平头哥

含光800人工智能芯片 Datasheet

【产品简介】

含光800是一款面向数据中心AI应用的人工智能推理芯片，采用台积电12nm工艺，集成高达170亿颗晶体管。其自研神经网络处理器(NPU)架构为AI推理专门定制和创新，包括专有计算引擎和执行单元、192M本地存储（SRAM）以及便于快速存取数据的核间通信，从而实现了高算力、低延迟的性能体验。在业界标准的MLperf测试中，Resnet-50推理性能达到78563 IPS，能效比达500 IPS/W。

含光800支持主流的深度学习框架，包括TensorFlow、MXNet、Caffe、ONNX等，能够以行业领先的性能和空前的效率来处理推理任务，为目前处于快速增长的数据中心领域提供了最优的解决方案。

参数指标

应用	推理
精度	INT8 / INT16
算力	INT8：825 TOPS； INT16：205 TOPS
功耗	276W
总线	PCIe4.0 x 16
深度学习框架	TensorFlow, MXNet, ONNX, Caffe, PyTorch* 备注：*PyTorch模型需要转换成onnx模型

【顶层架构】

含光800人工智能芯片包含命令处理器（CP），核间通信（xCORE-COMM），以及多个NPU内核。

- 命令处理器（CP）是内核驱动与NPU之间的接口，也用于在不同内核间做同步。
- 核的数目可以根据设计情况变化（最多4个），每个NPU核可以独立工作，包括下面子模块：
  - 多个张量计算阵列（Tensor Array），负责卷积、矩阵乘等张量操作（Tensor Operation）
  - 本地存储（Local Memory），与每个Tensor Array对应，负责保存相关数据
  - 向量计算引擎（Vector Engine），负责向量相关的操作
  - 控制单元（SEQ），负责指令的解码、调度和发射、各种事件同步，以及标量寄存器和运算
  - 直接内存存储引擎（DMA），负责提升本地存储与系统内存之间数据交互的传输效率

- 不但单个NPU内核在处理典型的推理（例如ResNet50 v1）时具有行业领先的性能和效率，而且多个NPU内核之间也可以通过核间通信（xCORE-COMM）紧密协作，以便以惊人的速度处理更大更复杂的任务（例如RestNet101，Mask R-CNN等）。

