300 Orchard City Dr, Suite #170
Campbell, CA 95008
(408) 412-8645

# TritonAI™ 64 Platform: Delivering AI at the Edge

Wave Computing's customizable TritonAI™ 64 platform combines a triad of powerful technologies in a single, licensable solution that enables efficient artificial intelligence (AI) at the edge. The scalable platform offers a flexible and efficient way for system on chip (SoC) developers to incorporate AI inferencing capabilities into their edge computing designs.

## Future-Proof Platform Design

The world of AI is constantly evolving, with both the algorithms and trained models for varying use cases changing frequently. The TritonAI 64 platform helps organizations future-proof their environments from continual change by delivering a flexible design using 8-to-32-bit integer-based, high-performance AI inferencing at the edge today.

Designing SoCs with embedded AI capabilities can be a complex undertaking requiring a highly flexible, adaptable architecture. The TritonAI 64 platform includes three powerful, scalable technologies developers can easily configure to address a broad range of AI use cases and computational requirements:

- MIPS® 64-bit SIMD engine
- WaveFlow™ dataflow engine
- WaveTensor™ processing engine

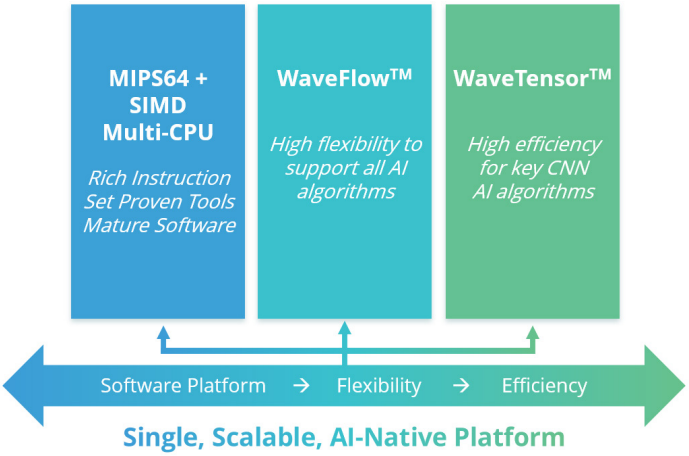
The TritonAI 64 platform delivers varying performance levels by incorporating additional compute elements from each of the three technologies in a modular and linear fashion. Designers can configure each of the three modules as needed to address the performance needs for varying AI use cases.

## Key Features:

- Mature IDE & tools
- Integrated driver layer for technology mapping
- Linux support
- Abstract AI framework via the Wave Runtime (WaveRT) API
- Optimized AI libraries for:
  - CPU/SIMD/WaveFlow/WaveTensor
  - TensorFlow-Lite Build support

The TritonAI 64 platform will automatically update with each software iteration to ensure customers' environments keep pace with rapidly evolving AI requirements.

## TritonAI 64 Platform for AI-enabled Edge SoCs




## Benefits:

- **Flexible support** for a broad range of AI use cases
- **Efficient** execution of current AI CNN algorithms
- **Easily configurable** allowing customers to scale performance to meet changing AI algorithmic needs and use cases
- **Comprehensive**, future-proof, licensable IP platform
- **Support** for 8- to 32-bit integer-based inferencing

300 Orchard City Dr, Suite #170
Campbell, CA 95008
(408) 412-8645

# TritonAI 64 Platform Components

## MIPS-64 RISC CPU Technology

The platform's MIPS-64 RISC CPU coupled with a mature integrated developer environment (IDE) provides a robust solution for developing AI applications, stacks and use cases. The IDE also includes:

- MIPS64r6 instruction set architecture
- 128-bit SIMD/FPU for INT/SP/DP ops
- Virtualization extensions
- Superscalar 9-stage pipeline w/SMT
- Caches (32KB-64KB), DSPRAM (0-64KB)
- Advanced branch prediction and MMU

**Multi-Processor Cluster:**

- 1-6 cores
- Integrated L2 cache (0-8MB, opt ECC)
- Power management (F/V gating/CPU)
- Interrupt control with virtualization
- 256b native AXI4 or ACE interface

## WaveFlow Dataflow Platform

The platform's WaveFlow subsystem features low latency, single batch sized, AI network execution with the flexibility to address concurrent AI network execution. The subsystem includes:

- Configurable IMEM and DMEM Sizes
- Overlap of communication & computation
- Compatible datatypes with WaveTensor
- Integer (Int8, Int16, Int32)
- Wide range of scalable solutions (2-1K tiles)
- Future-proof for all AI algorithms
- Flexible multi-dimensional tiling
- Supports signal and vision processing

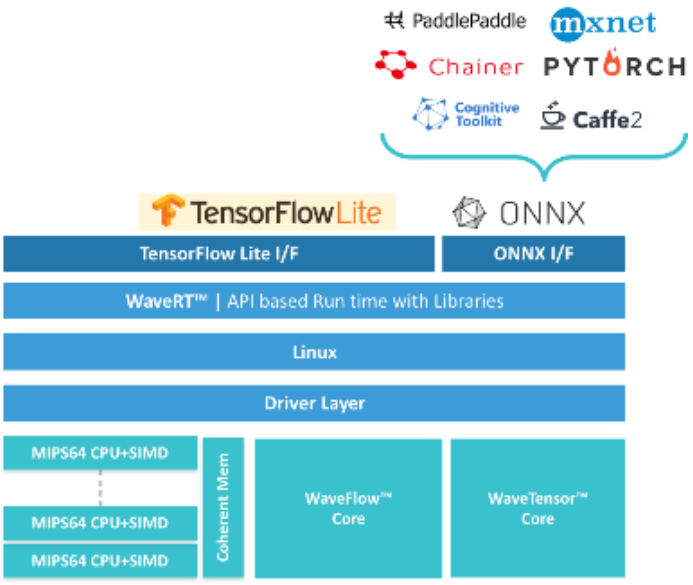
## WaveTensor Processing Engine

The WaveTensor architecture can scale up to a PetaOP of 8-bit integer operations on a single core instantiation for the highly efficient execution of today's key Convolutional Neural Network (CNN) algorithms. The subsystem features:

- Configurable MACs, accumulation & array size
- 4x4 (64 MAC) and 8x8 (512 MAC) base tiles
- Up to 32 tiles per slice & up to 32 slices per array
- Slices up to overlap of communication & computation
- Supports int8 for inferencing

## Application Programming Kit

The TritonAI 64 platform also includes an APK that allows developers to parse and execute appropriate AI tasks using the WaveFlow and WaveTensor acceleration engines. These engines are controlled in a heterogeneous programming environment managed by a unifying API platform, the Wave Run-Time (WaveRT). This software-centric approach abstracts the AI use case, AI framework and AI algorithms from the dedicated silicon executing the code, allowing developers to exploit algorithmic parallels for faster performance.




*The TritonAI 64 APK provides for a TensorFLow-Lite framework stack running on a Linux OS*

## About Wave Computing

Wave Computing is revolutionizing artificial intelligence (AI) with its dataflow-based systems and solutions. The company's vision is to bring deep learning to customers' data wherever it may be—from the datacenter to the edge—helping accelerate time-to-insight. Wave is powering the next generation of AI by combining its dataflow-based architecture with its MIPS embedded RISC multi-threaded CPU cores and IP.