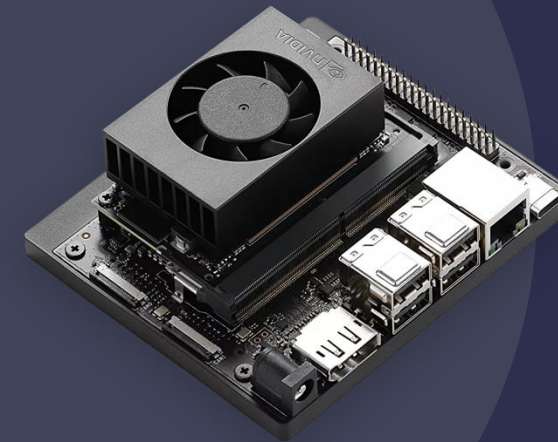# Affordable LLM Training & Inference

What Phison is Announcing for aiDAPTIV+ at GTC 2025

Demo of World's 1st LLM Training Laptop

Maingear laptop PC built for AI powered by Phison aiDAPTIV+

- Full Model Fine-Tuning
- Enhanced Inferencing

Support for NVIDIA Jetson IoT Devices

Edge computing and robotics powered by Phison aiDAPTIV+

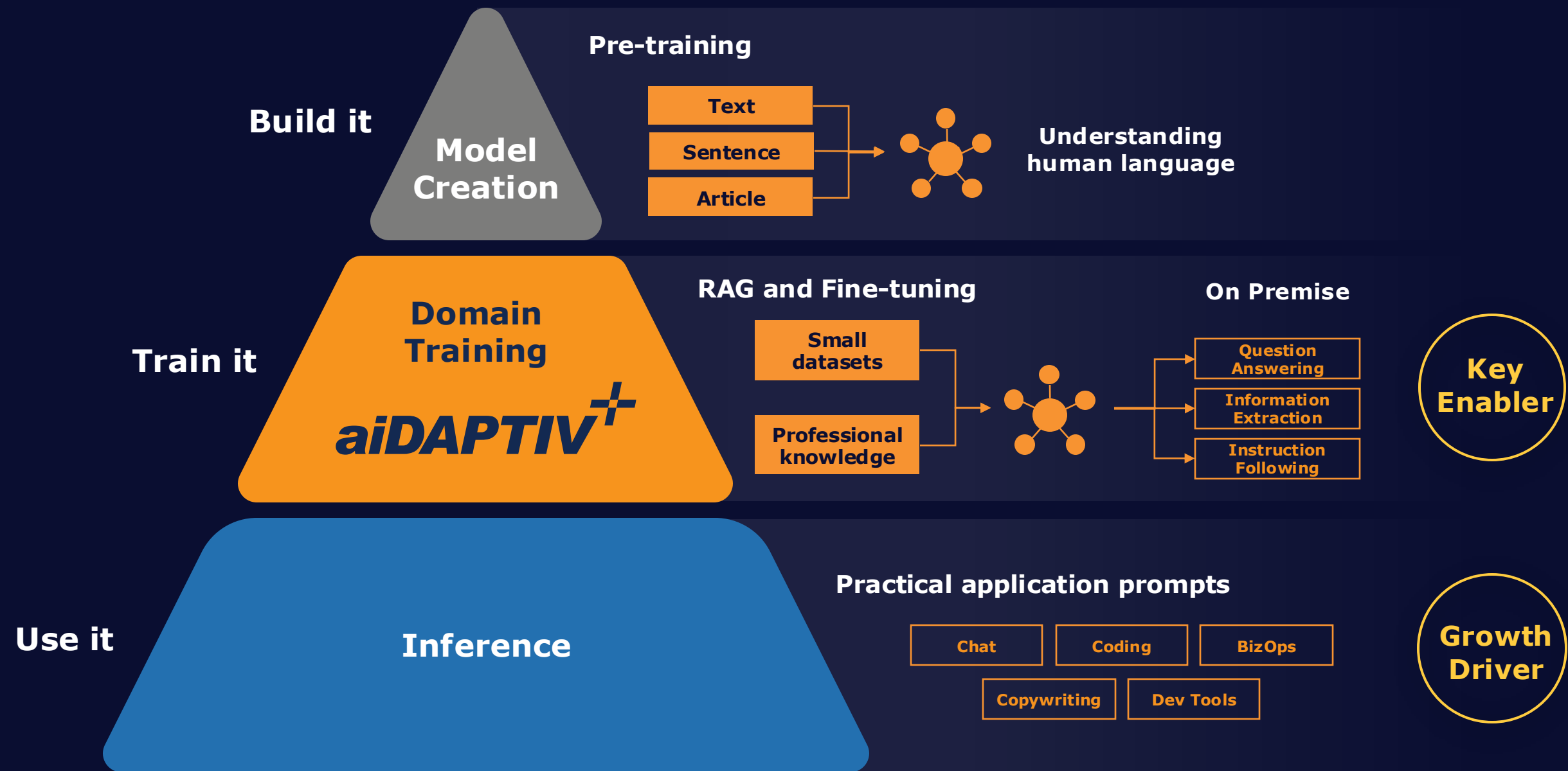- LoRA Model Fine-Tuning
- Enhanced Inferencing

