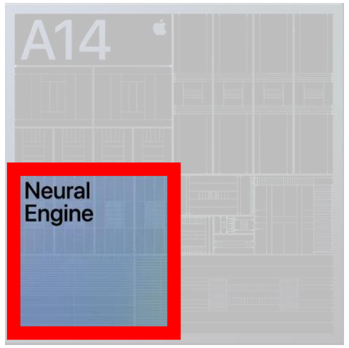
embedded VISION summit

A Novel Packet-Based Accelerator for Resource-Constrained Edge Devices

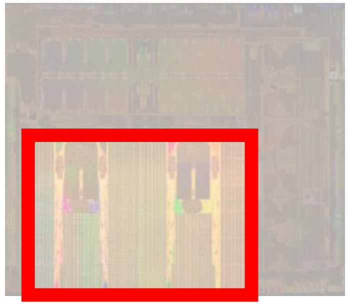
Sharad Chole, Chief Scientist
Expedera

AI Inference Solutions Are Not Sustainable

- Up to 40% of SoC resources are devoted to NPUs
  - Todays NPUs are performance, power, and area inefficient
  - Scaling only amplifies the problem – NPU power and area requirements have grown significantly compared to performance
- NPUs can be, and must be, better to enable the true promise of edge AI
- Alternative architectures can be deployed to support the rapidly increasing needs of edge AI
  - More efficient, highly utilized processors
  - Scalable designs for reuse across platforms
- **Let's show you how we can do this…**



Apple A14
-11 TOPS-



Tesla FSD
-74 TOPS-

Edge AI – Today vs Tomorrow

Current Solutions

- Multi-core architectures with complex compilers
  - Stagnated by low utilization & area efficiency
- Expensive for edge SoCs
  - Require too much DDR (performance/GBps)
  - Low power efficiency (performance/W)
- Applications are struggling to approach markets
  - Single tenant, high latency
  - Small resolutions & low network diversity

Future Needs

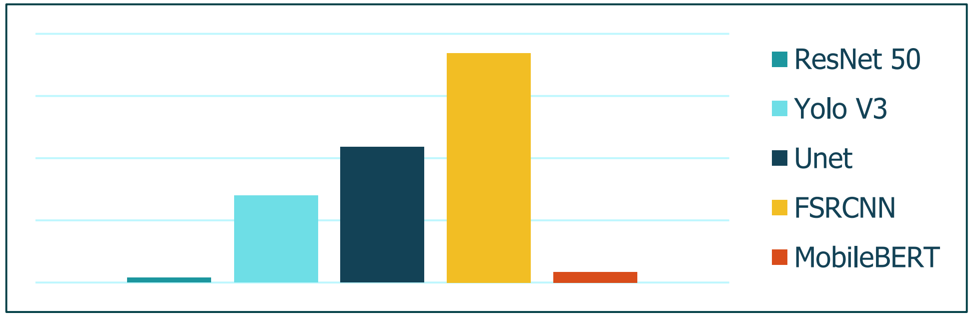
- Scalable monolith with robust software toolchain
  - High utilization & area efficiency
- Built for edge deployments
  - Reduced/no requirement for DDR
  - Enable lower thermal and battery constraints
- Model deployments with SLA guarantees
  - Multi-tenant real time processing, scalable to 4K and beyond

Accelerators must meet tight cost, bandwidth & power consumption constraints while still delivering high performance

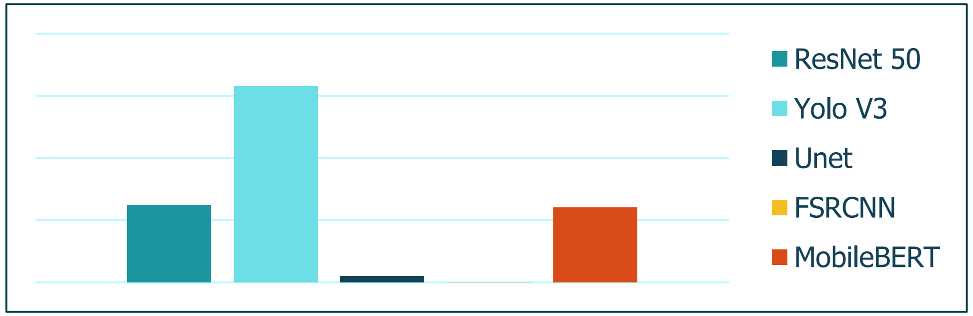
A Growing Diversity of Networks Demands Flexibility

