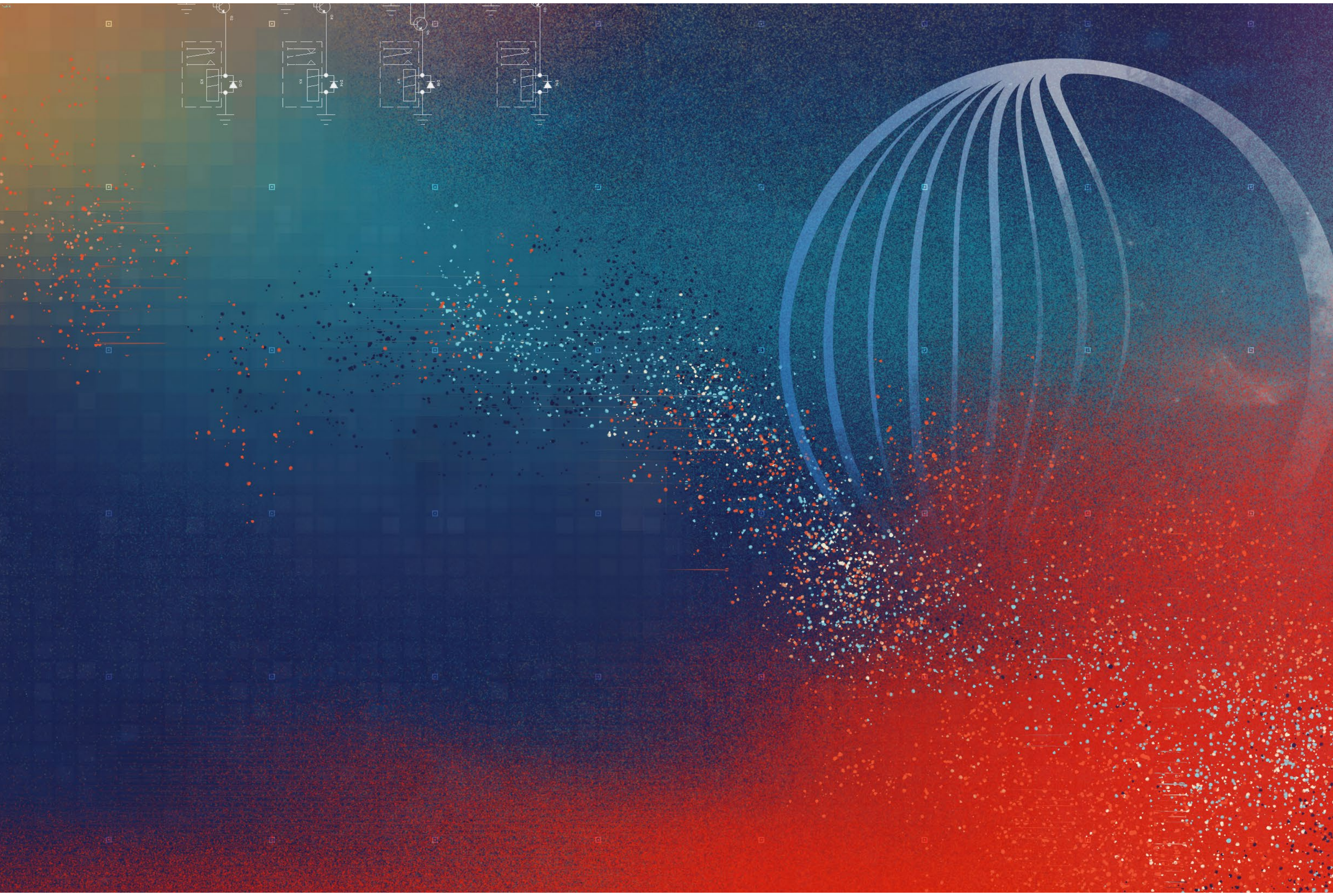
signal65



EXECUTIVE SUMMARY

Cost Efficient On-Premises AI Processing with Phison aiDAPTIV+

AUTHOR

Mitch Lewis
Performance Analyst | Signal65

IN PARTNERSHIP WITH

PHISON

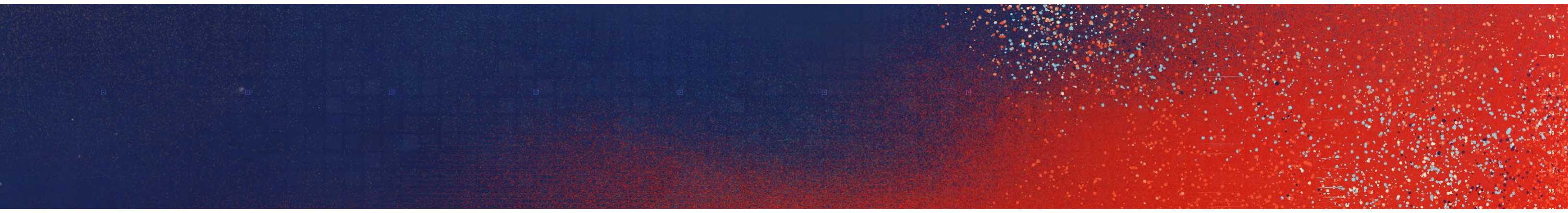
JULY 2025

Challenges of AI Fine-tuning: Privacy, Cost, and Infrastructure

The emergence of generative AI presents significant opportunities for innovation across organizations of all sizes and industries. Rapid development in the field of AI has resulted in powerful open-source large language models, including the Llama and Qwen model families, that have quickly gained widespread popularity. For custom AI applications, however, such models often require fine-tuning on additional data to meet company or industry specific needs. While this fine-tuning process unlocks powerful customized AI models, the IT organizations are challenged by technical, financial, and regulatory hurdles.

