昆仑芯 AI 加速卡 R200

产品手册

文档版本：V2.0

文档时间：2022-08-29

目　　录

1 版本声明

欢迎使用昆仑芯（北京）科技有限公司（以下简称“昆仑芯”）产品，在您即将使用昆仑芯的产品或服务前，您同意接受下列条款和条件的约束。

一、免责声明：

客户对昆仑芯产品或服务的使用承担风险，昆仑芯对此不作任何类型的担保，不论是明确的或隐含的。昆仑芯不对任何直接、间接、特殊及附带的损害承担责任，这些损害可能来自：未按本说明书要求使用昆仑芯产品，或客户自行变更设计等。

二、信息准确性

昆仑芯对通过本产品所获得的资源均按现状提供，对其准确性、内容、完整性、合法性、可靠性、可操作性或可用性不承担任何责任。昆仑芯对通过本产品所获得的产品或服务以及通过本产品任何链接而得到的任何讯息不做任何担保。昆仑芯保留在任何时候不经通知对本说明书或本说明书所述产品作出修改和变更的权利。

三、知识产权声明

本说明书中所有官方产品、技术、软件、程序、数据及其他信息（包括文字、图标、图片、照片、音频、视频、图表、色彩组合、版面设计等）的所有权利（包括著作权、商标权、专利权）均归昆仑芯及/或其关联公司所有，受法律法规保护。未经昆仑芯事先许可，任何人擅自使用上述内容、信息等，视为对昆仑芯知识产权等合法权益的侵犯，昆仑芯将会追究侵权者的法律责任。

四、版权声明

2 前言

2.1 文档概述

本文档主要讲解了昆仑芯 AI 加速卡 R200（以下简称 R200）的外观外形、技术规格参数、产品特性等。

2.2 读者对象

本文档的主要读者对象包括：

1. 企业终端用户
2. 企业管理员

2.3 版本历史

文档版本	日期	修订内容简述	备注
V1.0	2022-05-20	正式发布第一版	新建
V2.0	2022-08-29	更新发布第二版	更新

3 安全

3.1 安全信息

为了避免操作过程中对人和设备造成伤害，请在操作前，仔细阅读产品相关安全信息。实际操作中，包括但不限于本文描述的安全信息。

3.1.1 运行安全

- 仅专业工程师或我司授权人员才能安全操作该设备。
- 请保持设备清洁、无尘，请勿将设备放置在潮湿的地方或使液体进入设备。
- 设备上电前，请确保设备已可靠接地。
- 为确保充分散热，请勿堵塞设备散热孔。

3.1.2 电气安全

- 请仔细检查工作区域内是否存在潜在的危险，如地面潮湿、或接地不可靠等。
- 请勿在设备带电状态下进行维护操作。

3.2 安全注意事项

- 进行设备维护时，请将设备放在干净、平稳的工作台或地面上。
- 进行设备维护时，为避免设备过热造成人身伤害，请确保设备冷却后再操作。
- 使用工具进行维护时，务必按照正确的操作方式进行，以免危及人身安全

或损伤设备。

- 放置设备时，请勿用力过猛。

3.3 静电防护

3.3.1 防止静电释放

人体或其他导体释放的静电可能会损坏主板和对静电敏感的部件，由静电造成的损坏会缩短设备的使用寿命。为避免静电损害，请注意以下事项：

- 在运输和存储设备时，请将设备装入防静电包装中。
- 防静电包装中取出设备前，请先将设备放置在防静电工作台上，然后再取出。
- 在没有防静电措施的情况下，请勿触摸设备上的插针和电路元器件。

3.3.2 防止静电释放的接地方法

在取放设备时，用户可采取以下一种或多种接地方法以防止静电释放。

- 佩戴防静电腕带，并将腕带的另一端良好接地，请将腕带紧贴皮肤，且确保其能够灵活伸缩。
- 在工作区内，请穿上防静电服和防静电鞋。
- 请使用导电的现场维护工具。
- 使用防静电的可折叠工具垫和便携式现场维修工具包。

