NPU IP Hardware Shaped Through Software Insights and Use-Case Analysis

Ido Gus
Deep Learning Senior Team Leader, Ceva

November 2025

Company Overview

Trusted partner for over 2 decades
>19bn Ceva-powered devices shipped to date; >1.6bn annually

#1 worldwide in wireless connectivity IP, with 67% market share*

Edge AI focus with scalable NPUs for Embedded ML up to Gen-AI

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)

OUR MISSION

The partner of choice for transformative IP solutions for the Smart Edge

Typical Technical Requirements for Embedded ML Deployment

Memory Footprint

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

Power Consumption

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

Model Size

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

Computational Requirements

- Minimal computational resources for inference tasks >10GOPs
- 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

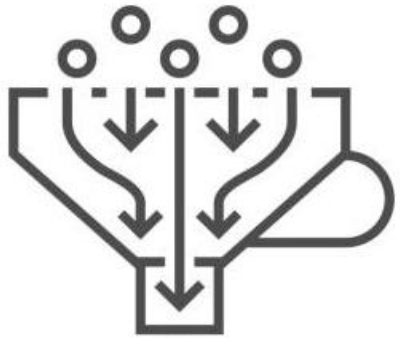
MCUs / DSPs not ML optimized ->
poor in power consumption and performance

Complex integration, SW, memory management

Embedded ML solutions require a flexible and scalable architecture that delivers the optimal balance of performance, size, & power efficiency together with a complete AI SDK

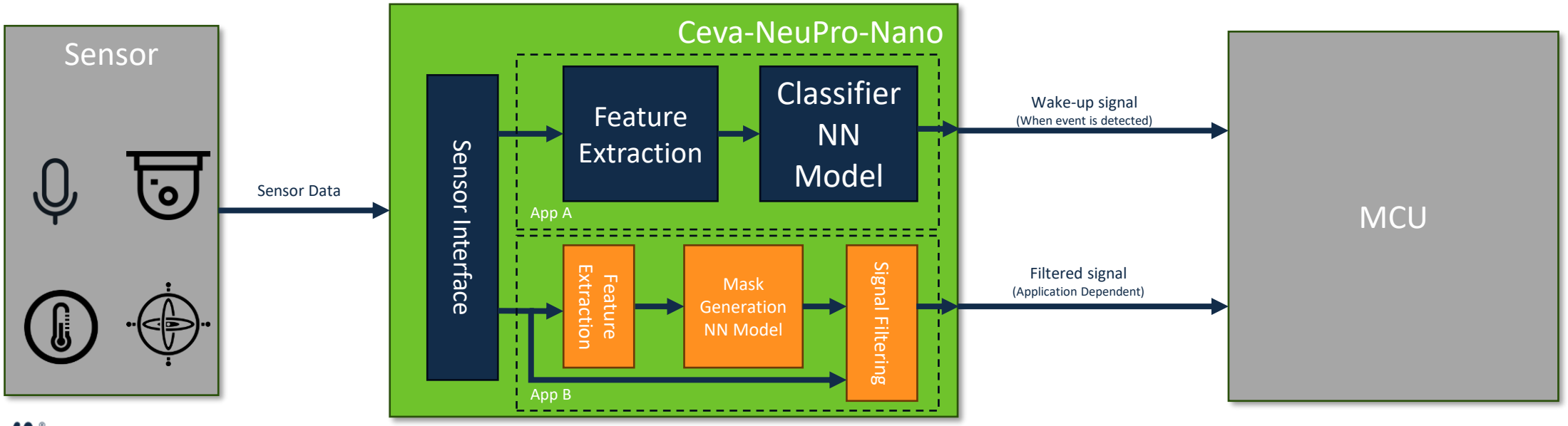
Ceva-NeuPro-Nano Embedded ML NPU: Design Guidelines

- Guidelines were shaped by a deep analysis of user perspectives, recognizing the need for a solution that is both **powerful** and **user-friendly**.
- The design philosophy was guided by focusing on **application-level** challenges rather than on the neural-net layer level challenges.
- The approach ensures that the 3 major workloads can be handled efficiently and seamlessly:
  - Neural network workloads
  - DSP workloads
  - Control workloads



