KALRAY

Manycore Processor for Next Generation Vehicles

Kalray MPPA®

Massively Parallel Processor Array

January 2020– Confidential Information

www.kalrayinc.com

KALRAY IN A NUTSHELL

We offer a new type of processor targeting the booming market of intelligent systems

International Presence

France	Grenoble (HQ), Sophia-Antipolis,
USA	Los Altos, California
Japan	Yokohama

Breaktrough technology from **10** years of development

23 patent families

Financial and industrial investors

cea investissement

INOCAPGestion

MBDA

RENAULT NISSAN MITSUBISHI

ACE

Pengpai

SAFRAN

bpifrance

EURONEXT

IPO in June 2018

(ALKAL)

MPPA® ROADMAP
SCALABLE PRODUCT FAMILY

SAMPLES

2018 · 2019 · 2021 · 2023

	BOSTAN	COOLIDGE v1	COOLIDGE v2[2]	DOLOMITES[2]
PROCESS	28 nm	16 nm	16 nm	12 nm or 7nm
PERFORMANCE	1.3 TOPS[1]	25 TOPS[1]	50 / 100 TOPS[1]	100 TOPS [1]/ 200 TOPS[1]
USE CASES / MARKET	40G Data Center Auto Prototypes	100G Data Center / EDGE Autonomous Vehicles	100G Data Center / EDGE Autonomous Vehicles	Next Gen Data Center / EDGE Autonomous Objects
CONSUMPTION (WATTS)	25W	25W	30W/60W	N/A
	PRODUCTION	AVAILABLE	UNDER DEVELOPMENT	UNDER DEFINITION

[1] 8-bit operations ("TOPS")
[2] Initial target – may change

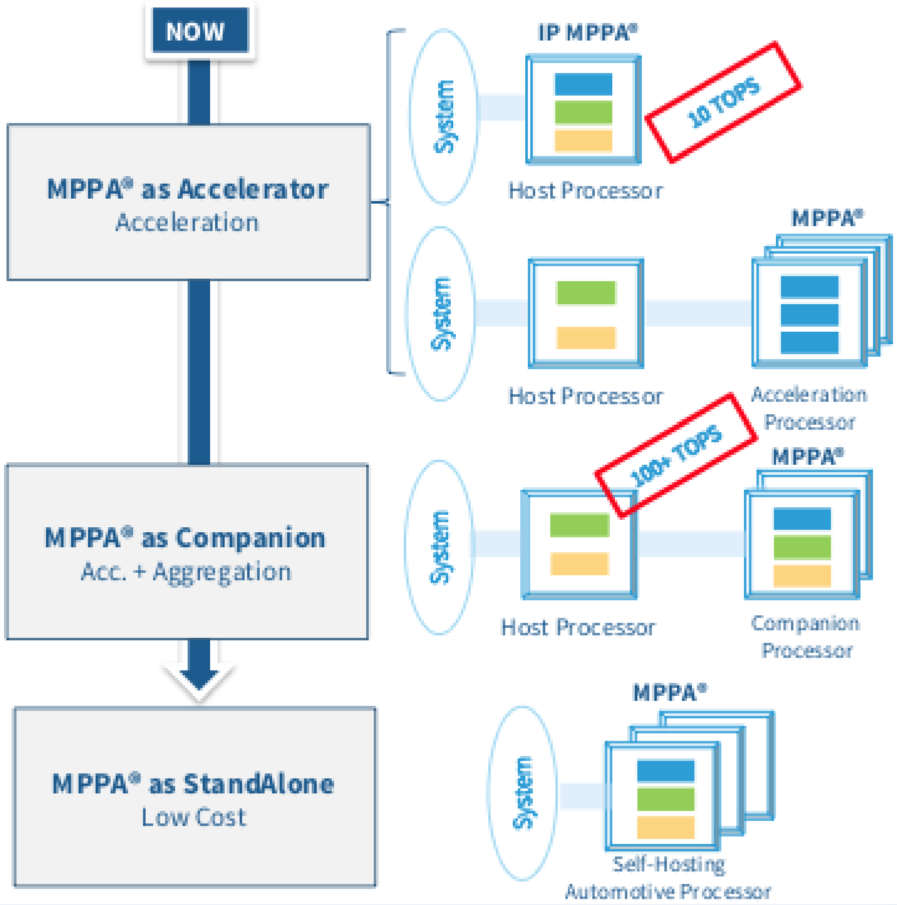
MPPA® ARCHITECTURE BENEFITS

- Computing power
- Data processing in real time
- Completion of dozens of critical tasks in parallel
- Low power consumption
- Programmable / Open system
- Security & Safety

