AIBOX QUA KIT

4-Card AI Matrix Development Kit (RK1828)



V0.1 2026-11-10

FIREFLY TECHNOLOGY

Product features



Multi-core Heterogeneous Design with Hardware Load Isolation

It adopts a split heterogeneous architecture with primary RK3588 and secondary RK1828, connected via high-speed PCIe. All AI inference is offloaded to RK1828, resulting in only 2.4%~2.6% CPU usage on RK3588. The primary module handles codec, task scheduling and data preprocessing, while the secondary module provides AI inference capability.

Multi-card Cascading & Elastic Compute Scaling

Supports PCIe multi-card cascading and model partitioning for distributed inference. Available configurations: 7B (single card), 12B (dual cards), 27B (4 cards). The motherboard accommodates up to four M.2 RK1828 modules with 3D stacked 5GB on-chip DRAM. It runs 27B LLMs and 35B MoE models at 13 TPS@27B and 27.86 TPS@35B for edge LLM deployment.

Agent Ecosystem Direct Connect, Plug-and-Play

It provides OpenAI /v1 compatible APIs and natively supports OpenClaw, Nanobot and other Agent frameworks for rapid local Agent deployment. Compliant with MCP, it dynamically extends capabilities such as device control and data acquisition via registered tools.

Supports Multiple Deep Learning Frameworks

It supports INT4/INT8/INT16 multi-precision computation, compatible with CNN, RNN, LSTM and mainstream frameworks including TensorFlow, PyTorch and PaddlePaddle for vision and time-series inference.

Product features



On-board Offline Inference, Perception Data Local Processing

LLM, VLM and ASR inference run locally on board. Raw audio and video data are processed within the private network. The device works offline with millisecond-level response. All AI data is stored and computed locally to meet private deployment and data compliance requirements.

8K Codec, Parallel Multi-stream Video & AI Inference

The primary module integrates high-performance VPU for 8K codec and hardware preprocessing, handling multiple 4K streams concurrently. AI inference is offloaded to the secondary NPU. Hardware isolation between codec and NPU avoids CPU contention, ensuring smooth preview for multi-camera edge vision applications.

Abundant Expansion Interfaces

Equipped with 5G Ethernet, Gigabit Ethernet, USB3.0 HOST, USB3.0 OTG, HD2.1, M.2 and TF Card interfaces for flexible peripheral expansion.

Wide Application Scenarios

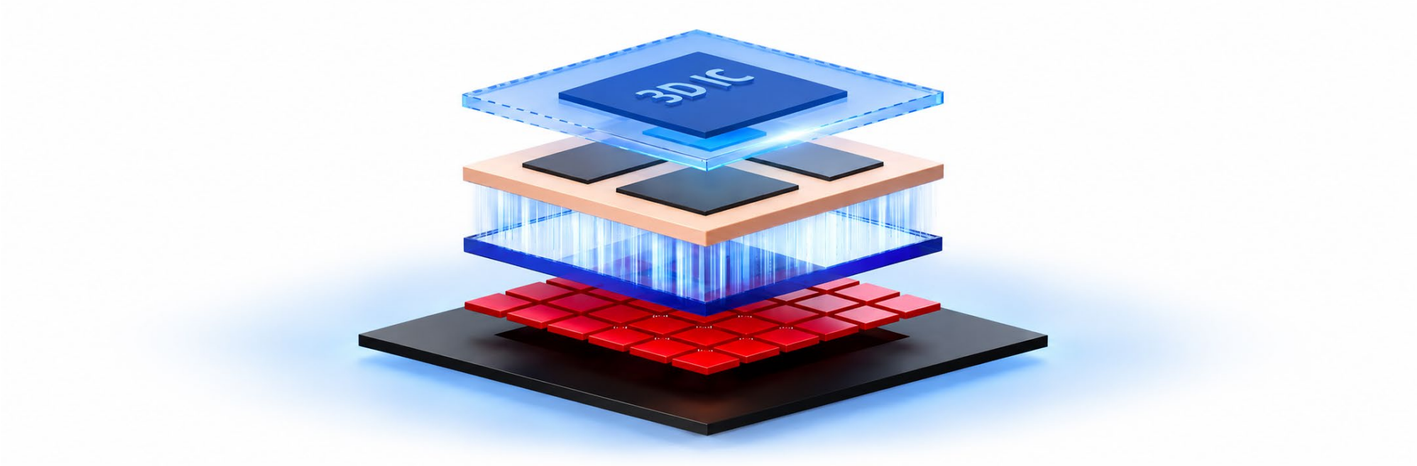
It applies to robotics, edge AI computing, industrial automation, intelligent security, autonomous driving, traffic management, smart city, home storage and retail automation.

