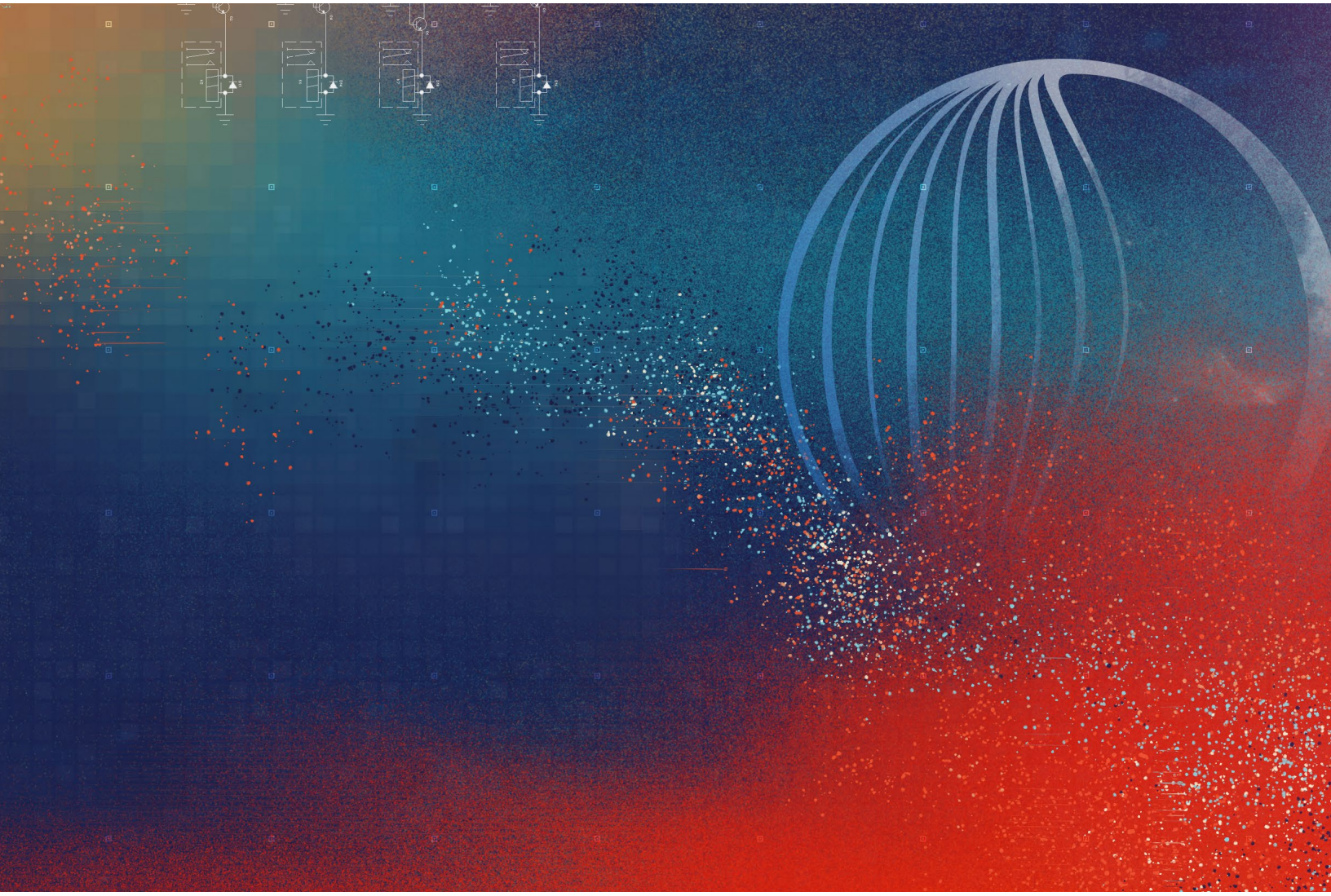
signal65



Cost Efficient On-Premises AI Processing with Phison aiDAPTIV+

AUTHOR

Mitch Lewis
Performance Analyst | Signal65

IN PARTNERSHIP WITH

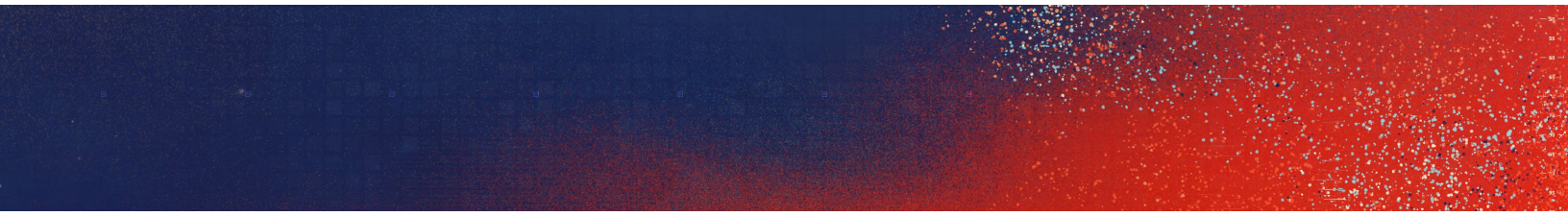
PHISON

JULY 2025

Executive Summary

The emergence of generative AI presents significant opportunities for innovation across organizations of all sizes and industries. However, the technical infrastructure required for both model fine-tuning and inference often poses a major challenge due to the reliance on expensive, high-performance GPUs. This paper examines the cost and data privacy limitations associated with traditional cloud and on-premises approaches and explores how Phison aiDAPTIV+ can help organizations overcome these barriers. Through hands-on testing, Signal65 validated the solution's capabilities and assessed its impact on easing AI infrastructure challenges. Key findings include:

- Successful fine-tuning of four distinct AI models under conditions where standard configurations failed due to memory constraints
- Fine-tuning of a 70-billion-parameter AI model on a single GPU 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
- Simplified AI development, enhanced data security, and improved inference performance



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

Generative AI has quickly become a key priority for organizations across all industries, forming the backbone of new intelligent applications. The possibilities of AI applications are broad, generating interest from organizations of all sizes. Creating effective AI applications, however, can be challenging due to privacy, cost, and infrastructure constraints – especially for smaller organizations with limited budgets.

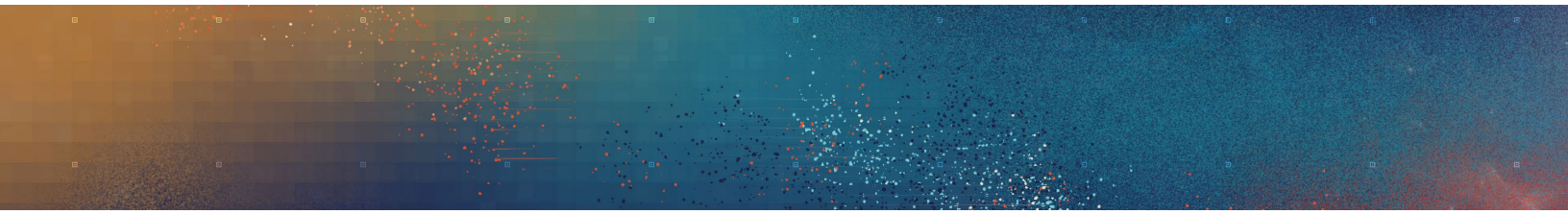
