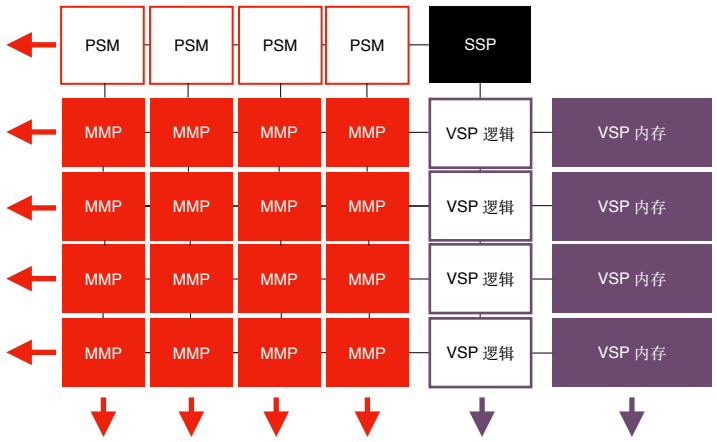
Origin™ E2 深度学习加速器 IP

在功率优化的封装中实现理想 AI 推理性能

艾伯德 Origin E2 专为手机和边缘节点等功耗敏感型设备中的人工智能 (AI) 应用而设计。E2 仅使用片上内存，消除了对外部 DRAM 访问的需求，在提高性能的同时节省了系统功耗，降低了延迟，并缩减了系统 BOM 成本。它可针对特定工作负载进行调整，从而为独特的应用要求提供理想的性能配置文件。

艾伯德 **Origin E2** 产品架构



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，提高带宽，节省功耗，并显著提高性能。

特征

- 高达 20 TOPS 性能
- 18 TOPS/W 典型功耗
- 2 至 9K INT8 MACS 的可扩展性能
- 高级激活内存管理
- 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 和双线性插值
- 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
- 可在片上处理实时高清视频和图像
- 以软 IP 提供：可移植到任何工艺

服务的市场

- 智能手机
- 边缘节点
- 安全

应用示例

- 降低视频流中的微光噪声
- 生物识别、面部检测和识别、增强和虚拟现实以及图像过滤器

规格

计算能力	2.25、4.5 或 9K INT8 MAC	功耗效率	高效的 18 TOPS/W (INT8)
神经网络支持	CNN 和其他架构	数据类型	INT8/INT16 激活值，INT8 权重
量化	整个通道，自定义	延迟	已进行了优化，提供确定保证
框架	TensorFLOW、TFLite、ONNX、TVM	工作负载示例	片上 4K 处理