AIBOX-9075

Industrial-grade Edge Computing Box

V1.0 2026-5-12

FIREFLY TECHNOLOGY

Product features

Qualcomm High-performance AI Processor IQ-9075

Integrated with an octa-core full-performance Kryo Gen6 CPU, with a maximum main frequency of 2.36 GHz and a general computing power of 230 KDMIPS. It has a built-in MCU-like security subsystem equipped with four real-time cores.

200 TOPS Flagship-grade NPU

Dual Hexagon NPU, delivering up to 200 TOPS (sparse INT8) / 100 TOPS (dense INT8). It supports hardware collaborative acceleration of CPU/GPU/NPU, enabling private deployment of edge-side large models and generative AI, with large model inference performance up to 12 Tokens/s@13B.

Powerful GPU and 8K Encode/Decode VPU

Adreno 663 GPU, compatible with graphics and computing interfaces including Vulkan 1.2 and OpenGL ES 3.2. The VPU supports concurrent encoding and decoding: 1×8K60, 4×4K60, 32×1080P30, as well as simultaneous encoding and decoding of 2×4K60 + 2×4K60.

36GB LPDDR5 + 128GB UFS2.2

Integrated with 36GB LPDDR5 (ECC) high-speed memory and 128GB UFS2.2 flash storage, it also supports PCIe 4.0×4 M.2 NVMe SSD expansion, fully ensuring the high-speed reading and writing as well as stable operation of edge-side large model loading, high-definition videos and massive data.

Product features

Full-scenario Network Connectivity

Equipped with industrial-grade dual 2.5G Ethernet and TSN, the device supports optional expansion of Wi-Fi 6, Bluetooth 5.2, 4G and 5G. With low latency and high reliability, it perfectly meets the application requirements of industrial automation, real-time control and high-speed data backhaul.

Industrial-grade Design

Supports wide-temperature operation from -40℃ to 85℃, equipped with 8×GMSL2, 2×CAN-FD, 2×RS485, 12×optically isolated DI/DO and other interfaces, meeting expansion requirements for industrial control, robotics and other scenarios.

Complete Software Ecosystem

Relying on the Qualcomm Linux software stack, it natively supports Ubuntu and Yocto systems, is compatible with deep learning frameworks such as TensorFlow, PyTorch and ONNX, and can connect to model marketplaces and massive model repositories to quickly complete model conversion, tuning and validation.