Network	Typical Tasks
ResNet 50	Image classification, feature extraction backbone
Yolo V3	Object detection
UNet	Image segmentation, denoising
FSRCNN	Embedded super resolution
MobileBERT	Language understanding

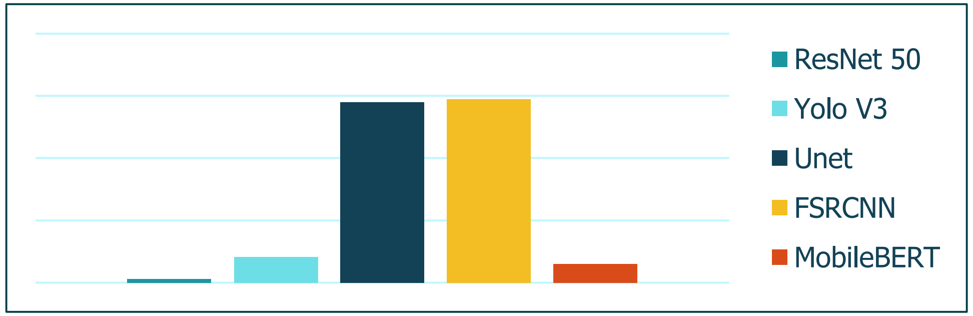
Operations (Billion)



Weights (Million)



Activations (Million)



Edge AI Needs a Different Definition of Compute

Two types of compute abstractions are used today:

Instruction
e.g. CPU, GPGPU, DSP

- Need to distribute workload and gather results back
- Heavy use of synchronization primitives for scalability
- Hierarchical memory systems cost power and area

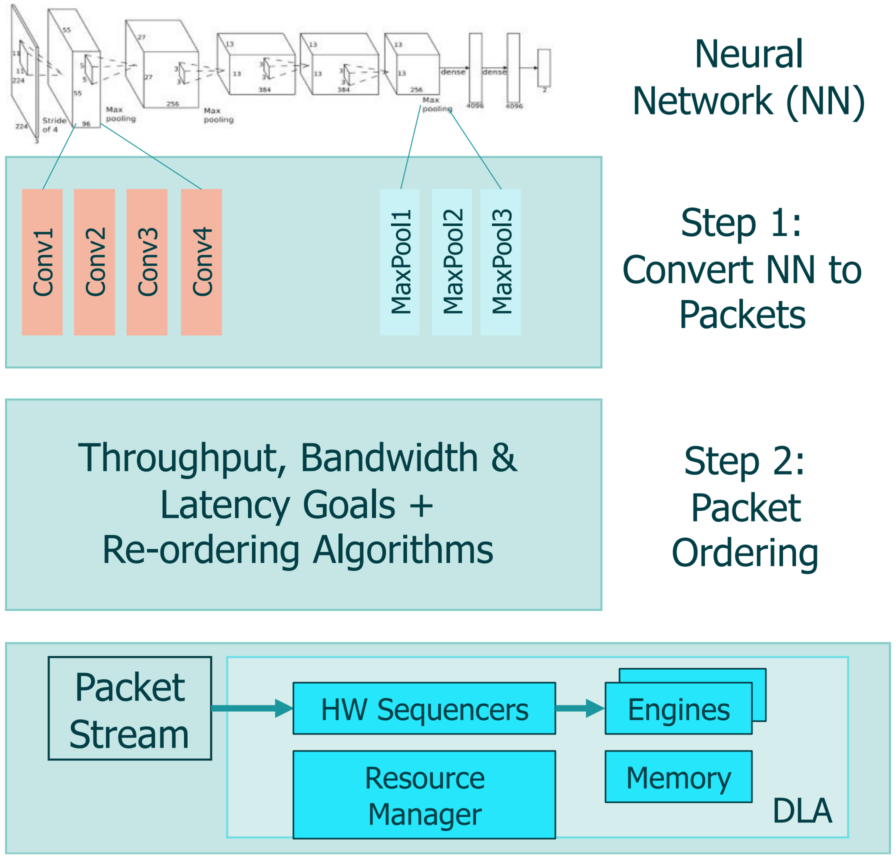
Neural Network Layer
e.g. Systolic Arrays, DNN APIs

- Need large on chip memory to process a layer
- Limited by layer ordering optimizations
- NN orchestration done through controller CPU

Current AI engines are limited by high cost of context switching and cannot break down layers into more granular pieces. Resulting in poor utilization and worse PPA.