Specifications

Specifications		
Basic Specifications	Host module	RK3588 module (Core-3588L): LPDDR4/LPDDR4x (4/8/16/32GB) + eMMC (32/64/128/256GB), 6 TOPS (INT8)
	Video encoding	8K@30fps H.265/H.264
	Video decoding	8K@60fps H.265/VP9/AVS2, 8K@30fps H.264/AVC/MVC, 4K@60fps AV1, 1080P@60fps MPEG-2/-1/VC-1/VP8
	Accelerator	4 × RK1828 M.2 module (RM1828MC0): On-chip 5GB DRAM, 20 TOPS (INT8), smoothly runs large models up to 8B.
	Storage expansion	1 × TF Card
	Motherboard power	DC 19V (5.5×2.1mm, supports 12V~24V wide voltage input)
	Accelerator power	4 × 12V (2 × 4Pin-1.25mm Wafer sockets)
	Power consumption	Normal: 13.3W(19V/700mA), Max: 79.8W(19V/4200mA), Min(Sleep): 8.74W(19V/460mA)
	OS	Linux
	Dimension	120.5mm × 124.1mm × 47.6mm
	Weight	≈567g
	Environment	Operating Temperature: -20℃ ~ 60℃, Storage Temperature: -20℃ ~ 70℃, Storage Humidity: 10%~90%RH (non-condensing)
Interface Specifications	Network	1 × 5G Ethernet (5Gbps/RJ45), 1 × Gigabit Ethernet (1000Mbps/RJ45) On-board Wi-Fi 6 & Bluetooth 5.2 combo module
	Video output	1 × HD2.1 (up to 8K@60Hz)
	USB	1 × USB3.0 HOST (USB-A, 1A current limit), 1 × USB3.0 OTG (USB-A, 1A current limit), 1 × Type-C (USB3.0 OTG)
	Button	1 × Power, 1 × Reset, 1 × Recovery, 1 × BOOT, 1 × User (customizable for any function)
	Others	3 × FAN (4Pin-1.25mm), 1 × Debug (3Pin-2mm), 1 × RTC battery holder

