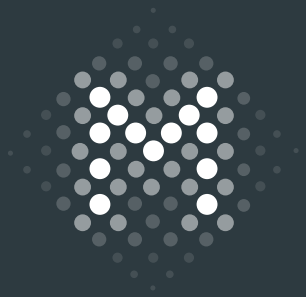

MOFFETT AI

墨芯人工智能
S4计算卡

产品简介

2022

概述

墨芯人工智能S4计算卡（下文简称S4计算卡）为数据中心的AI推理应用而打造。作为通用深度学习推理加速器，外形规格采用单槽PCIe Gen3 x16的半高半长的设计方式。S4计算卡支持20GB LPDDR4x内存，理论内存峰值带宽84 GB/s，最大功耗70 W。采取被动冷却板设计方式，通过系统气流使其在热限制值内进行计算工作。

S4计算卡基于墨芯人工智能Antoum®架构构建。通过软硬件紧密结合的架构设计，强调平衡的结构化稀疏性，支持高达32倍的高稀疏率。基于Antoum®架构，S4计算卡支持BF16和INT8计算。同时，S4计算卡支持包括集成模型稀疏器的软件工具链、编译器和运行时在内的端到端软件解决方案，确保主流AI推理作业可以快速实现。

硬件与软件紧密结合的设计使得Antoum®成为一个高效的人工智能片上系统处理器。此外，S4计算卡还支持硬件视频编解码器和JPEG解码器，使其能够处理各种视频和图像应用场景。同时，S4计算卡随设备发货时，为系统DDR开启ECC功能，防止内存出现可检测的错误。

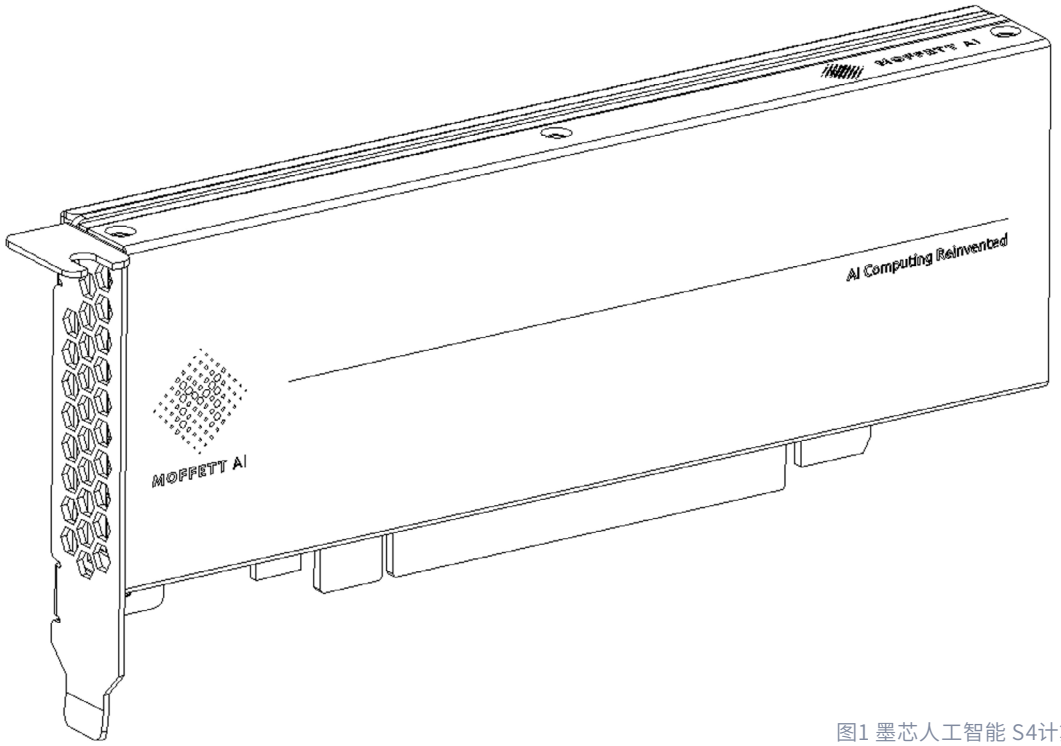

图1 墨芯人工智能 S4计算卡

规格

产品规格

S4计算卡的产品和内存规格如表1和表2所示。

规格	描述
产品SKU	KI1CYH01FF
总板卡功率	70W
Antoum®芯片SKU	XI1CAS0100
机械外形规格	单槽，半高半长
PCI IDs	Device ID: 0x7000 Vendor ID: 0x1F36 Sub-vendor ID: 0x1E5D Sub-system ID: 0x7000
VBIOS NOR-Flash大小	16 MB
热冷却解决方案	被动式
系统接口	PCIe Gen 3 x 16
板卡重量	316.8g