Packets: a Radical Approach to AI Inference Optimization

- Aggregate of work with notion of dependencies and performance deterministic execution, based on a network-centric approach
  - A contiguous fragment of a neural network layer with entire context of execution; layer type, attributes, priority
- Manage activations better/more intelligently
  - Reordering packets does not incur penalty – as DLA supports zero cost context switching
- Through reorganizing packets in the optimum order without hurting accuracy, Expedera produces the minimum number of moves
  - Greatly increases performance while lowering silicon power and area requirements



Neural Network (NN)

Conv1 Conv2 Conv3 Conv4 MaxPool1 MaxPool2 MaxPool3

Step 1: Convert NN to Packets

Throughput, Bandwidth & Latency Goals + Re-ordering Algorithms

Step 2: Packet Ordering

Packet Stream

HW Sequencers

Engines

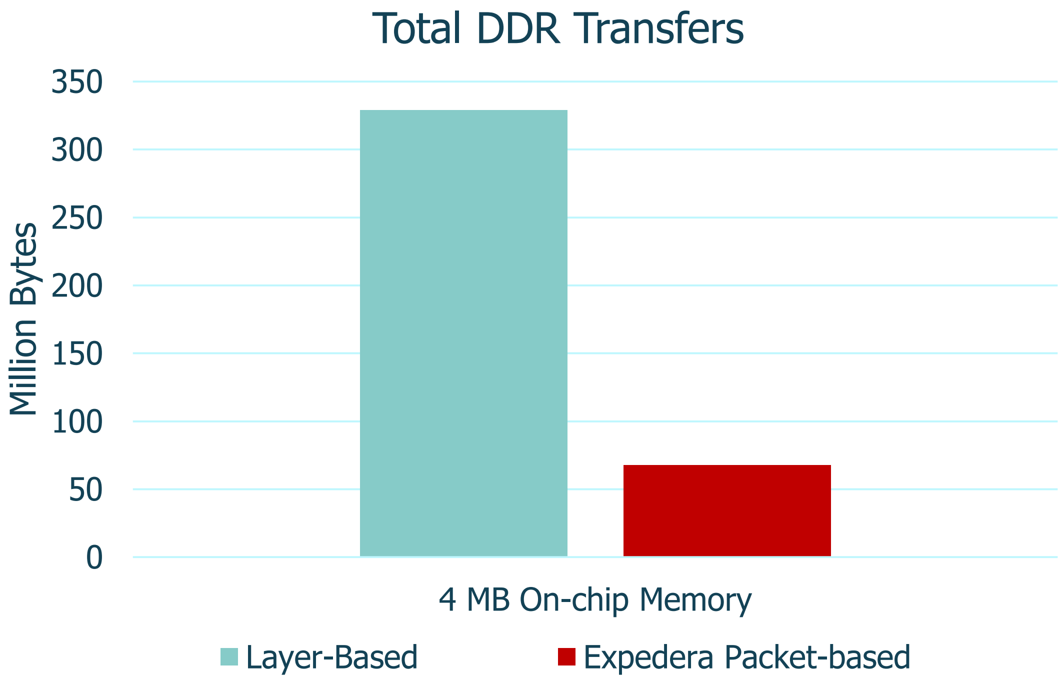
Resource Manager

Memory

DLA

Packets Save Memory and Bandwidth

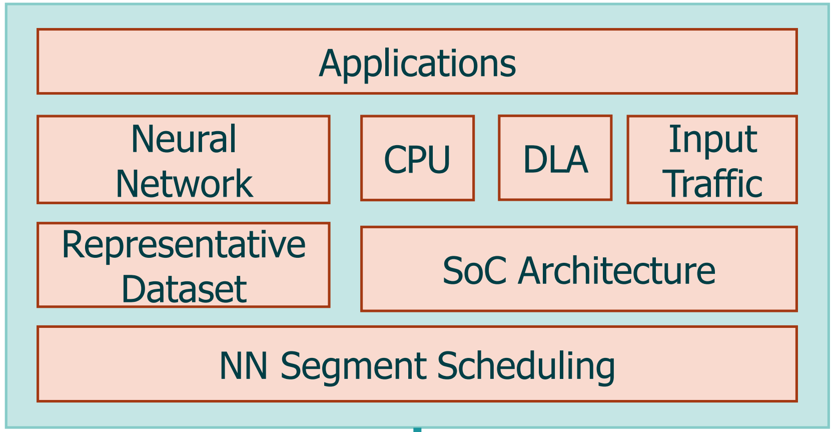
- Example shown:
  - YoloV3 608 x 608, batch of 2
  - Total 63 M weights, largest layer 4.3 M
  - 235 M activations, largest layer 24 M
  - 280 B operations
- Packets reduce DDR transfers by >5x
  - Less intermediate data movement, higher throughput
  - Lower system power, reduced BOM cost
- Uniformly spread-out bandwidth
  - Sustained utilization, tolerant towards latency variations



Expedera requires fewer DDR transfers compared to typical architectures:
Higher throughput, lower power consumption, less chip area required

Packets Allow Right-sizing DLA in Context for SoC

- Packet-stream guarantees cycle-accurate DLA performance
  - Packets are performance deterministic & complete - exact execution cycles as well as memory & bandwidth usage is known
- Quickly right-size DLA for AI workloads with visibility
- **Expedera Estimator** opens a DLA-centric view into AI workload performance of SoC components