INFERENCE

Enhanced Inferencing

A better user experience

- Longer Token Lengths delivers more accurate results
- Faster Time to First Token recall for quicker research

Today's LLM Market Segment

GPU Cards Requirements

Build it — Model Creation — Pre-training: Text, Sentence, Article → Understanding human language

>1000
Massive GPUs for High Computing Power

Train it — Domain Training — *aiDAPTIV+* — RAG and Fine-tuning: Small datasets, Professional knowledge → On Premise: Question Answering, Information Extraction, Instruction Following — **Key Enabler**

10~100
Depends on Memory Size
(GPU RAM ≥ 20x Model Size)

Use it — Inference — Practical application prompts: Chat, Coding, BizOps, Copywriting, Dev Tools — **Growth Driver**

1
Minimal requirement

Phison aiDAPTIV+ Delivers Affordable Memory Scaling

Current AI Algorithm Architecture

GPU + HBM/GDDR Scaling Limits
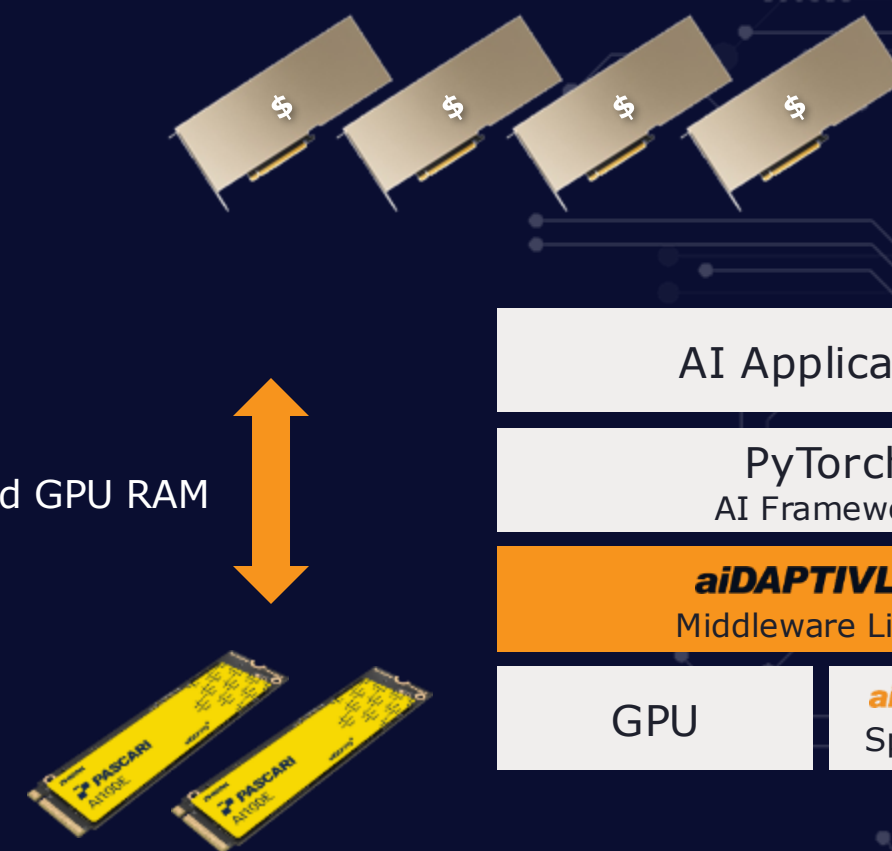
Lots of GPUs
to Run Mainstream LLM

