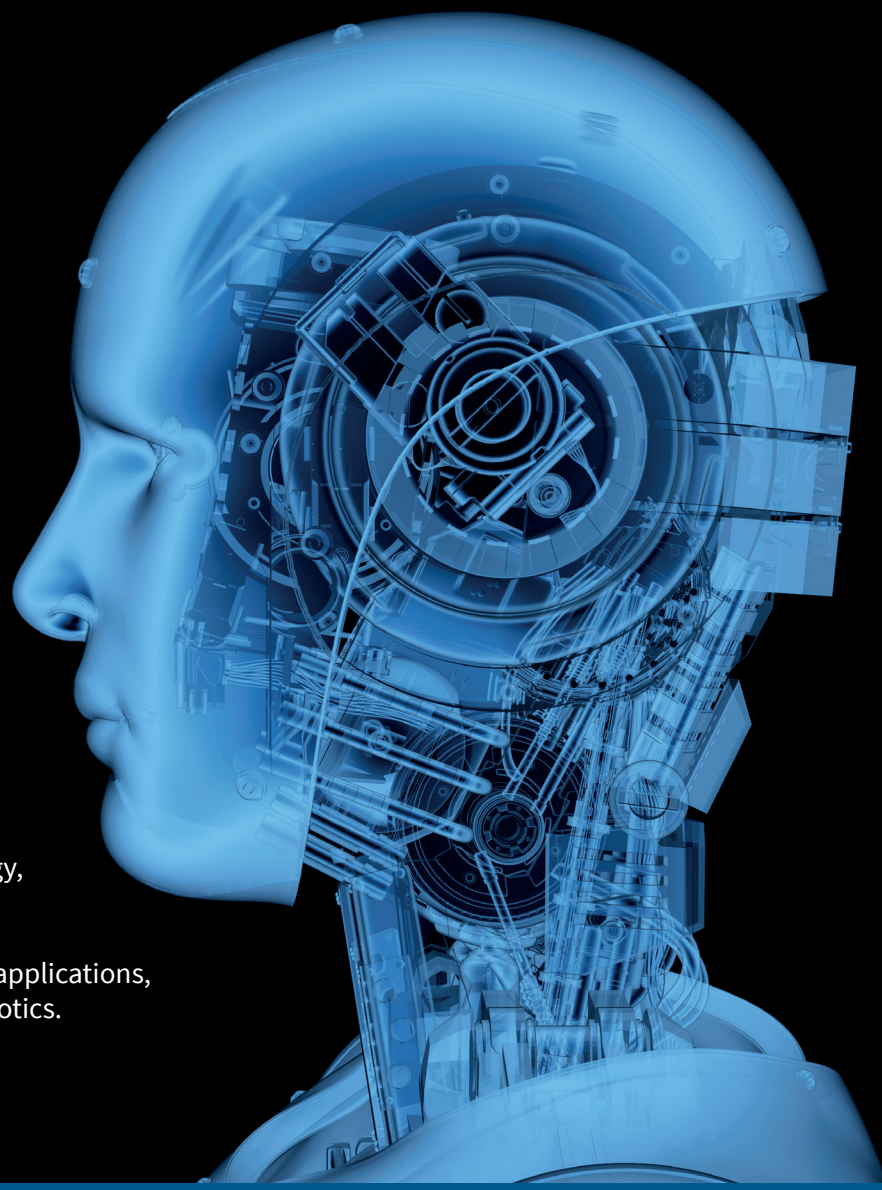
KALRAY

DEEP LEARNING
with KaNN™ Solutions

Kalray Neural Network
for AI applications



In a world where Artificial Intelligence (AI) algorithms have become the new standard in processing, the necessity for high-performance and low-latency processors has risen dramatically. Leveraging its unique parallel manycore technology, Kalray offers an all-in-one deep learning platform, designed to be at the heart of more reliable, more cost-effective and more energy-efficient AI-based applications, for intelligent data centers, self-driving cars or robotics.

The Kalray advantage

Kalray's MPPA® manycore processor has been optimized for managing highly demanding deep learning applications. Along with its processor, Kalray provides the Kalray Neural Network (KaNN™), a tool that allows users to take full advantage of the performance and flexibility offered by the MPPA®'s unique architecture.
For AI-based applications, this combined solution offers high performance, power efficiency and the ability to execute multiple applications in parrallel with freedom from interference.

Kalray's MPPA® processor offers:

- Multi-network parallel processing
- Dedicated high performance co-processor
- Large amount of on-chip memory
- On-the-fly reconfigurability
- Low latency
- High-bandwidth interfaces
- Real-time execution

www.kalrayinc.com

KaNN™
Kalray Neural Network

FRAMEWORKS

Caffe · TensorFlow · ONNX

Thanks to a dynamic network topology, KaNN™ is compatible with any framework.

NETWORKS

GoogLeNet, ResNet, Yolo V3, etc

- **Customizable**
  Inherent system adaptability enables users to add new layers that meet their needs.
- **Modular design**
  Functional partitioning allows users to run their preferred, custom deep learning network.

KaNN™
Kalray Neural Network

- **Built-in inference code generator**
  For high level optimization of the graph
- **Run-time library**
  Low level optimization to leverage MPPA® micro architecture
- **Easy to use**
  Simplified prototyping and accelerated CNN development

- **Low latency**
  Leveraging the MPPA®'s unique parallel processing capabilities, KaNN™ makes deep learning inference faster than ever.

- **Multi-application**
  Thanks to the native freedom from interference of MPPA® architecture, run deep learning networks while simultaneously process other applications, without sacrificing performance or reliability.

Servers and Data centers · Autonomous Vehicle · 5G Telecom Infrastructure · Industry 4.0, Robotics · Aerospace, Drones · Healthcare

contact@kalrayinc.com
www.kalrayinc.com

KALRAY