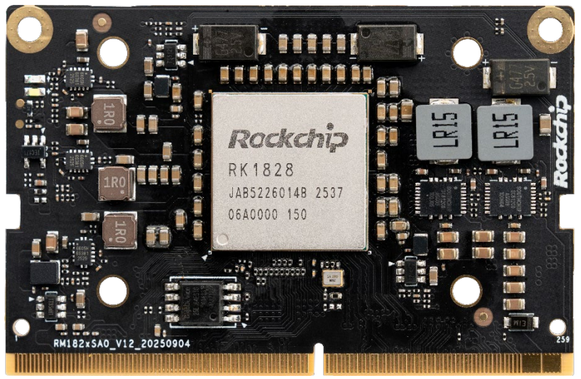
RK1828算力卡

- RM1828SA0（SODIMM模组）
- RM1828MC0（M.2模组）



RM1828SA0

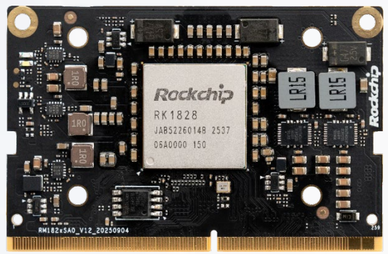


RM1828MC0

2026-10-29

天 启 集 团

产品特点 Product features



RM1828SA0



RM1828MC0

3D RAM 堆叠构架，高性能边缘 AI 加速

RM1828SA0 / RM1828MC0 为面向端侧大模型场景的 AI 推理模组，处理器采用 3D RAM 堆叠架构，搭载 5GB 高带宽 DRAM，大幅降低访存开销，端侧生成速率超 100 tokens/s，端到端延迟低至 0.1s，带来流畅低时延的本地大模型交互体验。

20TOPS 算力 NPU

NPU 算力最高 20TOPS（INT8），支持 INT4/INT8/INT16 多精度计算，兼容 CNN、RNN、LSTM 等传统网络架构，适配各类视觉与时序类算法推理。

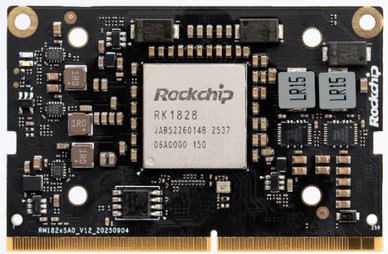
多卡级联扩展

支持灵活多卡级联扩展，单卡可实现 8B 模型轻量化推理，推理性能达 61.11TPS@8B LLM、70TPS@7B VLM；双卡承载 12B 模型、四卡支持 27B 模型、五卡可部署 35B MoE 大模型，算力规格可弹性扩展，适配不同规模大模型推理任务。

全面的模型支持

适配大语言模型（Qwen、GLM Edge）、视觉大模型（SigLIP、DINO）、CNN 传统网络（YOLO）、多模态模型以及视觉语言模型（MiniCPM、InternVL3），覆盖各类主流 AI 算法，满足多样化的端侧智能推理开发需求。

产品特点 Product features



RM1828SA0



RM1828MC0

支持 Agent 智能体与开发工具调用

本地大模型除基础问答能力外，可作为 Agent 推理核心；支持对接文件读取工具、Shell 及编译工具、本地代码仓库、MCP 服务、文档解析与知识库，可调用企业内部 HTTP API，搭配 OpenCode 代码 Agent 实现代码开发任务。

边缘计算低延迟，毫秒级响应

使用本地 AI 处理，减少网络依赖，不需要数据上云端，不受网络带宽和网速波动的影响，实现毫秒级响应和数据隐私保护。所有业务数据本地存储、本地运算，满足数据安全法规与行业隐私合规要求。