Applications

Neural Network

CPU

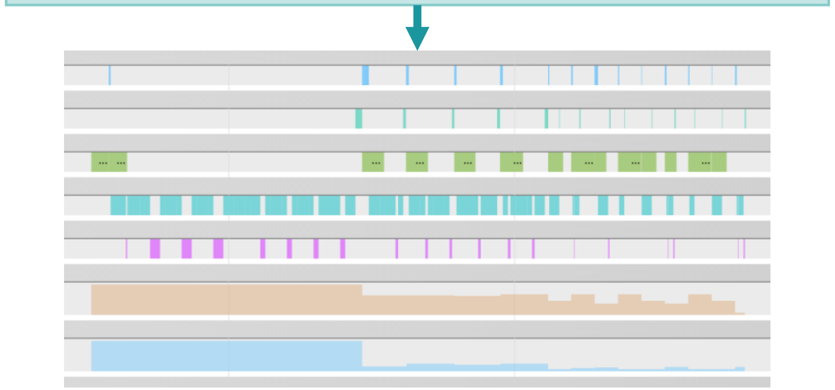
DLA

Input Traffic

Representative Dataset

SoC Architecture

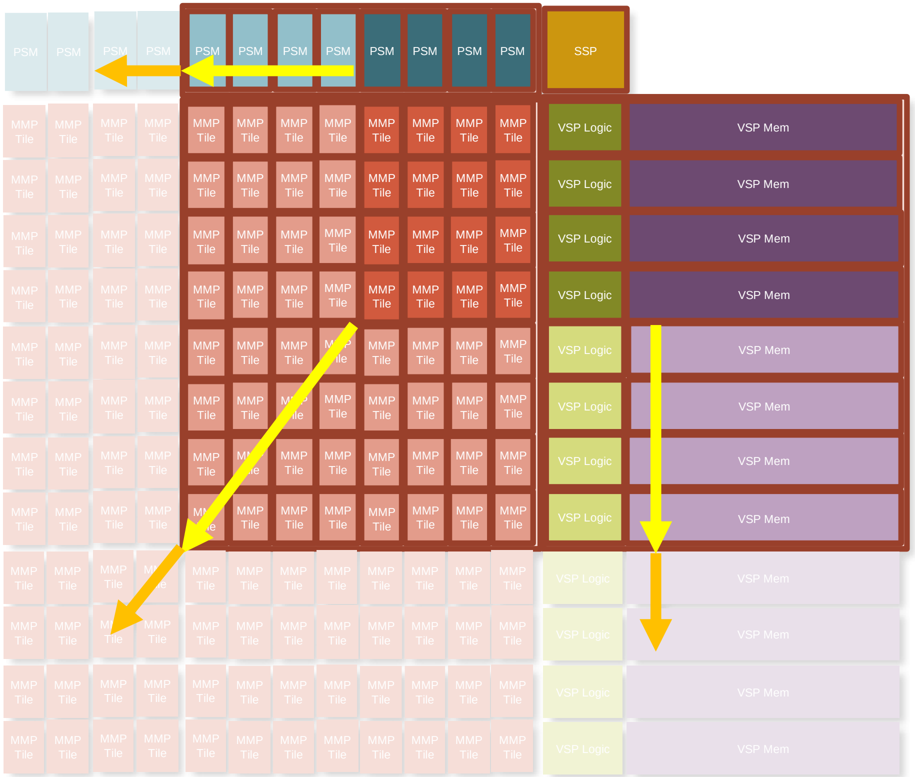
