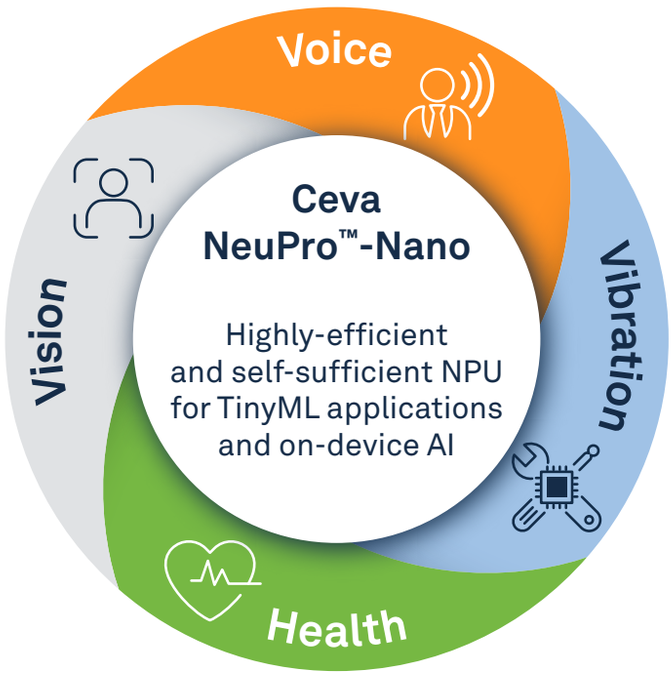
EDGE AI SENSING

CONTENT LIST

Ceva-NeuPro-Nano: TinyML Optimized NPUs for AIoT Devices

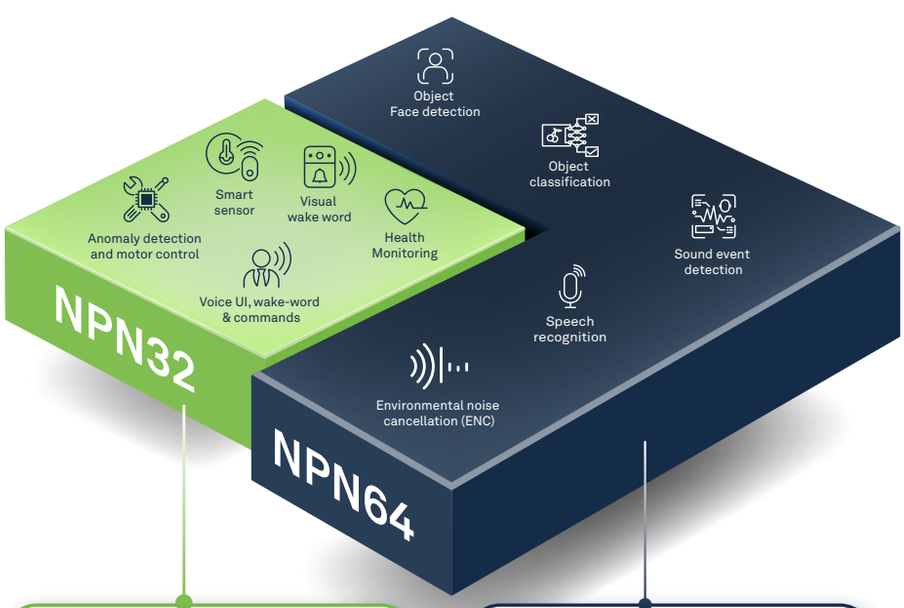
Ceva NeuPro-Nano™ is a highly efficient, self-sufficient Edge NPU designed for TinyML applications. Ceva-NeuPro-Nano NPUs deliver the optimal balance of ultra-low power and best performance in a small area to execute TinyML workloads across AIoT product categories, including Hearables, Wearables, Home Audio, Smart Home, Smart Factory, and more.

Edge NPU for TinyML Applications



Key Benefits

- Fully programmable and efficiently executes Neural Networks, feature extraction, control code and DSP code
- Supports most advanced machine learning data types and operators including native transformer computation, fast quantization and sparsity support with weight decompression and acceleration
- Self-sufficient architecture enables Ceva-NeuPro-Nano NPUs to deliver superior power efficiency, with a smaller silicon footprint, and optimal performance
- Enables on-device voice, vision, and sensing use cases across multiple end markets
- Two NPU configurations - Ceva-NPN32 and Ceva-NPN64 - to address a wide variety of use cases



- 32 4x8, 32 8x8, 16 16x8, 8 16x16, 4 32x32 MPY/MAC operations per cycle
- 80% memory footprint reduction using Weight De-compression
- Lowest cost implementation
- Highly optimized for TinyML workloads

- 128 4x8, 64 8x8, 32 16x8, 16 16x16, 4 32x32 MPY/MAC operations per cycle
- 2x performance acceleration using 50% weight sparsity (Sparsity Acceleration)
- Enhanced performance for complex networks