KALRAY
MPPA
MANYCORE

KALRAY

A STEP BY STEP & FLEXIBLE APPROACH



- IP Business Case

CLUSTER IP

Use Cases: **Accelerating** Perception & Decision

- Offload Inference to MPPA: Support of main CNN framework
- Offload Computer Vision: OpenCV, Open VX
- Offload Computing Algorithms: OpenCL, FFT, BLIS
- Offload of customer computing functions
- AEC-Q100

Use Cases: **Aggregating** functions for Perception & Decision (& Controls)

- Same as Accelerator mode
- + Execute applications: C/C++/RTOS
- + Execute Safety context: C/Safety RTOS
- + mixed criticality support

Use Cases: **Consolidating** functions for Perception & Decision (& Controls)

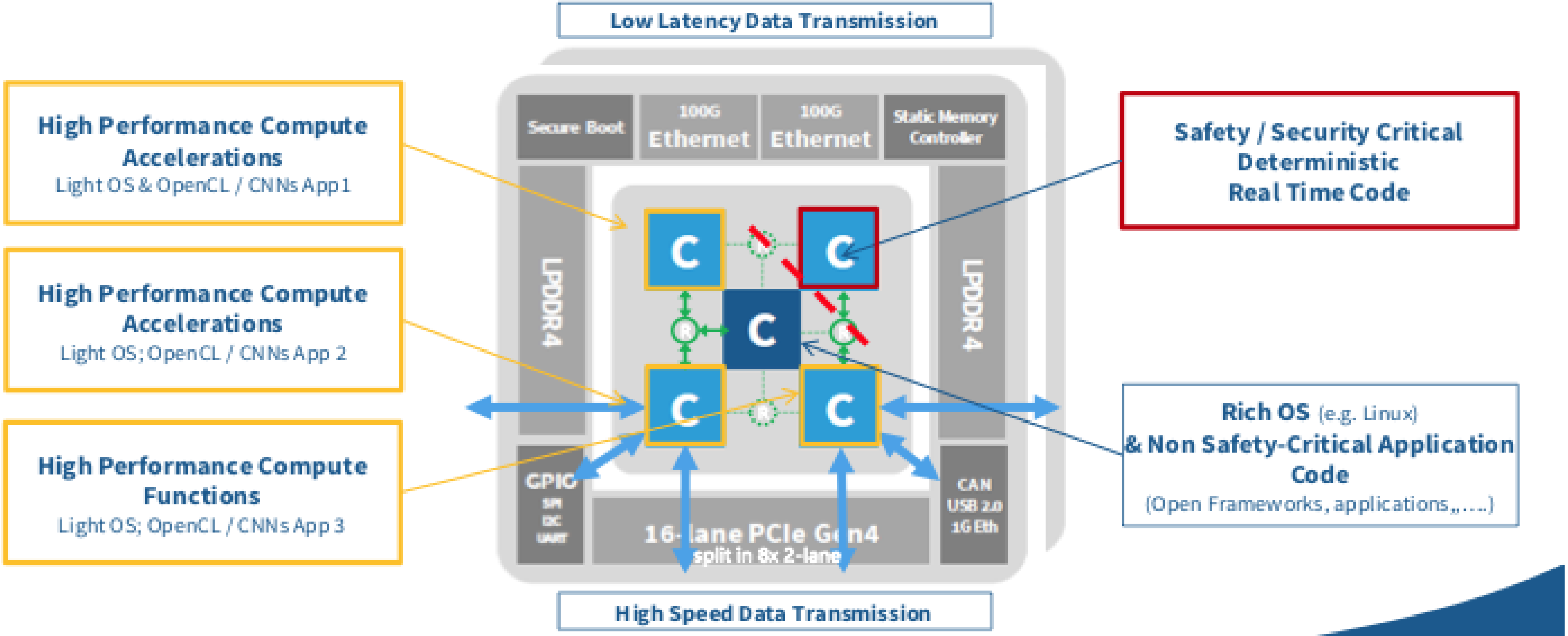
- Same as Companion Processor
- + self-hosting (no need of Host Processor)
- + Scalable consolidation by assignment to clusters

