Kalray TC4™ High-Performance Acceleration

Kalray TC4™ High-performance acceleration card

Leveraging the Massively Parallel Processor Array (MPPA®) processor architecture, Kalray patented ground-breaking manycore technology

Fully programmable card for a large variety of applications requiring efficient acceleration of heterogeneous tasks in parallel

- AI Inference
- Computer Vision
- Industry 4.0
- Server Infrastructure

The Kalray TC4 offers an unprecedented level of acceleration performance and power efficiency by integrating four MPPA® Coolidge™ v2 processors on a single PCIe card.

The TC4 card is designed to meet the needs of compute intensive applications deployed in servers on premises or in datacenters. It delivers a high-density of compute and process power for a combination of AI and non-AI workloads. It is designed to maximize the performance per server PCIe slot and is well suited for a large variety of use cases including AI inference, computer vision, signal processing, and offloading of complex algorithms.

The TC4 is an excellent choice for customers looking for the best performance per dollar and performance per watt for the most demanding industry solutions.

Specifications	
Main Component	4x MPPA®3-80 v2.0 @1GHz
Number of Cores	320 independent 64-bit VLIW cores 320 tensor accelerators
Supported Formats	INT8/16/32/64, FP16/32/64
Performances	INT8: 174 TOPS, FP16: 92 TFLOPS
Benchmarks	2.3 kDMIPS/MHz/Core 3.52 Coremark/MHz/Core
Memory	32GB DDR4 @ 3200MT/s 160MB total on-chip memory
Memory Bandwidth	200 GB/s
Main Input/Output	PCIe Gen4 16-lane
Format	Dual-slot, GPU Form Factor
Power	150W typical
Dissipation	Passive

Computer Vision Example Use Case

Application

Quality inspection process for manufacturing requiring object detection and image classification in large-scale industrial system

Customer challenge

- Achieve performance targets for system based on combination of customized AI model and classical computer vision algorithms
- Reduce the Total Cost of Ownership related to AI inferencing
- Maximize compute density to answer infrastructure footprint constraints

Kalray TC4 performance highlights

- Processing 1 Giga pixels/s per card
- 6,500 frames/s in FP16
- 9,990 frames/s in INT8

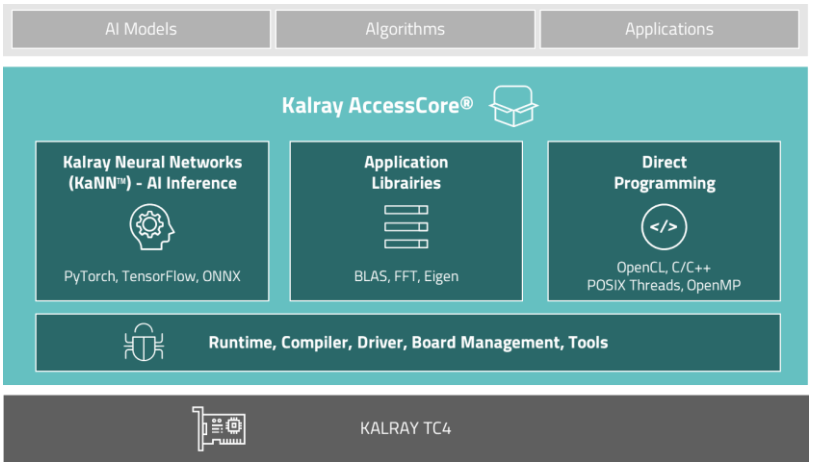
Benefits compared to leading GPU

- 3X compute efficiency
- 4X in performance per dollar
- Possibility to combine AI and computer vision algorithms on the same Kalray TC4 card

Kalray AccessCore® Software SDK

Kalray AccessCore® SDK offers a comprehensive standard software environment for development, testing, and deployment of applications on Kalray TC4 card. It includes:

- Kalray Neural Networks (KaNN™) for AI inferencing that supports a wide range of models as well as leading AI frameworks. KaNN enables fine-grain performance optimizations of AI workloads on Kalray TC4 card
- Optimized libraries to facilitate development of specific algorithms maximizing the utilization of Kalray hardware resources
- Direct programming capabilities using OpenCL and C/C++ standard programming models
- A complete toolchain based on open-source solutions for code generation (GCC and LLVM compilers), code execution, debug and analysis (GDB), enabling to minimize application development time
- Management tools for efficient deployment and operations of the Kalray TC4 card



contact@kalrayinc.com
kalrayinc.com

KALRAY
THE POWER OF MORE