The current AI model landscape includes a broad range of open-source Large Language Models (LLMs), including model families such as Llama, Gemma, Qwen, Mistral, and many others. These models span a wide range of parameter sizes, architectures, and training data sources, all of which impact their outputs and their overall usability for various applications. Using such open-source foundational models enables organizations to leverage general-purpose, flexible AI, without the overwhelming requirement of building full models themselves, which is typically beyond the technical scope of most organizations. These open-source foundational models typically deliver impressive general language and reasoning capabilities; however, they may not be trained on the industry or company specific information necessary to build custom AI applications.

To better enable industry specific AI use cases, organizations can fine-tune open-source foundational LLMs to train them to understand specific information or complete specific tasks. Although far less resource intensive than fully training a foundational model, on-premises fine-tuning still introduces additional cost. The fine-tuning process involves computationally expensive training steps, typically requiring multiple expensive GPUs with large memory (VRAM) capacities.

To quickly meet these GPU requirements, many organizations have utilized cloud resources for their initial AI development. The cloud provides easily accessible, scalable GPUs, capable of fine-tuning large, state-of-the-art LLMs. While the cloud provides ease of access to powerful infrastructure, it typically does so at an expensive hourly rate, often with unpredictable end-of-month billing amounts. These high, unpredictable costs are often not realistic for many smaller organizations such as small to medium-sized businesses, state and local government, and universities. Leveraging the cloud for AI fine-tuning additionally presents a challenge when considering data privacy and data sovereignty. Due to the sensitivity of their data and regulatory requirements, many organizations must retain data on-premises, making cloud-based fine-tuning infeasible.

