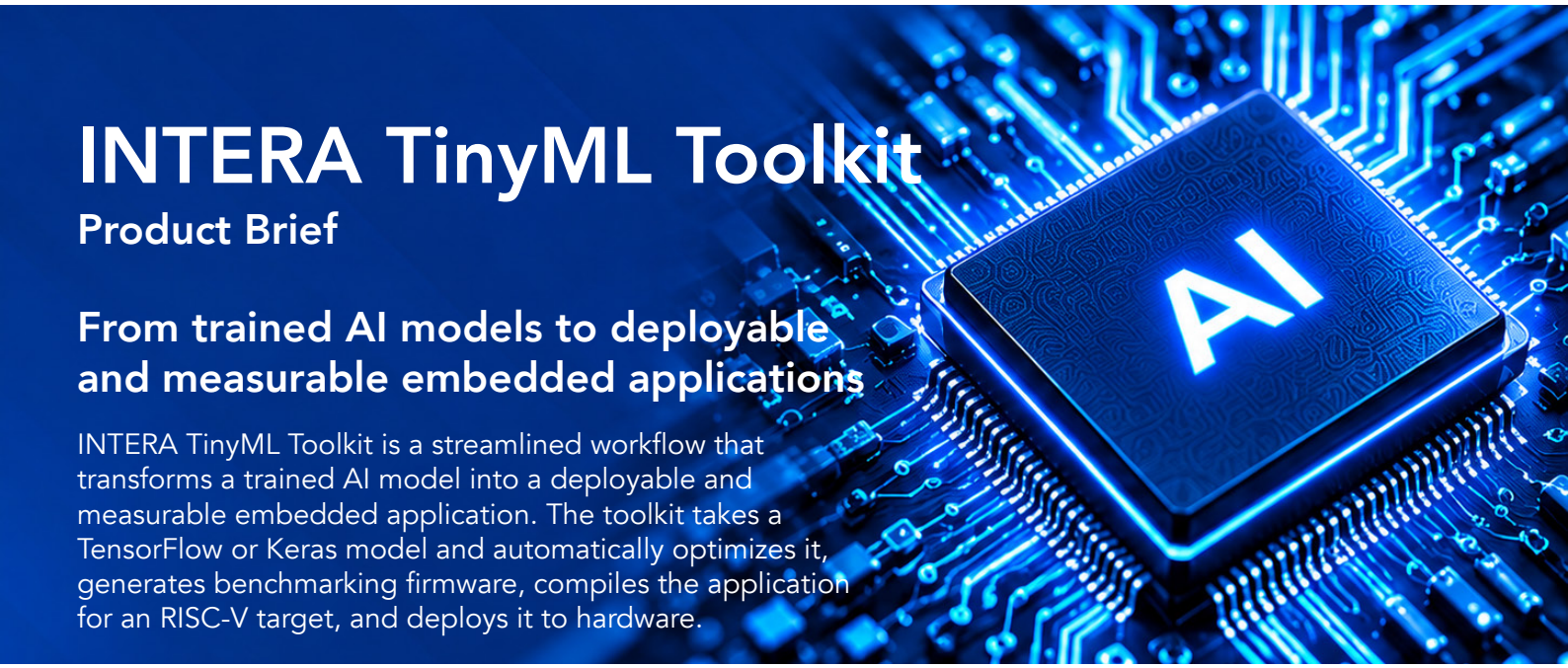
INTERA
SEAMLESS AI

INTERA TinyML Toolkit

Product Brief



From trained AI models to deployable and measurable embedded applications

INTERA TinyML Toolkit is a streamlined workflow that transforms a trained AI model into a deployable and measurable embedded application. The toolkit takes a TensorFlow or Keras model and automatically optimizes it, generates benchmarking firmware, compiles the application for an RISC-V target, and deploys it to hardware.

One workflow, from an AI model to measured results

Trained AI model
TensorFlow/ Keras

→

INTERA TinyML Toolkit:
Convert and optimize
Generate a project and benchmark firmware
Build for RISC-V

→

Deploy and measure
RISC-V hardware
Latency, memory, and cycles

---

01 Convert
Transform and quantize the trained AI model into TensorFlow Lite format.

→

02 Generate
Create a TensorFlow Lite Micro project with embedded model data and runtime integration.

→

03 Benchmark
Automatically generate firmware for functional validation and performance measurement.

04 Build
Compile the application using the RISC-V toolchain and build system.

→

05 Deploy
Program target hardware and execute inference workloads.

→

06 Analyze
Generate reports with memory usage, execution cycles, latency, and accelerator performance metrics.

C/Sabino de Arana 14-16, 2nd - Barcelona, Spain - 08028 info@intera-group.com www.intera-group.com

INTERA
SEAMLESS AI

INTERA TinyML Toolkit

Product Brief

Key Capabilities

AI Model Optimization

Keras and TensorFlow model support
TensorFlow Lite conversion
Quantization (float, dynamic, integer)
Accelerator-aware optimization
CNN, FCNN, MLP, RNN, LSTM, GRU

Automated Firmware Generation

TensorFlow Lite Micro integration
Automatic C model generation
Benchmark scaffolding
Memory arena configuration

Embedded Deployment

RISC-V toolchain support
FPGA and SoC deployment
OpenOCD + GDB integration
Automated build and flashing

Performance Analytics

Memory footprint analysis
Layer-by-layer profiling
Cycle counting
Accelerator vs CPU comparison
HTML reports

Supported INTERA Cognit AI Accelerators

INTERA Falcon AI Accelerator	Monolithic, streaming accelerator for efficient MxV computation
INTERA Hawk Advance AI Accelerator	RISC-V-based accelerator with custom instructions and scalable MxV / MxM
INTERA Hawk Pro AI Accelerator	RISC-V-based accelerator with higher performance and configurability
INTERA Eagle AI Accelerator	Near-memory computing accelerator for high energy efficiency
INTERA Harrier AI Accelerator	Monolithic accelerator for scalable matrix computation (MxM)

Automatic accelerator integration • Integer-quantized execution • Benchmark and profiling support

Typical Use Cases

Edge AI Prototyping

Move from a trained model to embedded execution in a reproducible workflow.

Accelerator Evaluation

Compare custom hardware accelerators against baseline CPU execution.

TinyML Benchmarking

Measure latency, memory usage and execution cycles.

Production Readiness

Generate deployable firmware and validation reports.

C/Sabino de Arana 14-16, 2nd - Barcelona, Spain - 08028 info@intera-group.com www.intera-group.com