DEEPX

Last Generation Paid Per Thought

# The DX-M2 Era

You are not Renting Intelligence Anymore. You Own It.

| | |
|---|---|
| Ultra-low Power<br>~5W | World's First 2nm<br>2 nm |
| Frontier-class Models. Off the Grid<br>~100 B | Faster than You Read (Single-Batch)<br>~30 TPS |
| NPU Performance<br>80 TOPS | Memory Bandwidth<br>153.6 GB/s |

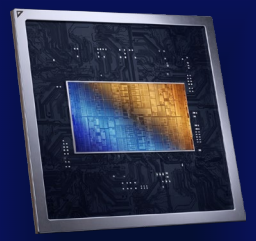
Chip Line-up

# How Big Should It Think?

Others count operations. We count how much mind fits on the die.

## DX-M2

FcBGA 16x16



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

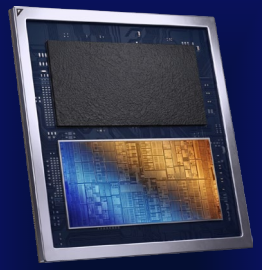
Power Consumption
5W (Only SoC)

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)

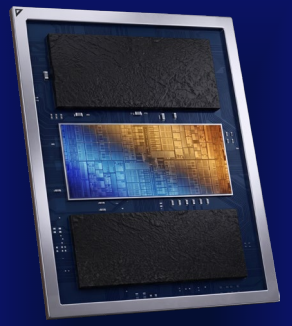
Power Consumption
5W (SoC 3W + DRAM 2W)

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

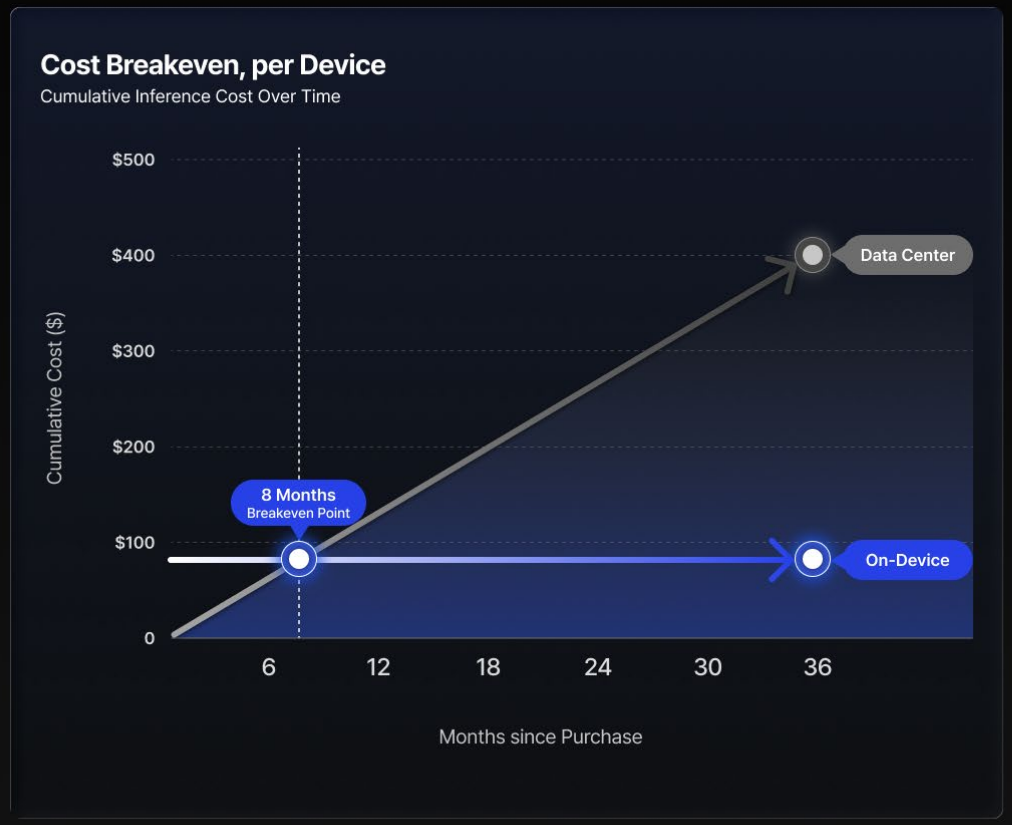
*Specifications are subject to change without notice during development.

DEEPX

## Every Deployment Breaks Even

# Here Sooner Than You Think.

The cloud bills you for thinking.
The DX-M2 bills you once.





### 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

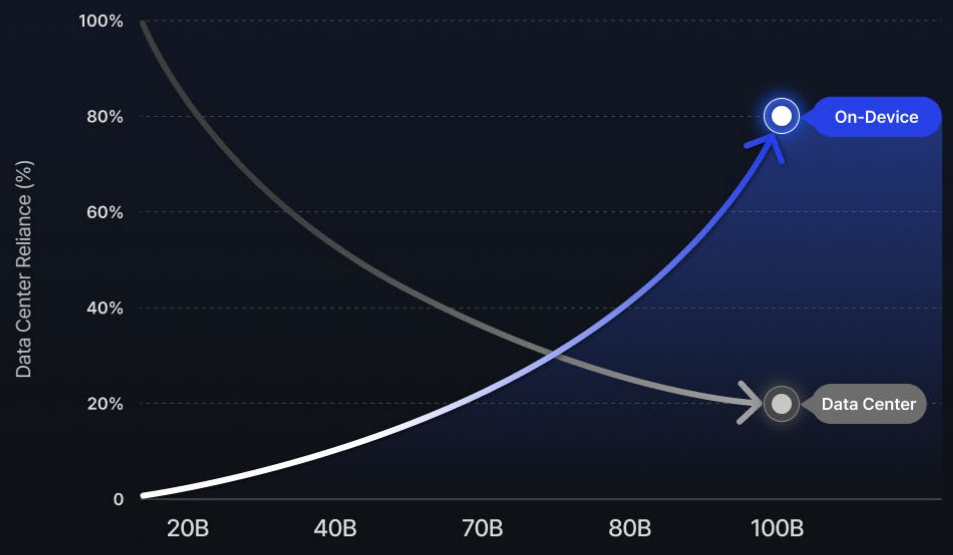
## Every Deployment Breaks Even

# Here Sooner Than You Think.

The cloud bills you for thinking.
The DX-M2 bills you once.

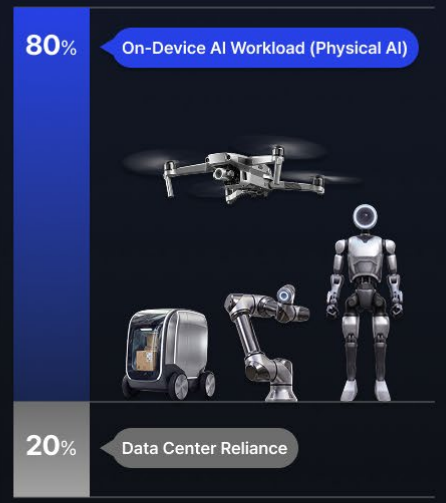
### The Cloud AI Breakeven Point

Dependency Trend by Model Size




### Shifting Workloads

Physical AI Takes the Edge




Deployment of Generative at 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

# Your Code Already Runs Here.

Nothing to port. Nothing to rewrite. The compiler does the work you were planning to do.

01 Connect

### Bring Your Own Stack

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

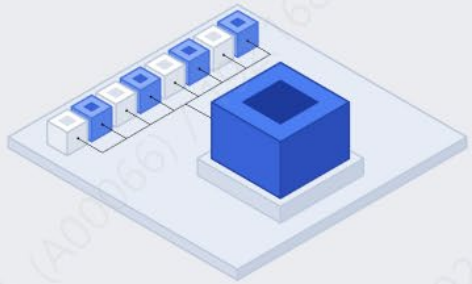
- OpenAI-compatible API
- PyTorch
- llama.cpp
- Ollama
- vLLM
- ExecuTorch

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 |
|---|---|---|---|
| **Qwen3-VL-2B**<br>Alibaba | Image-Text-to-Text | Text · Image · Video | Text |
| **Qwen3-VL-8B**<br>Alibaba | Image-Text-to-Text | Text · Image · Video | Text |
| **gpt-oss-20b**<br>OpenAI | Text Generation | Text | Text |
| **Llama-3.1-8B**<br>Meta | Text Generation | Text | Text |
| **DeepSeek-R1-Distill-Llama-8B**<br>DeepSeek | Text Generation | Text | Text |
| **Qwen3-4B**<br>Alibaba | Text Generation | Text | Text |
| **Qwen3-1.7B**<br>Alibaba | Text Generation | Text | Text |
| **Qwen3-0.6B**<br>Alibaba | Text Generation | Text | Text |
| **whisper-large-v3-turbo**<br>OpenAI | Speech Recognition | Audio | Text |
| **whisper-large-v**<br>OpenAI | Speech Recognition | Audio | Text |

*Specifications are subject to change without notice during development.

DEEPX