4 产品简介

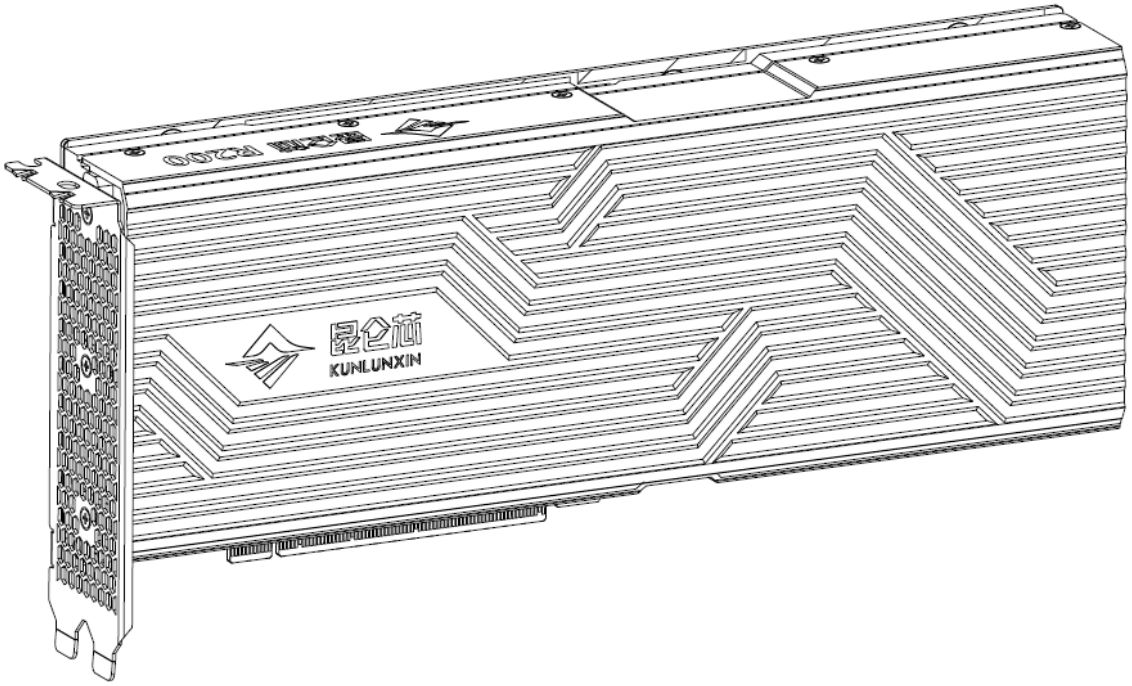
本章重点介绍 R200 的产品外观，硬件规格及产品架构。

4.1 简介

R200 是一款基于昆仑芯自研 XPU-R 架构的 2 代 AI 加速卡，可以提供高达 128T FP16 算力，并且通过支持 GDDR6 可提供高速带宽达 512GB/s。同时 R200 适合数据中心的高性能推理场景，适配服务器主流 X86 服务器，和一系列国产化服务器。R200 的突出表现在数据中心的 CV、NLP、语音和推荐等场景，并且在数据中心服务的严格延时的要求下，可以保证服务的高吞吐量。目前，R200 已大规模部署在包括互联网，智慧交通，智能制造，智算中心等各应用场景。