🟩 Applications
🟨 Safety Critical
🟦 Accelerations

STANDALONE SYSTEM BUILD ON MPPA

STANDALONE MODE

THE « **FREE OF INTERFERENCE** » and « **DETERMINISTIC** » MPPA® ARCHITECTURE ALLOWS FOR COMPLEX CRITICAL SYSTEMS INTEGRATION ON A SINGLE DIE, RUNNING DIFFERENT OPERATING SYSTEMS AND SOFTWARE ENVIRONMENTS AND DRAMATICALLY SIMPLIFYING CERTIFICATION EFFORT



Low Latency Data Transmission

High Performance Compute Accelerations
Light OS & OpenCL / CNNs App 1

High Performance Compute Accelerations
Light OS; OpenCL / CNNs App 2

High Performance Compute Functions
Light OS; OpenCL / CNNs App 3

Safety / Security Critical Deterministic Real Time Code

Rich OS (e.g. Linux) **& Non Safety-Critical Application Code**
(Open Frameworks, applications,….)

High Speed Data Transmission

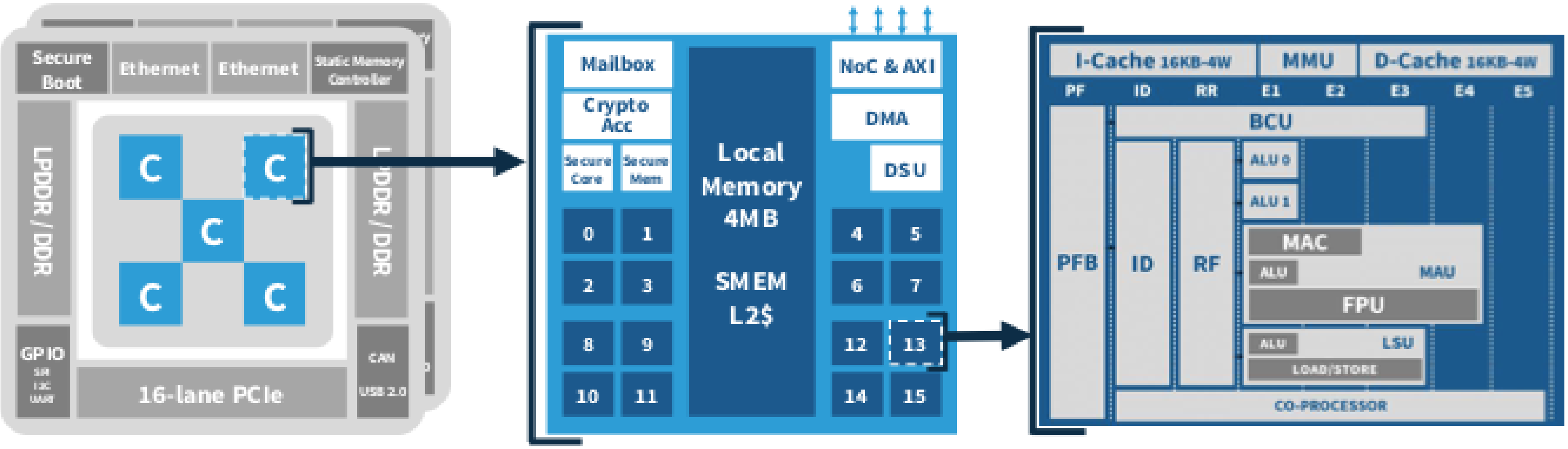
HARDWARE OVERVIEW

MPPA®

The Intelligent Processors

MPPA® COOLIDGE SCALABLE APPROACH
for Next Generation Embedded Systems

PATENTED



MANYCORE PROCESSOR

Architecture updates

- 80 or 160 CPU cores
- 600/900/1200MHz 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-processor
- Safety/Security 64-bit core

