PCP SOLUTIONS

AIT275060T

AI Training Computer

Product Introduction

Project code: FB613
PRO3 team | July 09, 2026

AIT275060T

Product Outlook

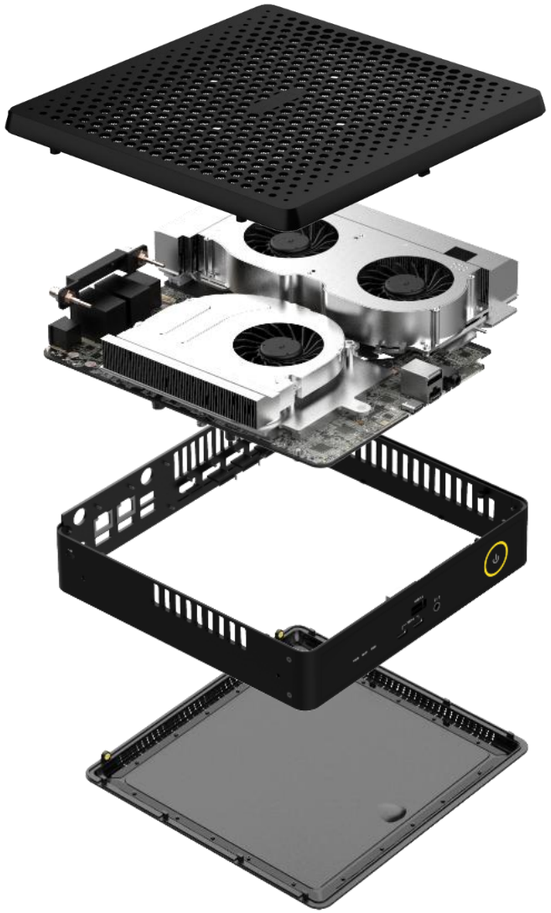
2.65 Litres

Back Panel

Top side

Left side

Right side



Front Panel

PCP SOLUTIONS

AIT275060T

Highlight Features

- The **AIT275060T** is an all-in-one, compact workstation engineered for small-scale full-lifecycle AI development, including LLM/VLM training, fine-tuning, and high-speed inference
- Data Sovereignty: Eliminates cloud dependency by providing a secure, on-premises environment for processing sensitive datasets with zero external data exposure.
- Powered by NVIDIA **GeForce RTX 5060 Ti 16GB** and high-endurance Phison AI SSD **aiDAPTIVCache** rated for 100 Drive Write Per Day (DWPD) to handle intensive AI workloads (*consumer NVMe: <0.3 DWPD*)
- A cost-effective solution to run / infer AI models typically requiring 24GB to 48GB+ VRAM by leveraging smart SSD offloading.
- Accelerate inference response times for long-context tasks by approximately 10 times compared to non-optimized SSD offloading.

Technical Specification	
Processor	Intel Core Ultra 7 Processor 255HX, 20-core, 5.2GHz, 30MB cache
Graphics	NVIDIA GeForce RTX 5060 Ti 16GB GDDR7 PCIe card
System memory	32 GB DDR5-6400 SO-DIMM
OS storage	2TB M.2 2280 NVMe SSD
AI cache	320GB Phison aiDAPTIVCache AI100 pSLC SSD rated for 100 DWPD
Video-out	1 x HDMI 2.1, 3 x DP 1.4 display output
Hi-Speed I/O	5 x USB 3.2 10Gbps, 2 x Thunderbolt 4
LAN	2 x 2.5Gbps Ethernet
Wireless	Intel Wi-Fi 7 IEEE 802.11be, Bluetooth 5.4
Dimension	210(L) x 203(W) x 62.2(H) mm^3, 2.65 Litres
Op. temperature	0˚C ~ +38˚C
Power Input	20V/330W AC-DC Power adapter
Supported OS	MS Windows 11, Ubuntu 24.04.3 or newer
Supported AI models	Llama-2-13b-hf, Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.1, gemma-2-9b-it, gemma-3-1b-it, Qwen2.5-14B/7B/3B/1.5B-Instruct, Qwen3-14B/8B, Phi-4, Phi-4-mini-instruct, gpt-oss-20b, gpt-oss-120b