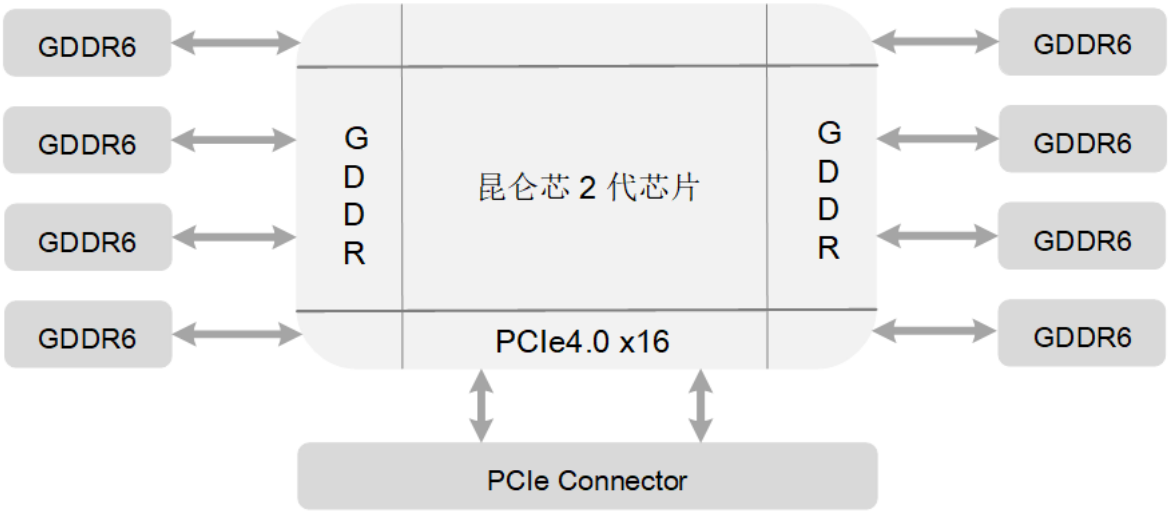
4.2 产品外观

图 4-1 R200 外观图



4.3 产品结构

图 4-2 R200 产品结构图



4.4 产品特性

- 软件定义的芯片架构

- R200 采用昆仑芯自研深度学习软件定义的芯片架构 XPU-R，具有强大的标量、向量和张量运算能力，FP16 算力达到 128 TFLOPS，INT8 算力达到 256 TOPS，通过软硬件联合设计优化，高效支持业界主流深度学习模型加速，性能达到业界主流 GPU 性能的 1.5 到 2 倍。

- 技术领先的存储设计
  - R200 是国内率先支持 GDDR6 的云端 AI 芯片产品之一，提供高达 512GB/s 的存储带宽，具有非常高的能效比和性价比，采用 GDDR6 能极大加速 AI 模型的推理性能。
- 高性能编解码能力
  - R200 支持 108 路高清视频（1080P@30FPS）解码和 27 路高清视频（1080P@30FPS）编码，支持 4K/8K 图片解码。
- 硬件支持多租户架构
  - R200 支持灵活的硬件切分方案，包括 AI 算力、显存和编解码等功能逻辑的切分，满足云计算资源池化业务需求，已率先支持百度智能云线上套餐。
- 高速系统互联接口
  - R200 提供高速的 PCIe 第四代接口，双向带宽可达 64GB/s，同时向下兼容 PCIe 3.0/2.0/1.0，可灵活搭配业界已上市 AI 服务器。

5 产品规格

5.1 规格参数

5.1.1 硬件参数

R200 的硬件参数如**表 5-1** 所示。

表 5-1 硬件参数

规格	说明
TDP	150W
核心频率	1.3 GHZ
工艺	7 nm
PCIe 标识	1d22:3684
供电接口类型	PCIe-8pin
总线类型	PCIe4.0 x16
散热方式	被动
规格	全高全长

最大重量	1.05kg
尺寸（仅板卡）	266.7mm*111.15mm*38.87mm

5.1.2 产品规格

R200 的产品规格如**表** 5-2 所示。

表 5-2 **产品规格**

规格	说明
架构	XPU-R
精度	INT8/INT16/INT32 FP16/FP32
算力	INT8：256 TOPS INT16：128 TOPS INT32：32 TOPS FP16：128 TFLOPS FP32：32 TFLOPS
内存	16GB GDDR6
访存带宽	512GB/s
视频解码	108 路 1080P@30FPS 解码
视频编码	27 路 1080P@30FPS 编码
ECC	支持
SMBUS (I2C)	支持

