NVIDIA GTC

PHISON

aiDAPTIV+

Affordable LLM Training & Inference

Executive Summary

Problem Statement

- High cost of entry and scaling
- Lack of control and data privacy
- LLM training not widely available

Target Market

- SMB – Small budgets & infrastructure
- Regulated, privacy conscious
- Need domain training of LLMs

Phison's aiDAPTIV+ Solution Enables…

- **Up to 405B parameter models on-prem**
- **IoT to PCs to Workstations to Servers**
- **8x Cost and Power Reduction**

Current
AI Training Architecture

GPU
HBM

CPU
DRAM
SSD

Phison aiDAPTIV+
AI Training Architecture

GPU
HBM or GDDR

CPU
DRAM
SSD

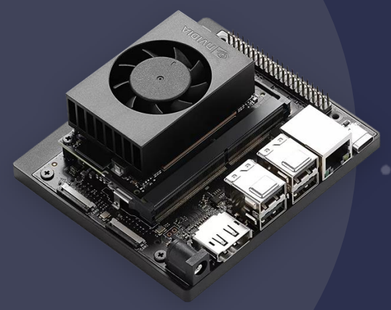
Extends GPU Memory

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

Quick Facts About Phison

Experience

23+

Years

$1.5B

2023 Revenue

80%+

R&D Expense / OPEX

4000+ HC

75% R&D

18%+

Global SSD Market Share

Business Focus

- Enterprise (SSD, Signal IC)
- AI
- Industrial
- Client

1. Phison Technical Marketing Apr 2024 Incl. Enterprise and Client

Phison Trusted by Highly Respected Companies

Enterprise Customers

SEAGATE — *Partnered with Phison to supply their Enterprise SSD brand*

AMPERE · Barracuda · ddn

FORTINET · FUJITSU · Hitachi Vantara

ixsystems · QNAP · QSAN

SPECTRA · SUPERMICRO

+More Qualifications In Process

NAND Wafer/SSD Controller Partners

KIOXIA

Micron

Kingston TECHNOLOGY

Western Digital

SK hynix

Space Missions Require: Zero Failures

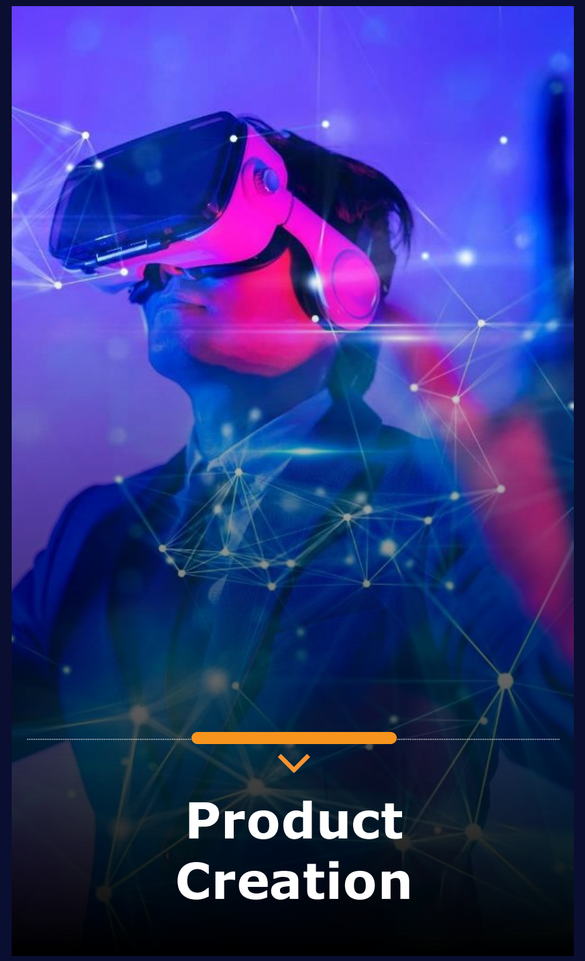
Mars Perseverance Rover Depends on Phison's SSD

