cadence

TECHNICAL BRIEF

# Cadence AI IP Platform

## Comprehensive IP platforms for edge to on-device AI

With the growing demand for AI-based tasks across various applications and vertical segments, the need for efficient on-device and edge AI processing is becoming increasingly critical. The diverse computational and power requirements of these AI tasks make it challenging for SoC architects to create solutions that meet market expectations.

At the heart of these requirements are power, performance, and area. The Cadence AI IP platform is designed to optimize each of these parameters, while allowing architects and designers the flexibility to make IP configuration decisions that can be optimized for a particular design target. By allowing configuration options across mac array sizes, data types, clock frequency, bandwidth, interface widths, memory sizes and more, Cadence AI IPs can enable a nearly endless combination of different PPA envelopes to meet almost every need.

The Cadence® AI IP portfolio offers a comprehensive solution of both IP and Software, enabling SoC developers to design and deliver high-performance, power-efficient solutions tailored to their specific KPIs and requirements.

# Neo NPU

The Cadence Neo NPU is Cadence's flagship advanced Neural Processing Unit (NPU) that offers energy-efficient and high-performance AI processing. The Neo NPUs target a wide variety of applications, including sensor, audio, voice/speech, vision, radar, and more. The

comprehensive performance range makes the Neo NPUs well-suited across ultra-power-sensitive applications such as IOT and hearables/wearables, up to high-performance systems in AR/VR, automotive, and more.

The product architecture natively supports the processing required for many network topologies and operators, allowing for a complete or near-complete offload from the host processor for both classic and generative AI networks. Depending on the application's needs, the host processor can be an application processor, a general-purpose MCU, or a Tensilica processor.

The Neo NPUs provide performance scalability with a range from 256 MACs up to 32k MACs (8x8-bit MACs per cycle) with a single core, suiting an extensive range of processing needs. Capacity configurations are available in power-2 increments, allowing for the right sizing in an SoC for target applications. Int4, Int8, Int16, and FP16 are all natively supported data types across a wide set of operations that form the basis of CNN, RNN, and Transformer-based networks, with mixed precision supported by the hardware and associated software tools allowing for the best performance and accuracy tradeoffs.

Additional features of the Neo NPUs include compression/decompression of weights (or coefficients) to minimize system memory space and bandwidth consumption for a network and energy-optimized compute hardware that leverages network sparsity for optimal performance and power profile.

The Neo NPUs have a target clock frequency of 1.25GHz in 7nm TSMC, yielding up to 80 TOPS performance in a single core. Customers can target higher or lower clock frequencies for specific product needs. Neo NPUs can also be configured in a many-core or multi-core fashion to scale beyond 80 TOPS to address advanced generative AI, large language models, and transformer-based state-of-the-art machine learning models.

| | **Neo NPU** |
|---|---|
| Tera Operations per Second (@ 1.25GHz) | 0.64 to 80+ |
| Decompression (Weights) | Y |
| Compression/Decompression (Activations) | Y |

| | | Neo NPU |
|---|---|---|
| Data Types | | Int4, Int8, Int16, FP16 |
| # AXI | | 3 (1 or 2 Initiator and 1 Responder) |
| AXI Width | | 128/256/512-bit |
| MAC for AI | 8x8 | 256 to 32k |
| | 8x16 | 128 to 16k |
| | 16x16 | 64 to 8k |
| | FP16 | 64 to 8k |