Memory

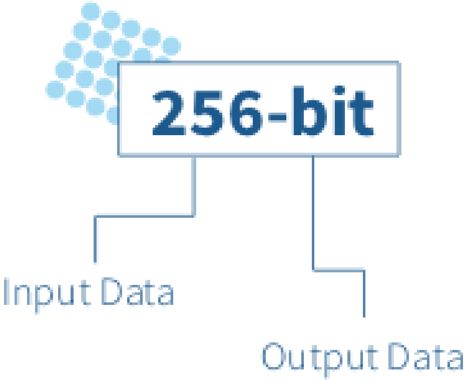
- L1 cache coherency (configurable)
- 4MB memory (BW = 614GB/s)

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

COOLIDGE CO-PROCESSOR
Tightly Coupled Accelerator (TCA)



256-bit

Input Data

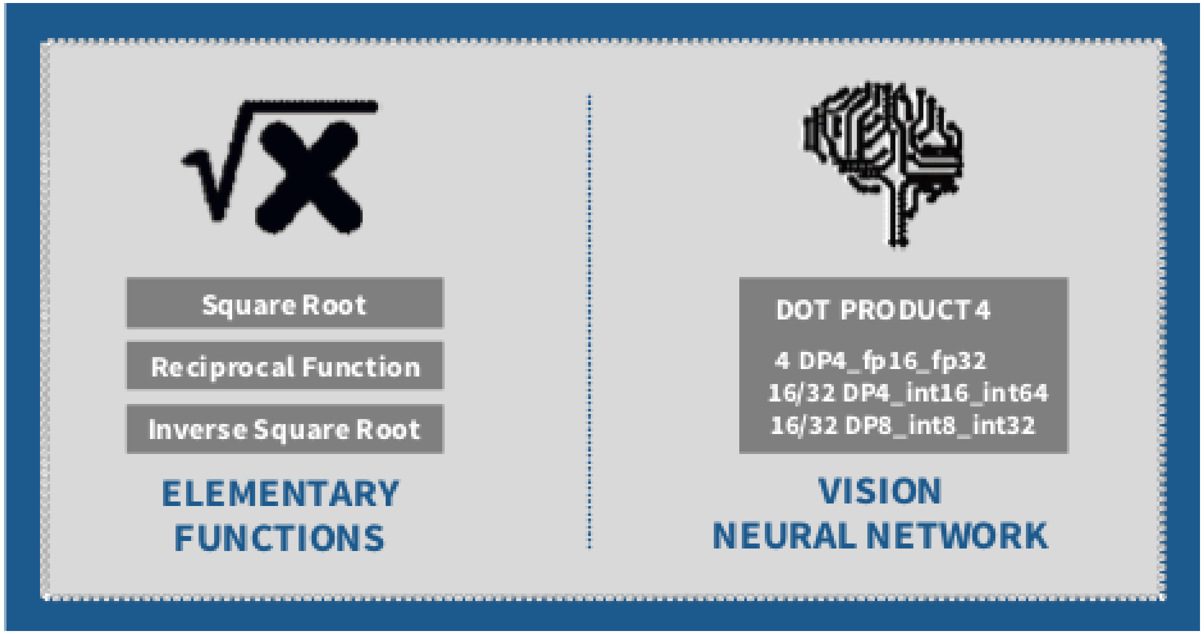
Output Data

UP TO

256

MACs

per cycle



Square Root

Reciprocal Function

Inverse Square Root

ELEMENTARY FUNCTIONS

DOT PRODUCT 4

4 DP4_fp16_fp32
16/32 DP4_int16_int64
16/32 DP8_int8_int32

VISION NEURAL NETWORK

Coolidge vs. Xavier Performance

		Coolidge-80 v1 @1.2 GHz	Coolidge -80 v2 @1.2 GHz	Coolidge -160 v2 @1.2 GHz	NVIDIA Xavier
INT8	Core	N/A	N/A	N/A	N/A
	Copro	24.6 TOPS	49.2 TOPS	98.4 TOPS	20 + 10
	TOTAL	**24.6 TOPS**	**49.2 TOPS**	**98.4 TOPS**	**30 TOPS**
INT16	Core	2 TOPS	2 TOPS	4 TOPS	
	Copro	12.3 TOPS	24.6 TOPS	49.2 TOPS	10 + 5
	TOTAL	**14.3 TOPS**	**26.6 TOPS**	**53.2 TOPS**	**15 TOPS**
FP16	Core	1.15 TFLOPS	1.15 TFLOPS	2.3 TFLOPS	
	Copro	3.05 TFLOPS	3.05 TFLOPS	6.1 TFLOPS	10 + 5
	TOTAL	**4.2 TFLOPS**	**4.2 TFLOPS**	**8.4 TFLOPS**	**15 TFLOPS**
FP32	Core	1.15 TFLOPS	1.15 TFLOPS	2.3 TFLOPS	
	Copro	N/A	N/A	N/A	1.3 TFLOPS
	TOTAL	**1.15 TFLOPS**	**1.15 TFLOPS**	**2.3 TFLOPS**	**1.3 TFLOPS**
Power		**25W**	**30W**	**60W**	**30W**

Architecture

Core

- 64-bit/32-bit architecture
- 6-issue VLIW
- 16KB instruction cache / 16KB data cache with MMU
- IEEE 754-2008 Floating Point Unit (FPU)
- Squareroot and reciprocal operations in floating single precision
- 64-bit integer multiplication (Asymmetric cryptography)
- Support up to 4 execution rings
- Support up to 256-bits per cycle Load/Store

Co-processor (one per Core)

- Support of INT8, INT16 or FP16 accuracy
- Up to 128 MAC per cycle

Cluster

- 16 x 64-bit Cores + 1 dedicated safety/security Core
- 4 MB of Memory / L2 Cache – 600GB/s Low Latency / High Speed
- Configurable cluster/chip cache coherency modes
