

A Frictionless AI Solution for Enterprise



Copyright © 2019 - 2024 Neuchips. All rights reserved.



Neuchips Inc.

Hsinchu | Taipei | Bay Area



www.neuchips.ai | contact@neuchips.ai



Transform Your Operations :

Empower Your Enterprise with Seamless Offline AI Integration!

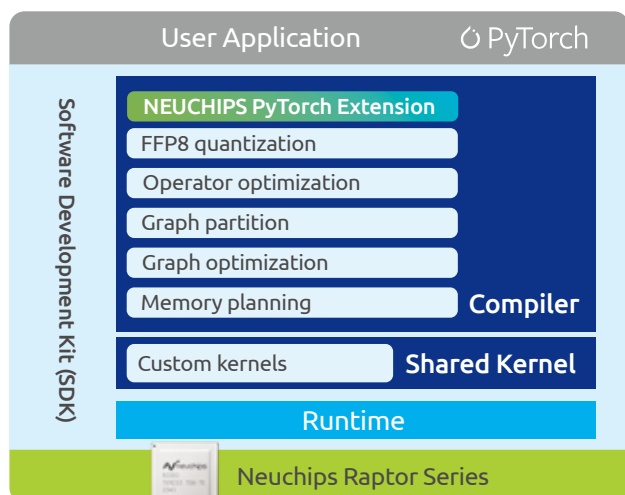
In today's rapidly evolving digital landscape, enterprises are constantly seeking innovative solutions to optimize their AI applications. With the advent of Neuchips Viper series, a cutting-edge technology designed for enterprise AI, the possibilities are endless. Let's delve into the features that make Viper series the perfect fit for your organization's AI needs.

Offline AI Solution	Marketing & Sales	Operations	IT & Engineering	Risk & Legal	HR	Utility & Manufacturing
Domain Focus AI Application	<ul style="list-style-type: none"> • Create Product Literatures • Analyze Customer Feedback • Customer Support Service... 	<ul style="list-style-type: none"> • Identify Production Yield Rate • Automatically Process & Agent • Document Analysis ... 	<ul style="list-style-type: none"> • Create Technical Documentation • Automatically Generate Data ... 	<ul style="list-style-type: none"> • Draft & Summarize • Legal Documents Summarize and Highlight ... 	<ul style="list-style-type: none"> • Assist in Interview • HR Training System • Candidate Assessment ... 	<ul style="list-style-type: none"> • Search & Question Answering • Optimize Employee Communication • Presentation Foil Creation • Document Extraction & Data Analysis ...
System Integrator	HW & SW System Integration Service Domain Focus API Management					
Software Service Provider / Application Interface	Device Management Platform Fine Tuning Document Partner RAG... etc.					
Open Sorce AI Models	Gen AI / LLM Models (TAIDE Llama2 Llama3 Mistral Breeze Phi2)					
System Hardware Manufacturer	PC IPC Workstation Server					
Neuchips Products	<div> <div>Raptor Series Gen AI Processor</div>  <div>Viper Series Gen AI PCIe Card</div>  </div>					

Raptor Series Specification

Embedded ARC HS48 Processors	<ul style="list-style-type: none"> ■ CPU0 - Quad Core with MMU Function ■ CPU1 - Dual Core
Embedded ARC EV72 Processors	<ul style="list-style-type: none"> ■ Vector DSP ■ FPU
Communication Interface	PCIe Gen 5 x 8
Memory	<ul style="list-style-type: none"> ■ 16 LPDDR5-6400 Controllers ■ 512KB Shared Memory
AI Accelerators	<ul style="list-style-type: none"> ■ Matrix Engines x 10 (Dynamic MLP Engine, DME) ■ Vector Engines x 2 (Feature Cross, FX) ■ Embedding Engine
Crypto Engines	Hardware Root of Trust with Processor, TRNG, and OTP Storage
Peripherals	QSPI Flash Controller, I ² C, SPI, UART, GPIO

