NVIDIA GTC | PHISON

AI Training Innovation

aiDAPTIV+

Scaling Boundaries, Not Budgets

Executive Summary

Problem Statement

- LLM training not easily accessible
- Cost of entry is high
- GPU memory cannot scale fast enough

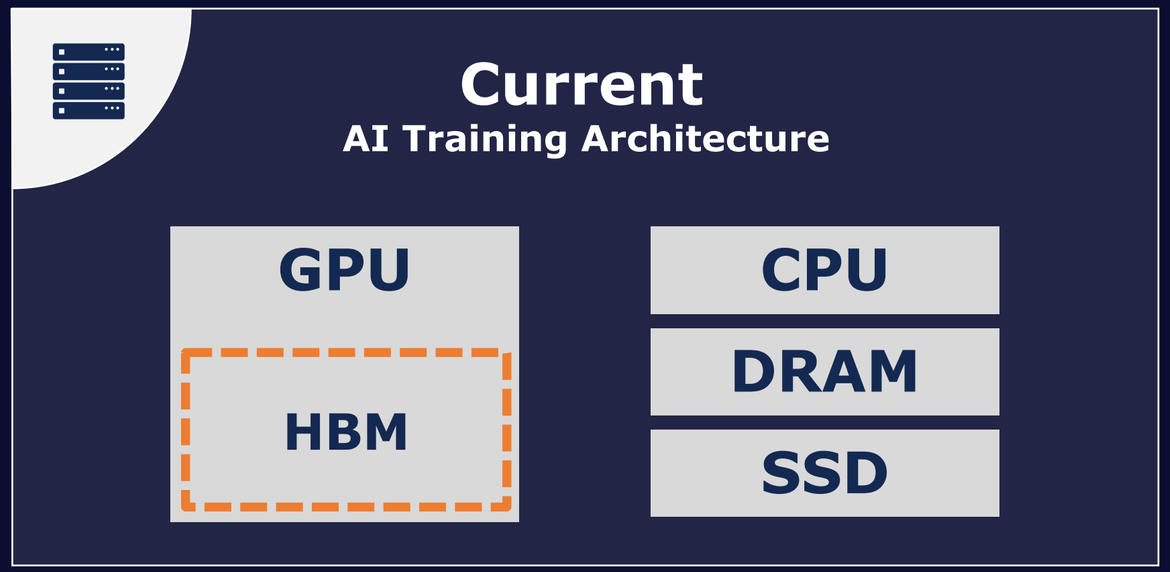
Target Market

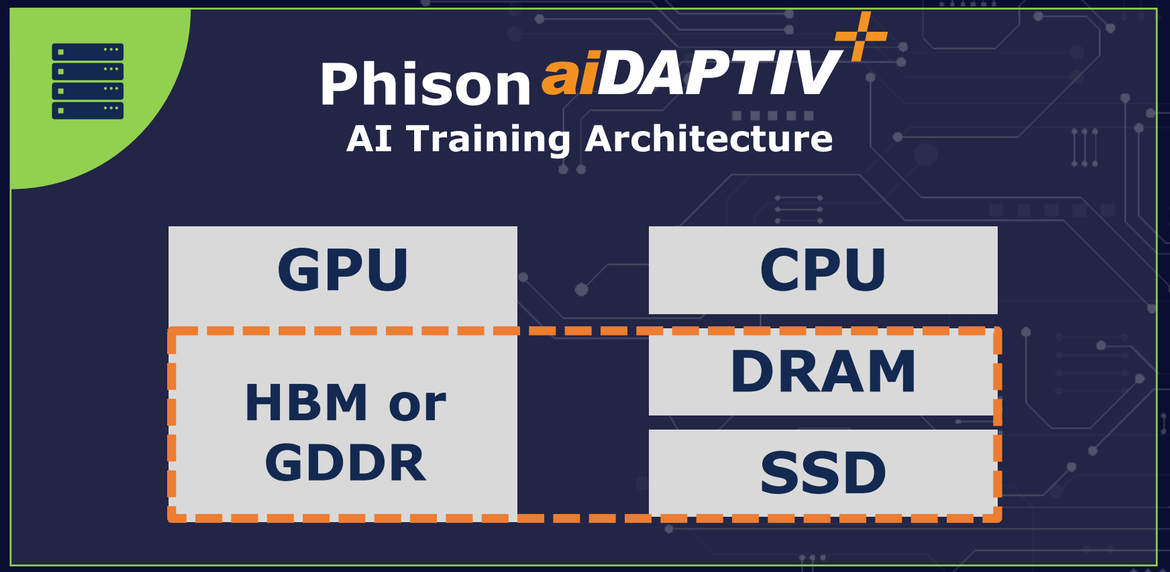
- SMB
- Domain training / Fine tuning
- Privacy Conscious
- Smaller Infrastructure (15a Circuit)

Phison's aiDAPTIV+ Solution Enables...