Phison's aiDAPTIV+ AI Algorithm Architecture

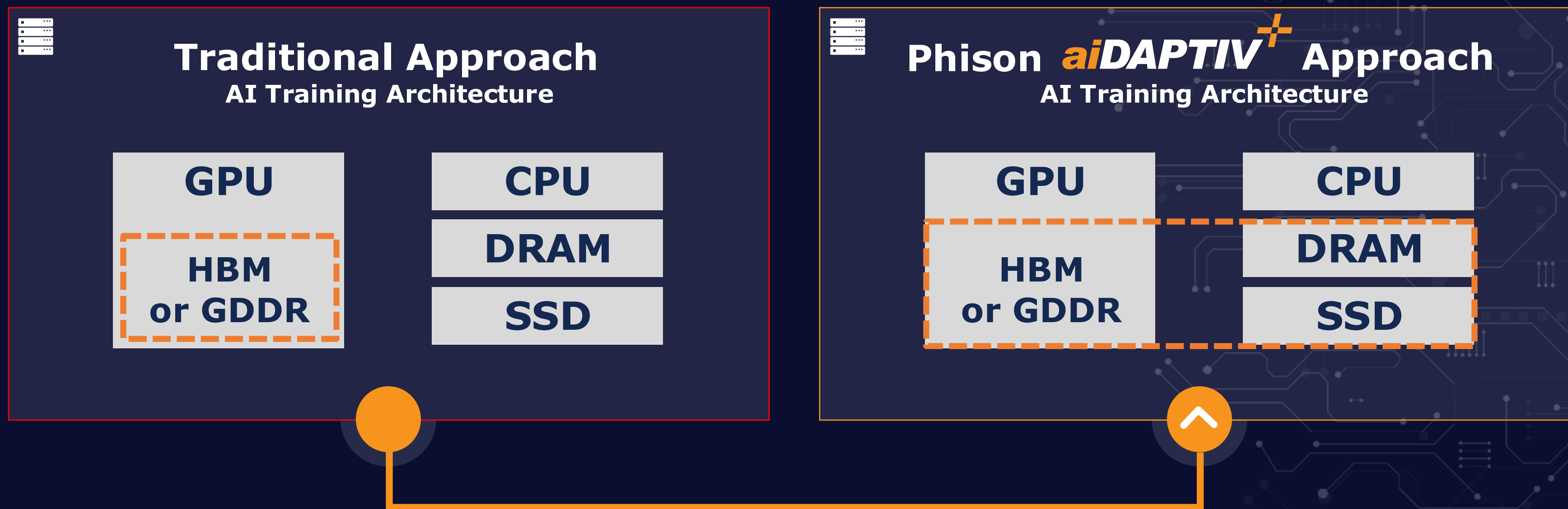**Expands Memory Scaling with NAND Flash**

8-10x Cost Savings

Extend GPU RAM

AI Application
PyTorch AI Framework
aiDAPTIVLink Middleware Library

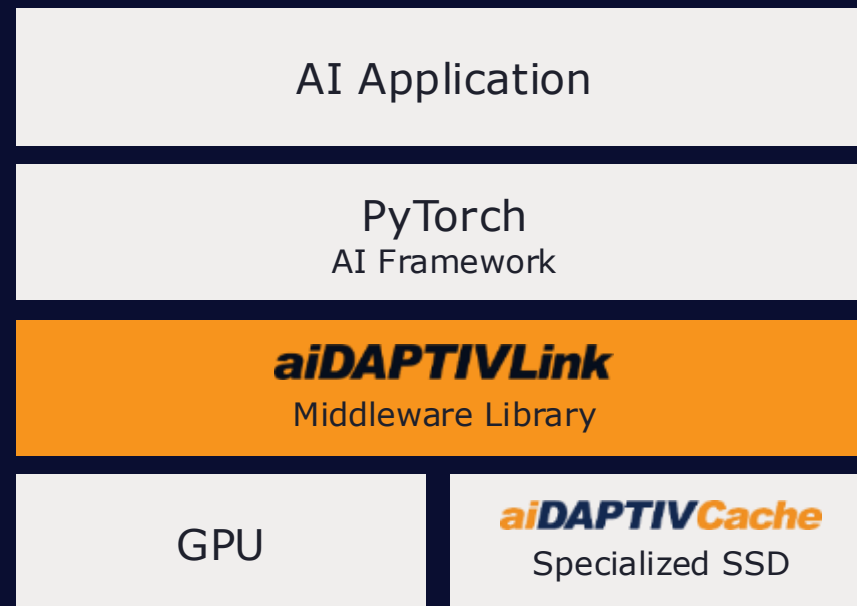| GPU	**aiDAPTIVCache** Specialized SSD

aiDAPTIVCache

1: Based on NAND capacity

How aiDAPTIV+ Does It

Builds a Larger Memory Pool for LLM Training
By Tiering **Expensive** VRAM and **Affordable** Flash Memory

