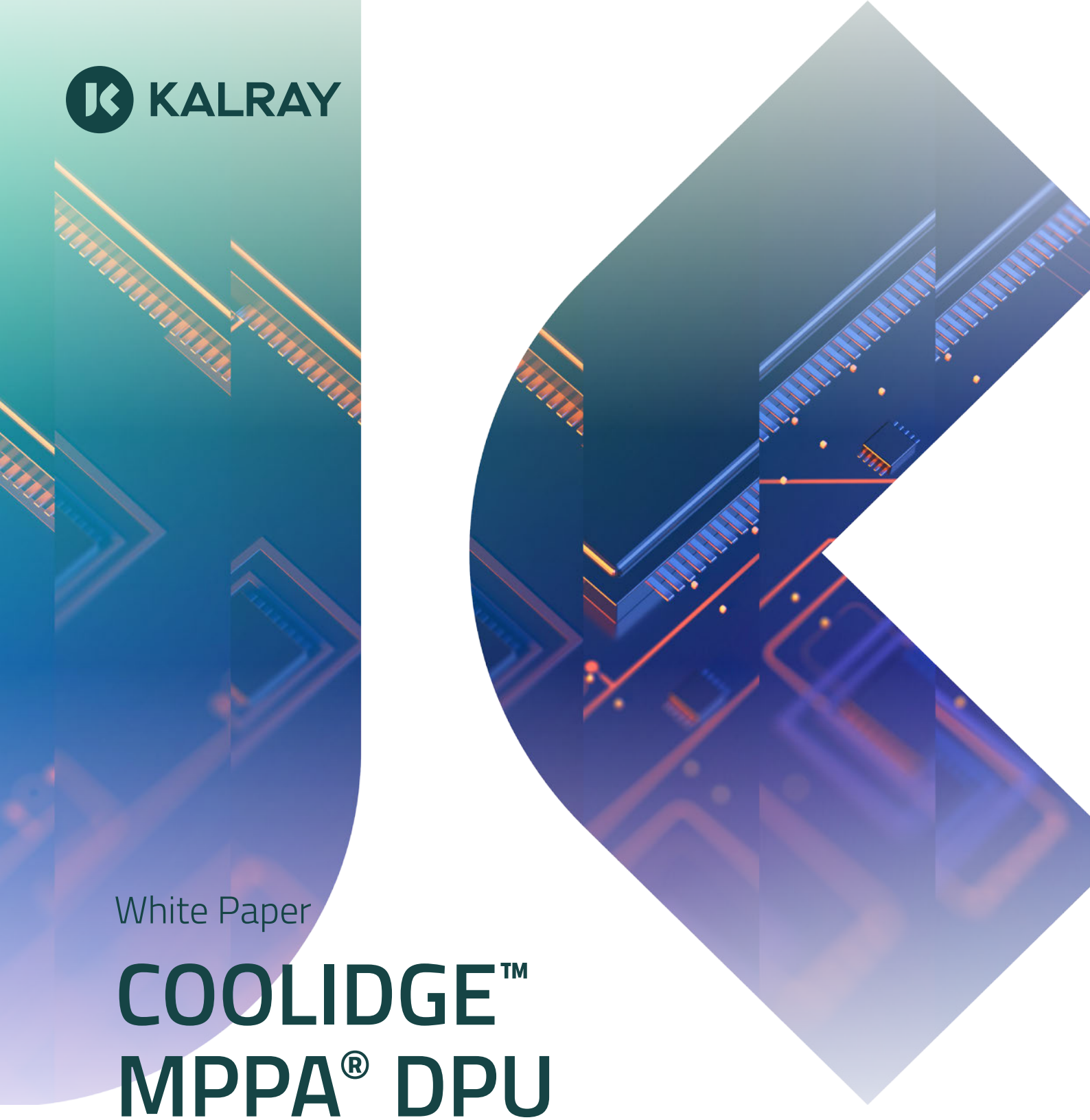
KALRAY



White Paper

COOLIDGE™ MPPA® DPU

Kalray's Unique Processor Architecture

CONTENTS

INTRODUCTION

The world is facing an “explosion of data”. Solving this challenge with current technologies is complicating designs, both in hardware and software, with all the cost impact that entails. Industry needs a new type of processor to address this new challenge in an efficient manner.