NN Segment Scheduling



Cycle Accurate DLA Trace

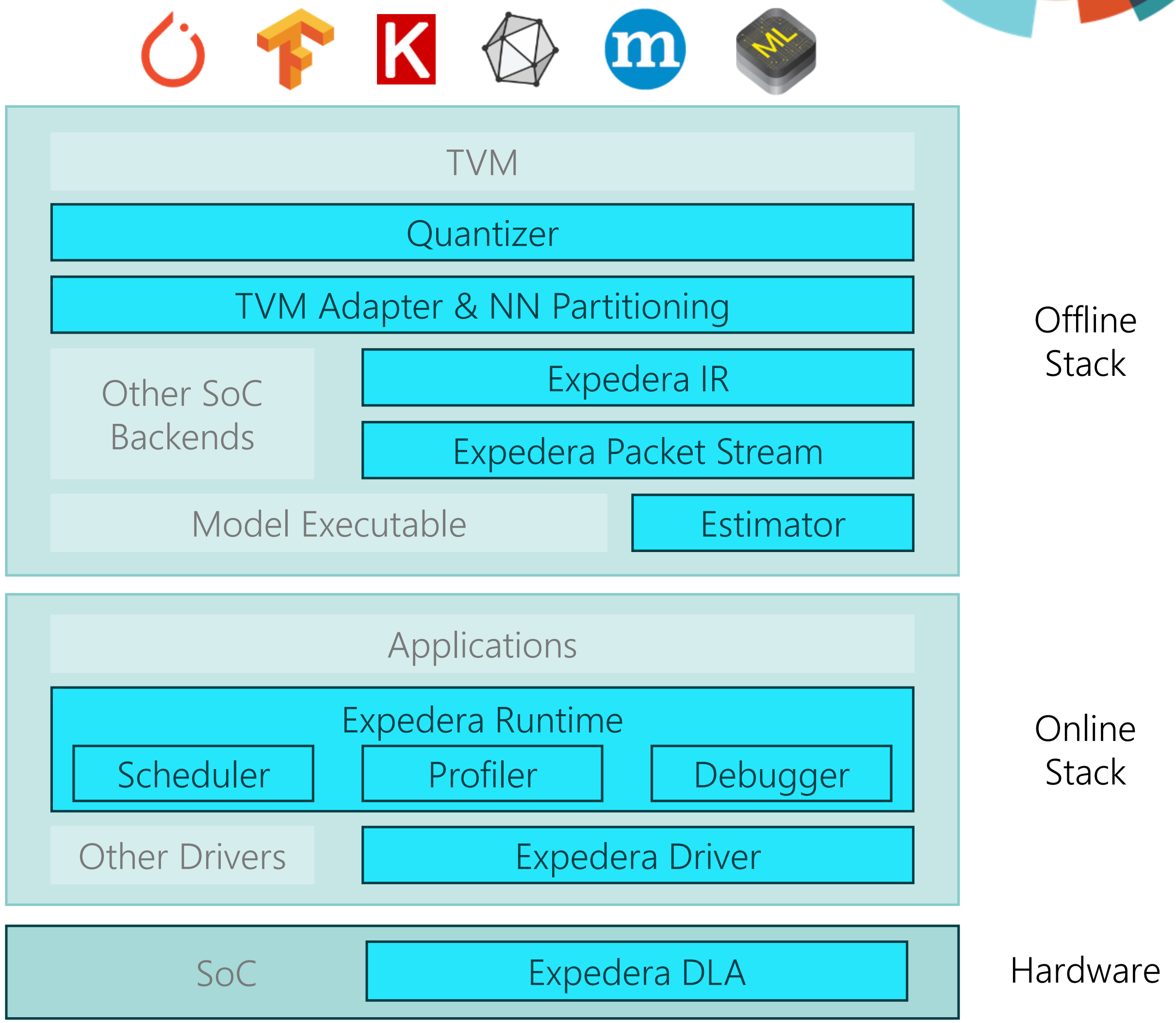
Expedera Origin^{TM} Deep Learning Accelerator (DLA)