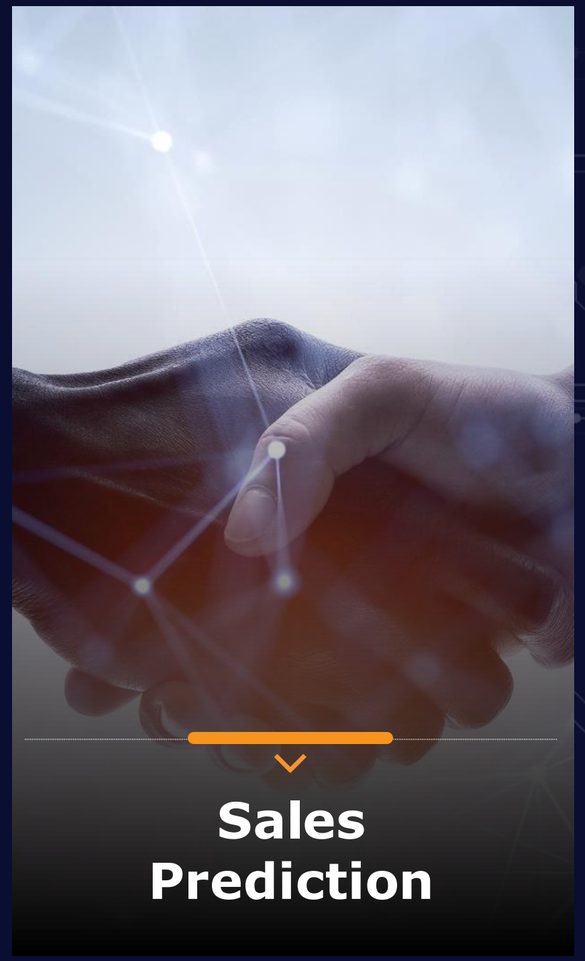
2 Missions to the ISS and Counting

New Missions to Lunar Surface

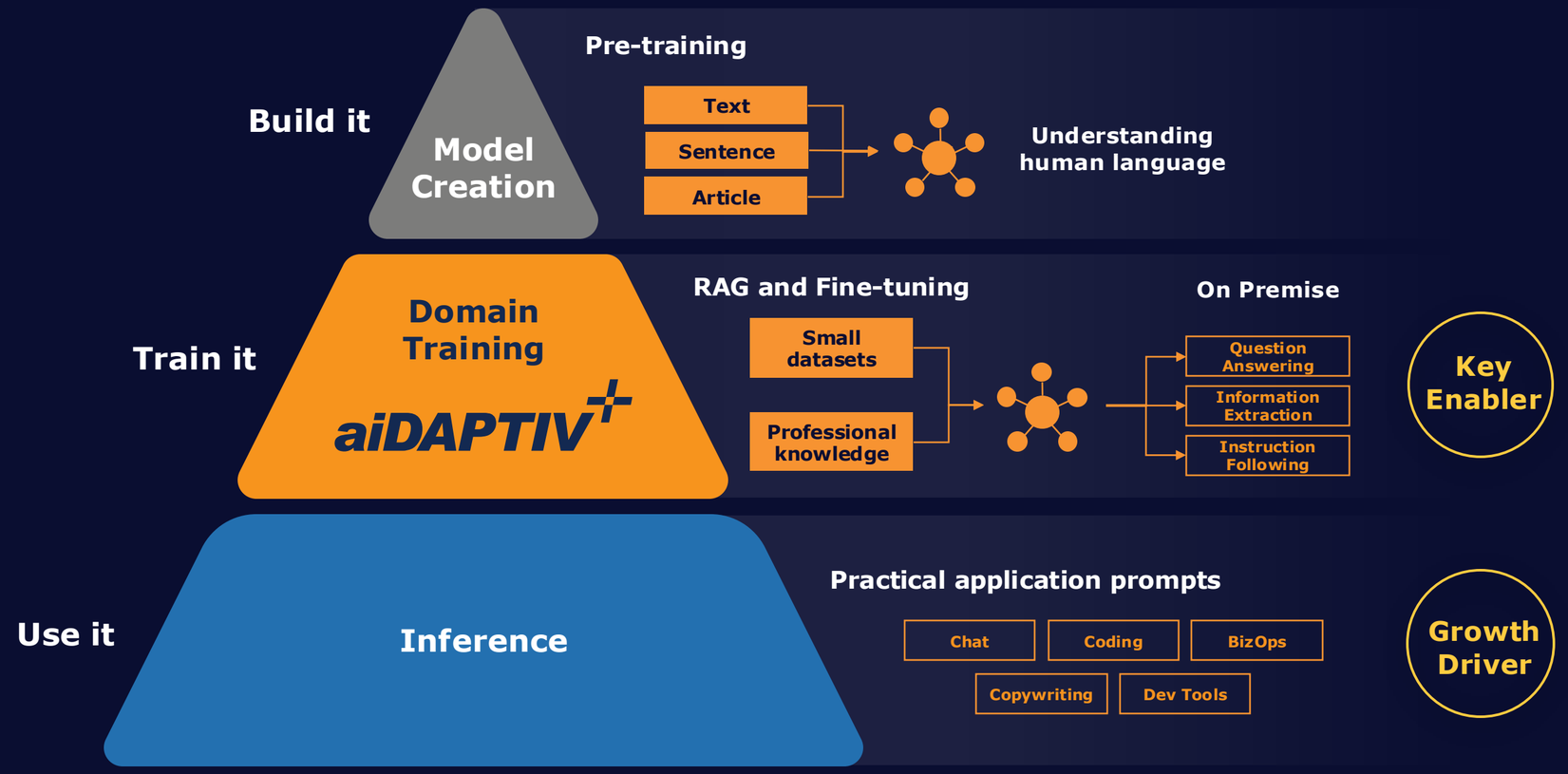
Corporations Need AI

AI isn’t just for individuals; more companies are using it to improve processes, marketing predictions, and business operations…





Today's LLM Market Segment



GPU Cards Requirements

>1000

Massive GPUs for High Computing Power

10~100

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

1

Minimal requirement

Technical Constraints for Generative AI On-Premises

Fine-Tuning

- Rapid growth in size of models
  - Insufficient memory capacity
  - Difficulty in scaling
- High machine costs slows adoption

Inference

- Insufficient memory for tokens leading to:
  - Limited context for chat and prompt
  - Slow responses hurt user experience