PCP SOLUTIONS

AIT275060T

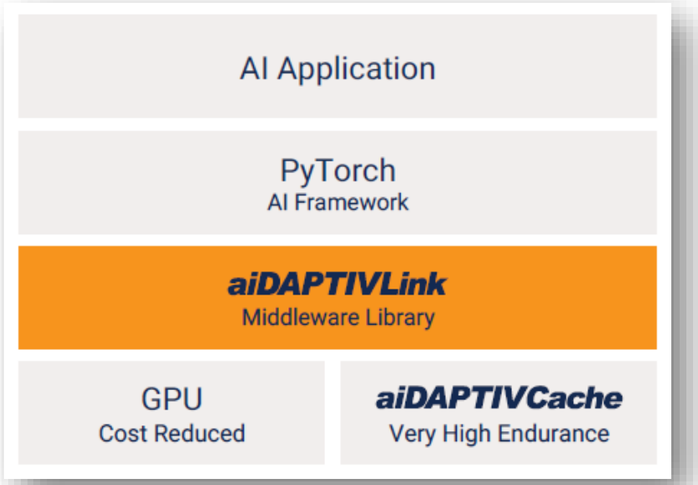
Use Cases

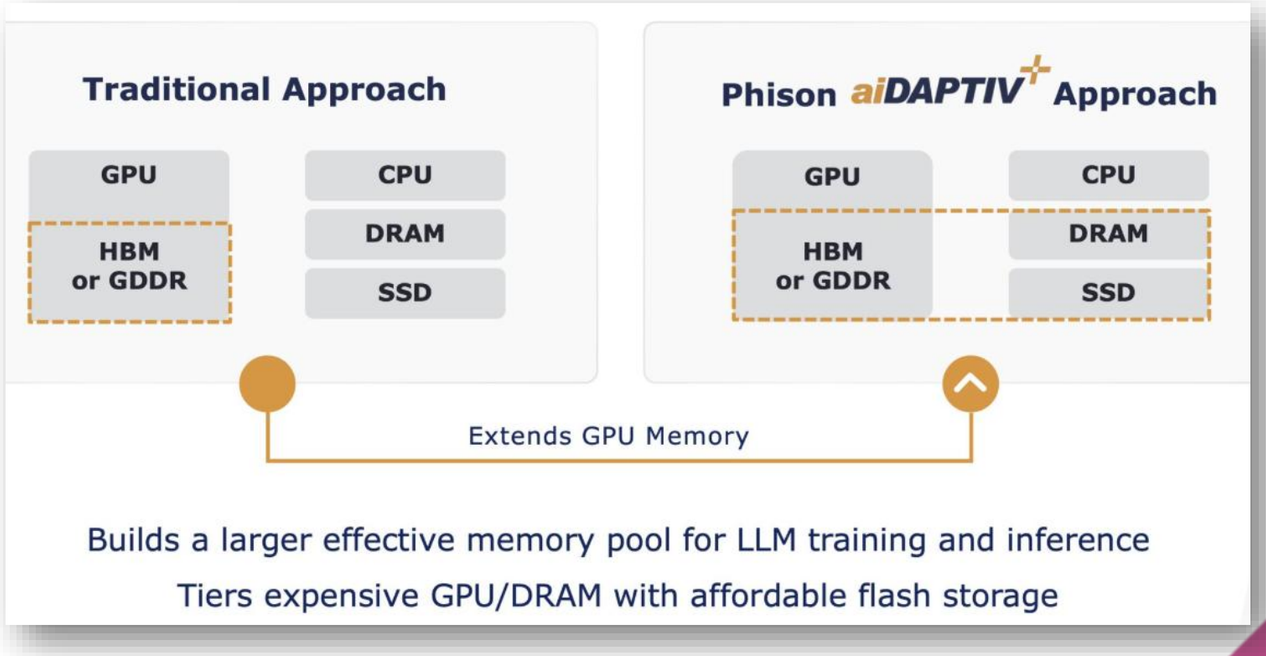
Knowledge RAG (Retrieval-Augmented Generation) inference	Serve assistants that are tuned to your business or curriculum using local data, without exposing that data to third-party clouds.
Domain-specific copilots and chatbots	Run retrieval-augmented generation pipelines on-prem to answer questions from internal documents, manuals, research, or records while keeping content private.
Coding assistants and tools	Host local code copilots that understand your repositories, build systems, and internal libraries — all from a secured workstation.
Agentic and long-context workflows	Support multi-step agents, longer session histories, and richer tool use by giving models more working memory without sacrificing latency.
Learning and experimentation	Give teams and students a hands-on environment to explore LLM behavior, safety, and evaluation using real workloads on local hardware.