Flexible, Scalable, and Highly Efficient NPU Architecture

- Two Configurations: Ceva-NPN32 with 32 int8 MACs, and the Ceva-NPN64 with 64 int8 MACs (native 128 MACs of 4x8)
- Self sufficient, single core solution with no companion MCU/CPU required
- Ceva-NetSqueeze™ proprietary AI compression technology
  - Up to 80% memory footprint reduction through direct processing of compressed model weights without need for intermediate decompression stage
- Future proof architecture that supports the most advanced ML data types and operators
  - 4-bit to 32-bit integer support
  - Native transformer computation
- Ultimate ML performance for all use cases using advanced mechanisms:
  - Sparsity acceleration
  - Acceleration of non-linear activation types
  - Fast quantization – up to 5 times acceleration of internal re-quantizing tasks

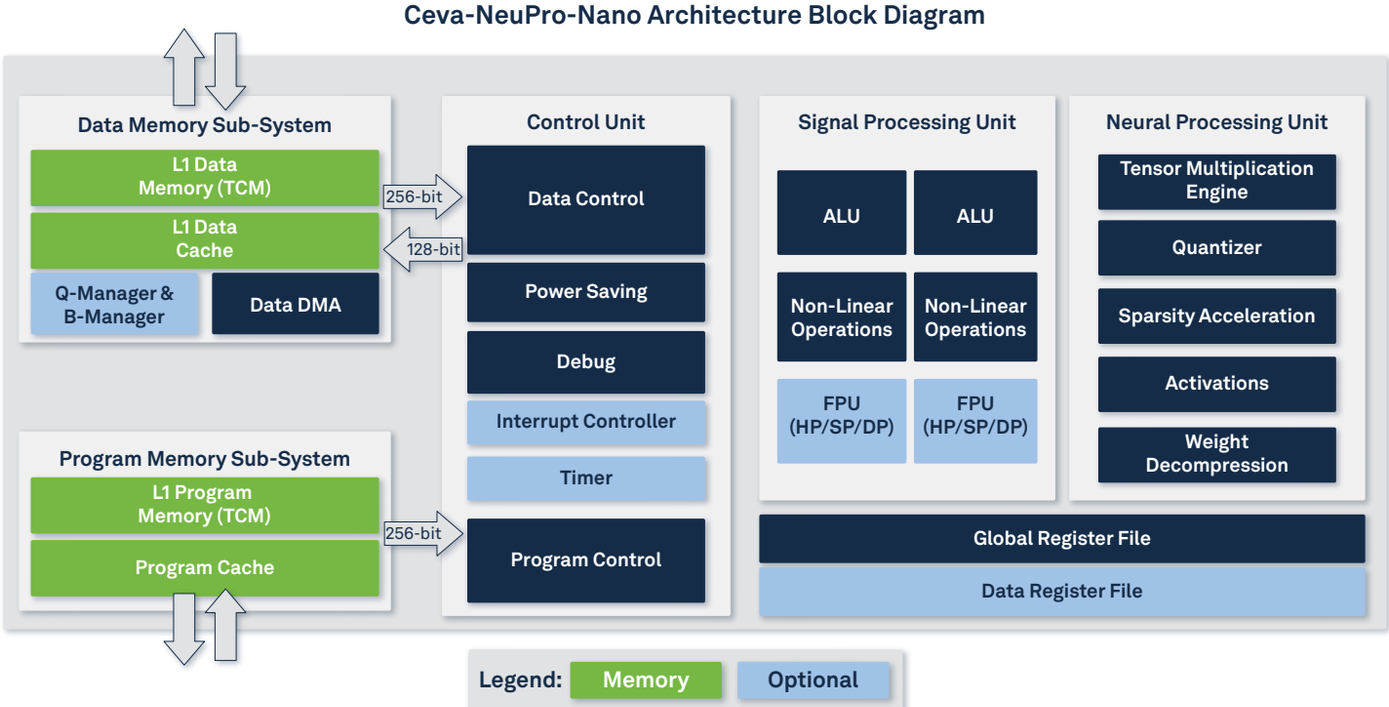
Ultra-low Energy Achieved Through Innovative Energy Optimization Techniques

- Dynamic Voltage and Frequency Scaling support - tunable for the use-case
- Dramatic energy and bandwidth reduction by distilling computations using weight-sparsity acceleration

Complete, Simple to Use AI SDK

- **Ceva-NeuPro Studio** provides a unified AI stack, with an easy click-and-run experience, for all Ceva-NeuPro NPUs, from the Ceva-NeuPro-Nano to the powerful Ceva-NeuPro-M
- Optimized to work seamlessly with leading, open AI inference frameworks including TFLM and µTVM
- Model Zoo of pretrained and optimized TinyML models covering voice, vision and sensing use cases
- Flexible to adapt to new models, applications and market needs
- Comprehensive portfolio of optimized runtime libraries and off-the-shelf application-specific software