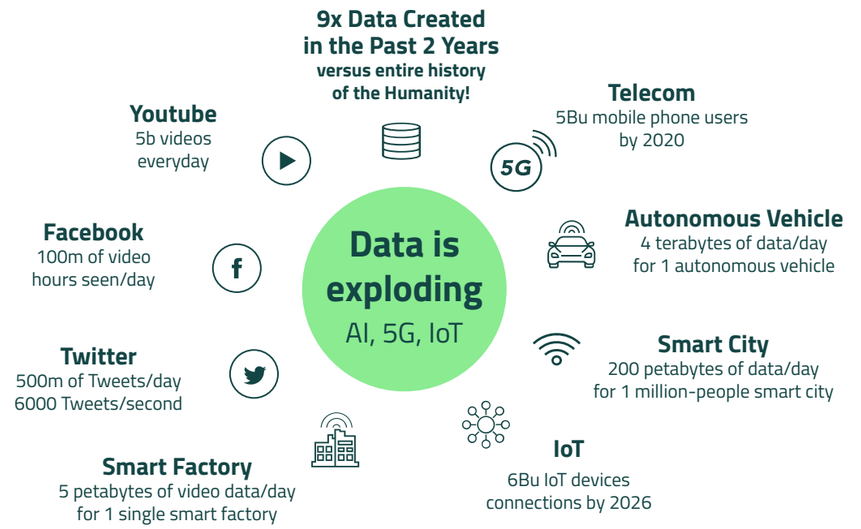
KALRAY

New Types of Processor for New Needs

The speed, the complexity and the quantity of data to analyze is increasing so much that a new generation of processor is needed to face this explosion, both on Cloud and at the Edge.

A few examples: 5 billion videos watched every day on YouTube, more than 1 gigabyte per second created in an autonomous car, more than 200 million gigabytes generated per day in one smart city and more than 5 million gigabytes per day in an intelligent factory.

According to the Cisco Global Cloud Index* 2018-2021, only 25% of data generated today reaches a centralized data center on the Cloud! More and more, data is transient and neither recorded nor stored. This data must therefore be processed on the spot, where it is created, at the Edge of the Cloud. **This is new!**



Computing at the edge (“Edge Computing”) is therefore booming and is complementary to computing in the Cloud. Edge Computing is required when local processing becomes critical for low latency issues (e.g. Augmented Reality /Virtual Reality), for privacy of data (e.g. manufacturing enterprises), for conservation of bandwidth (e.g. smart factory IoT) or for high performance localized compute (e.g. automotive).

Other strong evidence of the surge of the amount of data to analyze is the growth in artificial intelligence semiconductor solutions. According to McKinsey**, it will be multiplied by 4 from now to 2025. Today’s semiconductor market for AI is 99% in the Cloud ($ 5.5 billion) and 1% Edge. By 2025 the market will evolve into a market of $ 5.5 Billion in the Edge and will reach $14B on the Cloud ! Both Cloud and Edge demands are exploding.

**Cisco Global Cloud Index* 2018-2021*
***McKinsey White paper: AI hardware: New opportunities for semiconductor companies, December 2018*

To meet these new needs related to the data deluge, the industry requires a new type of processor that has the capability to capture and perform inline analysis of a very large amount of data, close to where data is generated, to extract usefull information from this flow of data and to react in real time based on this data. This is what we call an “intelligent” processor, because it has the capability to understand its environment and react accordingly. Intelligent processors will be at the heart of “intelligent systems”.

An autonomous vehicle is a very good example of an intelligent system. An autonomous vehicle must analyze a very large flow of data coming from its sensors such as cameras, lidars and radars. From this raw data, it will at the same time recognize a sign, understand in which direction the road is going and predict whether a pedestrian is likely to cross, etc. These are dozens of functions that must be performed in parallel or in sequence, which are very demanding in terms of compute capacity and which must be performed in a guaranteed maximum time for latency constraints.

There is a pressing need for intelligent systems: in the industrial world, surveillance, smart cities, healthcare, autonomous vehicles, telecoms / 5G, etc ... and in the core data centers themselves, where the data must be analyzed as quickly as possible on the fly, as in storage servers or during communication between servers and applications.

For that reason, an intelligent processor needs to support the simultaneous execution of different types of software task such as storage data services, network software stacks, mathematical algorithms, signal processing, data processing, artificial intelligence inference; in a nutshell, it must demonstrate a superior heterogeneous compute ability. Additionally, the intelligent processor must deliver this compute ability with guaranteed response time for each of the tasks, regardless of workload intensity on other tasks in bounded time.