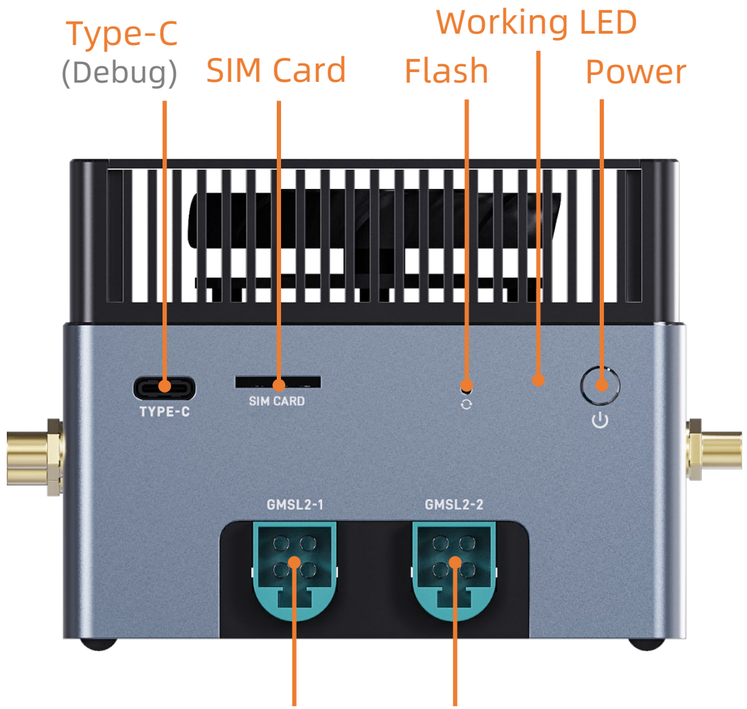
Wide Range of Application Scenarios

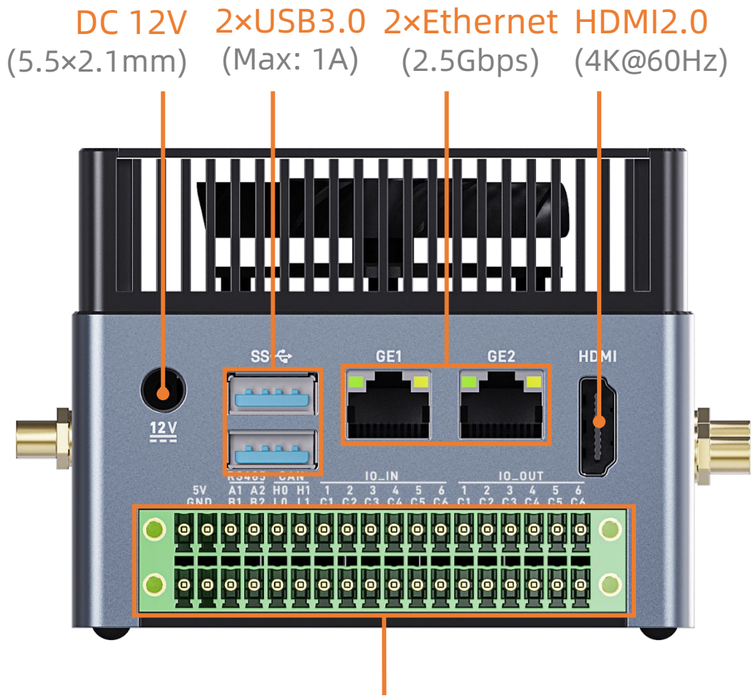
Widely applicable to edge computing, private LLM deployment, robotics, computer vision, intelligent security, industrial control, retail automation, medical self-service terminals and other products and fields.

Specifications

Specifications		
Basic Specifications	SOC	Qualcomm IQ-9075
	CPU	Qualcomm Kryo® Gen6 CPU: Octa-core 64-bit (8×Kryo Gold Prime), up to 2.36 GHz
	MCU	Integrated subsystem with quad-core Cortex-R52 CPU; each Cortex-R52 CPU runs at up to 1.85 GHz, supporting independent boot via the OSPI interface
	GPU	Adreno 663 GPU supports safe GPGPU compute, up to 840MHz · Graphics APIs: Vulkan 1.2, OpenGL ES 3.2 · Compute APIs: Vulkan 1.2, OpenCL 2.0 FP, Adreno NN Direct
	ISP	Qualcomm Spectra 690 ISP Image Signal Processor, featuring 2× Image Front End (IFE) + 5× Lightweight Image Front End (IFE_L). Supports 24-bit HDR Bayer processing, lens distortion correction, advanced tone mapping, offset correction, lens vignetting correction, dead pixel correction, directional scaling, color lookup table, color space conversion and noise reduction.
	NPU	Dual Hexagon Tensor Processor integrating Qualcomm Hexagon DSP, clocked up to 1.42 GHz. It incorporates 4 Hexagon Vector Extension (HVX) and 2 Hexagon Matrix Extension (HMX) cores, delivering up to 100 TOPS (Dense INT8) / 200 TOPS (Sparse INT8), and achieves 22 token/s for Llama2-7B inference.
	DPU	Dual Adreno DPU1199 supports image processing features including target scaler, exclusion rectangle extraction, inline rotation, 17×17×17 3D LUT (ViG/DSPP), HDR10 enhancement, wide gamut WCG, corner rounding, and CCCS fixed-point conversion. It also supports UBWC 4.0 and DSC v1.2 image compression technologies.
	Video codec	Video decoding: 1×8K@60fps, 2×8K@30fps, 4×4K@60fps, 8×4K@30fps, 16×1080p@60fps, 32×1080p@30fps AV1, H.264, H.265, VP9, MPEG2 Video encoding: 2×4K@60fps, 4×4K@30fps, 8×1080p@60fps, 16×1080p@30fps H.264, H.265, HEIF/HEIC
	RAM	36GB LPDDR5 (ECC supported)
	Storage	128GB UFS2.2
	Storage expansion	1 × M.2 (PCIe 4.0 x4 NVMe 2280 SSD expandable; located inside the computer)
	Power supply	DC 12V/5A (5.5 × 2.1mm)
	Power consumption	Normal: 12W(12V/1000mA), Max: 60W(12V/5000mA), Min(Sleep): 1.08W(12V/90mA)
	OS/Software	Ubuntu, Yocto Linux Software Stack
	Software support	Supports private deployment of ultra-large parameter models under the Transformer architecture, including the Deepseek-R1 series, Gemma series, Llama series, Qwen series, Phi series and other large language models. Supports private deployment of image generation models such as Stable Diffusion. Supports the QNN AI inference framework, as well as multiple deep learning frameworks including TensorFlow, TensorFlow Lite, PyTorch, ONNX, Keras and Caffe.
	Size	110.2mm × 102.25mm × 72.0mm (without mounting ears)
	Environment	Operating Temperature: -40℃ ~ 85℃, Storage Temperature: -40℃ ~ 90℃, Storage Humidity: 10% ~ 90%RH (non-condensing)
Interface Specifications	Ethernet	2 × 2.5G Ethernet (2.5Gbps/RJ45)
	Wireless network	Wi-Fi & Bluetooth: Supports Wi-Fi 6 and Bluetooth 5.2 (expandable via internal M.2 E-KEY) 4G/5G: Expandable via internal Mini PCIe
	Video input	8 × GMSL2 (2×4Pin Mini FAKRA)
	Video output	1 × HDMI2.0 (4K@60Hz)
	USB	2 × USB3.0 (Max: 1A), 1 × Type-C (Debug)
	Antenna	4 × 5G Antenna, 2 × Wi-Fi Antenna
	Button	1 × Power Button, 1 × Burn Button
	Others	1 × SIM Card, 1 × Phoenix terminal block (36Pin, 3.5mm pitch): 2 × RS485, 2 × CAN-FD, 12 × Opto-isolated DI/DO

