NPU IP Hardware Shaped Through Software and Use-Case Analysis

Yair Siegel

Senior Director Wireless and Emerging Markets

Ceva

Company Overview

Trusted partner for over 2 decades
>19bn Ceva-powered devices shipped
>2bn Ceva-powered chips shipped annually

#1 worldwide wireless connectivity IP,
67% market share*
Edge AI NPUs portfolio, scalable from
Embedded ML up to GenAI

NASDAQ:CEVA
>$100m annual revenue
$164m cash, no debt

40-50 licensing deals annually
>70 royalty paying customers
>100 active customers

>200 registered patents

~450 employees (~75% R&D)
HQ in Maryland, main R&D Centres:
U.S., France, Israel, Greece, Serbia

***Source: IPNest's latest Design IP report – 2023 (published May 2024)**

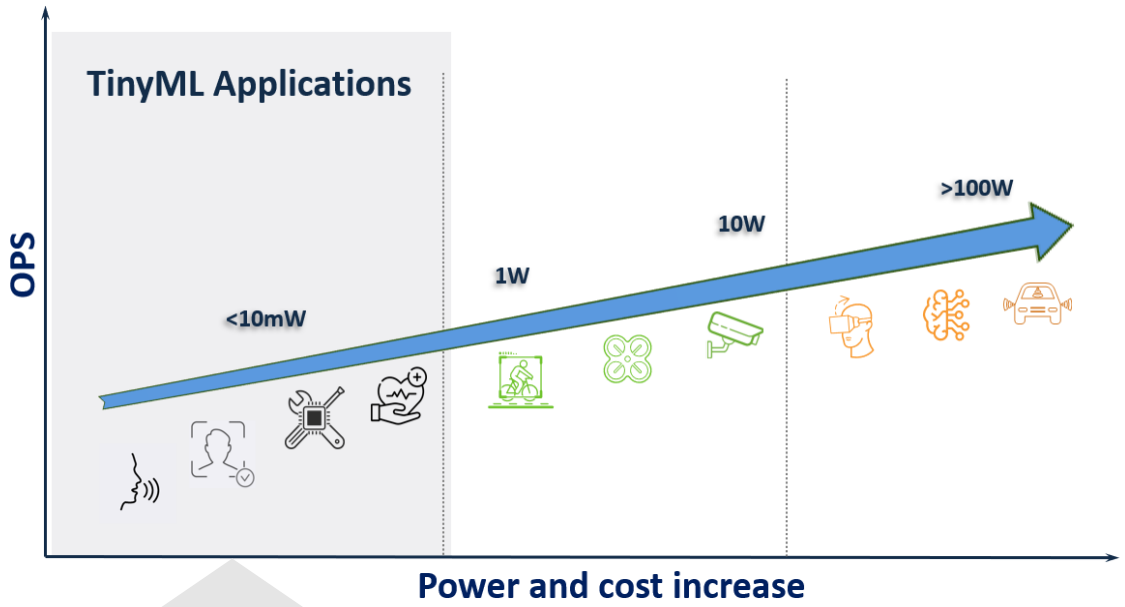
Embedded ML Applications for Consumer & Industrial IoT

Voice: keyword spotting (KWS), voice biometrics, sound detection & classification, environmental noise cancellation (ENC)

Vision: object detection, image classification, always-on human presence detection or similar contactless recognition

Predictive Maintenance: vibration, temperature, humidity, sound sensing for predictive maintenance

Health & Fitness Sensing: physical activity tracking, heart rate monitoring, sleep pattern analysis



Embedded ML applications typically consume <1 Watt and support 10's of GOPs of compute

End-User Devices:

Typical Technical Requirements for Embedded ML Deployment

Memory Footprint

- < 10 MB Flash/ROM/RAM size
- < 500 KB code + dynamic data memories

Power Consumption

- Optimized for Low Power < 10 mW
- Enable battery-powered devices
- Minimize device recharges