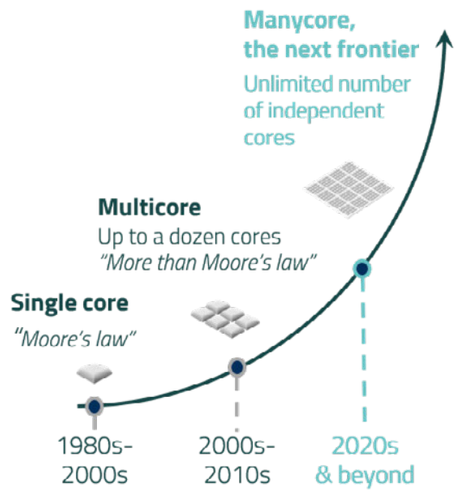
In addition, in systems called Cyber Physical Systems, the safety and security of operation are crucial because these systems control physical entities such as cars, aircraft, drones, robots, or even systems making critical decisions in energy, health and industry.

1.1 The Evolution Towards the Multiplication of Cores

Historically, increasing the performance of processors has meant increasing the frequency of the core, the core being the processing unit.

For the past twenty years (and the introduction of 65nm technology), the improvement in processor performance has been limited by a constraint on the power dissipated per unit area of silicon.

However, it is possible to get around this problem of energy density by performing the processing in parallel, by integrating an increasing number of cores, by specializing these calculation units and by ensuring physical proximity to the data.



A $ 5.5 Billion Market in 2025

Data Centers

Autonomous Vehicles, AD/ADAS

Machine Vision

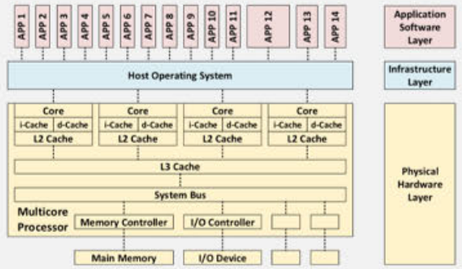
Aerospace & Avionics

Robotics, Industry 4.0

5G Telecom Infrastructure

1.2 Parallel Processing: Multicore Versus Manycore

Homogeneous Multicore Processor



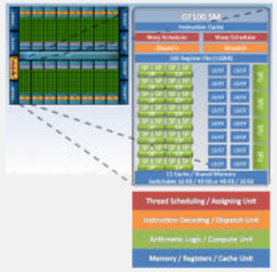
(1)

Multiple CPU Cores Sharing a Cache-coherent Memory Hierarchy
- Scalability by replicating CPU cores
- Standard programming models

Energy Efficiency Issues
- Global cache coherence scaling

Time-predictability Issues
- No scratch-pad or local memories

GPGPU Manycore Processor



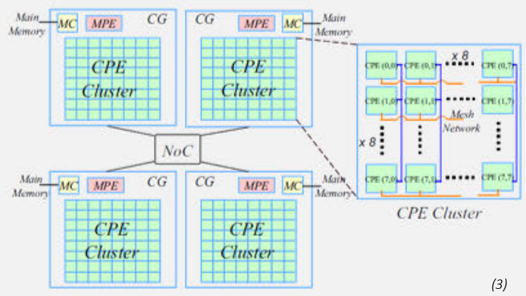
(2)

Multiple Streaming Multiprocessors
- Restricted programming models

Performance Issues of 'Thread Divergence'
- Branch divergence slow down the execution
- Memory divergence: non-coalesced accesses

Time-predictability Issues
- Dynamic allocation of thread blocks
- Dynamic scheduling of warps

CPU-Based Manycore Processor



(3)

Multiple "Compute Units" Connected by a Network-on-Chip (NoC)
- Scalability by replicating Compute Units
- Standard multicore programming inside a Compute Unit

Compute Unit
- Group of cores + DMA
- Scratch-pad memory (SPM)
- Local cache coherency

In the early 90s, the industry turned to multicore architectures.

A multicore processor refers to an integrated logic circuit that contains multiple cores, where these units can run application software. The cores can be identical or specialized, the first case corresponding to homogeneous multicore processors which are the most widespread choice for producing the Central Processing Units (CPUs) of computers. As in a single-core processor, a multicore processor surrounds its cores with a hierarchy of cache memories and busses providing access to the main memory and to the computer's I/O devices.

A cache memory refers to a hardware component located near a core comprising a fast memory associated with a management hardware block. Its function is to speed up the memory accesses of a core by keeping up to date and supplying copies of frequently addressed data in main memory on demand from the core.

(1) https://insights.sei.cmu.edu/sei_blog/2017/08/multicore-processing.html
(2) Nvidia, "NVIDIA's Next Generation CUDA Compute Architecture: Fermi," NVidia, 2009.
(3) Z. Xu, J. Lin et S. Matsuoka, "Benchmarking SW26010 Many-Core Processor," at IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2017.