Skills Constraint for AI Adoption On-Premises

> Only a handful of companies have the infrastructure and the resources in place to be able to train their LLMs.
>
> Davit Buniatyan
> CEO, Activeloop

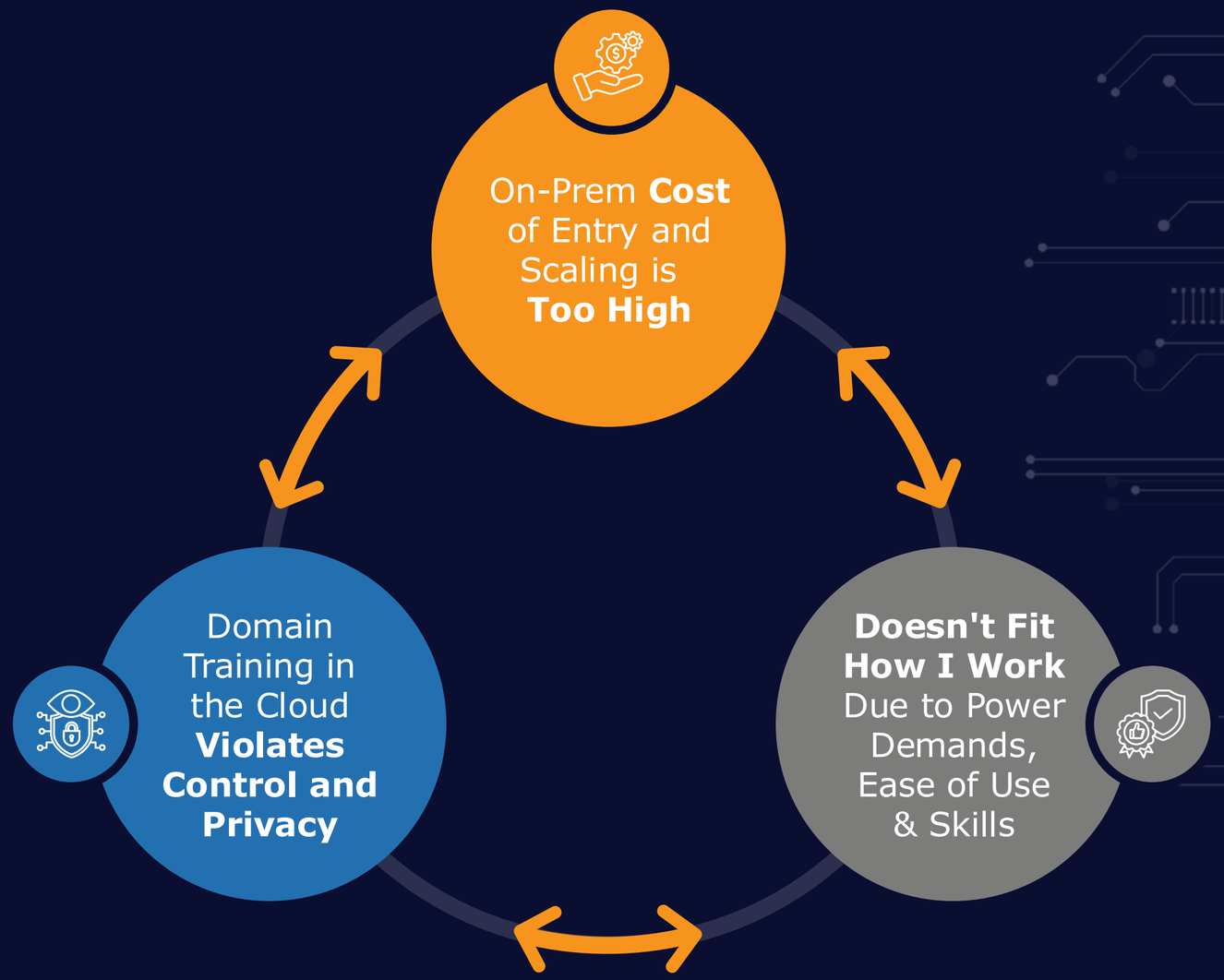
Source: Lack of LLM Developers Impacting AI Ecosystem itprotoday.com

Lack of LLM Developers Impacting AI Ecosystem

Demand for LLM developers is growing exponentially, but the pace at which they have been educated has not kept up.

1. **LLM deployments are limited by the lack of skilled LLM developers.**
2. **These developers lack affordable infrastructure on which to learn.**