- Low Power (600MHz) / Standard (900MHz) / High performance (1.2 GHz) modes

System-on-Chip

- 5 clusters (total of 80 Cores + 5 security Cores)
- Up to 1.15TFLOPs (SP) / 384 GFLOPs (DP)
- Up to 3 TFLOPs (16 bits) / 25 TOPs (8bits) for deep learning
- 40GB/s High Speed/Low Latency Network-on-Chip
- Support standalone or acceleration mode (X86 or ARM Host)
- Support of chip-to-chip connection to scale performance

PCIe Interface

- 16-lane PCIe GEN4 Endpoint (EP) or Root Complex (RC)
- N-furcation up to 8 downstream ports in RC mode
- SR-IOV up to 8 Physical Functions / 248 virtual functions
- Address translation and protection
- Up to 2048 MSI-X & 64 MSI interrupts
- Support for Hot plug
- Up to 512 DMAs for multiqueues / kernel bypass drivers
- Direct PCIe-to-clusters and PCIe-to-DDR transfers
- Multichip extension

DDR4 Interface

- 64-bit LPDDR4/DDR4-3200 channels with sideband/inline ECC
- Up to two ranks per DDR4 Channels
- Support for SoDIMM/UDIMM or soldered discrete devices
- Up to 2 DDR channels with channel interleaving
- Up to 16GB per DDR channel

Offloading

- RDMA protocols RoCEv1/v2
- NVMe-oF storage protocol
- TCP/UDP/IP stateless offload
- Checksum offload
- Interrupt coalescing
- MACSec / IPSec / SSL / TLS offload

Ethernet Interface

- 8x1/8x10/8x25/2x40/4x50/2x100 GbE
- RDMA using RoCEv1/v2
- NVMe-oF (RoCE) - NVMe/TCP
- Jumbo Frame Support (9.6KB)
- Support for PTP 1588v2
- Priority Flow Control (PFC), IEEE 802.1Qbb
- Checksum offload Header & Payload
- Line rate packet classification/load balancing
- Hash & Round-robin based dispatch policy

Security

- Secure Boot with authentication (ECDSA-256) & encryption
- True Random Number Generators (TRNG) compliant with FIPS 140-2, FIPS140-3, SP800-90B
- RSA, Diffe-Hellman, DSA, ECC, EC-DSA and EC-DH acceleration
- Non-volatile Asset store & Embedded Secure Element

Cryptography Accelerators (optional)

