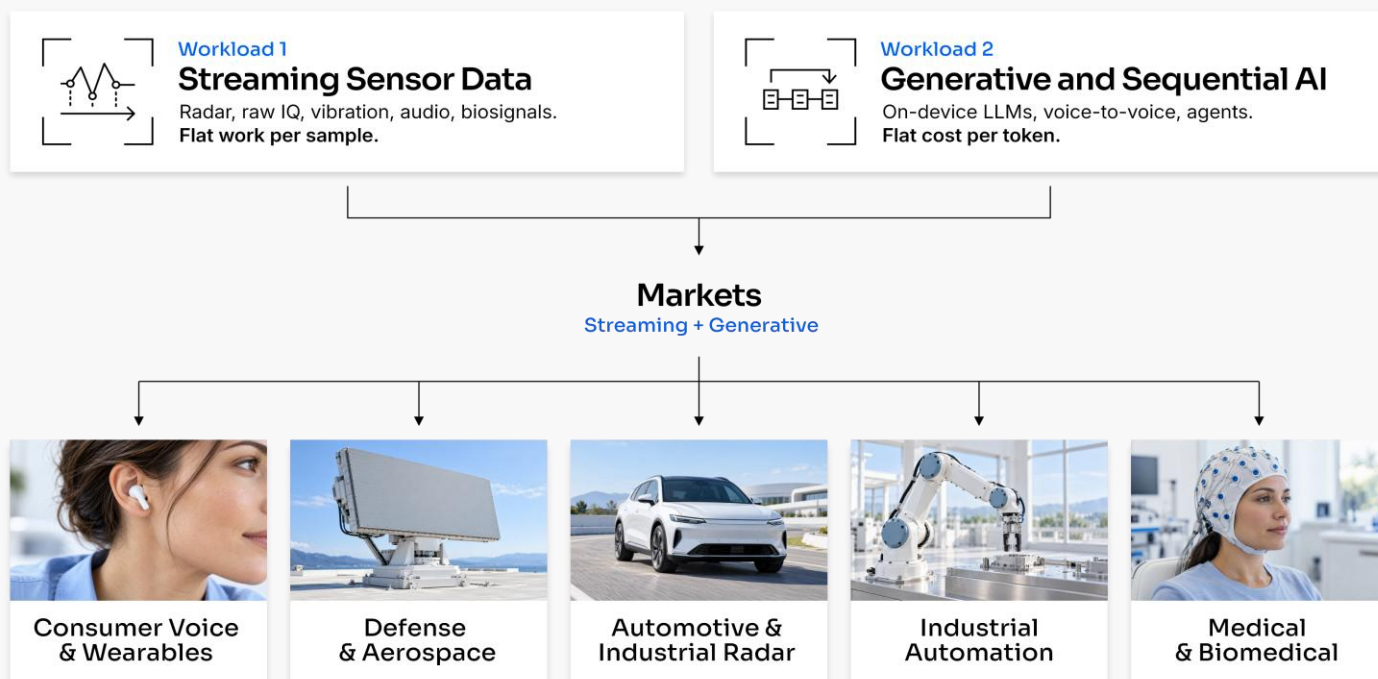


sequora™: A Generative AI Edge Processor

Sequora is a Generative AI edge processor, licensed as IP, built for combining both time-series and Generative AI workloads. Its state-space model (SSM) layers run both signal processing and neural inference on a purpose built multiply accumulate (MAC) array, to combine pre/post-processing with inference. Sequora optimizations for transformer-based-Attention and SSM operators minimize memory usage that dominates large language model (LLM) inference, enabling a complete voice-to-voice generative pipeline to run on the device without connectivity. It can support a wide range of applications in multiple markets including automotive, industrial IoT, defense and wearables where battery usage, latency and lack of connectivity are critical requirements. Sequora is a licensable core IP that is process-portable, fully synthesizable, verified RTL.

Combines Two Workloads in One Processor



Built for Generative AI at the Edge

The Challenge: Combining Edge AI with Generative AI

- ❖ **Conventional NPUs were built for a different workload.** They were designed for image recognition, where arithmetic was the bottleneck. Generative decode is memory-bound, and adding TOPS does not fix that.
- ❖ **Transformer memory grows with the conversation.** The KV cache stores data for every word so far, so each new word costs more memory and more time than the last.
- ❖ **Two blocks to do one job.** A DSP has to pre-process the sensor signal before the NPU can read it — two cores, toolchains and additional memory and network connectivity.
- ❖ **Memory traffic sets the power ceiling.** Loading weights from memory uses far more power than the computations. On a battery, that is what limits model size.
- ❖ **Runtime decisions make timing unpredictable.** Automotive, medical and avionics certification needs a guaranteed worst case, which run-time branching cannot provide.