Challenges for Organizations Who Want to Train LLMs



Phison aiDAPTIV+ Solves this Dilemma

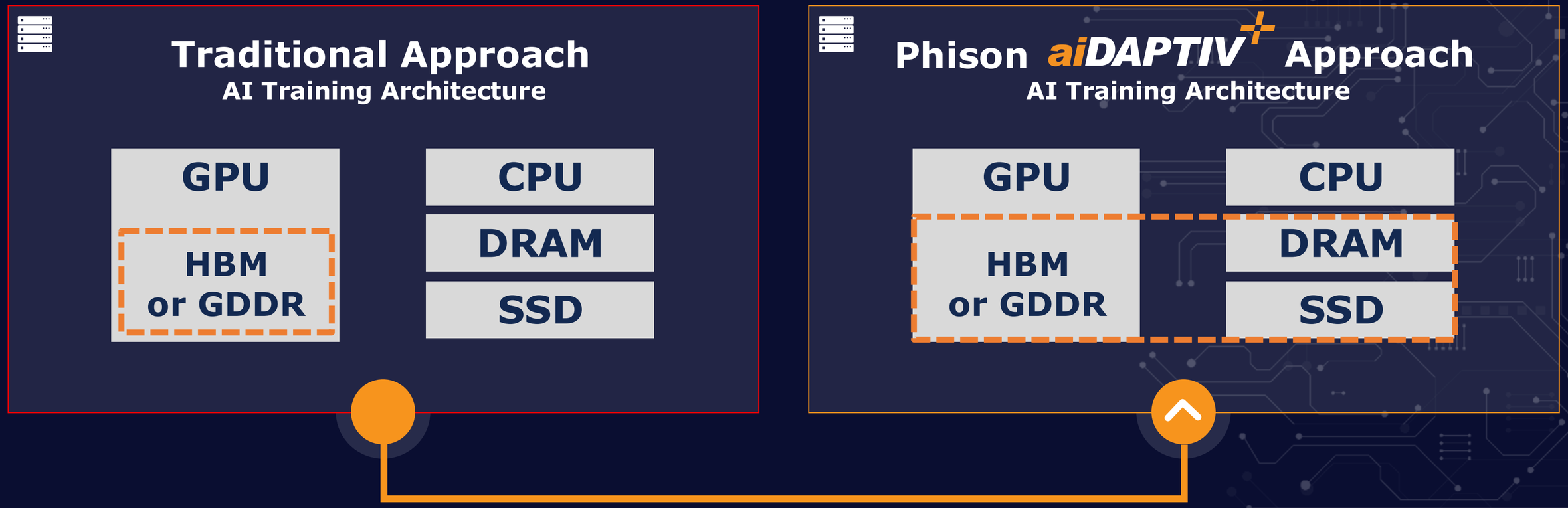
Makes AI Affordable by 8X Reduction in Cost and Power

Fits Home, Office or Classroom for Domain Training

Provides Full Control and Privacy of Your Data & IP

How aiDAPTIV+ Does It

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



Extends GPU Memory

Phison aiDAPTIV+ LLM Domain Training Solution

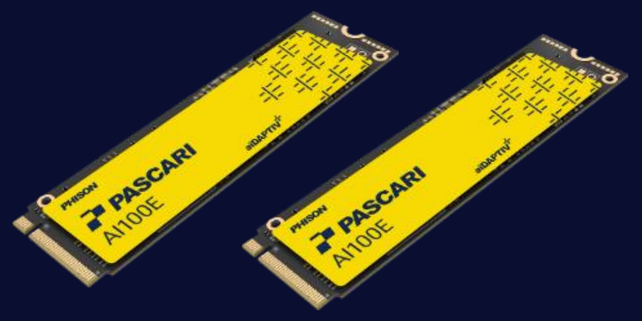
AI Application

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

aiDAPTIVLink Memory Management Middleware

Expands Capacity to Affordably **Fine-Tune** Growing Models

aiDAPTIVLink

Memory management middleware offloads fine-tuning data to aiDAPTIVCache SSD:

- Expands model size capacity
- Frees up GPU VRAM for handling more complex AI fine-tune processing

BENEFITS

- Transparent drop-in solution
- No need to change your AI application

High-Endurance Specialized SSD for LLM Training

aiDAPTIVCache

Pascari AI-Series SSD Family
@ 100 DWPD* / 5 years

*DWPD = Drive Writes Per Day
(The number of times per day an SSD can be filled without wearing out)

24/7 Model Training Demands

Pro GPU	4x 6000 ADA (Workstation)	4x W7900 Pro (Workstation)	8x 6000 ADA (Server)	8x W7900 Pro (Server)
24/7 Training DWPD*	15.71	16.96	31.42	33.92

Enterprise GPU	8x H100 (Server)	8x MI300 (Server)	8x Gaudi2 (Server)	8x Gaudi3 (Server)
24/7 Training DWPD*	61.39	67.84	30.70	63.29