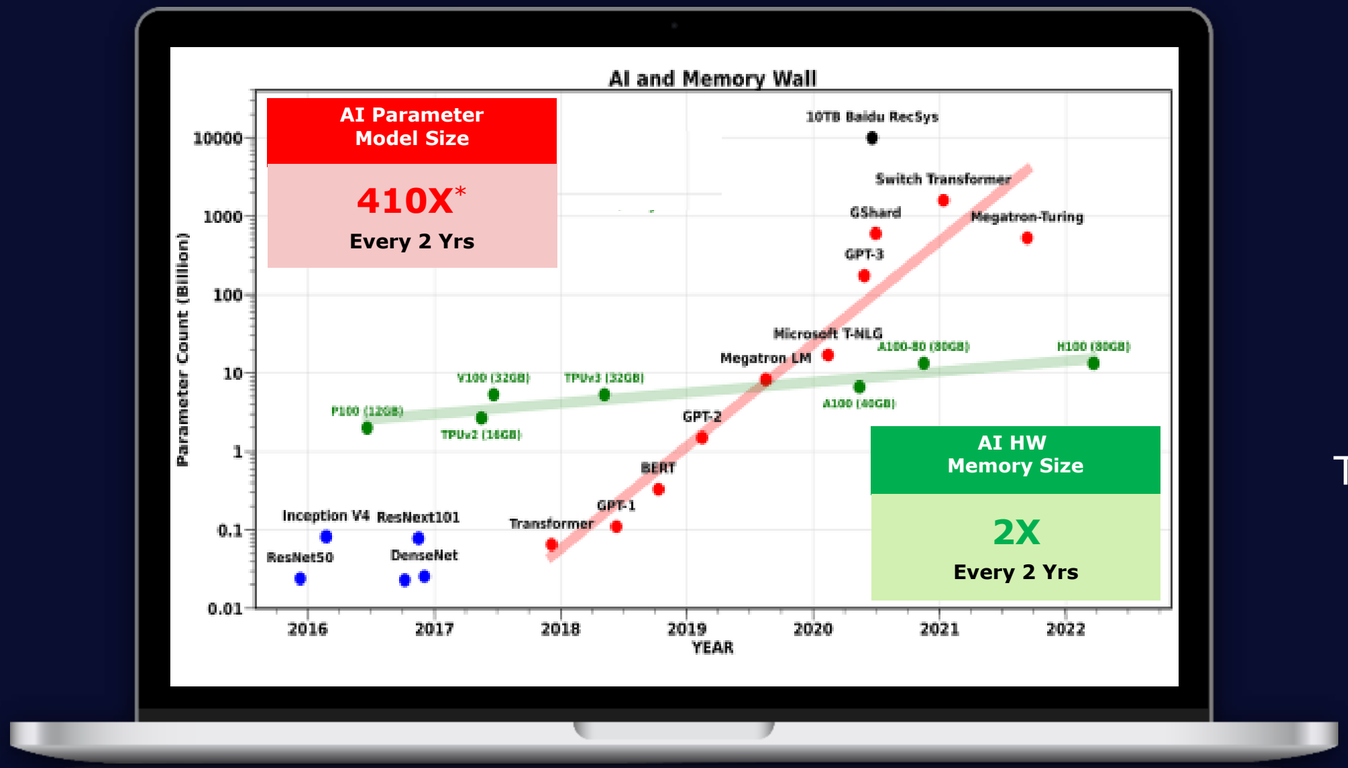
70b, 10M Tok LLM Training Workstation

- **Workstation-Class GPUs**
- **8x Cost and Power Reduction**
- **On Premise Training**
- **Less than 4.5 hours**





GPU RAM Rapid Growth of AI Model Parameters



DRAM scaling limits memory capacity expansion

RAM

The **GPU's limited memory** capacity becomes the bottleneck of AI model size*

RAM Size (Training)

LLaMA2 70B	1.4 TB
Falcon 180B	3.6 TB

Phison's aiDAPTIV+ Focus on SMB Enterprises

Large Cloud Service Providers

- Super fast performance
- Millions of users
- Priority GPU allocation
- Full size models

