

VDPU

Revolutionary Vector-Search Accelerator

Purpose-Built ASIC
for Complex Data Search



Vector Data Processing Unit

 Dnotitia

The Explosion of Unstructured Data

AI is generating and consuming data at unprecedented scale. To unlock its value, the next leap requires a new hardware architecture.

1 Data Surge



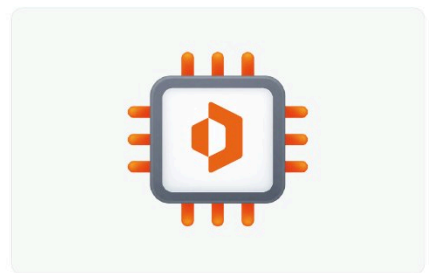
Multi-modal, multi-source unstructured data is growing exponentially.

2 Efficiency Ceiling



Legacy architectures have reached the limits of software optimization.

3 Need fundamental shift in hardware



A new hardware architecture is required to process and retrieve data at AI scale.

VDPU

Specialized Hardware for AI Agent Data Retrieval

Data-access of AI agents



AI uses a semantic interface (meaning-based vector retrieval)



AI agents run the core loop thousands of times, repeatedly retrieving vectors until they find grounded answers

Vector Data Processing Bottlenecks

Dnotitia VDPU & VDB



Phase 1

Accelerates Retrieval Memory

Vectors, Graphs



Phase 2

Extends to Execution Memory

KV cache, Contexts

**We Need an AI Revolution for the upcoming era of
"High Scale, High Precision, and AI Agentic Search"!**

Closing the **Memory Gap** Between CPU and GPU



CPU Inefficiency

- General-purpose CPUs excel at complex logic, but face utilization limits from continuous vector similarity calculations
- Offloading this specific task preserves CPU cycles for core application processes while optimizing the overall infrastructure thereby reducing TCO



GPU Latency Trap

- While powerful for training, GPUs are architected for massive batch processing, and inherently inefficient for real-time, individual vector queries (Frequent Random I/O)
- Unavoidable latency when immediate retrieval is required