Due to such privacy and data sovereignty concerns, many organizations have instead opted for an on-premises approach. While this approach ensures control and compliance over private data, it sacrifices the simplicity, accessibility, and scalability of utilizing the cloud. To fine-tune large models, organizations must acquire significant on-premises infrastructure. While this avoids the high hourly rate and often surprising monthly invoice amounts found in cloud services, it instead replaces it with a large upfront cost. Additionally, the high demand for numerous high-performance AI GPUs has made them difficult to obtain, resulting in long lead times and delaying AI development.

Ongoing model development further complicates these challenges, with larger models needing even greater overall infrastructure requirements, whether hosted in the cloud or deployed on-premises. As model sizes grow, GPU memory specifically becomes the highest cost factor. In fact, the VRAM on the GPU cards is often more expensive than the GPU itself. Large AI models can require hundreds of GBs of GPU memory, forcing organizations to purchase numerous expensive high-end AI GPUs in order to assemble a large enough memory pool to utilize the most advanced AI models. While AI applications offer clear benefits to virtually all organizations, many small and medium sized businesses, government agencies, and universities may be locked out by the infrastructure, cost, and privacy requirements typically associated with AI fine-tuning.



Expanding GPU Memory with Phison aiDAPTIV+

In an effort to assist organizations in achieving 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, while 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 software (referred to as “Pro Suite”)

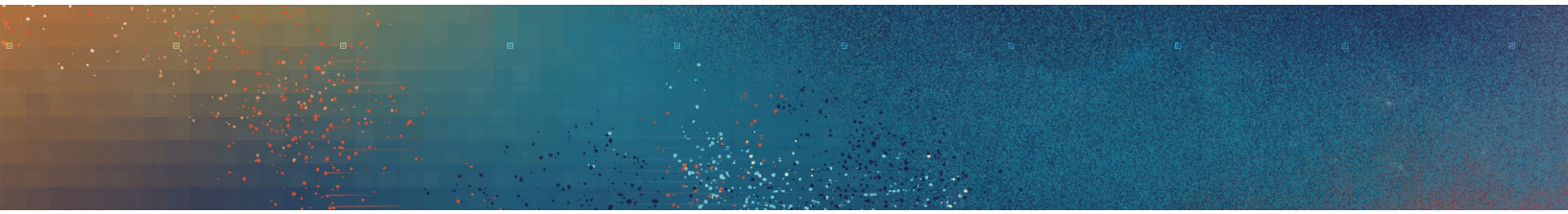
aiDAPTIV+ extends GPU memory to aiDAPTIVCache devices, a family of high endurance Phison SLC SSDs. Memory management is handled by aiDAPTIVLink, a middleware library that slices model weights and moves slices between

GPUs and aiDAPTIVCache devices. The solution is managed via Pro Suite software, which provides a full end-to-end toolset with an intuitive graphical user interface for model fine-tuning and evaluation.

By offloading model weights to SSD storage, Phison aiDAPTIV+ expands a system's overall training capabilities. This enables organizations to fine-tune large models on less advanced hardware than typically required.

Phison aiDAPTIV+ is compatible with edge devices, notebook PCs, desktops, workstations, servers, and storage arrays, greatly expanding the hardware options for organizations seeking to utilize large AI models. Phison currently supports the following model size fine-tuning for each device:

- Edge devices – up to 1 billion full parameter model training (70 Billion quantized parameter model training)
- AI notebook PC – up to 8 billion full parameters model training
- Desktop – up to 13 billon parameter model training
- Workstation PC – Up to 100 billion parameter model training
- Server – up to 671 billion parameter model training (projected to be over 2 trillion parameter model training by the end of 2026)



Testing Overview

To evaluate the functionality of Phison aiDAPTIV+, Signal65 performed a hands-on comparison of model fine-tuning, both with and without aiDAPTIV+. Testing was completed utilizing an HP Z8 Fury G5 workstation equipped with four NVIDIA RTX 6000 Ada Generation GPUs, each with 48 GB of VRAM.

Fine-tuning was performed using a creative writing dataset[1] consisting of 672 question and answer pairs on four distinct models to evaluate the system's ability to train models of various parameter sizes. Models tested included the following:

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

Both Llama and Qwen model families are well known for providing highly accurate, general-purpose models, suitable for a broad range of enterprise applications. The four models selected for this testing include some of the most widely used AI models currently available. Llama-3.1-8B and Qwen-2.5-7B are relatively small models, offering strong performance with relatively modest memory requirements. The larger Llama-3.1-70B and Qwen-2.5-72B models represent more capable models, offering greatly improved language understanding and complex task handling capabilities, but requiring significantly greater memory resources.