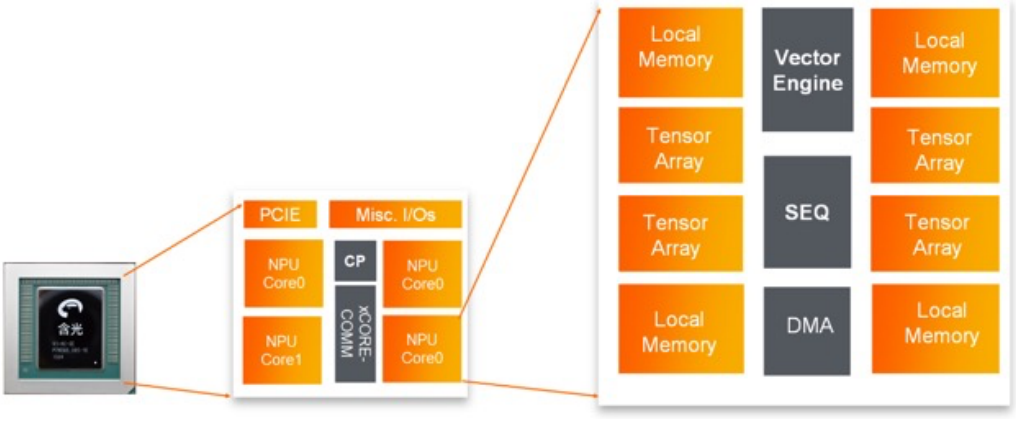

图1-含光800 NPU/NPU核 架构图

【性能】

图2，Resnet-50 V1 INT-8 推理结果

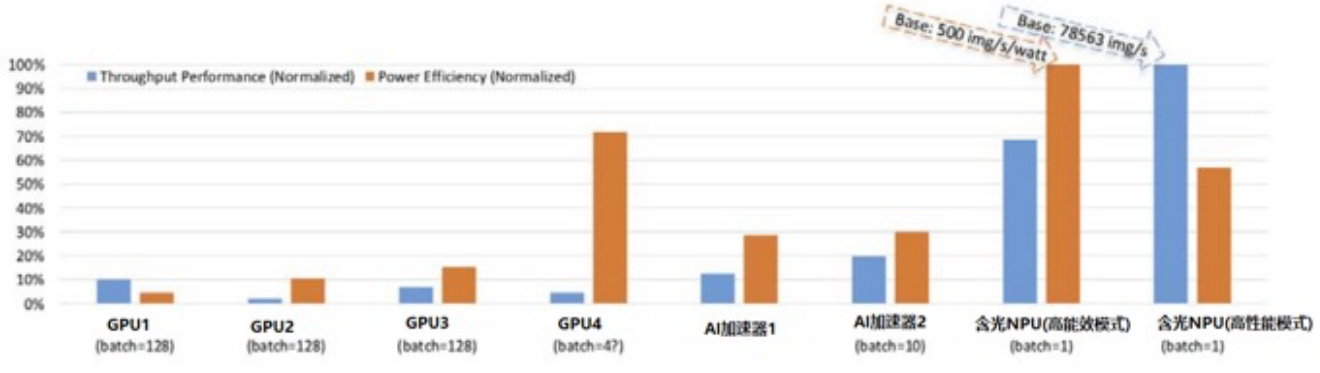

图2-a) 能效比对比

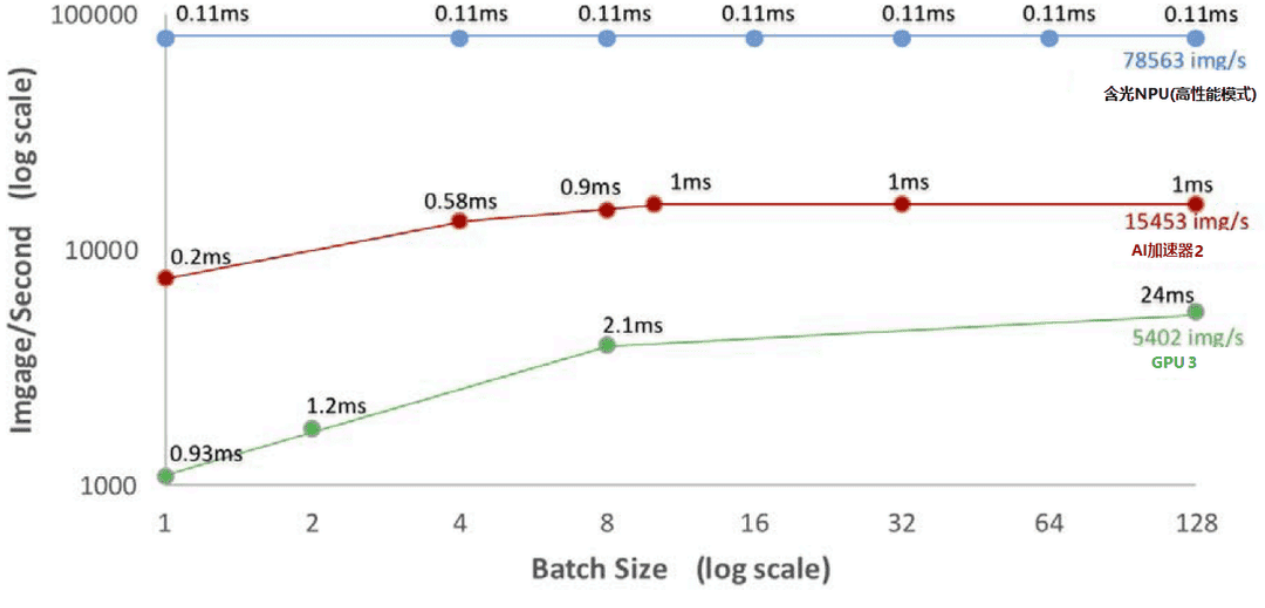

图2-b) 不同batch size下的延迟和吞吐量比较

【部署场景多样性】

各模式下还可以通过调频、调压进一步平衡功耗和性能

Resnet50 v1 (推理)	高性能模式	高能效模式	边缘模式（低功耗模式）		
频率（MHz）	700	475	475	475	475
NPU功耗(Watt)	276	108	75	50	25
性能(Img/S)	78,563	53,983	37,500	25,000	12,500

【配套软件全栈支持】

- 离线量化、编译、以及图优化
- 运行时资源分配、释放等
- 支持多模型动态部署、以及多设备管理
- 设备驱动、调试工具