CPU

Memory Capacity **Several TB possible**



Memory Bandwidth **Very Low**



GPU

Memory Capacity **~100GB Limit**



Memory Bandwidth **Very High(HBM)**



⊗ **Not Suitable for Vector Workloads**

VDPU

Perfect for Vector Workloads



Memory Capacity **256GB + Sufficient**

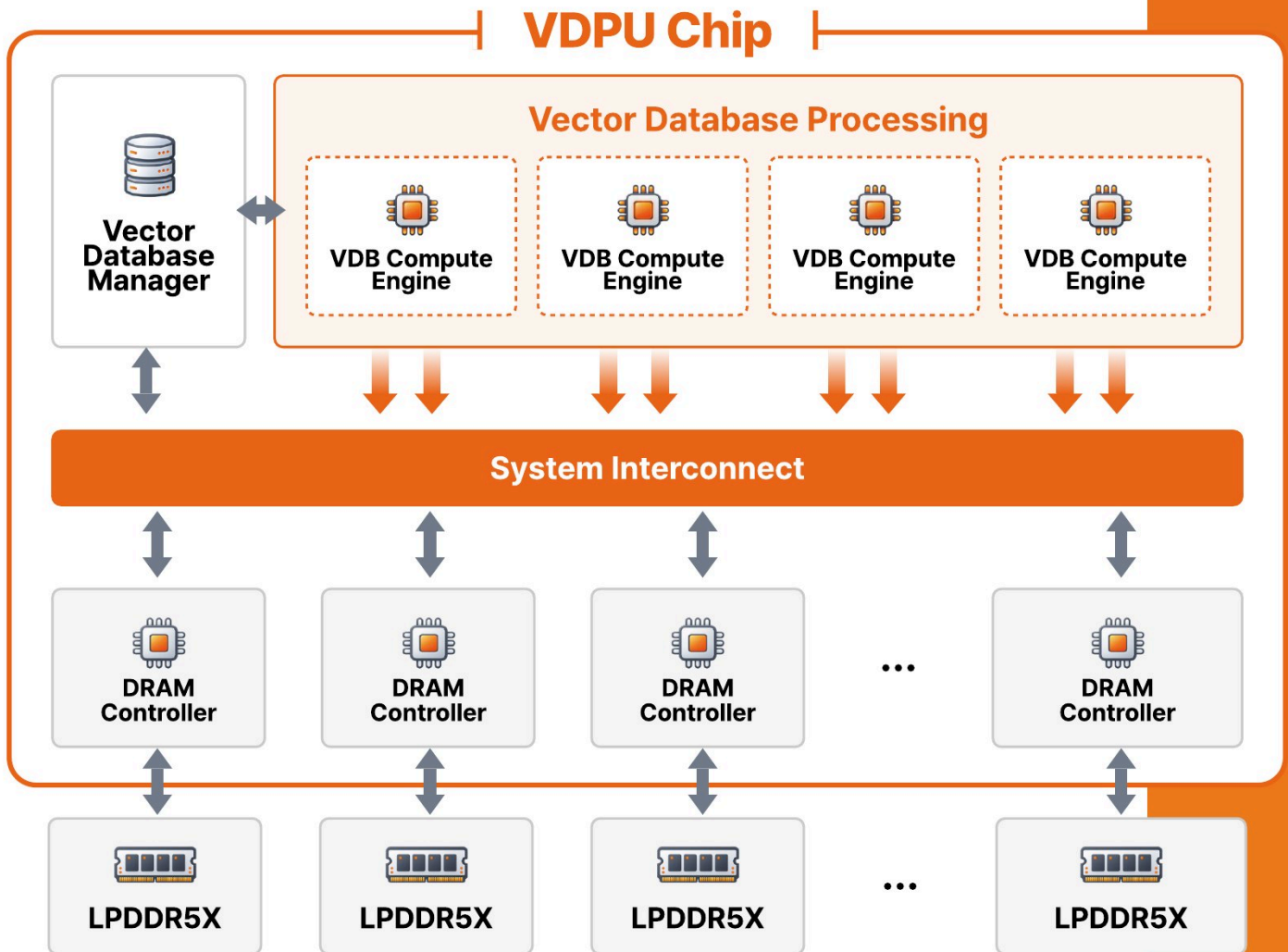


Memory Bandwidth **Balanced Support**



✓ **Optimized for both Memory Capacity and Bandwidth**

Dedicated Hardware for Vector DB Workloads



Uniqueness of VDPU Architecture

- ✓ Optimized DRAM architecture for maximum memory bandwidth utilization
- ✓ Near-Data Processing (NDP)-enabled AI storage architecture
- ✓ LP DDR-based batch processing with shared system memory
- ✓ Scalable design for large-scale data workloads
- ✓ Optimized for data-intensive random I/O (e.g., recommendations, graph search)
- ✓ Supports ENNS and graph/cluster-based ANNS indexing algorithms
- ✓ Compatible with major vector databases and libraries
- ✓ Seamless integration with AI software stacks, data ingestion, and agent workflows

Unbeatable Performance & Efficiency

More speed. Lower cost. Smaller footprint.
VDPU delivers a step change in vector search infrastructure.

10x
Faster speed

*Performance comparison per server

VDPU-based acceleration

Dnотitia

10x

Open source Vector DB

Company Z

Company W

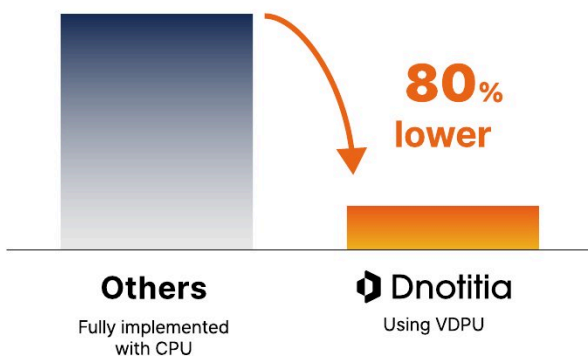
Commercial Vector DB

Company Z

Company E

80%
Reduced TCO

*Cost based on same performance



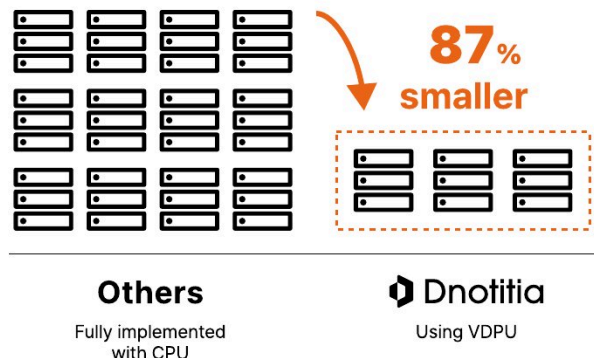
Others

Fully implemented with CPU

Dnотitia

Using VDPU

87%
Reduced floor space



Others

Fully implemented with CPU

Dnотitia

Using VDPU



Expected performance with ASIC card

VDPU is designed to maximize performance, efficiency, and scalability for real-world vector search workloads.

V110 — Our First VDPU on Silicon, Now in Bring-Up

First samples on the bench, and the software they will run is already proven on FPGA



V110 chip package and accelerator card



Taped out
First samples on the bench, with block-level function verified



Most of the retrieval stack is proven
Milvus, FAISS and hnswlib already run end-to-end on our FPGA platform



That same stack is now being brought up on silicon
The same libraries, running closer to the data



Chip-level characterization under way
Silicon performance will be evaluated using the same benchmarks



FMS 2026 Best of Show Award
AI Application award



CB Insights 2025
AI 100 — Vector databases
AI Agent Tech Stack — Context layer



PoCs under way
A Tier-1 silicon partner, global memory and storage vendors

Key Features of V110 and V200

VDPU approach	V110 in bring-up	V200 in design
In-storage vector semantic search acceleration Slow, inaccurate retrieval caps service quality	●	●
Zero-copy direct data path Transfers routed through CPU memory add latency	●	●
KV cache compression Growing context inflates GPU, storage and network cost	○	●
Inference context memory tiering Distributed inference and long context saturate GPU memory	○	●



Contact Us

Let's build the next generation of AI infrastructure together.

Website

dnotitia.com

E-mail

contact@dnotitia.com

Tel.

+82-2-588-1002