Ceva-NeuPro-Nano Architecture Block Diagram



USA (HQ)
15245 Shady Grove Road
Suite 400
Rockville
MD, 20850
Tel: +1 (240) 308 8328

Israel
2 Maskit Street
P.O. Box 4047
Herzelia 4612001
Tel: +972 9 961 3700l

France
Les Bureaux Green Side 5
400, Avenue Roumanille
06410 Biot
Sophia Antipolis
Tel: +33 4 83 76 06 00

Ireland
Unit A2, First Floor
Building 6500
Airport Business Park
Kinsale Road, Cork
Tel: +353 1 237 3900

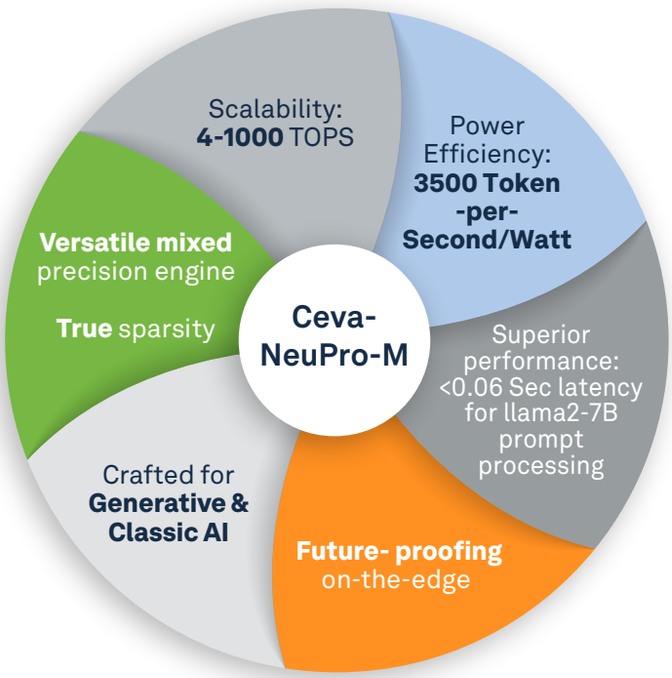
USA (West)
1174 Castro Street
suite 275
Mountain View
CA 94040
Tel: +1 (650) 417 7900

For more information:

Ceva-NeuPro-M: NPU IP family for generative and classic AI with highest power efficiency, scalable and future proof

Ceva-NeuPro-M NPU: A cutting-edge architecture with versatile mixed precision, true sparsity, and industry-leading power efficiency for unmatched performance and efficiency. Delivers scalable performance from 4 to 1000 TOPS, demonstrating power efficiency of 3500 tokens-per-second-per-Watt and 1800 tokens-per-second-per-mm² for llama2. Optimized to support diverse AI workloads across edge-to-cloud applications.

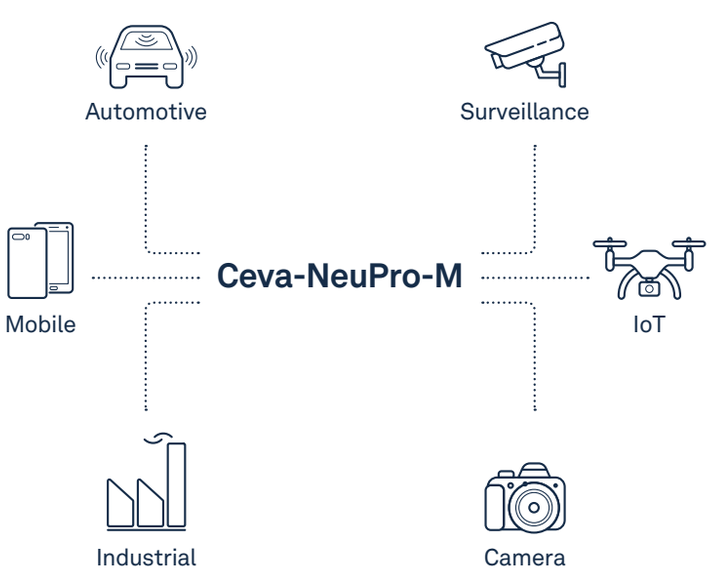
Ceva-NeuPro-M adaptive AI computing platform



Key Benefits

- Self-contained NPU architecture that concurrently processes diverse AI workloads with parallel coprocessing engines
- Future proof thanks to a programmable VPU (Vector Processing Unit) that enables any future layer or operator to be supported
- Outstanding energy efficiency of 3500 Token-per-Second/Watt
- Comprised of four Ceva-NeuPro-M NPU IPs for diverse application needs: NPM4K, NPM8K, NPM16K, NPM32K
- Up to 32 TOPS for a single engine NPM4K NPU
- Up to 256 TOPS for NPM32K NPU