Traditional Approach
AI Training Architecture

GPU		CPU
HBM or GDDR		DRAM
		SSD

Phison aiDAPTIV+ Approach
AI Training Architecture

GPU		CPU
HBM or GDDR		DRAM
		SSD

Extends GPU Memory

Phison aiDAPTIV+ LLM Domain Training Solution

AI Application

PyTorch
AI Framework

aiDAPTIVLink
Middleware Library

GPU | aiDAPTIVCache
Specialized SSD

aiDAPTIVLink

Middleware

Coordinates the swapping between HBM/DRAM and Flash Memory

aiDAPTIVCache

AI-Series SSD Family
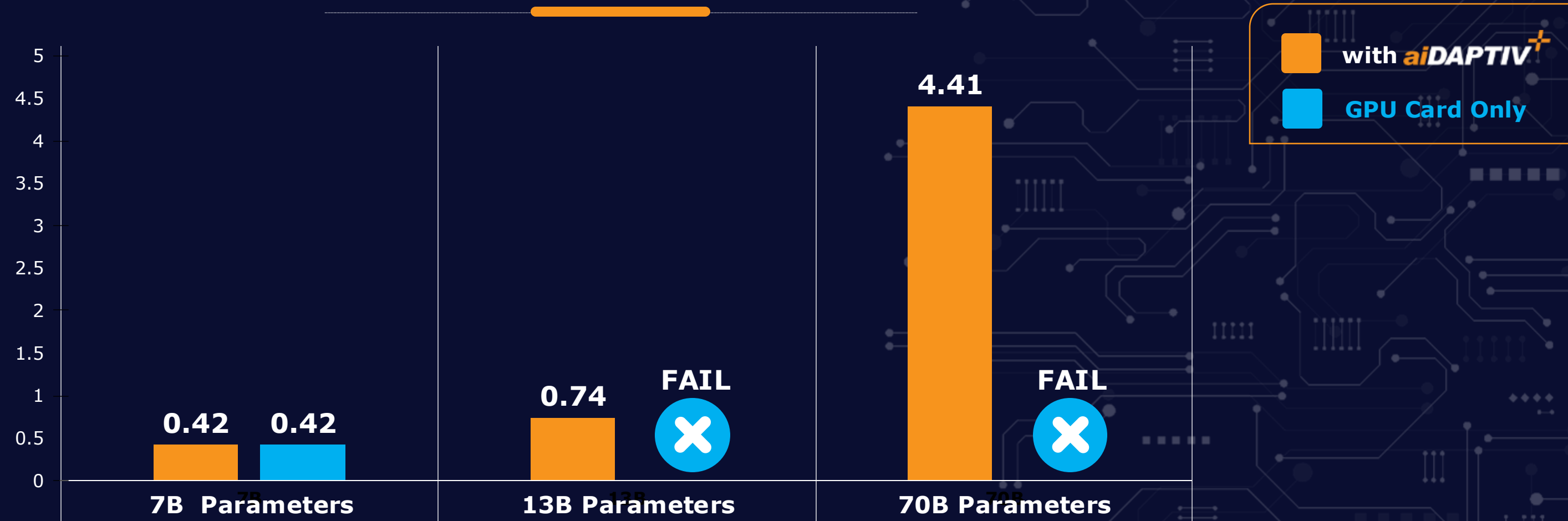
Seamless Integration with VRAM/DRAM

aiDAPTIV+

Pro Suite UI

End-to-End User Experience Spanning Training to Inference

aiDAPTIV+ Workstation Scaling

Makes Possible Fine-Tuning Beyond VRAM Capacity

Single node 4x GPU configuration comparing GPU only and GPU with *aiDAPTIV+*

Legend:
- ■ with *aiDAPTIV+* (orange)
- ■ GPU Card Only (blue)

Chart values:

	7B Parameters	13B Parameters	70B Parameters
with aiDAPTIV+	0.42	0.74	4.41
GPU Card Only	0.42	FAIL	FAIL

Y-axis: 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5

Fixed VRAM Capacity →

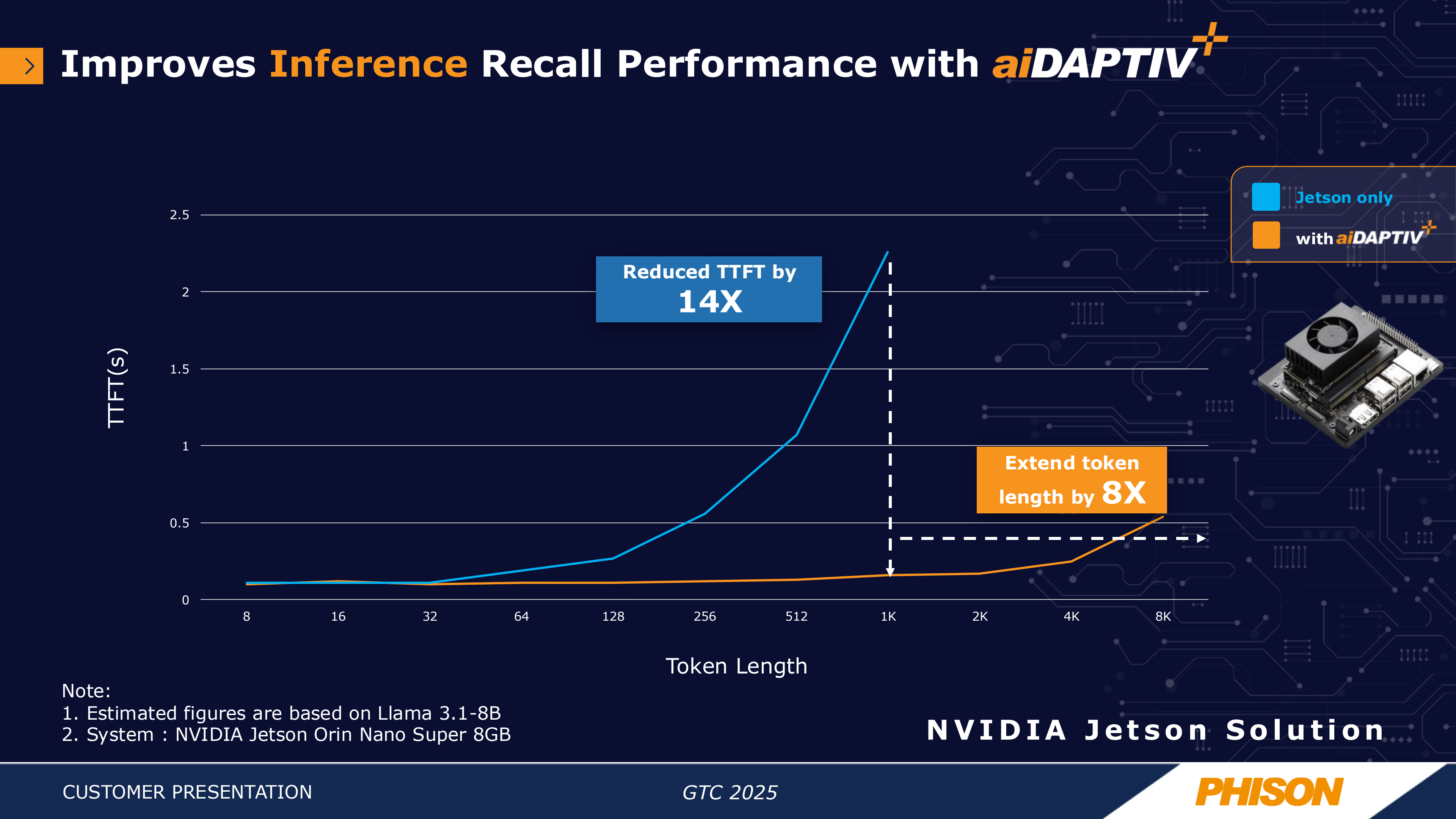|	7B Parameters	13B Parameters	70B Parameters
Training Set Size	140 GB	260 GB	1400 GB
HBM Pool (Usage%)	192 GB (73%)	192 GB (120%)	192 GB (729%)
Minimum GPU Count	4 / 4	4 / 6	4 / 30

