

AIワーク ロードをIBM Power上で実 行すべき5つ の理由

AI戦略を強化するための
信頼できる基盤



シームレスなAI統合により、Powerは総合的なビジネス・トランスフォーメーションを実現します。AIをトランザクションに組み込むにしろ、ハイブリッド環境をまたいで生成モデルの拡張やアプリケーションの導入を行うにしろ、IBM® Powerサーバーにはそれらすべてを、業務を中断させることなくこなせるだけの性能、セキュリティー、柔軟性が備わっています。それでは詳しく見ていきましょう。

1 Powerはデータが存在する場所にAIを導入

オンプレミス、クラウド、エッジと、データがあらゆる場所に存在する世界においては、スマートだけでなく戦略的なAIソリューションが必要です。PowerはAI推論を、データが生成され、存在する場所にもたらします。ネットワーク経由で機密データを移動させたり、効果で複雑なGPUを待たったりする必要はもうありません。Powerは、ミッションクリティカルなアプリケーションがまさに存在する場所にAI推論を組み込むことで、レイテンシー、セキュリティー・リスク、追加コストを削減できます。PowerのIBM Spyre AcceleratorとIBM watsonx.dataにより、最新のデータレイクハウスを高度なオフチップ・アクセラレーションと統合することで高速な洞察が獲得でき、信頼できるデータと強力な推論機能の架け橋が構築されます。

2 Powerの24時間365日セキュリティ

AIワークロードは価値のあるものですから、脅威に対して脆弱であってはなりません。Powerは、透過的なメモリー暗号化を用いてスタックの各レイヤーにセキュリティーを組み込み、パフォーマンスに影響を与えることなくインサイトを保護します。生成AIのような複雑なタスクのために、パフォーマンスに関する不安なしに、自身を持って推論を拡張できます。Powerサーバーでは、IBM Power Cyber Vaultにより1分未満でのランサムウェア検知¹、最大99.9999%のアップタイム²が保証され、レジリエントであらゆる用途に備えた常時稼働のAIが実現できます。

脅威検知の保証¹ アップタイムの信頼性²

3 Powerはハイブリッドの柔軟性でスムーズ

AIワークロードのデプロイに関しては、ハイブリッドITの柔軟性が非常に重要です。Powerサーバーは、ハイブリッドクラウド・エクスペリエンスを円滑に実現するための完全に最適化されたエンタープライズ・ソフトウェアを提供し、オンプレミス・システムとIBM Power Virtual Serverとの間におけるワークロードの移動を容易にします。クラウドにおけるモデルのトレーニングも、オンプレミスでの推論も、Powerではより多くのことを柔軟に行えます。さらに、130以上のオープンソースAIツールやパッケージをサポートするため、チームは障害にぶつかることなく構築、デプロイ、拡張できます。

4 Powerはサステナブルなパフォーマンスをオンデマンドで提供

AIワークロードではサステナビリティ要件は満たせないなどと言うのは誰でしょうか。Power11は、x86プロセッサベースのサーバーと比べてワットあたりの性能が2倍なので、同じワークロードを低消費電力で実行できます。³そしてPower11サーバーの新たなエネルギー効率モードにより、最大パフォーマンス・モードと比較して電力効率を28%改善できます。⁴

↑ 28%

電力効率に優れる
(X86 Serverとの比較)⁴

5 Powerは妥協なき高速化を実現

既存のワークロードを中断させることなく、大規模のAIをリアルタイムで実行したいとお考えでしょうか。IBM Powerサーバーは、まさにそのために構築されました。高度な並列処理、大容量メモリー、オンチップ・アクセラレーションにより、ワークフローにAIを直接組み込むよう設計されています。プラットフォームを使用するためにデータサイエンティストがリファクタリングを行う必要はなく、AIワークロードをそのまま実行できます。さらに、大規模言語モデルを実行する場合は、同時ユーザー数40人というピーク負荷時でも、Power S1022はx86プロセッサ搭載サーバーと比較して毎秒42%多くバッチ・クエリー処理を実行⁵でき、レイテンシーは1秒未満。⁶つまり、迅速な洞察、スムーズな操作、1秒未満の推論が実現できるということです。

↑ 42%

Power S1022で毎秒のバッチ・クエリー
処理量が増加 (X86 Serverとの比較)⁵

AIとIBM Powerのテクノロジーをより深く掘り下げる →

© Copyright IBM Corporation 2025

IBM、IBMのロゴ、IBM Spyre、Power、watsonx.dataは、米国およびその他の国または地域におけるIBMの商標です。その他の製品名およびサービス名はIBMまたは他社の商標である可能性があります。IBMの商標の最新リストは、ibm.com/jp-ja/legal/copytradeをご覧ください。

IntelとIntel Xeonは、米国およびその他の国におけるIntel Corporationまたはその関連会社の商標または登録商標です。

本書は最初の発行日時点における最新情報を記載しており、IBMにより予告なしに変更される場合があります。

IBMが事業を展開しているすべての国または地域において、すべての製品を利用できるわけではありません。

示されている例は、説明のみを目的として提供されています。実際の結果はお客様の設定や条件により異なるため、一般的に期待される結果を提供するものではありません。

IBM製品およびプログラムを使って他社製品またはプログラムの動作を検証する場合は、お客様の責任で行ってください。IBMは、他社の製品やプログラムに責任を負いません。

本書の情報は「現状のまま」で提供されるものとし、明示または暗示を問わず、商品性、特定目的への適合性、および非侵害の保証または条件を含むいかなる保証もしないものとします。IBM製品は、IBM所定の契約書の条項に基づき保証されます。

ITシステムや製品は完全に安全であると捉えるべきではなく、不適切な使用やアクセスを防止する上で

完璧な効果のある、製品、サービス、セキュリティー対策は1つもありません。IBMでは、いずれの当事者による不正行為または違法行為によっても、いかなるシステム、製品もしくはサービスまたはお客様の企業に対して影響が及ばないことを保証することはありません。

1. アラートを1分以内に表示することのみが保証の対象です。アラートが表示されない場合、修復措置として、対象製品の価格を上限としたドライブ交換を行います。諸条件が適用されます。詳細は[こちら](#)をご覧ください。
2. 単一のPower E1180システムにおける計画外ダウンタイムに基づいています。Power11 Processor-Based Systems RAS（セクション「99.9999%のアップタイム」参照）
<https://www.ibm.com/downloads/documents/jp-ja/10a999803d9afd7776>
3. 2025年5月15日時点のIDCの定量的パフォーマンス指数（QPI）データ（<https://www.idc.com/about/qpi>）および利用率に基づいています。IBM Power E1150（4x30コアPower11、3.0~4.1 GHz）のQPIは241,000Eであり、HPE Compute Scale-up Server 3200（4x60コアIntelコア、1.9 GHz）のQPI 208,898に対して優位です。利用率はIBM Power Performance Utilization Guaranteeに基づきE1150が75%、x86が40%とされています。
- エネルギー消費量は最大入力電力に基づいています。IBM Power E1050の最大電力は5,200Wです（<https://www.redbooks.ibm.com/redpapers/pdfs/redp5684.pdf>）。
- HPE Compute Scale Up Server 3200の最大電力は4,740Wです（https://www.hpe.com/psnow/doc/a50004268enw.html?jumpid=in_pdp-psnow-q3）。
4. フル構成のソケットとメモリー（E1180（10c×4 / 64GB DDIMM×64）、E1150（16c×4 / 32GB DDIMM×64）、S1124（16c×2 / 32GB DDIMM×32）、S1122（16c×2 / 32GB DDIMM×32）を搭載したPower11システムで、コンビューティング、ディスク、メモリーベースのワークロードを実行中に、最大パフォーマンス・モードとエネルギー効率モードを比較したサーバーのワットあたりのパフォーマンスのIBM測定に基づきます。
5. PrimeQAモデルを使用した質疑応答推論のIBM内部テストに基づく比較（<https://github.com/primeqa>、Dr. DecrモデルおよびColBERTモデルに基づく）。結果は2023年8月22日時点で有効で、実験室の条件下で実施されました。個々の結果は、ワークロードのサイズ、ストレージ、サブシステムの使用、その他の条件によって異なる場合があります。比較は、SMT 8を実行しているIBM Power S1022（1x20コア / 512GB）とIntel Xeon Platinum 8468ベース（1x48コア / 512 GB）システムでの毎秒のスコア（推論）の合計スレーブットに基づいています。テストは、Python 3.10やPyTorch 2.0を含むPython環境およびAnaconda環境で実施されました。使用したPythonライブラリーは、PowerとIntel用にプラットフォームを最適化しています。構成：バッチ・サイズ=60、同時ユーザー数=40人、torch.set_num_threads(int)は、さまざまな負荷レベルで最適化されています。
- IBM Power S1022（<https://www.redbooks.ibm.com/abstracts/redp5675.html>）：40人の同時ユーザーで毎秒6.26のバッチ・クエリー推論
- 比較対象のx86システム：Supremacy SYS-221H-TNR（<https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr>）：40人の同時ユーザーで毎秒4.4のバッチ・クエリー推論。
- IBM内部データのコーパスによってファイン・チューニングされたモデル。
6. PrimeQAモデル（Dr. DecrおよびColBERTモデルに基づく）を用いた質問応答推論のIBM社内テストに基づく。結果は2023年8月31日時点で有効、実験室条件下で実施され、ワークロードサイズ、ストレージサブシステムの使用状況などにより個別結果は異なる場合があります。テストはIBM Power S1022（2x20コア 2.9~4 GHz / 512GB）、2チップNUMAライン済み10コアPLPARで実施。PythonおよびAnaconda環境（Python 3.10、PyTorch 2.0パッケージ含む）で実施。使用したPythonライブラリーはPower向け最適化済みライブラリー。設定：SMT 2、torch.set_num_threads(16)、バッチ・サイズ=1。
- IBM Power S1022（<https://www.redbooks.ibm.com/abstracts/redp5675.html>）
- PrimeQAモデル（<https://github.com/primeqa>）