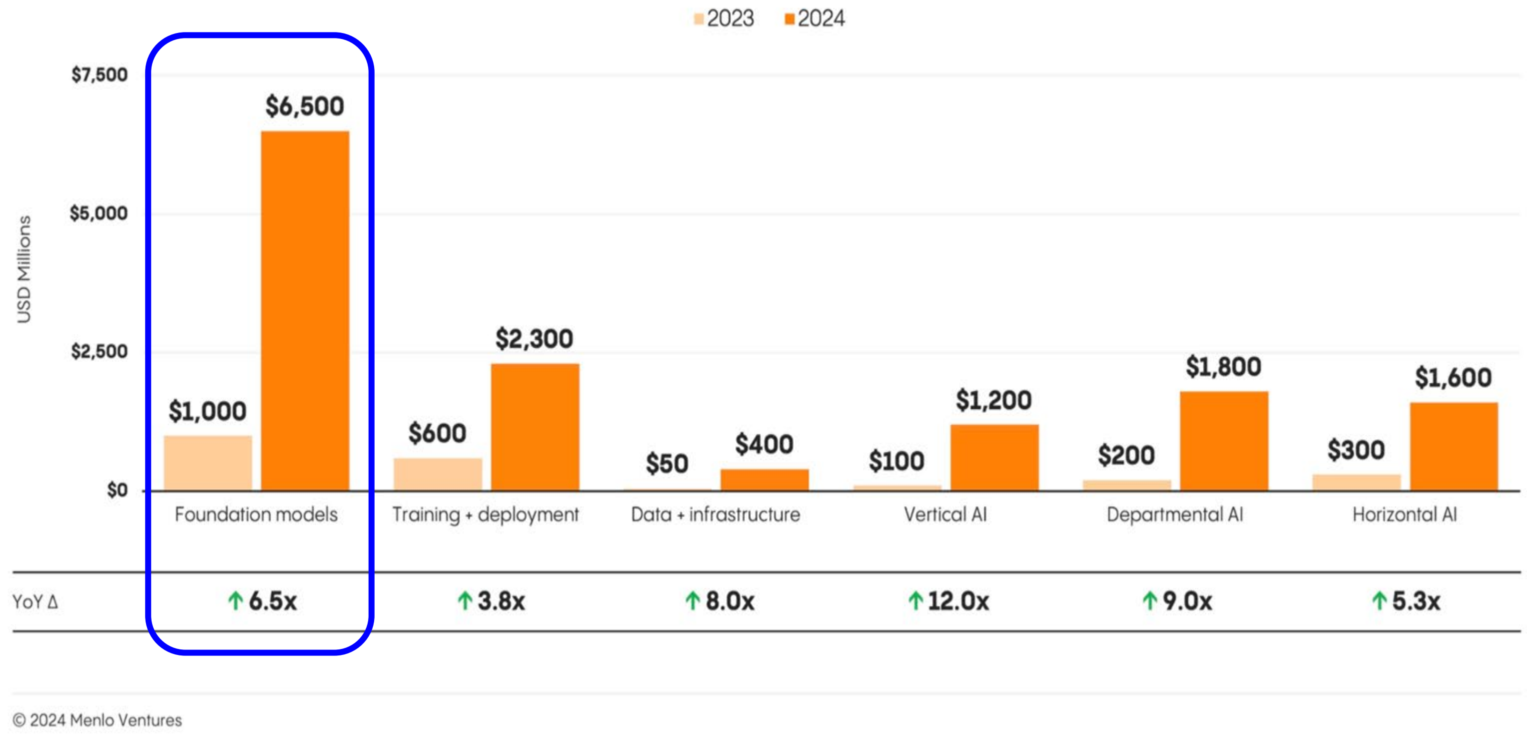
# AI Enterprise Market trends and opportunities