Notes:
1. *Scaling is linear based on GPU count and model size*
2. *Training is based on 10M tokens*

System Configuration
- RAM: 512 GB
- GPU: 4x RTX 6000 ADA
- GDDR: 192 GB

Improves **Inference** Recall Performance with *aiDAPTIV⁺*

Jetson only

with *aiDAPTIV⁺*

Reduced TTFT by
14X

Extend token length by 8X

TTFT(s)

2.5

2

1.5

1

0.5

0

8   16   32   64   128   256   512   1K   2K   4K   8K

Token Length

Note:
1. Estimated figures are based on Llama 3.1-8B
2. System : NVIDIA Jetson Orin Nano Super 8GB

NVIDIA Jetson Solution

PHISON

Train LLMs at Any Budget with *aiDAPTIV+*

IoT Device
Up to 64B Parameter
LoRA Model Training
$500-1,000

Laptop PC
Up to 8B Parameter
Full Model Training
$2,000-3,000

Desktop PC
Up to 13B Parameter
Full Model Training
$3,000-4,000

Workstation PC
Up to 100B Parameter
Full Model Training
$5,000-50,000

Server
Up to 405B Parameter
Full Model Training
$50,000+

aiDAPTIV+ The Affordable Path to **1 Trillion Parameter** Training

The largest size model training at 1/30 (<4%) of the cost

Year	Model	Parameters
2023	Llama 2	69 Billion
2024	Llama 3.1	405 Billion
2025	DeepSeek R3	671 Billion
2026	TBD	1 Trillion