Interface description

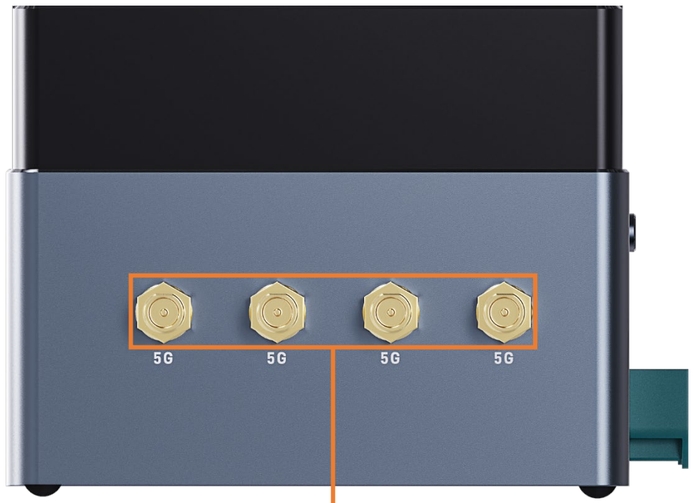




Phoenix terminal block (36Pin, 3.5mm pitch)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5V		RS485		CAN-FD		IO_IN						IO_OUT					
5V	5V	A1	A2	H0	H1	DIN1	DIN2	DIN3	DIN4	DIN5	DIN6	DOUT1	DOUT2	DOUT3	DOUT4	DOUT5	DOUT6
GND	GND	B1	B2	L0	L1	COM1	COM2	COM3	COM4	COM5	COM6	COM1	COM2	COM3	COM4	COM5	COM6