Model Size

- 0.01 MB to 10 MB memory required for the model weights (aka parameters)

Computational Requirements

- Typical computation: 10s of GOPS or more
- Deployable on resource constrained hardware (e.g., MCU)

Key requirement: Easily deployable on battery powered and resource limited devices, to reduce deployment costs and maximize value of Edge AI

Embedded ML Implementation Challenges

Key Challenges

Rapid Technology Evolution
New use cases, networks and data types

Low-Cost Expectations
Small memory size & die-size needed for proliferation

Ultra Low Power Requirements
Always-on, battery powered devices

Complex Software Infrastructure
AI frameworks, proprietary silicon, and varied networks

Existing Solutions

Full Hardwired NPUs
Can't cope well with new networks or data types

Made for very specific tasks with no upgrade path

MCUs or DSPs plus separate NPU
Multi-core solution yields sub-optimal area & cost

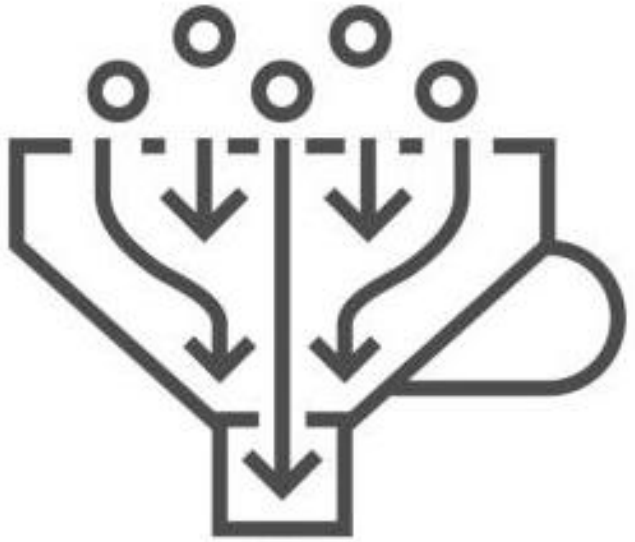
MCUs / DSPs not ML optimized ->
poor in power consumption and performance

Complex integration, SW, memory management

Embedded ML solutions require flexible and scalable architectures delivering optimal balance of performance, size, & power efficiency, with a complete AI SDK

Ceva-NeuPro-Nano Embedded ML NPU: Design Guidelines

- Shaped by deep analysis of user perspectives, recognizing need for **powerful** and **user-friendly** solution
- Design philosophy guided by **application-level** challenges, vs. neural-net layer level challenges
- Approach ensures 3 major workloads can be handled efficiently and seamlessly:
  - Neural network workloads
  - DSP workloads
  - Control workloads



Ceva-NeuPro-Nano: Software First Approach, Designed to Address Embedded ML Market Needs

1 Single Core

- End-to-end AI application on single core
  - Efficiently executes various NN architectures, operators, feature extraction, DSP and control code
- DSP built into NPU architecture
  - NeuPro-Nano not a HW accelerator DSP+HWA

2 Efficient & Flexible Compute

- Mixed math precision: supports 4/8/16/32-bit integer as well as floating point math
- Transformers & SLMs inherent support
- Weight de-compression on-the-fly
- Sparsity compression