- Unique "building block" architecture allows area-efficient DLA configurations matching customer needs
  - Scale compute independent of memory
- Unified pipeline architecture compute
  - 18 TOPS/W (average, not peak; TMSC 7 nm, ResNet50, 1 GHz, no sparsity required)
- Zero overhead context switching
  - 70-90% utilization across entire performance range
- Monolithic - single control system drives entire DLA


PSM PSM PSM PSM PSM PSM PSM PSM PSM PSM PSM PSM SSP
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem
MMP Tile ×12 VSP Logic VSP Mem

TVM-based Software Stack

- Ease of integration into SoC environment
  - DL framework supported through Apache TVM
  - Multi-target compilation
  - Neural network orchestration across SoC
- Extended features
  - Mixed precision quantization
  - Custom layer support
  - Multi-job APIs
- Exploration into FPS, latency & power metrics from architectural to deployment phase



TVM

Quantizer

TVM Adapter & NN Partitioning

Other SoC Backends

Expedera IR

Expedera Packet Stream

Model Executable

Estimator

Offline Stack

Applications

Expedera Runtime

Scheduler

Profiler

Debugger

Other Drivers

Expedera Driver

Online Stack

SoC

Expedera DLA

Hardware

Packet-based AI Processing Leads the Industry

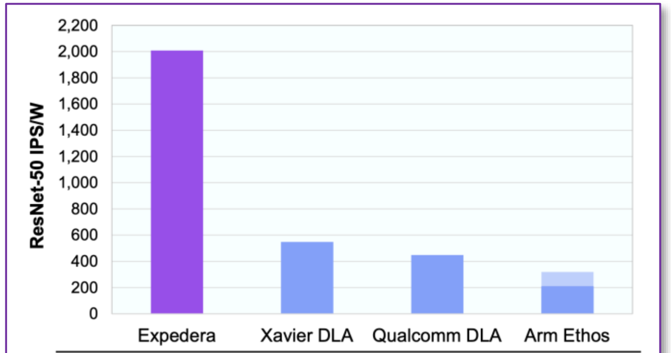


Figure 2. DLA-IP performance comparison. Expedera's performance and efficiency stand out against competitors'. The chart shows ResNet-50 v1.0 images per second (IPS) at best batch size versus typical power (W). (Source: vendors and The Linley Group estimates)

"Expedera Redefines AI Acceleration for the Edge"
- Microprocessor Report, April 2021

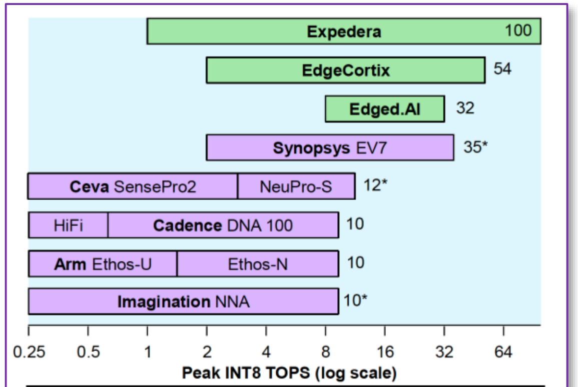


Figure 2. DLA-IP performance. Relative to established IP vendors, three startups pushed the upper limits of single-core performance, measured in trillions of 8-bit integer operations per second (INT8 TOPS). *Greater performance available in multicore configurations. (Data source: vendors)

"CPU-IP Vendors Chase High-end Wins"
- Microprocessor Report, January 2022

Competitive Benchmarks

6-10X

Maximum single-core performance (TOPS), versus IP from Cadence, CEVA, & Imagination

