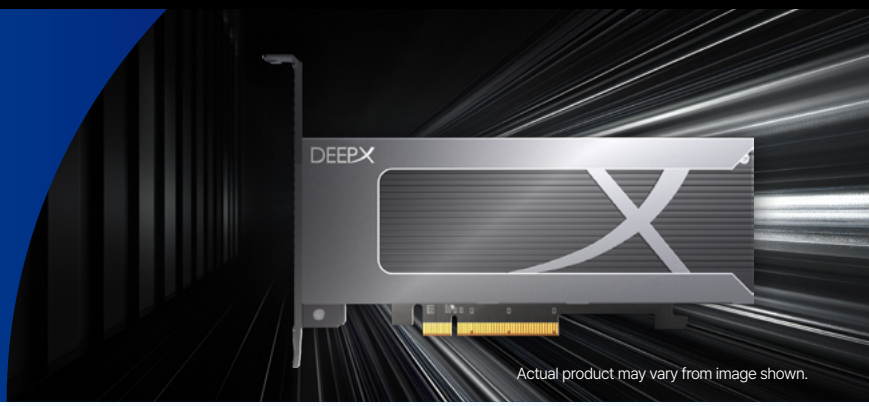
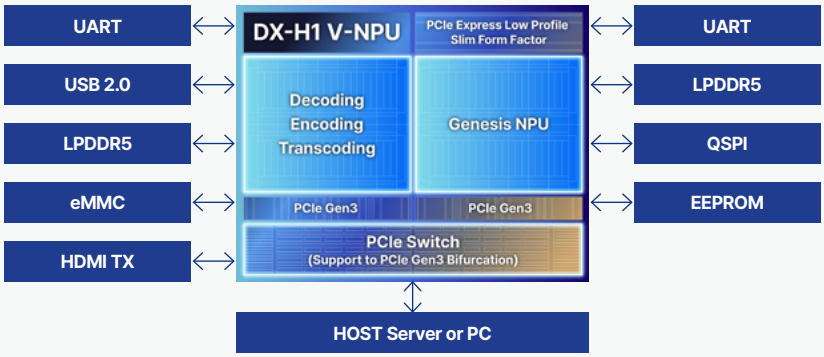
DX-H1 V-NPU

Integrated decoding, encoding, and transcoding turn the DX-H1 V-NPU into a complete foundation for real-time, large-scale monitoring. What once required rooms full of expensive, power-hungry servers can now be placed into the hands of any organization. **The barrier to advanced AI falls away, and what remains is possibility, accessible, sustainable, and inevitable.**



Power redefined — massive AI video at 40W, making intelligence sustainable at scale

Functional Block Diagram



Specifications

Feature	Details
AI Performance	50 TOPS (=400 eTOPS / INT8) for real-time AI inference
SRAM	Genesis NPU: Internal SRAM, Codec Chip: 256 KB to 512 KB
CPU	Codec Chip: 4x Cortex-A76 (up to 2.4 GHz)
	Codec Chip: 4x Cortex-A55 (up to 1.8 GHz)
Memory	Genesis NPU: Total = LPDDR5 8GB (4GB + 4GB)
	Codec Chip: Total = LPDDR5 16GB (8GB + 8GB)
Storage	eMMC 32GB
Power Consumption	40W / 12V
PCIe	PCI Express Gen3 16X Using Data x8 Lanes
Video Decoding	H.264 64CH (1080P @ 30FPS)
Video Encoding	H.264 32CH (1080P @ 30FPS)
Display	HDMI 2.0
Dimensions	167mm x 56mm x 18.8mm
Host Hardware	x86, ARM Based Architecture

DEEPX

Key Benefits

Cost-Effective: Drastically lowers hardware and electricity costs.

Integrated Design: All-in-one architecture ensures stable 24/7 video processing on a single card.

Efficient Monitoring: Reduces bandwidth and storage by sending only key events and enables fast AI-powered video searches.

Easy to Scale: Compact, fanless design allows high-density deployment for maximum efficiency.

Support DXNN® SDK

DXNN® SDK is a comprehensive SW development environment for deploying AI on DEEPX NPUs. It integrates tools for compiling, optimizing, simulating, and inferring the latest AI models, such as YOLO, ViT, and VLMs. And it provides an optimized, ready-to-use environment as the DX-All Suite package to support fast and efficient AI development.

