The Era of Free Thinking is Coming

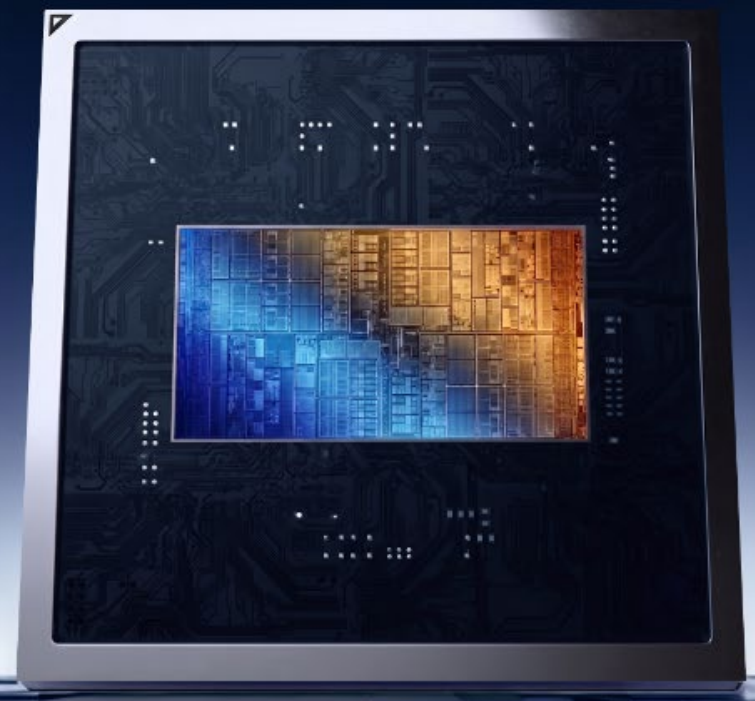
The DX-M2 Era

You are not renting intelligence anymore. You own it.

Ultra-low Power ~5W	World's First NPU 2 nm
Supported AI Model Size ~100 B	Single-Batch Performance ~30 TPS
NPU Performance 80 TOPS	Memory Bandwidth 153.6 GB/s

6 CES Innovation Award Winner 2024 · 2026 · 2027

CES Innovation Awards 2027 Honoree
Artificial Intelligence



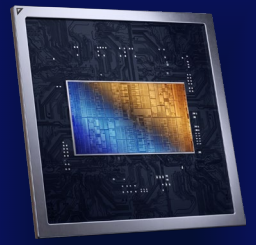
Chip Line-up

Choose Your Power.

Match your scale and accelerate your path to production with absolute confidence.

DX-M2

FcBGA 16x16



External Memory
Up to 96GB
(LPDDR5X 24GB x 4EA)

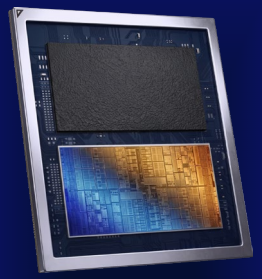
Power Consumption
5W (Only SoC)

Max Bandwidth
8-Channel (153.6 GB/s)

Form Factor
M.2 Module or PCIe Card

DX-M2M

FcMCM 19x19 (Tentative)



External Memory
Up to 24GB
(LPDDR5X 24GB x 1EA)

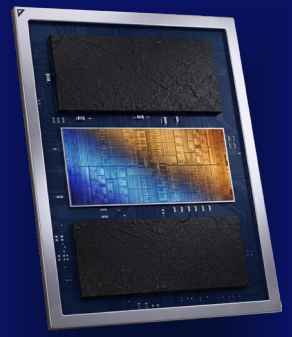
Power Consumption
5W (SoC 3W + DRAM 2W)

Max Bandwidth
4-Channel (76.8 GB/s)

Form Factor
M.2 Module

DX-M2M-Pro

FcMCM 29x21 (Tentative)



External Memory
Up to 48GB
(LPDDR5X 24GB x 2EA)

Power Consumption
8.5W (SoC 4.5W + DRAM 4W)

Max Bandwidth
8-Channel (153.6 GB/s)

Form Factor
M.2 Module

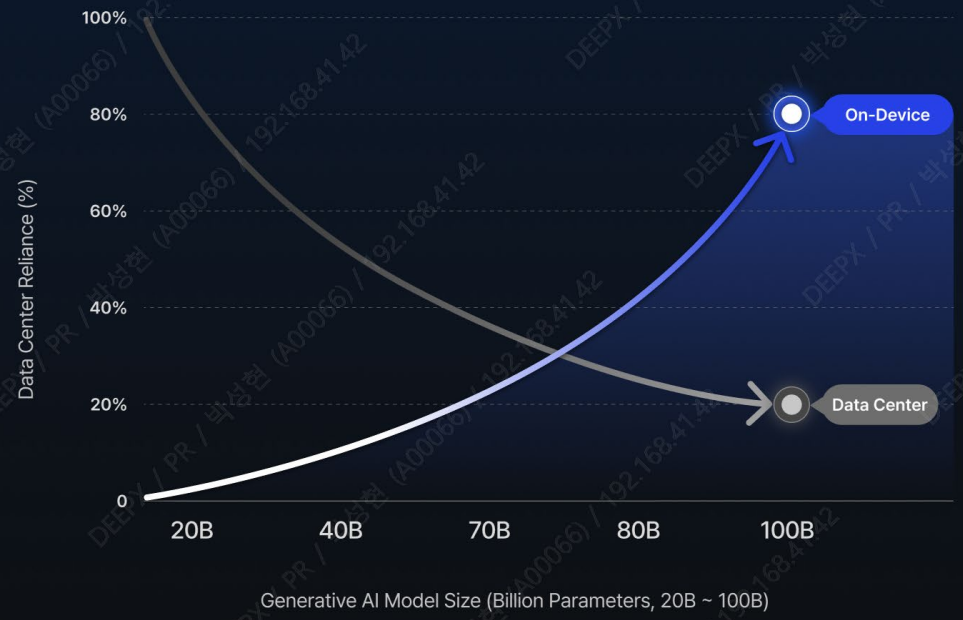
The Genesis of DX-M2

Breaking Cloud Ceilings with On-Device AI.

To scale generative AI, next-generation Physical AI must run natively on high-performance On-Device processors.

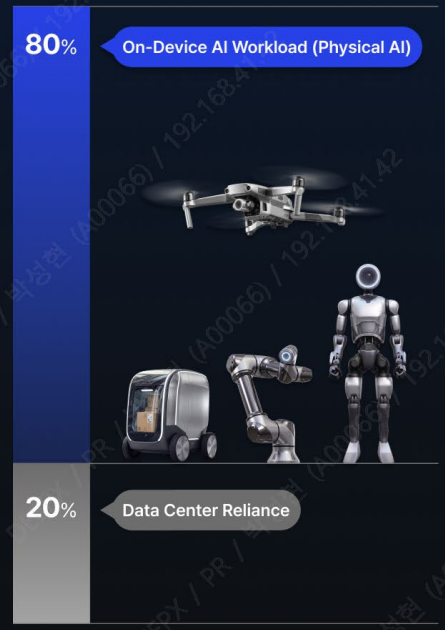
The Cloud AI Breakeven Point

Dependency Trend by Model Size



Shifting Workloads

Physical AI Takes the Edge



Cloud Bottlenecks

Infrastructure expansion is constrained by power grid volatility and skyrocketing cloud costs.

Native Edge Demand

The market demands native LLM execution on high-performance edge devices.

The DX-M2 Solution

DX-M2 resolves these constraints with architectural innovations for next-gen physical AI.

*Specifications are subject to change without notice during development.

DEEPX

DXNN SDK for DX-M2

Connect.
Optimize.
Unify.

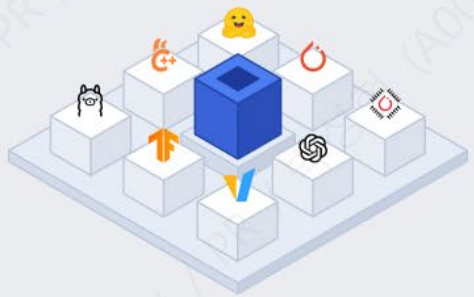
Engineered to seamlessly connect your existing stack, maximize M2 NPU performance, and unify deployment under a single runtime.

01 Connect

Bring Your Own Stack

Seamlessly run your existing frameworks and APIs on DX-M2 with zero code modification.

- OpenAI-compatible API
- PyTorch
- llama.cpp
- Ollama
- vLLM
- ExecuTorch

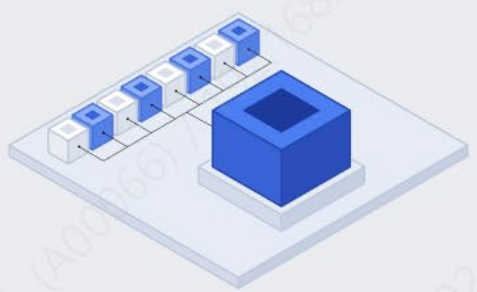


02 Optimize

Tuned to the Metal

Unlock peak hardware performance with an M2-native SDK that extracts every TOPS from silicon.

- DX-LLM: M2-native Runtime Framework
- Model / Memory / Compute Optimization
- NPU / DSP Compute Library



03 Unify

One Layer to Run It All

Consolidate standard frameworks and DX-LLM workloads under one unified middleware layer.

- M2 Middleware · Driver HAL
- Unified Compiler & Custom Kernel Build



*Specifications are subject to change without notice during development.

DEEPX

Universal Model Ecosystem

DX-M2 Optimized Model Zoo.

Every model is pre-optimized for the DX-M2 NPU and served through DX-LLM. These are the models we measured ourselves — quantization variants benchmarked against the bf16 baseline, on-device, with zero token fees.

Model	Task	Input	Output	Result
Qwen3-VL-2B Alibaba	Image-Text-to-Text	Text · Image · Video	Text	Deploy
Qwen3-VL-8B Alibaba	Image-Text-to-Text	Text · Image · Video	Text	Deploy
gpt-oss-20b OpenAI	Text Generation	Text	Text	Deploy
Llama-3.1-8B Meta	Text Generation	Text	Text	Deploy
DeepSeek-R1-Distill-Llama-8B DeepSeek	Text Generation	Text	Text	Deploy
Qwen3-4B Alibaba	Text Generation	Text	Text	Deploy
Qwen3-1.7B Alibaba	Text Generation	Text	Text	Deploy
Qwen3-0.6B Alibaba	Text Generation	Text	Text	Deploy
whisper-large-v3-turbo OpenAI	Speech Recognition	Audio	Text	Deploy
whisper-large-v OpenAI	Speech Recognition	Audio	Text	Deploy

*Specifications are subject to change without notice during development.

DEEPX