Model Performance

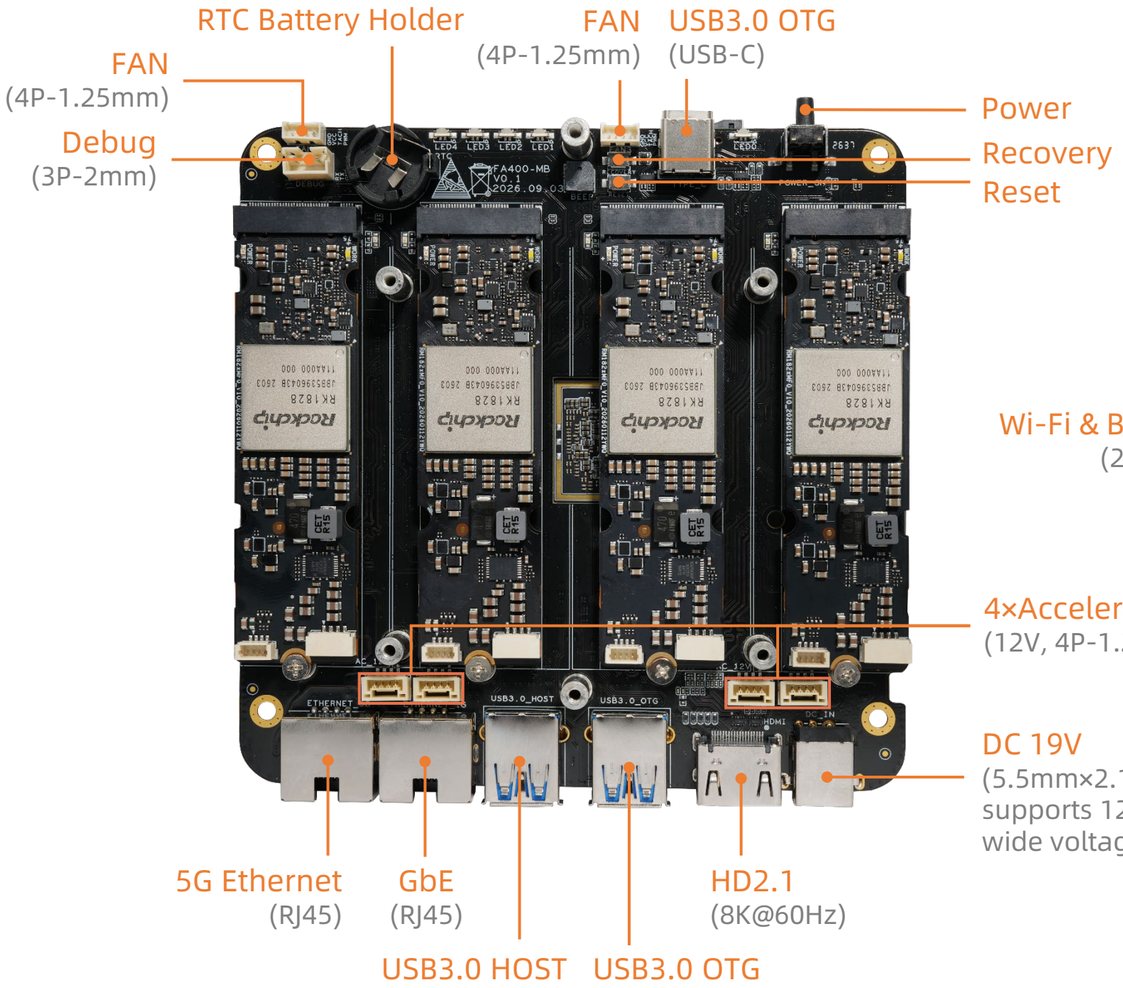
The RK1828 features an NPU architecture specifically designed for AI. Via its high-speed PCIe interface, it works collaboratively with the main SOC. Equipped with built-in high-bandwidth DRAM, it facilitates the local deployment and efficient operation of large models.