Complete End 2 End AI Application

- 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

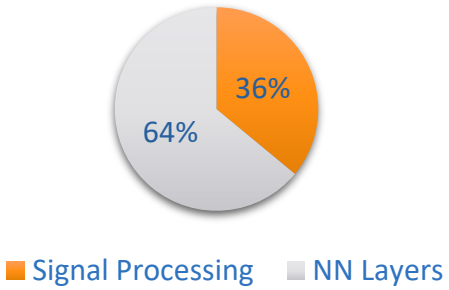


Complete AI Application on a Single Core (Examples)

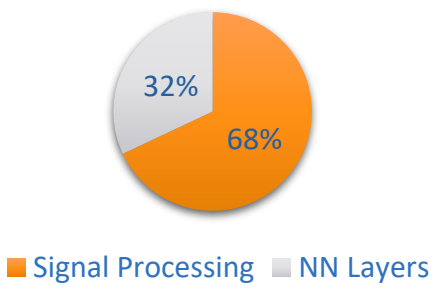
Ceva in-house complete AI based applications:

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

Ceva-ClearVox Control



Ceva-ClearVox ENC



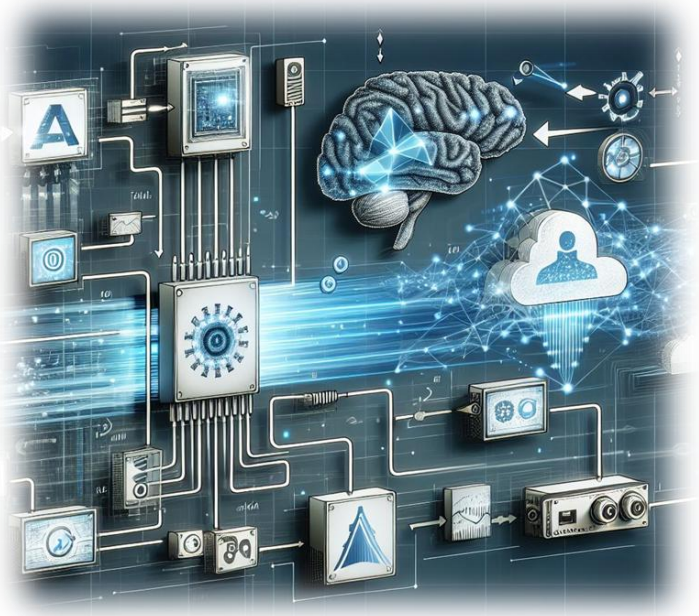
- NN layers include Fully Connected, RNN, Attention
- Feature extraction signal processing include: STFT, iSTFT and Mel Filter Banks (MFCC)

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

Software Centric Approach

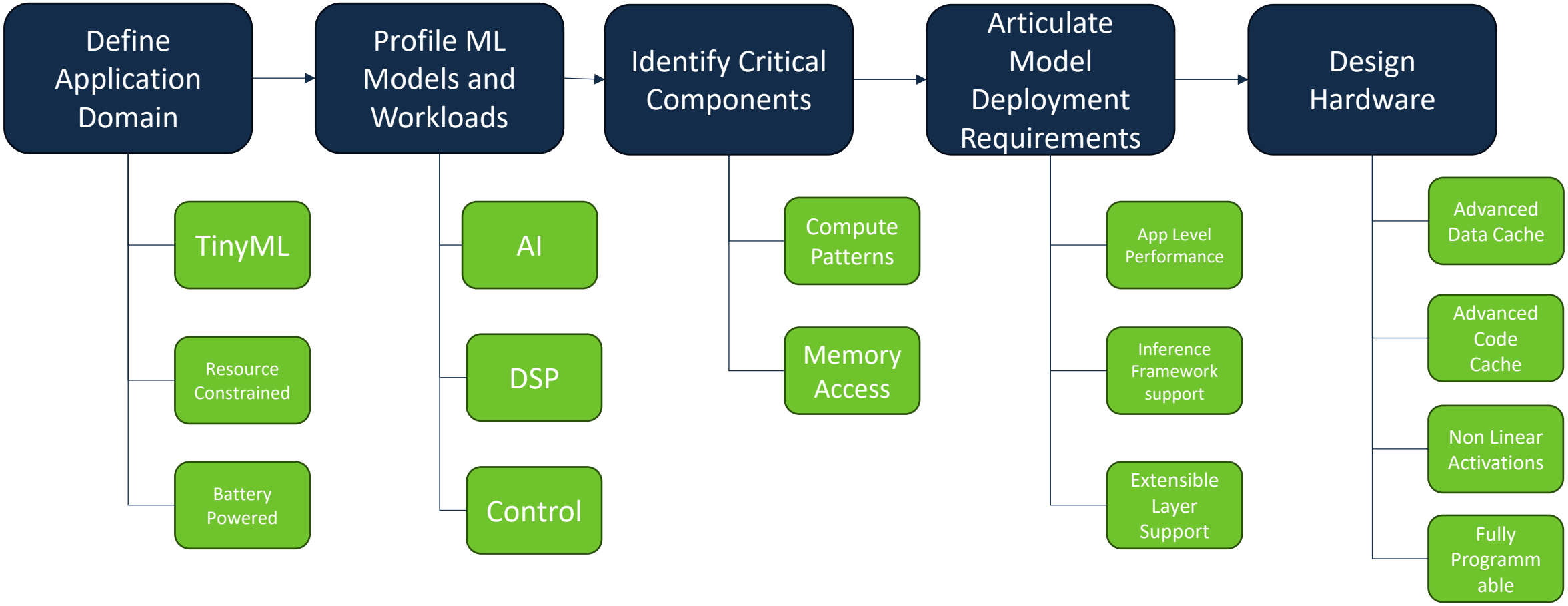
Three main principles were followed:

1. Software requirements drive hardware architecture decisions
2. Prioritize hardware flexibility and programmability over pure performance
3. Prioritize application-level performance over layer-level performance



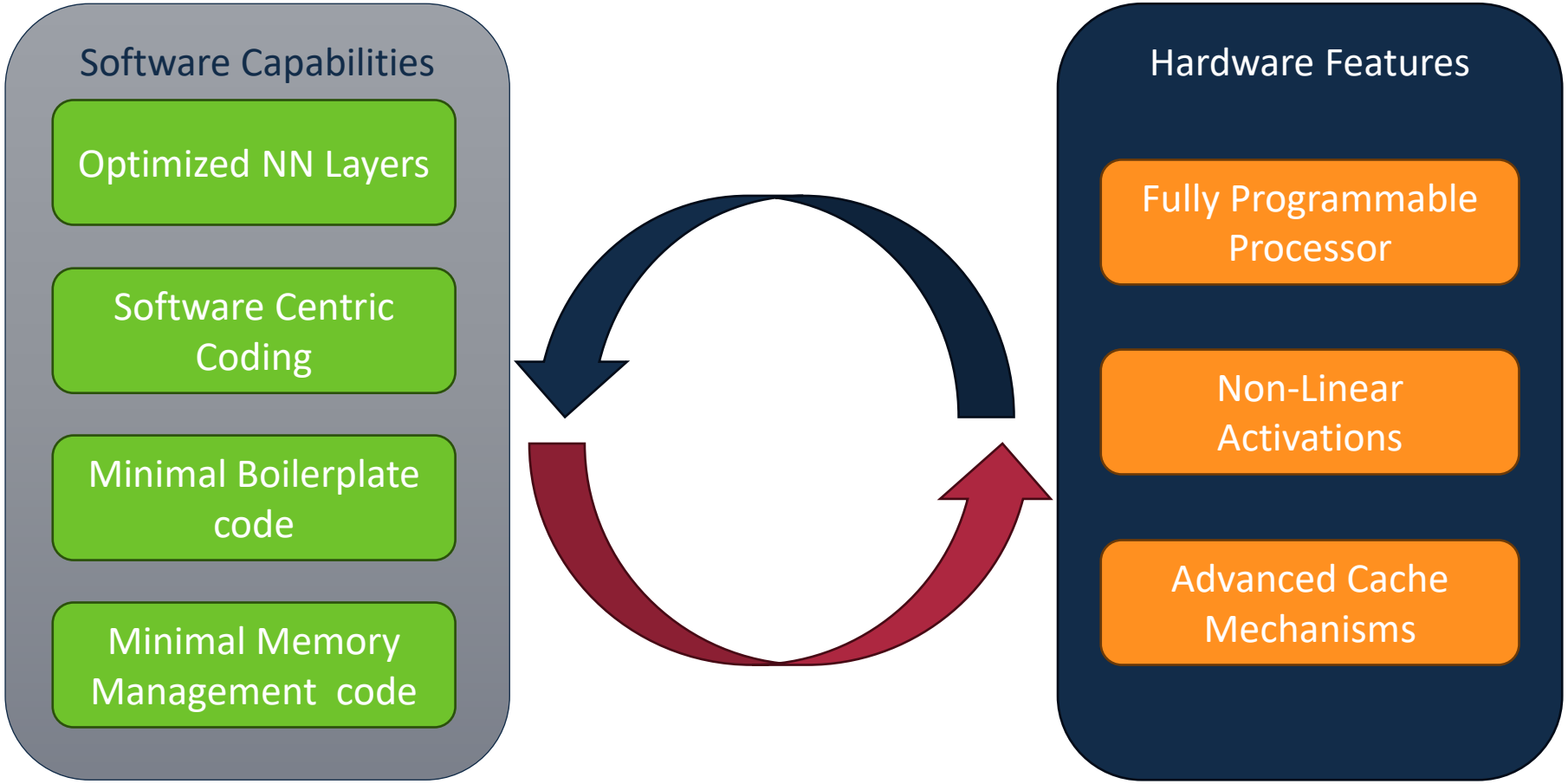
Software Centric Approach

- Driving hardware architecture decisions through software requirements



Software Centric Approach

Prioritize hardware flexibility and programmability over pure performance



