2025 年 1 月 23 日

プレスリリース

報道関係者各位

HPC システムズ株式会社
代表取締役 小野 鉄平
(コード番号：6597 東証グロース)
問合せ先 取締役管理部長 下川 健司
(電話番号：03-5446-5530)

HPC システムズ、AI 時代の“テキスト生成”にフォーカスした LLM フルファインチューニング環境を提供開始

－ Llama 3.x（8B～70B）、Gemma2（9B・27B）など最新モデルにも柔軟対応 －
－ 導入検討者向けオフラインセミナーで実機デモを公開予定 －

HPC システムズ株式会社（本社：東京都港区、代表取締役 小野 鉄平、以下 HPC システムズ）は、PHISON Electronics Corporation（台湾・苗栗県、以下「PHISON 社」）が開発した独自技術**「aiDAPTIV+」搭載 GPU サーバー・ワークステーションの提供を開始**いたします。

本ソリューションは**AI 時代のテキスト生成（NLP）**にフォーカスしており、Llama 3.x（8B～70B）や Gemma2（9B・27B）などの**大規模言語モデル（LLM）をフルファインチューニングできる環境を、PoC（概念実証）段階から低コストで導入可能にする点が特長**です。

GPU の VRAM 不足をシステムメモリと SSD で補う「aiDAPTIV+」によって、**従来は複数台の高価な GPU を必要としていた大規模モデルの学習を単一ノード構成から実現**。中小企業・研究機関が DX（デジタル変革）を進めるうえでの高い初期投資のハードルを大幅に下げます。さらに、導入時の不安を解消するため、オフラインセミナーでの実機デモンストレーションも予定しています。

【背景と課題】

生成 AI ブームを背景に、チャットボットや文章要約・翻訳などのテキスト生成タスクが幅広い業種・分野で注目を集めています。特に、モデル全体を再学習して専門領域に最適化する**「フルファインチューニング」**は、高度な精度とドメイン知識を内在化できることから、DX を加速する有力な方法とされています。一方で主に下記の課題があり実践できる組織は限られています。

- 高価な GPU や大容量 VRAM が必須
- マルチノード環境の構築・運用が複雑
- PoC 段階で大きな投資が必要

HPC システムズは、AI や HPC（ハイパフォーマンスコンピューティング）分野のノウハウを活かし、**テキスト生成にフォーカスしたより低コストで段階的に拡張できるフルファインチューニング環境を提供**。こうした障壁を取り除くことで、**中小企業や研究機関から大企業まで幅広い組織が DX の成果を迅速に享受できるよう支援**します。

【新ソリューションのポイント】

1. テキスト生成（NLP）にフォーカスした効率化

- マルチモーダル対応は行わず、テキスト処理に最適化した構成で高い学習効率を実現。
- チャットボット、文書要約、翻訳など、ビジネスニーズの高い NLP タスクを中心に DX 推進を加速。

2.「aiDAPTIV+」が VRAM 不足を補完

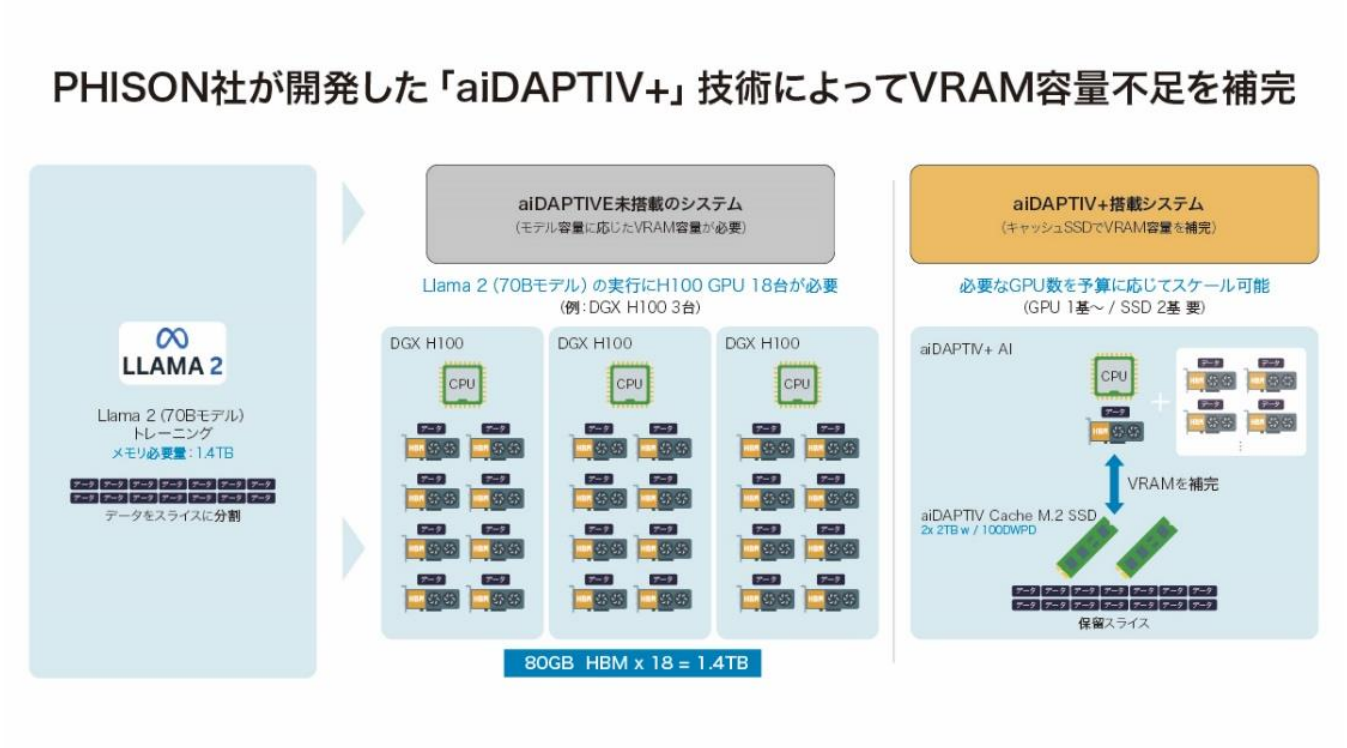
- PHISON 社の独自技術により、GPU VRAM を超える大規模モデルをシステムメモリと SSD で補いながら学習。
- Llama 3.x（8B～70B）、Gemma2（9B・27B）など多彩なモデルバリエーションにも柔軟対応。

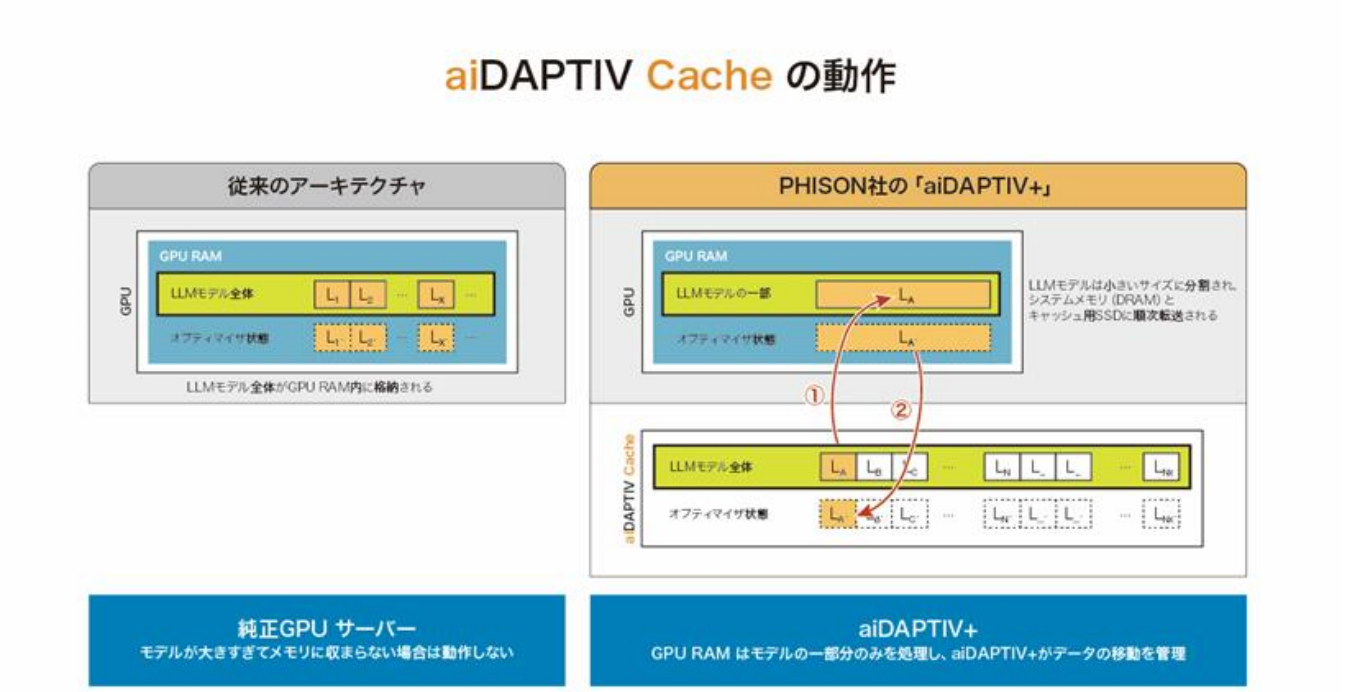
3. PoC 段階から大規模モデルまで段階的に拡張

- 単一ノード構成でまず PoC を行い、成果が確認できれば必要に応じて GPU やメモリを追加し、大規模 LLM にも移行可能。
- 巨額投資を初期段階から強いられないため、中小企業や研究機関でも無理なく AI 活用をスタート。

4. ノーコード GUI & Docker 環境による導入の容易化

- 複雑なセットアップやパラメータ調整をノーコード GUI や専用ミドルウェアが一体化してサポート。
- AI 分野に不慣れなチームでも、短期間でフルファインチューニングがスタートでき、DX 施策のスピードを損なわない。





性能比較例(PHISON社提供データ)

構成	ノード数	GPU	総VRAM容量	学習時間	コスト比	消費電力
NVIDIA DGX H100	3 台	24	1920 GB	約0.22 日	1	8.4KW
NVIDIA RTX 6000ada ×4 RAM:512GB aiDAPTIV+(2TBx2)	1 台	4	192 GB	約3.8 日	1/40 0.025	1.2KW

【製品構成例】

水冷式ワークステーション HPC3000-XSRGPU4TP-LC

Intel® Xeon® w5-3525 と **NVIDIA RTX 6000 Ada** をベースに、いずれも **NVMe Gen4 4TB** の SSD を標準搭載、主な違いはメモリ容量と GPU 枚数、aiDAPTIV+ キャッシュ用 SSD の容量です。用途や想定モデル規模に応じて**構成を柔軟にカスタマイズ**可能です。

※詳細は URL をクリックし、ご参照ください。

https://www.hpc.co.jp/product/hardware/hpc3000-xsrgpu4tp-lc/

構成	GPU 数	メモリ	aiDAPTIV+ SSD	想定モデル環境	主な用途
エントリー	1	64GB	2TB(1TBx2)	Llama 3.x 8B 、Gemma2 9B など	PoC・軽量モデル検証
ミドル	2	128GB	2TB(1TBx2)	Gemma2 27B など	本格運用（要約、翻訳等）
ハイエンド	2	256GB	4TB(2TBx2)	Llama 3.x（70B）	大規模 DX 施策、研究開発

【オフラインセミナー開催のご案内】

本ソリューションの導入を検討している企業・研究機関やメディア、システムインテグレーターの皆さまを対象に、**オフラインセミナー**を開催予定です。実機デモでノーコード GUI を用いたフルファインチューニング手順の実機デモンストレーションなどを直接ご覧いただけます。

日時・場所については、後日当社ウェブサイトにてご案内いたします。

【用語説明】

- **DX（デジタル変革）**
  業務プロセスやビジネスモデルを IT 技術で根本的に変革し、新たな価値や競争力を創出する取り組み。
- **LLM（Large Language Model）**
  大規模なデータをもとに学習した言語モデル。Llama や Gemma2 などが有名で、テキスト生成を中心とした高度な自然言語処理が可能。
- **フルファインチューニング**
  モデル全体を再学習し、ドメイン固有の知識を深く内在化する手法。精度が高くなる一方、GPU リソースを多く要する。
- **aiDAPTIV+**
  PHISON 社が開発した VRAM 補完技術。GPU のメモリ容量を超える大規模モデルも、システムメモリや SSD を活用することで学習を実行可能。

【HPC システムズについて】

HPC システムズは、ハイパフォーマンスコンピューティング（HPC）分野のニッチトップ企業です。
HPC 事業では、科学技術計算用高性能コンピュータとシミュレーションソフトウェア販売、科学技術計算やディープラーニング（深層学習）環境を構築するシステムインテグレーションサービス、シミュレーションソフトウェアプログラムの並列化・高速化サービス、計算化学ソフトウェア、マテリアルズ・インフォマティクスのプログラム開発・販売、受託計算サービス・科学技術研究開発支援、創薬研究開発や素材・材料研究開発分野向けサイエンスクラウドサービスをワンストップで提供しています。

また、CTO 事業では、顧客の用途、課題をヒアリングしながら、価格・性能・品質・高低温・防塵・防水・静電対策・過酷な環境に対する高耐久性など多種多様の対応が求められる、工場生産設備・製造装置・検査装置、制御機器や交通インフラ、自動運転、リテール店舗などのコントローラーとしての産業用コンピュータやエッジコンピュータの仕様提案から開発、生産、保守サポート、長期安定供給を実現しています。

社名 HPC システムズ株式会社　https://www.hpc.co.jp/
所在地 東京都港区海岸 3 丁目 9 番 15 号 LOOP-X 8 階
設立 2006 年 7 月 3 日
資本金 2 億 3,019 万円 (2024 年 9 月 30 日時点)
代表者 代表取締役 小野 鉄平

プレスリリースに関するお問い合わせ
https://www.hpc.co.jp/contact/company_form/