AIT275060T

- **How it works:** The **AIT275060T** uses specialized high-endurance SSD and proprietary middleware to act as an **AI cache** (swap space) for graphics memory.
- **Software Stack:** aiDAPTIVLink middleware designed to integrate with PyTorch for training and inference.
- **Supported LLM:** Designed to facilitate Llama-2/3, Mistral-7b, gemma-2/3, Qwen 2.5/3, Phi-4, gpt-oss-20b/120b, and other models on consumer grade GPU by allowing the LLM fine-tuning, training, and inference to utilize the SSD as an extension of the GPU memory.
- This capability is achieved by transparently offloading critical components to AI SSD with minimal performance impact, including: Gradient offloading, MoE (Mixture of Expert) offloading, and KV-cache offloading.

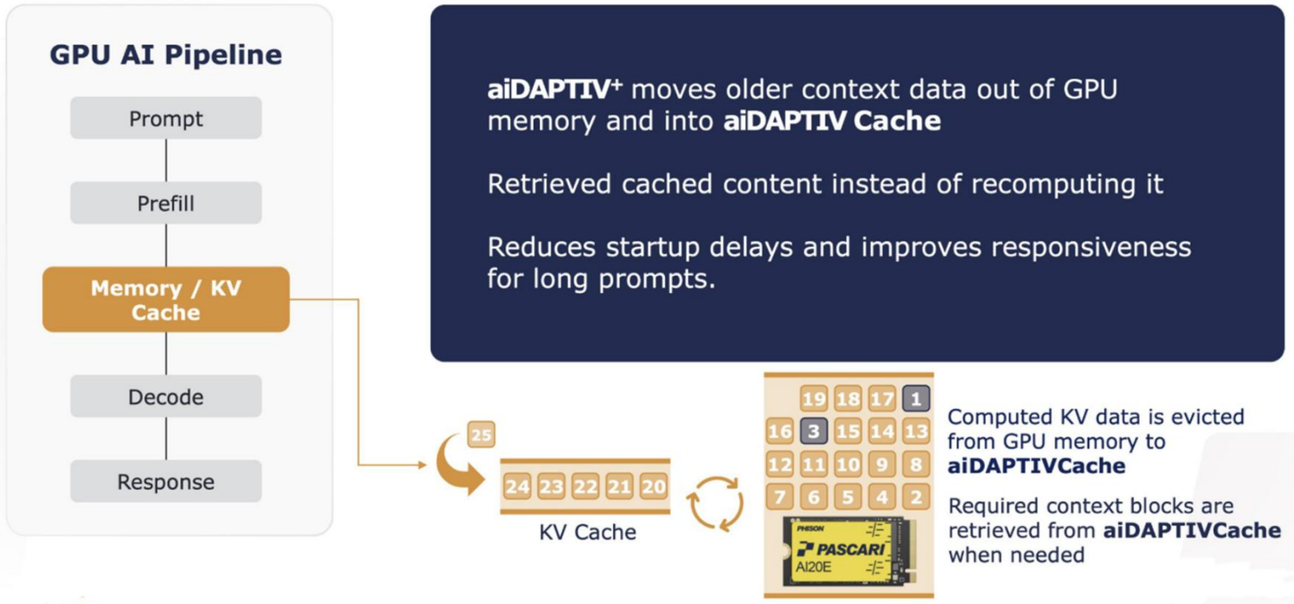
Key Benefit: While VRAM-only is always faster, aiDAPTIVCache makes it possible to **run** and **fine-tune** massive models on a single desktop PC that would otherwise require a high-end workstation or data-center-grade cluster.