A significant barrier for organizations seeking to implement AI applications is the high-performance GPU cards typically required. In particular, GPU cards with large VRAM capacity are often required to accommodate the large sizes of cutting-edge LLM models. For the largest and most advanced AI models, numerous high-end GPU cards are commonly required in order to assemble a large enough memory pool. As VRAM accounts for a significant cost of the overall GPU, such GPU cards are expensive and due to high demand, often difficult to obtain.

Cloud platforms offer organizations quick access to high end GPU cards, with a tradeoff of a high hourly rate, often leading to unpredictable monthly billing. Further, data privacy and regulatory concerns can prohibit many organizations from utilizing their data for fine-tuning in the cloud. Alternatively, an on-premises approach typically requires purchase of high-end GPU servers, involving high costs and long lead times. Both scenarios can be severely limiting to the AI goals of small to medium-sized businesses, state and local government, and universities.



GPU Memory Expansion with Phison aiDAPTIV+

In an effort to assist organizations achieve budget-friendly on-premises fine-tuning, Phison has developed a solution, known as aiDAPTIV+, to address the GPU memory constraints commonly found in AI. Phison aiDAPTIV+ lowers the infrastructure requirements of model fine-tuning by utilizing affordable SSDs as a cache to extend GPU memory. This approach enables organizations to fine-tune large models with modest on-premises infrastructure, avoiding high costs, and maintaining data privacy requirements.

The aiDAPTIV+ solution is comprised of three components:

- aiDAPTIVCache high endurance flash memory
- aiDAPTIVLink memory management middleware
- aiDAPTIVPro Suite all-in-one LLM training toolset software (referred to as “Pro Suite”)

The aiDAPTIV+ solution enables secure, on-premises AI processing, using modest GPU cards for a fraction of the price of typical AI datacenter infrastructure. Phison aiDAPTIV+ enables organizations to fine-tune large LLM models on a wide range of infrastructure, even with limited VRAM.

	Cloud	Traditional On-Premises	Phison aiDAPTIV+ On-Premises
Infrastructure	High-end GPU cloud instances	Specialized AI servers with high-end GPU cards	Broad Device Support: • Edge devices • Notebook PCs • Desktops • Workstations • Servers • Storage arrays
VRAM Requirement	High	High	Low
Data Privacy	Data Resides on Cloud	Full on-premises data control	Full on-premises data control
Cost	High ongoing (hourly) costs	High upfront capital expense	Flexible – budget friendly hardware

Figure 1: *Phison aiDAPTIV+ Fine-tuning Comparison*

Signal65 Testing Summary

Signal65 conducted hands-on testing of AI model fine-tuning to evaluate the impact of Phison aiDAPTIV+. To demonstrate the memory expansion capabilities of aiDAPTIV+, Signal65 conducted model fine-tuning on four AI models using a GPU workstation, both with and without aiDAPTIV+ enabled. The workstation utilized contained four NVIDIA RTX 6000 Ada Generation GPU cards, each with 48 GB of VRAM. Models tested include the following:

- Meta-Llama-3.1-8B-Instruct
- Meta-Llama-3.1-70B-Instruct
- Qwen-2.5-7B-Instruct
- Qwen-2.5-72B-Instruct

Fine-tuning workloads were run for each model on 1, 2, and 4 GPU card configurations, each with memory-intensive configurations. Key highlights from testing and subsequent cost analysis include:

- **Phison aiDAPTIV+ enabled successful fine-tuning of all four models tested under conditions where standard configurations failed due to memory constraints**
- **Fine-tuning of a 70-billion and 72-billion parameter AI models on a single GPU card with only 48 GB of VRAM when enabled by the aiDAPTIV+ GPU memory extension**
- **Up to 85% cost savings compared to traditional AI infrastructure deployment approaches**

Final Thoughts – Phison aiDAPTIV+ Lowers the Barrier for AI Applications

Testing of Phison aiDAPTIV+ demonstrated successful memory expansion, enabling each model tested to be fine-tuned on a single GPU card with as little as 48 GB of VRAM. This evaluation showcases ability of aiDAPTIV+ to unlock advanced AI workloads on modest, cost effective hardware. Additionally, aiDAPTIV+ enables organizations with regulatory concerns to easily deploy fine-tuning on-premises. Through this evaluation, Signal65 recognizes Phison aiDAPTIV+ as a technology that lowers hardware barriers and broadens access to AI development, enabling innovation across a wider range of organizations and use cases.

Important Information About this Report

CONTRIBUTORS

Mitch Lewis
Performance Analyst | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

IN PARTNERSHIP WITH

PHISON

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION

Signal65 I signal65.com