Market Solution

Buy more GPU and sell access

Small/Medium Organizations

- Private data
- Local hosting
- Limited GPU allocation
- Must Reduce models

Market Solution

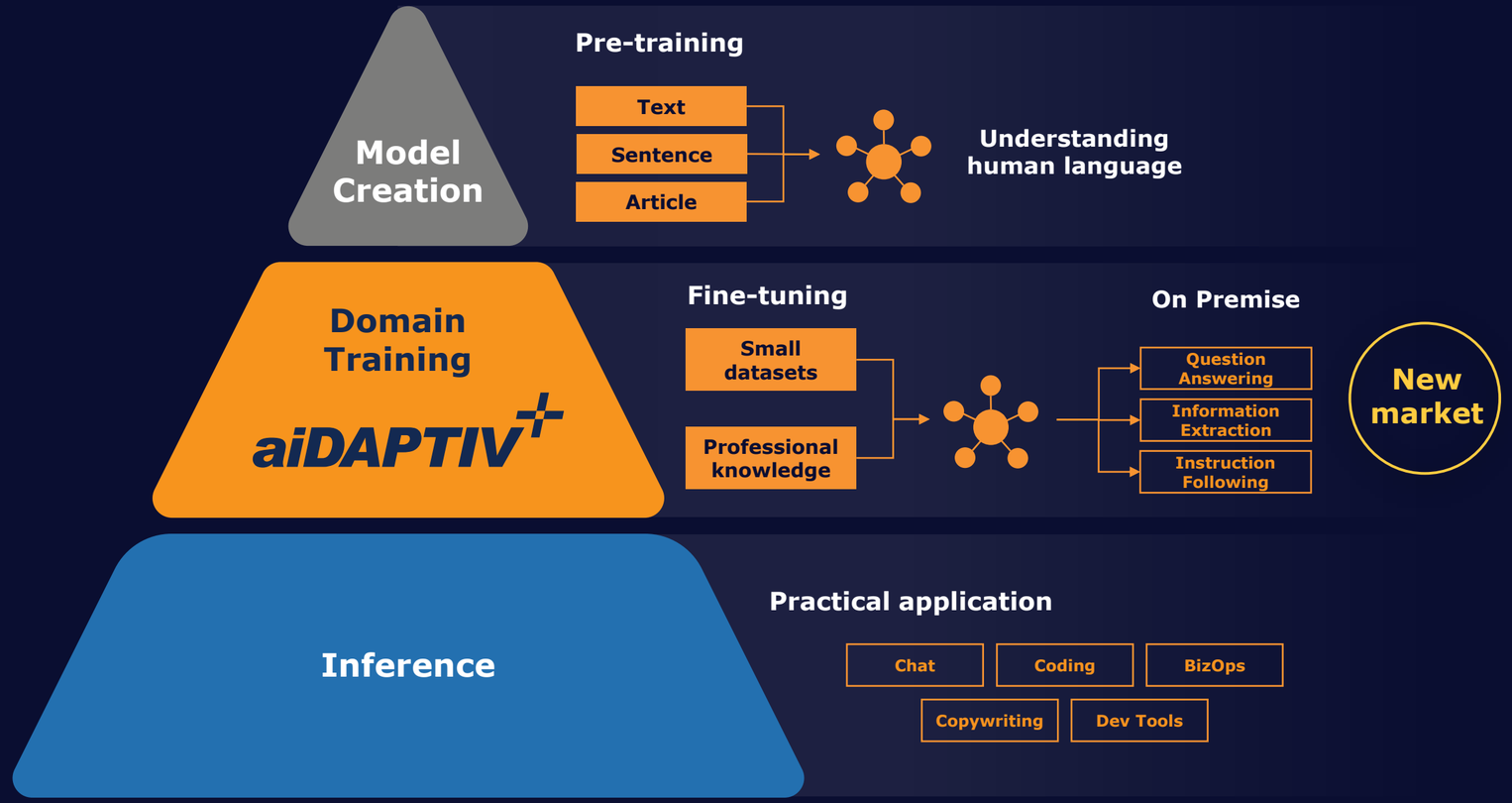
Lower model quality so it fits

aiDAPTIV+ Bridges the Gap for SMB Enterprise

➢ SMB Represent 44% of US GDP*

* https://advocacy.sba.gov/2019/01/30/small-businesses-generate-44-percent-of-u-s-economic-activity/

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

What Is Phison aiDAPTIV+ ?

Phison's aiDAPTIV+ Architecture Overview

NVIDIA GTC

LLAMA 2

Llama-2 70B Training

1.4TB memory space required

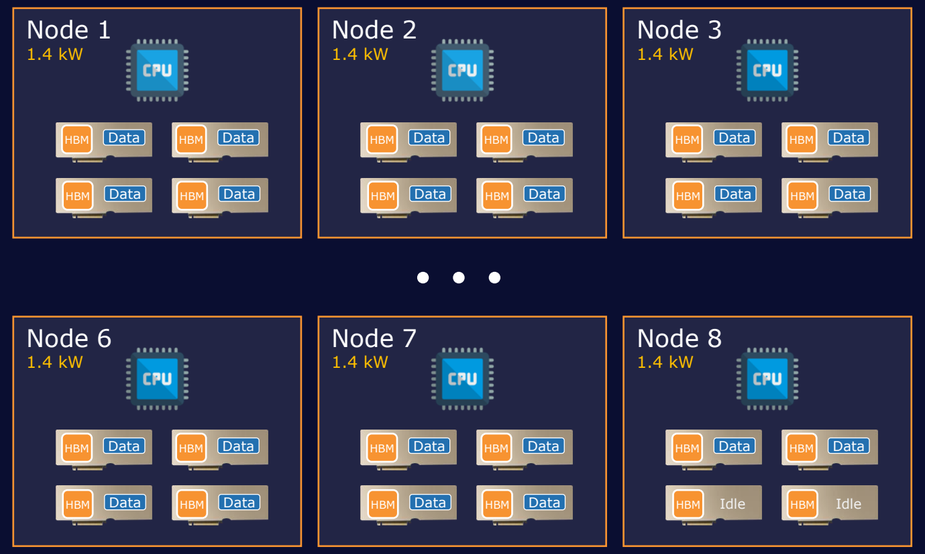
Data cut into slices

Non-aiDAPTIV+ Architecture

Limited by GPU Memory

30 GPU to train Llama-2 (70B)

(Requires 8 Workstations and Network Infrastructure)