[1] https://huggingface.co/datasets/lionelchg/dolly_creative_writing

It should be noted that Phison aiDAPTIV+ is not intended to increase system performance during the fine-tuning process. Instead, aiDAPTIV+ enables model fine-tuning to overcome memory constraints that would otherwise lead to an out of memory failure. To validate this capability, testing evaluated the completion of fine-tuning jobs under memory intensive conditions.

To stress system memory, testing was completed using both large sequence lengths and large batch sizes. Tests were run for a single epoch utilizing between one to four GPUs. Configuration details for all models tested are outlined below:

	Meta Llama-3.1-8B	Meta Llama-3.1-70B	Qwen-2.5-7B	Qwen-2.5-72B
Epochs	1	1	1	1
Max Sequence Length	12,000	12,000	12,000	12,000
Micro-batch Size	18	9	14	8
Gradient Accumulation	10	10	10	10
Learning Rate	0.000007	0.000007	0.000007	0.000007
GPUs	1, 2, 4	1, 2, 4	1, 2, 4	1, 2, 4

Figure 1: *Model Fine-tuning Configuration Details*

Each configuration was tested with Phison aiDAPTIV+ through the Pro Suite software, as well as without aiDAPTIV+ using the open-source Axolotl library. An overview of the testing results can be found below:

Model	GPUs	VRAM	With aiDAPTIV+	Without aiDAPTIV+
Llama-3.1-8B	1	48 GB	Complete – 1 hr, 8 min, 20 s	Failed – out of memory
	2	96 GB	Complete – 36 min, 31 s	Failed – out of memory
	4	192 GB	Complete – 20 min, 42 s	Failed – out of memory
Llama-3.1-70B	1	48 GB	Complete – 10 hr, 33 min, 56 s	Failed – out of memory
	2	96 GB	Complete – 5 hr, 30 min, 44 s	Failed – out of memory
	4	192 GB	Complete – 2 hr, 53 min, 10 s	Failed – out of memory
Qwen-2.5-7B	1	48 GB	Complete – 1 hr, 11 m, 5 s	Failed – out of memory
	2	96 GB	Complete – 38 min, 45 s	Failed – out of memory
	4	192GB	Complete – 22 min, 38 s	Failed – out of memory
Qwen-2.5-72B	1	48 GB	Complete – 11 hr, 1 min, 26 s	Failed – out of memory
	2	96 GB	Complete – 5 hr, 54 min, 4 s	Failed – out of memory
	4	192GB	Complete – 3hr, 19 min, 55 s	Failed – out of memory

Figure 2: *Model Fine-tuning Test Results*

Key Findings

Testing demonstrated that by utilizing Phison aiDAPTIV+, the system was capable of fine-tuning all models tested, even when stressing the system with large batch sizes and long sequence lengths. Notably, fine-tuning of all models could be achieved on a single NVIDIA RTX 6000 GPU with 48 GB of VRAM.

Without aiDAPTIV+, the system was unable to complete the fine-tuning jobs for any of the comparative test configurations due to GPU memory constraints. It should be noted that additional testing found the system capable of fine-tuning Llama-3.1-8B and Qwen-2.5-7B when configured with less memory intensive settings outside of those outlined in the pre-determined test plan. This included configurations with smaller batch sizes and shorter max sequence lengths. Fine-tuning tests for the two larger models encountered errors due to GPU memory constraints for every configuration tested without aiDAPTIV+.

The test configurations were purposely chosen to stress the system's GPU memory, highlighting how aiDAPTIV+ can overcome common memory challenges while using modest hardware. For smaller models, such as Llama-3.1-8B and Qwen-2.5-7B, the model can be fine-tuned on such a system either with or without aiDAPTIV+, depending on the configuration. The chosen test configurations demonstrate, however, that for more complex fine-tuning jobs, such as utilizing large batch sizes or long sequence lengths, Phison aiDAPTIV+ enables fine-tuning where the system would otherwise fail. This capability provides organizations with greater flexibility to run complex fine-tuning workloads. Larger batch sizes can be more efficient when training with large datasets, while long sequence lengths can achieve more accurate models for tasks requiring long contexts, such as document summarization or code generation.