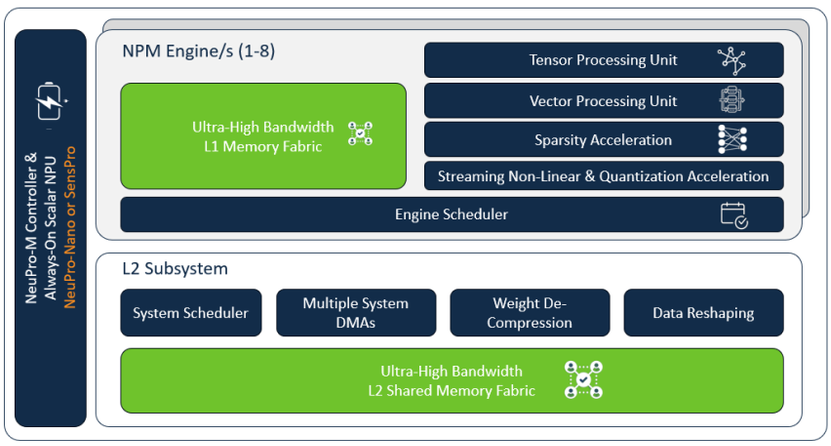
Target Markets



NPU IP Architecture

- Supports generative AI networks, transformers (self-attention), CNN, RNN, Multi modality, as well as 3D convolution and fully-connected layers
- Programmable and scalable VPU for handling any future new neural networks
- CSTL (Complementary Stream Logic) and VPU running in parallel to MAC array and multiple coprocessors to achieve ultimate utilization
- Support 4/8/16 bit integer and FP16 data-types
- Supports multi-level memory system hierarchy that enables multi-core scalability
- Weight compression for reduced DDR bandwidth
- True structured, semi-structured and unstructured sparsity to avoid operations with zero-value weights or activations
- Automotive ISO 26262 ASIL-B functional safety compliant
- Arch Planner profiling tool for IP evaluation and system prototyping, supporting over 150 AI networks, including generative AI and transformers, for applications such as segmentation, classification, detection, NLP, super resolution and more
- Ceva-NeuPro-M studio - a comprehensive AI software stack, which enables profiling performance and accuracy, optimizing AI models, graph compiling, AI inference simulation - based on open source platforms

Ceva-NeuPro-M Architecture Block Diagram



Ceva-NeuPro-M Core Configurations

Product Name	TOPS	Number of Engines	Security	Safety	Max. Frequency
NPM4K	32	1	Optional	Yes: NPM4KS	2GHz
NPM8K	64	1	Optional	Yes: NPM8KS	2GHz
NPM16K	128	2	Optional	Yes: NPM16KS	2GHz
NPM32K	256	4	Optional	Yes: NPM32KS	2GHz

USA (HQ)
15245 Shady Grove Road
Suite 400
Rockville
MD, 20850
Tel: +1 (240) 308 8328

Israel
2 Maskit Street
P.O. Box 4047
Herzelia 4612001
Tel: +972 9 961 3700l

France
Les Bureaux Green Side 5
400, Avenue Roumanille
06410 Biot
Sophia Antipolis
Tel: +33 4 83 76 06 00

Ireland
Unit A2, First Floor
Building 6500
Airport Business Park
Kinsale Road, Cork
Tel: +353 1 237 3900

USA (West)
1174 Castro Street
suite 275
Mountain View
CA 94040
Tel: +1 (650) 417 7900

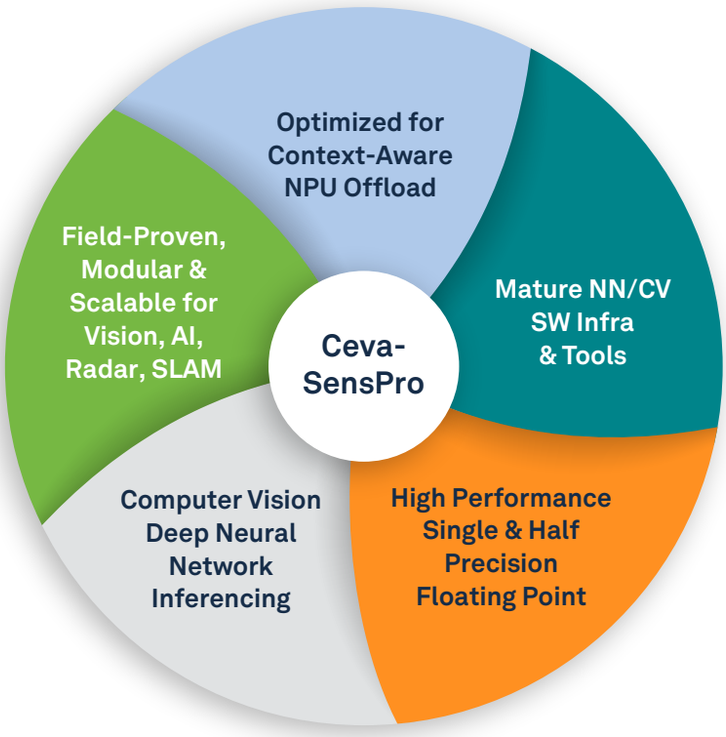
For more information:

Ceva-SensPro2 – 2nd Generation High-Performance Vision AI DSP Architecture

Ceva-SensPro2

Ceva-SensPro2 is a powerful and flexible Vector DSP architecture, crafted for advanced AI pre/post-processing, computer vision, and signal processing workloads.

Ceva-SensPro2



SensPro2 is Crafted For:

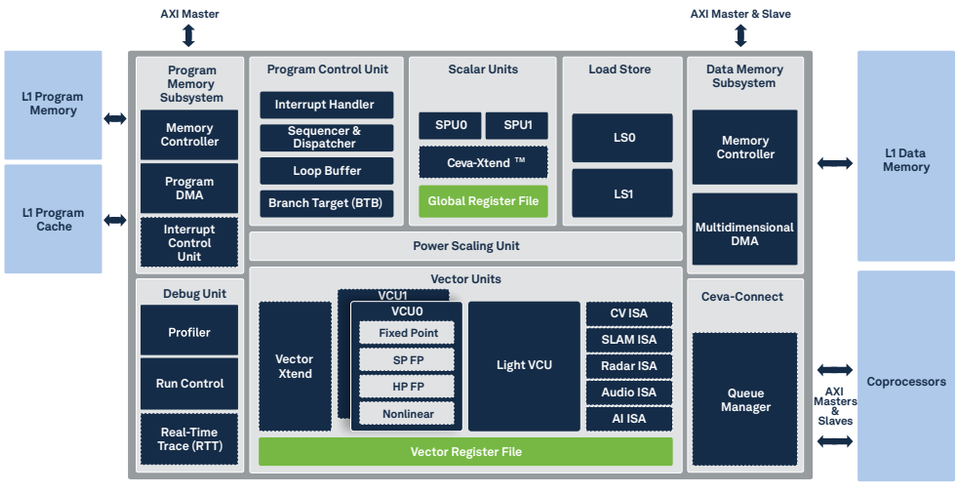
- NPU Companion
  - Computer vision kernels & algorithms
- Exceptional performance for AI foundation operations
- Supplements ML and AI pre- & post-processing tasks
- Easy to offload from the NPU
- Best-in-class efficient 2D image processing, enabling simultaneous multi-pixel, multi-operation handling
- Efficient linear-algebra operations critical for vision and signal processing pipelines
- Best Vector Processor Performance, balanced, scalable, and configurable vector architecture
- Optimized for Performance/Watt and Performance/mm² trade-offs
- Built-in ASIL B (random faults) and ASIL D (systematic fault) support for automotive-grade safety
- Highly scalable across a wide range of configurations — from low-power edge devices to high-performance vision and automotive systems
- Deployed in multiple field-proven production programs across vision and automotive systems, robotics, and industrial applications

Target Markets

Ceva-SensPro2 Highlight Features

- 8-way VLIW
- 1.25GHz@7nm Max frequency
- 400 GFLOPS for floating point arithmetic
- Up to 3.2 TOPS for 8x8 networks inferencing
- Memory Architecture
  - 400 GByte/second Data Bandwidth
  - 4-Way Instruction Cache
  - DMA and Queue and Buffer Managers
- Complementary software stack including Ceva AI, Ceva-CV and Ceva-SLAM SDK
- SmartFrame2.0: unified runtime for AI, CV, SLAM, and DSP
- Halide & Vec-C: optimized offline kernel development
- SensPro2 Graph Compiler: plug-in for TVM backend for model partitioning

Ceva-SensPro2 Architecture Block Diagram



Ceva-SensPro2 Processors Family

Ceva-SensPro2 Core	MAC Configuration			Target Application
	INT8	INT16	FP32	
SP100	128	32	Optional	
SP250	256	64	Optional	
SP500	512	128	Optional	
SP1000	1024	256	Optional	
SPF2	-	-	32	
SPF4	-	-	64	

USA (HQ)
15245 Shady Grove Road
Suite 400
Rockville
MD, 20850
Tel: +1 (240) 308 8328

Israel
2 Maskit Street
P.O. Box 4047
Herzelia 4612001
Tel: +972 9 961 3700l

France
Les Bureaux Green Side 5
400, Avenue Roumanille
06410 Biot
Sophia Antipolis
Tel: +33 4 83 76 06 00

