CSC2-N48R3588S(M50)

Power Server

V1.0 2026-6-11

FIREFLY TECHNOLOGY

Product features

48 High-performance distributed computing nodes

Built-in 48 high-performance distributed computing nodes: Rockchip RK3588S, max. frequency 2.4GHz, peak computing power 6TOPS (INT8). Node quantity is configurable.

Expandable with 160 TOPS Houmo M50 accelerator card

Built-in 48 M.2 interfaces, expandable with 1-48 Houmo M50 accelerator cards as required. Each card delivers peak computing power of 160 TOPS (INT8) and 100 TFLOPS (bFP16). It supports smooth operation of large models within 35B parameters, with inference speed ranging from 10~35 Token/s.

48GB High-bandwidth LPDDR5 large memory

The Houmo M50 accelerator card supports up to 48GB LPDDR5 memory with a bandwidth of 153.6GB/s, delivering powerful data throughput.

Private deployment of large models on edge side

Supports edge-side private deployment of mainstream large language models (Llama, Qwen, etc.), vision models (SAM, ViT, etc.) and image generation models (Flux, Stable Diffusion, etc.).

Product features

Integrated network management, High Speed, Stability & Reliability

Features 4 10G ports (SFP+) and 1 MGMT (BMC management port). It supports independent BMC out-of-band management, and integrates Layer 3 routing & switching, VLAN partitioning and QoS traffic control to optimize transmission efficiency and enhance overall network security.

Innovative structural & Appearance design

Adopts high-density 2U rack-mount server chassis, equipped with one touch display. It real-time shows key operating data including chassis temperature, operational efficiency, fan speed, network IP, date and time for users to monitor device status conveniently.

Equipped with BMC management system

Equipped with BMC intelligent management system, it can easily complete real-time monitoring, software configuration, hardware management, troubleshooting, system upgrade, and can provide secondary development.

Wide range of application scenarios

It is widely used in intelligent computing servers, edge computing, large model localization, smart cities, smart healthcare, smart industry, intelligent security and other types of products and fields.

Specifications

Specifications		
Technical Specifications	Server form	2U rack-mounted computing power server
	Architecture	ARM architecture
	Compute nodes	48 distributed computing nodes (RK3588S): Octa-core 64-bit processor, max. frequency 2.4GHz
	AI Accelerator card expansion	48 × Houmo M50 Modules (48GB LPDDR5, 160 TOPS(INT8), 100 TFLOPS(bFP16), 153.6GB/s high bandwidth. Supports smooth running of large models within 35B parameters, with performance of 10~35 Token/s@35B)
	Control nodes	1 control node (RK3588): Octa-core 64-bit processor, max. frequency 2.4GHz
	Codec	Video Encoding: 1×8K@30fps/16×1080P@30fps H.265/H.264 Video Decoding: 8K@60fps/4K@120fps H.265/VP9/AVS2, 8K@30fps H.264/AVC/MVC, 30×1080P@30fps H.265/H.264
	RAM	16GB LPDDR5 (4GB/8GB/16GB/32GB) × 48 (number of computing nodes)
	Storage	256GB eMMC (16GB/32GB/64GB/128GB/256GB) × 48 (number of computing nodes)
	Power	2 AC redundant power supplies (Hot-swappable supported)
	Screen	1 touch display, capable of real-time displaying device information such as chassis temperature, energy efficiency, fan speed, network IP, date and time
	Fan module	12 high-speed cooling fans
	BMC	The BMC management system is integrated with the web-based management interface, supporting Redfish, VNC, NTP, monitoring advanced and virtual media, and the BMC management system can be redeveloped
Physical Specifications	Size	Standard 2U rack servers: 724.0mm × 430.0mm × 88.8mm
	Installation requirements	IEC 297 Universal Cabinet Installation: 19 inches wide and 800 mm deep and above Retractable slideway installation: The distance between the front and rear holes of the cabinet is 543.5mm~848.5mm
	Full weight	Net weight of the server: 23.1kg, total weight with packaging: 25.3kg
	Environment	Operating Temperature: 0°C ~ 35°C, Storage Temperature: -40°C ~ 60°C, Operating Humidity: 5% ~ 80%RH (non-condensing)
Interface Specifications	Internet	4 × 10G Ethernet (SFP+), 1 × Gigabit Ethernet (RJ45, MGMT used as BMC management network)
	Console	1 × Console (RJ45, BMC debug serial port, baud rate 115200)
	Display	1 × HDMI (Maximum resolution 1080P, BMC management display)
	USB	2 × USB3.0 (The lower USB is USB3.0 OTG, and the BMC can be upgraded OTG using a USB flash drive)
	Button	1 × Reset button, 1 × Power button, 1 × Restart BMC button

LLM Edge inference performance

The Houmo M50 delivers AI inference performance of 160 TOPS@INT8 and 100 TOPS@bFP16. A single node runs 35B-A3B open-source models smoothly with an inference speed of 10~35 Token/s. Combined with RK3588S computing nodes, it is deeply optimized for lightweight AI inference and distributed compilation, well adapted for large model edge inference. The device supports multi-node array deployment for flexible scaling and load balancing, meeting the demands of multi-user concurrent inference.

Model	Size	Ctx(k)	Input(k)	Prefill(tps)	Decode(tps)
Qwen3	26b_a4b	128	4	1116.98	26.25
			16	773.49	19.46
Gemma4			1	1401.93	26.02
			4	1342.04	23.16
			8	1340.28	19.94
			16	1241.37	15.34

Interface description

Console
BMC debug serial port

MGMT
BMC management network

Touch display

Power button
Blue (solid): The server is in standby
Blue (flashing): BMC management system is starting
Green (solid): The server is powered on
Off: The server is not powered on

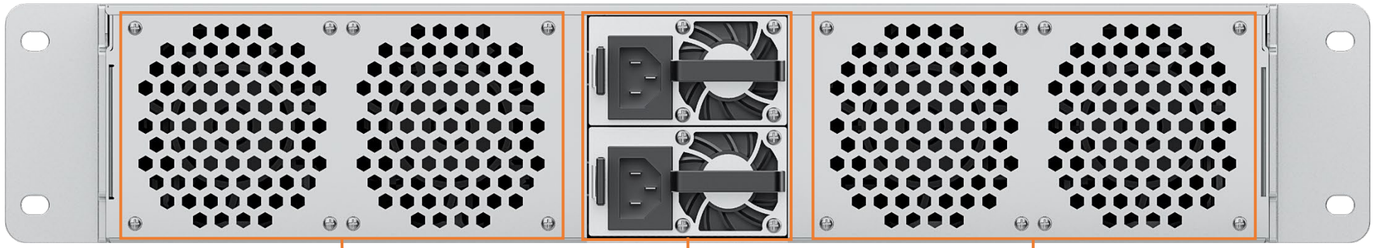
HDMI
1080P

2×USB3.0
Lower: OTG

4×10GE
SFP+

Restart BMC button
Short press

Reset button
Reset password: Press and hold for 5s until the BS light flashes slowly
Factory reset: Press and hold for 10s until the BS light flashes
Wrong press to recover: Press and hold until the BS light returns to solid on



Fan module

2×Power module

Fan module

Dimension

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.