全爱 QA200EP AI 加速卡

技术白皮书（型号 3000）

文档版本 01

发布日期 2021-12-28

全爱科技（上海）有限公司

全爱科技

版权所有　全爱科技（上海）有限公司2021。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明

全爱科技 后羿 JTDS 二郎神

和其他全爱商标均为全爱科技（上海）有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受全爱科技商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，全爱公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

全爱科技（上海）有限公司

地址：　　上海市闵行区剑川路930号D栋3层 邮编：200240

网址：　　www.quanaichina.com

前言

概述

本文详细介绍 QA200EP 加速卡（型号 3000），包括外观特点、性能参数和配置应用等，让用户对QA200EP加速卡（型号 3000）有一个深入细致的了解。

读者对象

本文档主要适用于以下工程师：

- 售前工程师
- 渠道伙伴售前工程师
- 企业售前工程师

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
危险	表示如不避免则将会导致死亡或严重伤害的具有高等级风险的危害。
警告	表示如不避免则可能导致死亡或严重伤害的具有中等级风险的危害。
注意	表示如不避免则可能导致轻微或中度伤害的具有低等级风险的危害。
须知	用于传递设备或环境安全警示信息。如不避免则可能会导致设备损坏、数据丢失、设备性能降低或其它不可预知的结果。 “须知”不涉及人身伤害。
说明	对正文中重点信息的补充说明。 “说明”不是安全警示信息，不涉及人身、设备及环境伤害信息。

版权所有 © 全爱科技（上海）有限公司

修改记录

文档版本	发布日期	修改说明
01	2021-12-28	第三次正式发布。 修改**3.1 基本规格**中的“AI算力”。

版权所有 © 全爱科技（上海）有限公司

目 录

版权所有 © 全爱科技（上海）有限公司

1 产品简介

1.1 概述

QA200EP加速卡（型号 3000）采用1个海思昇腾310 AI处理器（Ascend 310 AI处理器）的PCIE HHHL卡，配合主设备（ARM和X86各种服务器，已适配intel、鲲鹏、海思、飞腾等品牌），实现快速高效的处理推理、图像识别及处理等工作。

说明

昇腾310是一款全爱专门为图像识别、视频处理、推理计算及机器学习等领域设计的高性能、低功耗AI芯片。芯片内置2个AI core，可支持128位宽的LPDDR4X，最高可提供22TOPS INT8/PCS 的计算能力。