Target Applications

- Security
- Smart Buildings
- Smart Factory
- Smart Cities
- Smart Safety Monitoring
- Content Delivery Network (Smart Search)
- Logistics & Warehouses
- Smart Transportation
- Smart Retail

DEEPX HQ
5F, 20 Pangyoyeok-ro 241beon-gil, Seongnam-si, Gyeonggi-do, South Korea

USA
1735 Technology Drive Suite 740. San Jose, CA U.S.A

Taiwan
Nanjing E. Rd., Zhongshan Dist., Taipei City 104, Taiwan (R.O.C.)

Europe
Coming Soon

Buy Now! - DX TechBridge Program

Contact Sales - sales@deepx.ai

DXNN® SDK

DXNN® (DEEPX Neural Network) SDK streamlines AI deployment on DEEPX NPUs by integrating version-aligned tools for compilation, optimization, simulation, and inference. For **efficient development**, it's offered as the **DX-AS (All Suite), a fully integrated and optimized package.**

How It Works: 4-Step AI Deployment with DXNN® SDK

Step 1

Prepare Your ONNX Model

Convert a pre-trained AI model from frameworks such as TensorFlow, PyTorch, or scikit-learn into the ONNX format, which enables model interoperability.

AI Model Traning

TensorFlow | TensorFlow Lite | Keras | XGBoost | PyTorch

↓

User Must Convert to ONNX Before Input

Step 2

Convert Model for DX-NPU

Use the DX-COM compiler to convert the ONNX model into a format optimized for DEEPX NPU hardware, ensuring efficient execution on the target NPU.

User Must Convert to ONNX Before Input

↓

DX-COM™

↓

DEEPX AI Model

Step 3

Real-time Operating DX-RT

Deploy the optimized model using the DX-RT runtime API. Input data is fed into the model, and inference results are produced in real-time on-device.

DEEPX AI Model | Input Tensor | Output Tensor

↓ ↓ ↑

DX-RT™ API

↕

DEEPX AI Products

Step 4

Run Your Application

Apply NPU-generated AI inference results to applications like Image & Video Analytics, Object Detection, Robotics and Automation, and Smart Homes and Factories.

DEEPX AI Products Applications

↓

Image & Video Analytics | Object Detection

Robotics and Automation | Smart Homes and Factories

DXNN® Full Stack Architecture

DXNN® Full Stack Architecture streamlines AI model deployment onto DEEPX products using its two-stage AI Model Compile and Runtime Environments.

1) AI Model Compile Environment

AI Model Input Sources

TensorFlow | TensorFlow Lite | PyTorch | Keras | XGBoost

↓

ONNX User Model
User Must Convert to ONNX Before Input

↓

AI Model Compiler

DX-COM®
NPU Model Compiler

DX-TRON
Graphical Model Viewer

↓

DX-Model Zoo
DEEPX Pre-compiled Models

or

Compiled User Model
Optimized Binary for DEEPX NPU

2) AI Model Runtime Environment

AI Model Input Sources

DX-APP
SDK Usage Examples

User App
Your Custom Application

AI App Dev Tool

DX-STREAM™

gstreamer

Ecosystem

aws AWS IoT Greengrass | WNDRVR | NetworkOptix

CVEDIA | VCA | Baidu 百度 | 飞桨

DX-RT™

C++-based DX-RT™ API | Python-based DX-RT™ API

Core Runtime Library

Device Drivers (Windows or Linux)

↕ PCIe

DEEPX AI Products

DX-M1™	DX-M1M	DX-H1	DX-V3	DX-M2™
AI Accelerator	AI Accelerator	AI Accelerator	AI Vision SoC	GenAI Accelerator

DEEPX HQ
5F, 20 Pangyoyeok-ro 241beon-gil, Seongnam-si, Gyeonggi-do, South Korea

USA
1735 Technology Drive Suite 740. San Jose, CA U.S.A

Taiwan
Nanjing E. Rd., Zhongshan Dist., Taipei City 104, Taiwan (R.O.C.)

Europe
Coming Soon

Buy Now! - DX TechBridge Program

Contact Sales - sales@deepx.ai