5.2 R200 接口

R200 接口如**图** 5-1 所示，具体含义如**表** 5-3、**表** 5-4 和**表** 5-5 所示。

图 5-1 R200 接口

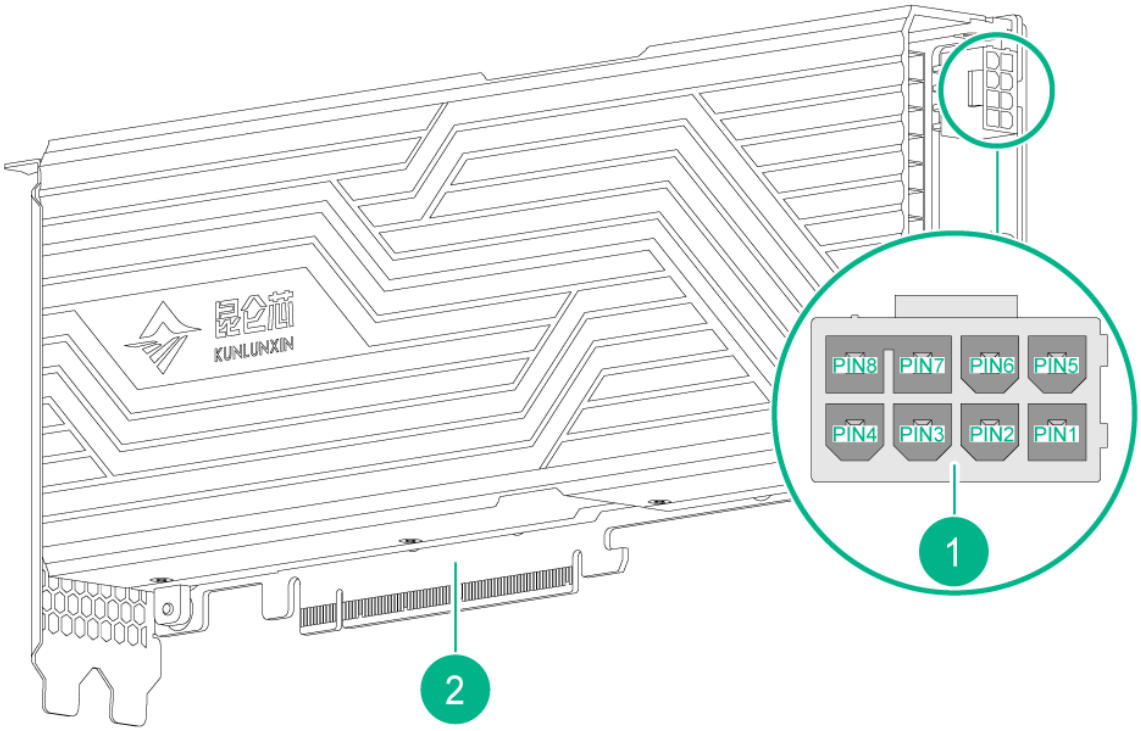


表 5-3 接口说明

管脚（PIN）	信号描述
1	电源连接器（具体管脚描述见表 5-4）
2	PCIe4.0 x16

表 5-4 电源连接器管脚描述

管脚（PIN）	信号描述
PIN1	+12V
PIN2	+12V
PIN3	+12V
PIN4	Sense1（具体 Sense 管脚描述见表 5-5）
PIN5	Ground
PIN6	Sense0（具体 Sense 管脚描述见表 5-5）
PIN7	Ground
PIN8	Ground