AIT275060T

How does aiDAPTIVCache fix AI inference limitations?



aiDAPTIVCache vs Standard VRAM-Only Inference

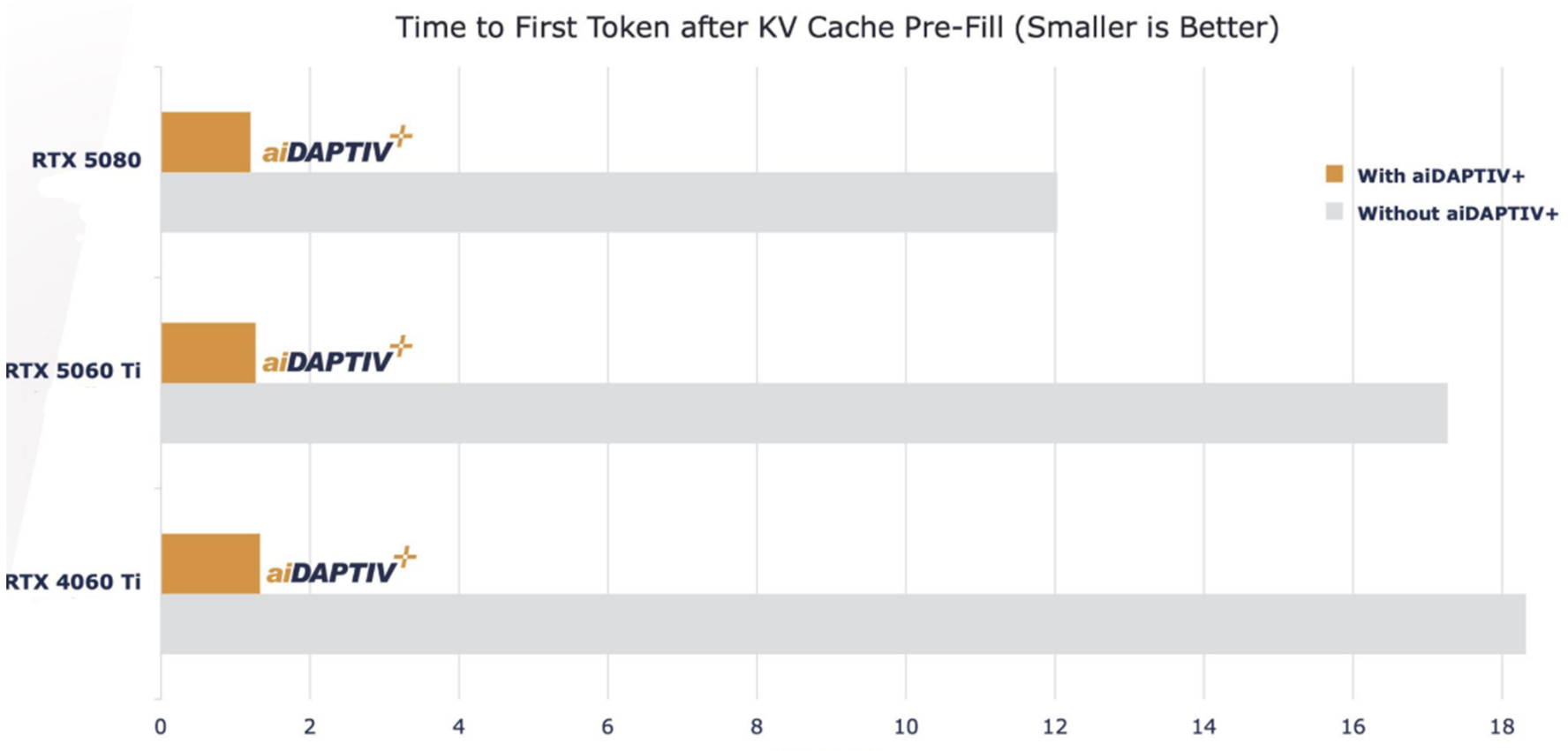
Feature	Standard VRAM-Only	aiDAPTIV AI Cache Drive
Model Size Limit	Hard-capped by physical VRAM (e.g. 32GB for an RTX 5090, 16GB for an RTX 5080/5060 Ti)	Virtually extended; can run **120B models** on a single consumer GPU with 16GB VRAM and 32GB system memory.
Performance	Highest throughput and lowest latency; no "swapping" overhead.	Lower throughput than pure VRAM, but up to **10x faster** than standard system-RAM-only offloading.
Cost	High; requires multiple data-center GPU for large models.	**Cost-effective**; uses affordable SSDs to mimic high-end GPU memory pools.
Data Handling	Model must fit entirely in VRAM for full speed.	Model is "sliced" and swapped between SSD cache and GPU VRAM as needed.