High-performance processors have up to three cache levels, with a first level usually specialized for the 2 program instructions (i-Cache) and for the processed data (d-Cache).

In the case of a homogeneous multiprocessor, part of the cache hierarchy is common between the cores, from level 2 when the cores are arranged in a cluster, at level 3 otherwise. The main advantage of sharing caches between cores, besides pooling memory capacity, is that this arrangement ensures data consistency as seen by the cores sharing that cache.

Otherwise, logic for maintaining consistency between the cache memories becomes necessary. Indeed, the programming standards for multicore processors such as the "POSIX threads" libraries or the 'OpenMP' compilation directives impose purging of the caches at the synchronization points of the cores when consistency is not ensured, which severely degrades performance.

Maintaining a consistent cache hierarchy on a multicore processor effectively limits the number of cores to a few dozen.

Beyond that, energy efficiency degrades, while interference to cores due to shared resources in the memory hierarchy preclude any accurate prediction of execution times.

The answer to these cache memory problems lies in realizing a processor according to an architecture known as manycore.

A manycore processor is characterized by an apparent grouping from the software point of view of cores and their portion of the memory hierarchy into computing units. This grouping can delimit the scope of cache consistency and inter-core synchronization operations, include explicitly addressed local working memories (as opposed to caches), or even specific data movement engines and other accelerators.

Computing units interact and access external memories and processor I/O through a communication device that can take the form of a Network-on-Chip (NoC).

The advantage of the **manycore architecture** is that a processor can **scale to massive parallelism** by replicating the computing units and extending the network on chip, whereas for a multi-core processor the replication applies to the core level.

1.3 Need for AI, but also for Heterogeneous Multi-processing

Machine learning techniques applied in intelligent systems, often referred to as "artificial intelligence", belong, for the majority of computational needs, to the category of "Artificial Neural Networks" (ANN). This category designates a generic model of realization of non linear classifiers, where the elementary operator realizes a weighted sum of the inputs, then applies a nonlinear function called activation in similarity with biological neurons. The multiplying factors and the bias of the weighted sum are called parameters.

The main specialization of artificial neural networks applies the principles of deep learning to convolutional neural networks (CNN). Deep learning enables models composed of multiple processing layers to learn representations of data at multiple levels of abstraction. Convolutional networks include a majority of processing layers characterized by the invariance of parameters according to certain dimensions of the layer data (such as height and width in image processing).

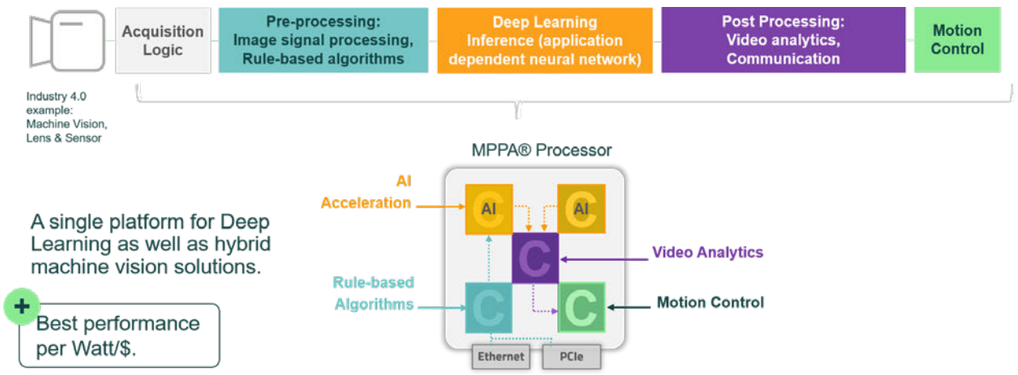
The application of machine learning is a two-step process. The first step adapts the parameters to the expected function by training a model with reference labeled dataset, supervised, unsupervised, or by reinforcement. In the second step called inference, the trained model is deployed and applied to new data, to enable decision making.

From an implementation standpoint, training takes place offline, in data centers on servers accelerated by GPGPU-like processors or specialized hardware, such as Google's Tensor Processing Units. On the other hand, inference can be done online on embedded systems and in particular on intelligent systems.

The AI part is a necessary part of intelligent processors as it is an extremely efficient way to analyse data. But it is far from sufficient. A processor dedicated to AI has only one function: to accelerate calculations related to artificial intelligence. In most of the edge applications, this is clearly not enough.

Heterogeneous Multi-processing on MPPA® DPU Processor: Example of Machine Vision for Industry 4.0

End-to-end use cases usually involve multiple stage processing that can be now processed on a single MPPA® DPU processor.