aiDAPTIVCache's endurance exceeds even the harshest model training demands

Ease of Use with aiDAPTIV+ Pro Suite

Optimized End-to-End User Experience Spanning Data Ingest to Inference

aiDAPTIV+

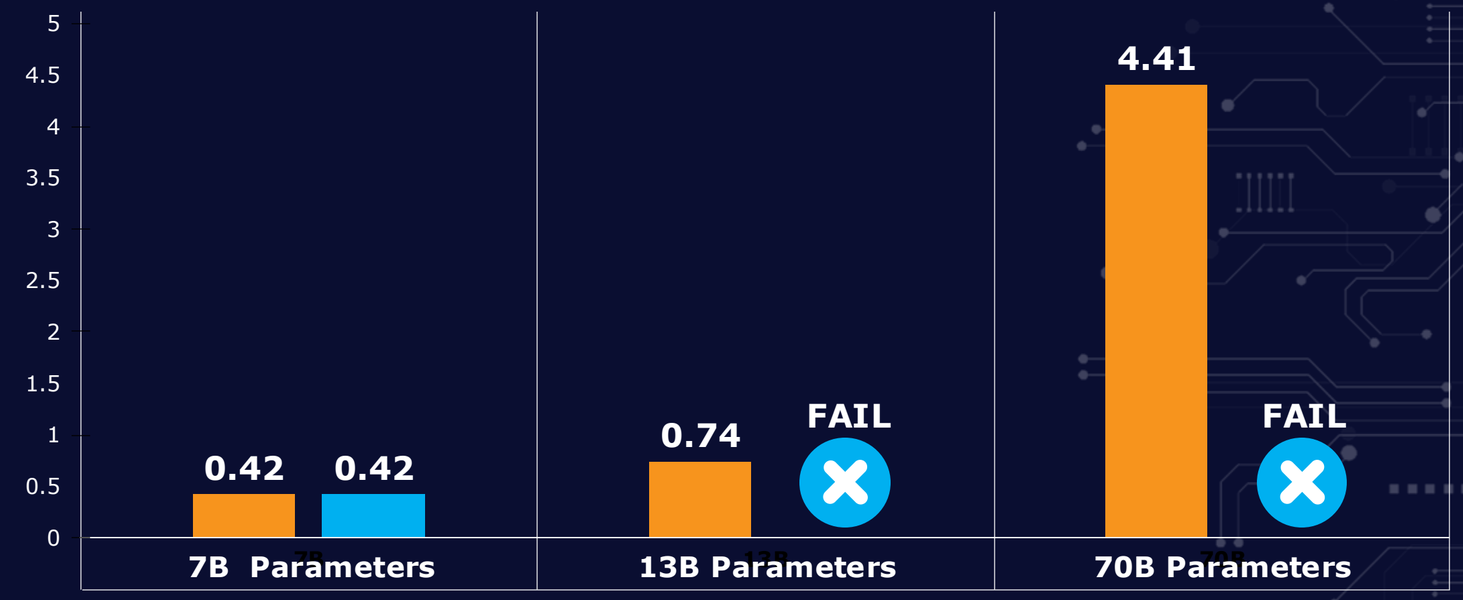
Pro Suite UI

Data Ingest → RAG → Fine Tune → Monitor → Validate → Inference

aiDAPTIV+ Workstation Scaling

Makes Possible Fine-Tuning Beyond VRAM Capacity

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



	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

Fixed VRAM Capacity →

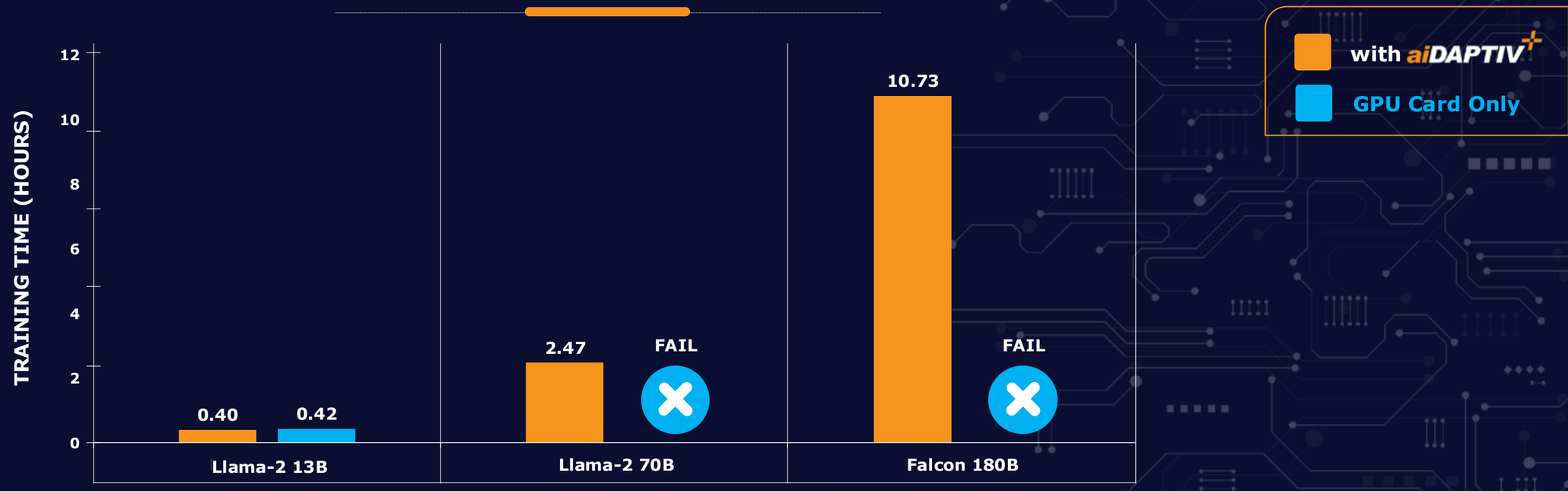
Notes:
1. *Scaling is linear based on GPU count and model size*
2. *Training is based on 10M tokens*