Nilesh Shah
VP Business Development
ZeroPoint Technologies

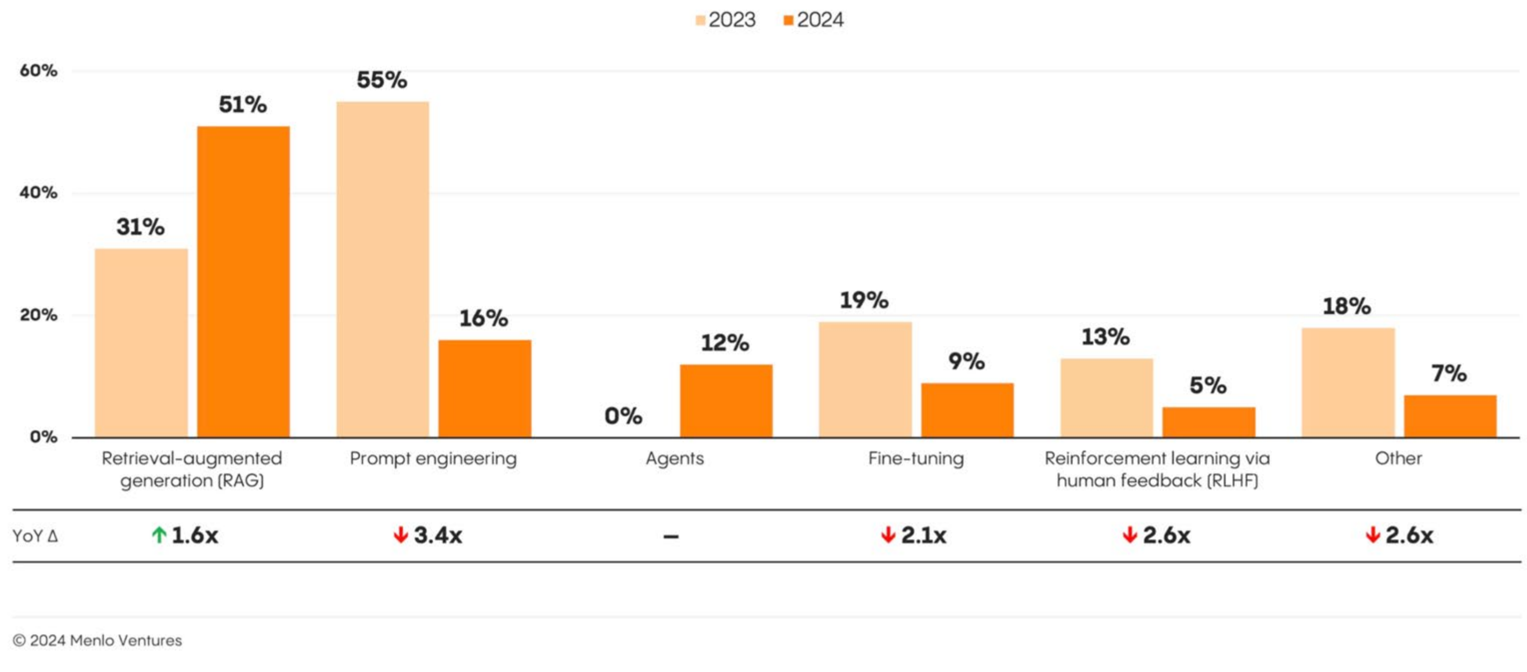
# Enterprise Gen AI: Trends

Inference spend dominates



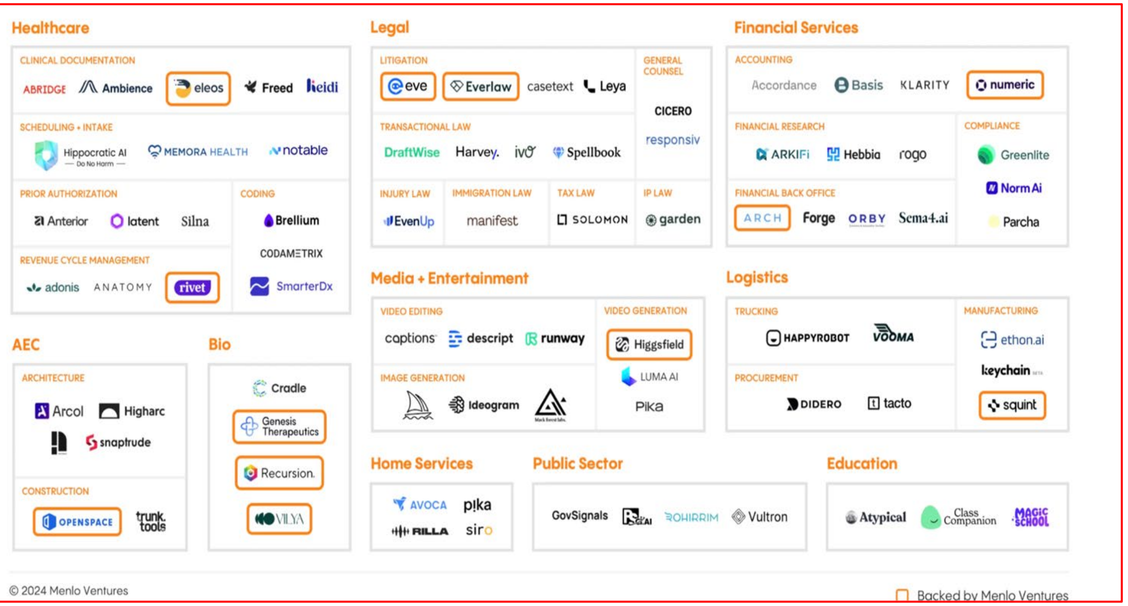
© 2024 Menlo Ventures

RAG use growing

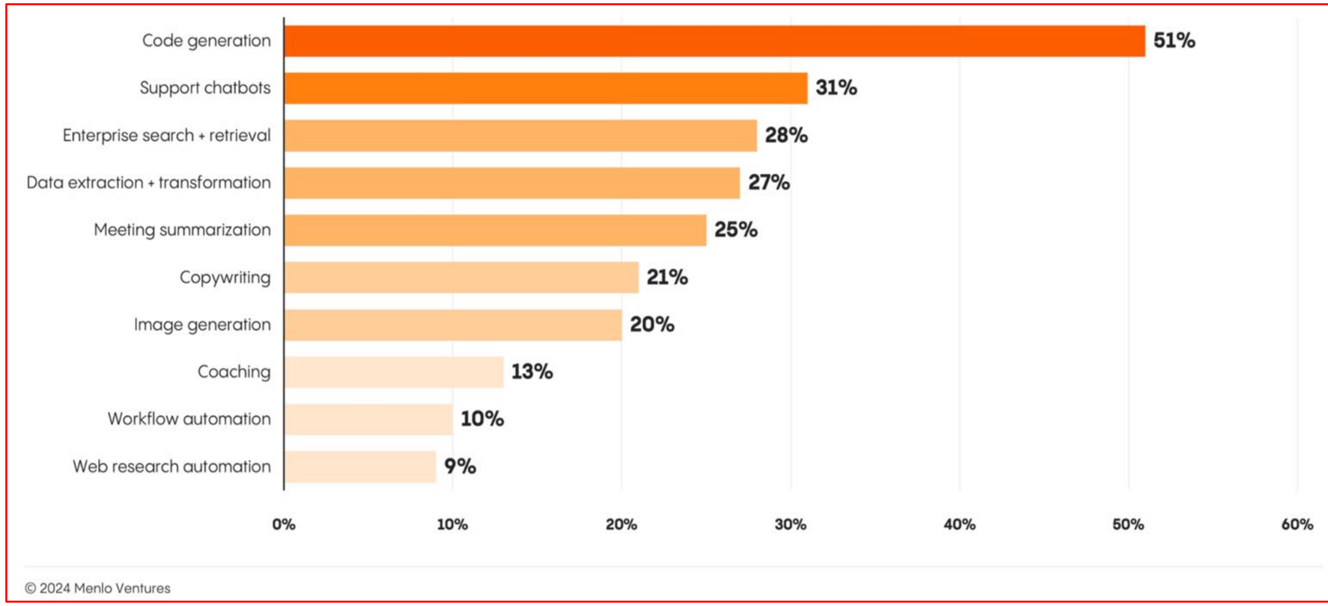


© 2024 Menlo Ventures

Source: https://menlovc.com/2023-the-state-of-generative-ai-in-the-enterprise-report/