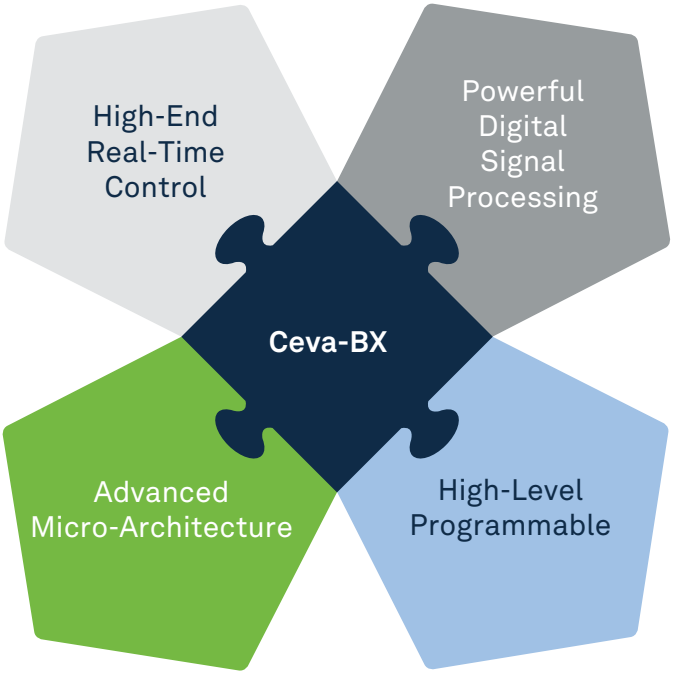
Ireland
Unit A2, First Floor
Building 6500
Airport Business Park
Kinsale Road, Cork
Tel: +353 1 237 3900

USA (West)
1174 Castro Street
suite 275
Mountain View
CA 94040
Tel: +1 (650) 417 7900

For more information:

High level programmable, modern processor architecture for a broad range of signal processing and control workloads

Ceva-BX2 is a multipurpose hybrid DSP and Controller designed for the inherent low power requirements of DSP kernels with high-level programming and compact code size requirements of a large control code base.



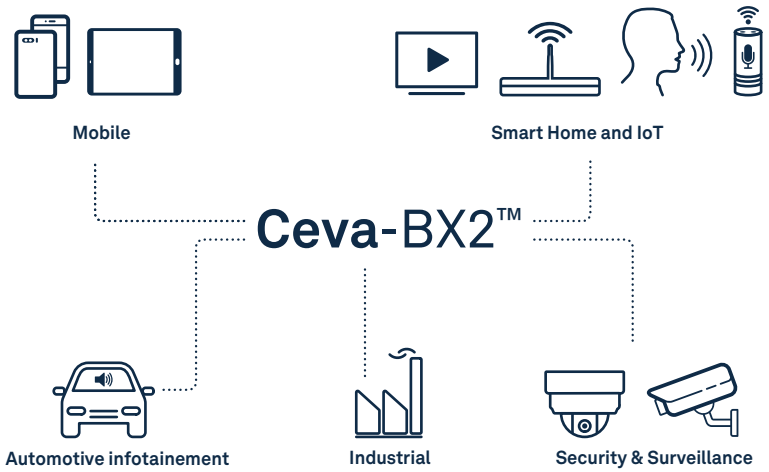
Key Benefits

- **Perfect alternative** for special purpose DSPs and MCUs with DSP coprocessors, allowing efficient handling of today's diverse algorithm needs
- **Fast software bring up** using a comprehensive tool chain and DSP libraries
- **Advanced Neural Network** support for deep learning at the edge
- **Compact code size** and **low power** using inbuilt mechanisms

Tools and Software Availability

- **Application specific ISA and software:**
  - ClearVox - multi-mic noise reduction
  - WhisPro - speech recognition
  - Cellular IoT and GNSS ISA
- **DSP and neural network compute libraries**
- **Common neural network frameworks support**
- **Advanced LLVM compiler**
- **Eclipse based debugger**
- **RTOS**