When testing the larger Llama-3.1-70B and Qwen-2.5-72B models, aiDAPTIV+ enabled successful fine-tuning of both models, overcoming the memory constraints that otherwise resulted in errors. Although models of this scale typically require substantial hardware resources, they are often highly desirable for enterprise applications. Large models, such as Llama-3.1-70B and Qwen-2.5-72B offer advanced reasoning capabilities, broader domain and language understanding, and support for longer context windows, making them especially well-suited for complex, enterprise-grade workloads. While all fine-tuning attempts without aiDAPTIV+ failed, it is notable that with aiDAPTIV+ enabled, the system was able to successfully fine-tune the large 70B and 72B parameter models on as little as a single GPU.

Although training times were recorded for transparency, testing was not intended to showcase performance characteristics. As expected, training times varied depending on model size and the number of GPUs available. Notably, fine-tuning of both the 70B and 72B parameter models required over 10 hours to complete on a single GPU. While this represents a significant amount of training time, it underscores the challenge of fine-tuning large models on smaller hardware configurations. For practical applications, organizations should assess how often fine-tuning is required, their model selection, and their overall resource availability to evaluate an acceptable balance between resource utilization and performance. Use of more GPUs would result in faster fine-tuning.

Key Takeaways

Lower Infrastructure Costs

The key value of Phison aiDAPTIV+ is enabling organizations to utilize large AI models, on relatively modest hardware. By enabling organizations to run models that would otherwise require more substantial infrastructure, aiDAPTIV+ can both empower organizations to achieve more innovative AI applications, and lower overall infrastructure costs.

Rapid advancement in AI model development brings promising potential for powerful new AI applications; however, the new AI models are continually growing in size, requiring more powerful hardware. Datacenter servers with high performance GPUs and accelerators are expensive, and with the high demand for AI, they are often difficult to obtain. This presents a hurdle for many organizations, especially small to medium-sized businesses, state and local government, and universities that do not have the ability to build large AI datacenters. By utilizing Phison aiDAPTIV+, organizations can instead utilize easily obtainable, cost-effective hardware, such as workstation focused GPUs.

The following provides a comparison between a GPU-based workstation with aiDAPTIV+, such as the one used in this testing, an AI datacenter server, and an AWS cloud instance. Both the AI datacenter server and the AWS cloud instance utilize 8 NVIDIA H100 GPUs, representing common configurations for running large, state of the art AI models, either on-premise or in the cloud. All pricing information was collected from publicly available sources. The GPU-based workstation includes the added price of two Phison Pascari AI100E M.2 2280 2TB SSDs to enable the aiDAPTIV+ capabilities.

	HP GPU-based Workstation	On-premises NVIDIA H100 Server	AWS EC2 p5.48xlarge
GPUs	4x NVIDIA RTX 6000 ADA Generation	8x NVIDIA H100	8x NVIDIA H100
aiDAPTIV+	Yes	No	No
VRAM	192 GB	640 GB	640 GB
Cost	$45,499.78	$301,402.00	$55.04 / hr ($482,150.40 /yr)

Figure 3: *Cost Comparison (Source: HP Workstation Pricing, Pascari AI100E Pricing, NVIDIA H100 Server Pricing, AWS EC2 Pricing)*

When considering cost, the GPU-based workstation holds a clear advantage over the AI datacenter server, with an 85% lower cost. While these two systems would not typically warrant a direct comparison, the addition of Phison aiDAPTIV+ enables GPU-based workstations to achieve fine-tuning of large AI models that typically require more powerful AI datacenter GPUs, as seen in the fine-tuning test results for Llama-3.1-70B and Qwen-2.5-72B.

The GPU-based workstation additionally offers a far greater cost efficiency than cloud hosted GPUs, as can be seen when compared to an AWS EC2 p5.48xlarge instance. This instance provides 8 NVIDIA H100 GPUs, similar to the on-premises AI datacenter server, for an hourly price of $55.04. While hourly pricing can be efficient for short term experimentation, for ongoing AI development, such pricing quickly exceeds that of the on-premises GPU-based workstation. The total cost of this cloud instance, will surpass the total purchase cost of the GPU-based workstation in slightly over a month. For one year, the total price of the cloud instance amounts to $482,150.4, over 10x more expensive than the GPU-based workstation. It's important to note that this calculation assumes continuous operation of the cloud instance. While many organizations have variable cloud usage needs, this example illustrates how cloud costs can accumulate quickly with sustained usage.