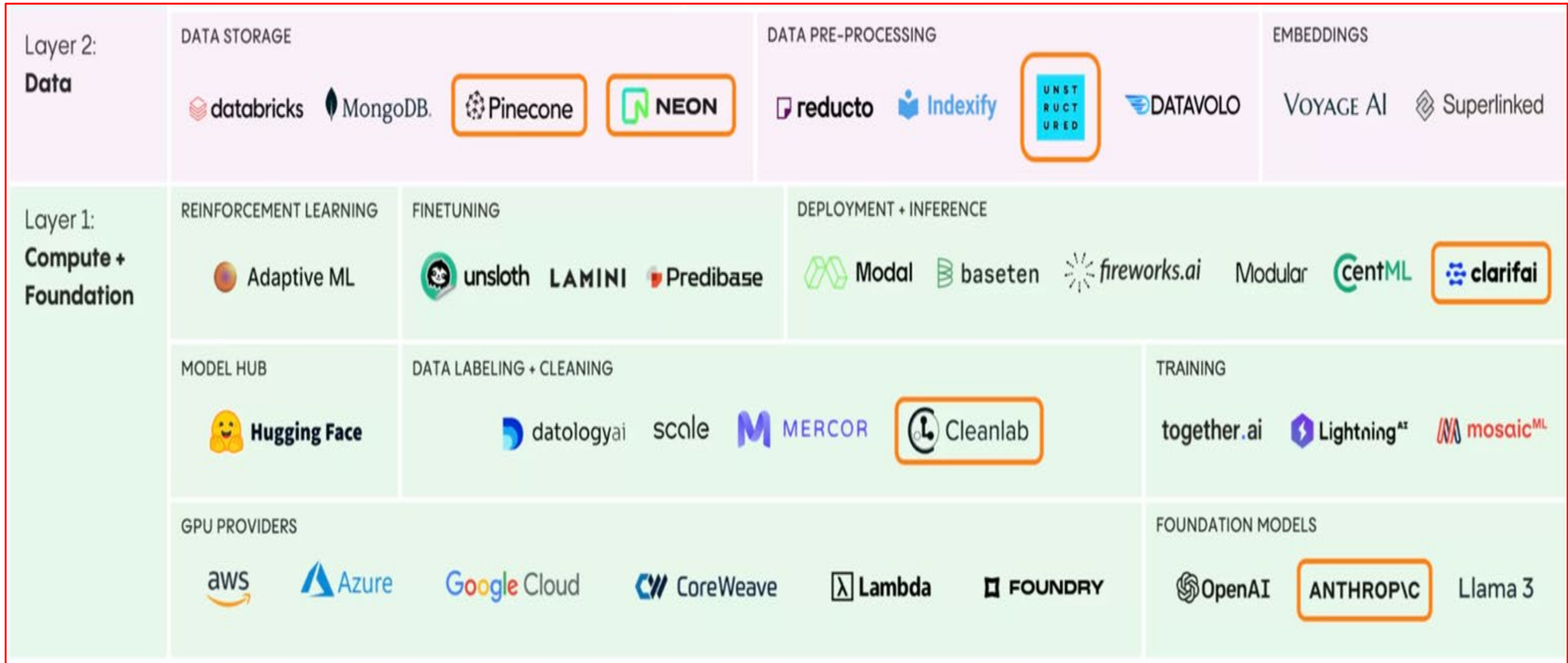


Early Adopters: Healthcare, Legal, Financial



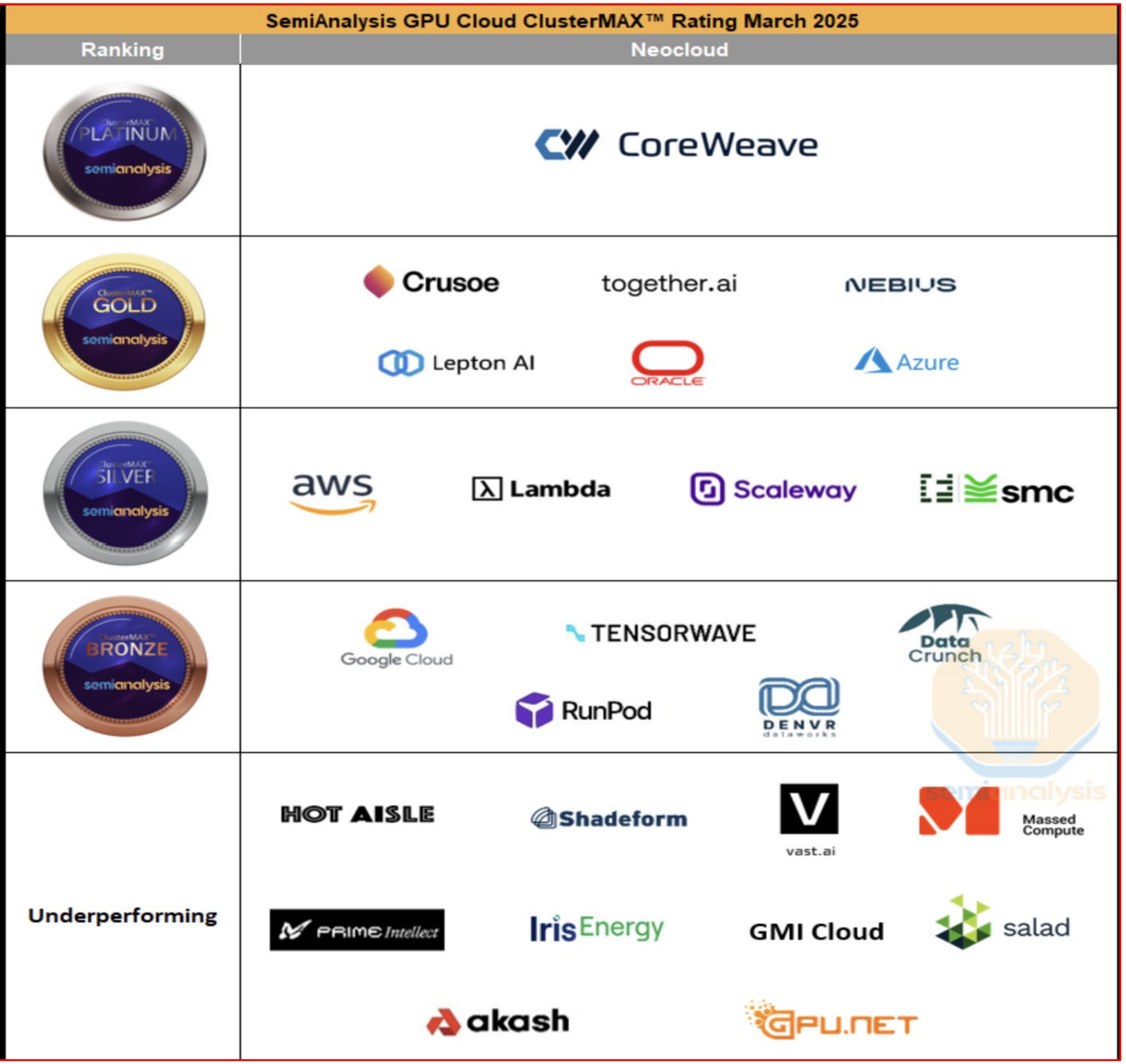
Dominant Use cases: Code generation, Chatbots

Data Stack: New Entrants



FMS
*the Future of Memory and Storage*

# Where 90%+ Gen AI executes: NeoCloud



| SemiAnalysis GPU Cloud ClusterMAX™ Rating March 2025 | |
|---|---|
| Ranking | Neocloud |
| ClusterMAX PLATINUM semianalysis | CoreWeave |
| ClusterMAX GOLD semianalysis | Crusoe<br>together.ai<br>NEBIUS<br>Lepton AI<br>ORACLE<br>Azure |
| ClusterMAX SILVER semianalysis | aws<br>Lambda<br>Scaleway<br>smc |
| ClusterMAX BRONZE semianalysis | Google Cloud<br>TENSORWAVE<br>Data Crunch<br>RunPod<br>DENVR dataworks |
| Underperforming | HOT AISLE<br>Shadeform<br>vast.ai<br>Massed Compute<br>PRIME Intellect<br>IrisEnergy<br>GMI Cloud<br>salad<br>akash<br>GPU.NET |

source: https://semianalysis.com/2025/03/26/the-gpu-cloud-clustermax-rating-system-how-to-rent-gpus/

# Infrastructure Scale: API vs NeoCloud vs On prem

| Deployment Model | Provider / Client | End Users (est.) | GPUs / Servers Used | Inference vs Training Share |
|---|---|---|---|---|
| API services (ChatGPT) | OpenAI via CoreWeave | **Millions** | **~10K+ H100 GPUs** | ~95 % inference / 5 % training |
| Hybrid (MSFT + NeoCloud) | Microsoft + CoreWeave | **Hundreds K+ users** | **~150K GPUs share** | ~90 % inference |
| Public hyperscaler overflow | Azure, GCP using NeoCloud | 10k+ enterprise developers | **Thousands GPUs burst** | Inference heavy use mainly |
| On-prem / colocation | Enterprise in NeoCloud DCs | Hundreds to thousands | **Hundreds GPU scale clusters** | ~50–70% inference, RAG-heavy |