Target Markets



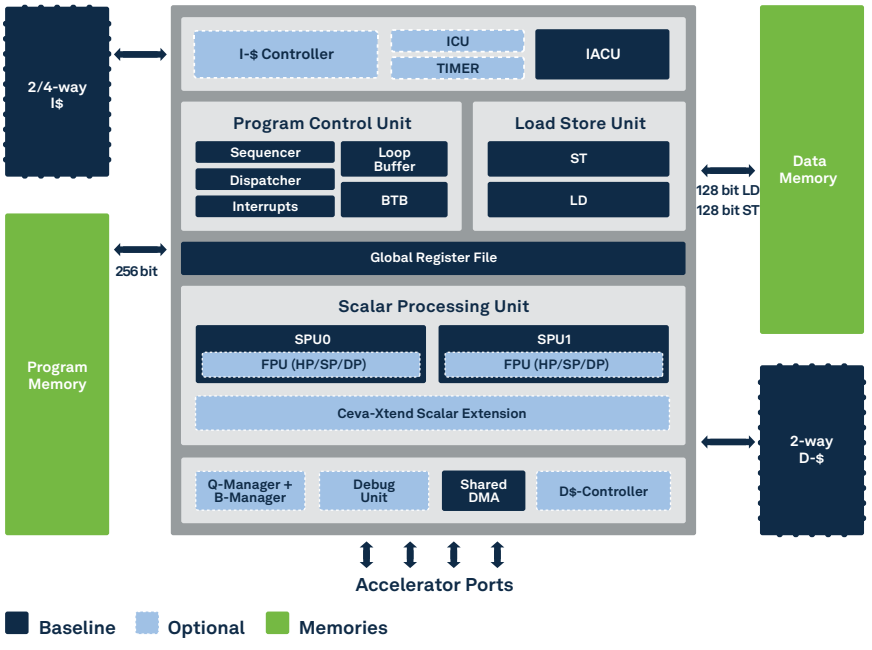
Multipurpose DSP/Controller

The Ceva-BX2 architecture delivers excellent all-round performance for a new generation of smart devices by offering a high performance hybrid architecture that is a single compute island for all DSP and control workloads.

Ceva-BX™2

Architecture Highlights

- **Powerful DSP performance**
  - 5-way VLIW
  - 8/16/32/64-bit data types
  - Quad 32x32 MACs
  - Octal 16x16 MACs
  - 16x8 and 8x8 Neural Network support
  - Half, single and double precision IEEE floating point units
- **High performance controller**
  - 4.5 CoreMark/MHz
  - Dynamic branch prediction
  - Full RTOS support
  - Secured execution modes
  - Compact code size
- **Advanced System Control**
  - Automatic buffer management
  - High QoS with queue managers
  - Dedicated HW accelerator ports
  - Advanced multi-way instruction and data cache subsystem



USA (HQ)
15245 Shady Grove Road
Suite 400
Rockville
MD, 20850
Tel: +1 (240) 308 8328

Israel
2 Maskit Street
P.O. Box 4047
Herzelia 4612001
Tel: +972 9 961 3700l

France
Les Bureaux Green Side 5
400, Avenue Roumanille
06410 Biot
Sophia Antipolis
Tel: +33 4 83 76 06 00

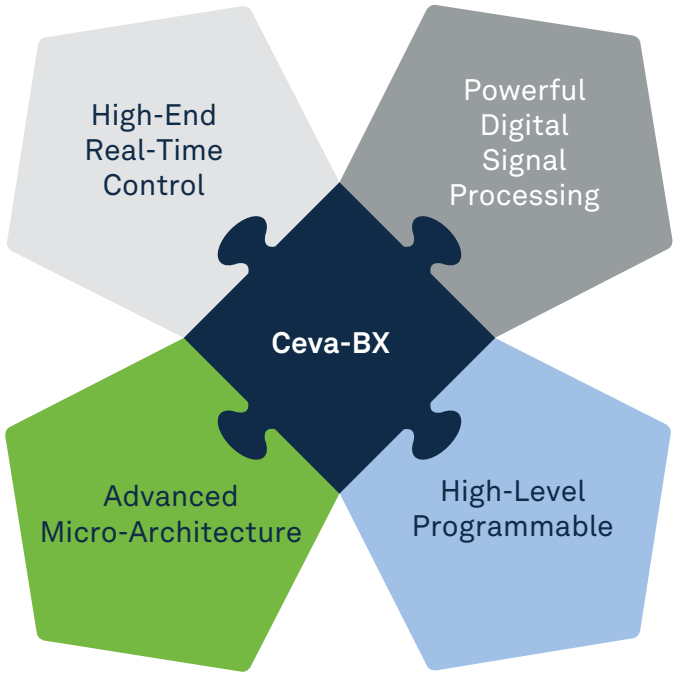
Ireland
Unit A2, First Floor
Building 6500
Airport Business Park
Kinsale Road, Cork
Tel: +353 1 237 3900

USA (West)
1174 Castro Street
suite 275
Mountain View
CA 94040
Tel: +1 (650) 417 7900

For more information:

High level programmable, modern processor architecture for a broad range of signal processing and control workloads

Ceva-BX1 is a multipurpose hybrid DSP and Controller designed for the inherent low power requirements of DSP kernels with high-level programming and compact code size requirements of a large control code base.