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

aiDAPTIV+ Server Scaling

Makes Possible Fine-Tuning Beyond VRAM Capacity

Single node 8x GPU configuration comparing GPU and GPU with aiDAPTIV+



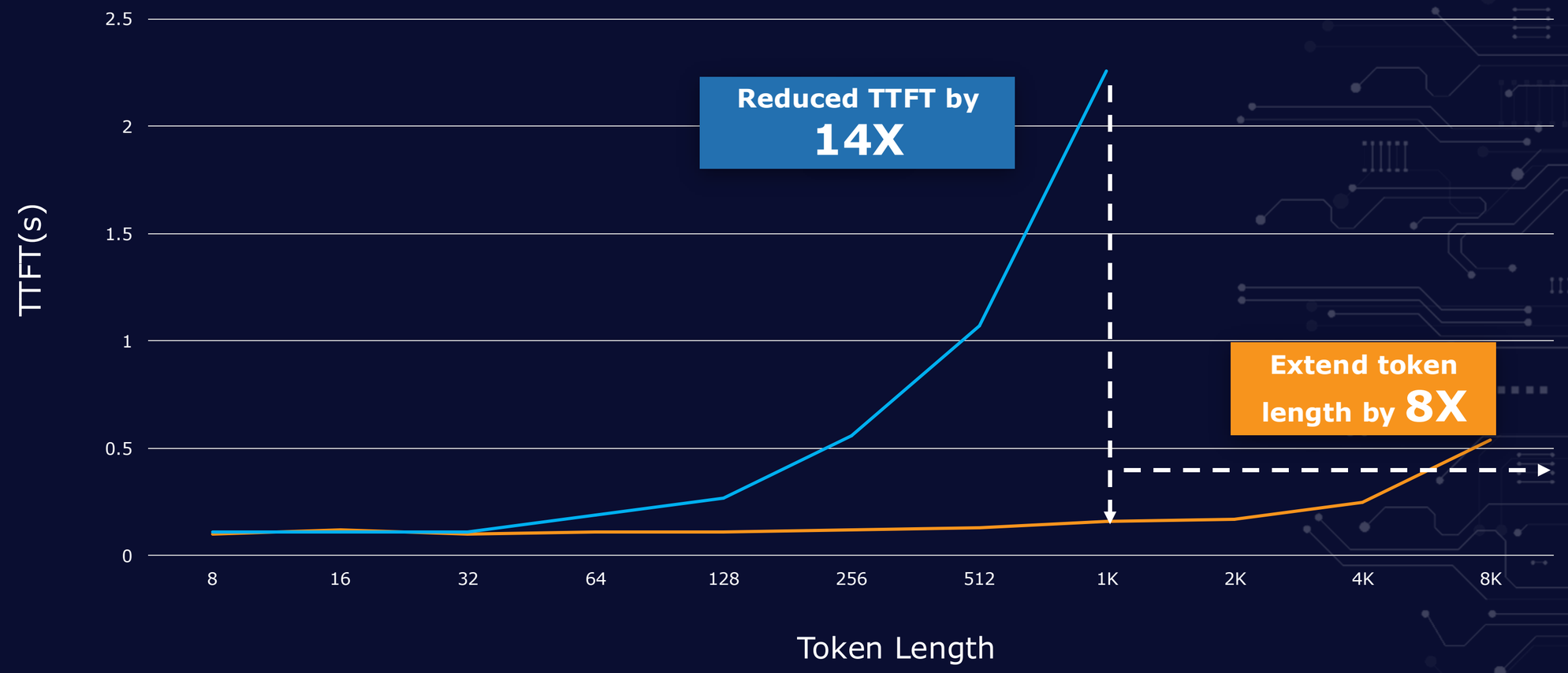
	Llama-2 13B	Llama-2 70B	Falcon 180B
Model Size (Training)	260 GB	1400 GB	3600 GB
HBM Pool (Usage%)	384 GB (68%)	384 GB (365%)	384 GB (938%)
Minimum GPU Count	8 / 8	8 / 30	8 / 75

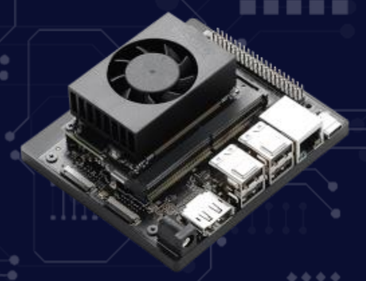
Fixed VRAM Capacity → HBM Pool (Usage%)

Note: Scaling is linear based on GPU count and model size

System Configuration
- RAM: 512 GB
- GPU: **8x RTX A6000**
- GDDR: 384 GB

Improves Inference Recall Performance with aiDAPTIV+





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

NVIDIA Jetson Solution

aiDAPTIV+ "Teaching" PC for LLM Domain Training

Powered by PHISON aiDAPTIV+

aiDAPTIV+ AI Teaching PC

"Learn how to train an LLM beyond just using an LLM...