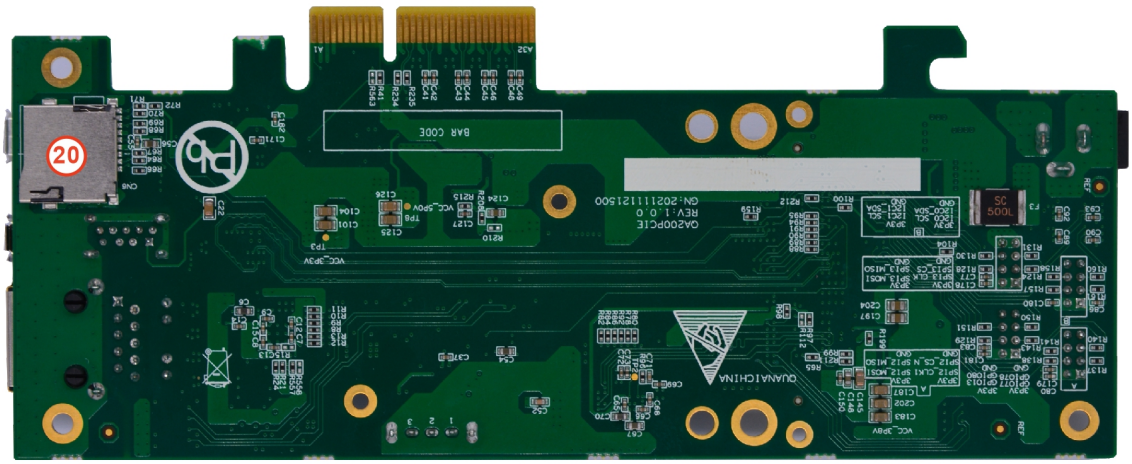
1.2 外观

QA200EP加速卡（型号 3000）外观如图1-1和图1-2所示。

图 1-1 QA200EP 加速卡（型号 3000）上视外观图

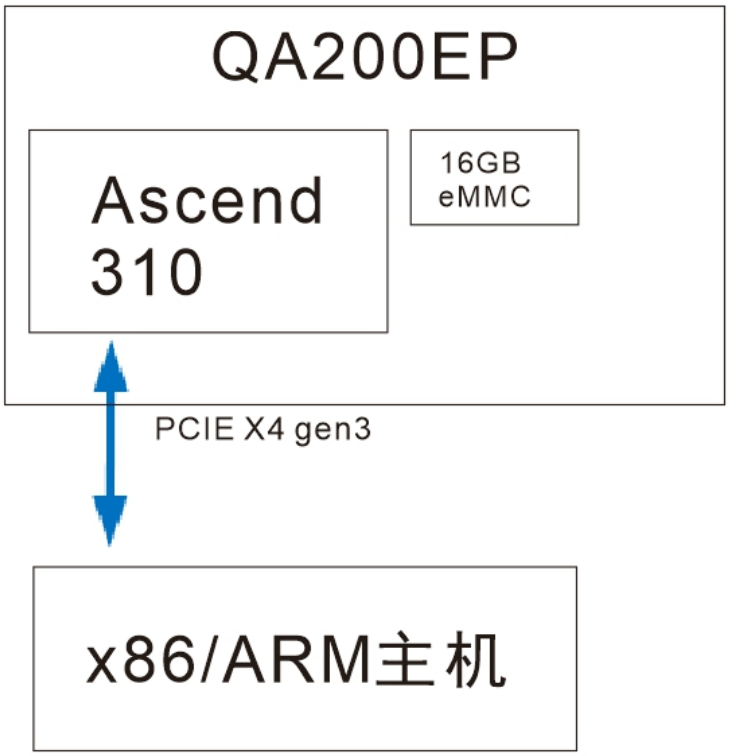
版权所有 © 全爱科技（上海）有限公司

图 1-2 QA200EP 加速卡（型号 3000）底视外观图



1.3 系统框图

QA200EP加速卡（型号 3000）的系统框图如图1-3所示。**图 1-3** QA200EP 加速卡（型号 3000）系统框图



版权所有 © 全爱科技（上海）有限公司

2 产品特点

2.1 性能特点

- 采用1个高性能低功耗的海思昇腾310 AI处理器，最高可提供22TOPS INT8/PCS Ascend 310的计算能力。
- 支持多种规格的H.264、H.265视频编解码，适用于用户不同的视频处理需求。

2.2 可维护性特点

- 支持带内的在线升级，方便客户进行日常维护。
- 支持带内及带外获取温度、电压、功耗等设备状态信息，图形界面让管理更简单。
- 完备的命令行管理功能，用户可以通过各种命令进行日常的设备管理。
- 支持带内及带外资产管理功能，提供生产日期、序列号等信息，方便资产管理。

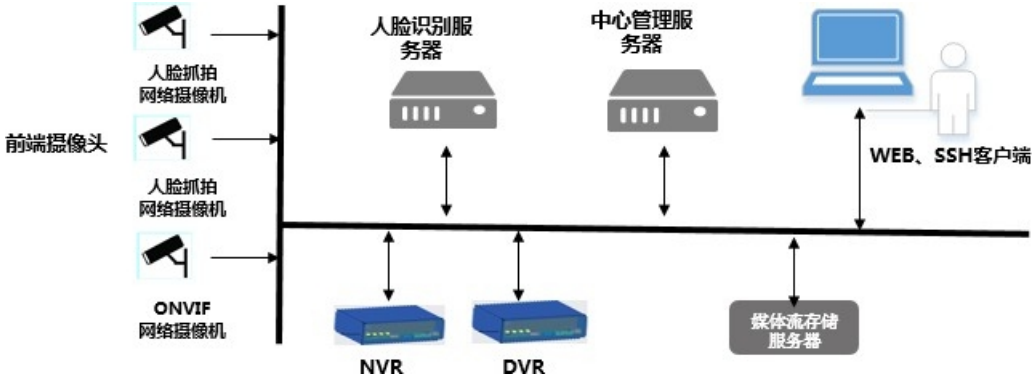
2.3 典型应用场景

QA200EP加速卡（型号 3000）典型应用场景为人脸识别系统，系统主要采用了人脸检测算法、人脸跟踪算法、人脸质量评分算法以及高速人脸对比识别算法，实现了实时人脸抓拍建模、实时黑名单对比报警和人脸后台检索等功能。

