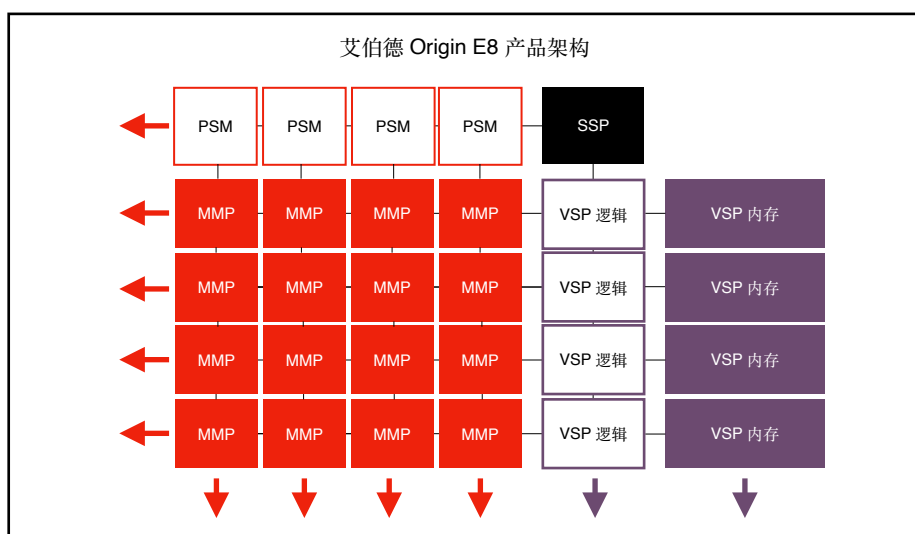


Origin™ E8 深度学习加速器 IP

高性能 AI 引擎，适用于要求最苛刻的推理应用

Origin E8 专为以性能为主要设计目标的应用而设计，包括汽车/ADAS 和数据中心应用。艾伯德的高级内存管理可确保持续的 DRAM 带宽和理想的总体系统性能。Origin E8 深度学习加速器 (DLA) 具有 32 至 128 TOPS 性能，实际利用率高达 90%（在运行 ResNet 等常见工作负载的片上测量），在图像相关的任务中表现出色，如计算机视觉、图像分类和目标检测。



艾伯德的可扩展设计基于 Tiles，包括单个控制器 (SSP)、多个矩阵数学单元 (MMP)、累加器 (PSM)、向量引擎 (VSP) 和用于存储网络的内存。具体配置取决于独特的应用要求。统一计算流水线架构支持高效的硬件调度和高级内存管理，从而实现优秀的端到端低延迟性能。已获得专利的该架构已在数学上被证明，为神经网络 (NN) 执行使用的内存最少。这将大幅降低芯片面积，减少 DRAM 访问，提高带宽，节省功耗，并显著提高性能。

规格

计算能力	16K 至 64K INT8 MAC	功耗效率	高效的 18 TOPS/W (INT8)
神经网络支持	CNN、RNN、LSTM 和其他	数据类型	INT8/INT16、FB16/FB32、BFloat
量化	整个通道，自定义	内存	系统内存分配和调度
框架	TensorFLOW、TFLite、ONNX、TVM	工作负载示例	大型 DNN 网络 (4K/8K 视频)

特征

- 32 至 128 TOPS 性能
- 高效的 18 TOPS/W 典型功耗
- 支持多达 8 个作业
- 高达 64K INT8 MACS 的可扩展性能
- 片上、L3 和 DRAM 协同工作以提高带宽
- 低延迟
- 可针对特定工作负载进行调整
- 硬件调度器
- 支持标准的神经网络功能，包括卷积、反卷积、全连通、激活函数、Reshape、Concat、Elementwise、池化、softmax、双线性插值等
- 可编程通用浮点函数，包括 Sigmoid、Tanh、Sine、Cosine、Exp 等
- 训练好即可处理模型；无需软件优化
- 使用熟悉的开源平台，包括 TFLite
- 作为软 IP 交付；可移植到任何工艺

服务的市场

- 汽车
- 数据中心

应用示例

- ADAS/自动驾驶
- 视频存储优化
- 自然语言处理 (NLP)