表 5-5 Sense 管脚描述

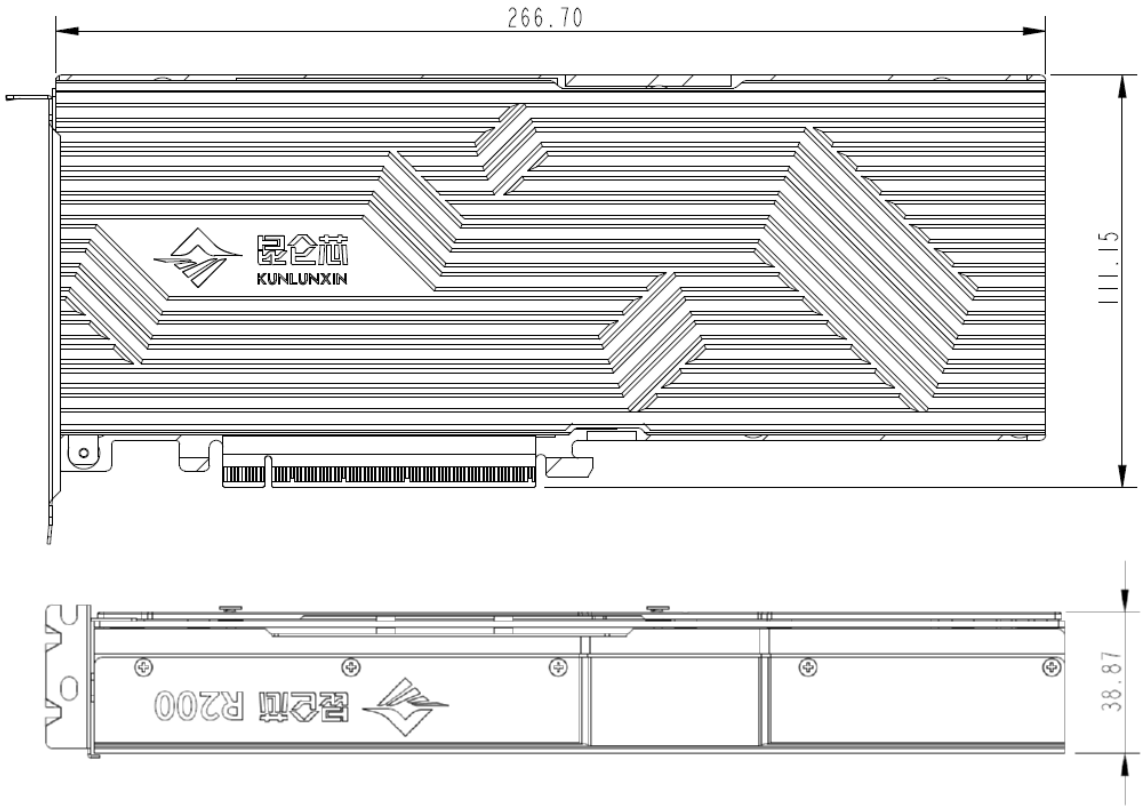
Sense1	Sense0	描述
Ground	Ground	插入的电源线缆为 2*4PIN
Ground	Open	预留

Sense1	Sense0	描述
Open	Ground	插入的电源线缆为 2*3PIN
Open	Open	未插入电源连接器

5.3 R200 尺寸

R200 的尺寸如**图 5-2** 所示。

图 5-2 R200 尺寸



长：266.70mm	宽：111.15mm	高：38.87mm

5.4 输入电源规格

R200 的输入电源规格如**表 5-6** 所示。

表 5-6 R200 输入电源规格

管脚（PIN）	信号描述
PCIE_P12V0	PCIe 接口供电 12V 电源，设计功耗：50W
ATX_P12V0	电源连接器供电 12V 电源，设计功耗：250W
PCIE_P3V3_STBY	PCIe 接口供电 3V3_STBY 电源，为 EEPROM 供电

5.5 运行环境

R200 的运行环境要求如**表** 5-7 所示。

表 5-7 运行环境要求

项目	要求
温度	工作环境温度：0°C～55°C 贮存环境温度：-40°C～70°C
湿度	工作环境湿度：8%～90%RH（无冷凝） 贮存环境湿度：5%～95%RH（无冷凝）