Note:
Key-Value (KV) cache is a crucial optimization technique in transformer-based LLMs used to accelerate the inference (generation) phase

PCP SOLUTIONS

AIT275060T

10X faster inference performance with aiDAPTIVCache (1/2)

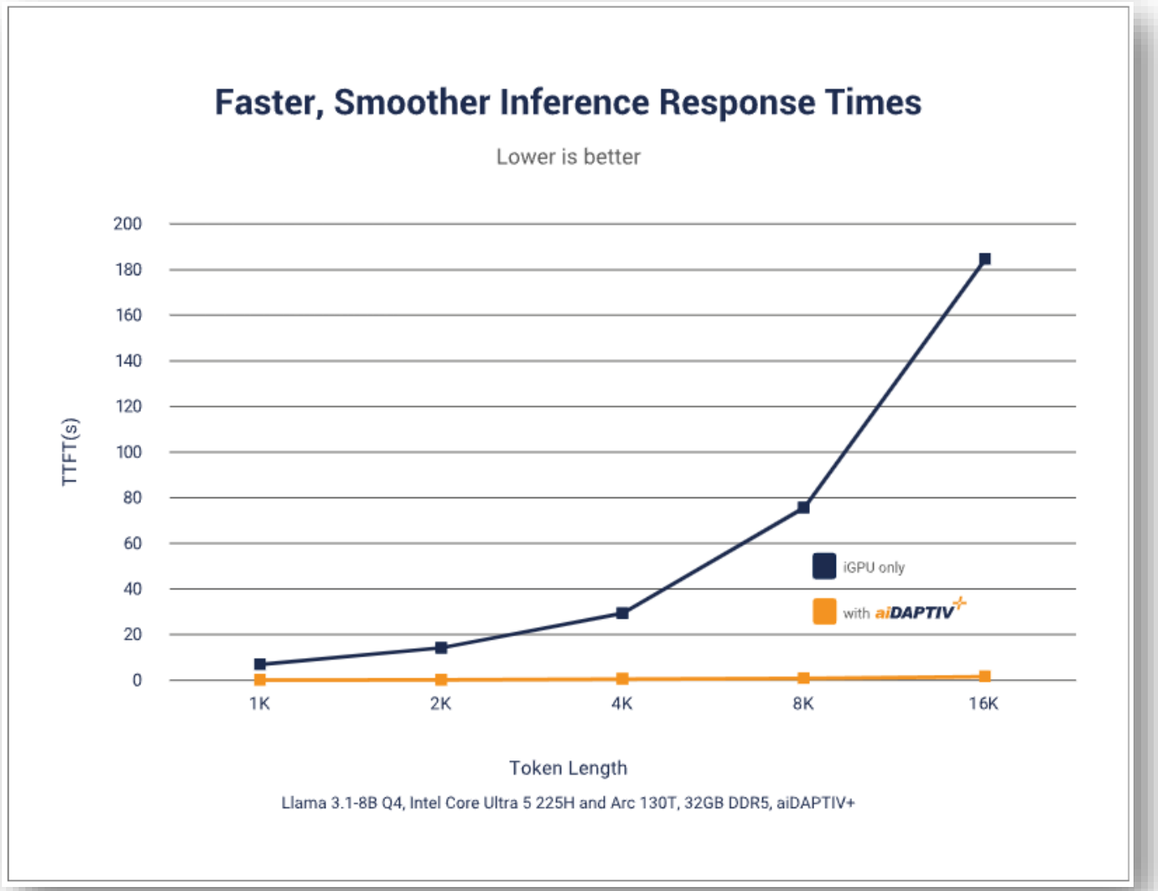


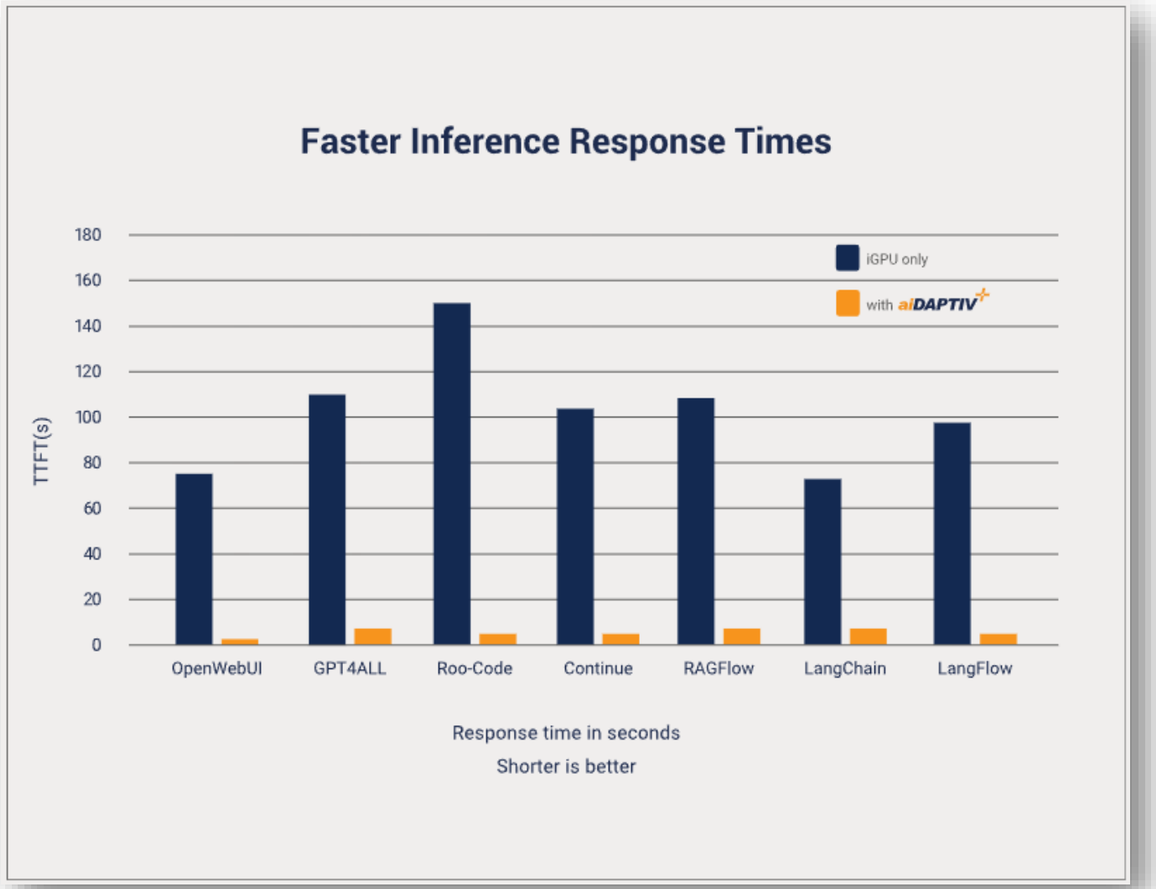
PCP SOLUTIONS

AIT275060T

10X faster inference performance with aiDAPTIVCache (2/2)