Powered by PHISON aiDAPTIV+
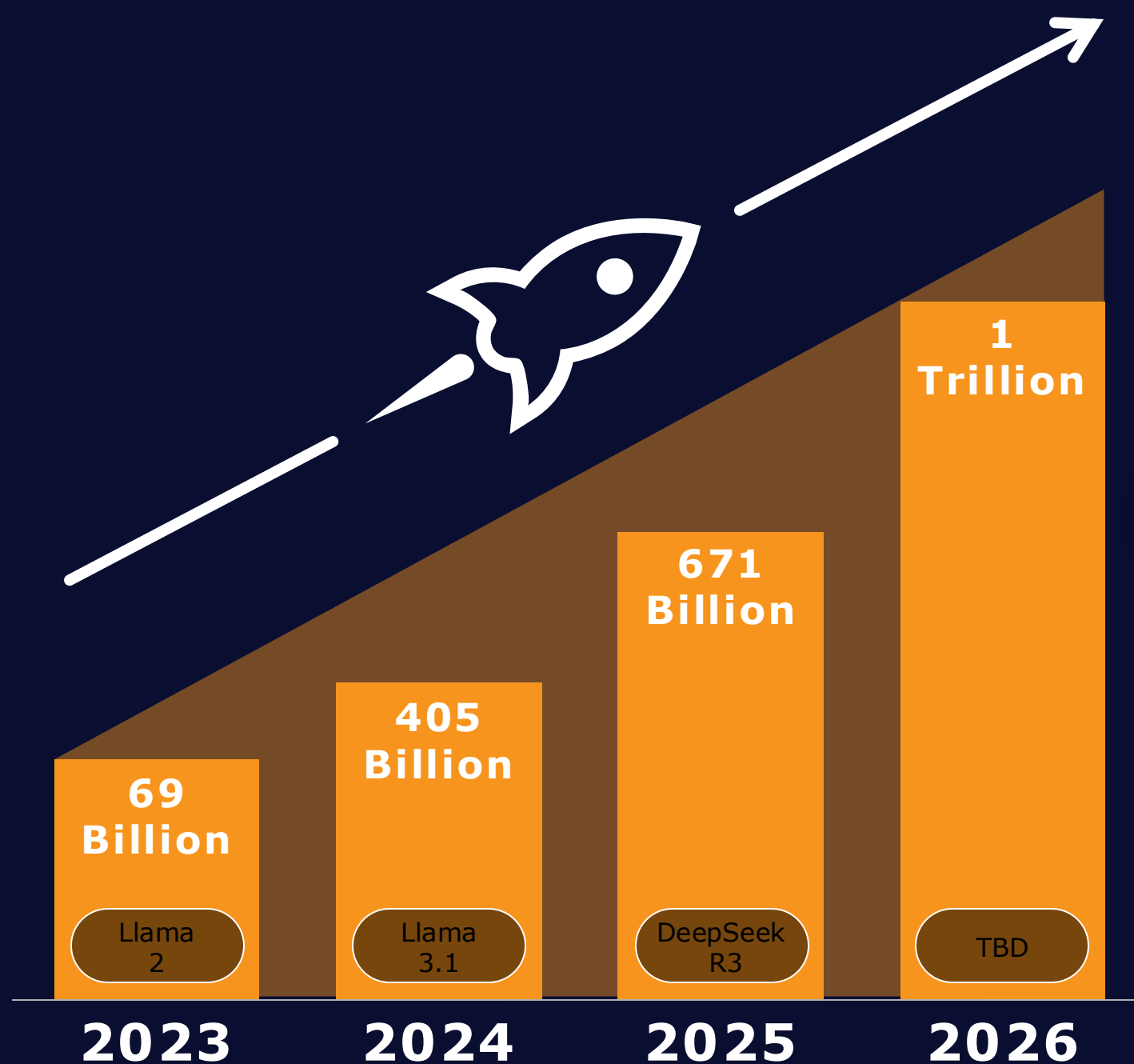
2U AI Training Appliance

- NVIDIA GH200 Superchip x 1
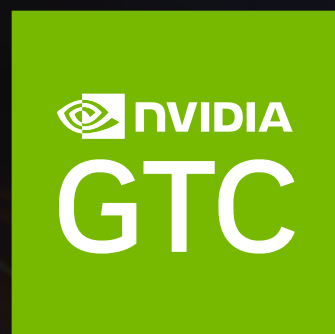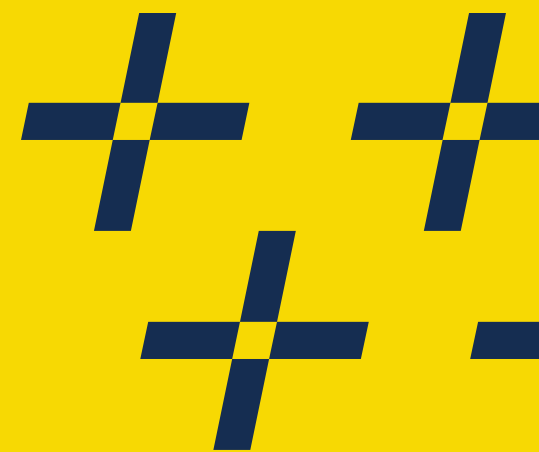- 480GB DRAM
- Phison aiDAPTIV+ MW/SW & aiDAPTIVCache 4TB SSD x 4

$3M

$100K

PHISON

Thank You!

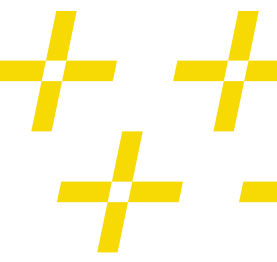# PASCARI

Enterprise SSD

PHISON

Phison Enterprise SSD Product Line-up

High Performance	Data Center	SATA	Boot Drive	Artificial Intelligence
X-Series	**D-Series**	**S-Series**	**B-Series**	**AI-Series**
PCIe 4.0/5.0	PCIe 4.0/5.0	SATA III	PCIe 4.0 / SATA III	PCIe 4.0
Up to 30.72TB	Up to 122.88TB	Up to 15.36TB	Up to 960GB	Up to 8TB
1/3/60 DWPD	0.3/1/3 DWPD	0.4/1/3 DWPD	1 DWPD	Up to 100 DWPD
U.2 / U.3 / E3.S	E3.S / E3.L / U.2 / E1.S / M.2	2.5"	2.5" / M.2	U.2 / M.2

D200V/D205V Series Product Overview
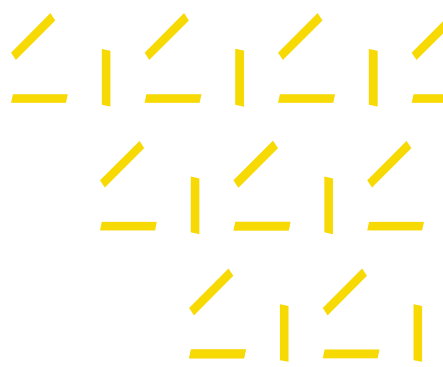
D200V/D205V QLC SSD

Phison's D200V/D205V series utilize QLC technology to meet the high storage density demand driven by AI. With an impressive 122TB per SSD, D205V will help drive the trend of efficient data storage with improved space utilization and reduced power consumption.