# Storage Implications: API vs NeoCloud vs On prem

| Feature | NeoCloud | Public Cloud | On-Prem / Colocation | API-driven Services |
|---|---|---|---|---|
| Hot Tier (Flash/NVMe) | Yes (GPU-affine NVMe nodes) | Yes (EBS, gp3) | Yes (local NVMe/Optane) | Abstracted from user |
| Cold Tier (Object/HDD) | Yes, optional object scale-out | Yes (S3, Blob) | Optional via NAS or tape | Abstracted |
| Vector DB Integration | Built-in or orchestration-ready | Managed vector DB services | Manually deployed systems | Encapsulated in endpoint |
| KV Cache Tiering | NVMe-oF offload with GPUDirect | Limited caching layers | Custom tiered caches possible | Opaque |
| Shared Multi-Tenancy | Tenant-aware orchestration | Platform-level isolation only | Full control per enterprise | Not exposed |
| Latency Guarantees | ~1–10 ms via NVMe-fabric | ~5–100 ms across regions | ~0.5–5 ms locally | Depends on provider |
| Custom Embedding Support | Full control & custom layout | API-specific restrictions | Fully programmable | Limited or hidden |

# Are all LLMs Transformers? Emerging Model Categories 2025

State of the Art

| Category Name | What it Represents | Examples |
|---|---|---|
| **Sparse MoE LLMs** | Scalable, expert-gated Transformer | DeepSeek-V2, Mixtral, AlexaTM, Switch |
| **State Space Hybrids** | Linear-time sequence models | JAMBA, Mamba, RWKV, RetNet |
| **Structured Token-Free** | Non-token, graph, patch, or recurrent | Gemini 1.5, Hyena, CoLT5, MEGA |

GPT style transformers are so 2024!

Memory / Storage Implications

| Metric | Sparse MoE (DeepSeek, Mixtral) | SSM Hybrid (JAMBA, Mamba) | Structured/Tokenless (Gemini, Hyena) | Dense Transformer (GPT-style) |
|---|---|---|---|---|
| **VRAM per Inference Session** | ~10 GB | ~6 GB | ~8 GB | ~20 GB |
| **Tokens/sec per 1MW Power** | ~800K | ~1.2M | ~1.0M | ~500K |
| **Storage (100M docs, full RAG)** | ~100 TB | ~80 TB | ~90 TB | ~120 TB |
| **Concurrent Users per 1MW** | ~2,000 | ~3,500 | ~3,000 | ~1,000 |

FMS the Future of Memory and Storage

# Model storage requirements

| Component | Sparse MoE *(DeepSeek, Mixtral)* | SSM Hybrid *(JAMBA, Mamba)* | Structured / Tokenless *(Gemini, Hyena)* | Dense Transformer *(GPT-style)* |
|---|---|---|---|---|
| **Vector DB** | ~1.0 TB | ~0.8 TB | ~0.9–1.0 TB | ~1.5–2.5 TB |
| **KV Cache (active)** | ~1.1–1.5 TB | ~0.6–0.8 TB | ~0.8–1.2 TB | ~2–3 TB |
| **Embedding Store** | ~0.8 TB | ~0.5 TB | ~0.6–0.9 TB | ~1.2 TB |
| **Total (approx.)** | ~3.0–3.3 TB | ~1.9–2.1 TB | ~2.3–3.1 TB | ~4.9–6.7 TB |