Easy-to-Use SDK



Viper Series Specification

Specification	Description
Support LLM Model	Llama2, Llama 3, Mistral, Breeze, TAIDE, Phi2
Total Board Power	55W
Thermal Solution	Passive
Mechanical Form Factor	HHHL-SS (half-height, half-length, single-slot)
PCI Device IDs	Device ID: 0x1000 / Vendor ID: 0x1FD9
Processor Clock Speed	1 GHz
FX Engine	2
Embedding	1
TFLOPS (BF16)	32
TOPS (INT8)	206
PCI Express Interface	PCI Express 5.0 x 8 Lane and Polarity Reversal Supported
Support OS	Linux / Windows
Memory Type	LPDDR5
Memory Size	Up to 64GB
Memory Clock	6400 Mbps
Ambient Operating Temperature	0 °C to 50 °C
Storage Temperature	-40 °C to 75 °C
Operating Humidity	5% to 85% Relative Humidity
Storage Humidity	5% to 95% Relative Humidity



Ultra Low Power Consumption



Offload 90% AI Workload from CPU



Precision Redefined: 10x Efficiency with Embedding Engines



Expand Possibilities with Extra 48GB Vector Database



Support Multiple Open -Sourced LLM Models



100% Made in Taiwan

Harnessing the Power of Specialized Processors

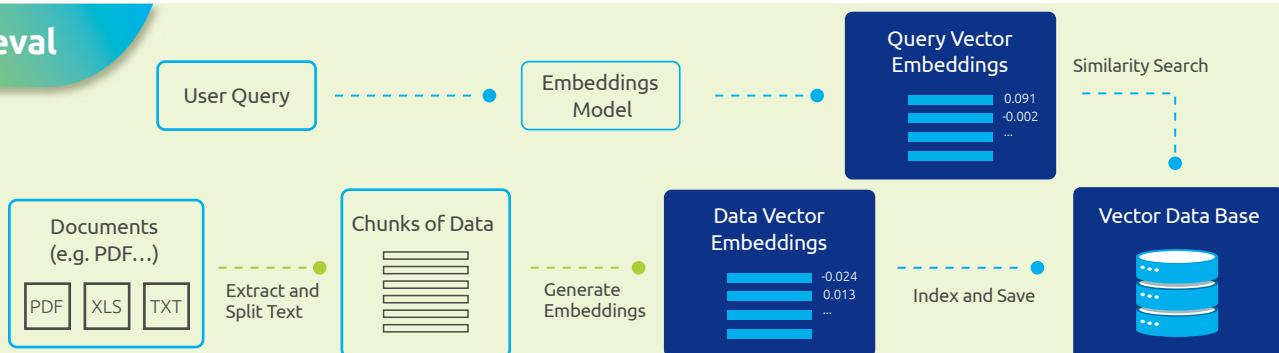
For enterprises seeking a seamless offline AI solution prioritizing data privacy, **Neuchips Viper series** offer a comprehensive answer. With the integrated AI application, users can confidently utilize the entire system within their existing office environment, ensuring sensitive data remains secure.

Furthermore, **Neuchips Gen AI processor Raptor** built-in with efficient embedding engine, effectively reducing communication resources and overhead between Neuchips Viper series and host CPU. This optimization is particularly beneficial for enterprise offline applications leveraging RAG (Retrieval Augmented Generation) support across specialized domains and industries. By offloading RAG and LLM workload and enhancing processing speeds, Neuchips Viper series deliver unparalleled efficiency.

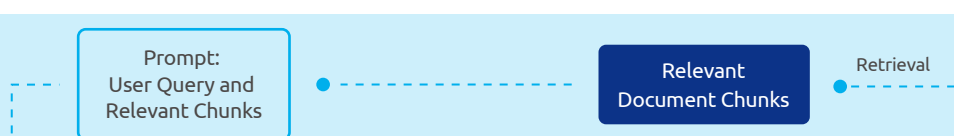
Moreover, the Viper integrates an innovative quantization technique, which efficiently compresses vector data size by more than 4 times. This technique eliminates vectors and expands the database capacity, thereby revolutionizing storage efficiency. With onboard extra 48GB memory capacity, Viper facilitates a vector database directly on the board, enlarging memory capacity and ensuring the security of sensitive data locally. This is particularly beneficial when employing AI technologies that demand extensive data processing. By eliminating the need for additional overhead between the CPU and Raptor, Viper significantly enhances performance and productivity.

The Role of Hardware Acceleration on Neuchips Raptor

Retrieval



Augmented



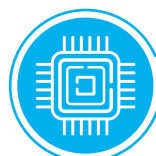
Generation



Maximize Savings,
Minimize AI Costs



Transform AI
Operation Instantly



Efficiently Compact,
No Hardware Hassle



Streamline Operations,
Maximize Output



Offline Security,
Zero Data Leaks



Low Power,
High Sustainability