Software Centric Approach

- Prioritize application-level performance over layer-level performance
  - Efficient execution of diverse workloads (Control, DSP, AI)
  - Support seamless integration with existing software frameworks and toolchains
  - Design for end-to-end system efficiency, including data transfers and memory hierarchies

Application-Level vs Layer-Level Optimization

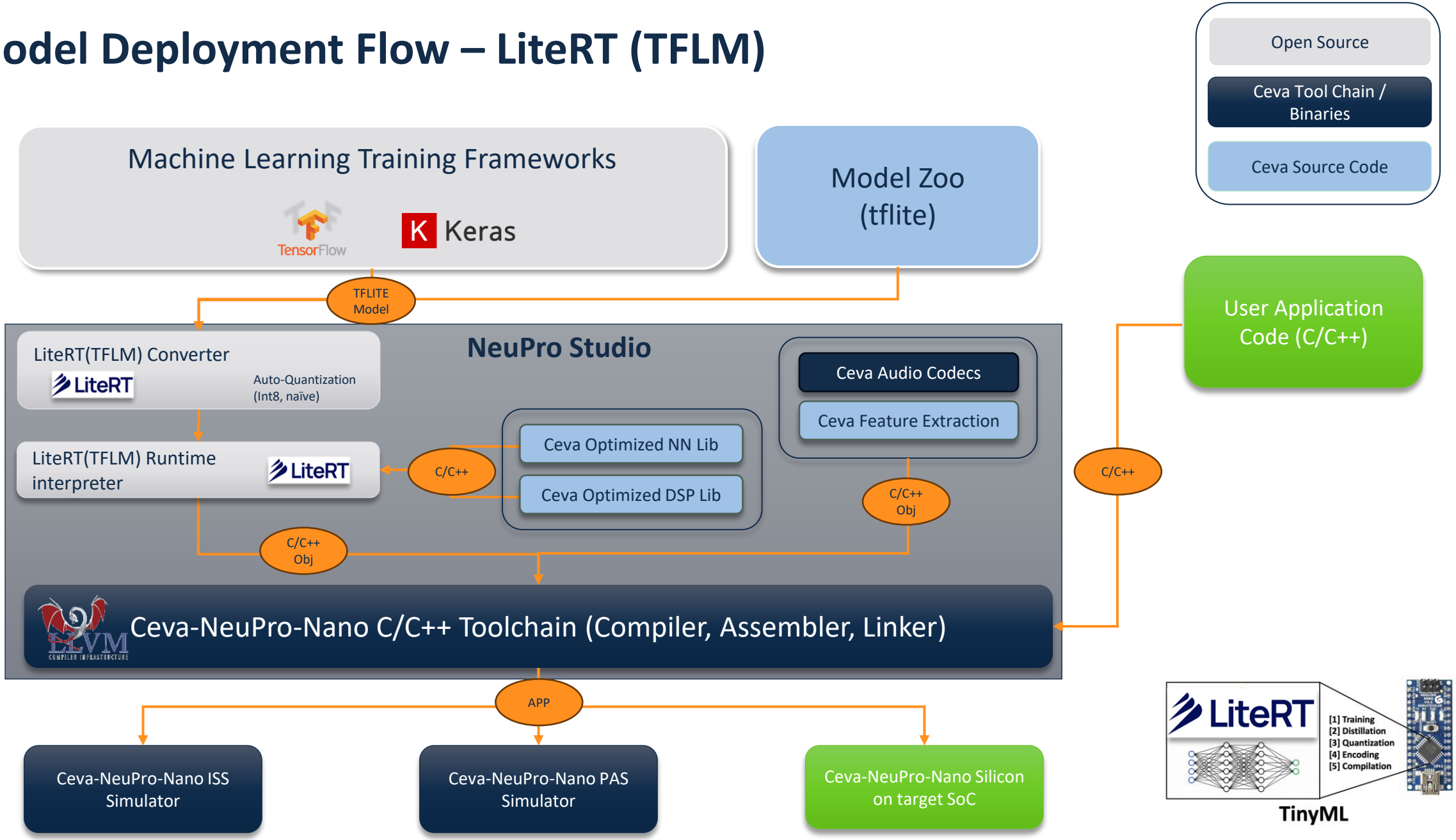
Application-Level (typical for processor)