The Solution: Sequora IP

- ❖ **Memory efficient decode.** Programmable parameter quantization and a memory-efficient decode path mean throughput follows parameter memory bandwidth, not MAC count.
- ❖ **Bounded memory growth for KV cache while using transformers.** Our transformer model optimizations reduce memory usage and helps transformer/attention operators to run more efficiently.
- ❖ **Separate DSP not required.** SSM-DSP layers run the frequency analysis on the same array: Streaming from the first sample, Full Spectrum with FFT - identical output, Selective Frequency at up to 20x fewer MACs.
- ❖ **Eliminate memory growth by using SSM models.** SSM/R-TENN operators hold history in a fixed-size state, so word-1,000 costs the same as word-1. SRAM stays constant whether the input lasts a second or an hour.
- ❖ **No connectivity required.** ASR · NLU · TTS and LLM decode run locally — no round trip, no per-query cost, data never leaves the device.
- ❖ **No runtime surprises.** Graphs are compiled at build time and issued as microcode in order — no instruction fetch, no runtime branching. Every run takes the same path, so the worst case can be certified.

brainchip™ Sequora IP Product Brief

Key Features

ISA-based architecture: programmable like a CPU, efficient like fixed-function silicon — new algorithms ship as software.

Native state-space processing: SSM / R-TENNs layers hold history in a fixed-size state, so memory is constant for any signal length.

Memory-efficient: Patented non-linear 4-bit codes use 4x less memory. SSM decode cuts per-token memory traffic to a level an edge platform can utilize.

Unified MAC array: FFT- and MFCC-equivalent analysis runs on the same array as the neural layers — no DSP preprocessing stage.

Built for time series: audio, vibration, EEG and radar at constant compute per sample, with no frame buffering.

Frozen frequency layers: physics-specified values that need no retraining, so the deployed model stays certifiable.

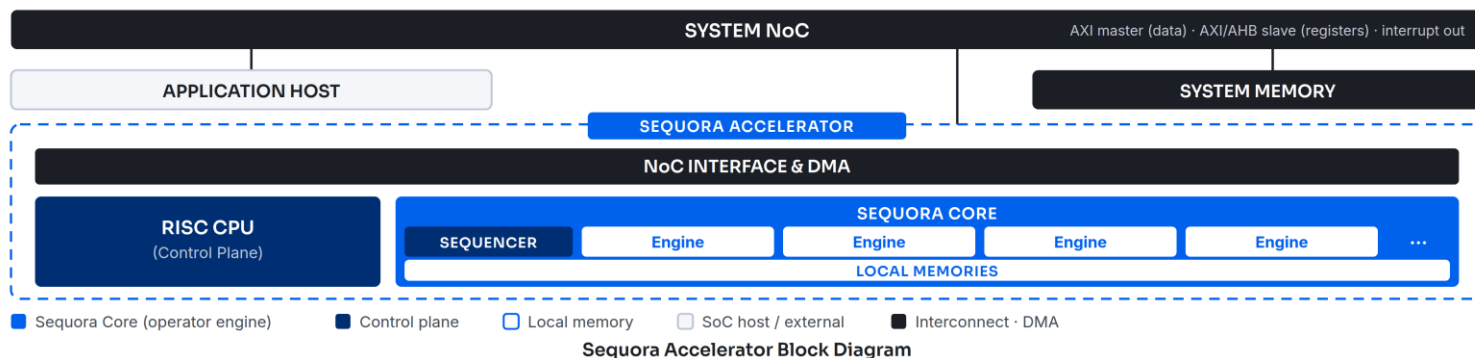
Platform

Self-Contained SSM Processor

- Executes SSM, TENNs-Recurrent and transformer-class layers natively
- Control-plane CPU dispatches compiled graphs to the engine
- Fixed-size state memory, independent of sequence length

SSM-DSP Operating Modes

- **Full Spectrum:** FFT-identical output using only MAC operations
- **Selective Frequency:** up to 20× fewer MACs for physics-specified bands
- **Streaming:** a spectral estimate from the first sample



Measured — SSM-DSP vs a Conventional CPU-DSP Pipeline*

59% lower latency (12.5 → 5.1 ms) · 14× lower jitter (± 0.18 ms) · 71% less CPU memory · +3.1% pAUC vs librosa MFCC
Generative: under 100M 4-bit params at 8 inf/s (400 MB/s) · under 420M at 20 tok/s (4.2 GB/s) · 20 tokens per billion params/joule

Specifications

Architecture	ISA-based programmable MAC array; signal processing, inference, generative neural layers execute on the same hardware
SSM-DSP modes	Streaming · Full Spectrum (FFT-identical output) · Selective Frequency (up to 20× fewer MACs)
Layer support	SSM / R-TENNs state-space layers and transformer-class layers
Precision	Int4, int8 and FP16 native
Memory	Fixed-size SSM state; SRAM footprint independent of sequence length. Standard RAMs, no specialty physical macros
Generative AI	Bandwidth-bound: under 100M 4-bit params at 8 inferences/s (400 MB/s); under 420M at 20 tokens/s (4.2 GB/s)
Reference models	Radar, EW and SSM-DSP models, including an all-SSM FMCW pipeline (160K parameters, ~211M MACs)

Software Development

- **Standard frameworks:** PyTorch / ONNX, TensorFlow / Keras, Edge Impulse
- BrainChip MetaTF for model development; model-, network- and OS-agnostic
- **Model zoo:** pre-built radar, EW and SSM-DSP models with weights and scripts

IP Delivery

- Process-portable, fully synthesizable soft RTL, packaged for your application
- Verification IP and testbench with simulation results
- RTL synthesis scripts and timing constraints
- Static graph compiler + runtime C++ library

Next Steps

- Available Now.
- **Request the evaluation package:** RTL access, testbench and reference models.
- Contact your BrainChip representative or sales@brainchip.com

*All the data is preliminary, and subject to change without notice.