百度昆仑芯片

Product Brief

版本更新记录

版本号	日期	变更说明
1.0	2019-10-11	初版发布

目录

第一章 概述……1

第二章 芯片规格书……2

第三章 板卡规格书……3

第四章 软件规格书……5

第五章 环境规格书……6

第一章 概述

百度的昆仑芯片是一款高性能的AI SoC芯片，支持推理（818-100）和训练（818-300）。昆仑芯片采用百度的先进AI架构，非常适合常用的深度学习和机器学习算法的云端计算需求，并能适配诸如自然语言处理、大规模语音识别、自动驾驶、大规模推荐等多种终端场景的计算需求。

昆仑板卡是支持PCIe Express Gen4 x8的通用AI推理或训练加速器，分推理卡（K100）和训练卡（K200）两个型号。其中，推理卡是全高半长的PCIe卡，有8GB HBM内存，而训练卡是全高全长的PCIe卡，有16GB HBM内存。昆仑板卡是作为被动冷却板提供的。作为百度AI平台的一部分，昆仑板卡可无缝支持百度飞桨以及灵活支持TensorFlow等主流计算框架，并支持丰富的AI模型，可提供出色的性能和效率，充分发挥大规模部署的效用。

具体的芯片架构和使用说明，请联系百度商务获取详细的技术手册。

第二章 芯片规格书

表1展示了两款昆仑芯片的主要规格参数。

表1 芯片规格参数

规格	昆仑818-100（推理芯片）	昆仑818-300（训练芯片）
架构	XPU[1]	
精度	INT4/8 FP32 XFP16/32[2]	
算力	INT8: 128TOPS XFP16: 32TOPS XFP32: 8TOPS[3]	INT8: 256TOPS XFP16: 64TOPS XFP32: 16TOPS
Base Clock (MHz)	1000	
HBM Bandwidth(GB/s)	256	512
HBM Memory Size(GB)	8	16
PCIe Express	Gen 4.0 x8	
PCIe IDs	Device ID: 0x1D22 Vendor ID: 0x3684	
Thermal Cooling	Passive	
制程 (nm)	14	
Package	2.5D	

注：

1. XPU是百度AI芯片的架构：Processing Unit for Diverse Workloads。
2. XFP16/32是指XPU FP16/32，是一种百度昆仑芯片自定义的数据格式，对软件提供标准的FP16/32接口，但能实现比标准FP16/32更高的计算精度。
3. TOPS是Tera Operations Per Second。