5.6 散热需求

5.6.1 风道方向

R200 支持双风道通风，具体的风道方向如**图** 5-3 和**图** 5-4 所示。

图 5-3 R200 风道方向（一）

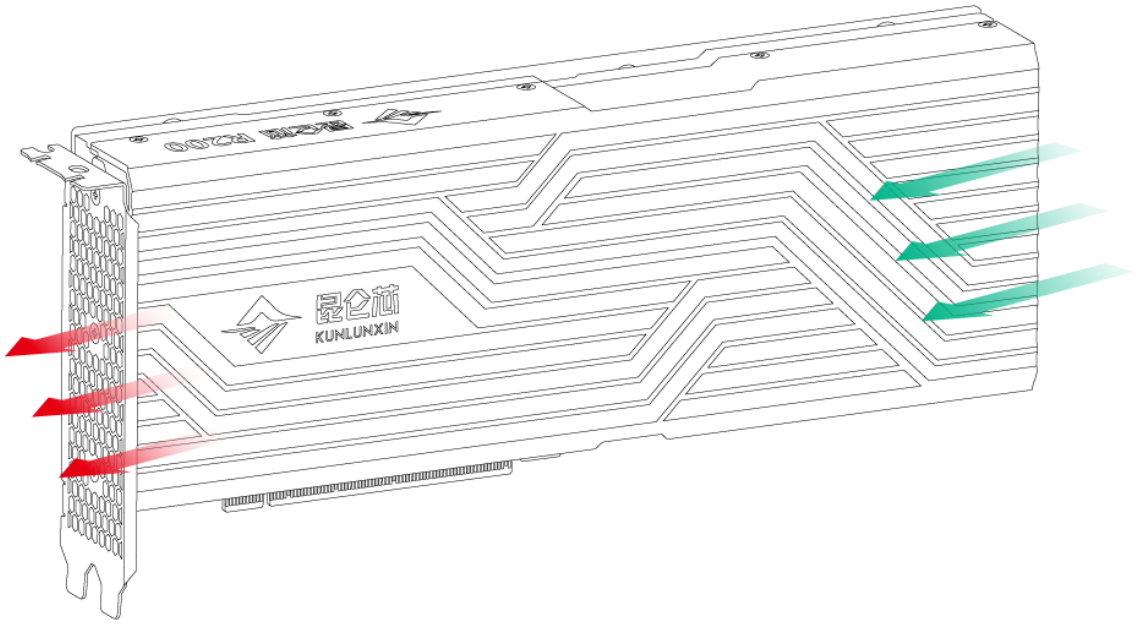
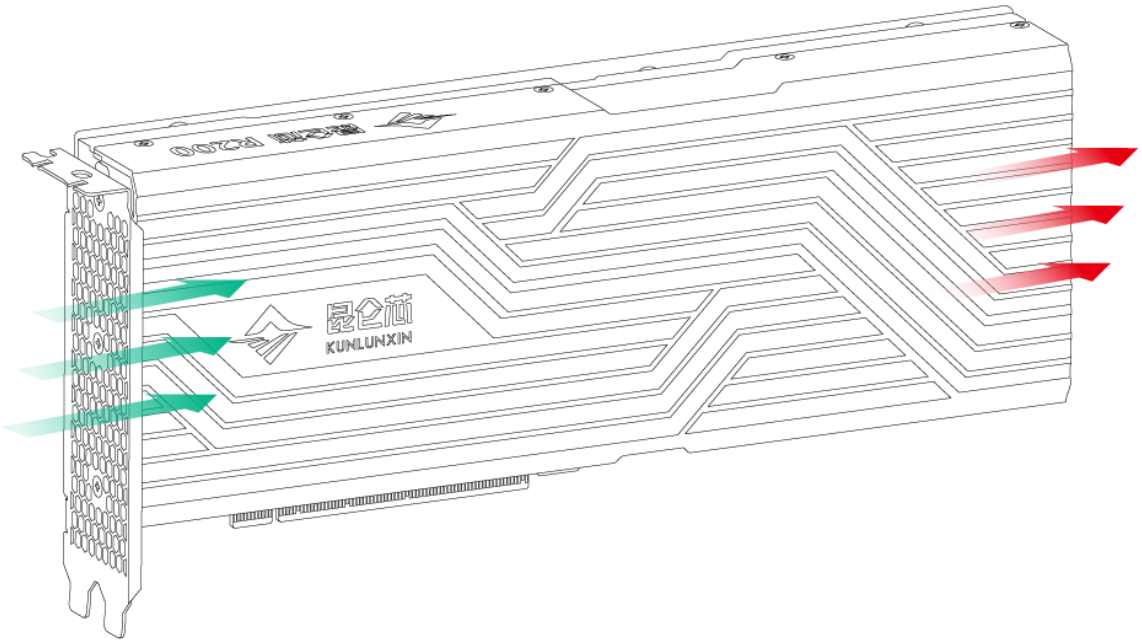