Specification	Features	Capacity
• Interface: PCIe 5.0 x 4 • Protocol: NVMe 2.0 • Form Factor: U.2 / E3.S / E3.L • DWPD: 0.3 DWPD • MTBF: 2.5 million hours • Warranty: 5 years	• 3D QLC NAND • Dual-port design • Power loss Protection • Namespaces: 128	**D200V**: • 30.72TB • 61.44TB **D205V**: • 122.88TB

X200 Series Product Overview

X200 Series SSD

The best PCIe Gen5 performance, features, endurance, and economics for enterprise applications. The X200 shows Phison's dedication to developing advanced SSD technology to lead the industry in density, performance, and power efficiency for all mass-capacity storage providers.

Specification

- Interface: PCIe 5.0 x 4
- Protocol: NVMe 2.0
- Capacity: Up to 30.72TB
- Form Factor: U.2 / E3.S
- DWPD: 1 and 3 DWPD
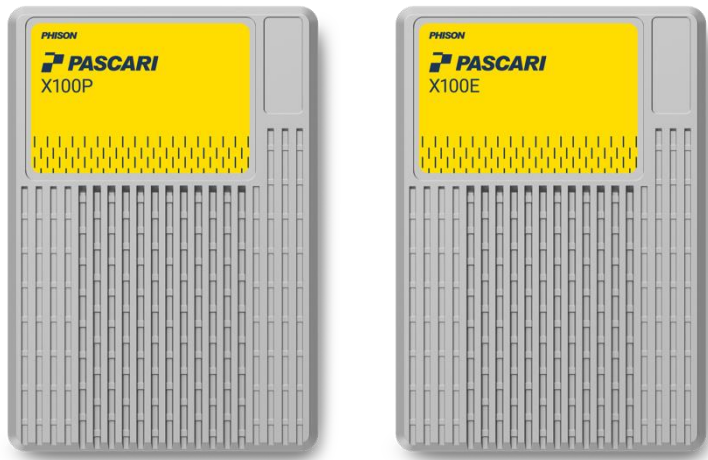- MTBF: 2.5 million hours
- Warranty: 5 years

Features

- Dual-port design
- Ultra-low latency
- Power loss Protection
- MF-QoS
- Namespaces: 128

Performance

Seq. Read	14,800 MB/s
Seq. Write	8,700 MB/s
Ran. Read	3,200K IOPS
Ran. Write	930K IOPS

X100 Series Product Overview

X100 Series SSD

Phison's X100 series SSD is highly customizable and marks the spot for enterprises demanding faster and smarter global infrastructures. Featuring best-in-class performance, the X100 enables enterprises to reduce the total cost of ownership through higher storage density, lower power consumption and higher performance.

Specification
• Interface: PCIe 4.0 x4
• Protocol: NVMe 1.4
• Capacity: 2TB to 32TB
• Form Factor: U.3/U.2, 2.5" x 15mm
• DWPD: Up to 3 DWPD
• MTBF: 2.5 million hours
• Warranty: 5 years

Features
• Dual-port design
• Ultra-low latency
• Power loss protection (PLP)
• End-to-end data path protection (E2EDPP)
• Phison 5th Gen LDPC ECC engine
• Self-encrypting drive (SED) & FIPS 140-3[1]
• Optimized for 24/7 enterprise workload

Performance

Seq. Read	7,400 MB/s
Seq. Write	6,900 MB/s
Ran. Read	1,750K IOPS
Ran. Write	470K IOPS

[1] Based on customer's requirement.