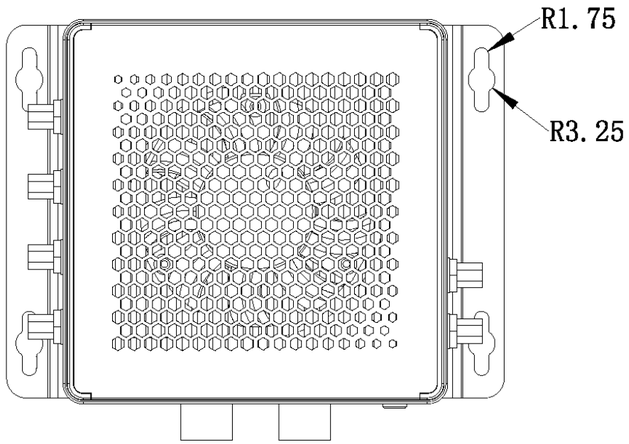
Interface description

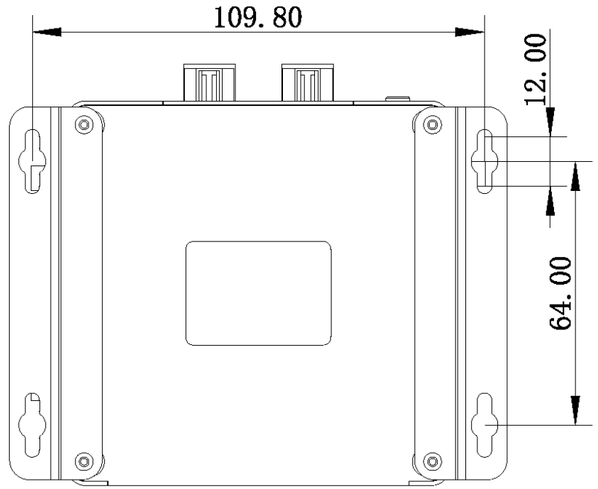


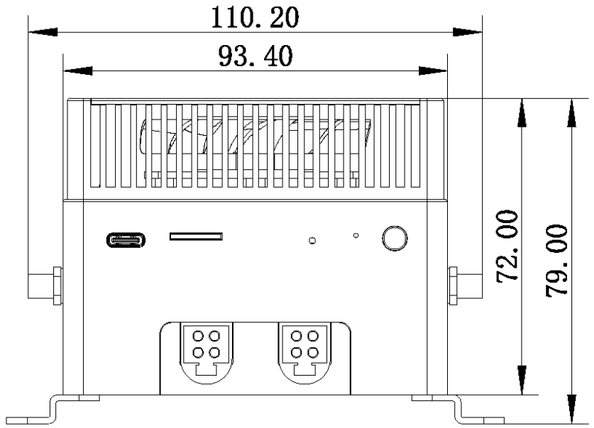
4×5G Antenna

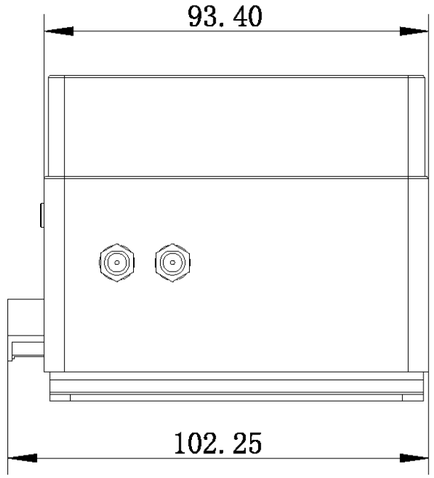
2×Wi-Fi Antenna

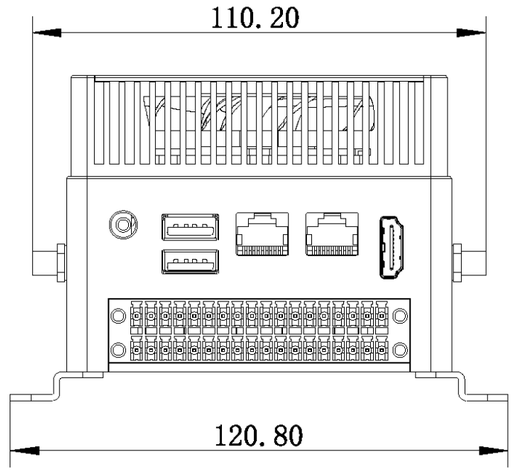
Dimension


R1. 75
R3. 25


109. 80
12. 00
64. 00


110. 20
93. 40
72. 00
79. 00


93. 40
102. 25


110. 20
120. 80

FIREFLY TECHNOLOGY

Contact Us
(+86)18688117175

E-mail
global@t-firefly.com

Website
https://en.t-firefly.com/

Address
Room 2101, Hongyu Building, #57 Zhongshan 4Rd, East District, Zhongshan, Guangdong, China.