- AES-128/192/256 (ECB/CBC/ICM/CTR/GCM/GMAC/CCM)
- AES-XTS for storage application
- MD5/SHA-1, SHA-2, SHA-3
- Kazumi/Snow 3G, ZUC

Management/Control Interfaces

- GPIOs/UARTs/SPI/I2C/CAN/PWM
- SSI Controller for serial NOR Flash with optional boot
- SDCARD UHS-I / eMMC 4.51 memory controller
- 2x USB 2.0 OTG ULPI
- JTAG IEEE 1149.1
- 16-bit Parallel Trace Interface

Safety & Predictability

- Mix criticality support
- Lockable critical configuration
- Capability to bank memory and caches for non interferent & time predictable execution
- LI Cache coherency enabling/disabling

AccessCore® SDK & Programmability

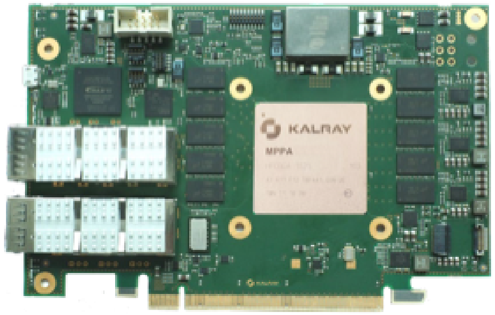
- Support of Eclipse IDE
- Support of open toolchain: GCC, GDB, LLVM
- Programmable core C99 and C++ 2014
- Operating System: Linux and configuration to port any RTOS
- Deep Learning code generator supporting popular AI frameworks (KaNN™, Kalray Neural Network)
- Support of Open Standard Computing Language for deployment of parallel code
- Support of OpenCV with optimized backend
- Support of Mathematical Libraries: FFT, BLAS, LAPACK
- Debug and Trace per cluster and core
- Simulators and profiling tools
- Standard NIC drivers for host CPUs (ibVerbs, VIRTIO, SPDK, ...)

Accelerator card for AI and beyond

Bringing AI & massive computing to the Edge

Limitations of cloud-based AI, such as bandwidth, latency, power consumption, security and cost, have become bottlenecks for the deployment of AI in appliances close to data sources. The Edge is becoming the perfect destination for local processing of data, such as machine learning models trained in the cloud, enriched with context-adaptation and local complex intelligent processing.

Kalray addresses the challenge of distributing the AI and data processing across the Edge, bringing massive localized compute capability to the nodes.



K200™ card with MPPA® processor

APPLICATIONS

Data Centers

Autonomous Vehicles

Video Surveillance

Aerospace & Robotics

Industry 4.0

5G Telecom Infrastructure

KEY BENEFITS

Single chip for whole app acceleration with open APIs

- Standard programmability (C/C++/Open Standard API)
- Open APIs

High performance AI, low latency, low power

- 25 to 100 TOPS
- Multiple concurrent applications: Multi-CNN, computer vision, parallel pre/post processing …
- Low power consumption <20W (MPPA®)
- Real-time data processing
- Scalable multi-MPPA® system support

Safety and high security

- Versatile high perf crypto-engine
- Hardware isolation and freedom of interference
- Secure Boot and root of trust