SODIMM / M.2 标准封装

模组提供 SODIMM、M.2 B+M-KEY 两种接口，其中 SODIMM 版本兼容 NVIDIA Orin Nano 线序，兼容性强。

广泛的应用场景

广泛适用于机器人、无人机、AI 边缘计算、工业自动化、智能安防、智能驾驶、交通管理、智慧城市、家庭存储、零售自动化等产品和领域。

规格参数　Specifications

		RM1828SA0（SODIMM 模组）	RM1828MC0（M.2 模组）
基本参数	SOC	Rockchip RK1828	
	RGA	内置2D图形引擎，支持 BT.601/BT.709 像素格式转换、抖动运算，最大分辨率：源图像 8192×8192，目标图像 4096×4096	
	NPU	NPU 算力 20 TOPS（INT8），支持 INT4/INT8/INT16/FP8/FP16/BF16 混合运算	
	编解码	JPEG 编码：支持基线（DCT 顺序）、JPEG文件交换格式（JFIF）1.02，支持图像尺寸范围从16×16到65520×65520，支持YUV400/YUV420/YUV422/YUV444 JPEG 解码：支持基线（DCT 顺序）、JPEG 文件交换格式（JFIF）1.02，支持图像尺寸范围从 48×48 至 65520×65520，支持YUV400/YUV420/YUV422/YUV440/YUV411/YUV444/RG888/RGB565	
	内存	片内5GB DRAM	
	电源	DC 5V	DC 12V
	系统	Linux	
	软件支持	支持YOLO目标检测算法 支持Transformer架构下超大规模参数模型的私有化部署，如Gemma系列、ChatGLM系列、Qwen系列、Phi系列等大语言模型 支持CNN、RNN、LSTM等传统网络架构，支持RKNN模型导入导出	
	接口类型	金手指（SODIMM 260P 标准接口，0.5mm 间距）	M.2 B-KEY / M.2 M-KEY
	尺寸	69.6mm × 45.0mm × 4.85mm（不带散热器） 69.6mm × 45.0mm × 20.67mm（带散热器）	22mm × 80mm × 6mm（不带散热器） 25mm × 81mm × 23.05mm（带散热器）
	重量	不带散热器：21g，带散热器：75g	不带散热器：11g，带散热器：53g
	环境	工作温度：-20℃ ~ 60℃，存储温度：-20℃ ~ 70℃，存储湿度: 10% ~ 90%RH（无凝露）	
接口参数	风扇接口	无	1 × FAN（4Pin-1.25mm）

模型性能 Model Performance

RK182X LLM 运行性能

模型名称	加速芯片	TTFT(ms)	TPOT(ms)	Decode TPS
Qwen2.5-3B	RK182X	83.44	9.80	102.01
Qwen3-1.7B	RK1828	52.16	7.20	138.88
Qwen3-8B	RK1828	177.87	16.36	61.11

* 数据基于 RKNN3 V1.0 SDK，所有模型使用 W4A16 G32 量化

RK182X VLM 运行性能

模型	加速芯片	Vision分辨率	Vision(ms)	LLM TTFT(ms)	LLM Decode TPS
FastVLM_1.5B_stage3	RK182X	512 * 512	144.13	47.99	148.47
InternVL3-2B	RK182X	448 * 448	190.80	47.93	148.26
Qwen2.5-VL-7B	RK1828	392 * 392	279.34	159.42	70.02
Qwen3-VL-4B	RK1828	384 * 384	158.89	108.29	89.69
MiniCPM-V_4	RK1828	448 * 448	237.55	94.94	106.62

* 数据基于 RKNN3 V1.0 SDK，所有模型使用 W4A16 G32 量化

RK182X CNN模型运行性能

Model	Resolution	Single-core Performance (FPS)	Multi-core Performance (FPS)
YOLOv5s	640*640	35.4	212.6
ResNet50V2	224*224	113.7	851.3
DINOv3	224*224	54.8	331.0
DepthAnythingV2	448*252	/	27.3

* 数据基于 RKNN3 V1.0 SDK，所有模型使用 W8A8 量化

RK1828 全模态模型运行性能

模型	加速芯片	Vision分辨率	Vision(ms)	Audio(ms)	LLM TTFT(ms)	LLM Decode TPS
Qwen2.5-Omni-3B (Thinker)	RK1828	392 * 392	310.86	98.91	84.83	102.63