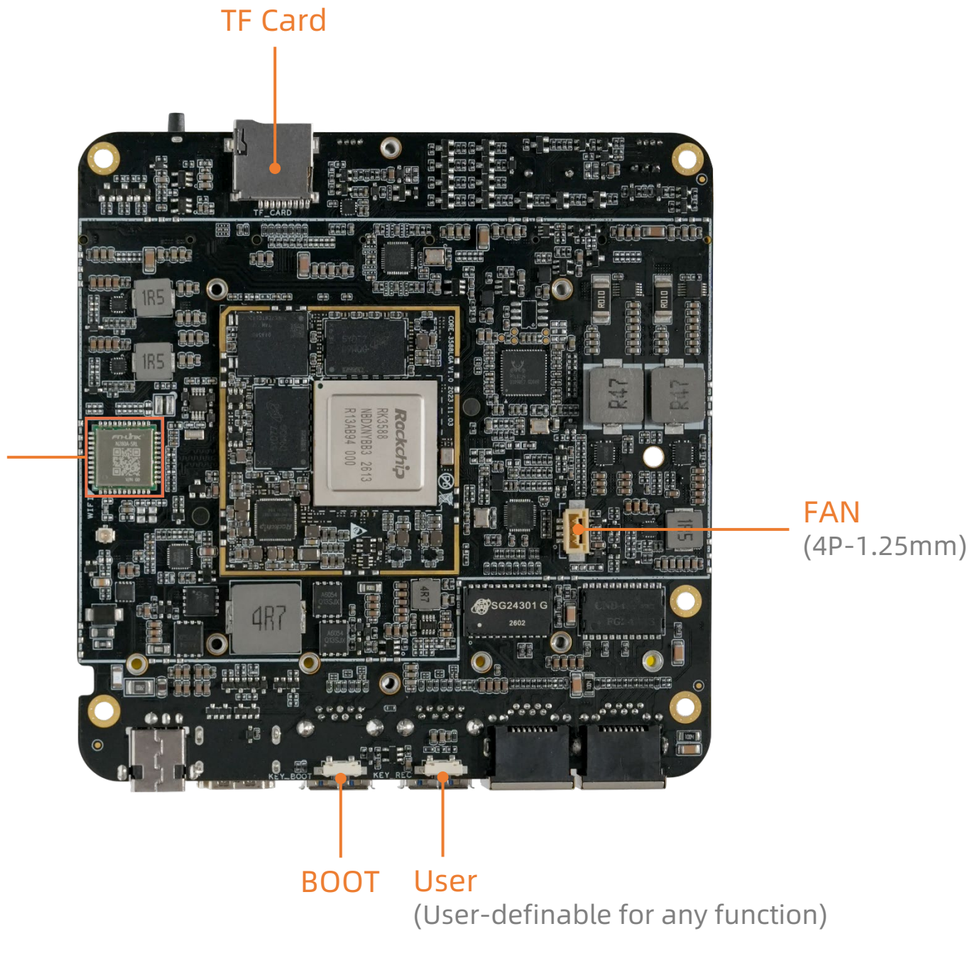


Test Conditions: RK3588 Host · Performance Mode Enabled · input=5120, output=128 · Model Quantization: W4A16

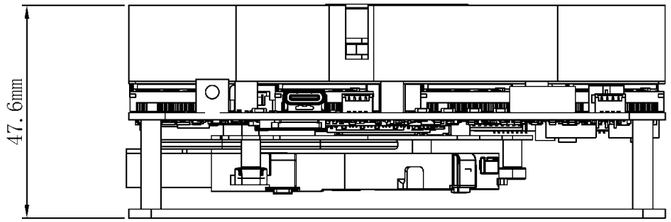
Model	Accelerator Card	Inteface	Stage	Total Time(ms)	Tokens	TPOP(ms)	TPS
Qwen3.8-27B	4 × RK1828	PCIe	Prefill	9510.68	5120	1.86	538.34
			Decode	9998.55	127	78.73	12.7