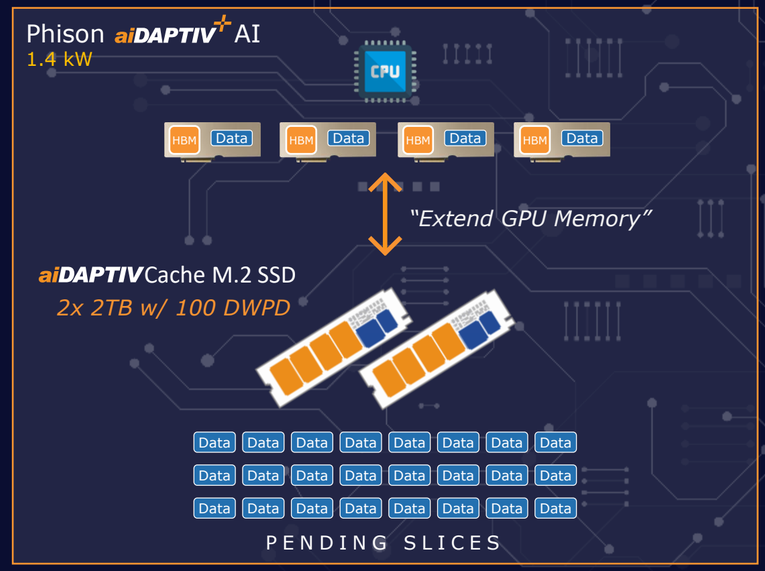
Note: Assumes 48GB / GPU

aiDAPTIV+ Architecture

Flexible Model Size

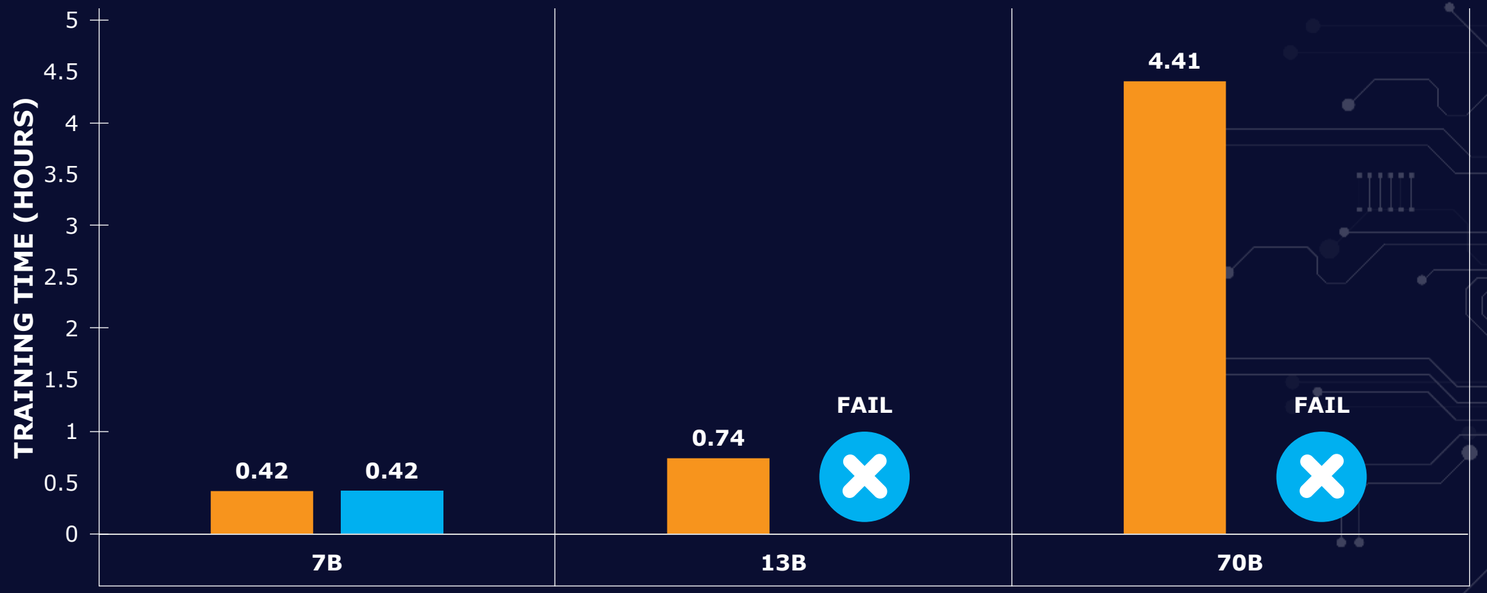
4 GPU to train Llama-2 (70B)