3 Programmable & Extensible

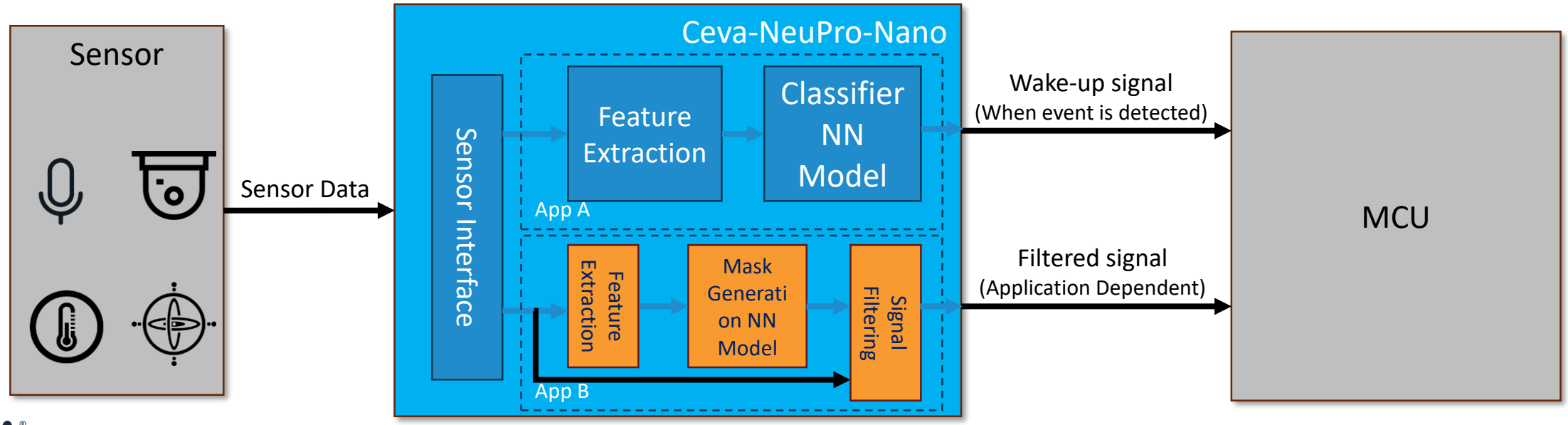
- Enables support for new AI frameworks, NN architectures and operators
- Supports future new DSP and NN enabled applications
- All in software → no hardware changes

4 Fast TTM to Deploy

- Robust AI SDK solution, Ready for 3rd party IDE/SDK integration
- No-friction business model for deployment
- Strong ecosystem of AI software and development companies

Complete End-to-End AI Application on a Single Core

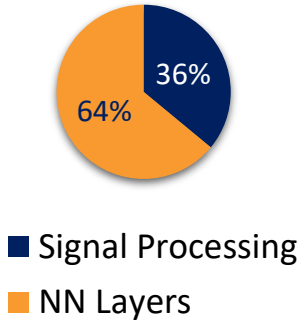
- Typical Embedded ML applications constructed from **feature extraction & NN layers**
  - Each block consumes substantial resources
- **Single core Edge NPU for complete Embedded ML applications**
  - Handles control code, **NN layers** and **feature extraction (Signal Processing - MFCC)** on same processor



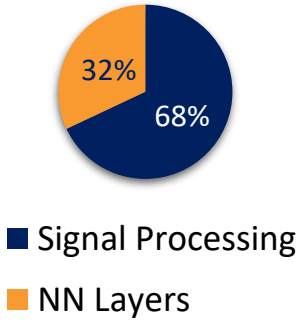
Complete AI Application on a Single Core

- Architecture minimizes die size and memory utilization by efficiently processing all application workloads on a single core
  - NN layers include Fully Connected, RNN, Attention
  - Signal Processing: pre & post-processing (e.g., vision networks), feature extraction (STFT, iSTFT and MFCC)

Ceva-ClearVox Control (cycles partition)



Ceva-ClearVox ENC (cycles partition)



Ceva's complete AI based applications:

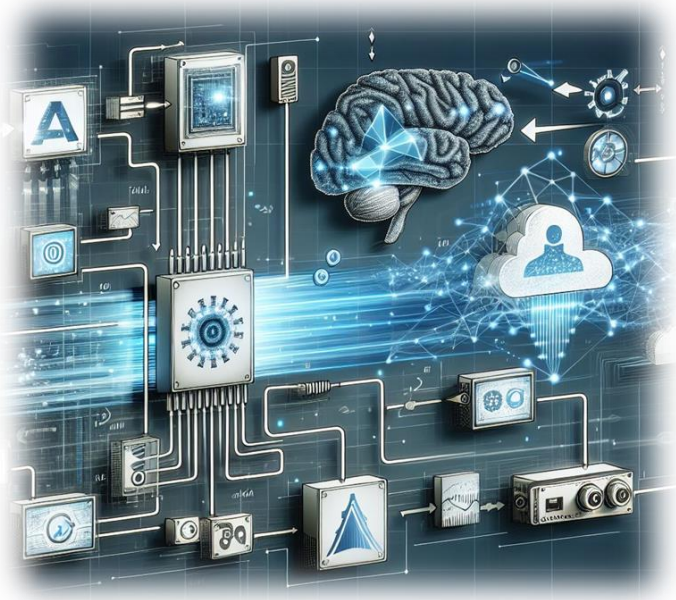
- **Ceva-ClearVox™ Control -** Wake Word and Commands (Amazon AVS qualified)
- **Ceva-ClearVox™ ENC** - Environmental Noise Cancellation for crisp calls in any conditions

Single core, future compatible NPU ensures high efficiency on NN layers and Feature Extraction workloads

Software First Approach Principles

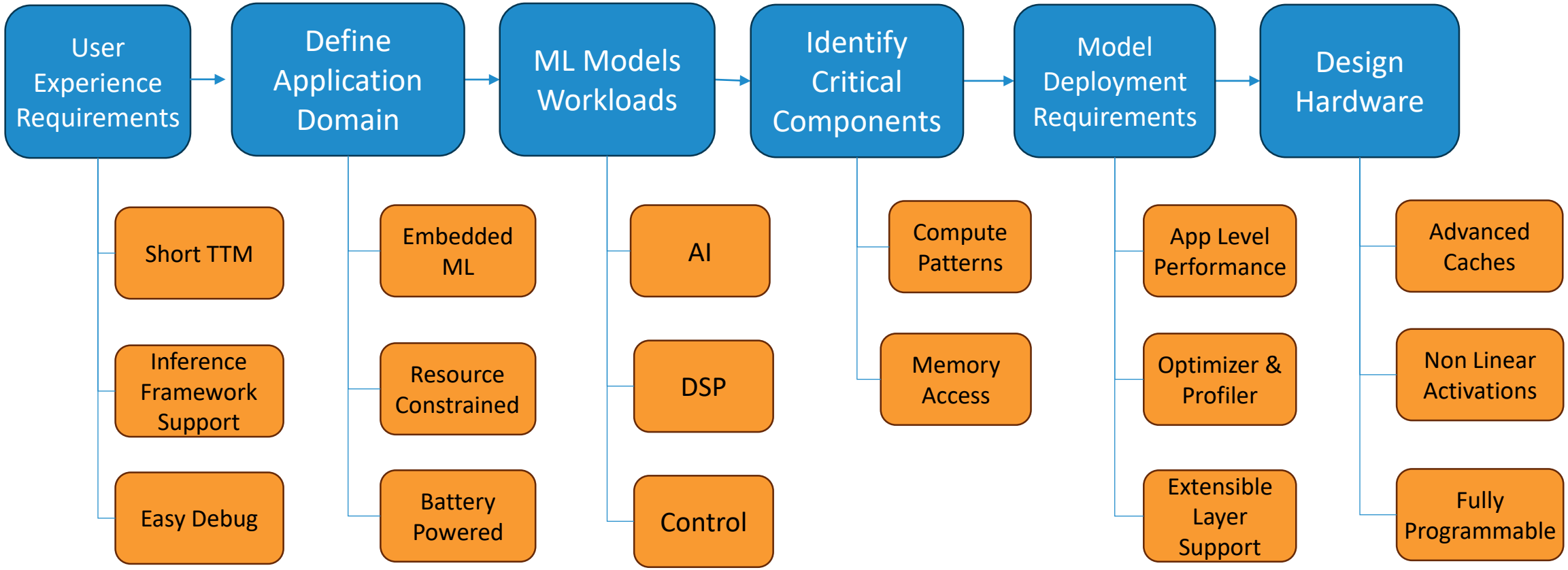
Main principles followed:

1. Software requirements drive hardware architecture
2. Prioritize hardware flexibility and programmability
3. Design for end-to-end system efficiency
   - Prioritize application-level performance
   - Consider data transfers and memory hierarchies



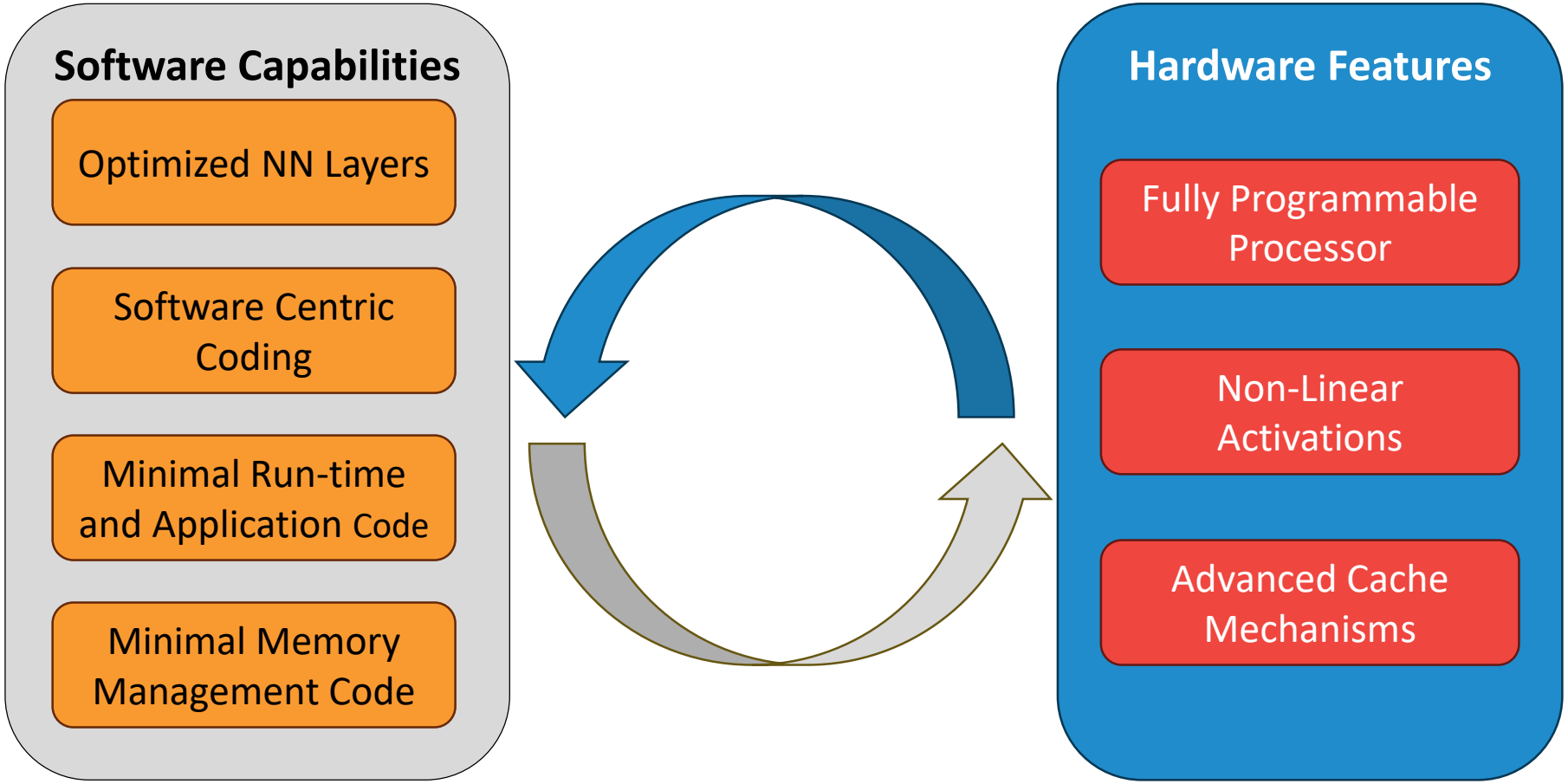
Software First Approach Requirements

Driving hardware architecture decisions through software requirements



Software First Approach

Prioritize hardware flexibility and programmability over pure performance



Application-Level vs Layer-Level Optimization

Application-Level (typical for processors)

- Minimize **total application compute**
- Control and DSP workflows are major compute consumers, handled **within** NPU
- Easy & efficient new operators support via software
- Unsupported operators not a bottleneck

Layer-Level (typical for accelerators)

- Minimize **layer level compute**
- Control and DSP workflows are major compute consumers, handled **outside** NPU (adding latency)
- New operators support requires **hardware modification** or executed by **MCU offloading**
- Unsupported operators become a bottleneck (MCU may incur severe compute penalty and memory transfers latency)

Ceva-NeuPro Studio Overview

Comprehensive AI SDK uniquely accelerating OEM and semiconductor ML product design & deployment

EMBEDDED VISION SUMMIT

Interfaces Leading Industry AI Tools

- BYOM
- MATLAB connection
- Edge Impulse & NVIDIA TAO

Model Optimizations & Deployment

- Graph compiler & runtime
- AI model & app profiling
- Model optimization & services
- Bring your own operator

Complete Development Tools

- Eclipse / Visual Studio IDE
- AI Model Viewer - Netron