Although aiDAPTIV+ enables GPU-based workstations to run large models at a fraction of the cost of specialized AI datacenter servers or cloud instances, it should be noted that it does not provide the same performance. Higher end or more numerous AI GPUs with more VRAM will achieve higher performance and lower overall training times compared to utilizing aiDAPTIV+ with modest hardware configurations. Although training time may be crucial for some organizations with large training requirements, for many organizations the lower infrastructure costs and enhanced training capabilities, along with enabling the data privacy and data sovereignty provided by aiDAPTIV+, will far outweigh additional training time. This is especially true for organizations where service levels only require fine-tuning a few times per day, at most.

Enables Secure On-Premises Fine-tuning

Phison aiDAPTIV+ enables organizations to easily achieve model fine-tuning on-premises. While cloud services offer easy access to GPUs, high hourly costs and data privacy concerns along with monthly billing surprises can make cloud usage undesirable. As more organizations seek to fine-tune AI models utilizing their own private data, the requirement for on-premises AI has grown.

While on-premises fine-tuning can mitigate data privacy concerns, the large infrastructure requirements can become cost prohibitive for many organizations. By reducing GPU memory constraints, Phison aiDAPTIV+ enables organizations to fine-tune models on-premises utilizing less expensive hardware, maintaining control over their private data while also better fitting their budgets.

Simplified AI Development with Phison Pro Suite

In addition to the memory extension capabilities achieved with aiDAPTIV+, Signal65 noted the overall simplicity of the Phison Pro Suite software throughout the testing process. Pro Suite provides an intuitive GUI interface that enables otherwise complex processes, such as model fine-tuning, to be achieved in a simple point-and-click manner.

Beyond launching fine-tuning jobs, the Pro Suite software offers several other features capable of simplifying AI development.

Pro Suite includes a data formatting tool called aiDAPTIVGuru, which automatically generates fine-tuning datasets based on uploaded documents. Properly formatting datasets for fine-tuning can often be a complex and time-consuming task, with the potential to delay the overall fine-tuning processes. With aiDAPTIVGuru, organizations can simply upload documents and quickly obtain a formatted dataset ready for fine-tuning, reducing the time and complexity of the overall process.

Pro Suite also includes functionality to easily evaluate and contrast models. Pro Suite includes a side-by-side inferencing interface, allowing users to select two models and easily compare their responses. This functionality allows a straightforward comparison of models, enabling organizations to understand how their models are performing, and which models are most well suited for deployment.

Enhanced Inferencing

While this evaluation primarily tested model fine-tuning, it should be noted that Phison aiDAPTIV+ can additionally be utilized to enhance AI inferencing. During inferencing, the expanded memory provided by aiDAPTIV+ extends the Key Value (KV) Cache resulting in quicker time to first token recall for faster results. The expanded GPU memory also enables longer token lengths for greater context windows, enabling complex workloads that may require long inputs and outputs to obtain more accurate results.

Final Thoughts

While the rapid advancement of AI presents organizations with significant opportunities for innovation and increased efficiencies, implementing the technology involves significant challenges when considering infrastructure, cost, and data privacy. Phison aiDAPTIV+ provides a compelling solution to overcome these challenges, allowing organizations to maintain data privacy on-premises while utilizing cost-effective, easily accessible hardware.

Signal65 testing of Phison aiDAPTIV+ has validated that the technology is capable of extending GPU memory during model fine-tuning, enabling successful fine-tuning of large AI models in configuration that would otherwise fail. This ability can enable organizations to overcome cost and infrastructure challenges and enable organizations of any size to utilize large, innovative AI models. Signal65 found model fine-tuning and evaluation processes to be significantly simplified with the use of the Phison Pro Suite interface. In addition to model fine-tuning, Phison aiDAPTIV+ notably provides the ability to increase the performance and capability of AI inferencing.

As AI models continue to grow in size and complexity, infrastructure requirements and GPU memory constraints will increasingly become a barrier for organizations looking to implement the latest AI technology. As these challenges continue, IT organizations should look to new technology, such as Phison aiDAPTIV+, to overcome their infrastructure and cost restraints, as well as ensure they stay in compliance with data privacy and data sovereignty mandates.

Appendix

System	HP Z8 Fury G5
Operating System	Ubuntu 22.04
CPU	Intel® Xeon® w5-3435X Processor
Memory	512 GB DDR5
Storage	2x Pascari AI100E M.2 2280 2TB PCI-Express 4.0 Solid State Disk Enterprise
GPUs	4x NVIDIA RTX 6000 Ada Generation
Phison aiDAPTIV Pro Suite	2.0.5

Figure 4: *Test System Specifications*

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

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