- Minimize **total application compute**
- Control and DSP workflows are major compute consumers, handled **within** the NPU
- Add support for new operators through software
- Unsupported operators do not become a compute bottleneck

Layer-Level (typical for Accelerator)

- Minimize **layer level compute**
- Control and DSP workflows are major compute consumers, handled **outside** of the NPU
- New operator support require **hardware modification** or through **MCU offloading**
- Unsupported operators become a compute bottleneck (MCU may incur a severe compute penalty)

NN Model Deployment Flow – LiteRT (TFLM)



Ceva-NeuPro Studio Overview

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

Interfaces Leading Industry AI Tools

BYOM - various model formats
- Via TVM/uTVM/LiteRT Micro
- Optimized backends

MATLAB connection – combine AI w/Signal Processing, co-debug

Edge Impulse platform - Collect data, design & deploy NN models

Model Development & Deployment

Graph Compiler & Runtime Inference
- Utilize TVM/uTVM/LiteRT Micro

AI Model to complete application

AI model & application profiling

Model optimization

Complete Dev. Tools

Eclipse / VS Code IDE
Integrating all components, AI & C/C++

AI Model Viewer - Netron

C/C++ toolchain: LLVM Compiler, Simulation & Emulation Debugger

Extendible: Connect CPU / external HWA

Pre-Optimized Software

Model Zoo: optimized, ready-for-use models

NN Libraries: Optimized operators, CMSIS-NN compatible

Domain specific libraries & algorithms (e.g. DSP Libs, Spatial Audio, ENC)

DSP Libraries

Value

- Optimized functions for digital signal processing (DSP) crucial for efficient signal processing tasks
- Seamlessly works with Ceva-NeuPro processors
- Integrated into Ceva-NeuPro Studio SDK, ensuring compatibility and ease of use
- Updated periodically - increasing contents and improving performance

Content

- Main library functions
  - Key filters: FIR, IIR, FFT
  - Math operations (Div, Sqrt, Log, Power)
  - Trig operations (Cos, Sin, Tan, …)
  - Vector operations (Vecadd, Vecdot,…)
  - Matrix operations (Mat Conj, Mat trans,…)
- Fixed-point & Floating-point implementations
- Supports popular data types & algo methods
- Source code for developer adaptation

Seamlessly combine Signal Processing and AI workflows for optimal performance

Neural Network Libraries

- Provide optimized runtime libraries and off-the-shelf application-specific software
- Highly efficient, work seamlessly with inference frameworks (e.g. LiteRT Micro, uTVM)
- Designed to handle memory-intensive tasks efficiently to improve performance
  - Direct processing of compressed model weights
  - Advanced data cache system
- Integrated into Ceva-NeuPro-Studio, ensuring compatibility and ease of use
- Updated periodically with new capabilities and improved performance

Accelerate AI Model Development

Summary

- Hardware design balancing power, performance, and ease of use can be achieved through deep internalization of software requirements:
  - Real world applications and use cases
  - Emerging technologies and trends
  - Programmer pain points
- Ceva-NeuPro-Nano is a prime product of a software centric NPU Design

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

Ido Gus

Deep Learning Senior Team Leader, AIDIV

Ido.Gus@ceva-ip.com

www.ceva-ip.com