...and make your AI PC actually useful"

Universities, Researchers, Students, Developers and Enthusiasts

- Universities need available platforms for teaching access
- Researchers and Software Developers need to keep up with the latest AI technology
- Students and Enthusiasts want to learn AI

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

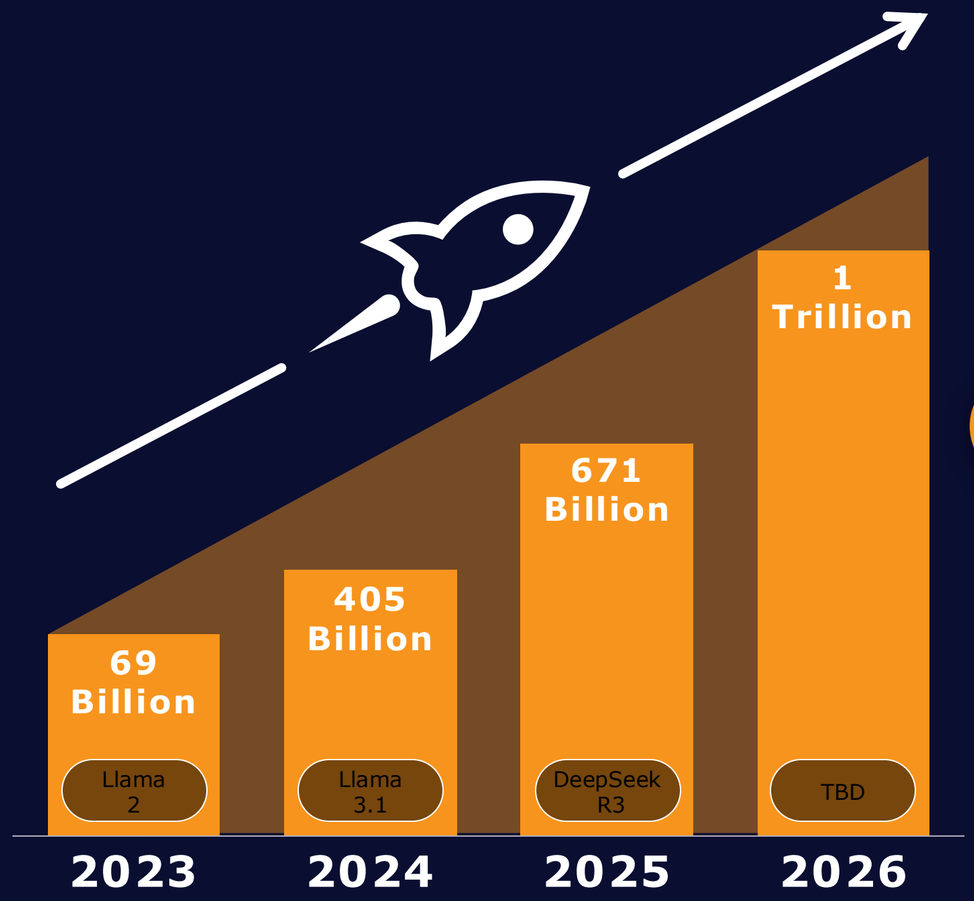
Desktop PC
Up to 13B Parameter
Full Model Training
$3,000-4,000

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



2U AI Training Appliance

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

$3M

$100K

Two Ways to Fine-Tune AI models

On-Prem/Edge

Challenges

Current All GPU/VRAM AI Equipment Too Expensive

Cloud-Based

Challenges

Data Privacy

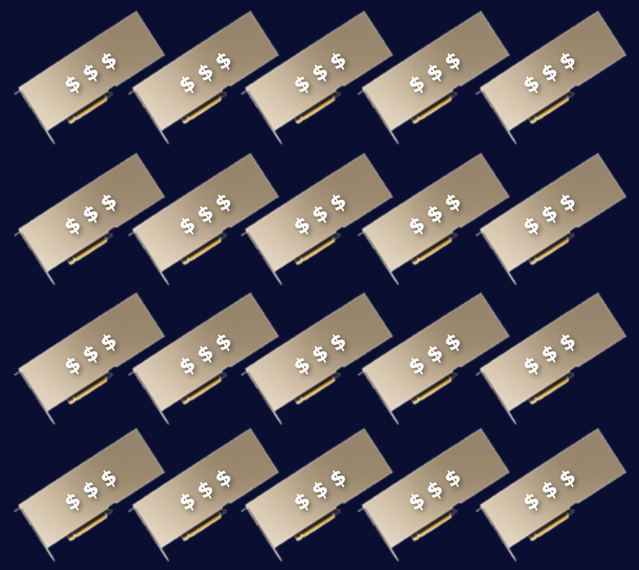
Unpredictable/Unlimited Costs

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

aiDAPTIVCache

AI Application

PyTorch
AI Framework

aiDAPTIVLink
Middleware Library

GPU

aiDAPTIVCache
Specialized SSD

1: Based on NAND capacity

LLM Domain Training Cost: On-Premises vs Cloud

Name (sorted by cost)	GPU	Node @ Cost/hr[1]	Duration (hour)	Session Cost	tk/sec	$/min	$/Mtk[3]	5 Year Training Cost		
								15/month	30/month	120/month
Phison: aiDAPTIV+ v1.03	4x 6000ada	1	4.34	$6.94[2]	640	$0.03	$0.69	$50,000		
Generic: GPU Only	4x 6000ada	8	0.54	$6.94[2]	5,120	$0.21	$0.69	$400,000		
Azure: ND96isr H100 v5	8x H100	3 @ $98.32	0.36	$106	7,630	$4.92	$10.62	$95,567	$191,134	$764,536
AWS: p5.48xlarge	8x H100	3 @ $98.32	0.36	$106	7,630	$4.92	$10.62	$95,567	$191,134	$764,536
Azure: ND96asr A100 v4	8x A100	3 @ $27.19	0.80	$65	3,469	$1.36	$6.53	$58,730	$117,461	$469,843

Key Points:

1. **Recommended training sets** for high quality fine-tune domain training (10 million tokens)
2. The main benefit of **higher tier** GPU for LLM is **faster execution**, but at a **huge premium**
3. Cloud Rental is a **recurring expense** that greatly **exceeds the CAPEX**

1: *Online Pricing July 2024, VM only*
2: *Assuming 120 sessions / month over 5 years*
3: *Mtk represents Million tokens*

Thank You!