Table 1: Features of the Neo NPU

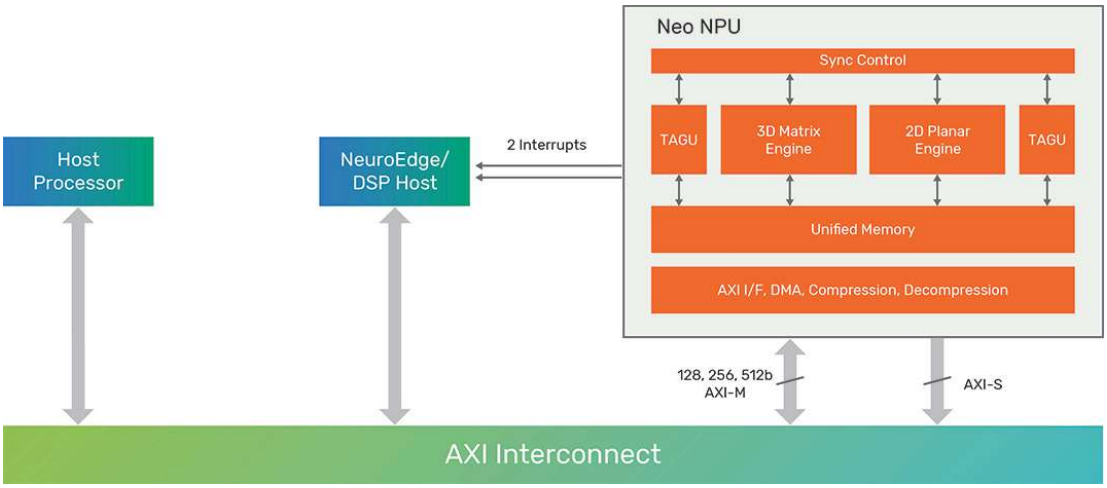


Figure 1: Product architecture of the NEO NPU

# Tensilica NeuroEdge AI Co-Processor

The NeuroEdge AI Co-Processor (AICP) is a new class of processor in the Tensilica family. This innovative processor is specifically designed to be paired with an NPU to create a robust AI subsystem, enabling users to efficiently handle AI workloads of today and tomorrow. The NeuroEdge AICP's versatility allows for seamless integration with Cadence's own Neo NPUs, a customer's homegrown NPU, or even third-party NPU IP, making it an attractive solution for a wide range of applications.

Since the latest AI networks of transformers and GenAI require more than just MAC operations, comprehensive layer and kernel support is needed for end-to-end AI model execution. The NeuroEdge AICP handles this task of executing layers and operations that are not suited for the NPU, e.g., sigmoid, tanh, relu, eltwise, non-linear operations, to name a few, or operations/layers that require proprietary implementations. By working in tandem with the NPU, the NeuroEdge AICP enables a flexible yet robust AI subsystem.

Built on the mature Xtensa architecture, the NeuroEdge AICP is based on VLIW and SIMD architectures, complete with instructions that are optimized to handle AI workloads with
`e. Furthermore, the NeuroEdge AICP offers a wide range of configuration options, allowing

it to strike the perfect balance between area and performance outperforming CPU and GPU-based options. This flexibility makes the NeuroEdge AICP an ideal choice for developers seeking to create high-performance yet efficient AI subsystems.

| | | **NeuroEdge 130** |
|---|---|---|
| MAX SIMD Width | | 512 |
| AI Enhancements | | Neural Network Quantization Acceleration<br>Non-Linear Operator Acceleration |
| MAC for AI | 8x8 | 512 |
| | 8x16 | 128 |
| | 16x16 | 128 |
| | FP16 | 64 |

Table 2: Features of the Tensilica NeuroEdge AI Co-Processor

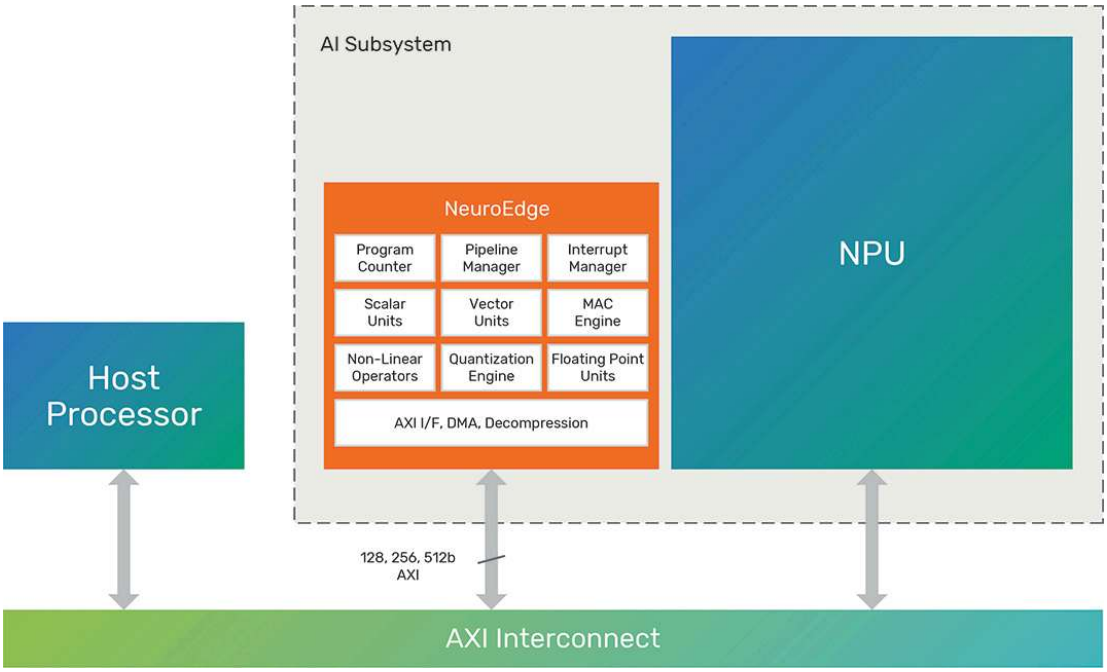


Figure 2: Product architecture for the Tensilica NeuroEdge AI Co-Processor

# Tensilica DSPs

Tensilica DSPs are capable of efficiently executing AI workloads in addition to traditional signal processing. The Tensilica DSP portfolio offers a wide range of performance, from 8GOPS to up to 2TOPS of AI performance, and includes the popular Tensilica HiFi DSPs for audio/voice and Vision DSPs for imaging/computer vision.

The DSPs are based on the VLIW and SIMD architectures, with instruction sets optimized for specific domains. Tensilica AI platform customers benefit from the domain-specificity, extensibility, and configurability they have come to expect from the trusted, mature Tensilica DSP architecture. SoC designers can extend the base architecture to address specific workload requirements and create differentiation.

These DSPs also offer a scalable multiplier accumulator (MAC) block that can run custom AI workloads efficiently, as well as an optimized NN library and comprehensive software support.

| | | Vision 110 | Vision 130 | Vision 230 | Vision 331 | Vision 240 | Vision 341 |
|---|---|---|---|---|---|---|---|
| MAX SIMD Width | | 128 | 512 | 512 | 512 | 1024 | 1024 |
| Xtensa Processor | | LX | LX | NX | NX | NX | NX |
| MAC for AI | 8x8 | 128 | 256 | 512 | 512 | 1024 | 1024 |
| | 8x16 | 32 | 128 | 128 | 128 | 256 | 256 |
| | 16x16 | 32 | 64 | 128 | 128 | 256 | 256 |
| | 32x32 | 2 | 8 | 16 | 16 | 32 | 32 |

Table 3: Tensilica Vision DSPs

| | HiFi1s | HiFi3z | HiFi4 | HiFi5s |
|---|---|---|---|---|
| MAX SIMD Width | 64 | 64 | 64 | 128 |
| Xtensa Processor | LX | LX | LX | LX |

| | | HiFi1s | HiFi3z | HiFi4 | HiFi5s |
|---|---|---|---|---|---|
| MAC for AI | 8x8 | 8 | - | - | 32 |
| | 8x16 | 8 | - | - | 32 |
| | 16x16 | 4 | 8 | 8 | 16 |
| | 32x16 | 2 | 4 | 8 | 16 |
| | 32x32 | 1 | 2 | 4 | 8 |

Table 4: Tensilica HiFi DSPs

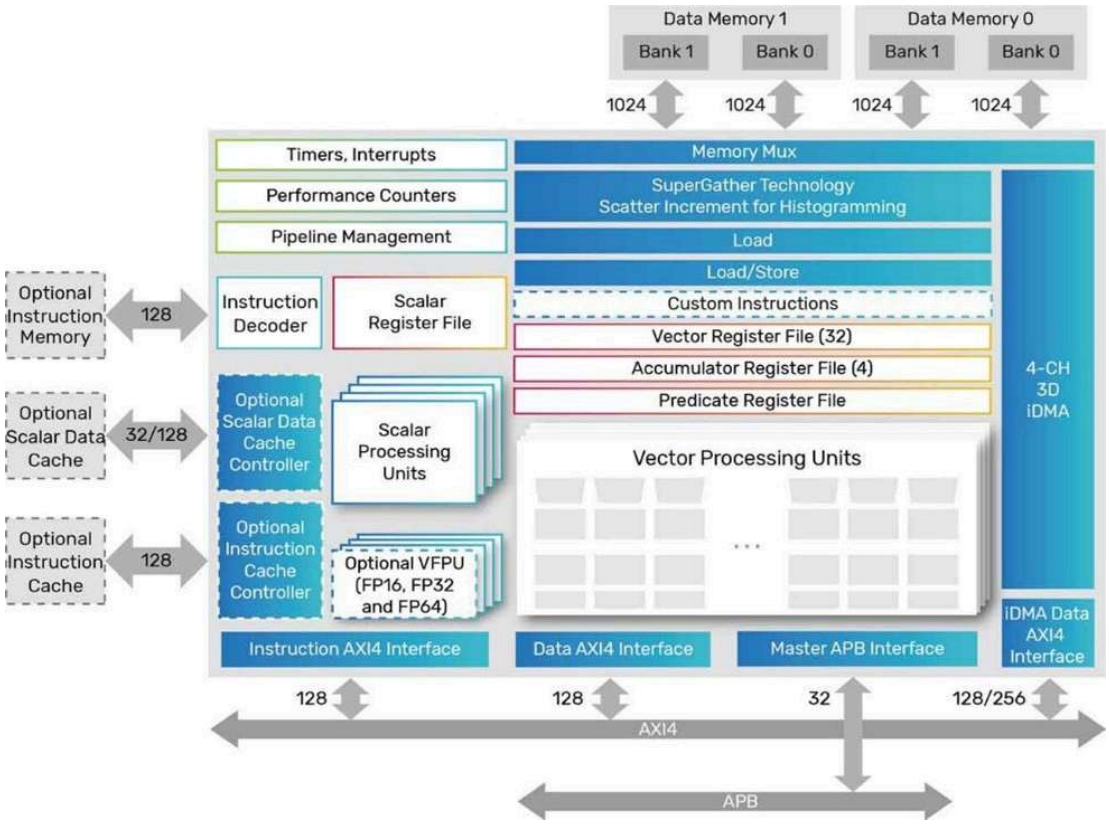


Figure 3: Product architecture for Tensilica DSPs

# NeuroWeave SDK

The Cadence NeuroWeave Software Development Kit (SDK), a single SDK used across all of Cadence's AI IPs, is what sets the Cadence solution apart. Leveraging the TVM stack, the NeuroWeave SDK is easy to use and allows architects to tune, optimize, and deploy their AI models on Cadence's AI IPs. The NeuroWeave SDK is further capable of quantizing floating point models to fixed point, pruning, and compressing convolution weights to reduce computation, memory traffic, and storage.

The NeuroWeave SDK has two major modes of operation:

- Compiler mode using Xtensa Neural Network Compiler (XNNC)
- Interpreter mode using TensorFlow Lite for Microcontrollers (TFLM)

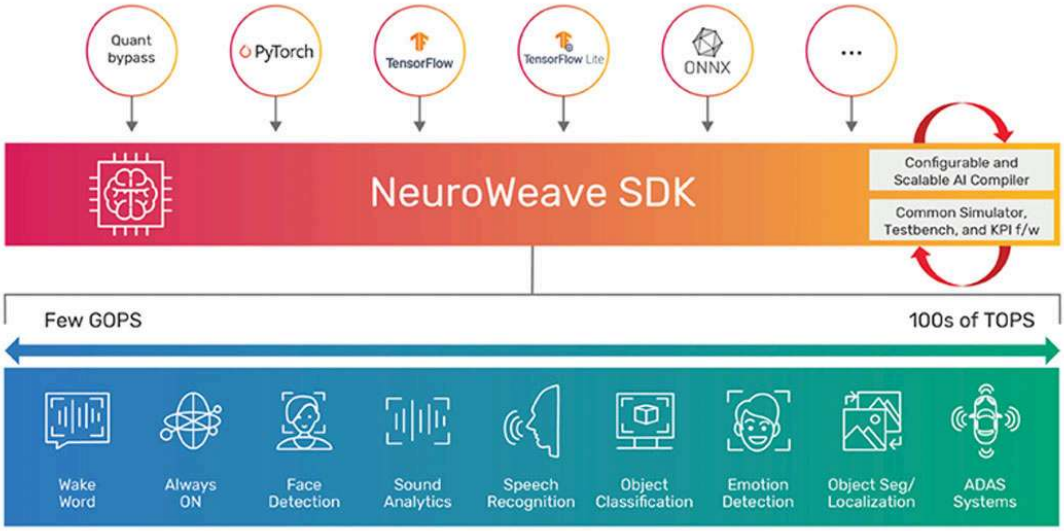


Figure 4: One AI Software Compiler Toolchain

# Compiler mode (XNNC)

- Includes ahead-of-time compiler toolchain, including TVM network importing
- Can perform light pruning and compression to reduce computation, memory traffic, and storage needs
- Can be used as command line tools and as a GUI that can be easily invoked in a customer's software pipeline flow
- Programmable and extensible framework adjusts to the target IP solution
- Is flexible and future-proof, as compute layers change rapidly over time
- Supports mixed-precision data formats (4-bit/8-bit/16-bit integer and 16-bit floating point)
- Provides concurrency support by running more than one workload at the same time