人脸识别系统架构如**图2-1**所示，主要部件有系统前端高清网络摄像机或人脸抓拍机、媒体流存储服务器（可选）、人脸智能分析服务器、人脸比对搜索服务器、中心管理服务器、客户端管理软件等组成。QA200EP加速卡（型号 3000）部署在人脸智能分析服务器中，主要实现视频解码/预处理、人脸检测推理，人脸对齐（矫正）和人脸特征提取推理功能。

版权所有 © 全爱科技（上海）有限公司

图 2-1 典型的人脸识别系统架构图



版权所有 © 全爱科技（上海）有限公司

3 产品规格

3.1 基本规格

表 3-1 QA200EP 加速卡（型号 3000）规格

特征	规格
形态	Low Profile标卡，支持全高和半高两种拉手条
AI处理器	昇腾310 AI处理器，一个 • 2个DaVinci AI Core • 8个A55 Arm Core（最大主频1.6GHz）
AI算力	• 半精度（FP16）：11 TFLOPS • 整数精度（INT8）：22 TOPS
内存	• LPDDR4X • 容量：8G • 位宽：128bit • 速率：3200Mbps • 总带宽：51.2GByte/s • 支持ECC

版权所有 © 全爱科技（上海）有限公司

特征	规格
编解码能力	• 支持H.264 Decoder硬件解码，16路1080P 30FPS（8路3840 x 2160 60FPS），YUV420 • 支持H.265 Decoder硬件解码，16路1080P 30FPS（8路3840 x 2160 60FPS），YUV420 • 支持H.264 Encoder硬件编码，1路1080P 30FPS，YUV420 • 支持H.265 Encoder硬件编码，1路1080P 30FPS，YUV420 • JPEG解码能力1x 1080P 256FPS，编码能力1x 1080P 64FPS，最大分辨率：8192 x 4320 • PNG解码能力1x 1080P 24FPS，最大分辨率：4096 x 2160
PCIe接口	PCIe 3.0 x4，兼容PCIe 2.0/PCIe 1.0 **说明**
功耗	典型功耗为15W
尺寸（长x高）	168.9 mm x 68.9mm x 27 mm
净重	210g

3.2 环境条件

QA200EP加速卡（型号 3000）硬件应用环境条件如表3-2所示。**表**

3-2 QA200EP 加速卡（型号 3000）硬件应用环境条件

环境指标	规格
工作温度	0℃～55℃（32℉～131℉）
存储温度	-20℃～+70℃
工作湿度（RH，无冷凝）	5%～90%
存储湿度（RH，无冷凝）	5%～95%
海拔高度	小于3000m。高于900m使用时，海拔每升高300m最高温度规格降低1℃。

3.3 时钟要求

QA200EP加速卡（型号 3000）遵从标准PCIe标卡协议（PCI Express® Card Electromechanical Specification Revision 3.0），整卡只需要提供标准PCIe 3.0（可向下兼容 2.0及1.0）差分时钟，信号质量满足PCIe规范。

版权所有 © 全爱科技（上海）有限公司

3.4 热插拔

QA200EP加速卡（型号 3000）不支持热插拔。

3.5 电源管理

QA200EP加速卡（型号 3000）遵从标准PCIe标卡协议（PCI Express® Card Electromechanical Specification Revision 3.0），整卡最大功耗25W，要求对应QA200EP加速卡（型号 3000）槽位可提供3A@12V及0.5A@3.3V标准供电能力。

3.6 散热规格

3.6.1 散热要求

QA200EP加速卡（型号 3000）用于带风扇的主动散热环境，支持双向进风出风，风量必须满足散热要求。

表 3-3 QA200EP 加速卡（型号 3000）散热要求

入风口平均温度/℃	需求最低风量/CFM	压降/Inch H_2O
55	6.7	0.53
50	5.3	0.37
45	4.4	0.28
40	3.7	0.21
35	3.3	0.18
30	3.0	0.16
任何场景	3.0	0.16