NVIDIA GTC | PHISON

PASCARI

GTC 2025

Enterprise SSD

PHISON

Phison Enterprise SSD Product Line-up

High Performance

X-Series

PCIe 4.0/5.0
Up to 30.72TB
1/3/60 DWPD

U.2 / U.3 / E3.S

Data Center

D-Series

PCIe 4.0/5.0
Up to 122.88TB
0.3/1/3 DWPD

E3.S / E3.L / U.2 / E1.S / M.2

SATA

S-Series

SATA III
Up to 15.36TB
0.4/1/3 DWPD

2.5’’

Boot Drive

B-Series

PCIe 4.0 / SATA III
Up to 960GB
1 DWPD

2.5’’ / M.2

Artificial Intelligence

AI-Series

PCIe 4.0
Up to 8TB
Up to 100 DWPD

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

- 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

Features

- 3D QLC NAND
- Dual-port design
- Power loss Protection
- Namespaces: 128

Capacity

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

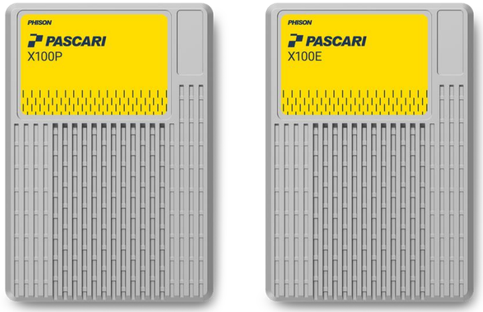
Features

- 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

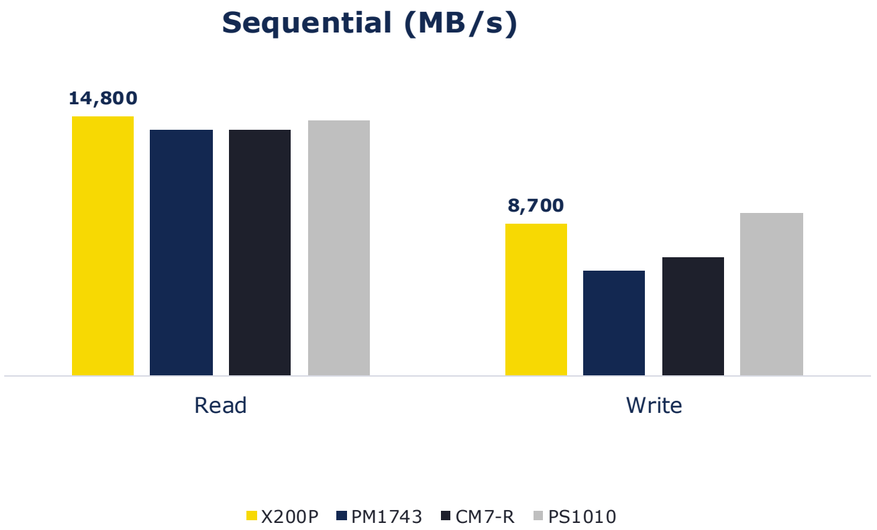
PM1743

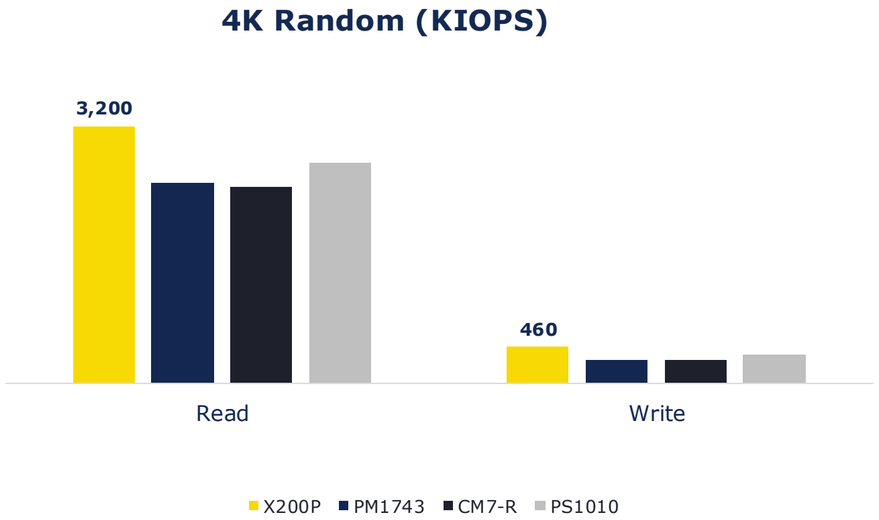
KIOXIA

CM7-R

SK hynix

PS1010*





*PS1010 performance is not based on 7.68TB