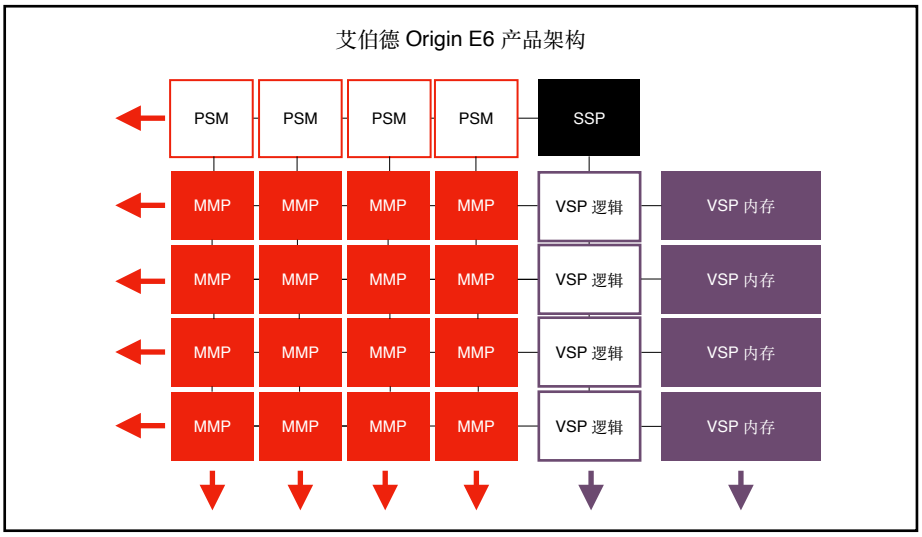
# Origin™ E6 深度学习加速器 IP

## 用于推理应用的高性能 AI 引擎

Origin E6 专为以性能和功耗为主要设计目标的应用而设计，包括智能手机、平板电脑和边缘服务器。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 E6 DLA 具有 16 至 32 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。此外，它还能够执行与 NLP（自然语言处理）相关的任务，如机器翻译、句子分类和生成。




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。

### 特征

- 16 至 32 TOPS 性能
- 18 TOPS/W 典型功耗
- 双作业支持
- 高达 18K 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
- 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
- 以软 IP 提供：可移植到任何工艺

### 服务的市场

- 智能手机
- 平板电脑和计算
- 工业和商业
- 安全

### 应用示例

- 图像和视频优化：微光增强、8MP 实时处理、多摄像头监控
- 自然语言处理

## 规格

| | | | |
|---|---|---|---|
| 计算能力 | 4.5、9 或 18K 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） |