说明

- 需求的最低风量为通过QA200EP加速卡（型号 3000）散热器风量。
- 散热器入口环境温度为进风口的平均温度。
- 需求的风量是建议值，不同系统提供给QA200EP加速卡（型号 3000）的风量和温度可能存在差异，需要根据实际系统进行实测确定。
- QA200EP加速卡（型号 3000）上电状态，需要有风量进行散热，需求的最低风量为3.0CFM。

3.6.2 散热规格

QA200EP加速卡（型号 3000）支持入口温度为0℃～55℃，内部有温度监控点，带内及带外均可对Ascend 310、存储芯片进行实时监控，确保该卡在工作过程中，系统的散热情况需保证该卡的温度值低于规格值。

版权所有 © 全爱科技（上海）有限公司

表 3-4 关键器件温度规格

规格	Ascend 310温度 ℃	存储芯片温度 ℃
下电温度	106	100
降频温度	101	90
长期工作温度	100	85

3.6.3 过温保护

QA200EP加速卡（型号 3000）支持带外及带内通道检测Ascend 310及存储芯片等关键器件的结温，同时也支持检测整板温度，QA200EP加速卡（型号 3000）日志中会记录最高温度和过温的次数及总时间等信息。

QA200EP加速卡（型号 3000）的主要器件Ascend 310及其存储芯片最高支持入风口温度为55℃，为保证可靠工作，外界需提供散热需求的风量，并设计了如下温控策略，QA200EP加速卡（型号 3000）使用了2级预警机制：

- 第一级为严重告警，Ascend 310芯片的严重告警阈值为101℃，存储芯片的严重告警阈值为90℃。当芯片结温或环境温度达到该值，固件就会限制设备的性能。
- 第二级为致命告警，Ascend 310芯片的致命告警阈值为106℃，存储芯片的致命告警阈值为100℃。当芯片结温或环境温度达到该值，QA200EP加速卡（型号 3000）会启动自身下电。

版权所有 © 全爱科技（上海）有限公司

4 维护管理

QA200EP加速卡（型号 3000）提供了丰富的维护管理功能，包括运行在OS中的带内管理命令集和通过BMC提供的带外管理功能。

说明

如果AI芯片没有加载驱动，则带外管理无法准确识别AI芯片是否真正发生故障，因此带外对AI芯片失效场景不做告警提示。带内管理只提供AI芯片健康状况的查询，如果上层业务需要对AI芯片失效场景做实时告警，需要上层业务调用DCMI API中相关接口，并做相关处理。

4.1 带内管理

带内管理的功能有：

- 在线升级功能，升级Firmware，方便用户的设备维护。
- 资产管理功能，提供生产日期、序列号等信息，方便用户进行资产管理。
  具体资产管理操作请参见《**QA200EP加速卡 用户指南（型号 3000）**》中“npu-smi工具”章节。

4.2 带外管理

QA200EP加速卡（型号 3000）提供SMBUS接口，支持服务器的带外管理功能。BMC提供带外管理功能，包括资产信息及监控QA200EP加速卡（型号 3000）温度、电压、实时功耗及芯片监控状态等信息。同时BMC能够接管QA200EP加速卡（型号 3000）的对应告警信息。

- QA200EP加速卡（型号 3000）的具体带外管理功能请参见Atlas300I推理卡/所属产品的BMC用户指南。
- QA200EP加速卡（型号 3000）的具体告警信息请参见所属产品的Atlas300I推理卡/所属产品的BMC告警参考。

版权所有 © 全爱科技（上海）有限公司

5 维保

保修期限1年，详细信息请参见《**维保与保修信息**》。

版权所有 © 全爱科技（上海）有限公司

A 缩略语

A		
AI	Artificial Intelligence	人工智能
B		
BMC	Baseboard Management Controller	主板管理控制单元
C		
CFM	Cubic Feet Per Minute	立方英尺每分钟
E		
ECC and	Error Checking Correction	误差核对与改正
O		
OS	Operating System	操作系统
P		
PCIe	Peripheral Component Interconnect Express	快捷外围部件互连标准
S		
SMbus	System Management Bus	系统管理总线

版权所有 © 全爱科技（上海）有限公司