(Requires 1 Workstation)



aiDAPTIV+ Trains larger models w/ linear scaling

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



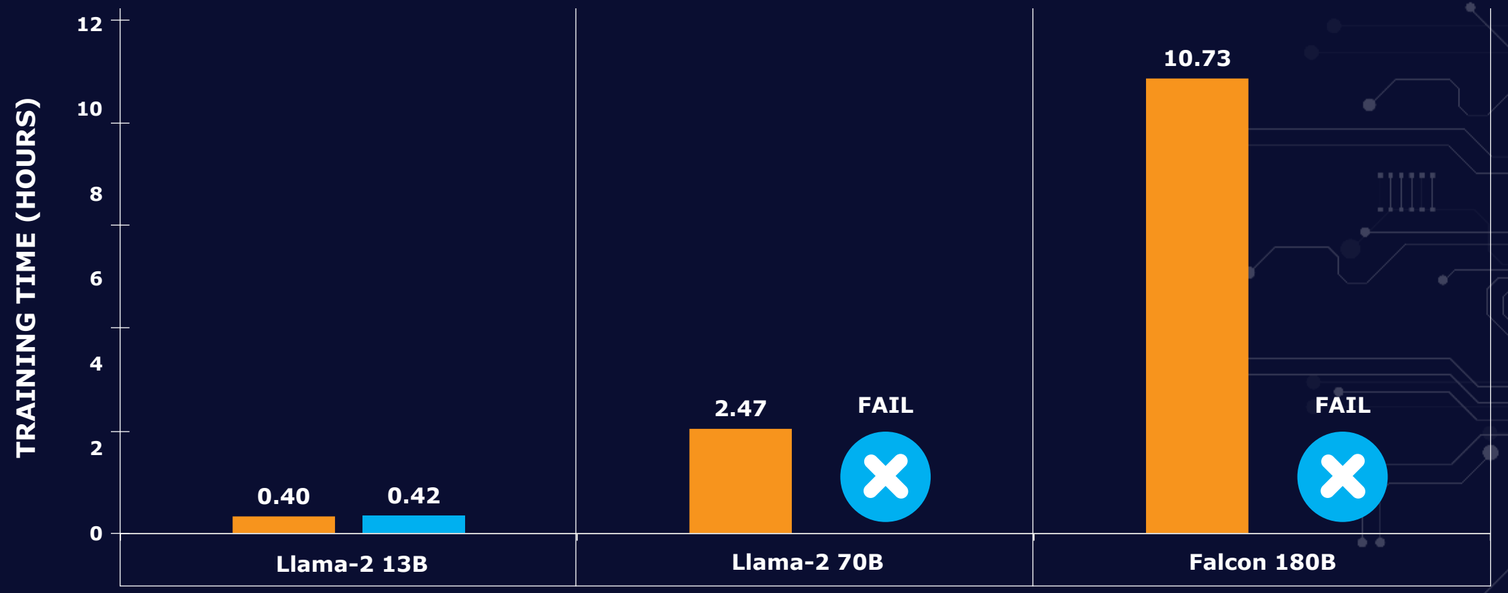
	7B	13B	70B
Model Size (Training)	140 GB	260 GB	1400 GB
HBM Pool (Usage%)	192 GB (73%)	192 GB (120%)	192 GB (729%)
Minimum GPU Count	4 / 4	4 / 6	4 / 30

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

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

aiDAPTIV+ Beyond workstations – server scaling

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

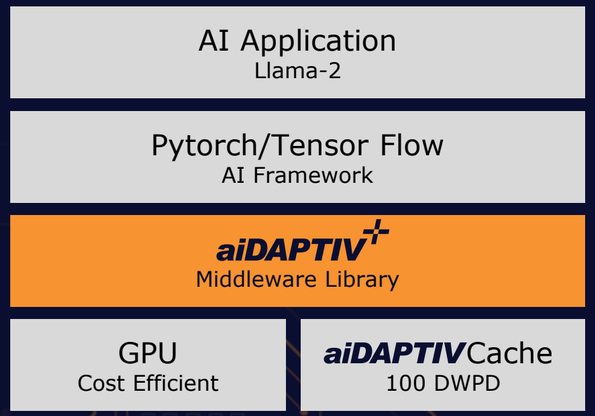


	Llama-2 13B	Llama-2 70B	Falcon 180B
Model Size (Training)	260 GB	1400 GB	3600 GB
HBM Pool (Usage%)	384 GB (68%)	384 GB (365%)	384 GB (938%)
Minimum GPU Count	**8** / **8**	**8** / **30**	**8** / **75**