# Interpreter mode (TFLM)

- Tensor Flow Lite for Microcontrollers includes TFL model generation, compile time tools for the TFLM reference library and interpreter and runtime support for TFLM applications and interpreter.
- Includes NN libraries, which provide highly optimized implementations of TFLM operators on Tensilica products
- Daily regression testing by the TFLm team ensures TFLm works robustly on Tensilica IP

# Library Support

Tensilica products offer a highly optimized neural network library. This library provides optimized implementations of the most common layers and kernels for the target hardware. A programmer can leverage these functions to develop highly efficient neural networks in a short time. Tensilica continually adds and develops new operators and thoroughly tests this library to ensure state-of-the-art performance and robustness for AI development.

# Tensilica DSP Common Toolchain

Our processors are delivered with a complete set of software tools:

- A high-performance C/C++ compiler with automatic bundling and vectorization support for the VLIW and SIMD capabilities
- Linker, assembler, debugger, profiler, and graphical visualization tools
- A comprehensive instruction set simulator (ISS) allows you to quickly simulate and evaluate performance
- When working with large systems or lengthy test vectors, the fast, functional TurboXim simulator achieves speeds that are 40X to 80X faster than the ISS for efficient software development and functional verification
- Tensilica Xtensa Modeling Protocol (XTMP) for system modeling in C and Xtensa SystemC (XTSC) for system modeling in SystemC® provide for full-chip simulations, and the pin-level XTSC model offers co-simulation of the SystemC model at the pin level for fast, cycle-accurate system simulations

- All major back-end EDA flows are supported

# Protocol IP for AI

Cadence Protocol IP for AI, including PCI Express® (PCIe®), UALink, CXL, Universal Chiplet Interconnect Express (UCIe™), and advanced memory interfaces such as HBM and GDDR, is engineered to optimize AI applications across markets. These technologies provide high-performance, low-latency interconnects for efficient data transfer and seamless communication, supporting various data speeds and configurations. Our AI systems handle intensive computational tasks and large datasets with enhanced performance, power efficiency, and scalability.

- **PCIe and CXL for AI:** Cadence's PCIe and CXL products for AI deliver robust, high-speed interconnect solutions that facilitate efficient data transfer and communications. Our cutting-edge PCIe and CXL products support a wide range of configurations with processing nodes. By leveraging PCIe and CXL, AI developers can achieve faster data processing, reduced latency, and improved system reliability in their AI applications.

- **High-Speed SerDes for AI:** Cadence's high-speed SerDes products for AI enable ultra-fast data transfer and communication. These solutions support high data rates and maintain long-reach performance, ensuring efficient connectivity for AI applications. With industry-leading power efficiency and compatibility across various process nodes, our SerDes products help AI developers achieve superior system performance and reliability.

- **UCIe PHY and Controller for AI:** Cadence's UCIe products advance AI capabilities by delivering high-performance, low-latency interconnect solutions for chiplets within a package. These products enable seamless communication between chiplets and support a variety of data speeds and channel lengths. Available across multiple process nodes and configurations, they allow compatibility with various manufacturing technologies. Using UCIe, AI developers can now achieve higher performance, improved power efficiency, and greater scalability in their AI systems.

- **Memory Interface and Storage IP for AI:** Cadence's memory solutions are critical for AI advancements, offering high-performance, high-bandwidth memory systems such as DDR5, LPDDR5X, GDDR7, and HBM3E. These solutions are engineered to handle the intensive data throughput and low-latency requirements of AI training and inference. By optimizing data access patterns and minimizing bottlenecks, our memory solutions enable efficient processing

of large datasets and complex models, resulting in enhanced system performance and reliability.

## Cadence Services and Support

- Cadence Tensilica application engineers can answer your technical questions and provide technical assistance and custom training.
- Cadence certified instructors teach a series of courses on Tensilica IP and bring their real-world experience into the classroom.
- Internet Learning Series (iLS) online courses allow you the flexibility of training at your own computer via the internet.
- The Cadence Tensilica IP support site gives you 24x7 online access to a knowledge base of the latest solutions, technical documentation, software downloads, and more at support.