MOBILINT

MLA400

PCIe Card

mobilint Mobilint, Inc.

v1.0

mobilint

PCIe Card

MLA400

MLA400 is the world's leading AI accelerator PCIe card for edge computing. It delivers outstanding AI performance with low energy consumption to overcome the limitations of existing processors and unleash the full potential of AI technologies. MLA400 supports various computation structures with DNN-based architectural blocks and enables seamless operation with existing models with an advanced compiler stack. Empower yourself with MLA400 to accelerate the implementation of AI algorithms in your solution today.

MLA400

Specification

Performance	Approx. 320 TOPS (TBU)
Host Interface	PCI Express Gen5.0 16-lane
Memory Capacity	16 GB x4 (Optional 32 GB x4) LPDDR4, 4X
Memory Bandwidth	66.7 GB/s
Power (TDP)	120 W
Product Size	268 mm (L) × 107 mm (W) × 19 mm (H)

Model Compatibility

Mobilint's NPU ensures comprehensive support for diverse AI workloads, from vision AI to large language models.

For more information, access our Model Zoo comprising 400+ pre-optimized models.

Computer Vision | Vision Transformer | Language Models | Multi-Modal Models

KEY FEATURES: Break free from GPU DEPENDENCY today.

Cost effective

Most cost-effective options for engineers prioritizing high price-performance ratio.

Flexibility

Deployment-ready architecture allows concurrent execution of up to 32 different deep learning models.

High accuracy

World-leading lightweight technology maintains 99 % accuracy of existing models.

Easy to use

User-friendly full-stack SDK that supports major ML frameworks including Tensorflow, Pytorch, ONNX.

Eco friendly

Eco-friendly solution that maximizes data reuse and minimizes memory access for high energy efficiency.

Programmable

Supports more than 400 deep learning models, including SOTA models, with excellent performance.

Full-Stack SDK Support

SDK qb

'SDK qb' is a powerful tool designed to streamline the development of AI applications on NPUs, enabling rapid and efficient deployment while maintaining over 99% of the original model's accuracy. With compatibility across more than 400 AI models and proven stability and accuracy, SDK qb empowers developers to quickly implement NPU-accelerated deep learning solutions for a wide range of use cases.

Access 400+ Models

Supported Frameworks

PyTorch, TensorFlow, TFLite, ONNX, and KERAS.

Mobilint Compiler Suite

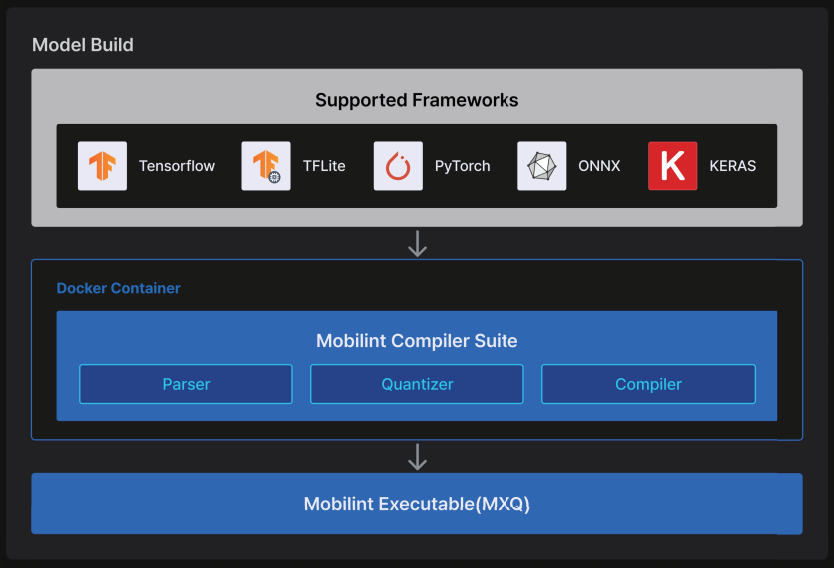
Mobilint compiler suite's proprietary techniques optimize the original model while retaining its accuracy.

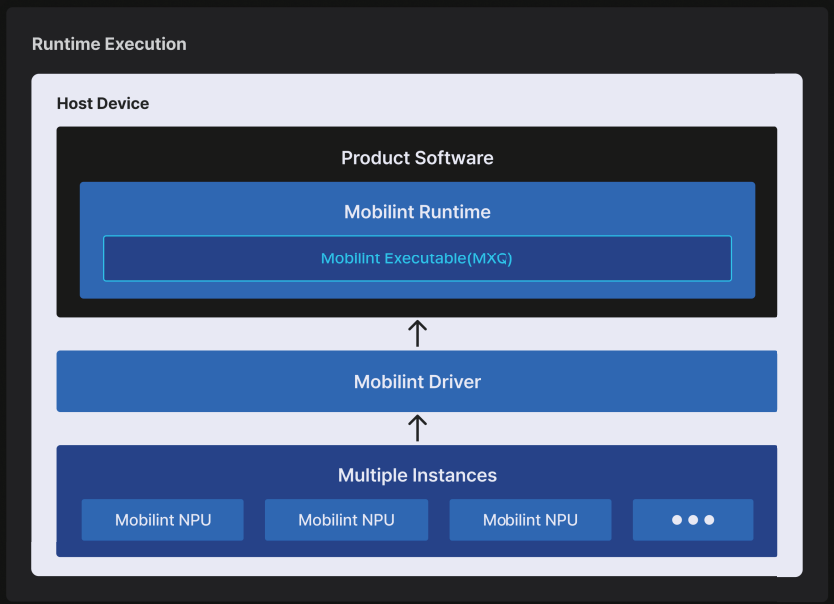
Mobilint Executable (MXQ)

An NPU-ready format for Mobilint runtime.

Multiple Instances

Mobilint NPUs can be plugged into the host device in multiple instances depending on the application.

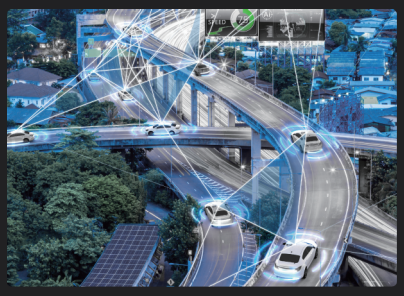




MLA400

Applications

Mobilint offers comprehensive solutions for a wide range of applications by integrating real-time artificial intelligence into edge environments such as smart cities, smart factories, healthcare systems, robotics, and more. Our technology delivers high performance, low power consumption, minimal heat generation, space efficiency, and cost competitiveness, enabling innovative and efficient solutions across industries.



SMART CITIES

SMART FACTORY

SMART FARM

HEALTHCARE

EDGE DATACENTER

AI SECURITY

Please **contact us** for inquiries about our products and solutions.
contact@mobilint.com

mobilint

AI Starts Here with Mobilint.

We provide accelerated AI chip solutions through the vertical integration of co-optimized algorithms, software, and hardware technologies.