- C/C++ toolchain
- Extendable tools

Pre-Optimized Models

- Model zoo
- Ecosystem optimized models
- NN libraries
- Domain specific libs & algos

Ceva-NeuPro Studio AI Model Deployment Flow

EMBEDDED VISION SUMMIT

NN Model/s → **AI Application**

Model Zoo

BYOM

PyTorch · ONNX · TensorFlow · Keras

NeuPro Studio

C/C++ User Code

Profile

Arch Planner

Model/s

Compile

C/C++ Compiler

NN Graph Compiler

Runtime Libs

Ceva Optimized Libraries (DSP, NN, domain specific)

LiteRT Micro

Application

Infer

Ceva-NeuPro™

Execute

Debug

Profile

Data Set **HW Config**

Summary

- Hardware design balancing power, performance, and ease of use achieved through deep internalization of software requirements:
  - Real world applications and use cases
  - Emerging technologies and trends
  - Programmer pain points

Ceva-NeuPro-Nano: Software First NPU Design

Resources

Ceva-NeuPro-Nano Information - https://www.ceva-ip.com/product/ceva-neupro-nano/

Ceva-NeuPro Studio SDK - https://www.ceva-ip.com/product/ceva-neupro-studio/

Tech Insights, Microprocessor Report by Dylan McGrath - https://www.ceva-ip.com/wp-content/uploads/Ceva-NPU-Core-Targets-TinyML-Workloads.pdf

Google LiteRT for Microcontrollers - https://ai.google.dev/edge/litert/microcontrollers/

Edge AI Foundation - https://www.edgeaifoundation.org/

Alliance Dev Tools - https://www.edge-ai-vision.com/resources/technologies/development-tools/

Ceva-NeuPro-Nano NPU Already Won Industry Awards

2024 IoT Edge Computing Excellence Award

The Best IP/ Processor of the Year 2024 award at the prestigious EE Awards Asia event

Q&A

Thank You!

For more info: yairs@ceva-ip.com