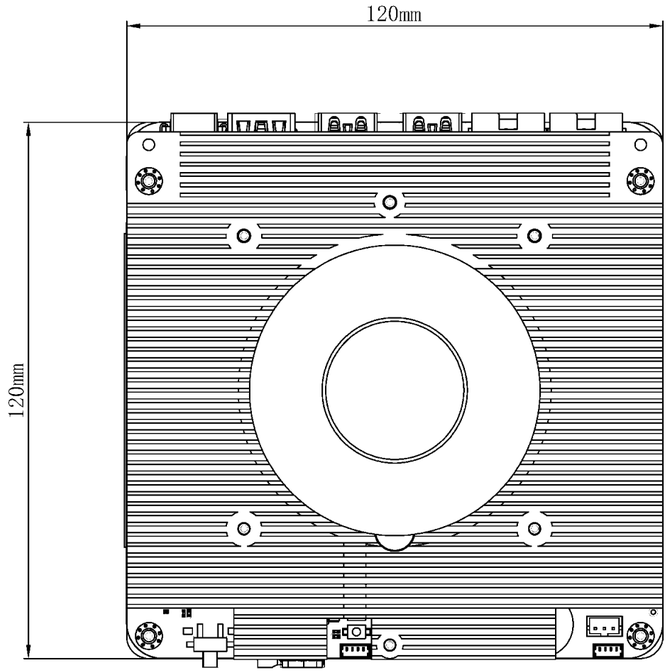
Interface description

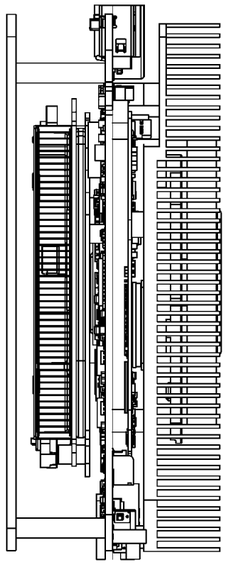


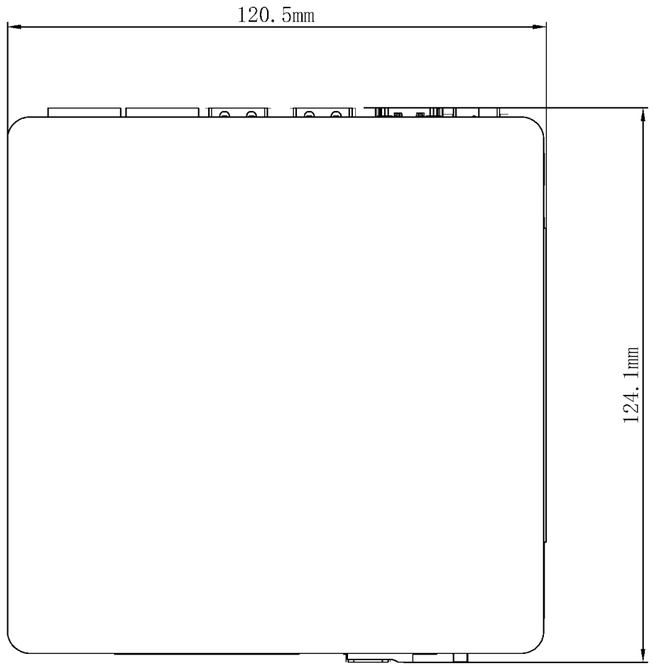


Dimension



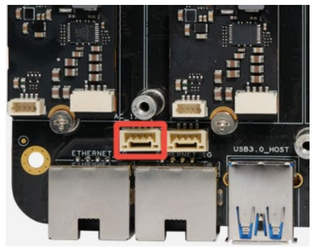






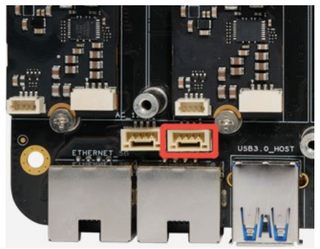
Interface definition

1. (J14) Accelerator power: 4PIN 1.25mm Wafer Socket



NO.	Definition	Power/V	NO.	Definition	Power/V
1	VCC_12V	12	3	GND	
2	VCC_12V	12	4	GND	

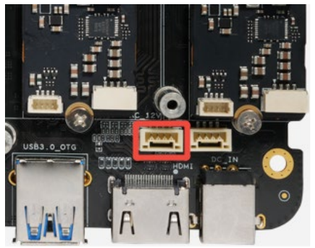
2. (J16) Accelerator power: 4PIN 1.25mm Wafer Socket



NO.	Definition	Power/V	NO.	Definition	Power/V
1	VCC_12V	12	3	GND	
2	VCC_12V	12	4	GND	

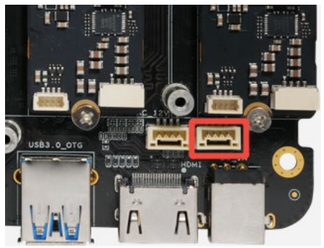
Interface definition

3. (J12) Accelerator power: 4PIN 1.25mm Wafer Socket



NO.	Definition	Power/V	NO.	Definition	Power/V
1	VCC_12V	12	3	GND	
2	VCC_12V	12	4	GND	

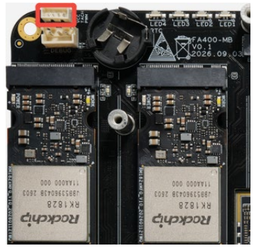
4. (J15) Accelerator power: 4PIN 1.25mm Wafer Socket



NO.	Definition	Power/V	NO.	Definition	Power/V
1	VCC_12V	12	3	GND	
2	VCC_12V	12	4	GND	

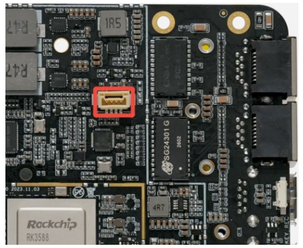
Interface definition

5. (J1) FAN2: 4PIN 1.25mm Wafer Socket



NO.	Definition	Power/V	NO.	Definition	Power/V
1	GND		3	FAN2_TACH	
2	FAN2+	12V	4	FAN2_PWM	

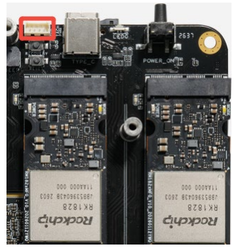
6. (J9) FAN1: 4PIN 1.25mm Wafer Socket



NO.	Definition	Power/V	NO.	Definition	Power/V
1	GND		3	FAN1_TACH	
2	FAN1+	12V	4	FAN1_PWM	

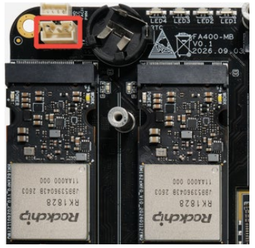
Interface definition

7. (J17) FAN3: 4PIN 1.25mm Wafer Socket



NO.	Definition	Power/V	NO.	Definition	Power/V
1	GND		3	FAN3_TACH	
2	FAN3+	12V	4	FAN3_PWM	