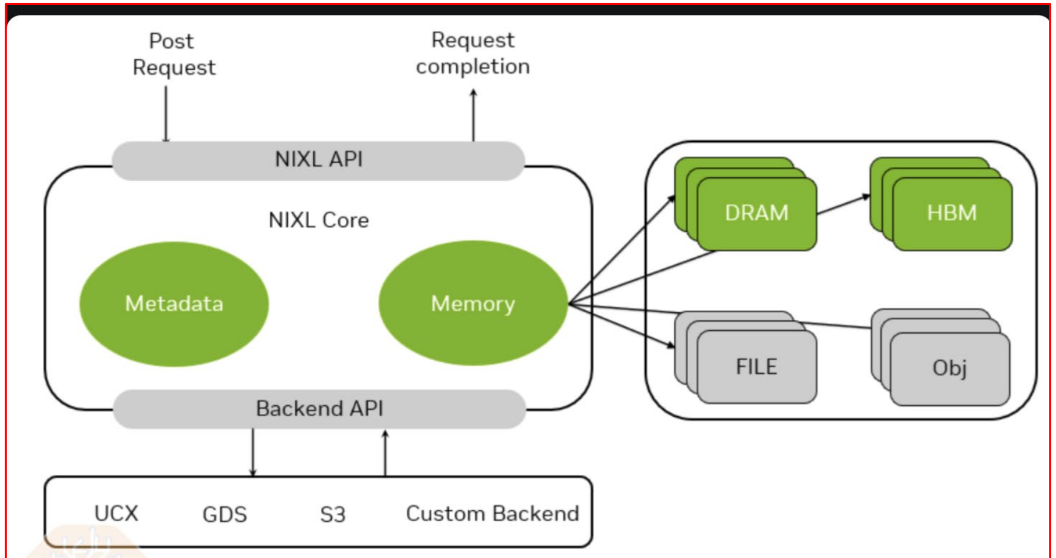


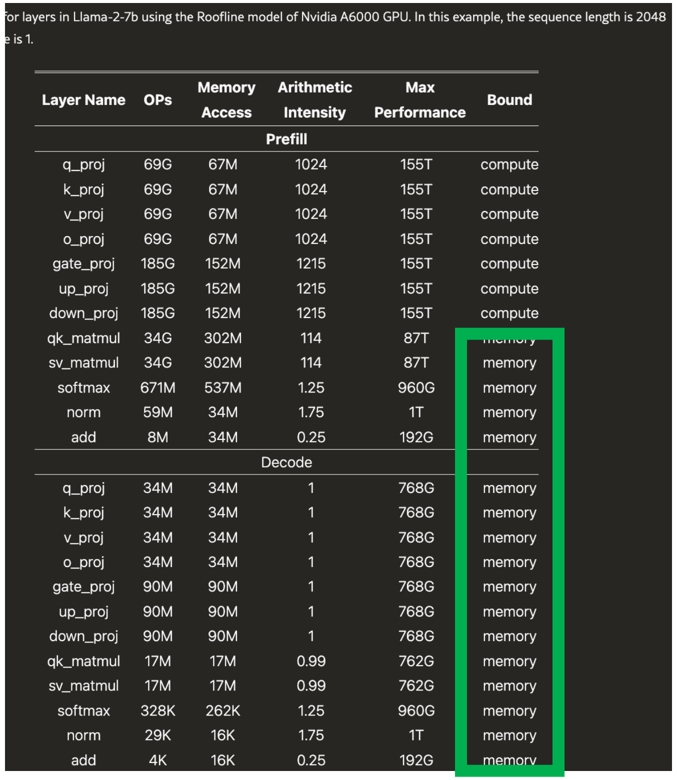
*Figure 7. NVIDIA Inference Transfer Library (NIXL) abstracts the complexity of data movement across heterogeneous memory and storage devices*

Inference Latency matters: Optimizations, Parallel File systems are a MUST HAVE

the Future of Memory and Storage

# Innovations

## LLMs: Memory Bound

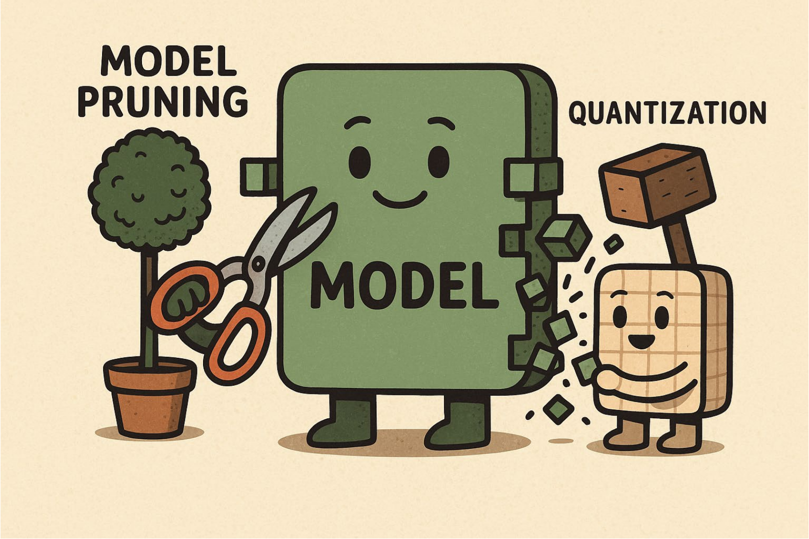


for layers in Llama-2-7b using the Roofline model of Nvidia A6000 GPU. In this example, the sequence length is 2048
e is 1.

| Layer Name | OPs | Memory Access | Arithmetic Intensity | Max Performance | Bound |
|---|---|---|---|---|---|
| | | | Prefill | | |
| q_proj | 69G | 67M | 1024 | 155T | compute |
| k_proj | 69G | 67M | 1024 | 155T | compute |
| v_proj | 69G | 67M | 1024 | 155T | compute |
| o_proj | 69G | 67M | 1024 | 155T | compute |
| gate_proj | 185G | 152M | 1215 | 155T | compute |
| up_proj | 185G | 152M | 1215 | 155T | compute |
| down_proj | 185G | 152M | 1215 | 155T | compute |
| qk_matmul | 34G | 302M | 114 | 87T | memory |
| sv_matmul | 34G | 302M | 114 | 87T | memory |
| softmax | 671M | 537M | 1.25 | 960G | memory |
| norm | 59M | 34M | 1.75 | 1T | memory |
| add | 8M | 34M | 0.25 | 192G | memory |
| | | | Decode | | |
| q_proj | 34M | 34M | 1 | 768G | memory |
| k_proj | 34M | 34M | 1 | 768G | memory |
| v_proj | 34M | 34M | 1 | 768G | memory |
| o_proj | 34M | 34M | 1 | 768G | memory |
| gate_proj | 90M | 90M | 1 | 768G | memory |
| up_proj | 90M | 90M | 1 | 768G | memory |
| down_proj | 90M | 90M | 1 | 768G | memory |
| qk_matmul | 17M | 17M | 0.99 | 762G | memory |
| sv_matmul | 17M | 17M | 0.99 | 762G | memory |
| softmax | 328K | 262K | 1.25 | 960G | memory |
| norm | 29K | 16K | 1.75 | 1T | memory |
| add | 4K | 16K | 0.25 | 192G | memory |

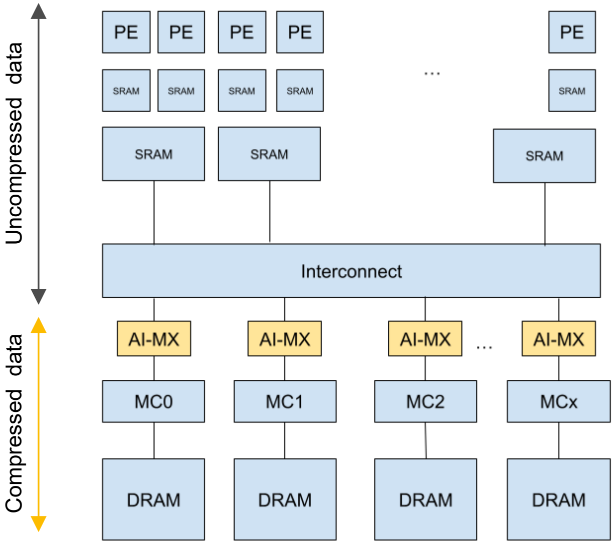
[LLM Inference Unveiled: Survey and Roofline Model Insights]