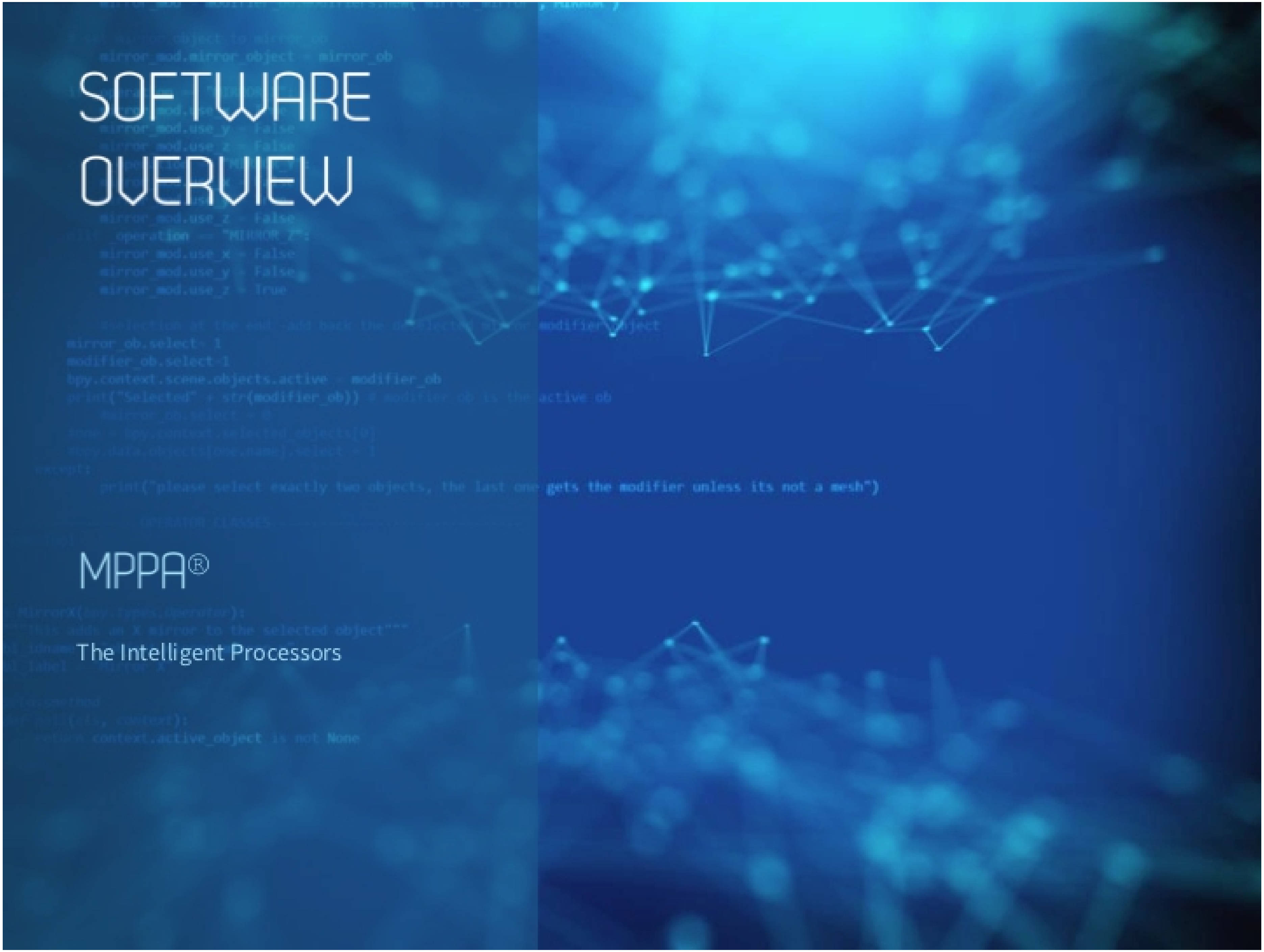
Seamless integration in existing AI ecosystems

- Same software environment as GPU/CPU
- Minimal application changes required
- No new training required
- Frameworks: Caffe, Tensorflow, …
- Networks: GoogLeNet, ResNet, Yolo (v2, v3), Faster-RCNN, custom networks, etc

TurboCard K200^{TM} Specification and Performance

HW specifications	
Power consumption	20W (MPPA®) / 35W (board)
Interface	PCIe Gen4 x16, 2xQSFP28
DDR	2x 4GB DDR4-3200
TOPS	25 TOPS per card (8-bit)
Floating point operations	3 TFLOPS (16b), 800 GFLOPS (32b)
MIPS	192 KDMIPS (Coolidge-80, 1.2 Ghz)
Examples of neural network performance	
GoogLeNet	3025 fps
Faster-RCNN (VGG16)	302 fps
Yolo v3	310 fps

