MPPA® Manycore Processor: At the heart of Intelligent Systems

Benoît Dupont de Dinechin, CTO

June 2019

www.kalrayinc.com

Kalray at Glance

We design processors at the heart of new intelligent systems

3

sites

Grenoble (France) ,
Los Altos (USA), Tokyo (Japan)

~85

Staff Members

~ 75 engineers & PhDs

23

patent families

Financial and industrial shareholders

cea investissement AMORÇAGE TECHNOLOGIQUE
INOCAPGestion
MBDA MISSILE SYSTEMS
RENAULT NISSAN MITSUBISHI
ACE PRIVATE EQUITY
Pengpai
SAFRAN
bpifrance

EURONEXT

~ 48M€

raised at IPO in June 2018

Outline

Intelligent Systems

Manycore Processors

Kalray MPPA® Processors

Deep Learning Inference

Model-Based Design

Applications & Outlook

Intelligent Systems

Cyber-Physical Systems

- Information processing and physical processes are tightly integrated
- Time constraints associated with information manipulation
- Distributed systems
- Functional safety
- Cyber-security

Intensive Computing

- Numerical computing
- Signal processing
- Image processing
- Graph computing

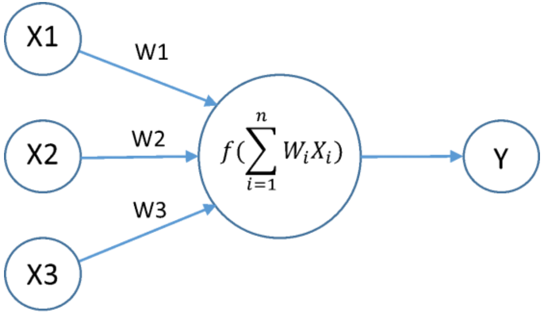
Artificial Intelligence

- The science and engineering of creating intelligent machines (J. McCarthy, 1956)
- Mostly represented by the Machine Learning field, in particular Deep Learning
- Association causation level: “the objective of curve-fitting is to maximize fit, while deep learning tries to minimize over-fit” (J. Pearl 2018)

Machine Learning (ML)

Give computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

- **Rule Extraction** Goal is to identify statistical relationships in data
- **Clustering** Group similar data together, while increasing the gap between the groups
- **Classification & Regression** Map a set of new input data to a set of discrete or continuous valued output, respectively
- **Artificial Neural Networks (ANN)** General implementation model for nonlinear classifiers, trained by using back-propagation algorithms



$$\boldsymbol{Y} = \boldsymbol{f}\Big(\overbrace{(W_1 \times \boldsymbol{X_1}) + (W_2 \times \boldsymbol{X_2}) + (W_3 \times \boldsymbol{X_3})}^{\text{Weighted sum of inputs}}\Big)$$

$\boldsymbol{f}(\,)$ is the “activation function”

$f(x) = \max(0, x)$, “Rectified Linear Unit”

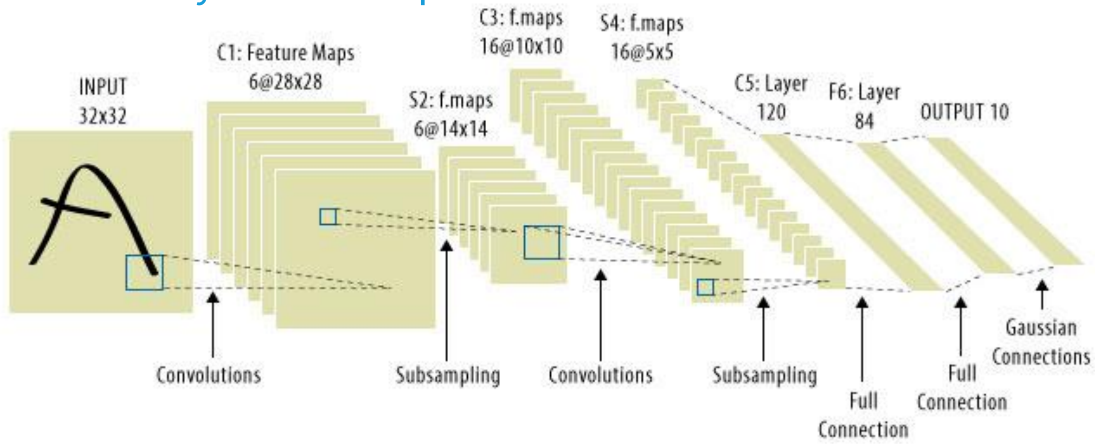
$f(x) = \tanh(x)$

$\ldots$

Deep Learning (DL)

Computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction (Yann Le Cun et al., 2015)

- **Convolutional Neural Networks (CNN)** Networks where most filtering operations performed by feature maps are discrete convolutions



- **Recurrent Neural Networks (RNN)** Networks with feedback loops

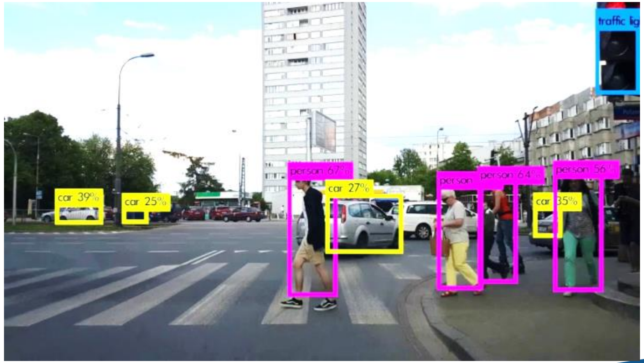
Machine Learning Steps

Training (datacenter)

- Learning part or Machine Learning
  - Supervised (classification & regression)
  - Unsupervised (clustering)
  - Reinforcement (decision-making)
- Off-line processing of large data sets
- Floating-point 32-bit arithmetic

Inference (intelligent system)

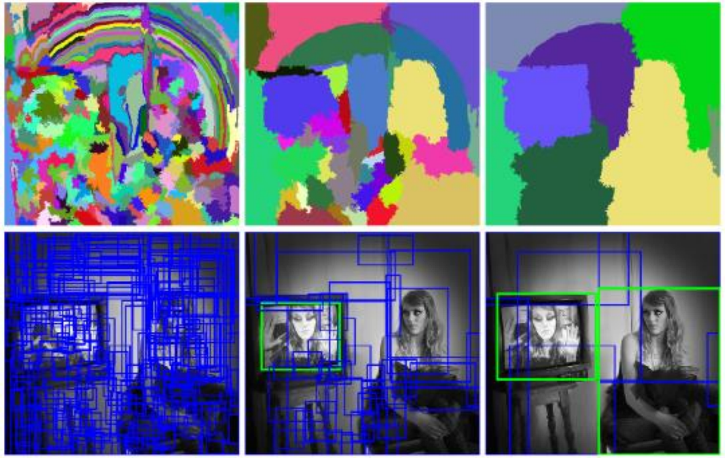
- Classification / Segmentation / Detection
- On-line / Real-time data stream processing
- Floating-point 16.32-bit or “bfloat16” arithmetic
- Integer 8.32 arithmetic (quantization) for CNN

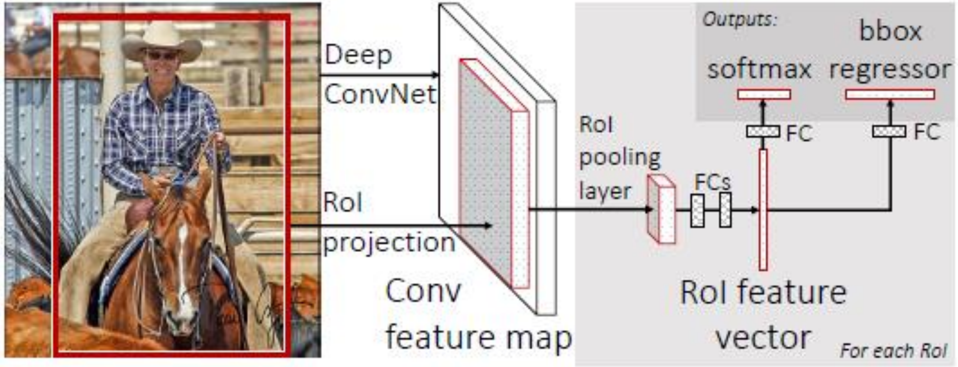


R-CNN, Fast & Faster R-CNN (Girshick & Ren, 2014-2016)

Regional CNN and improvements use two steps for object detection

1) Proposal of candidate regions (initially by segmentation, then by neural computing)
2) Classification of candidate regions (neural computing and refinment steps)

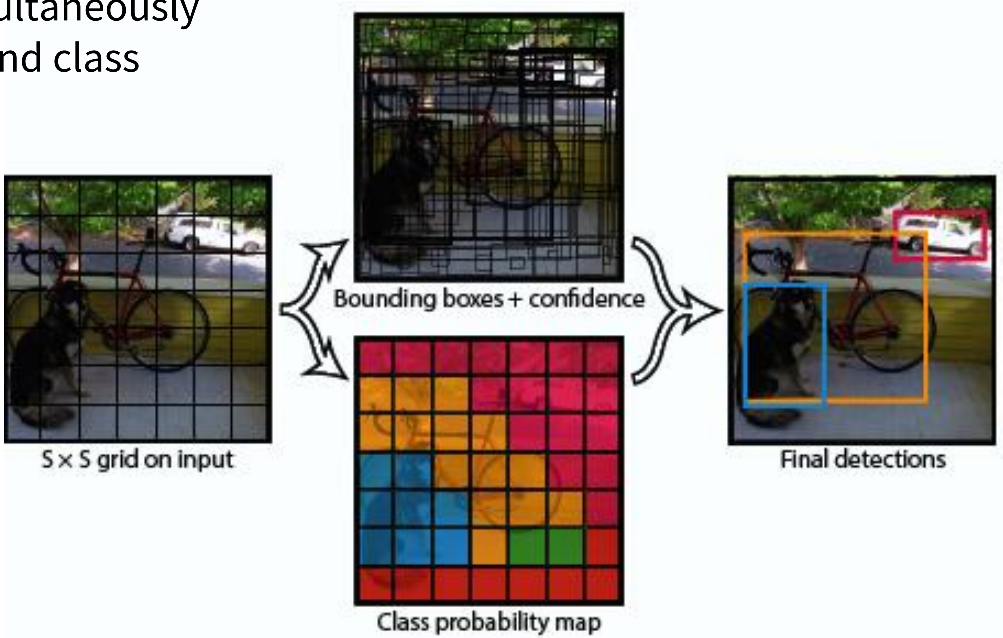




YOLO v1-3 « You Only Look Once » (Redmon 2016-2018)

Single-step method (unlike « R-CNN » family)

- A single convolutional network simultaneously predicts multiple bounding boxes and class probabilities for those boxes



Cyber-Security Requirements

	Defense	Avionics	Automotive
Hardware root of trust (HSM)	✓	✓	✓
Authenticated software	✓	✓	✓
Encrypted boot firmware	✓	✓	
Encrypted application code	✓	✓	✓
Event data record encryption		✓	✓
Secured communication	✓	✓	✓
Physical attack protection	✓		✓

Secure Boot for Trusted Software Deployment

Measured boot

- Enables external agent to attest the platform state after the boot process
- Provides a secure measurement and reporting chain to external agent
- Detect modified boot code, settings and boot paths
- Typically used on servers, by associating UEFI and a TPM chip

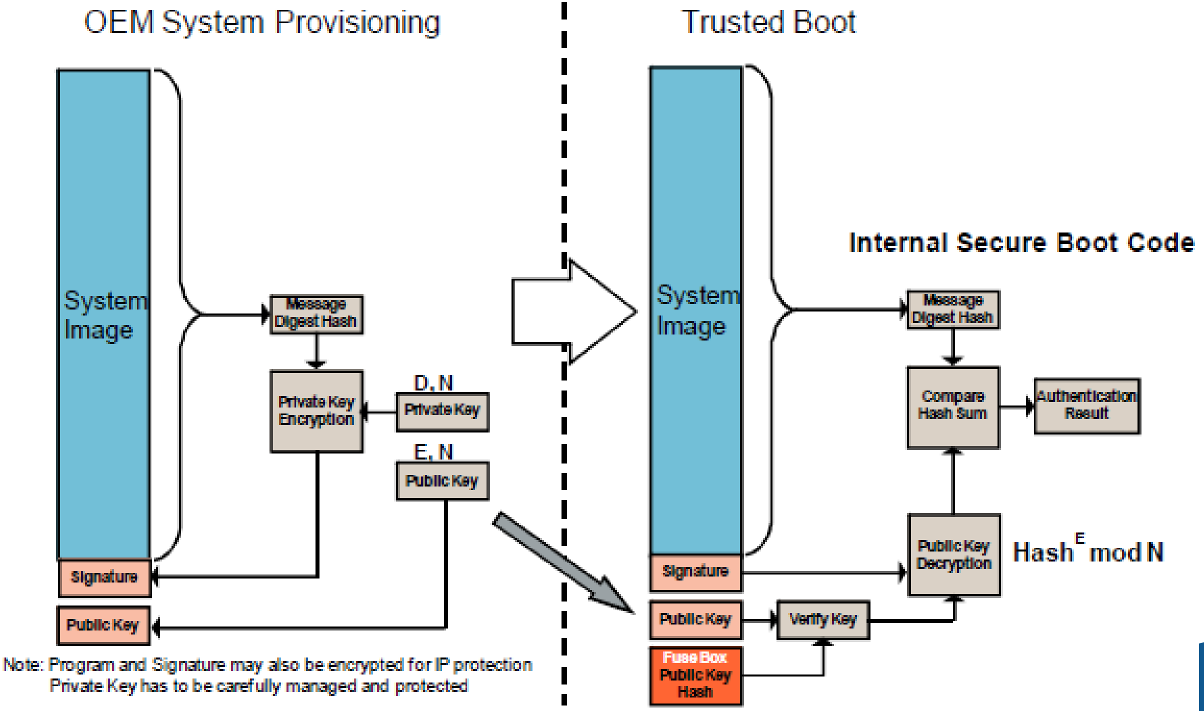
Trusted boot

- Unbroken chain of trust across all stages of the OS boot:
  - Phase-0 (internal ROM) check platform, validate & launch Phase-1
  - Phase-1 (Flash) initialize peripherals, validate & launch Phase-2
  - Phase-2 (network or disk), validate & launch operating system
- Typically used on embedded systems

FreeScale QorIQ Trusted Boot

QorIQ processors target consumer, industrial, medical, networking

- OEM public keys and the intent to secure (ITS) bit in immutable storage (fuses)
- When the ITS bit is set, jump to internal boot ROM (IBR) for Phase-0
- Phase-1 firmware digitally signed using OEM private signature key
- Phase-0 verifies firmware signature using public key



Note: Program and Signature may also be encrypted for IP protection
Private Key has to be carefully managed and protected

Outline

Intelligent Systems

Manycore Processors

Kalray MPPA® Processors

Deep Learning Inference

Model-Based Design

Applications & Outlook

Homogeneous Multicore Processor

Multiple cores sharing a cache-coherent memory hierarchy

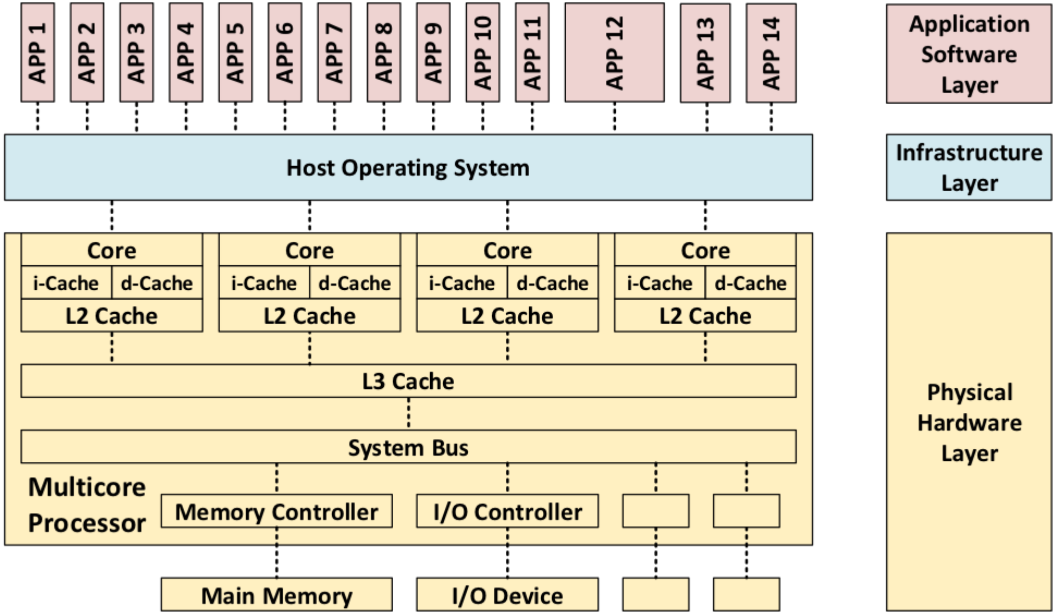
- Private L1 i-cache and d-cache
- Shared or clustered L2 cache
- Shared L3 cache

Application programming

- C/C++, Python, Java
- Pthreads, std::thread, OpenMP
- Rich operating system (Linux)

Application partitioning

- Virtual machine monitor



https://insights.sei.cmu.edu/sei_blog/2017/08/multicore-processing.html

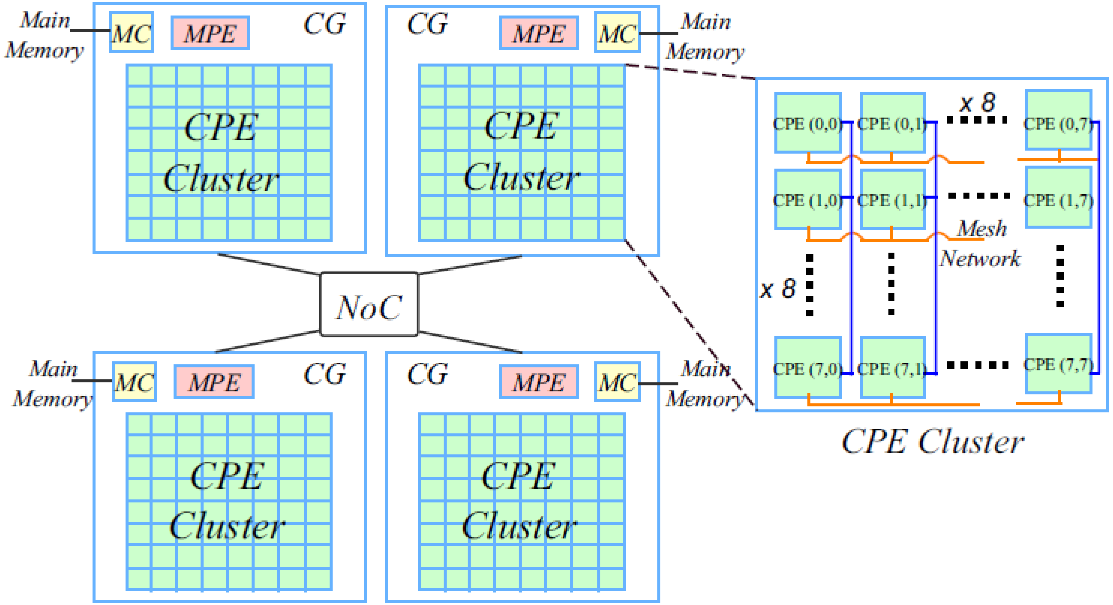
Manycore Processors

Multiple 'Compute Units' connected by a network-on-chip (NoC)

- Group of cores + DMA engine
- Scratch-pad memory (SPM)
- Software-managed caches
- Local cache coherency

SW26010 Manycore processor

- Node of the Sunway TaihuLight supercomputer (#1 TOP 500 in 2016)
- 4 'core groups' with MPE core, CPE core cluster, collective DMA engine
- 64KB SPM per CPE core



Z. Xu, J. Lin, S. Matsuoka, «Benchmarking SW26010 Many-Core Processor» IPDPS 2017

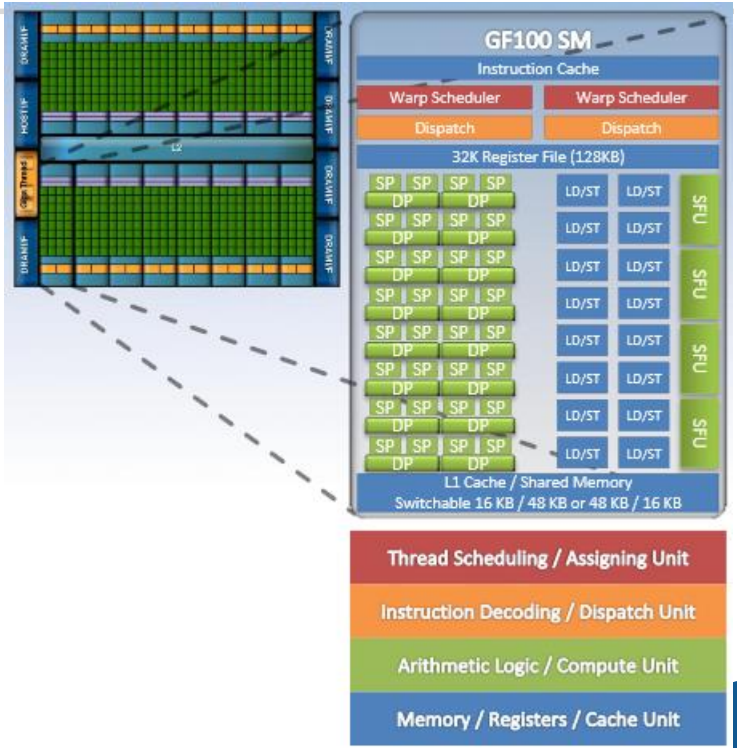
GPGPUs as Manycore Processors (NVIDIA)

Classic GPGPU: NVidia Fermi architecture

- GPGPU ‘compute units’ are called Streaming Multiprocessors (SM)
- Each SM comprises 32 ‘streaming cores’ or ‘CUDA cores’ that share a local memory, caches and a global memory hierarchy
- Threads are scheduled and executed atomically by ‘warps’, which execute the same instruction or are inactive at any given time
- Hardware multithreading enables warp execution switching on each cycle, helping cover memory access latencies

GPGPU programming models (CUDA, OpenCL)

- Each SM executes ‘thread blocks’, whose threads may share data in the local memory and access a common memory hierarchy
- Synchronization inside a thread block by barriers, local memory accesses, atomic operations, or shuffle operations (NVIDIA)
- Synchronization between thread blocks through host program or global memory atomic operations in kernels


DRAM I/F
HOST I/F
Giga Thread
DRAM I/F
DRAM I/F
DRAM I/F
DRAM I/F
DRAM I/F
L2
GF100 SM
Instruction Cache
Warp Scheduler
Warp Scheduler
Dispatch
Dispatch
32K Register File (128KB)
SP
DP
LD/ST
SFU
L1 Cache / Shared Memory
Switchable 16 KB / 48 KB or 48 KB / 16 KB
Thread Scheduling / Assigning Unit
Instruction Decoding / Dispatch Unit
Arithmetic Logic / Compute Unit
Memory / Registers / Cache Unit

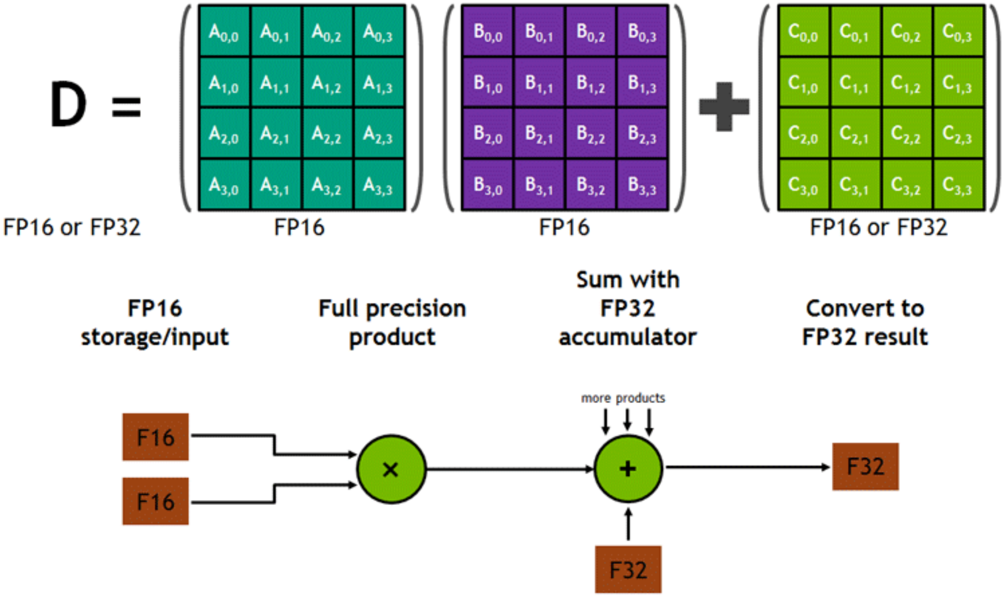
GPGPU Tensor Cores for Deep Learning (NVIDIA)

NVidia Volta architecture

- 64x FP32 cores per SM
- 32x FP64 cores per SM
- 8x Tensor cores per SM

Tensor core operations

- Tensor Core perform D = A x B + C, where A, B, C and D are matrices
- A and B are FP16 4x4 matrices
- D and C can be either FP16 or FP32 4x4 matrices
- Higher performance is achieved when A and B dimensions are multiples of 8
- Maximum of 64 floating-point mixed-precision FMA operations per clock



$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

Limitations of GPGPUs

Restrictions of GPGPU programming

- CUDA is a proprietary programming environment
- Writing OpenCL programs implies writing host code and device code, then connecting them through a low-level API
- GPGPU kernel programming lacks standard features of C/C++, such as recursion or accessing a (virtual) file system

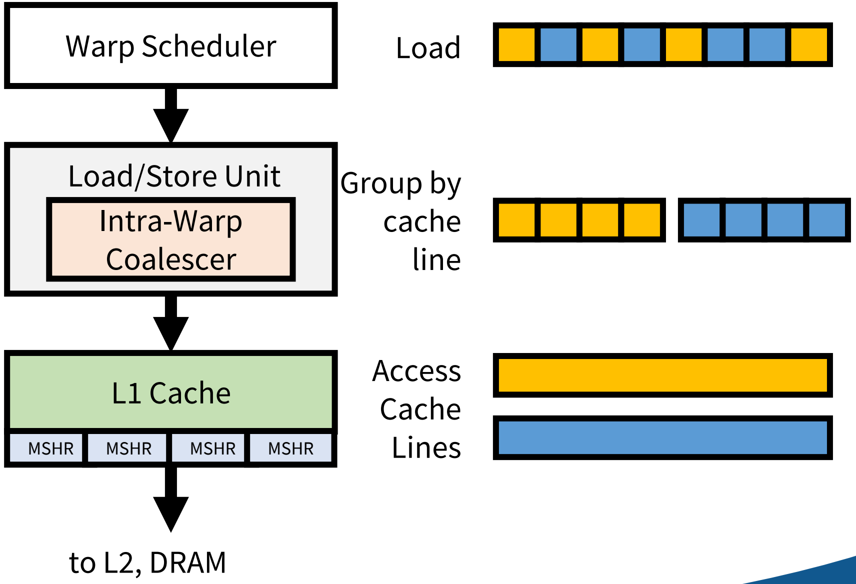
Performance issues with ‘thread divergence’

- Branch divergence: if…then…else construct will force all threads in a warp to execute both the "then" and the "else" path
- Memory divergence: when hardware cannot coalesce the set of warp global memory accesses into one or two L1 cache blocks

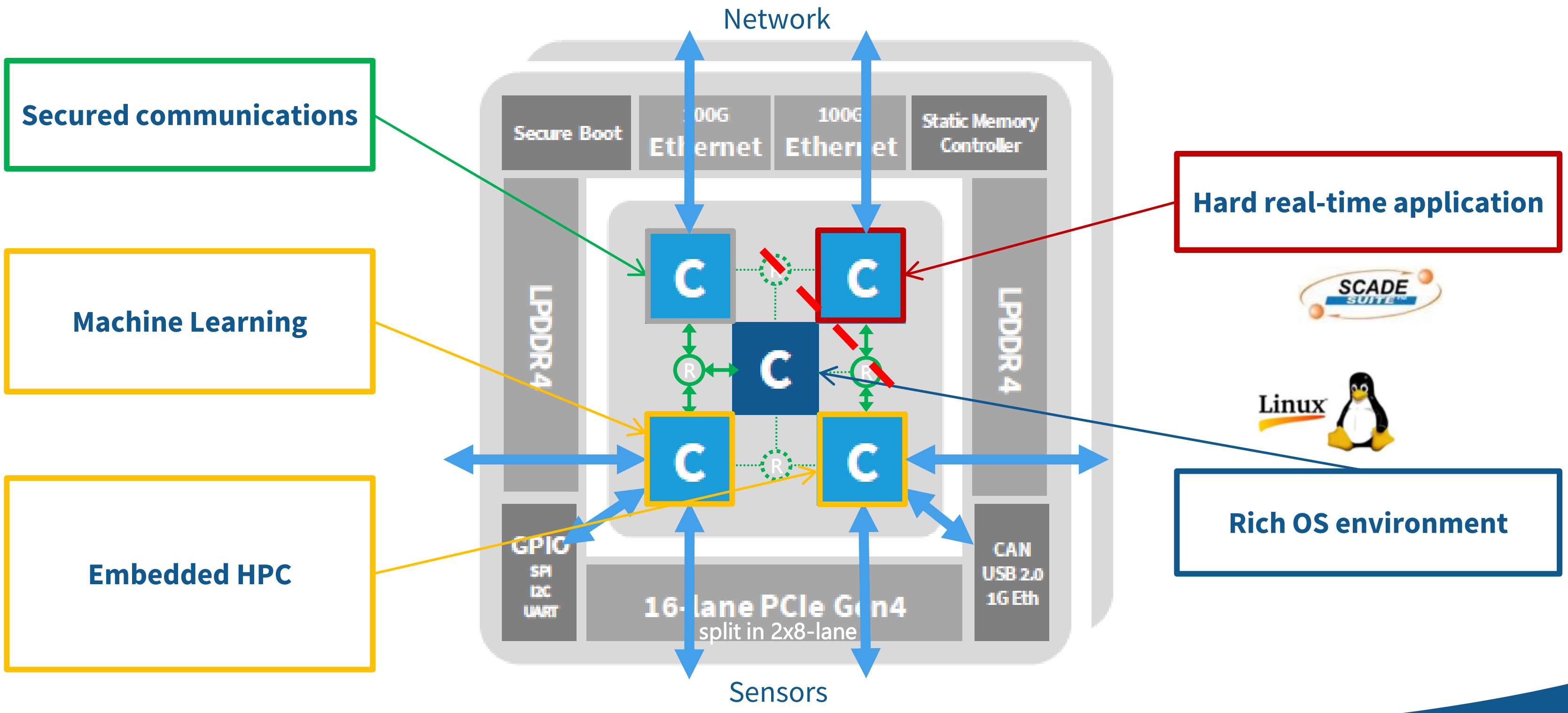
Time-predictability issues

- Dynamic allocation of thread blocks to SMs
- Dynamic warp scheduling and out-of-order execution on a SM

Memory access coalescing (Kloosterman et al.)



Mapping Intelligent System Functions to Compute Units



Secured communications

Machine Learning

Embedded HPC

Hard real-time application

Rich OS environment

Outline

Intelligent Systems

Manycore Processors

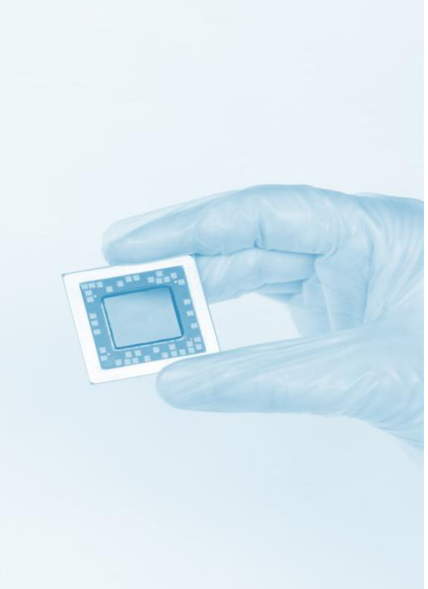
Kalray MPPA® Processors

Deep Learning Inference

Model-Based Design

Applications & Outlook

Kalray’s MPPA® Manycore Architecture



MPPA® (Massively Parallel Processor Array) Platform

Hardware

Manycore CPU architecture

Compute clusters of 16 high-performance CPU cores with local memory

DSP-like timing predictability

‘Fully timing compositional’ cores for accurate static timing analysis

Service guarantees of local memory system and network-on-chip

FPGA-like I/O capabilities

Software

CPU programming

Standard C/C++/OpenMP/OpenCL, OpenVX

Library code generators (MetaLibm, KaNN)

Model-based (SCADE Suite®, Simulink®)

MPPA® Processor Family and Roadmap

COMMERCIAL LAUNCH

2018 2019 2020 2021

	BOSTAN	COOLIDGE -1	COOLIDGE -2	Dx
PROCESS	28 nm	16 nm	16 nm	12 nm or 7nm
PERFORMANCE	1 DL TOPS 700 MFLOPS SP	24 DL TOPS 1 TFLOPS SP 3 TFLOPS HP	48 DL TOPS / 96 TDL OPS	100 TOPS / 200 TOPS
USE	Boards SC (40G) Prototypes	Boards and storage chip controllers (100G) Accelerator intelligent car	Qualification Car Market DC - NFV	DC
CONSUMPTION (WATTS)	8W – 25W	5W – 15W	5W – 20W	2W – 10W

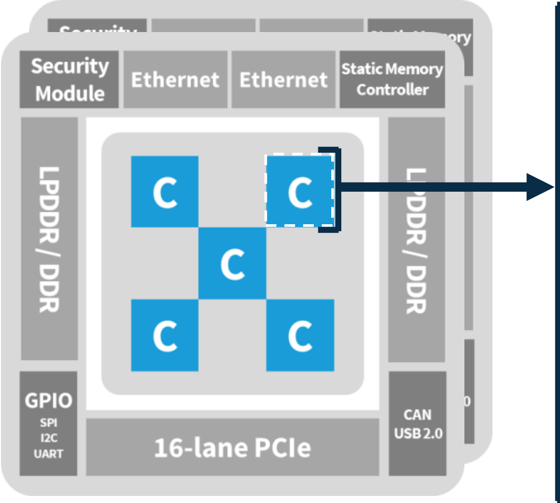
UNDER DEVELOPMENT

UNDER DEFINITION

MANYCORE TECHNOLOGY THAT ENABLES PROCESSOR OPTIMIZATION BASED ON EVOLVING MARKET REQUIREMENTS

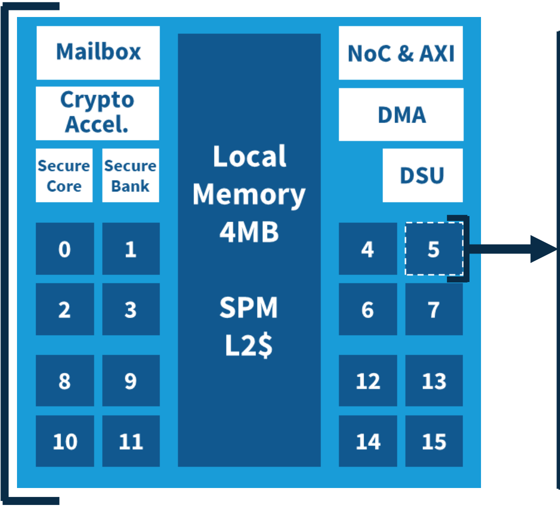
MPPA3-80 Processor (TSMC 16FFC, 1.2GHz)

1TFLOPS FP32, 3TFLOPS FP16.32, 24 DLTOPS INT8.32



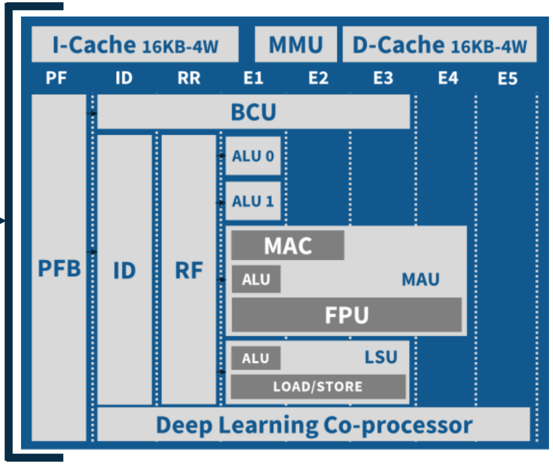
COOLIDGE PROCESSOR

5 compute clusters at 1200 MHz
2x 100Gbps Ethernet, 16x PCIe Gen4



COMPUTE CLUSTER

16+1 cores, 4 MB local memory
NoC and AXI global interconnects



6-ISSUE VLIW CORE

64x 64-bit register file
128MAC/c tensor coprocessor

Network-on-Chip for Global Interconnects

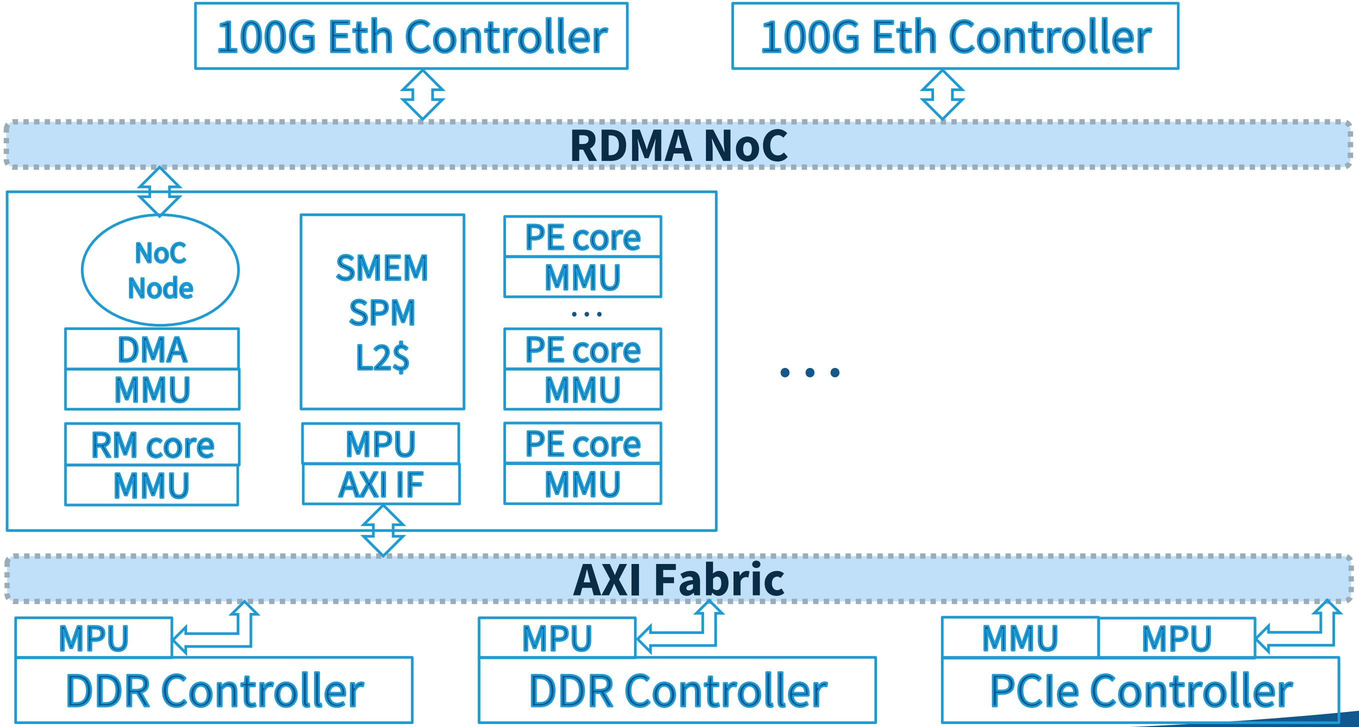
NoC as generalization of busses

- Connectionless
- Address-based transactions
- Flit-level flow control
- Implicit packet routing
- Inside a coherence domain
- Reliable communication
- Coherency protocol messages
- Coordinate with DDR memory controller front-end (Ex. Arteris FlexMem Memory Scheduler)

NoC as integrated macro-network

- Connection-oriented
- Stream-based transactions
- [End-to-end flow control]
- Explicit packet routing
- Across address spaces (RDMA)
- [Packet loss or packet reordering]
- Traffic shaping for QoS (application of DNC)
- Terminate macro-network (Ethernet, InfiniBand)
- Support of multicasting

MPPA3 Global Interconnects

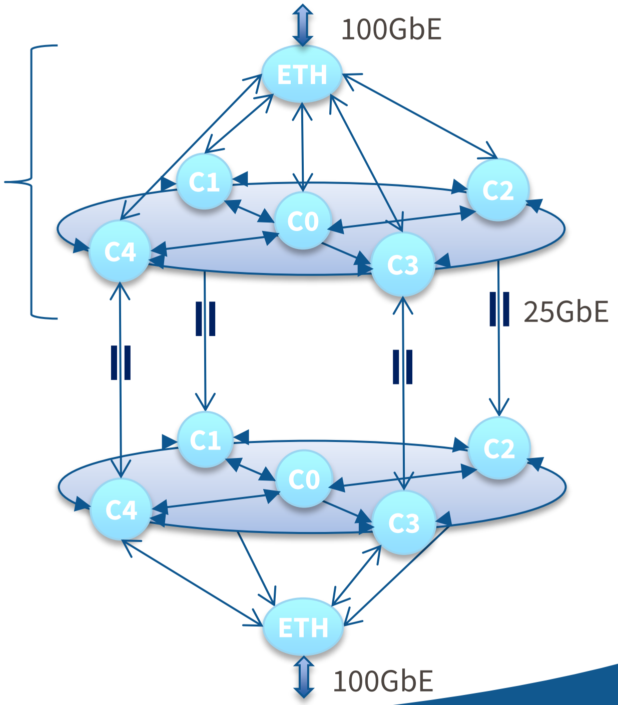


MPPA3 RDMA NoC

MPPA3 NoC architecture

- Wormhole switching with source routing
- 2 virtual channels, 4x TX DMA channels
- RDMA, remote queues, remote atomics
- 128-bit flits, up to 17 flits/packet (256B payload)

MPPA3-80 Processor



4x 25Gbps Ethernet lanes reused for NoC extension

- NoC packet encapsulation into IEEE 802.1Q standard for VLAN
- Designed for direct connections between 2 to 4 chips (using FEC)
- VCs map to IEEE 802.1Qbb Priority-based Flow Control (PFC) classes

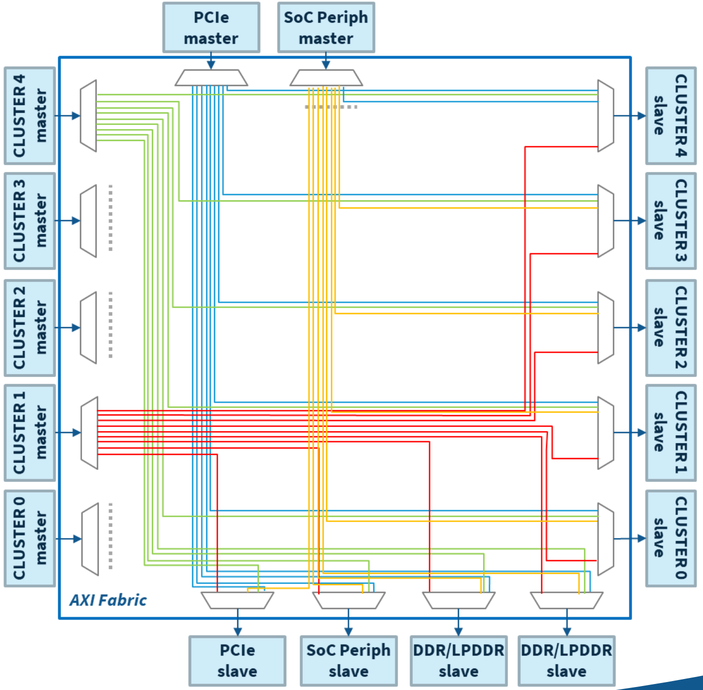
MAC dst 6 bytes	MAC src 6 bytes	VLAN etype 0x8100 2 bytes	VLAN TCI PFC (3 bits) / CFI (1 bit) NoC pkt nb (12 bits) 2 bytes	NoCX etype 0xB000 2 bytes	NoC pkt0	NoC pkt1	FCS 4 bytes

MPPA3 AXI Fabric

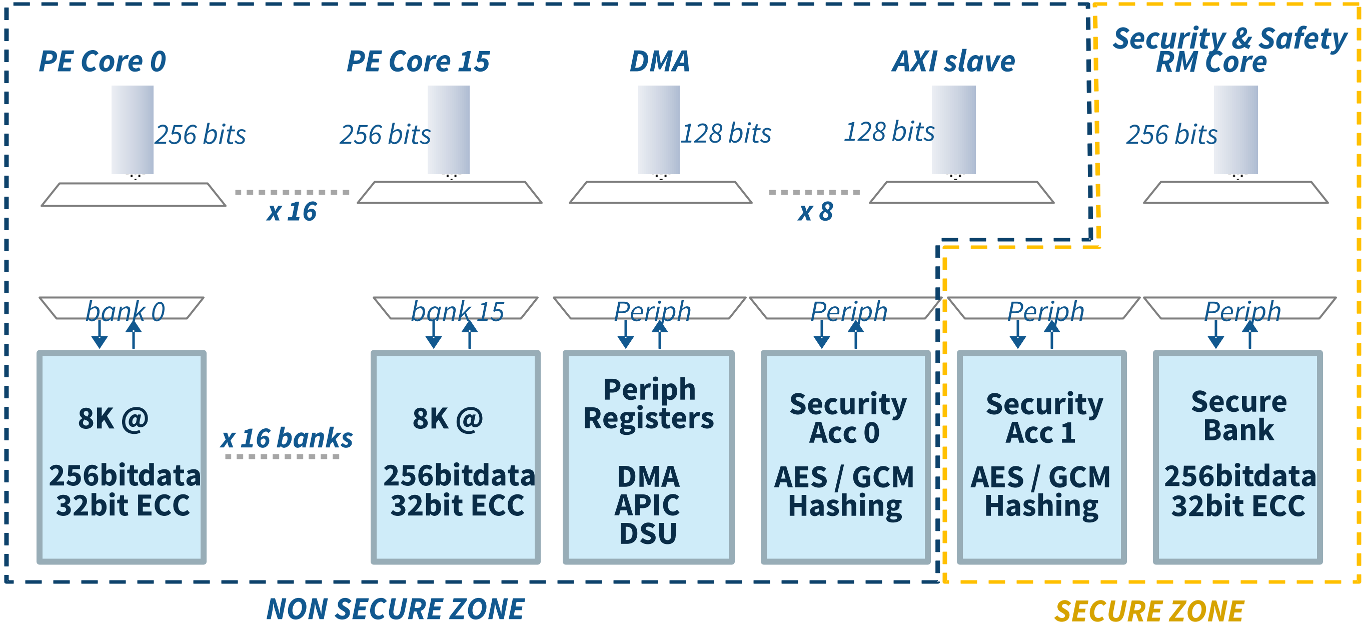
Deficit Round-Robin (DRR) Arbitration

- Assing a ‘quantum’ of flits Q_1 … Q_n to each input
- Associate a ‘deficit counter’ in flits DC_1 … DC_n to each input
- Iterate on the non-empty inputs; for each input i:

1. DC_i += Q_i
2. Transfer packets to output while cumulative flit count ≤ DC_i
3. DC_i -= transferred cumulative flit count
4. DC_i := 0 if input is empty



MPPA3 Compute Cluster



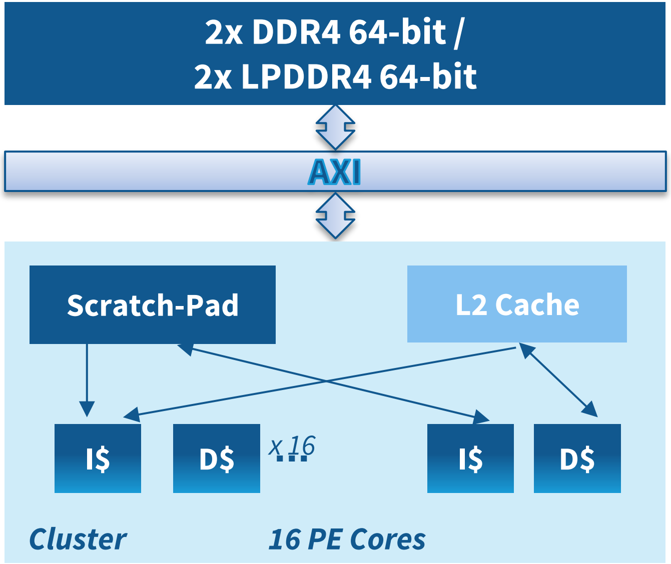
MPPA3 Memory Hierarchy

VLIW Core L1 Caches

- 16KB / 4-way LRU instruction cache per core
- 16KB / 4-way LRU data cache per core
- 64B cache line size
- Write-through, write no-allocate (write around)
- Coherency configurable across all L1 data caches

Cluster L2 Cache & Scratch-Pad Memory

- **Scratch-pad from 2MB to 4MB**
  - **16 independent banks, full crossbar**
  - **Interleaved or banked address mapping**
- L2 cache from 0MB to 2MB
  - 16-way Set Associative
  - 256B cache line size
  - Write-back, write allocate



L1 cache coherency	L2 cache coherency
enable /disable	enable /disable

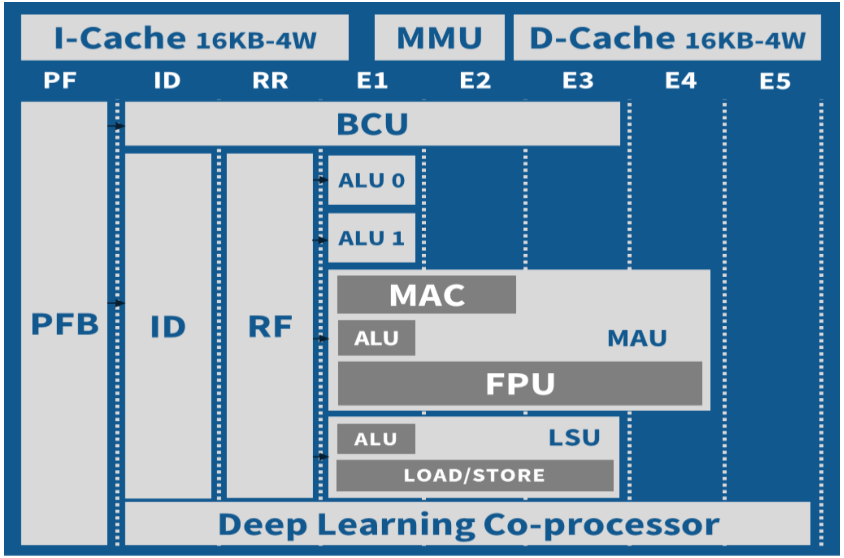
MPPA3 64-Bit VLIW Core

Vector-scalar ISA

- 64x 64-bit general-purpose registers
- Operands can be single registers, register pairs (128-bit) or register quadruples (256-bit)
- Immediate operands up to 64-bit, including F.P.
- 128-bit SIMD instructions by dual-issuing 64-bit on the two ALUS or by using the FPU datapath

FPU capabilities

- 64-bit x 64-bit + 128-bit → 128-bit
- 128-bit op 128-bit → 128-bit
- FP16x4 SIMD 16 x 16 + 32 → 32
- FP32x2 FMA, FP32x4 FADD, FP32 FMUL Complex
- FP32 Matrix Multiply 2x2 Accumulate



K1C VLIW CORE PIPELINE

MPPA3 Tensor Coprocessor

Extend VLIW core ISA with extra issue lanes

- Separate 48x 256-bit wide vector register file
- Matrix-oriented arithmetic operations (CNN, CV …)

Full integration into core instruction pipeline

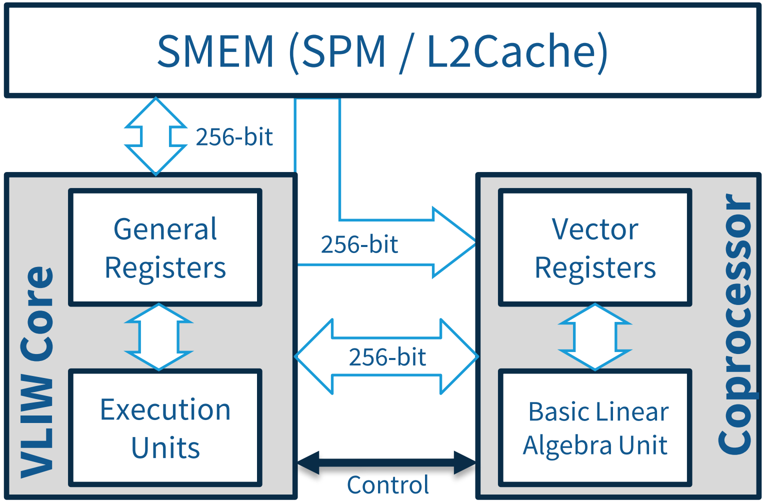
- Move instructions supporting matrix-transpose
- Proper dependency / cancel management

Leverage MPPA memory hierarchy

- SMEM directly accessible from coprocessor
- Memory load stream aligment operations

Arithmetic performances

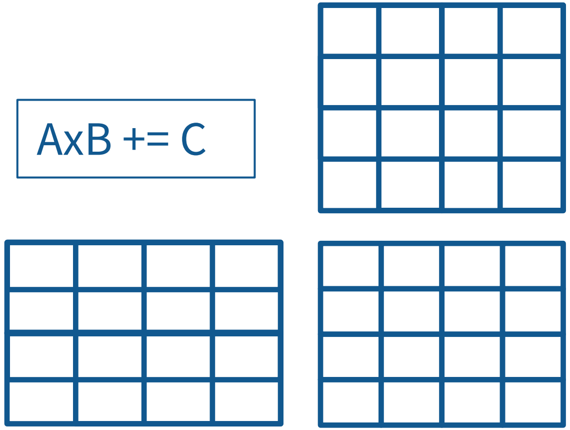
- 128x INT8→INT32 MAC/cycle
- 64x INT16→INT64 MAC/cycle
- 16x FP16→FP32 FMA/cycle



MPPA3 Coprocessor Matrix Operations

- INT16 to INT64 convolutions:

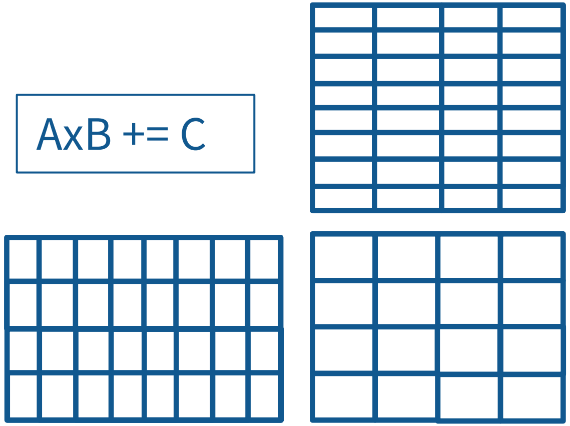
$(4x4)_{\text{int16}} \cdot (4x4)_{\text{int16}} += (4x4)_{\text{int64}}$



AxB += C

- INT8 to INT32 convolutions

$(4x8)_{\text{int8}} \cdot (8x4)_{\text{int8}} += (4x4)_{\text{int32}}$



AxB += C

Outline

Intelligent Systems

Manycore Processors

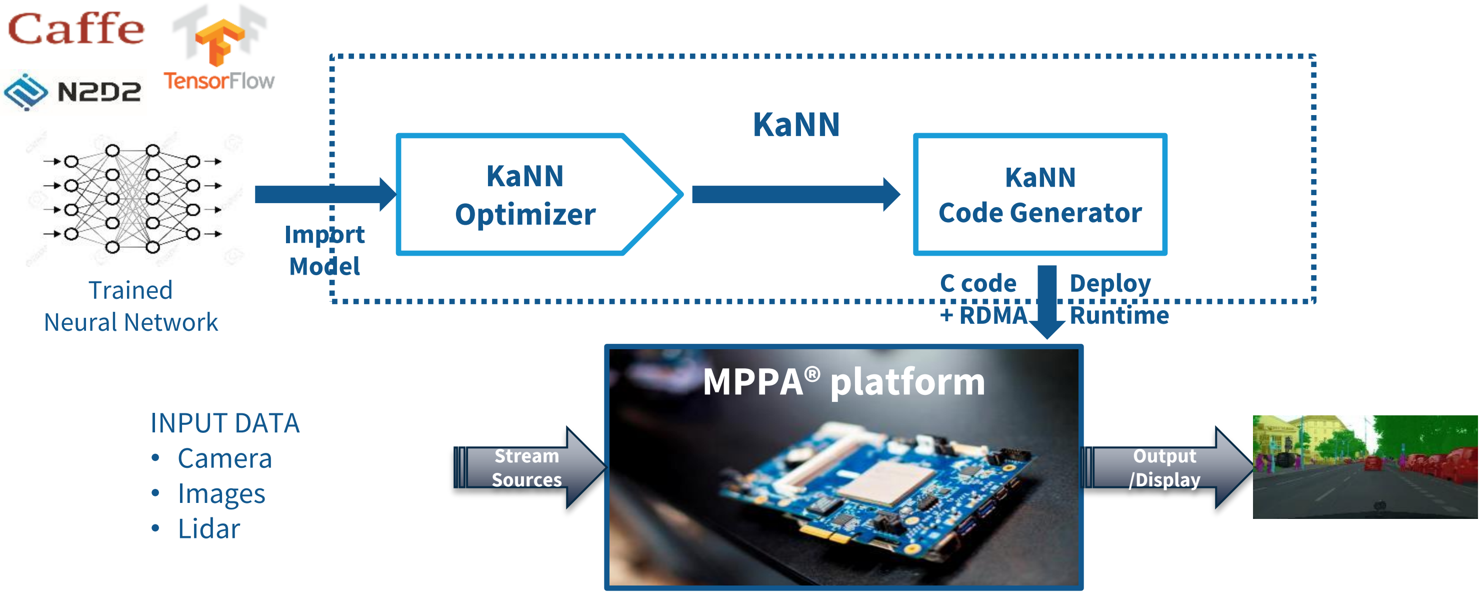
Kalray MPPA® Processors

Deep Learning Inference

Model-Based Design

Applications & Outlook

KaNN (Kalray Neural Network) Inference Code Generator



INPUT DATA

- Camera
- Images
- Lidar

RESULTS

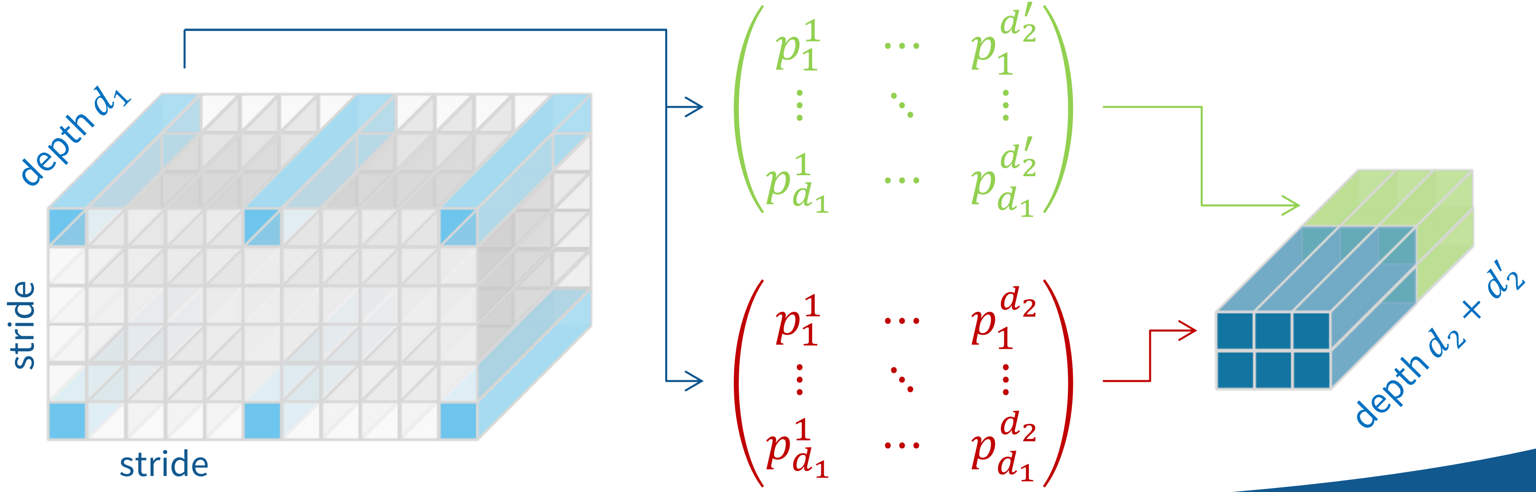
- Classification
- Segmentation
- Detection

CNN Inference on a MPPA Processor (1)

Compute one DNN layer at a time in topological sort order of the network

Decompose NxN convolutions as accumulations of N^2 1x1 convolutions

- Pixels layout is sequential along depth (channels) for dense memory accesses



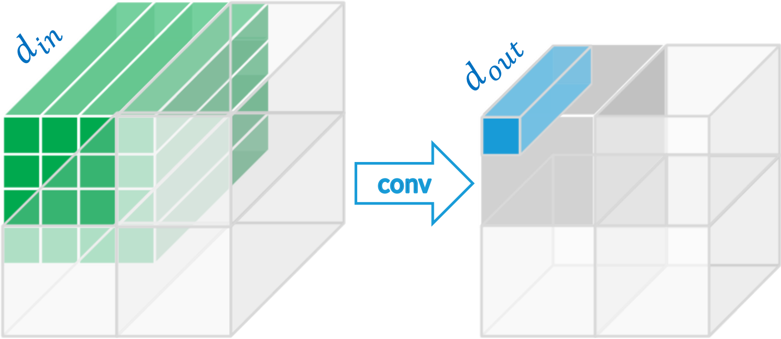
$$\begin{pmatrix} p_1^1 & \cdots & p_1^{d'_2} \\ \vdots & \ddots & \vdots \\ p_{d_1}^1 & \cdots & p_{d_1}^{d'_2} \end{pmatrix}$$

$$\begin{pmatrix} p_1^1 & \cdots & p_1^{d_2} \\ \vdots & \ddots & \vdots \\ p_{d_1}^1 & \cdots & p_{d_1}^{d_2} \end{pmatrix}$$

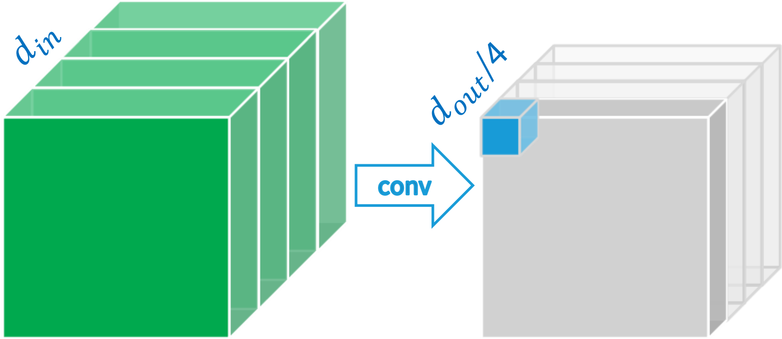
CNN Inference on a MPPA Processor (2)

Distribute activations across clusters SPMs, splitting along spatial and/or depth dimensions

- Spatial dimension splitting requires that the full set of parameters be loaded from external memory
- Channel dimension splitting requires access to the whole input image and a subset of the parameters
- Leverage NoC multicasting of parameters from external memory in case of spatial dimension splitting



$[3][3][d_{in}][d_{out}]$

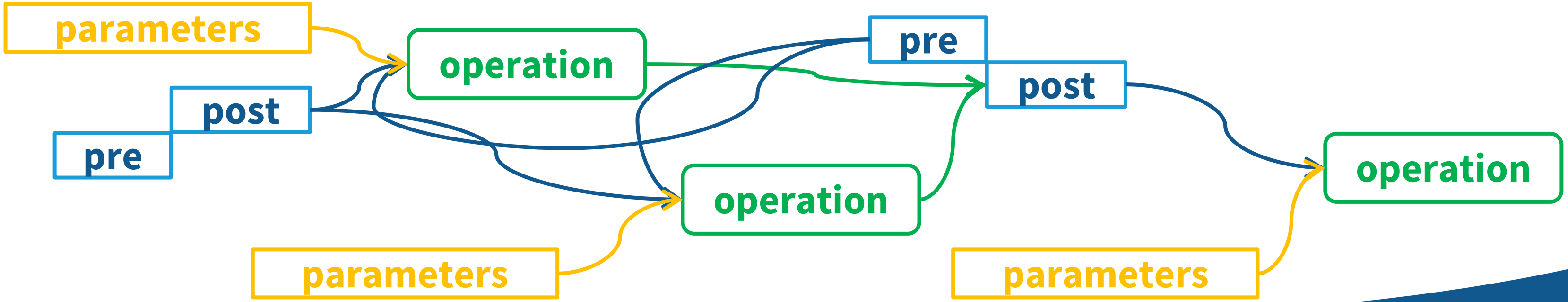


$[3][3][d_{in}][d_{out}/4]$

CNN Inference on a MPPA Processor (3)

SPMD (Single Program Multiple Data) execution that leverages NoC multicasting of parameters

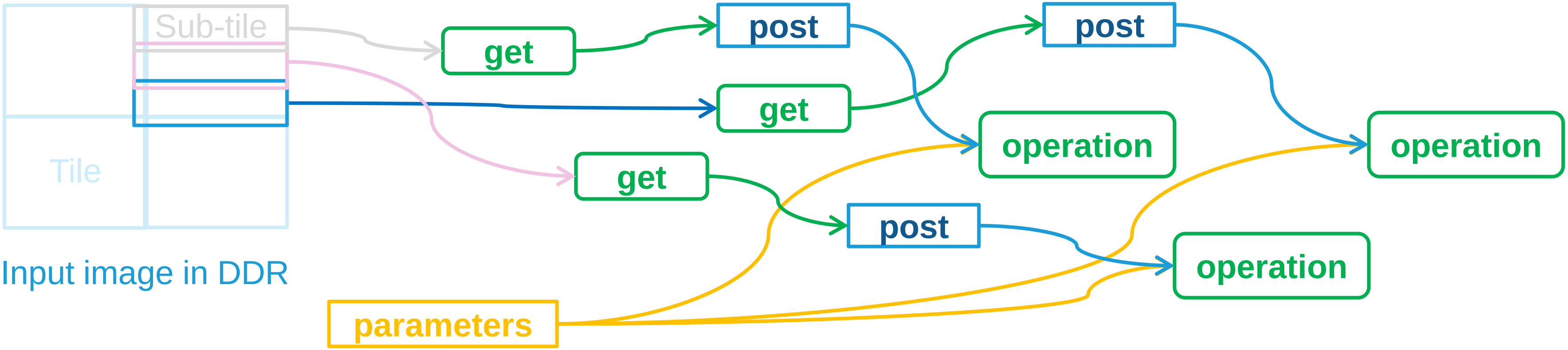
- Build a local memory buffer allocation and task execution schedule in each cluster
- Overlap parameter transfers from external memory with computations on local memory
- Allocation and scheduling are performed on the CNN network
  - an image corresponds to pre and post tasks,
  - layer compute operations corresponds to a malleable task
  - pre tasks load biases from external memory into the local memory buffer



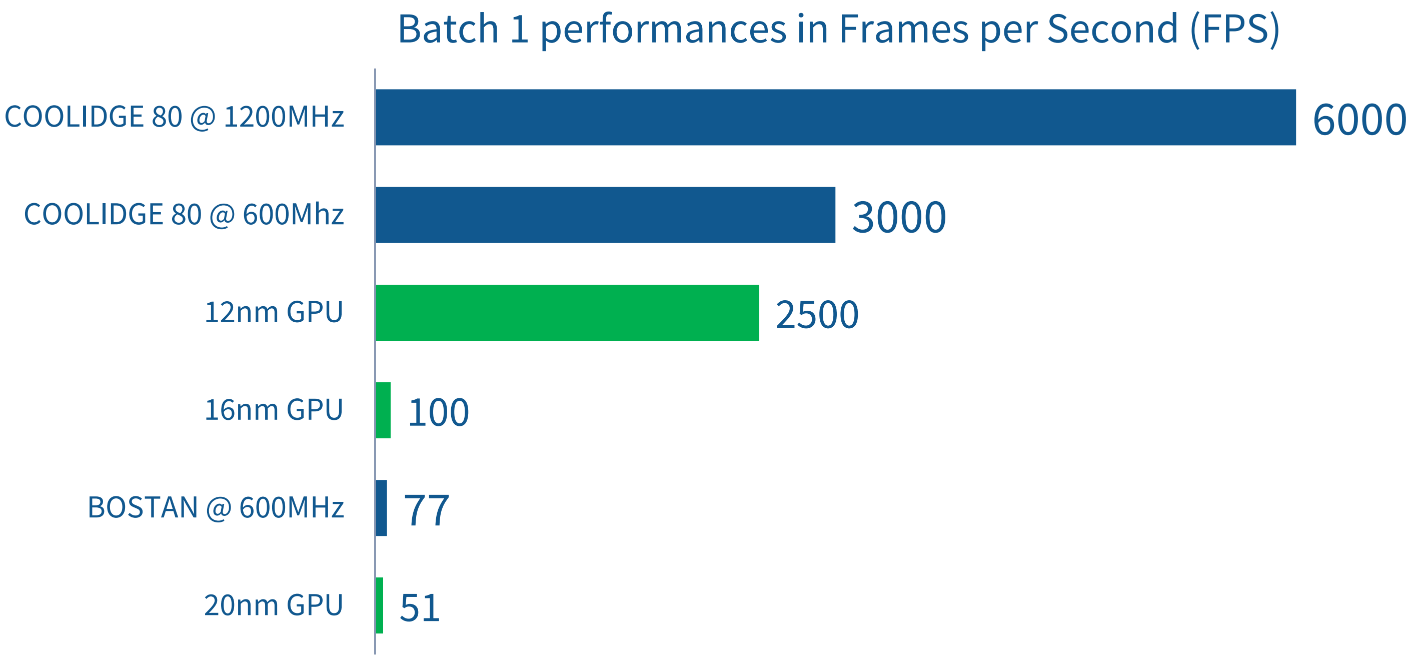
CNN Inference on a MPPA Processor (4)

For layers where images do not fit on-chip, stream sub-tiles from DDR memory

- All clusters remote write their tile of output image to DDR memory, then enter a synchronization barrier
- After clusters leave the barrier, they pipeline the remote read from DDR / operate / put to DDR of sub-tiles
- Larger sub-tiles factor more control overhead but reduce the amount of pipelining



Deep Learning Inference on Caffe GoogLeNet



Outline

Intelligent Systems

Manycore Processors

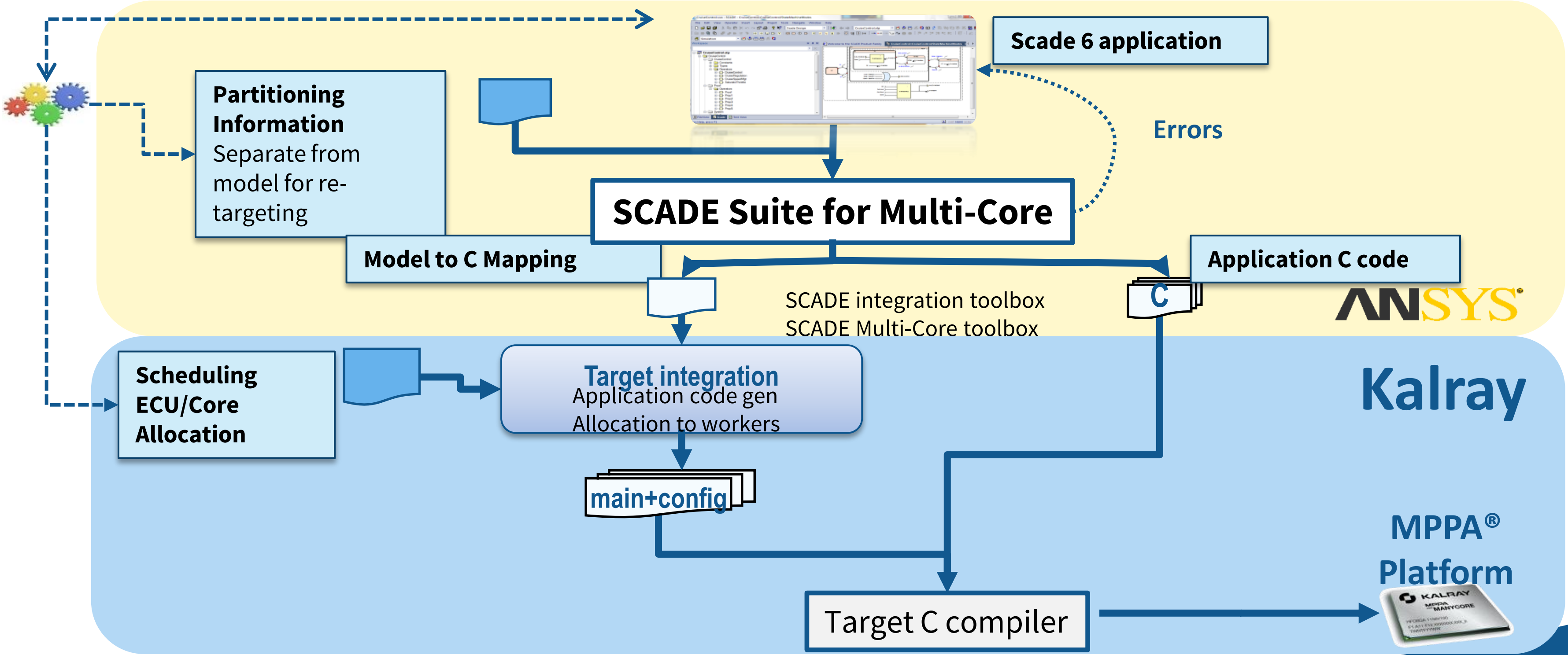
Kalray MPPA® Processors

Deep Learning Inference

Model-Based Design

Applications & Outlook

SCADE Suite Multi-Core Code Generation Flow



Scade 6 application

Partitioning Information
Separate from model for re-targeting

Errors

SCADE Suite for Multi-Core

Model to C Mapping

Application C code

C

SCADE integration toolbox
SCADE Multi-Core toolbox

ANSYS

Scheduling ECU/Core Allocation

Target integration
Application code gen
Allocation to workers

main+config

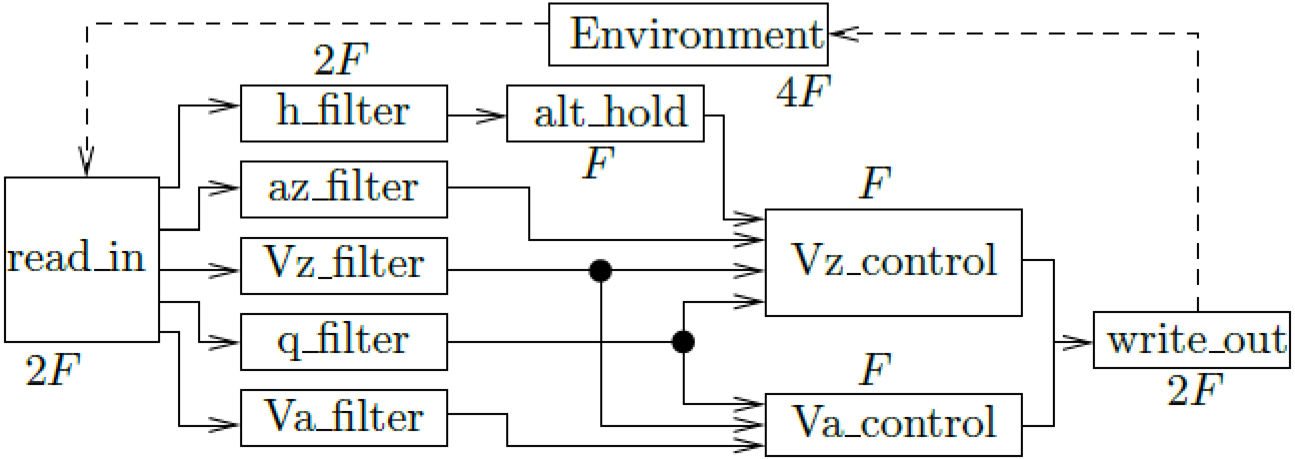
Kalray

MPPA®
Platform

Target C compiler

ROSACE Demonstration Application

- Simplified controller for the longitudinal motion of a medium-range civil aircraft in en-route phase: cruise and change of cruise level sub-phases



- Original application has 3 harmonic periods: F, 2F, 4F

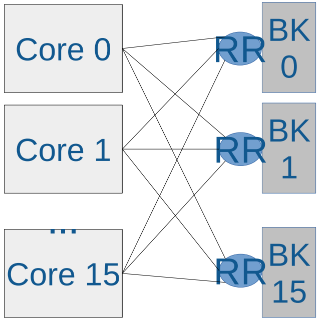
SCADE Suite MCG Code Generation (1)

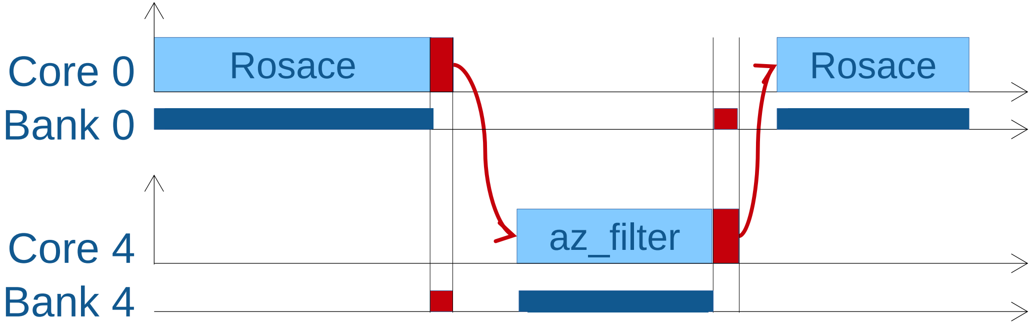
- MCG generates a set of tasks communicating trough one-to-one channels:
  - The root task executes the root operator of the input model
  - One task for each operator instance annotated in the input model
  - Each task receives data on an input channel, calls the operator and then sends the result on an output channel
  - Channels are single-producer, single-consumer FIFOs of size one

- The platform provider (Kalray) integrates MCG generated code by:
  - Providing *workers*, each able to execute sequentially a set of tasks
  - Implementing communication *channels* with their *send/recv* methods
  - Applying the prescribed scheduling and mapping of tasks to workers

SCADE Suite MCG Code Generation (2)

- Exploit the MPPA cluster configuration for 'high-integrity' execution
  - Enable the cluster local memory mapping of one bank per core





- Precisely compute the task WCETs (Worst-Case Execution Times)
  - Static analysis or measurement for the WCET of tasks in isolation
  - Refine the WCET with interferences using fixed-point [Rihani RTNS'16]

Outline

Intelligent Systems

Manycore Processors

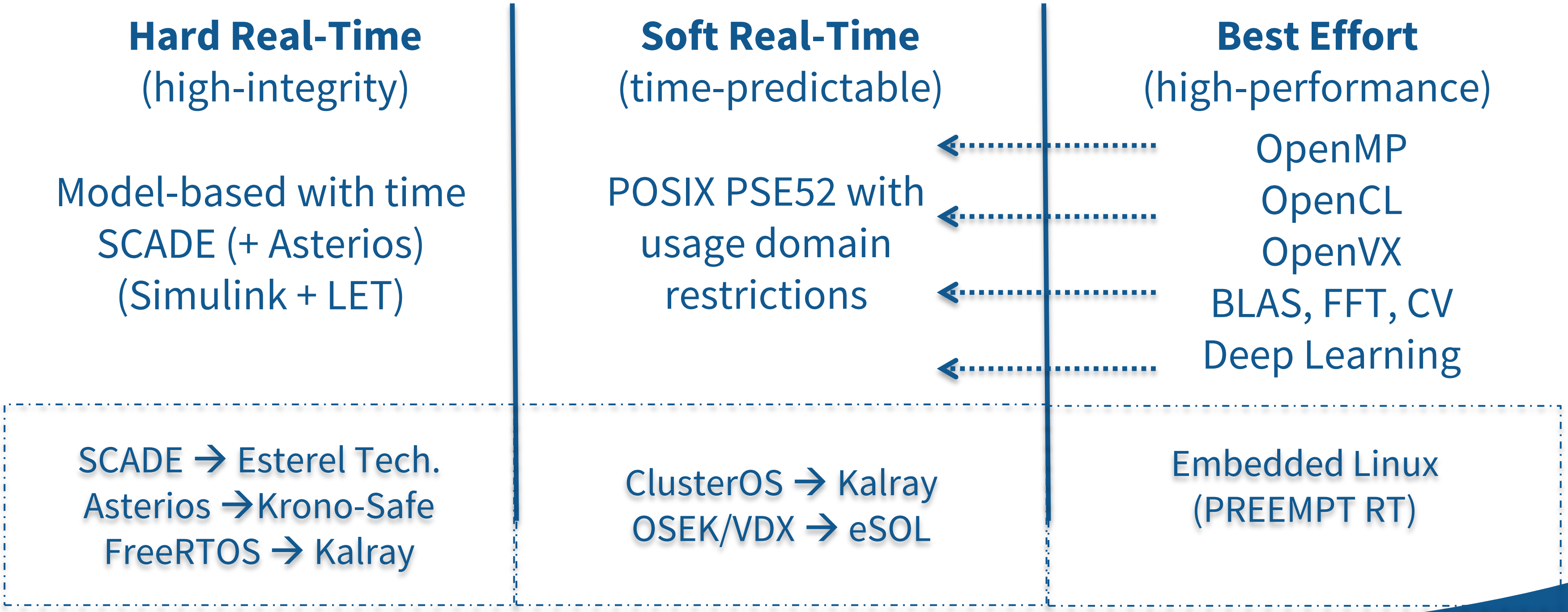
Kalray MPPA® Processors

Deep Learning Inference

Model-Based Design

Applications & Outlook

MPPA® Embedded Platform



Hard Real-Time
(high-integrity)

Model-based with time
SCADE (+ Asterios)
(Simulink + LET)

SCADE → Esterel Tech.
Asterios →Krono-Safe
FreeRTOS → Kalray

Soft Real-Time
(time-predictable)

POSIX PSE52 with
usage domain
restrictions

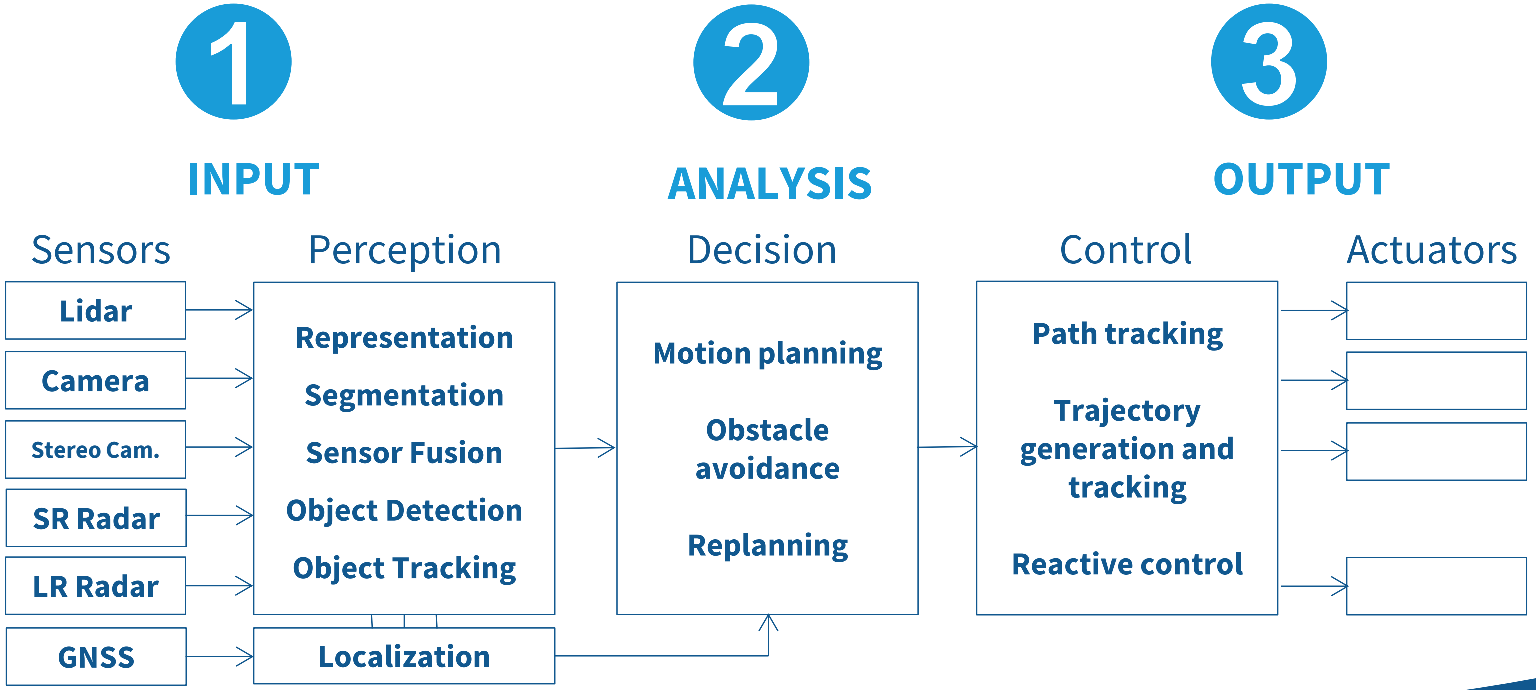
ClusterOS → Kalray
OSEK/VDX → eSOL

Best Effort
(high-performance)

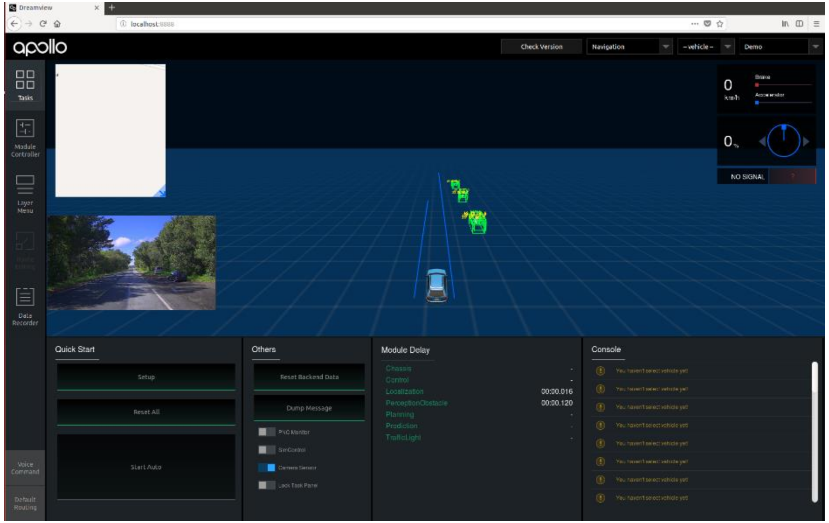
OpenMP
OpenCL
OpenVX
BLAS, FFT, CV
Deep Learning

Embedded Linux
(PREEMPT RT)

Autonomous Driving System



KaNN Integration into 3rd Party Autonomous Software Platforms



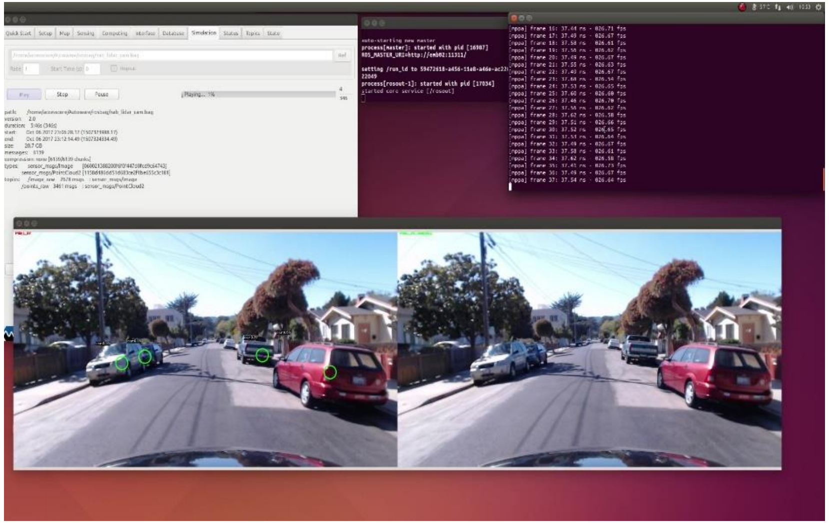
Module Delay

Chassis	-
Control	-
Localization	00:00.016
PerceptionObstacle	00:00.120
Planning	-
Prediction	-
TrafficLight	-

Console

You haven't select vehicle yet
You haven't select vehicle yet
You haven't select vehicle yet
You haven't select vehicle yet
You haven't select vehicle yet
You haven't select vehicle yet
You haven't select vehicle yet
You haven't select vehicle yet

MPPA2 Processing of BAIDU Apollo (Perception)



```
path:        /home/autoware/Autoware/rosbag/nab_lidar_cam.bag
version:     2.0
duration:    5:46s (346s)
start:       Oct 06 2017 23:06:28.17 (1507331988.17)
end:         Oct 06 2017 23:12:14.49 (1507324334.49)
size:        20.7 GB
messages:    6139
compression: none [6139/6139 chunks]
types:       sensor_msgs/Image       [060021388200f6f0f447d0fcd9c64743]
             sensor_msgs/PointCloud2 [1158d486dd51d683ce2f1be655c3c181]
topics:      /image_raw    2678 msgs    : sensor_msgs/Image
             /points_raw   3461 msgs    : sensor_msgs/PointCloud2
```

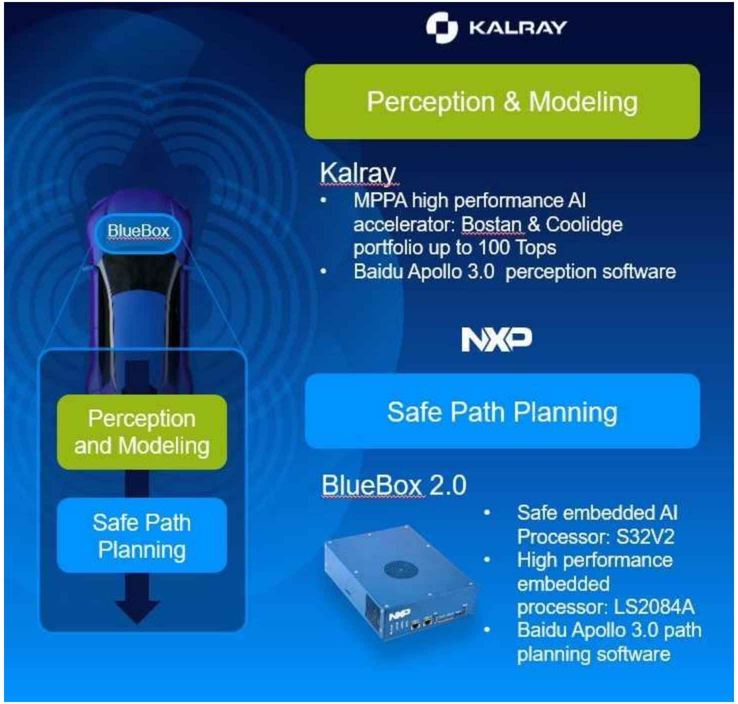
```
auto-starting new master
process[master]: started with pid [16987]
ROS_MASTER_URI=http://emb02:11311/

setting /run_id to 59472618-a456-11e8-a46e-ac22
22049
process[rosout-1]: started with pid [17034]
started core service [/rosout]
```

```
[mppa] frame 16: 37.44 ms - 026.71 fps
[mppa] frame 17: 37.49 ms - 026.67 fps
[mppa] frame 18: 37.58 ms - 026.61 fps
[mppa] frame 19: 37.56 ms - 026.62 fps
[mppa] frame 20: 37.49 ms - 026.67 fps
[mppa] frame 21: 37.55 ms - 026.63 fps
[mppa] frame 22: 37.49 ms - 026.67 fps
[mppa] frame 23: 37.68 ms - 026.54 fps
[mppa] frame 24: 37.53 ms - 026.65 fps
[mppa] frame 25: 37.60 ms - 026.60 fps
[mppa] frame 26: 37.46 ms - 026.70 fps
[mppa] frame 27: 37.56 ms - 026.62 fps
[mppa] frame 28: 37.62 ms - 026.58 fps
[mppa] frame 29: 37.51 ms - 026.66 fps
[mppa] frame 30: 37.52 ms - 026.65 fps
[mppa] frame 31: 37.51 ms - 026.64 fps
[mppa] frame 32: 37.49 ms - 026.67 fps
[mppa] frame 33: 37.58 ms - 026.61 fps
[mppa] frame 34: 37.62 ms - 026.58 fps
[mppa] frame 35: 37.41 ms - 026.73 fps
[mppa] frame 36: 37.49 ms - 026.67 fps
[mppa] frame 37: 37.54 ms - 026.64 fps
```

MPPA2 Processing of Autoware (Perception)

Kalray News from CES 2019 (EETimes)

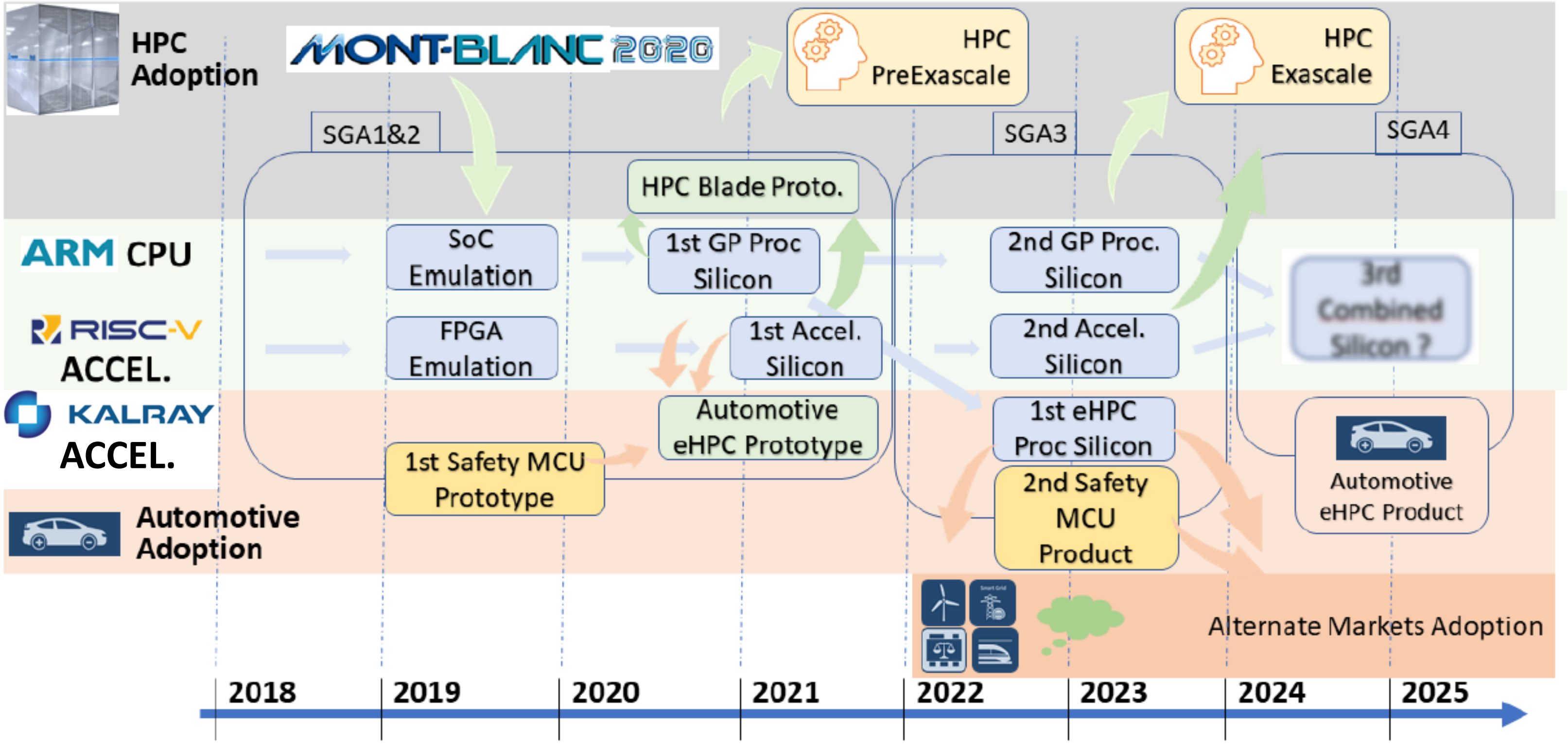

KALRAY
Perception & Modeling
Kalray
• MPPA high performance AI accelerator: Bostan & Coolidge portfolio up to 100 Tops
• Baidu Apollo 3.0 perception software
NXP
Safe Path Planning
BlueBox 2.0
• Safe embedded AI Processor: S32V2
• High performance embedded processor: LS2084A
• Baidu Apollo 3.0 path planning software
BlueBox
Perception and Modeling
Safe Path Planning

The Dutch semiconductor company revealed at the Consumer Electronics Show here that it has chosen a French startup called Kalray to fill in a void created by Qualcomm when it walked away last summer from a $44 billion deal to buy NXP.

Under their new partnership, Kalray and NXP are developing a central computing platform that combines Kalray's MPPA processors with NXP's S32 processors.

At CES, the companies demonstrate Kalray's MPPA and NXP BlueBox running together on Baidu's Apollo open automotive software.

Mont-Blanc 2020 and EPI Projects



HPC Adoption

MONT-BLANC 2020

HPC PreExascale

HPC Exascale

SGA1&2

SGA3

SGA4

HPC Blade Proto.

ARM CPU

SoC Emulation

1st GP Proc Silicon

2nd GP Proc. Silicon

RISC-V ACCEL.

FPGA Emulation

1st Accel. Silicon

2nd Accel. Silicon

3rd Combined Silicon ?

KALRAY ACCEL.

Automotive eHPC Prototype

1st eHPC Proc Silicon

1st Safety MCU Prototype

2nd Safety MCU Product

Automotive eHPC Product

Automotive Adoption

Alternate Markets Adoption

2018

2019

2020

2021

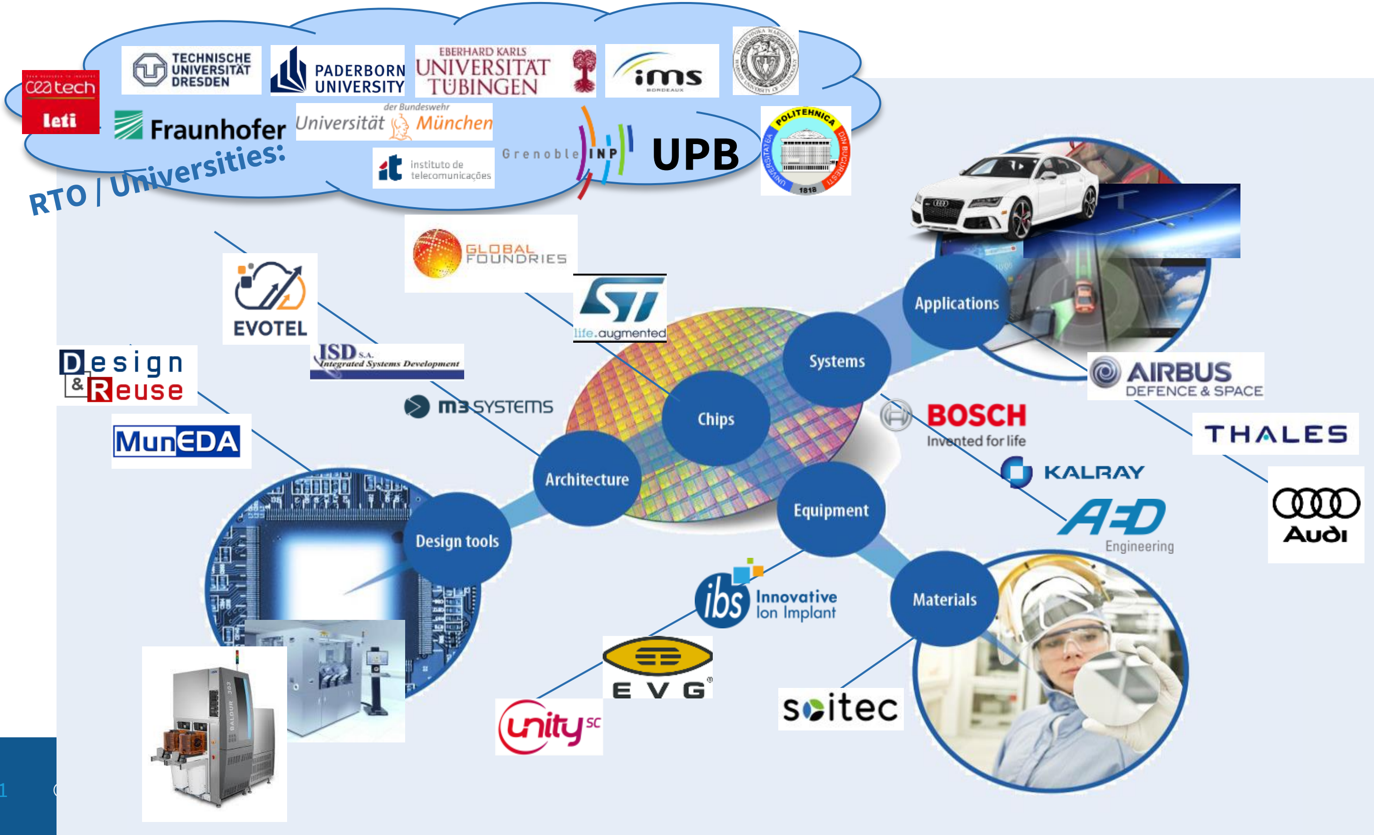
2022

2023

2024

2025

OCEAN 12 ECSEL Project
Opportunity to Carry European Autonomous driviNg further with FDSOI technology up to 12nm node



RTO / Universities:

Design tools

Architecture

Chips

Systems

Applications

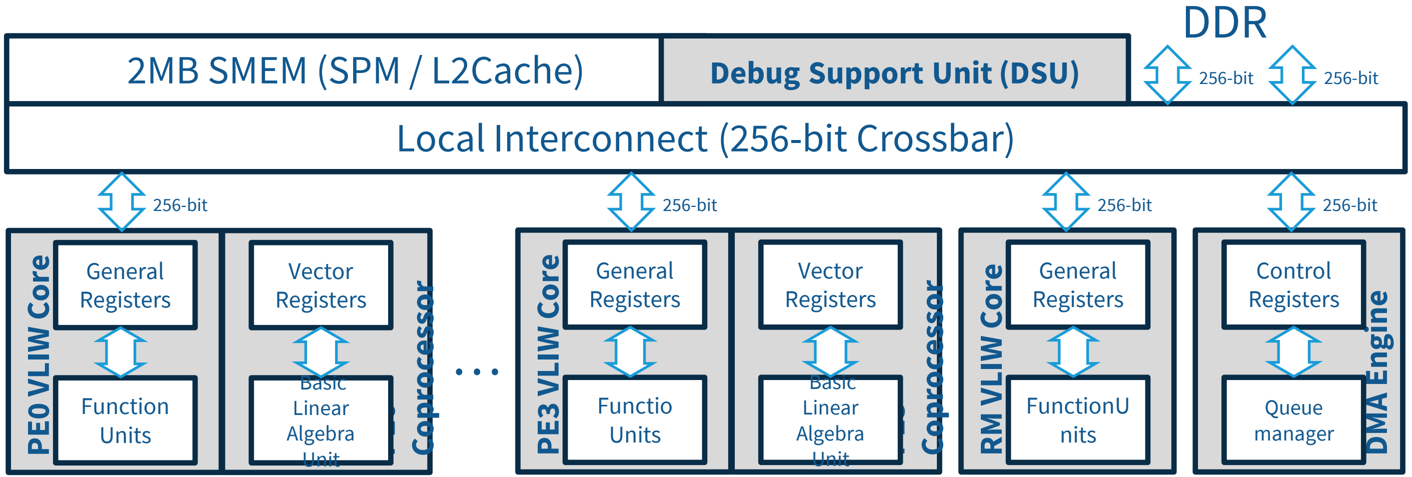
Equipment

Materials

OCEAN12 Work Package 3 (IP Factory)

Task 3.1: High-Performance Computing & Vision Signal Processing (Kalray, CEA, ISD, M3S)

- MPPA Cluster tile IP designed and running on FPGA emulation (Altera Stratix-10)
- Running deep learning inference (KaNN) and computer vision (dense Optical Flow)



DDR

256-bit

256-bit

2MB SMEM (SPM / L2Cache)

Debug Support Unit (DSU)

Local Interconnect (256-bit Crossbar)

256-bit

256-bit

256-bit

256-bit

PE0 VLIW Core

General Registers

Function Units

Vector Registers

Basic Linear Algebra Unit

Coprocessor

. . .

PE3 VLIW Core

General Registers

Functio Units

Vector Registers

Basic Linear Algebra Unit

Coprocessor

RM VLIW Core

General Registers

FunctionU nits

Control Registers

Queue manager

DMA Engine

Conclusions

CPU-based manycore accelerators

- C/C++/POSIX/OpenMP/OpenCL programmability
- Energy efficiency & time predictability

Kalray manycore accelerators

- Learn from GPGPU and computer vision processors
- High compute intensity comes from 2D operations
- Leverage local memories with RDMA engines

Programming environments

- Optimized application library generators
- Deep learning and graph-based frameworks
- High-performance using OpenCL and OpenMP
- Model-based programming for safety-critical

European Projects Mont-Blanc 2020, EPI and OCEAN12

Conclusions

SAFETY

- Hardware partitioning
- Software partitioning
- Hypervisor support
- ISO26262 ASIL B/C

SECURITY

- Hardware root of trust
- Secure boot
- Authenticated debug
- Trusted execution environment
- Encrypted application code

DETERMINISM

- Fully timing compositional cores
- Banked on-chip memory
- Interference-free local interconnect
- Network-on-Chip (NoC) service guarantees

PERFORMANCE

- High-end floating-point and bit-level processing
- DSP-style energy efficiency
- Scalability by replicating clusters

STANDARDS

- Standard programming environments (C/C++, OpenMP, POSIX, OpenCL, OpenVX)
- Standard development tools (Eclipse, GCC, GDB, LLVM, Linux)

SCALABLE

- Adaptability to E/E architecture
- Low range to high range car lines
- Allow distribution of functions

THANK YOU

KALRAY S.A. - GRENOBLE - FRANCE
180 avenue de l’Europe,
38 330 Montbonnot - France
Tel: +33 (0)4 76 18 09 18
email: info@kalray.eu

KALRAY INC. - LOS ALTOS - USA
4962 El Camino Real
Los Altos, CA - USA
Tel: +1 (650) 469 3729
email: info@kalrayinc.com

Pictures credits: Kalray, ©Fotolia

KALRAY