Time To First Token vs Token Length



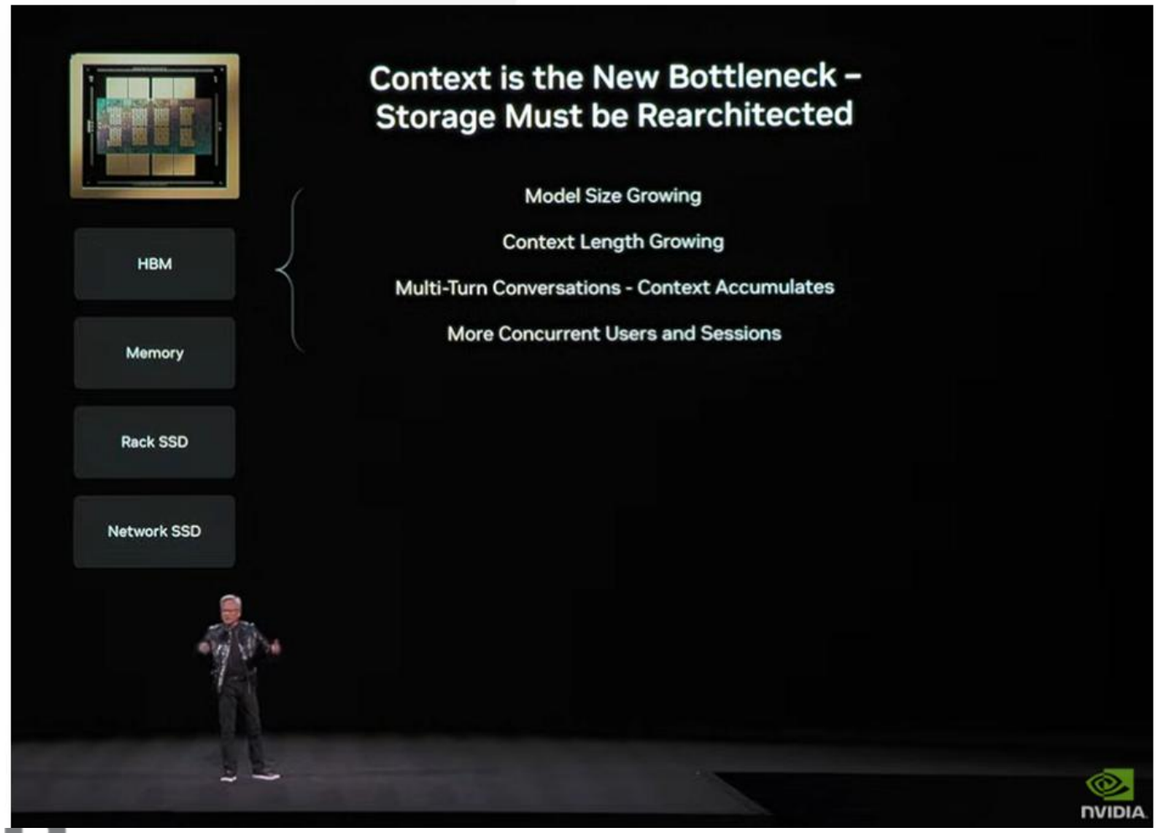


AIT275060T

Quoted from NVIDIA Keynote at CES 2026

KV-Cache, the Cache of AI

How to keep KV Cache is significant important during AI inference stage



"AI doesn't use SQL, AI uses semantic information"

"It creates "temporary knowledge" or "temporary memory" called **KV cache**"

"It's basically **the cache of AI, the working memory of AI**"

- Jensen Huang, CEO of NVidia

PCP SOLUTIONS

AIT275060T

DISCLAIMER

PC Partner Limited does not warrant the accuracy, completeness or reliability of information, materials and other items contained on this document No liability is assumed with respect to the use of the information contained herein. When using this document, users acknowledge that PC Partner Limited will not be liable in any event for any damages arising out of the use of this document.

While every effort has been taken to ensure accuracy and completeness in the preparation of this document, we assume no responsibility for errors or omissions. Products shown in this document are provided for general reference purposes only. All product information herein is subject to change without notice.