表1 产品规格

内存规格

规格	描述
最大内存时钟	4200 MHz
内存大小	20 GB
内存总线宽度	160-bit
理论内存带宽峰值	84 GBytes/s

表2 内存规格

环境和可靠性规格

S4计算卡的环境指标规格如表3所示。

规格	描述
操作环境温度	0°C - 50°C
存储温度	-40°C - 70°C
操作环境湿度	5%~95% 相对湿度
存储湿度	5%~95% 相对湿度

表3 环境指标规格

气流方向的支持

S4 PCIe卡采用双向散热的设计，实现灵活散热。它可以接受从左到右或者从右到左的气流，如图2所示。

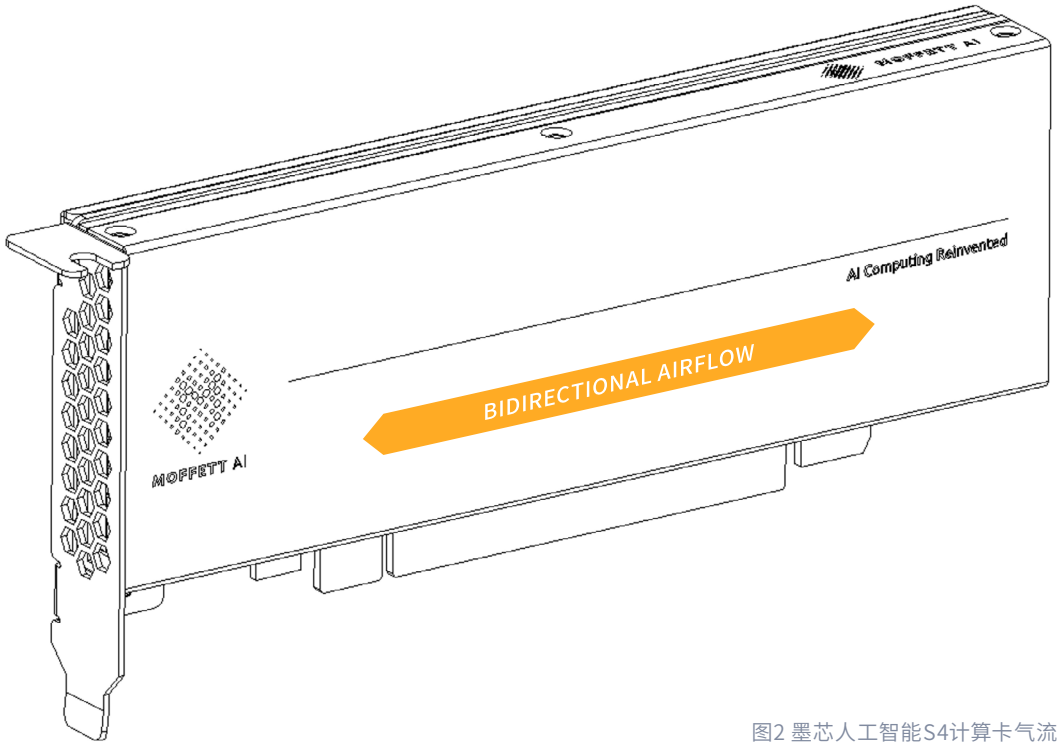

图2 墨芯人工智能S4计算卡气流方向

产品功能

PCIe 接口规格

墨芯人工智能S4 PCIe卡的接口规格如下所述：

PCIe速率支持

S4 PCIe卡支持PCIe Gen3.0

PN翻转和Lane翻转支持

S4 PCIe卡支持PCIe规范中定义的PN翻转和Lane翻转。当翻转PCIe通道时，Rx通道和Tx通道的顺序都必须颠倒。

硬件信任根

S4 PCIe计算卡通过片上硬件安全引擎和ARM CPU信任区域技术支持硬件根信任。信任根的基本功能包括安全引导和安全固件升级。S4　PCIe卡可以通过加密和认证进一步保护用户的AI模型，由强大的密钥管理系统和硬件信任根支持。