Data based on Coolidge MPPA®3-80

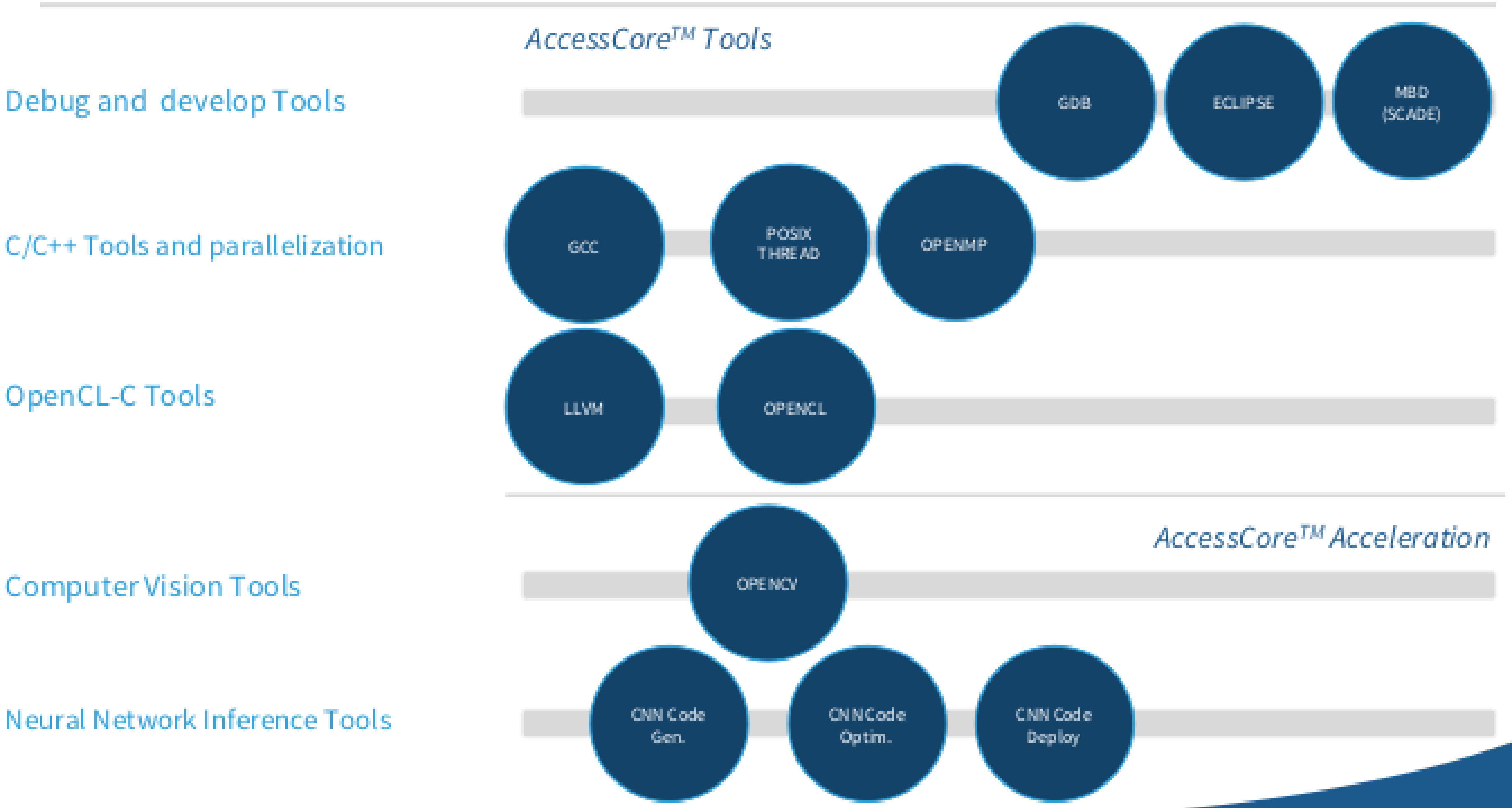


SOFTWARE OVERVIEW

MPPA®

The Intelligent Processors

STANDARD DEVELOPMENT TOOLCHAINS



AccessCore™ Tools

Debug and develop Tools

GDB

ECLIPSE

MBD (SCADE)

C/C++ Tools and parallelization

GCC

POSIX THREAD

OPENMP

OpenCL-C Tools

LLVM

OPENCL

AccessCore™ Acceleration

Computer Vision Tools

OPENCV

Neural Network Inference Tools

CNN Code Gen.

CNN Code Optim.

CNN Code Deploy