## Model Pruning + Quantization

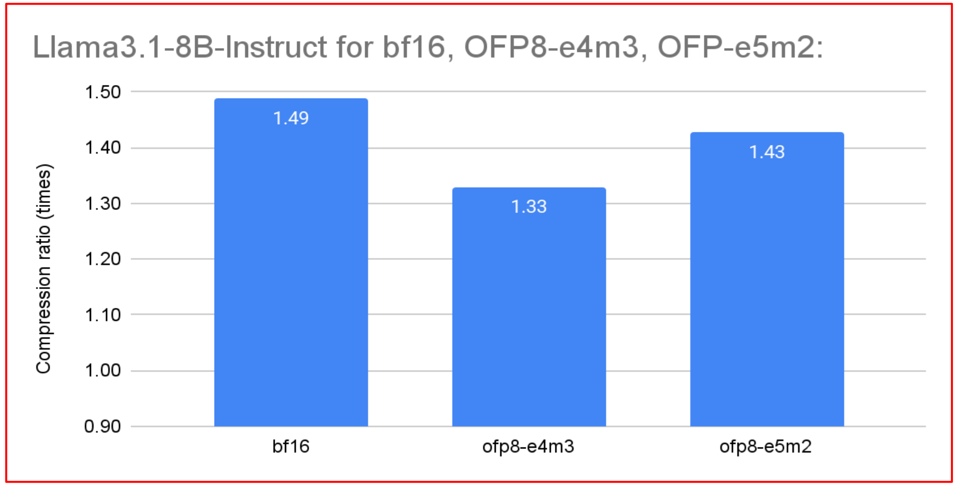


Accuracy Loss,
expensive Retraining

## Lossless HW accelerated (de)compression



1.5X Model compression

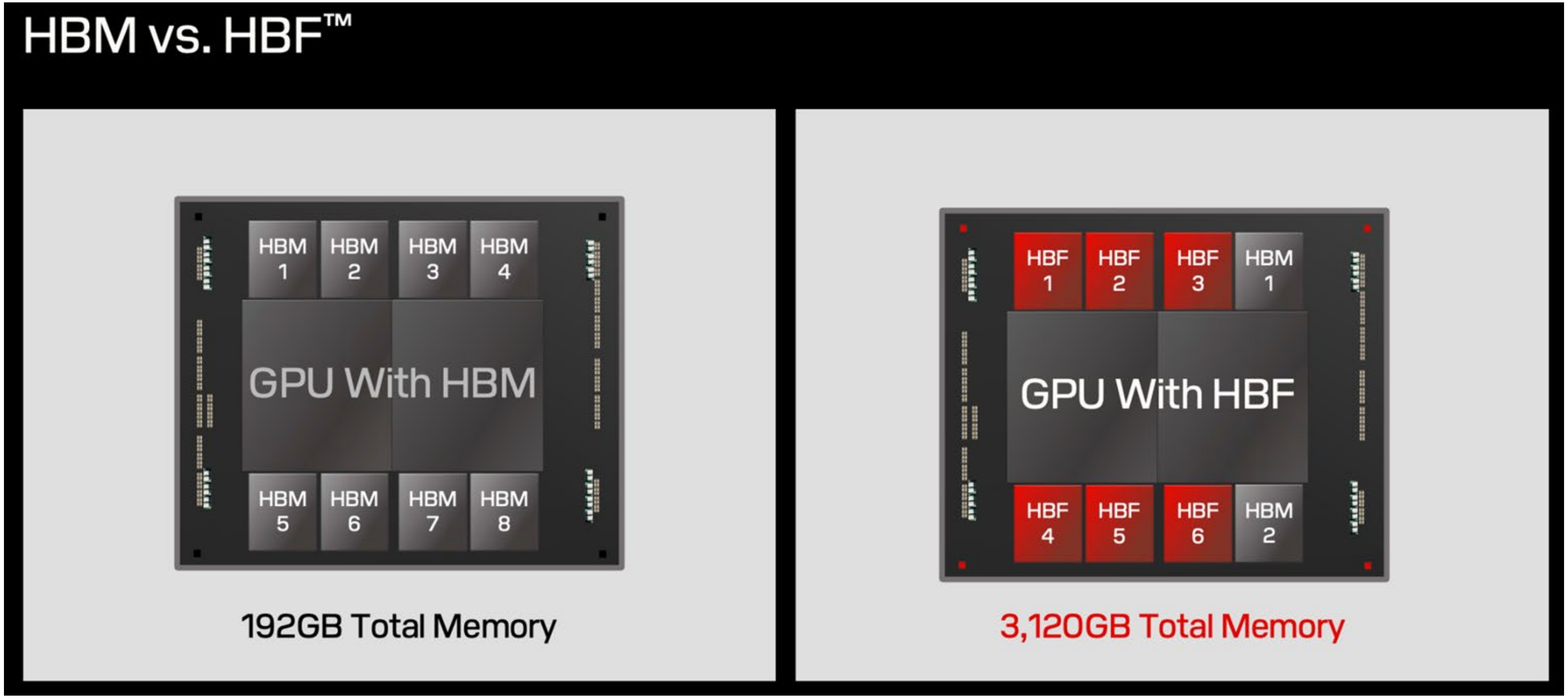


Llama3.1-8B-Instruct for bf16, OFP8-e4m3, OFP-e5m2:

# Innovations

High Bandwidth Flash

Combine compression with HBF to go from 3TB to 6TB?