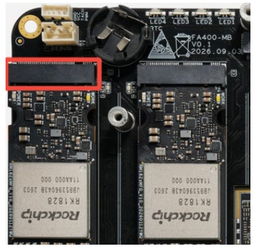
8. (J13) DEBUG: 3PIN 2.0mm Wafer Socket



NO.	Definition	Power/V	NO.	Definition	Power/V
1	DEBUG_RXD	3.3	3	GND	
2	DEBUG_TXD	3.3			

Interface definition

9. (U8) PCIE M.2 NGFF M-KEY SOCKET



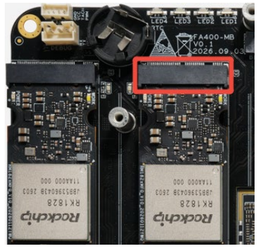
NO.	Definition	Power/V	NO.	Definition	Power/V
1	GND		2	VCC3V3_M2-2	
3	GND		4	VCC3V3_M2-2	
5	NC		6	NC	
7	NC		8	NC	
9	GND		10	DAS/DSS#[pull up resistor10K]	
11	NC		12	VCC3V3_M2-2	
13	NC		14	VCC3V3_M2-2	
15	NC		16	VCC3V3_M2-2	
17	NC		18	VCC3V3_M2-2	
19	NC		20	NC	
21	GND		22	NC	
23	NC		24	NC	
25	NC		26	NC	
27	GND		28	NC	
29	NC		30	NC	

Interface definition

31	NC		32	NC	
33	GND		34	NC	
35	NC		36	NC	
37	NC		38	DEVSLP[pull up NC/10K]	
39	GND		40	NC	
41	PCIE30_PORT1_RX3N		42	NC	
43	PCIE30_PORT1_RX3P		44	NC	
45	GND		46	NC	
47	PCIE30_PORT1_TX3N(Series capacitor 220nF)		48	NC	
49	PCIE30_PORT1_TX3P(Series capacitor 220nF)		50	PCIE30X1_1_PERSTN_M1[pull up NC/47K]	
51	GND		52	PCIE30X1_1_CLKREQN_M1[pull up NC/47K]	
53	PCIE30_M2_SSD3_REFCLKN		54	PCIE30X1_1_WAKEN_M1[pull up NC/47K]	
55	PCIE30_M2_SSD3_REFCLKP		56	NC	
57	GND		58	NC	
67	NC		68	NC	
69	GND		70	VCC3V3_M2-2	
71	GND		72	VCC3V3_M2-2	
73	GND		74	VCC3V3_M2-2	
75	GND				

Interface definition

10. (U7) PCIE M.2 NGFF M-KEY SOCKET



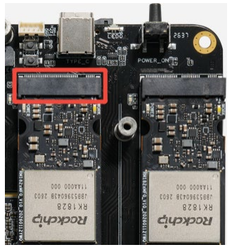
NO.	Definition	Power/V	NO.	Definition	Power/V
1	GND		2	VCC3V3_M2-2	
3	GND		4	VCC3V3_M2-2	
5	NC		6	NC	
7	NC		8	NC	
9	GND		10	DAS/DSS#[pull up resistor10K]	
11	NC		12	VCC3V3_M2-2	
13	NC		14	VCC3V3_M2-2	
15	NC		16	VCC3V3_M2-2	
17	NC		18	VCC3V3_M2-2	
19	NC		20	NC	
21	GND		22	NC	
23	NC		24	NC	
25	NC		26	NC	
27	GND		28	NC	
29	NC		30	NC	

Interface definition

31	NC		32	NC	
33	GND		34	NC	
35	NC		36	NC	
37	NC		38	DEVSLP[pull up NC/10K]	
39	GND		40	NC	
41	PCIE30_PORT1_RX2N		42	NC	
43	PCIE30_PORT1_RX2P		44	NC	
45	GND		46	NC	
47	PCIE30_PORT1_TX2N(Series capacitor 220nF)		48	NC	
49	PCIE30_PORT1_TX2P(Series capacitor 220nF)		50	PCIE30X2_PERSTN_M3[pull up NC/47K]	
51	GND		52	PCIE30X2_CLKREQN_M3[pull up NC/47K]	
53	PCIE30_M2_SSD2_REFCLKN		54	PCIE30X2_WAKEN_M3[pull up NC/47K]	
55	PCIE30_M2_SSD2_REFCLKP		56	NC	
57	GND		58	NC	
67	NC		68	NC	
69	GND		70	VCC3V3_M2-2	
71	GND		72	VCC3V3_M2-2	
73	GND		74	VCC3V3_M2-2	
75	GND				

