KaNN™ Low Latency Neural Networks Inference Solutions

- Kalray MPPA®
  Massively Parallel Processor Array

December 2020 – Confidential Information

www.kalrayinc.com

Agenda

1. Kalray in a Nutshell
2. MPPA® at the Heart of Next Gen. Cars
3. Market Deployment
4. Hardware Overview
5. Software Overview
6. Schedule
7. Conclusion

KALRAY IN A NUTSHELL

Kalray offers a new type of processor targeting the booming market of intelligent systems.

A Global Presence

- France (Grenoble, Sophia-Antipolis)
- USA (Los Altos, CA)
- Japan (Yokohama)
- Canada (Partner)
- China (Partner)
- South Korea (Partner)

Leader in Manycore Technology

3rd generation of MPPA® processor

~€85m R&D investment

30 Patent families

Industrial investors

NXP

RENAULT NISSAN MITSUBISHI

SAFRAN

MBDA

EURONEXT

- Public Company (ALKAL)
- Support from European Govts
- Working with 500 fortune companies

*Financial investors: CEA Investissement, Bpifrance, ACE, INOCAP Gestion, Pengpai

INTELLIGENT SYSTEMS / EDGE COMPUTING

At the Heart of Next Decade Industry

Next Gen. Embedded Systems

Next Gen. Data Center

Compute and AI Intensive Critical Systems

MPPA® Processors

PCIe Cards & Modules

Acceleration Solutions for Storage, Networking and Compute

STATE-OF-THE ART SUPPLY CHAIN

High-Quality Partners

tsmc

GUC Global Unichip Corporation

wistron

NXP

Built for High Volumes & High Quality Products

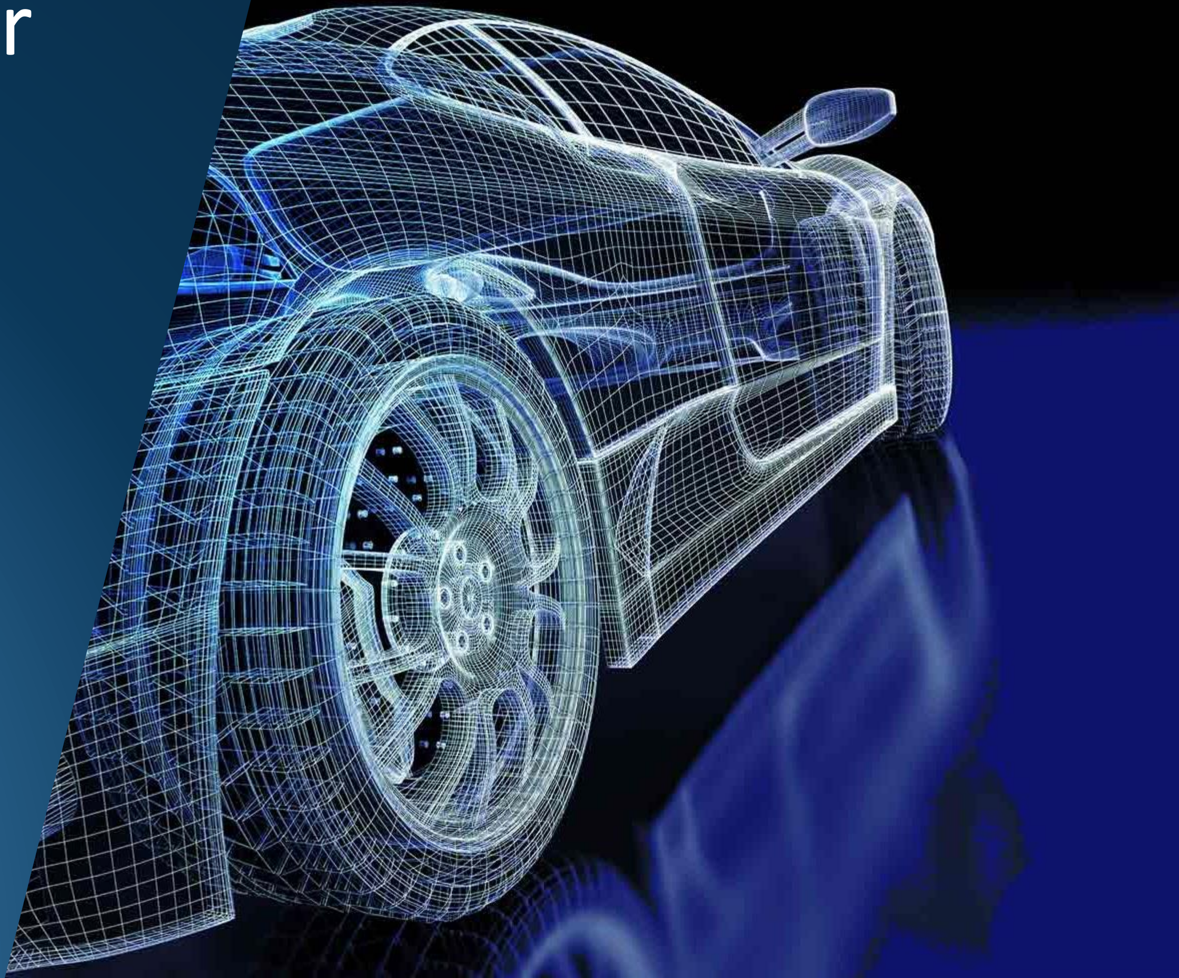
- Fabless model – fully scalable
- Qualified Global Supply Chain
- Support highly demanding customers in Automotive & Aerospace
- Partnership with NXP on safety-centric systems and technologies

CERTIFIED ORGANIZATION ISO 9001

iol Certified by the University of New Hampshire InterOperability Laboratory

CERTIFICATION IN PROGRESS ISO 26262

MPPA® Processor



MPPA®

The Processor at the Heart
of Intelligent Systems

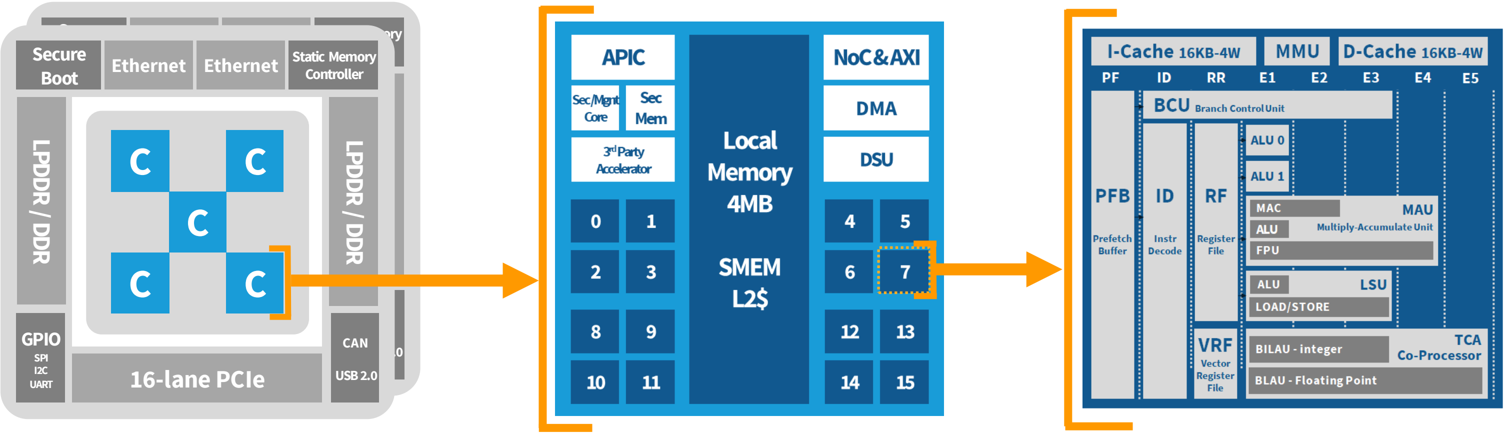
MPPA® PROCESSOR FAMILY & ROADMAP

SAMPLES AVAILABILITY		2019	2020 (IP3) / 2021 (IC)	2021 (IP4) / 2022 (IC)
	BOSTAN2	COOLIDGE1-80	COOLIDGE2 – 80	DOLOMITES (CLUSTER IP4)
PROCESS	28 nm	16 nm	16 nm	UNDER NDA
FIXED POINT OPERATIONS	1.3 TOPS	25 TOPS (8bit)	50 TOPS (8bit)*	
FLOATING POINT OPERATIONS	512 GFLOPS	4 TFLOPS (16bit)	12 TFLOPS (16bit)*	
DMIPs	250 KDMIPS	190 KDMIPS	190 KDMIPS	
CONSUMPTION (Typ.)	8 – 25W	25W / 30W	30W / 40W	
FEATURES	• 288 Kalray VLIW Cores • 128 Crypto Copro • 2xDDR3 • 8x 1/10G GbE • 2xPCle 8 lane Gen3	• 80 Kalray 64-bit cores • 80 Co-processors for vision and learning • 2 x LP/DDR4 • 8x 1/10/25GbE • 16-lane PCle Gen4	• 80 Kalray 64-bit cores • 80 Co-processors for vision and learning • 2 x LP/DDR4 • 8x 1/10/25GbE • 16-lane PCle Gen4	
QUALIF/CERTIF	Industrial (-20/+85C°)	• AEC-Q100 / QM	• ASIL B / ISO 26262	
TARGET MARKET	• DATA CENTER • AUTO (proto)	• DATA CENTER • AUTOMOTIVE	• DATA CENTER • AUTOMOTIVE	
		AVAILABLE (IC and IP)	UNDER DEVELOPMENT (IC and IP)	UNDER DEFINITION

MPPA® COOLIDGE™ SCALABLE APPROACH

5 Cluster Implementation

PATENTED



MANYCORE PROCESSOR

Architecture updates

- 80 or 160 CPU cores
- 600 to 1200 MHz frequency modes

Memory

- L2 refill in DDR and Direct access to DDR from clusters

COMPUTE CLUSTER

Architecture updates

- 16 CPU 64-bit cores
- 16 Co-processors
- Safety/Security 64-bit core

Memory

- L1 cache coherency configurable)
- 4MB memory (BW > 500 GB/s)

3RD GENERATION VLIW CORE

Architecture updates

- 64-bit core
- 6-issue VLIW architecture
- MMU + I&D cache (16KB+16KB)
- 16-bit/32-bit/64-bit IEEE 754-2008 FPU
- Vision/CNN Co-processor (TCA)

ARCHITECTING TIGHTLY COUPLED ACCELERATOR
for CNN and Computer Vision Acceleration

Leverage MPPA® Architecture

- Extend core ISA with « generic » SIMD extension
- 64x 256-bit wide vector register file
- Matrix-oriented arithmetic operations (CNN, CV …)

Leverage MPPA® memory architecture

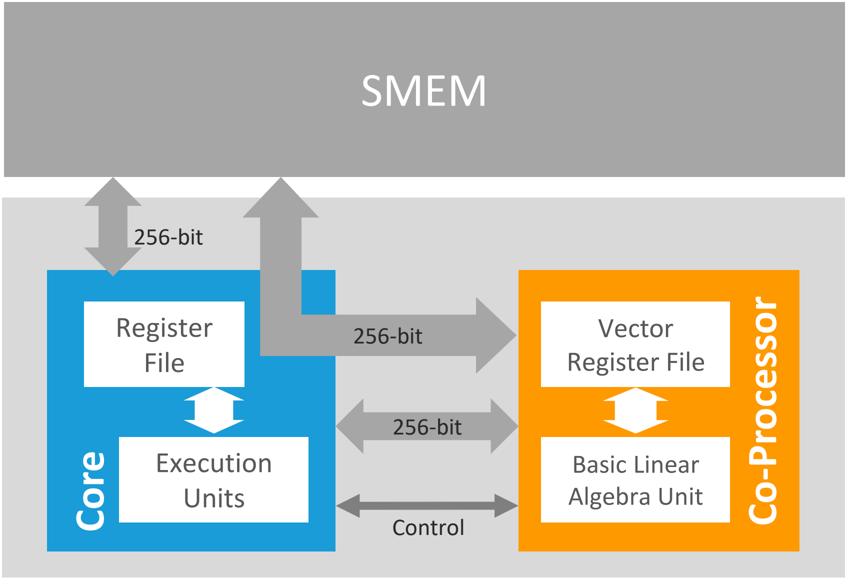
- Large generic internal memory (SMEM)
- Low latency, directly accessible from cores and Copros
- Optimized Memory stream alignment operations

Full integration into core instruction pipeline

- Move instructions supporting matrix-transpose
- Proper dependency / cancel management

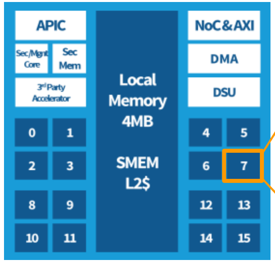
High Arithmetic performances

- 16 x FP16 → FP32 FMA/cycle
- 64 x INT16 → INT64 MAC/cycle
- 128 x INT8 → INT32 MAC/cycle

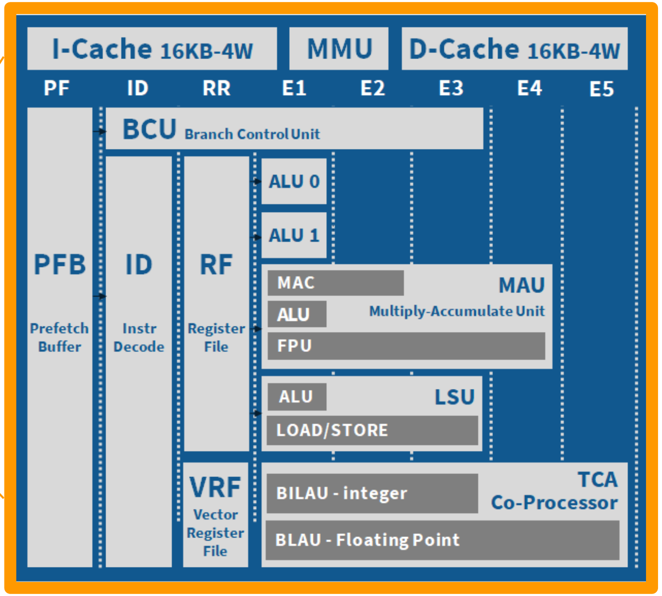


A UNIQUE TIGHTLY COUPLED ARCHITECTURE

For a Large Spectrum of Arithmetic Use Cases



Compute Cluster

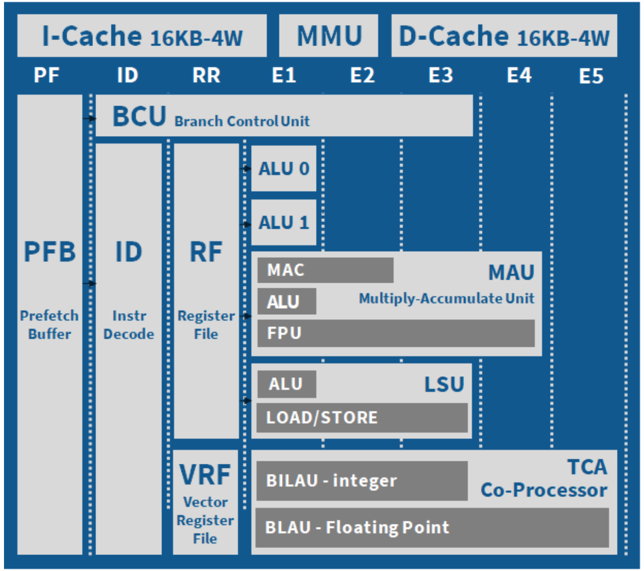


Core and Co-processor

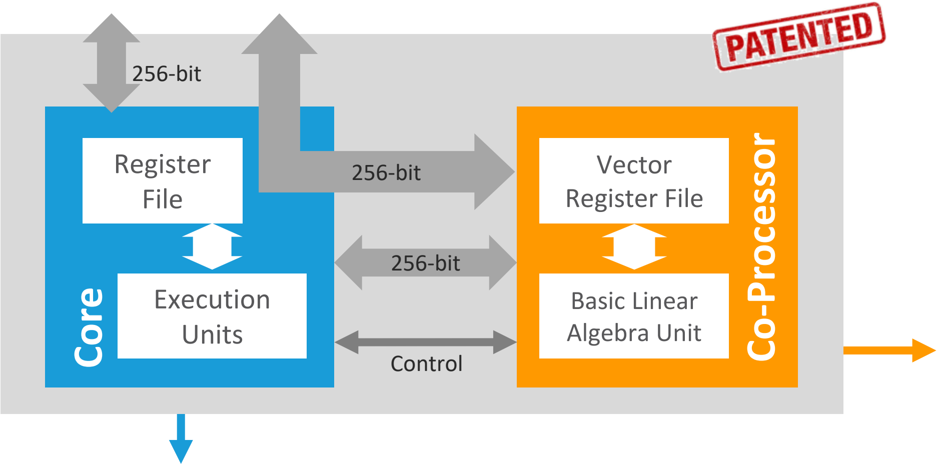
- CNN
- Computer Vision
- LAPACK
- BLAS
- FFTW

Much more than AI!
Multiple arithmetic applications

2 UNITS FOR EXTENDED ARITHMETIC CAPABILITY



2 UNITS FOR EXTENDED ARITHMETIC CAPABILITY



Multiple units

- Various formats
- Various properties
- Covering different use cases

Co-processor unit

- Heavy lifting operators
- Matrix registers (256-1024 bits)
- Matrix operations
- Large vector operation

Floating-point unit

- Flexible operators
- Linear vector register (64 and 128-bit)
- Medium-sized vector operations

INCLUDES COOLIDGE™ CO-PROCESSOR

High Perf/Low Power Tightly Coupled Accelerator (TCA)

Coolidge™ TCA co-processor PATENTED

UP TO
256
MACs
per cycle

ELEMENTARY FUNCTION

- **Square Root**
- **Reciprocal Function**
- **Inverse Square Root**

**VISION
NEURAL NETWORK**

DOT PRODUCT 4

4DP4_fp16_fp32
16/32 DP4_int16_int64
16/32 DP8_int8_int32

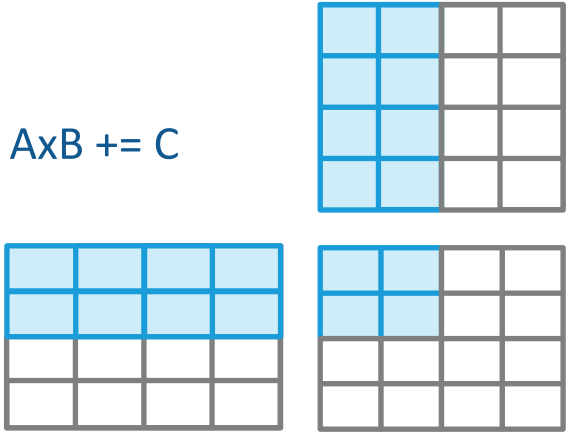
A unique & patented TCA co-processor tailored for Kalray’s core with optimal performance, bandwidth, high processing computing.

MPPA®3 V1: TCA V1 MATRIX OPERATIONS

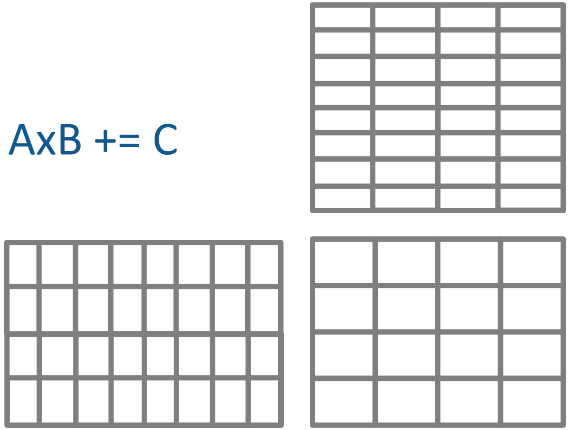
- **16-bit fp convolutions:**
  $(2x4)_{fp16} \cdot (4x2)_{fp16} \mathrel{+}= (2x2)_{fp32}$

- **16-bit int convolutions:**
  $(4x4)_{int16} \cdot (4x4)_{int16} \mathrel{+}= (4x4)_{int64}$

- **8-bit int convolutions:**
  $(4x8)_{int8} \cdot (8x4)_{int8} \mathrel{+}= (4x4)_{int32}$

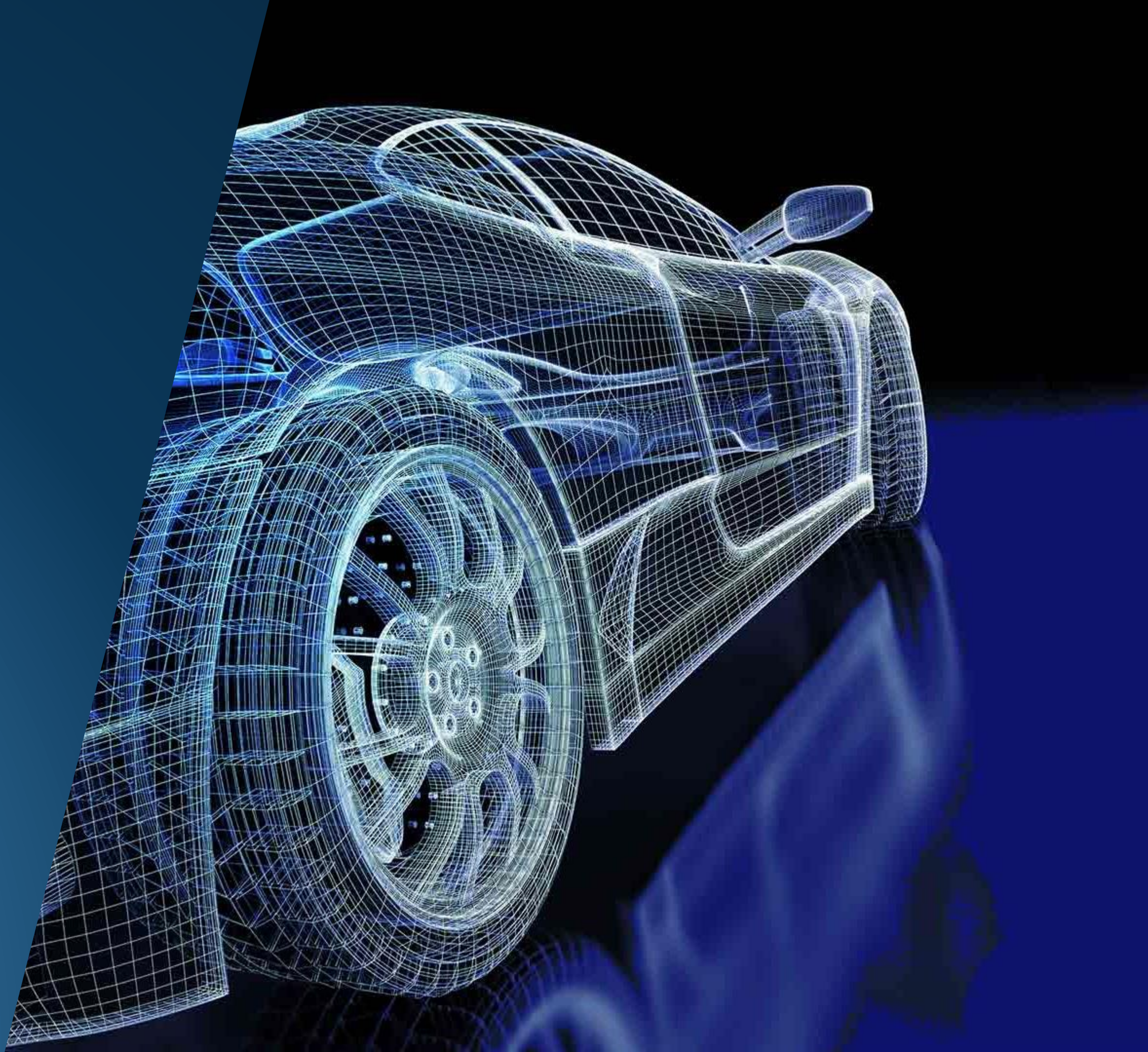


AxB += C



AxB += C

KaNN™ Solution



MPPA®

The Processor at the Heart
of Intelligent Systems

ACCESSCORE® SOFTWARE SUITE

A Complete Toolchain & Set of Libraries

	ACCESSCORE® SDK	ACCESSCORE® RUNTIME	OPTIMIZED LIBRARIES
STANDARD PROGRAMMING ENVIRONMENT **(C/C++/OpenCL)**	**COMPILER, SIMULATOR, DEBUGGER & SYSTEM TRACE**	**OPERATING SYSTEMS & LIBRARIES** **(Linux / ClusterOS)**	**DEEP LEARNING MATHEMATICS COMPUTER VISION**
POSIX THREAD, OPENMP OpenCL Eclipse	GCC , GDB, LLVM Optimization	Optimized exokernel, Open Source RTOS, Linux, Communication Libs	OPEN CV, BLAS, LAPACK CNN Code Gen.

AccessCore®
for a seamless
integration

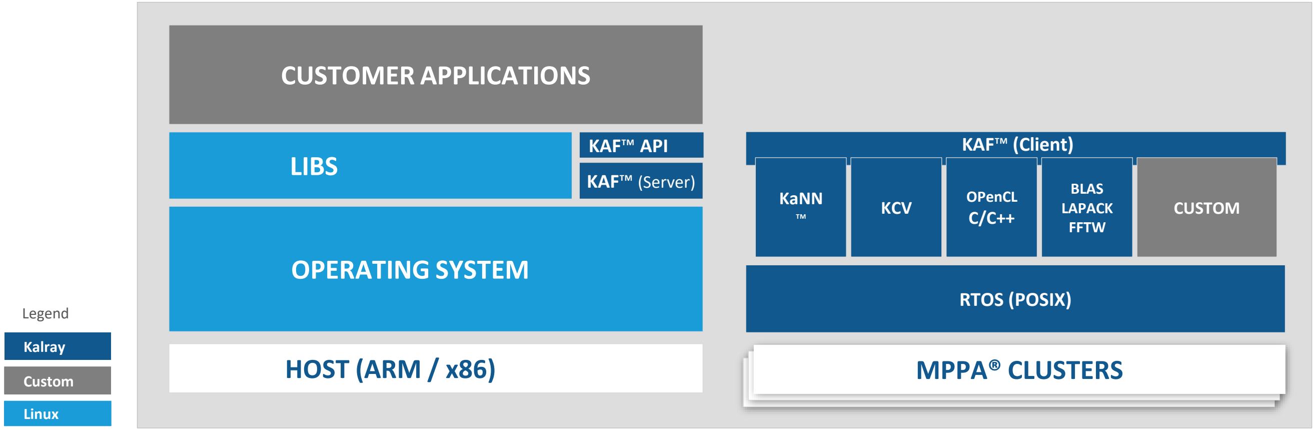
SDK

KAF™	Software framework targeting high computing function offloading
KANN™	CNN Code generator tool compatible with standard CNN frameworks (Caffe , TensorFlow ...)
3rd Party OS	RTOS
3rd Party Tools	Model-Based Development

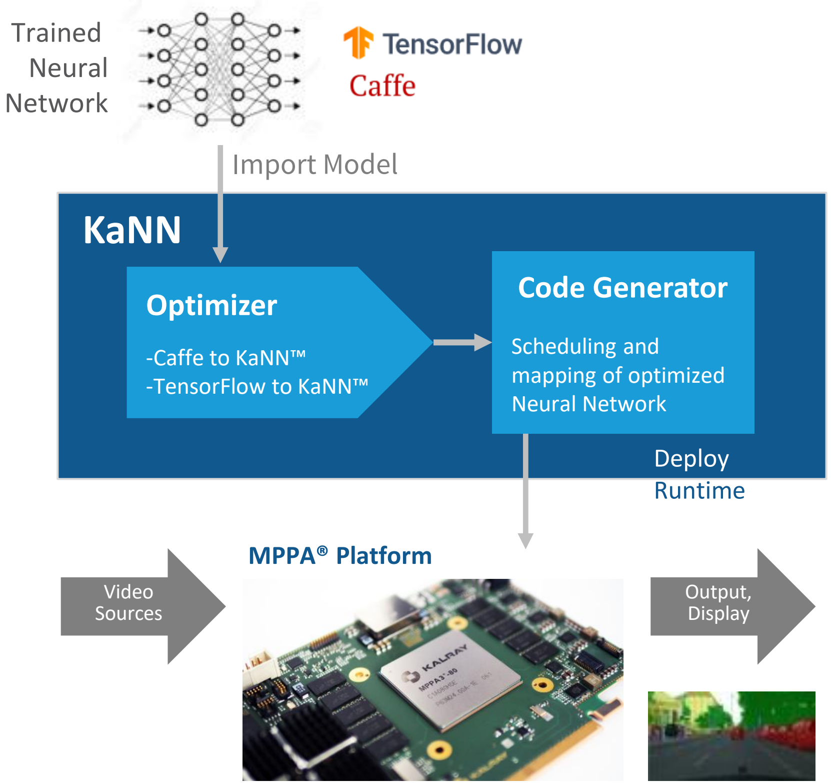
KAF™
Software Kalray Acceleration Framework

KAF™, for easy integration with host system

- Facilitated integration within customer environment
- Usage of Open Standards and expose Open Standards APIs
- Provide capability to add custom high compute processes (C/C++/OpenCL™)



KaNN™
Kalray Neural Network



Trained Neural Network

TensorFlow

Caffe

Import Model

KaNN

Optimizer

-Caffe to KaNN™
-TensorFlow to KaNN™

Code Generator

Scheduling and mapping of optimized Neural Network

Deploy Runtime

MPPA® Platform

Video Sources

Output, Display

- **Deep Learning Inference Code Generator**
  - Optimization of neural networks for MPPA®
  - Deployment of neural networks on MPPA®
- **Deep Learning Inference Runtime**
  - Image-based neural networks running on MPPA®
- **Support of major frameworks**

  Caffe | TensorFlow
- **Support of major networks**
  - Classification | Detection | Segmentation

A comprehensive Neural Network offer, from standard CNN frameworks to code generation, setup & multiple CNN execution

KaNN™
Inference Code Generator

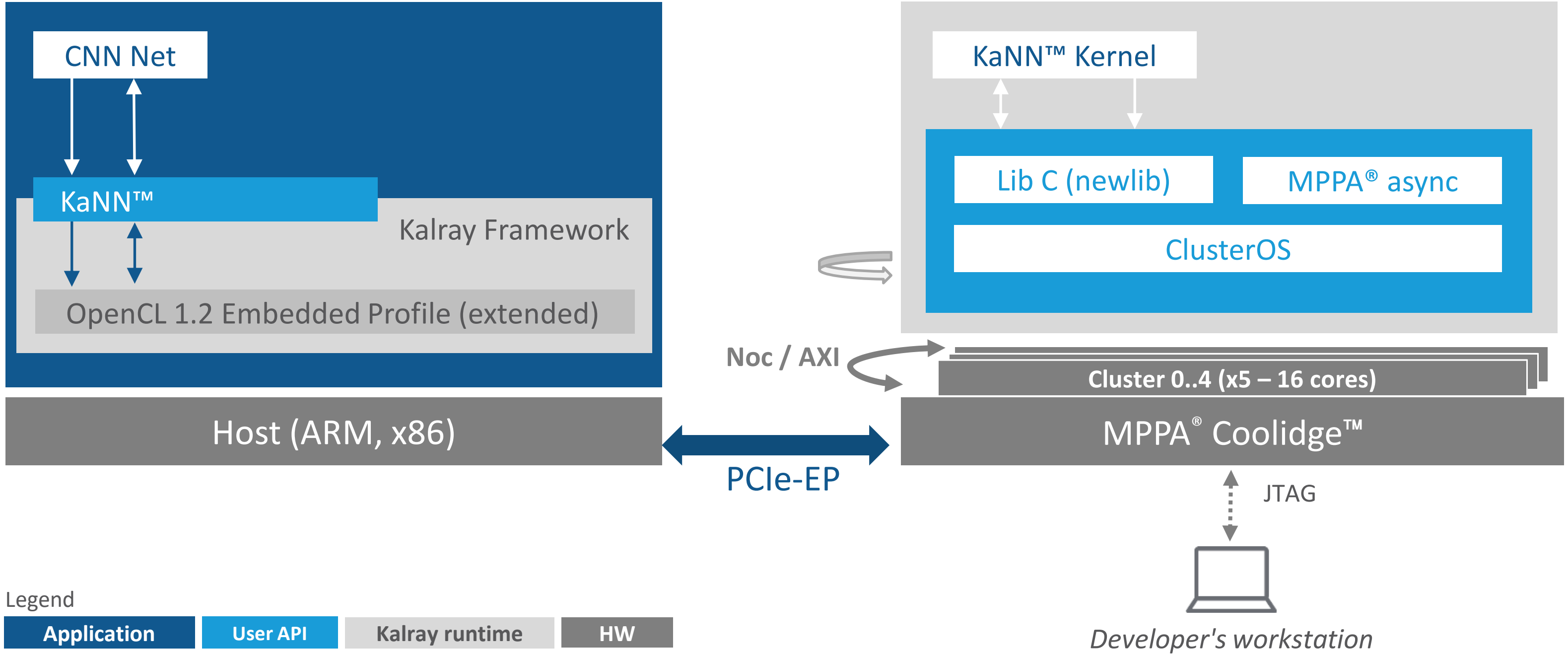
KaNN™ **Optimizer**

- From IR generated by parsing model
- Manages local memory allocation
  - Selecting data layout
  - Splitting output images
  - Inserting Global transfers
  - Inserting pre-fetching
- Merging / Simplifying layers
  - Fusion of ReLu layers
- Convolutions padding
- Dummy Quantization
- Copy collapse
- Scale layers folding
- Fusion of Element-wise layers

KaNN™ **Compiler**

- Memory allocation
- Scheduling
- Command buffer generation
- Parameters (weights) generation
- Static profiling

KaNN™
Software Architecture



KaNN™
Networks and Layers

Supported Networks

Classification Networks

- ResNet50-ILSVRC2012
- MobileNet_v1-MLPerf
- MobileNet_v2-ILSVRC2012

Bounding box based detection networks

•SSD_MobileNet_v1-MLPerf
•YOLO_v3-COCO
•YOLO_v3_320-COCO

Supported Layers

List of all the currently supported CNN layers, valid for any CNN framework supported by KaNN (TensorFlow and TensorFlow lite, Caffe, ONNX, etc.)

- Convolution / InnerProduct
- Deconvolution
- Depthwise convolution
- ReLU / ReLU6 / pReLU
- Pooling: max, average
- Padding (TensorFlow only)
- Softmax
- LRN
- BatchNorm
- Concatenation (only in depth)
- Element wise: addition, multiplication, subtraction, division, minimum, maximum, multiply-and-add
- Tanh / TanhLeCun
- Scale
- Copy
- Logistic

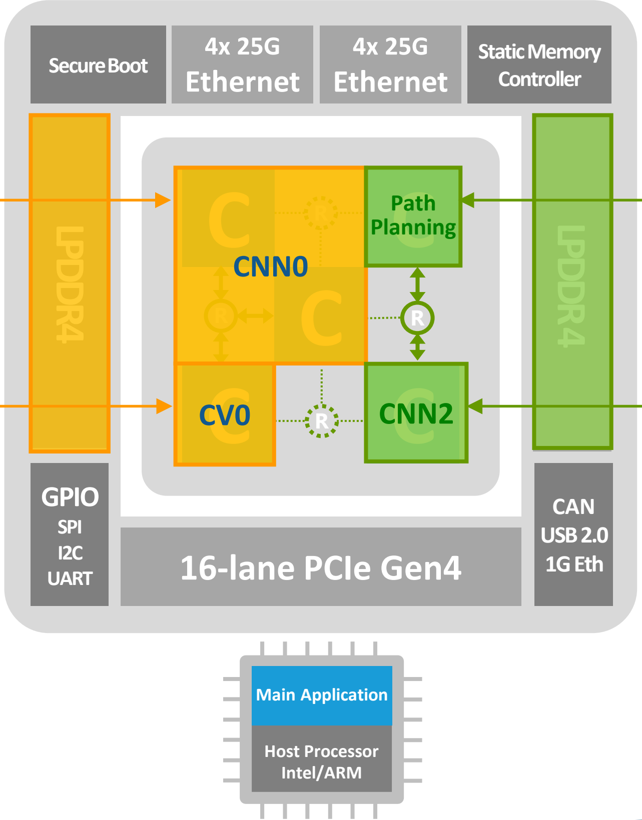
➕ Use our KaNN™ Extensivity tool to implement your own layer !!! (KaNN 4.1)

PERFORMANCE & AGGREGATION ON MPPA® PROCESSOR

Run Multiple High Performance Acceleration Functions Simultaneously

High Performance Compute Accelerations
CNN

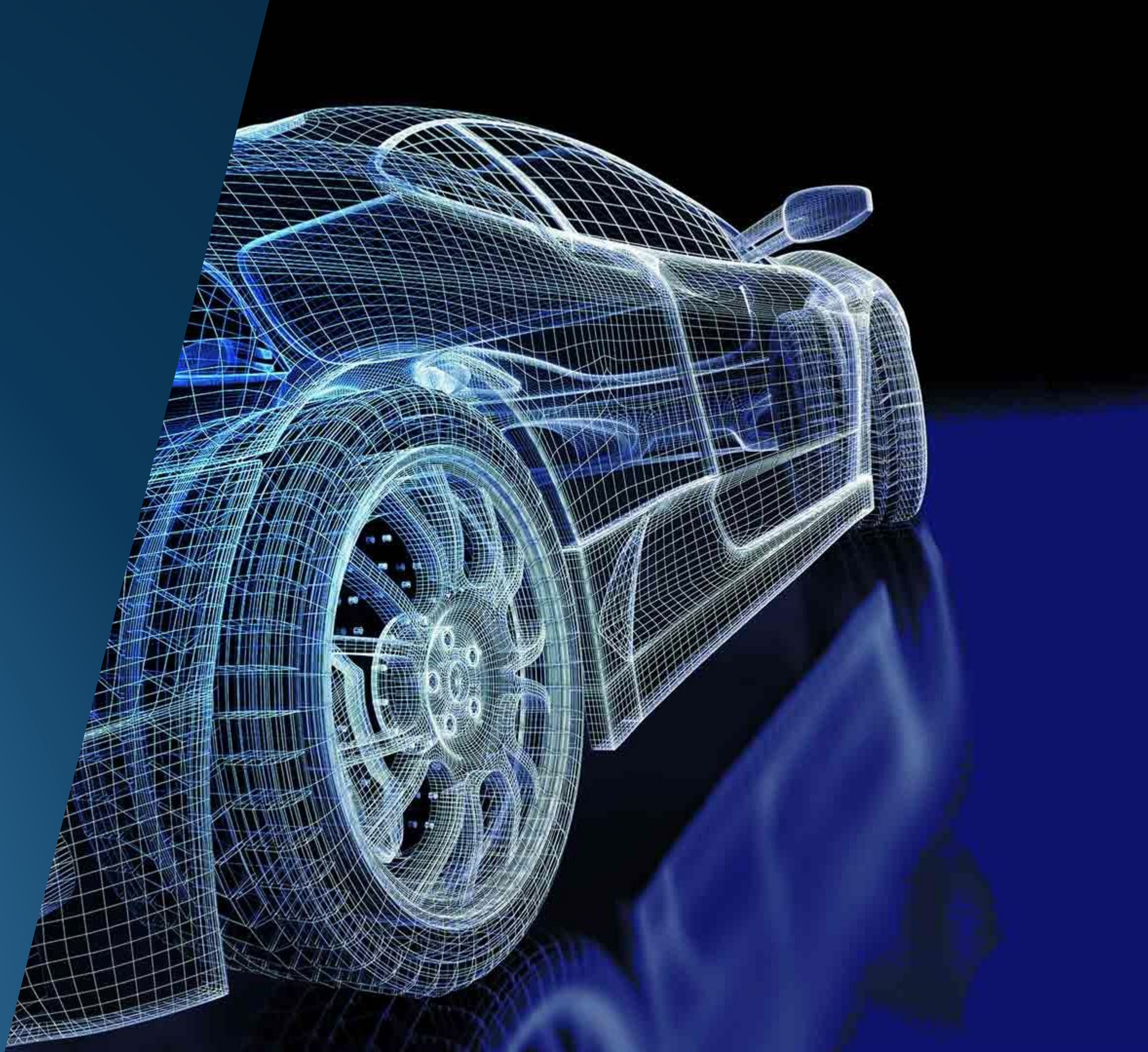
High Performance Compute Accelerations
Computer Vision



High Performance Compute Acceleration
Path Planning

High Performance Compute Acceleration
CNN2

MPPA®
Performance



MPPA®

The Processor at the Heart
of Intelligent Systems

COOLIDGE™ Performances

		Coolidge-80 v1 @1.2 GHz	Coolidge -80 v2 @1.2 GHz
INT8	Core	N/A	N/A
	Copro	24.6 TOPS	49.2 TOPS
	TOTAL	**24.6 TOPS**	**49.2 TOPS**
INT16	Core	2 TOPS	*1.9 TOPS*
	Copro	12.3 TOPS	*12.3 TOPS**
	TOTAL	**14.3 TOPS**	***14.2 TOPS***
FP16	Core	1.15 TFLOPS	*1.15 TFLOPS*
	Copro	3.05 TFLOPS	*12.2 TFLOPS**
	TOTAL	**4.2 TFLOPS**	***13.4 TFLOPS***
FP32	Core	1.15 TFLOPS	*1.15 TFLOPS**
	Copro	N/A	*N/A*
	TOTAL	**1.15 TFLOPS**	***1.15 TFLOPS***
Power		25W	30W

#INFERENCE / WATT PERFORMANCE$^{(1)}$

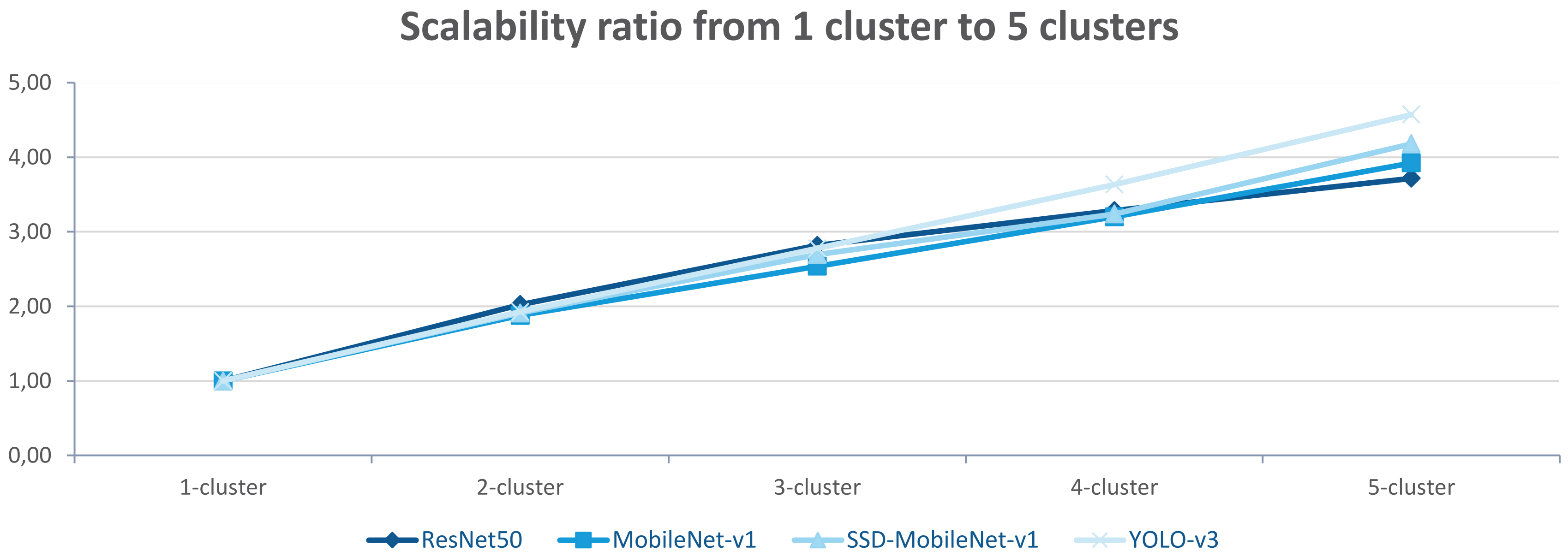
Better #Inference / W ratio

			GoogLeNet	ResNet-50			
Provider	**Product**	**Overall Power**	**FPS/Watt Batch1**	**FPS/Watt batch1**	**Thermal**	**PCIe Interface**	**DDR Capacity (GB)**
NVIDIA Jetson [(1)] (Inference/Device)	Xavier Xavier (2020)	27 W 29 W	38 79	21 48	Passive Passive	Gen4 Gen4	8/16 8/16
Xilinx (Inference)	Alveo 200 Alveo 250	89 W 112 W	35 37	* *	Pass./Active Pass./Active	Gen3+Gen4 Gen3+Gen4	* *
Kalray (Inference)	MPPA3 v1 80 MPPA3 v2 160	25 W 60 W	**80** **127**	**32** **50**	Passive Passive	Gen4 Gen4	4/8/16/32 4/8/16/32

Better Compute Efficiency

				Yolo v3 (416x416)		
Provider	**Product**	**Overall Power**	**FP16 TFLOPS**	**FPS**	**Efficiency**	**DDR Capacity (GB)**
NVIDIA Jetson [(1)] (Inference/Device)	Xavier	30 W	10 + 5	18	10%	8/16 8/16
Kalray (Inference)	MPPA3 v1 80	25 W	3 + 1	20	**35 ~ 40%**	4/8/16/32 4/8/16/32

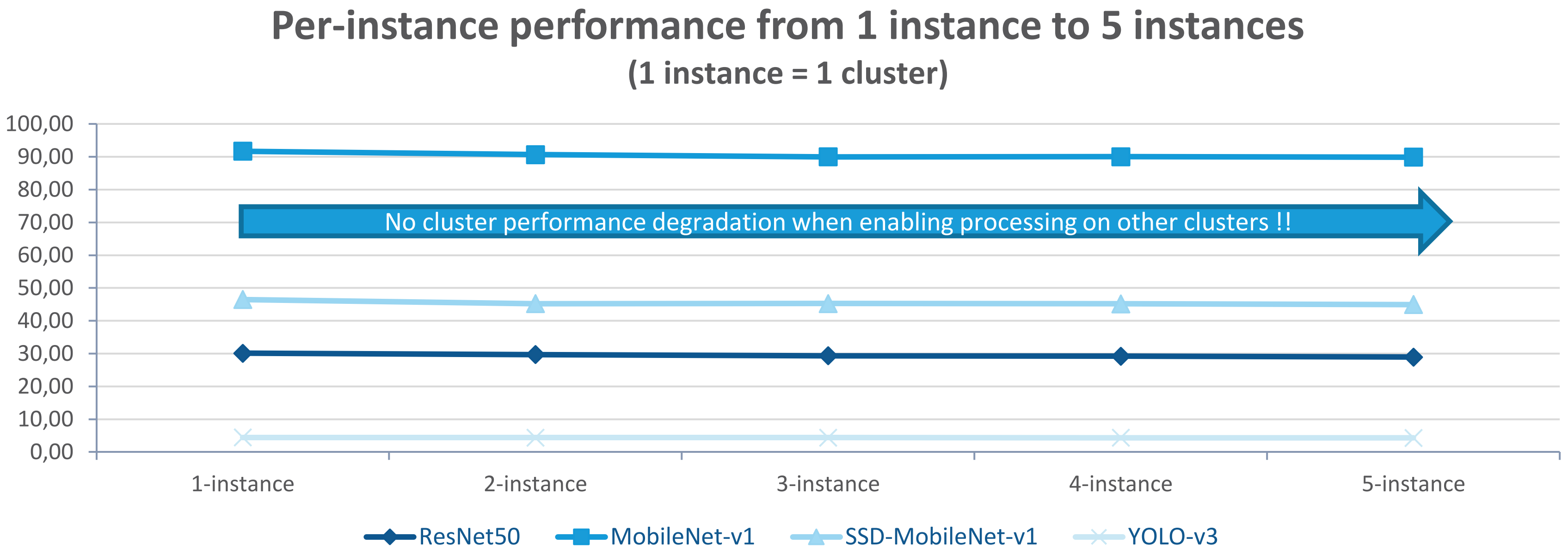
MPPA® COOLIDGE™ SCALABILITY



Scalability depends on computing power:

- ➢ Classification networks show lower scalability than object-detection networks
- ➢ Running small networks on multiple clusters prevent hiding communcations with processing

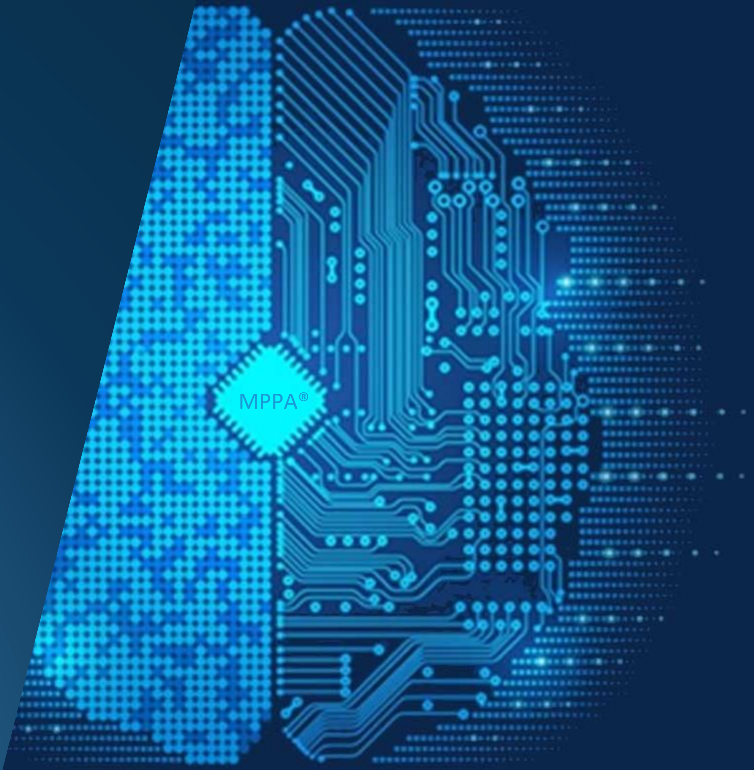
MPPA® COOLIDGE™ - MULTI-INSTANCE UNIQUE CAPABILITY



Allows full flexibility during developments

- ➢ Each application can be developed and optimized independently
- ➢ Heterogeneous applications can be easily executed simultaneously on Coolidge

Practical Generation Flow



MPPA®

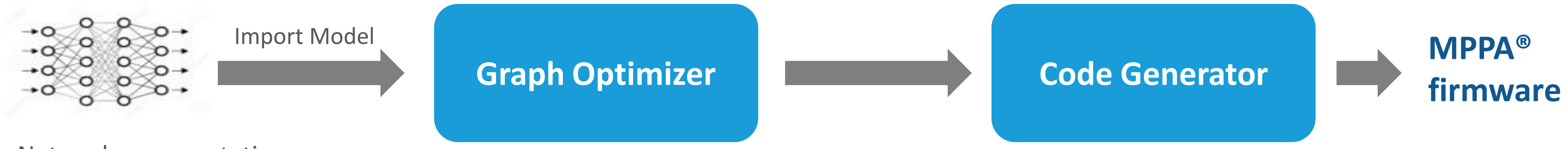
The Processor at the Heart

of Intelligent Systems

KANN™ COMPILATION FLOW (1/2)

I/O and internal formats supported by KaNN™:

- 16 and 32-bit floating-point.
- 8-bit integer (e.g. symmetric quantization with null zero point)



Network representation :

TensorFlow (Lite)
Caffe
N2D2
ONNX (soon)

- Dummy Quantization
- Copy collapse
- Scale layers folding
- ReLu layer merging
- Convolution padding
- Fusion of Element-wise layers

- Memory Allocation
- Scheduling
- Command buffer generation
- Parameter Buffer allocation
- Static Profiling

KANN™ COMPILATION FLOW (2/2)

A race for efficiency

Generic Graph simplifications

- Precision conversion (32 to 16-bit floating-point)
- Copy and concatenation elimination
- Merging layers together
  - ReLu can almost effortlessly be computed at the end of a computation
  - Folding of batch normalizations, scalings, additions, intoa single point-wise fused multiply-add operator

Maximizing use of efficient kernels

- Move part of the control flow out of the execution (into the codegen)
- Padding image / params with **zeros** if the gain in regularity exceeds the extra computation

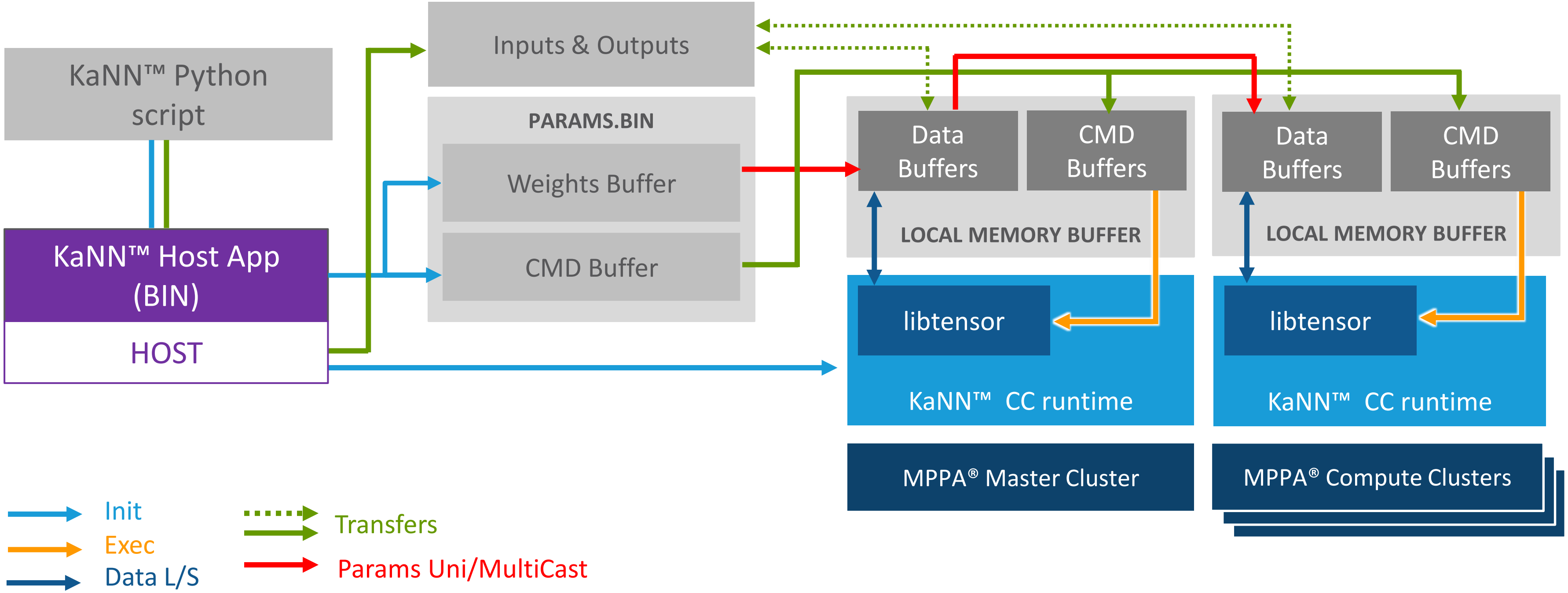
=> Target specific optimization flow

Legalization and kernel selections

- Mapping to **libtensor** kernels : library of tensor computer kernels optimized for MPPA3 core/cluster architecture (TCA, FPU, …)

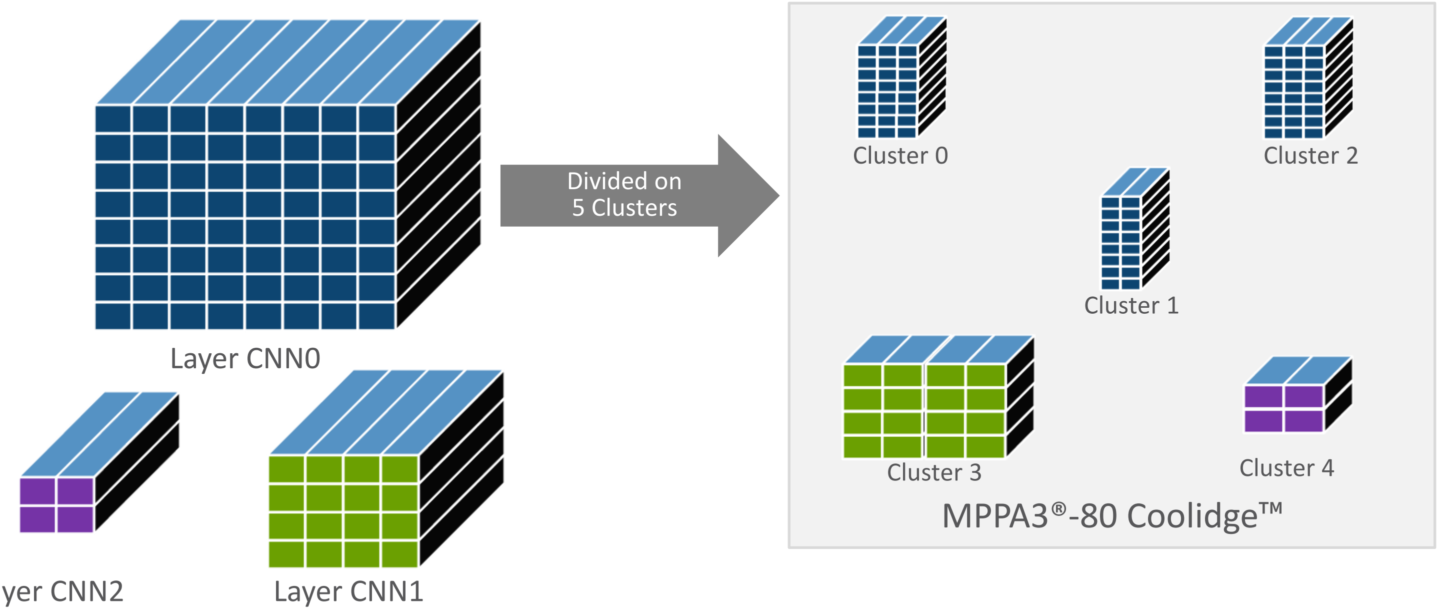
Byte-code generation

KANN™ RUNTIME ARCHITECTURE



MULTI-CNN ON MPPA® COOLIDGE™

Independent execution of multiple networks through spatial partitioning (implemented using OpenCL sub device)



KANN™ PROFILING

Inference Performance Tuning

Static profiling: extracting information during compilation

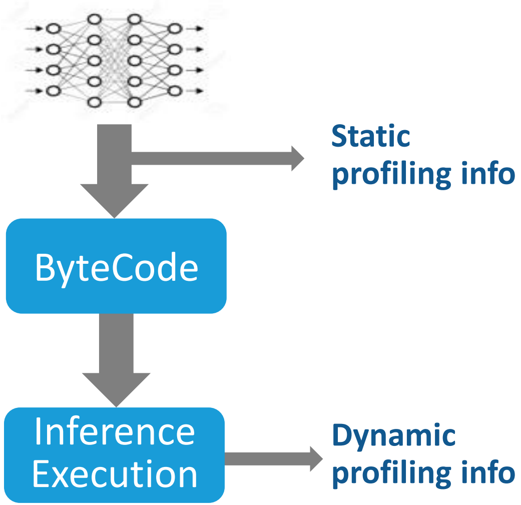
- Number of command executed (by cluster)
- Amount of memory exchanged (from/to DDR, between clusters)
- Evaluation of expected processing latency (future version)

➕ ***Useful to forecast performance and check initial load balancing***

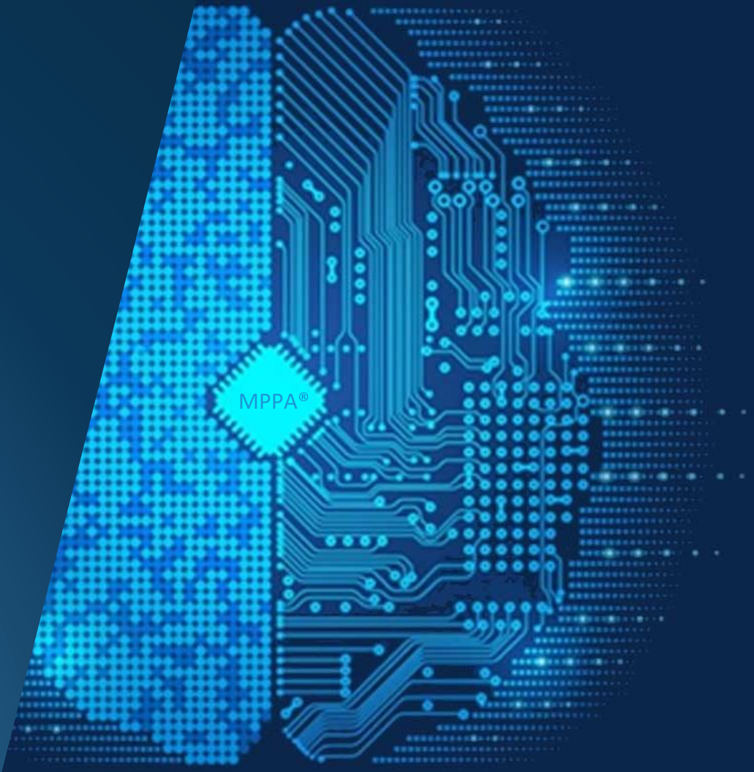
Dynamic profiling: extracting information at runtime

- Accumulated execution time by kernel
- Accumulated execution time by layer (future version)

➕ ***Useful to check real performance and actual load balancing***



Conclusion



MPPA®

The Processor at the Heart

of Intelligent Systems

MANYCORE FOR DEEP LEARNING

Optimized architecture for Deep Learning

- For Vision processing
- For Speech recognition
- For Deep learning

High compute Power

- Support both floating or integer operations
- VLIW core with FPU
- Better performance/power ratio than competition

Dedicated co-processor for vision & learning

- x14 more performance

A fully programmable solution

- Enabling Custom Neural Networks
- Floating, Fixed point, Recurrent, …

Leverage on MPPA® internal memory

- High bandwidth memory access
- Data transfer from cluster to cluster
- Limited DDR access for higher performance

Thank You

KALRAY S.A.
Corporate Headquarters
180, avenue de l'Europe
38 330 Montbonnot, France
Phone: +33 (0)4 76 18 90 71
contact@kalrayinc.com

CERTIFIED ORGANIZATION ISO 9001

CERTIFICATION IN PROGRESS ISO 26262

KALRAY INC.
America Regional Headquarters
4962 El Camino Real
Los Altos, CA - USA
Phone: +1 (650) 469 3729
contact@kalrayinc.com

KALRAY JAPAN - KK
Represented by MACNICA Inc. Strategic Innovation Group
Macnica Building, No.1, 1-6-3 Shin-Yokohama
Kouhoku-ku, Yokohama 222-8561, Japan
Phone: +81-45-470-9870

KALRAY S.A.
Sophia-Antipolis
1047 allée Pierre Ziller
Business Pôle – Bâtiment B, Entrée A
06560 Sophia-Antipolis, France
Phone: + 33(0) 4 76 18 09 18

www.kalrayinc.com

DISCLAIMER

Kalray makes no guarantee about the accuracy of the information contained in this document. It is intended for information purposes only, and shall not be incorporated into any contract. It is not a commitment to deliver any material, code or functionality, and should not be relied upon in making purchasing decisions. The development, release and timing of any features or functionality described for Kalray products remains at the sole discretion of Kalray. Trademarks and logos used in this document are the properties of their respective owners.