* 数据基于 RKNN3 V1.0 SDK，所有模型使用 W4A16 G32 量化

模型性能 Model Performance

多卡模型性能

测试条件：RK3588 Host · RK1828 SO-DIMM · 模型量化：W4A16

模型	加速芯片	Input	Output	接口	TTFT (s)	Prefill TPS	Decode TPS
Qwen3.5-9B	2x RK1828	4K	128	PCIe	4.80	853.7	33.19
		4K	128	USB3.0	4.90	836.0	33.52
		8K	128	PCIe	9.73	841.7	32.33
		8K	128	USB3.0	9.94	824.0	32.65
Qwen3.8-27B	4x RK1828	4K	128	PCIe	6.92	592.1	13.02
		4K	128	USB3.0	7.15	573.2	13.02
		8K	128	PCIe	13.67	599.4	12.74
		8K	128	USB3.0	14.08	581.7	12.83

注：当前 4K / 8K 实测下，PCIe接口 在 TTFT 和 Prefill 上约有 2%–3% 优势

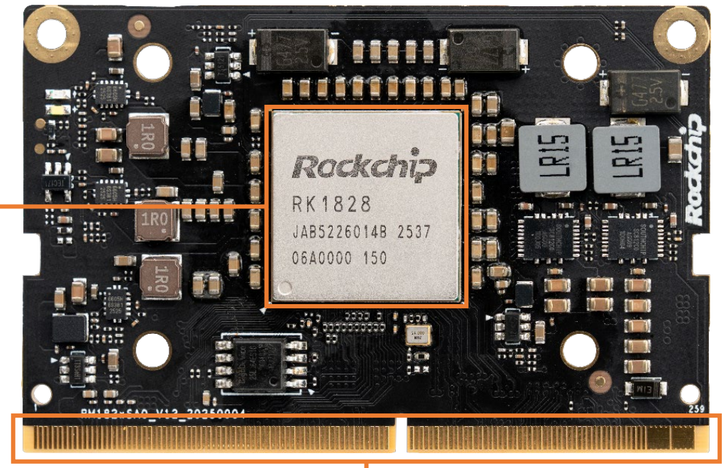
HOST系统开销

测试条件：RK3588 Host · RK1828 SO-DIMM · CPU 以四核 A76 满载 100% 为例

配置	加速芯片	接口	CPU 平均	内存平均	内存峰值
Qwen3.5-9B	2x RK1828	PCIe	2.4%	179 MB	310 MB
		USB3.0	2.0%	181 MB	312 MB
Qwen3.8-27B	4x RK1828	PCIe	2.6%	191 MB	524 MB
		USB3.0	2.2%	187 MB	523 MB