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

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

aiDAPTIV+ System Family



ai100 M.2 SSD

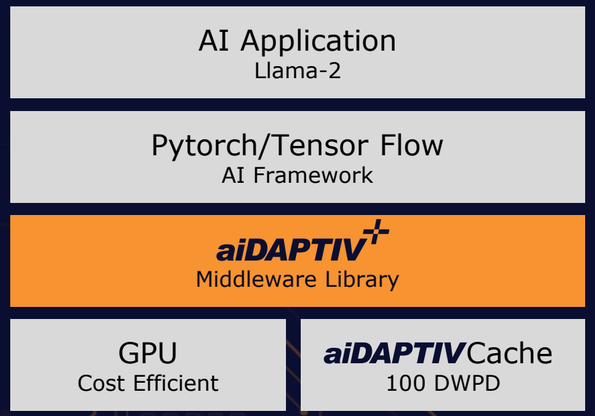
aiDAPTIV+ Middleware

Drop-in solution for all existing AI applications

aiDAPTIVCache Family

Seamless Integration with GPU Memory

aiDAPTIV+ System Family



aiDAPTIV+ Middleware

Drop-in solution for all existing AI applications

BENEFITS

- Transparent drop-in
- No need to change your AI Application
- Reuse existing HW or add nodes

aiDAPTIV+ MIDDLEWARE

- Slice model, assign to each GPU
- Hold pending slices on aiDAPTIV Cache
- Swap pending slices w/ finished slices on GPU

SYSTEM INTEGRATORS

- Access to ai100E SSD
- Middleware library license
- Full Phison support to bring up

aiDAPTIV+ System Family

SEAMLESS INTEGRATION

- Optimized middleware to extends GPU memory capacity
- 2x 2TB aiDAPTIV Cache to support 70B model
- Low latency

HIGH ENDURANCE

- Industry-leading 100 DWPD over 5 years
- SLC NAND with advanced NAND correction algorithm

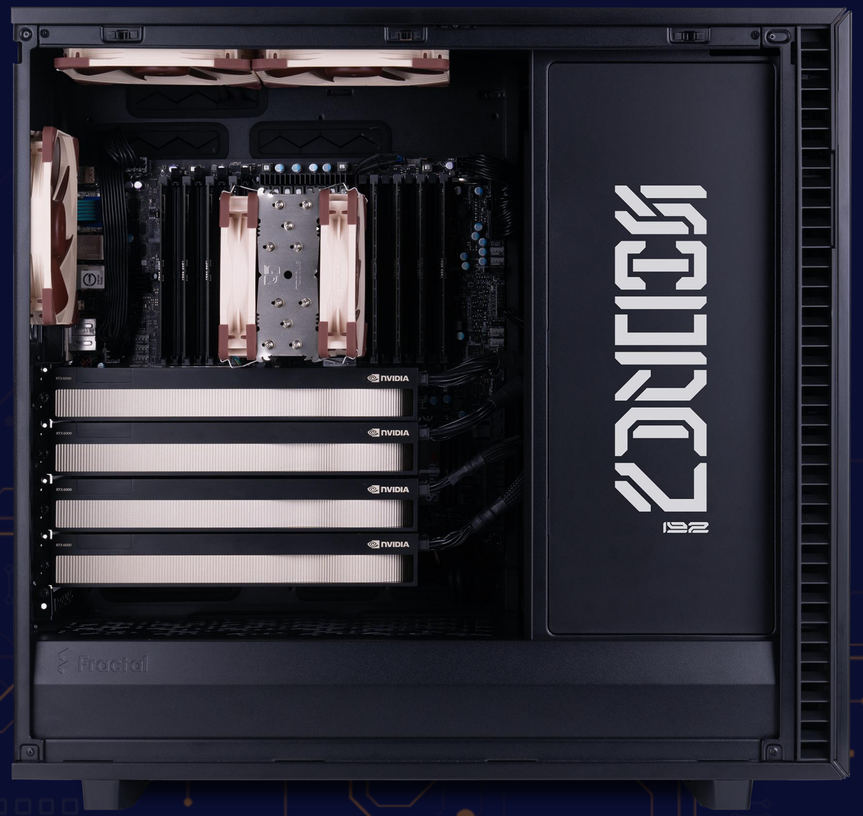
ai100 M.2 SSD

aiDAPTIVCache Family

Seamless Integration with GPU Memory

Go to Market aiDAPTIV+

GTM Configuration – Workstation



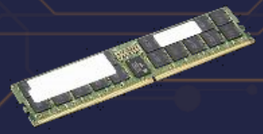
MAINGEAR AI Pro Workstation

70B LLM Workstation
Core Components

NVIDIA GPUs x 4
RTX 6000 ada

Phison ai100 M.2 SSD
2TB x 2

System DRAM
512GB



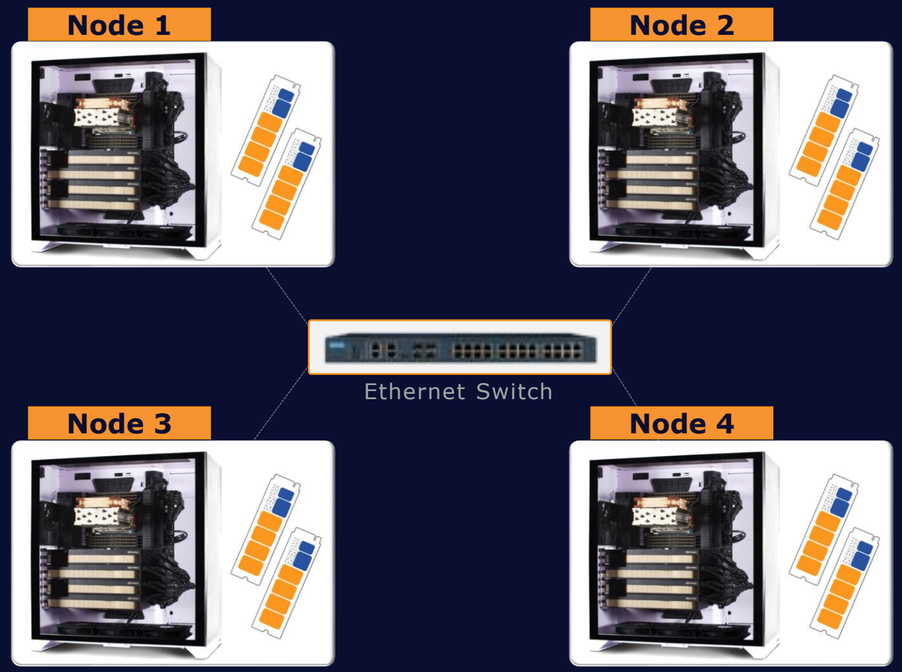
Phison
aiDAPTIV+ Middleware

aiDAPTIV+
Middleware Library

Multi-node LLAMA-2 Training

Computing power and performance for fine-tuning

aiDAPTIV+ Cluster



Multiple workstations connected with Star Topology

Low Risk Entry with Flexibility to grow over time

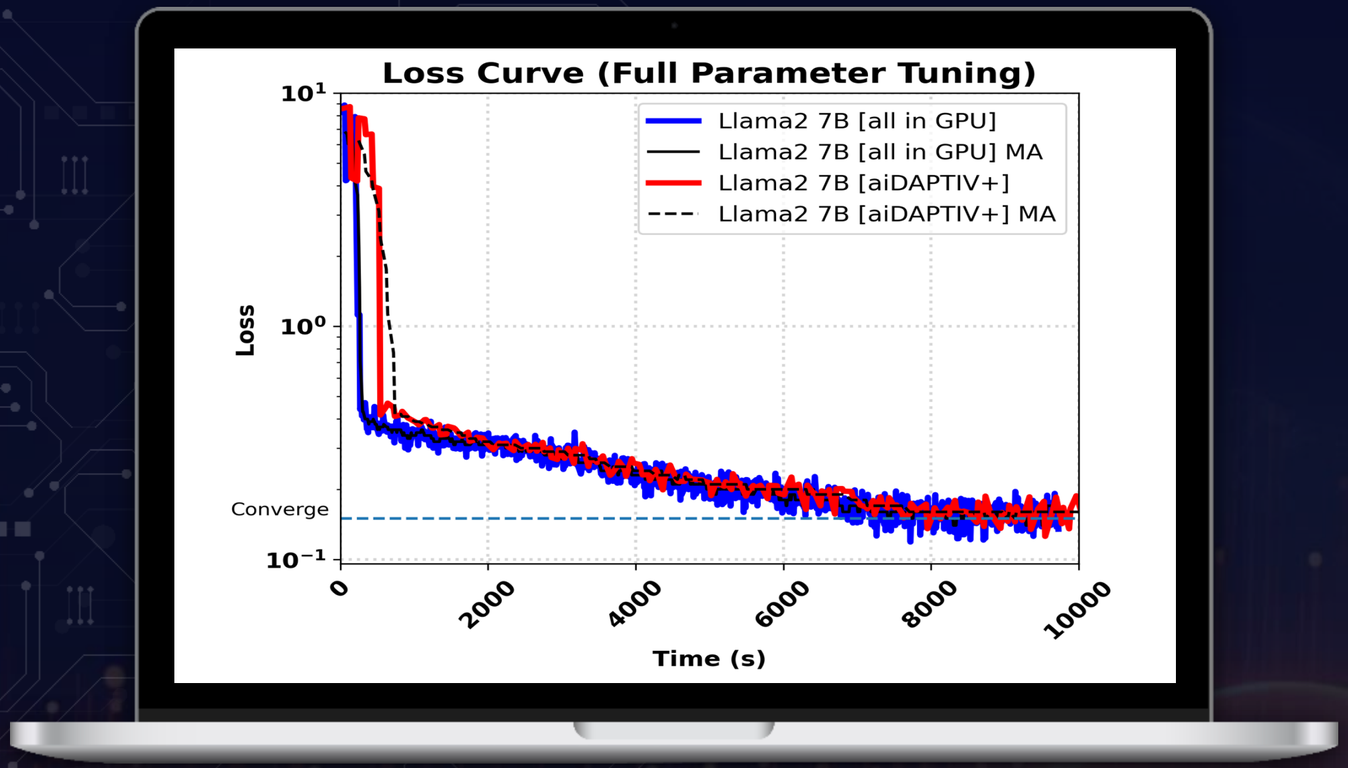
Lower the Cost of Entry
Llama-2 70B model w/ 10M Token

Up to 90% cost saving
Less than half a day

Architecture	Node Quantity	GPU # / GDDR RAM	System Cost (USD)	Training Time (Hours)
Baseline RTX 6000ada	8	30 / *1440 GB*	$480K	0.8
aiDAPTIV+ RTX 6000ada	1	4 / *128 GB*	$60K	4.41
	4	16 / *512 GB*	$240K	1.2

aiDAPTIV+ Node Configuration
- **CPU: Intel Xeon W5-3435x**
- **RAM: 512GB**
- **GPU: 4x NVIDIA RTX 6000 ada**
- **SSD: 2x 2TB Phison ai100E aiDAPTIVCache**

GenAI Language Model Training Loss Curve Comparison



Kaggle Scoring: 1000 Q&A Questions

Model	Approach	Score (%)
Llama-2-7B	Original	34
	All in GPU	65
	aiDAPTIV+	**64**

Original: without fine-tuning
all in GPU: full parameter fine-tuning in GPU
aiDAPTIV+: full parameter fine-tune with NVMe

Launching Partners

Partners & Agents

GAXIN

工業技術研究院
Industrial Technology Research Institute

hoodisk
富迪微科技

NYCU
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

MEDIATEK

yudu
御督科技有限公司
Service Technology

System Integrators

MAINGEAR

DeepMentor

GIGABYTE™

msi®

ASUS

Phison X200 PCIe Gen5 SSD

X200 Series SSD

X200 Series

Extraordinary, no ordinary. X200 overcomes your latency-sensitive workloads with the power of PCIe Gen5.

The X200 leverages Phison's proprietary CPU architecture to enable optimal performance-per-watt giving you more options to customize performance at the drive level that meets requirements while reducing operating costs at scale.

Specifications

- PCIe Gen5 x4 Dual Port
- Sequential Read - 14GB/s
- Sequential Write - 8.6GB/s
- Random Read - 3,200K IOPS
- Random Write - 800K IOPS
- Capacities - 1.92TB up to 30.72TB
- Form Factors - U.2 and E3.S

Phison
Retimer/Redriver

Retimer

Retimer PS7201

16-lane Retimer enables single IC placement for full-slot bandwidth to the edge

Interface

- PCIe 5.0
- Backward compatible with existing PCIe generation transfer rates

CXL 2.0
Support

Number of Lanes
16-lanes
*1 lane = 1 pair of Tx/Rx

Function
Retimer

Supply Voltage
0.9V, 1.2V, 1.8V

Max EQ Boosting Range
42dB @ 16GHz

Retimer PS7202

The diverse Retimer that spans all markets for challenging applications

Interface

- PCIe 5.0
- Backward compatible with existing PCIe generation transfer rates

CXL 2.0
Support

Number of Lanes
8-lanes
*1 lane = 1 pair of Tx/Rx

Function
Retimer

Supply Voltage
0.9V, 1.2V, 1.8V

Max EQ Boosting Range
42dB @ 16GHz

Redriver

Redriver PS7101

Dual-lane Redriver for value-add features in Gen5 platforms

Interface
- PCIe 5.0
- Backward compatible with existing PCIe generation transfer rates
- Also compatible with DP, SAS, SATA, and XF

Number of Lanes
2-lanes
*1 lane = 1 pair of Tx/Rx

Function
Mux/De-mux Redriver

Supply Voltage
3.3VMax

Max EQ Boosting Range
3~19dB @ 16GHz

Max Output Linear Range
1200mVppd

Redriver PS7102

8-channels to maximize the performance of hardware-assisted applications

Interface
- PCIe 5.0
- Backward compatible with existing PCIe generation transfer rates
- Also compatible with DP, SAS, SATA, and XF

Number of Channels
8-channels

Function
Pure Redriver

Supply Voltage
3.3V

Max EQ Boosting Range
0 ~ 28.5dB @ 16GHz

Redriver PS7103

16-lanes to enable high-bandwidth devices away further from the host

Interface
- PCIe 5.0
- Backward compatible with existing PCIe generation transfer rates
- Also compatible with DP, SAS, SATA, and XF

Number of Lanes
16-lanes*
1 lane = 1 pair of Tx/Rx

Function
Pure Redriver

Supply Voltage
3.3V

Max EQ Boosting Range
0 ~ 28.5dB @ 16GHz

Thank You!