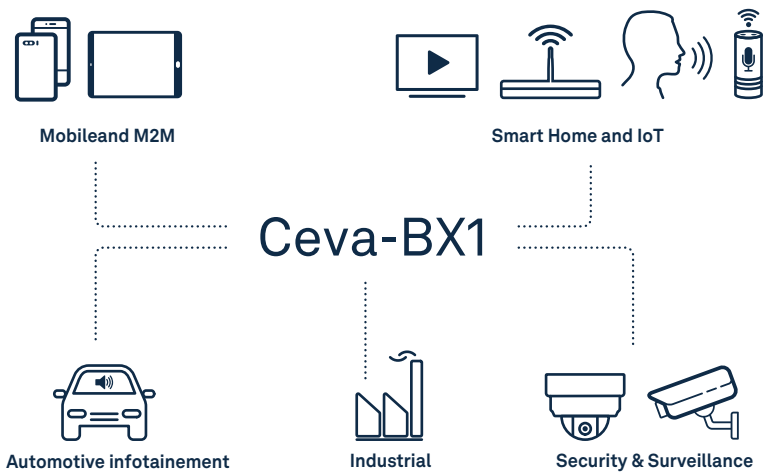
Key Benefits

- **Perfect alternative** for special purpose DSPs and MCUs with DSP co-processors,allowing efficient handling of today's diverse algorithm needs
- **Fast software bring up** using a comprehensive tool chain and DSP libraries
- **Advanced Neural Network** support for deep learning at the edge
- **Compact code size** and **low power** using inbuilt mechanisms

Tools and Software Availability

- **Application specific ISA and software:**
  - **ClearVox** - multi-mic noise reduction
  - **WhisPro** - speech recognition
  - **Cellular** IoT and **GNSS** ISA
- DSP and neural network compute libraries
- Common neural network frameworks support
- Advanced LLVM compiler
- Eclipse based debugger
- RTOS

Target Markets



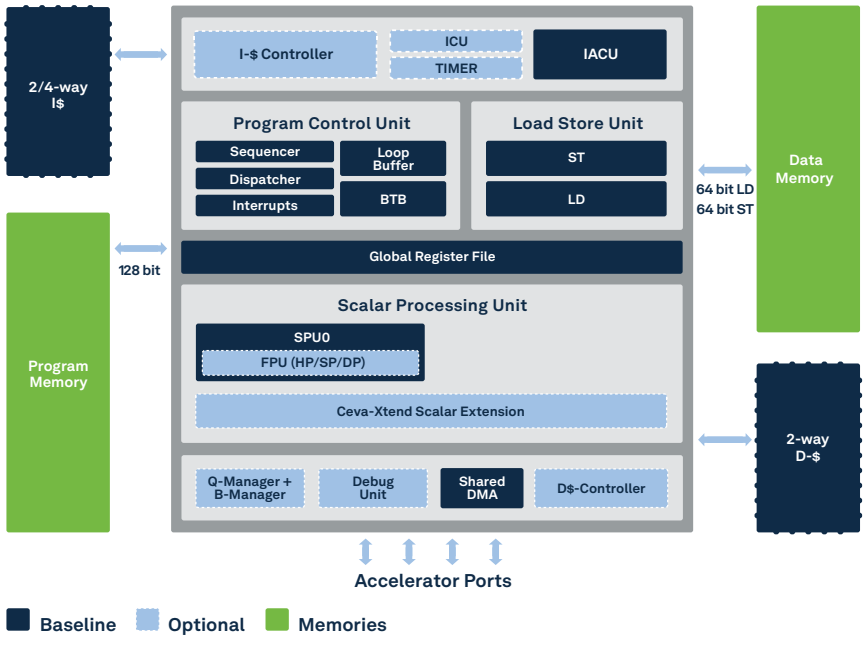
Multipurpose DSP/Controller

The CEVA-BX1 architecture delivers excellent all-round performance for a new generation of smart devices by offering a high performance hybrid architecture that is a single compute island for all DSP andcontrol workloads.

Ceva-BX™1

Architecture Highlights

- **Powerful DSP performance**
  - 4-way VLIW
  - 8/16/32/64-bit data types
  - Dual 32x32 MAC
  - Quad 16x16 MAC
  - 16x8 and 8x8 Neural Network support
  - Half, single and double precision IEEE floating point units
- **High performance controller**
  - 3.73 CoreMark/MHz
  - Dynamic branch prediction
  - Full RTOS support
  - Secured execution modes
  - Compact code size
- **Advanced System Control**
  - Automatic buffer management
  - High QoS with queue managers
  - Dedicated HW accelerator ports
  - Advanced multi-way instruction and data cache subsystem



USA (HQ)
15245 Shady Grove Road
Suite 400
Rockville
MD, 20850
Tel: +1 (240) 308 8328

Israel
2 Maskit Street
P.O. Box 4047
Herzelia 4612001
Tel: +972 9 961 3700l

France
Les Bureaux Green Side 5
400, Avenue Roumanille
06410 Biot
Sophia Antipolis
Tel: +33 4 83 76 06 00

Ireland
Unit A2, First Floor
Building 6500
Airport Business Park
Kinsale Road, Cork
Tel: +353 1 237 3900

USA (West)
1174 Castro Street
suite 275
Mountain View
CA 94040
Tel: +1 (650) 417 7900

For more information: