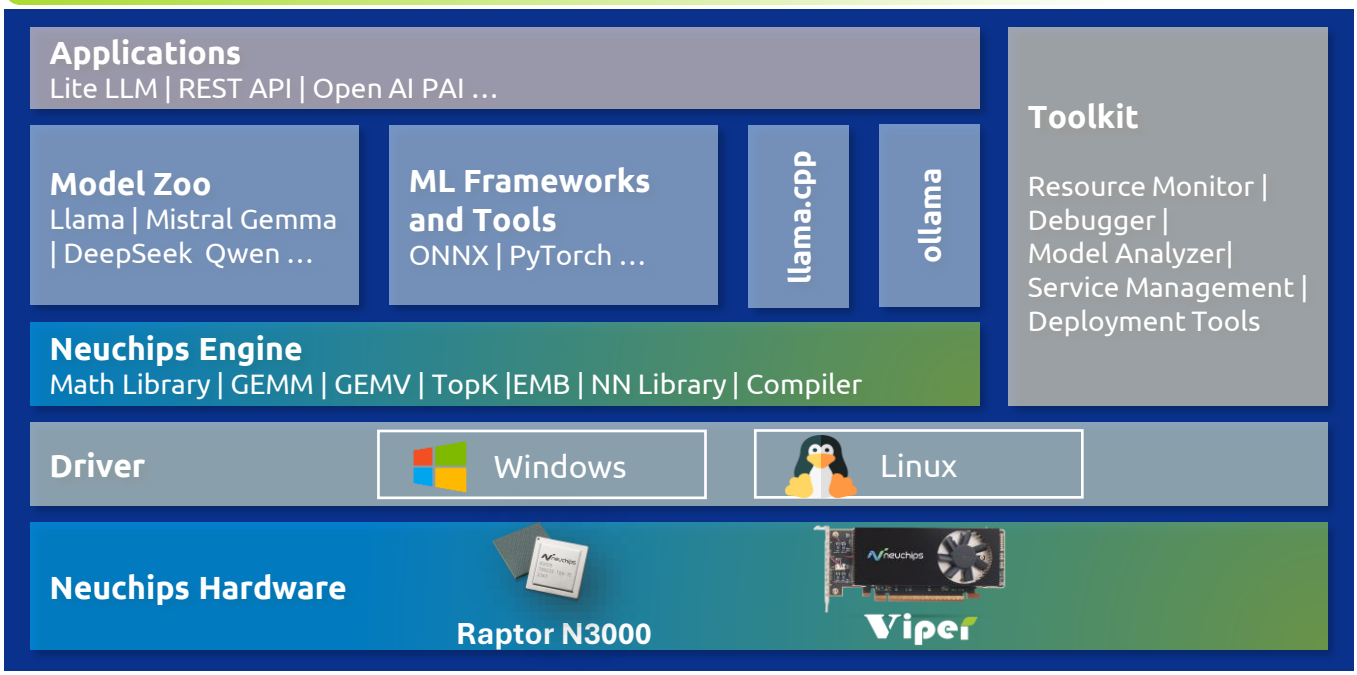
次世代の企業内推論を加速する ——
法人向けオンプレミス生成AI特化型アクセラレーター

仕様	説明
対応 LLM モデル (単体カードで最大 14B まで)	Llama, Mistral, Gemma, Phi, TAIDE, Qwen, DeepSeek Distilled Models
最大消費電力 (TDP)	45W
冷却方式	アクティブ冷却対応
形状 / フォームファクタ	HHHL-SS (ハーフハイト・ハーフレングス、シングルスロット)
PCI デバイス ID	Device ID: 0x1000 / Vendor ID: 0x1FD9
プロセッサ動作周波数	1 GHz
PCI Express インターフェース	PCI Express 5.0 x 8 (レーンおよび極性反転対応)
FX エンジン	2
エンベディング (Embedding)	1
演算性能 (BF16)	32 TFLOPS
演算性能 (INT8)	206 TOPS
対応 OS	Linux / Windows
メモリタイプ	LPDDR5
メモリ容量	64GB
メモリクロック	6400 Mbps
動作環境温度	0 °C ～ 50 °C
保存環境温度	-40 °C ～ 75 °C
動作環境湿度	5% ～ 85% 相対湿度
保存環境湿度	5% ～ 95% 相對濕度



Neuchips 生成 AI ソフトウェアスタック

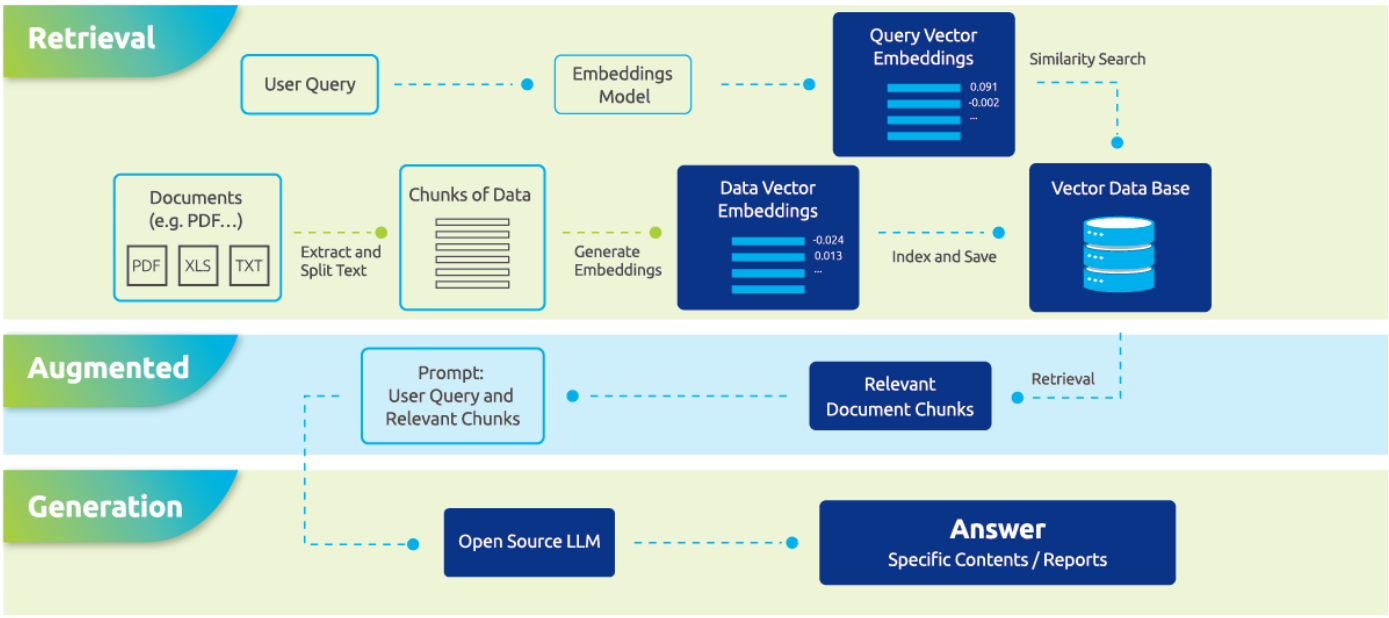


Neuchips | © 2019 - 2026 | All Rights Reserved

専用プロセッサの力を活用する

データプライバシーを最優先し、シームレスなオフラインAIソリューションを求める企業にとって、**Neuchips Viperシリーズ**は包括的な答えを提供します。統合されたAIアプリケーションにより、ユーザーは既存のオフィス環境内でシステム全体を安心して利用でき、機密データの安全性を確保することが可能です。
さらに、Neuchips Gen AIプロセッサ「Raptor N3000」には、効率的なエンベディング・エンジンが内蔵されています。これにより、ViperシリーズとホストCPU間の通信リソースとオーバーヘッドを効果的に削減します。この最適化は、専門分野や特定の業界においてRAG（検索拡張生成）を活用する企業のオフラインアプリケーションに特に大きなメリットをもたらします。RAGおよびLLM（大規模言語モデル）のワークロードをオフロードし、処理速度を向上させることで、Viperシリーズは比類のない効率性を実現します。また、Viperには革新的な量子化技術が統合されており、ベクトルデータのサイズを4倍以上に効率よく圧縮します。この技術によりベクトルを削減し、データベースの容量を拡張することで、ストレージ効率に革命をもたらします。
さらに、ボード上に48GBの追加メモリー容量を搭載しているため、Viperはボード上で直接ベクトルデータベースを構築することが可能です。これにより、メモリー容量を拡大させつつ、機密データのローカルでの安全性を保証します。これは、膨大なデータ処理を必要とするAI技術を採用する際に特に有効です。CPUとRaptor間の不要なオーバーヘッドを排除することで、Viperはパフォーマンスと生産性を大幅に向上させます。

Neuchips Viper における生成 AI ハードウェア加速化の役割



- 節約を最大化し、AIコストを最小化
- AI運用を即座にトランスフォーム
- 効率的なコンパクト設計、ハードウェアの煩わしさゼロ
- オペレーションを効率化し、アウトプットを最大化
- オフラインセキュリティで、データ漏洩ゼロ
- 低消費電力で、高いサステナビリティを実現

Neuchips | © 2019 - 2026 | All Rights Reserved