多实例SPU支持

S4 PCIe卡支持最多4个多实例SPU (稀疏处理单元)。墨芯人工智能MIS (Multi-Instance SPU) 技术可以将S4计算卡划分为多个单个实例，每个实例与自己的DDR、片上存储器、AI计算核心、视频编解码器和JPEG解码器完全隔离，从而实现计算资源供应和服务质量的优化。

外形规格

墨芯人工智能S4计算卡采取半高半长的设计，标称尺寸如图3所示。

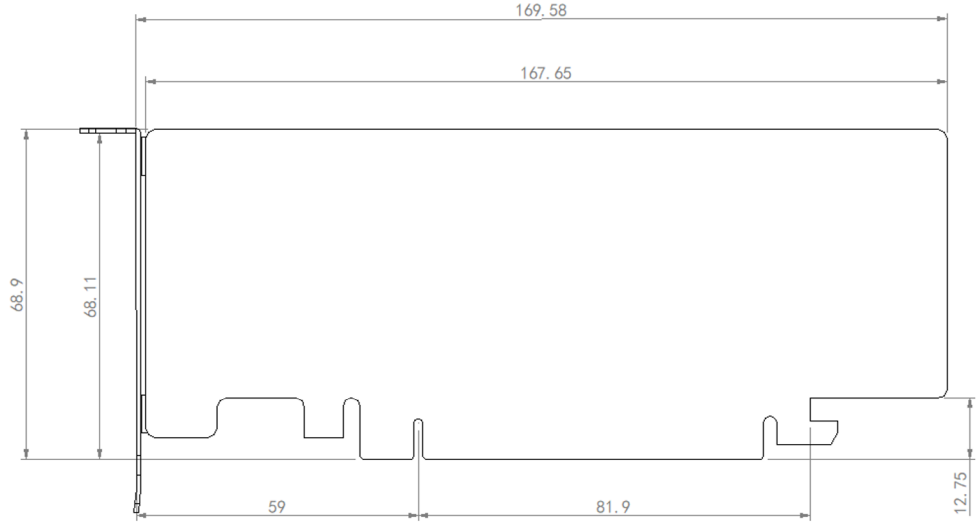

图 3 MOFFETT S4 PCIe 卡尺寸

支持信息

证书

EMC RoHS Reach CE

声明

本文档仅供参考之用，不应被视为对产品的某种功能、条件或质量的保证。墨芯人工智能科技 (深圳) 有限公司 (以下简称墨芯) 对本文档中所包含信息的准确性或完整性不作任何明示或暗示的声明或保证，对本文档中所包含的任何错误不承担任何责任。墨芯对该信息的使用可能导致的对第三方专利或其他权利的任何侵犯不承担责任。本文档不作为对开发、发布或交付任何材料、代码或功能的承诺。墨芯保留随时对本文件进行更正、修改、增强、改进和任何其他变更的权利，恕不另行通知。

客户应在下订单前获得最新的相关信息，并应核实该等信息是否最新完整。除非墨芯授权代表与客户签署的个人销售协议另有约定，墨芯产品的销售受订单确认时提供的墨芯标准销售条款和条件的约束。本文件不直接或间接构成任何合同义务。

墨芯产品的设计、授权或担保不用于军事、医疗或生命支持设备，也不适用于由产品故障导致人身伤害、死亡或财产或环境损害的应用。墨芯不承担在该设备或应用中包含和/或使用墨芯产品的责任，因此其使用的风险由客户自行承担。

本文件未授予任何明示或暗示墨芯专利权、版权或其他知识产权许可。只有在墨芯事先书面批准的情况下，在完全遵守所有适用的法律法规，并伴有所有相关条件、限制和通知的情况下，才允许复制本文件中的信息。

商标

墨芯、墨芯人工智能、Antoum®是墨芯在中国和其他国家的商标和/或注册商标。

版权

MOFFETT AI

扫码关注公众号

扫码进入官网

深圳 (总部)
地址：深圳市南山区粤海街道微软科通大厦24D
电话：0755-86700125

上海
地址：上海市徐汇区漕宝路650号桂林高智科技大楼1101-1102

北京
地址：北京市朝阳区融科望京中心B座2202A