接口描述 Interface description

Rockchip RK1828
（主频最高850MHz）



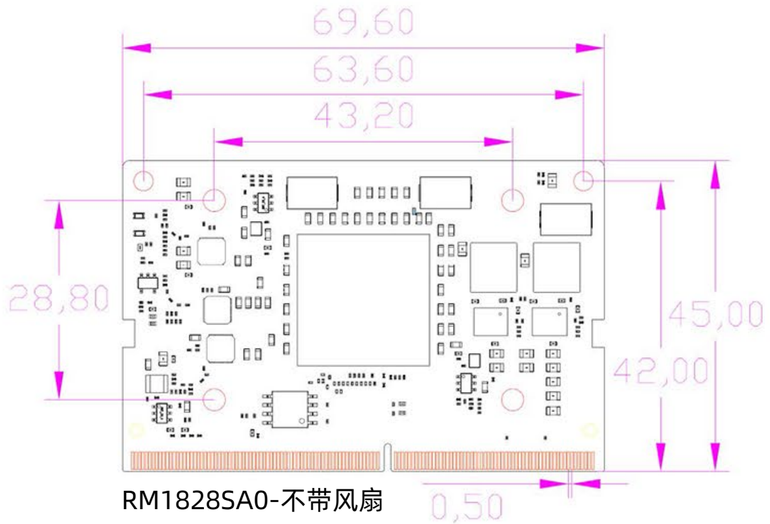
金手指
（260 PIN，SODIMM，0.5mm间距）



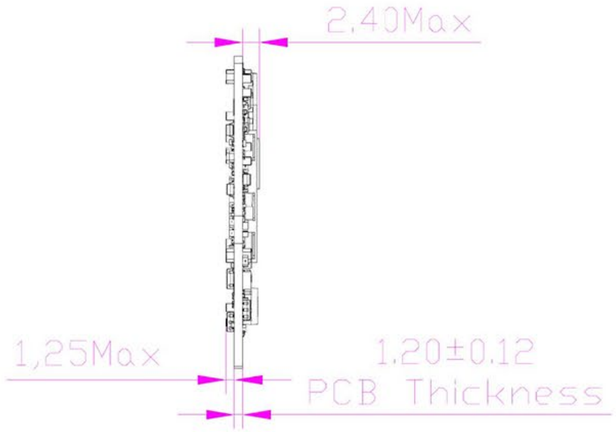
Rockchip RK1828
（主频最高850MHz）

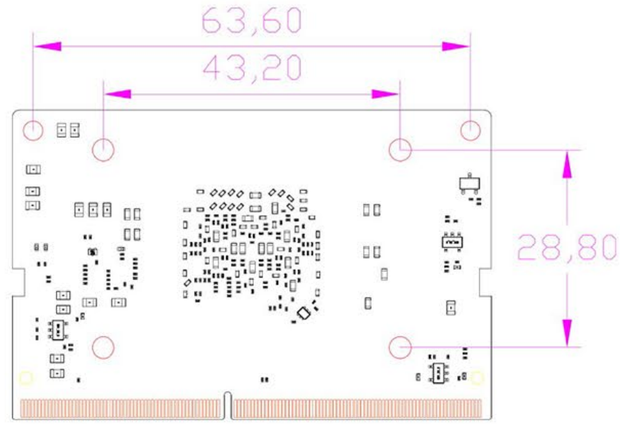
M.2 B-KEY / M.2M-KEY

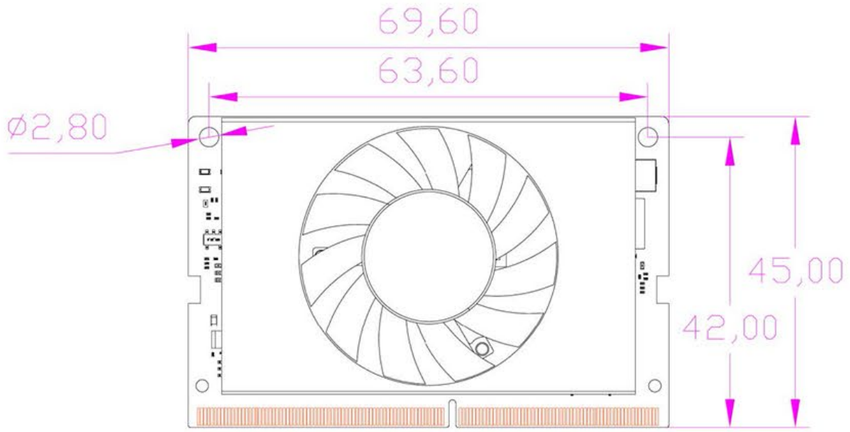
产品尺寸 Dimension



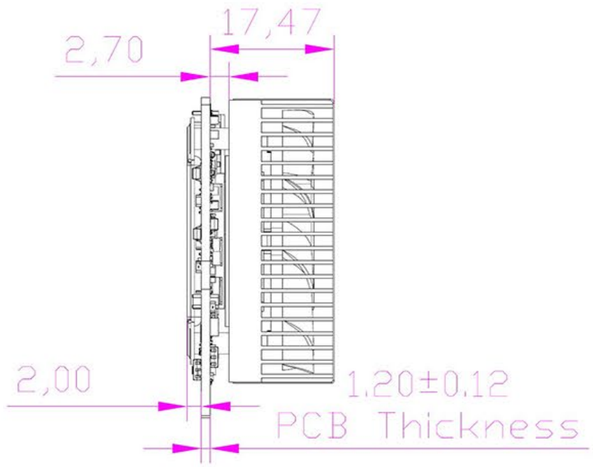
RM1828SA0-不带风扇

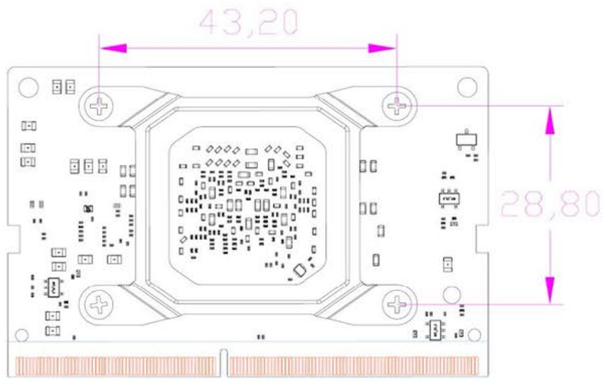






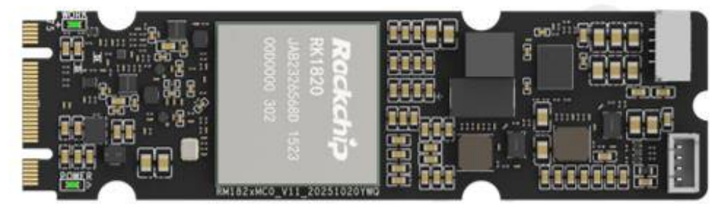
RM1828SA0-带风扇

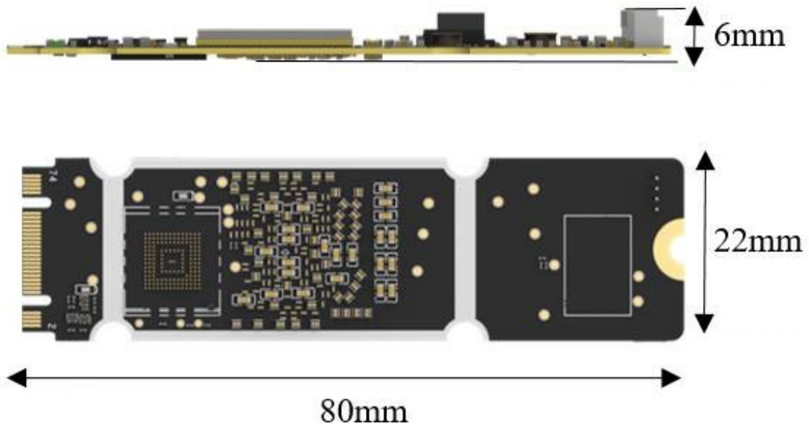




产品尺寸 Dimension

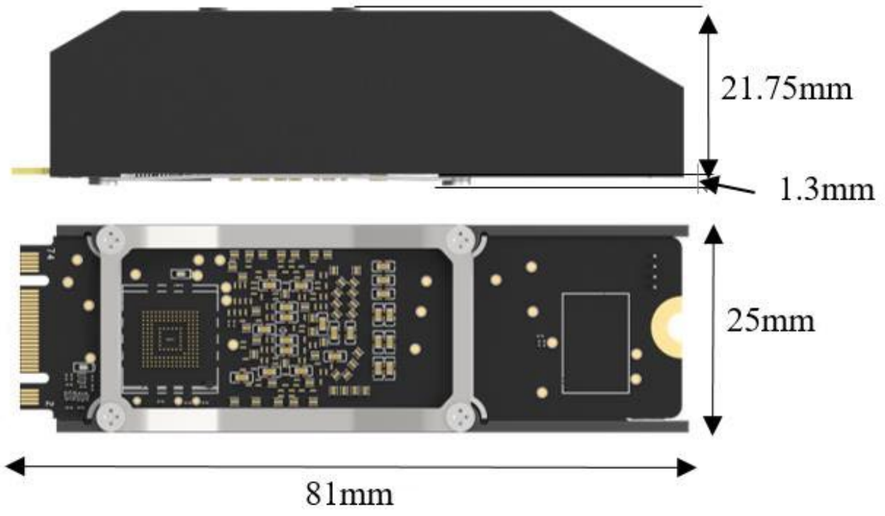
RM1828MC0-不带风扇





RM1828MC0-带风扇







天启集团

联系方式
(+86)18688117175

邮箱
sales@t-firefly.com

官方网址
www.t-firefly.com.cn

公司地址
广东省中山市东区中山四路57号宏宇大厦1座2101