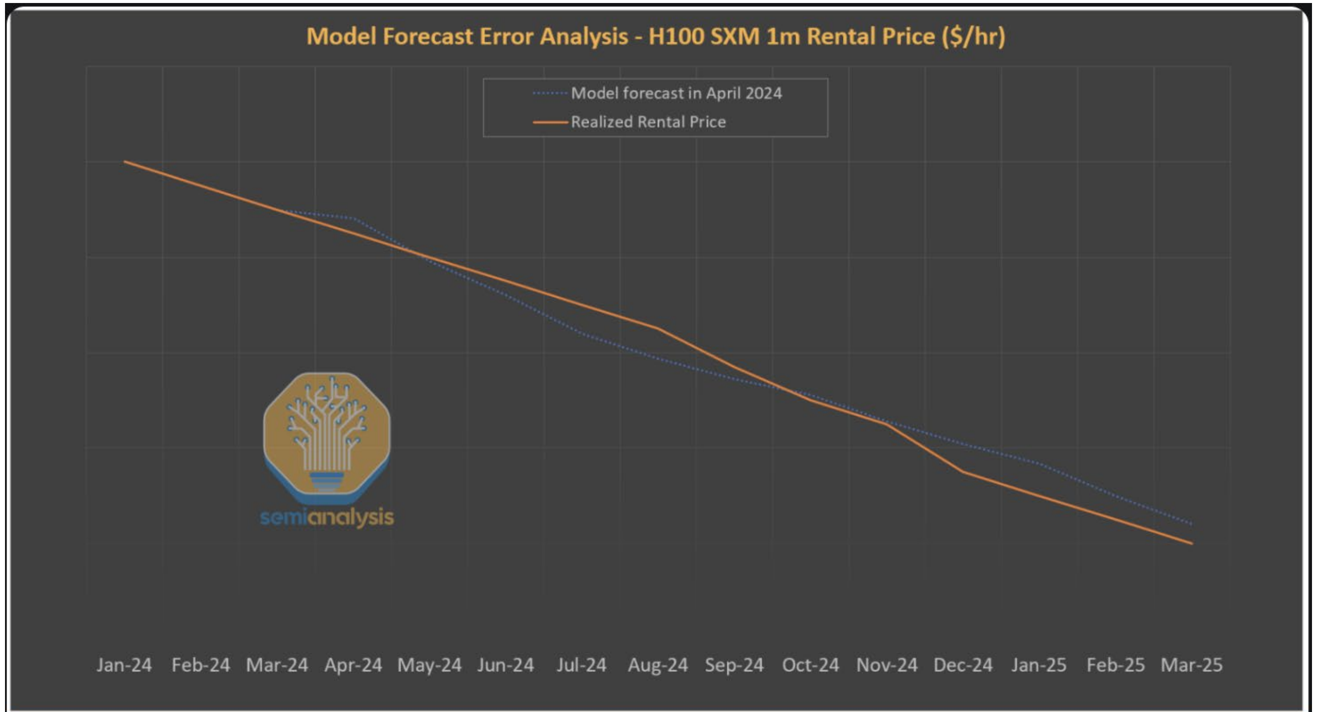


https://investor.sandisk.com/events/event-details/future-fwd-sandisk-2025-investor-day

# Innovations

Neocloud GPU rental margins race to the bottom

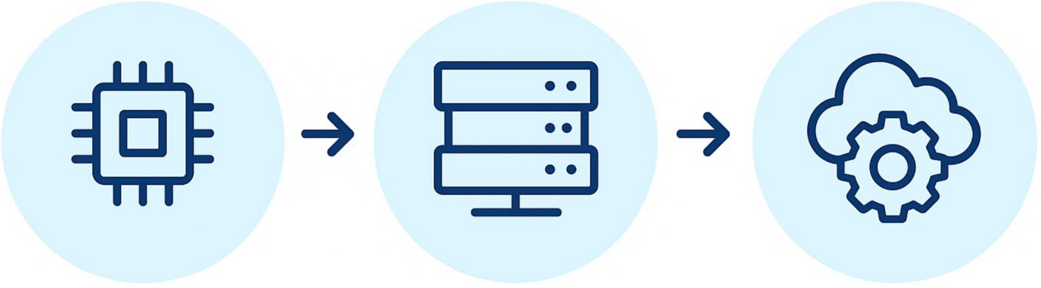
storage component providers buried in the value chain



Forecast Error Analysis

| | Jan-24 | Feb-24 | Mar-24 | Apr-24 | May-24 | Jun-24 | Jul-24 | Aug-24 | Sep-24 | Oct-24 | Nov-24 | Dec-24 | Jan-25 | Feb-25 | Mar-25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Actual Realized Pricing | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] |
| Forecast from April-24 Model | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] |
| *% Realized vs Forecast* | | | | | *0.3%* | *1.1%* | *2.3%* | *2.4%* | *1.0%* | *-0.4%* | *-0.2%* | *-2.4%* | *-2.9%* | *-2.1%* | *-1.8%* |

Business Model Innovation required
Move up value chain



**Component Provider** **Solution Provider** **Service Provider**

the Future of Memory and Storage

# Summary/ Call to Action

Summary:

- Inference /RAG dominates Enterprise AI deployments

- Storage /memory technologies emerging to match use cases

Call To Action:

- Partner to jointly innovate around new storage/memory technologies, business models