Interface definition

11. (U11) PCIE M.2 NGFF M-KEY SOCKET



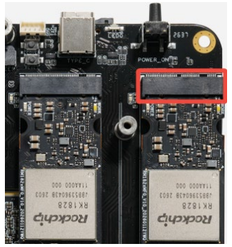
NO.	Definition	Power/V	NO.	Definition	Power/V
1	GND		2	VCC3V3_M2-1	
3	GND		4	VCC3V3_M2-1	
5	NC		6	NC	
7	NC		8	NC	
9	GND		10	DAS/DSS#[pull up resistor10K]	
11	NC		12	VCC3V3_M2-1	
13	NC		14	VCC3V3_M2-1	
15	NC		16	VCC3V3_M2-1	
17	NC		18	VCC3V3_M2-1	
19	NC		20	NC	
21	GND		22	NC	
23	NC		24	NC	
25	NC		26	NC	
27	GND		28	NC	
29	NC		30	NC	

Interface definition

31	NC		32	NC	
33	GND		34	NC	
35	NC		36	NC	
37	NC		38	DEVSLP[pull up NC/10K]	
39	GND		40	NC	
41	PCIE30_PORT0_RX0N		42	NC	
43	PCIE30_PORT0_RX0P		44	NC	
45	GND		46	NC	
47	PCIE30_PORT0_TX0N(Series capacitor 220nF)		48	NC	
49	PCIE30_PORT0_TX0P(Series capacitor 220nF)		50	PCIE30X4_PERSTN_M3[pull up NC/47K]	
51	GND		52	PCIE30X4_CLKREQN_M3[pull up NC/47K]	
53	PCIE30_M2_SSD0_REFCLKN		54	PCIE30X4_WAKEN_M3[pull up NC/47K]	
55	PCIE30_M2_SSD0_REFCLKP		56	NC	
57	GND		58	NC	
67	NC		68	NC	
69	GND		70	VCC3V3_M2-1	
71	GND		72	VCC3V3_M2-1	
73	GND		74	VCC3V3_M2-1	
75	GND				

Interface definition

12. (U26) PCIE M.2 NGFF M-KEY SOCKET



NO.	Definition	Power/V	NO.	Definition	Power/V
1	GND		2	VCC3V3_M2-1	
3	GND		4	VCC3V3_M2-1	
5	NC		6	NC	
7	NC		8	NC	
9	GND		10	DAS/DSS#[pull up resistor10K]	
11	NC		12	VCC3V3_M2-1	
13	NC		14	VCC3V3_M2-1	
15	NC		16	VCC3V3_M2-1	
17	NC		18	VCC3V3_M2-1	
19	NC		20	NC	
21	GND		22	NC	
23	NC		24	NC	
25	NC		26	NC	
27	GND		28	NC	
29	NC		30	NC	

Interface definition

31	NC		32	NC	
33	GND		34	NC	
35	NC		36	NC	
37	NC		38	DEVSLP[pull up NC10K]	
39	GND		40	NC	
41	PCIE30_PORT0_RX1N		42	NC	
43	PCIE30_PORT0_RX1P		44	NC	
45	GND		46	NC	
47	PCIE30_PORT0_TX1N(Series capacitor 220nF)		48	NC	
49	PCIE30_PORT0_TX1P(Series capacitor 220nF)		50	PCIE30X1_0_PERSTN_M2[pull up NC47K]	
51	GND		52	PCIE30X1_0_CLKREQN_M2[pull up NC47K]	
53	PCIE30_M2_SSD1_REFCLKN		54	PCIE30X1_0_WAKEN_M2[pull up NC47K]	
55	PCIE30_M2_SSD1_REFCLKP		56	NC	
57	GND		58	NC	
67	NC		68	NC	
69	GND		70	VCC3V3_M2-1	
71	GND		72	VCC3V3_M2-1	
73	GND		74	VCC3V3_M2-1	
75	GND				



FIREFLY TECHNOLOGY

Contact Us
(+86)18688117175

E-mail
global@t-firefly.com

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

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