第三章 板卡规格书

昆仑板卡是通用的AI推理或训练加速器，支持PCIe Express Gen4 x8。

昆仑推理卡-K100的板型是全高半长的PCIe卡，有8GB HBM内存，尺寸及功耗规格如图1所示。

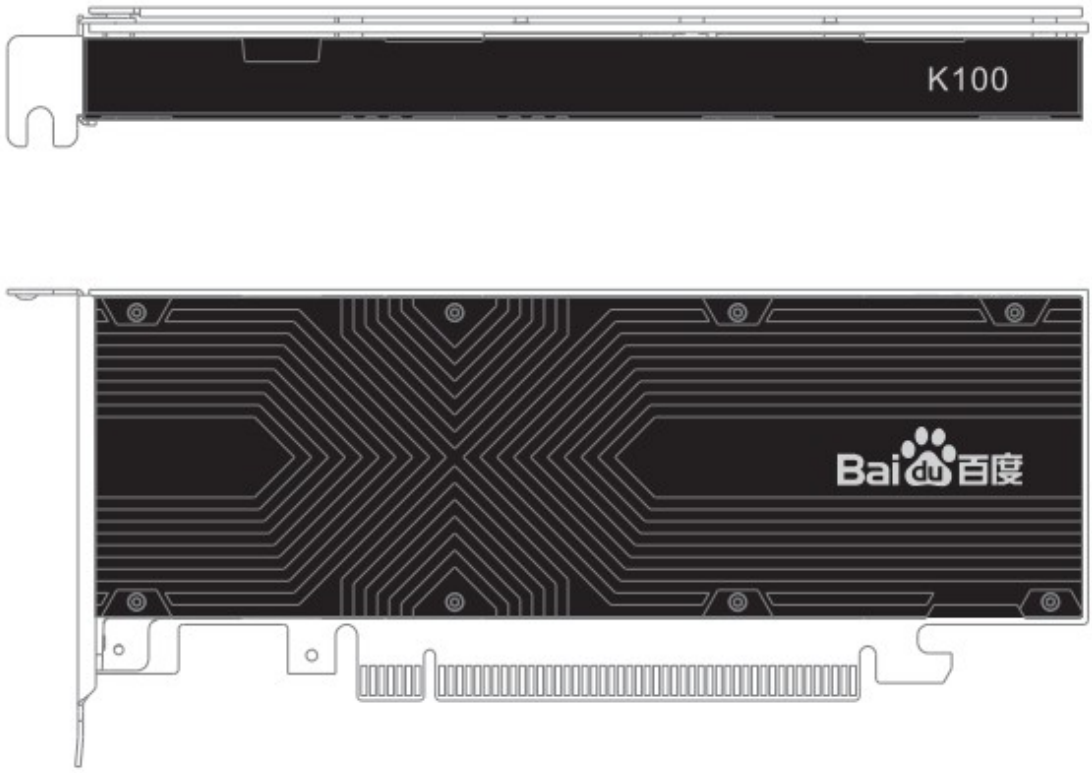

图1 昆仑推理卡-K100

尺寸：167.64mm x 111.15mm x 38.17mm （长 x 高 x 厚）

功耗：75W

昆仑训练卡-K200的板型是全高全长的PCIe卡，有16GB HBM内存，尺寸及功耗规格如图2所示。

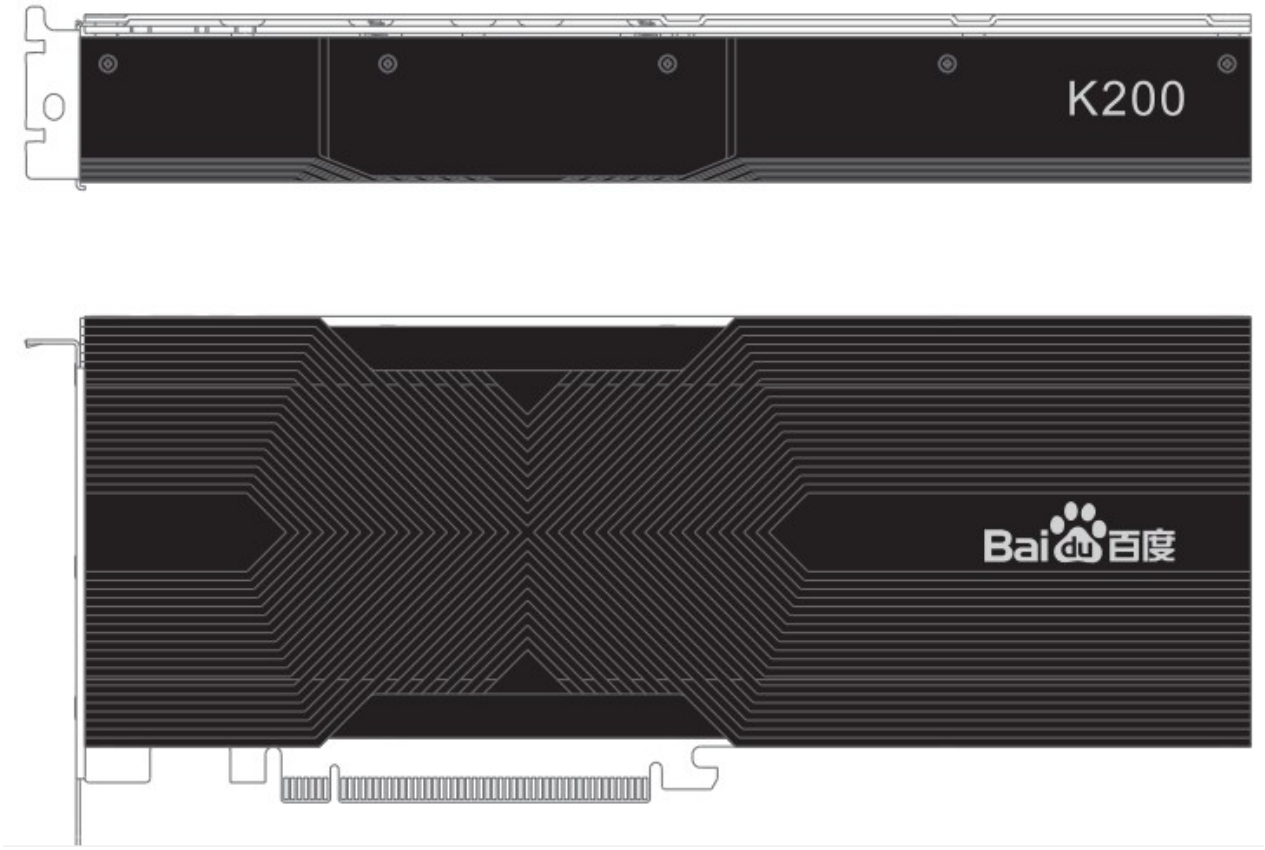

图2 昆仑训练卡-K200

尺寸：266.7mm x 111.15mm x 38.17mm （长 x 高 x 厚）

功耗：150W ~ 200W

第四章 软件规格书

表2展示了昆仑的主要软件规格参数。

表2 软件规格参数

规格	描述
PCIe Base Address (Configurable by Firmware)	Bar 0: 256MB Bar 2: 16MB Bar 4: 128MB
Message Signaled Interrupts	支持MSI 不支持MSI-X
OS Supported	CentOS 6.3 & up Ubuntu 16.4 & 18.4
Runtime	提供库或源码
Driver	提供库或源码
通讯库（训练）	提供库或源码
框架	提供与飞桨的无缝连接 灵活支持TensorFlow及其他主流框架
动态升降频	Firmware控制
软件升级	支持Firmware升级

第五章 环境规格书

表3展示了昆仑板卡的主要环境规格参数。

表3 环境规格参数

规格	工况
环境温度	0 - 50°C
环境湿度	5% - 90%
芯片温度	-25 - 95°C