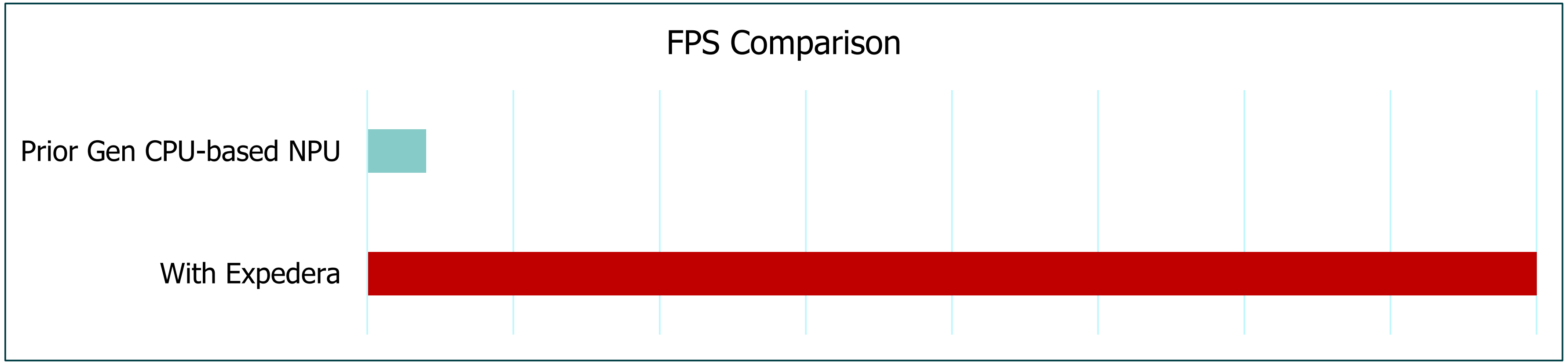
6X

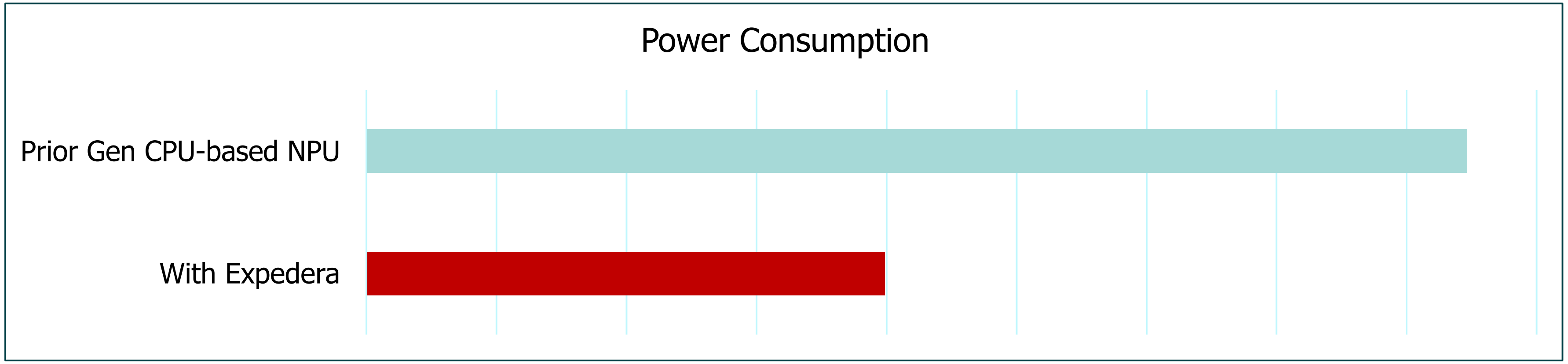
IPS Performance per Watt, versus ARM Ethos

2.7X

TOPS Performance per silicon area, versus Apple A13

Real-World Customer Measured Results





- Data from worldwide top 5 OEM, device available for purchase
- Expedera Origin deployed for 4K video low light denoising
- 18 TOPS total capacity

20X improved FPS while consuming less than half the power over prior CPU-based NPU

Expedera Resources

Contact Us

Within the Alliance
https://www.edge-ai-vision.com/companies/expedera/

Website, including technical briefs
https://www.expedera.com/products-overview/

Email
info@expedera.com

Social: @ExpederaInc

/ExpederaInc

/Company/Expedera/

2022 Embedded Vision Summit

Visit us at booth #320