Industry's Most Advanced PCIe 5.0 Enterprise SSD (7.68TB, 1 DWPD)
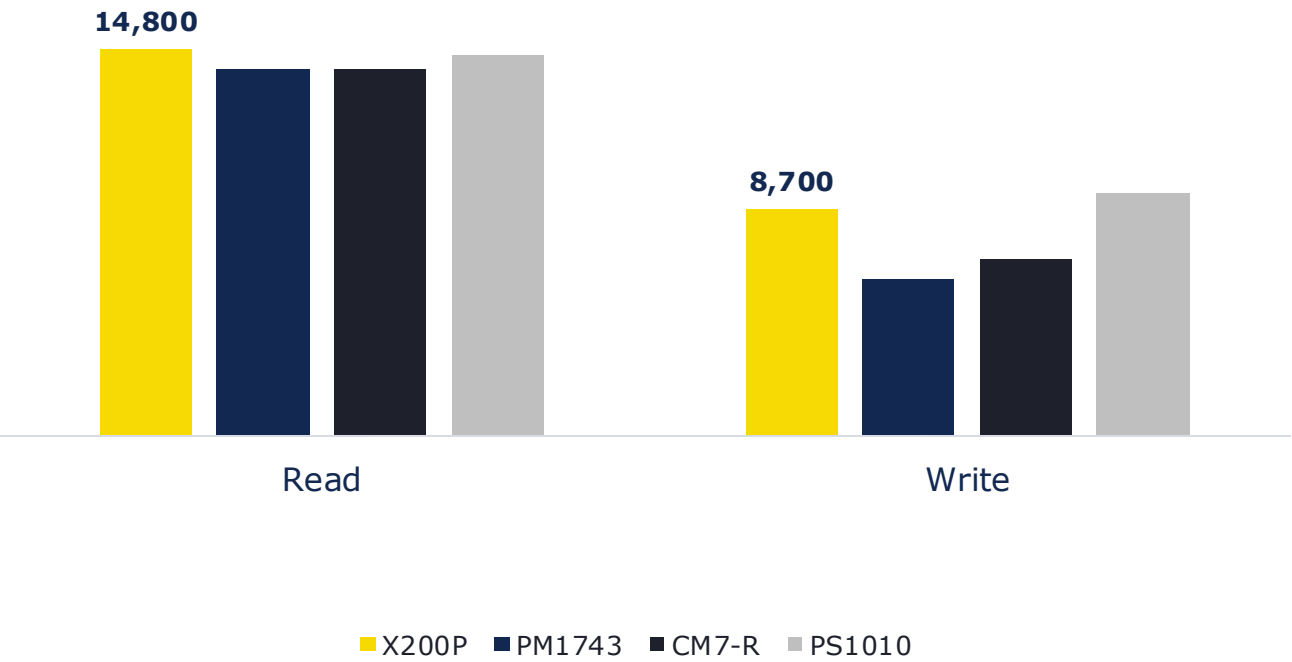
PHISON
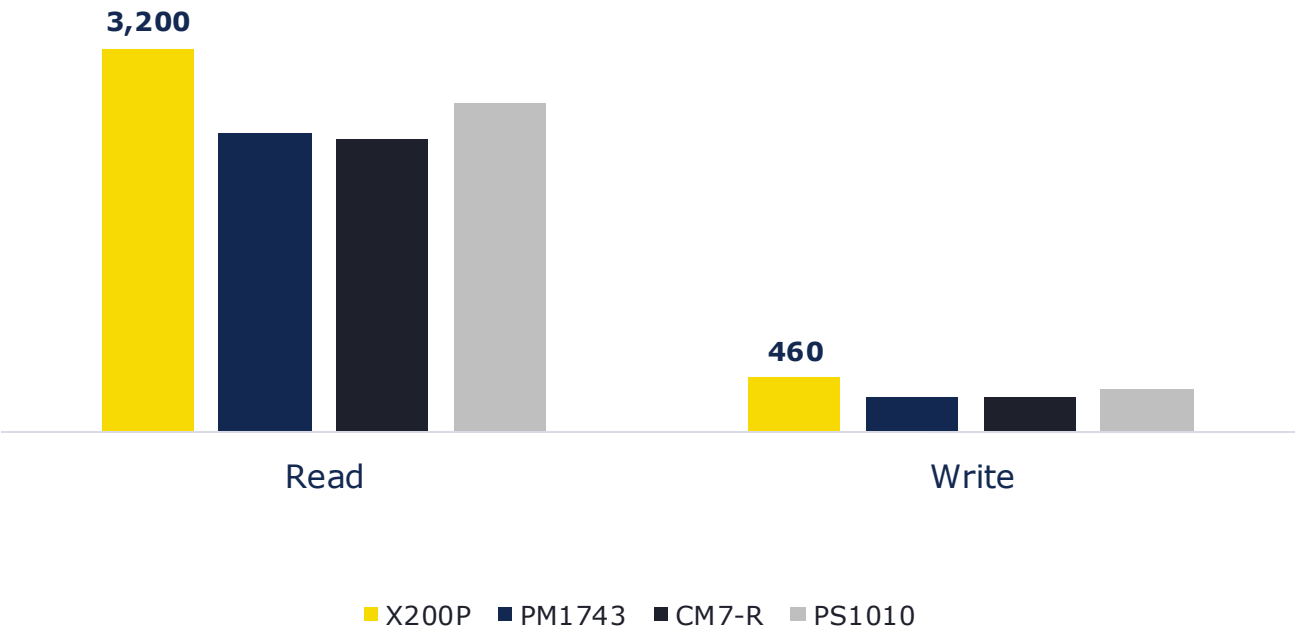X200P

SAMSUNG
PM1743

KIOXIA
CM7-R

SK hynix
PS1010*

Sequential (MB/s)

14,800

8,700

Read · Write

■ X200P ■ PM1743 ■ CM7-R ■ PS1010

4K Random (KIOPS)

3,200

460

Read · Write

■ X200P ■ PM1743 ■ CM7-R ■ PS1010

*PS1010 performance is not based on 7.68TB

PASCARI · GTC 2025 · PHISON