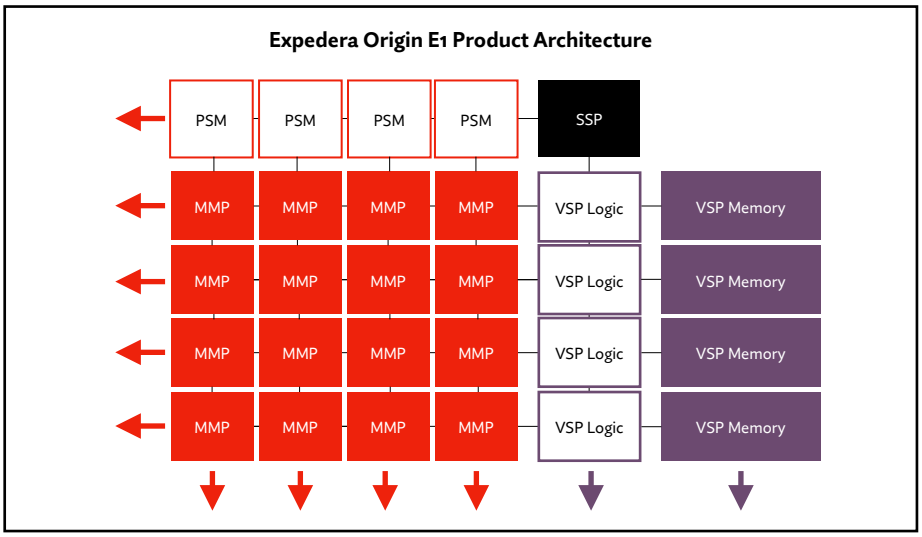
Origin™ E1 Artificial Intelligence Accelerator IP

Neural network-optimized AI processing for power-constrained edge devices

The Expedera OriginTM E1 NPU (Neural Processing Unit) is a series of Artificial Intelligence processing cores which are individually optimized for a class of specific neural networks commonly found in edge, smartphone, and consumer devices. Neural networks have their own unique workload and processing requirements, which often differ significantly from others. These unique differences lead to significant inefficiencies and overhead when non-optimized processing is applied. By tailoring Origin E1 IP to specific neural networks, Expedera provides an NPU that consumes the lowest possible amount of silicon area and external bandwidth while delivering optimum performance and utilization. This insures ideal PPA (power, performance, area) deployment in power-constrained devices.



Expedera's silicon-proven, scalable architecture includes a controller (SSP), matrix math units (MMP), accumulators (PSM), vector engines (VSP) and memory. The unified compute pipeline architecture enables highly efficient hardware scheduling and advanced memory management to achieve unsurpassed end-to-end low-latency performance.

Features

- 1 TOPS performance
- Up to 18 TOPS/W typical power consumption (7nm)
- >80% average utilization
- Tuned for specific neural network workloads
- Minimal or no off chip memory required
- Processes model as trained; no need for software optimizations
- Advanced activation memory management
- Low latency
- Hardware scheduler
- Support for standard NN functions including Convolution, Deconvolution, FC, Activations, Reshape, Concat, Elementwise, Pooling, Softmax, Bilinear & others
- Use familiar open-source platforms including TFlite
- Delivered as GDS or RTL: portable to any process

Markets Served

- Edge nodes
- Television and displays
- Smart doorbells
- Home appliances
- IoT endpoints
- Consumer devices

Example Neural Networks Supported

ResNet 50 V1	MobileNet V1	PicoDet
NanoDet	MobileNet SSD	Unet
Inception V3	BERT	Tiny YOLO V3
RNN-T	EfficientNet	ShuffleNet2