图 5-4 R200 风道方向（二）



5.6.2 入风口温度与风量需求

R200 的入风口温度和风量需求如表 5-8 所示。

表 5-8 R200 入风口温度与风量需求

R200 入风口温度（℃）	风量需求（CFM）
30℃	11.78CFM
35℃	15.34CFM
40℃	19.49CFM
45℃	28.06CFM
50℃	40.25CFM

6 附录

6.1 已完成认证

- 《关于限制在电子电气设备中使用某些有害成分的指令》（RoHS 指令）
- 《化学品注册、评估、许可和限制》（EU）－（REACH 认证）

6.2 术语及缩略语

术语及缩略语	说明
AI	Artificial Intelligence 人工智能
ATX	Advanced Technology Extended 英特尔公司在 1995 年制定的主板规格
CFM	cubic feet per minute 立方英尺每分钟，是气体流量单位。
CV	Computer Vision 计算机视觉，开发赋予计算机视觉能力的技术
ECC	Error Correcting Code 是一种能够实现“错误检查和纠正”的技术
EEPROM	Electrically Erasable Programmable read only memory 带电可擦可编程只读存储器
FCBGA	Flip Chip Ball Grid Array 倒装芯片球栅格阵列的封装格式，也是图形加速芯片最主要的封装格式。
FP16	半精度浮点数，即 16bit 的浮点数，用 1bit 作标识，用 5bit 表示指数，10bit 表示小数。
FP32	单精度浮点数，即 32bit 的浮点数，用 1bit 作标识，用 8bit 表示指数，23bit 表示小数。
FPS	Frames Per Second 每秒传输帧数，通俗来讲就是指动画或视频的画面数。
GDDR	Graphics Double Data Rate，GDDR DRAM 是专为图形处理器(GPU)和加速器设计的存储器
GDDR6	Graphics Double Data Rate, version 6，属于第六代版图形用双倍数据传输率存储器（即显卡的缓存）
GPU	graphics processing unit 图形处理单元
INT8/INT16/INT32	进行 8 位、16 位、32 位定点运算。
I2C	Inter－Integrated Circuit，是由 Philips 公司开发的一种简单、双向二线制同步串行总线
NLP	Nature Language Processor 自然语言处理
PCIe	Peripheral Component Interconnect Express，是一种高速串行计算机扩展总线标准
PCIe 4.0 x16	PCIe 4.0 一种最新的协议传输标准，PCIe 4.0×16 指最高的 PCIe 4.0 能够支持 16GT/s 的传输速率。
PIN	管脚，是从集成电路（芯片）内部电路引出与外围电路的接线，所有的管脚构成了这块芯片的接口。
RH	Relative Humidity 相对湿度，指空气中水汽压与相同温度下饱和水汽压的百分比
SMBUS	System Management Bus 系统管理总线
TDP	Thermal Design Power 散热设计功耗
TFLOPS	teraFLOPS，每秒浮点运算多少万亿次

TOPS	Tera OPS, Tera Operation Per Second，处理器计算能力单位，1TOPS 代表处理器每秒钟可进行一万亿次基本运算操作。
X86	泛指一系列基于 Intel 8086 且向后兼容的中央处理器指令集架构。
XPU-R	昆仑芯纯自主研发的芯片硬件架构，XPU-R 用在昆仑芯 2 代系列产品。