WHAT IS AN INTELLIGENT PROCESSOR?

An intelligent processor is much more than just an AI compute accelerator, although it integrates these functions very effectively.

It can execute multiple concurrent heterogeneous multi-processing in real time, in a guaranteed timeframe, regardless of overall processor workload.

Manycore technology offers a great approach to support **both efficient AI computing** and **heterogeneous** multi-processing support.

DATA CENTER USE CASE

Data Centers are undergoing a very important revolution with rapidly evolving network architectures and technologies.

KALRAY

The Challenges of Intelligent Systems: Example of Data Centers

2.1 Data Centers Need a Model

Data Centers are undergoing a very important revolution with rapidly evolving network architectures and technologies. This is mainly due to the explosion of usages, the surge of data to be processed and the resulting exponential growth of the number of machines to be managed to support this growth.

There is a multitude of challenges in terms of diversity of users, data, type of applications and associated processing, size, location and purposes, while trying to cope with always-increasing needs in term of performance, pressure on cost and return-on-investment.

To continue to bring high value to their demanding customers, data centers are seeking composable, highly efficient and heterogeneous multi-processing and adaptive solutions, both on storage nodes and compute nodes.

2.2 Need for DPU - Data Processing Unit and More

General purpose CPU's and OS' have now shown their limitations for decades. As an example, ~25% of the server power is today spent in data centric computation like storage stacks or network stacks or crypto, showing their inefficiency for data centric computation. In addition, general purpose CPUs can only support single threaded user applications preventing massivey parallel workloads being run properly.

GPGPU does exploit this parallelism approach. Primarily optimized for graphic, they have demonstrated their value in supporting matrix operations for artificial Intelligence but they have not been designed for dealing with multiple workloads and their data efficiently.

There is therefore a clear need for a new class of processing accelerator for these predominantly data-centric heterogeneous processing tasks and offload the main CPU (x86 or ARM).

Intelligent Processors have the capability to process a huge flow of data in parallel and to apply heterogeneous type of data processing on such a flow. They are the perfect candidates to address the need of DPU. And they can do more…

2.3 Need for Intelligent Processor on Storage

Until now, storage has been one major bottleneck of the increasing performance needs in Data Centers. The introduction of flash memory-based drives (Solid-State Drive or SSD) with hyper-fast communication protocols such as NVMe® and NVMe-oF™, offers breakthrough solutions to the industry to massively scale up and increase performance of existing data centers. Smart storage adapters and associated composable storage appliances, such as Fabric attached just a Bunch of Flash or "FBOF", are at the heart of this revolution, targeting the new generations of hyper-fast storage solutions

and data intensive application such as AI, Data Analytics or IoT.

However, to sustain the peak performance of this new generation of memory, there is a need to displace the data processing such as encryption, data reduction, compression from the main CPU to a much more efficient "data cruncher" processor, close to the data.

More generally, whereas the conventional approach was to move data from where its stored to where its processed, in the main CPU, new generation architectures tend to rely on "in-situ" processing or computational storage: compute data where it resides.

The Intelligent Processor will be at the center of this evolution of the data center architectures.

Smart storage adapter cards, based on Intelligent Processors, can therfore be configured in order to offload the main CPU (x86) from demanding functions such as NVMe, NVMe-oF, Erasure Coding or KVS, in-line security services offloading (e.g. SSL-TLS or IPSEC), mathematical algorithms (e.g. FFT, Monte Carlo) or AI with the support of multi-CNN.

Performance, resiliency, compute capabilities for offloading data services are important, but the need for an open and easily programmable platform is also a must have to provide the requested flexibility data center architects have been looking for. For that reason, Intelligent processors have a clear advantage compared to other alternatives such as GPU and FPGA, in helping system architects to build their next generation data center solutions.

Build Next-Gen Storage Today with Kalray's Smart Cards

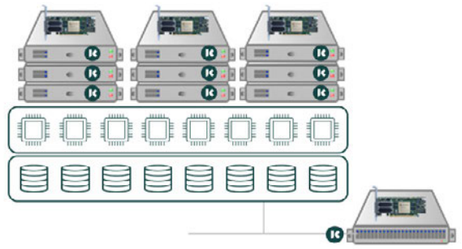
❶ Standard NVMe-oF TCP & RoCE

❷ Easy integration on x86 nodes via SR-IOV NVMe Emulation

❸ Drive access via PCIe (RC or P2P)

❹ Management via GbE

❺ AccessCore® Storage Framework on MPPA® to deliver Data Services



Example of Composable Disaggregated Infrastructure (CDI) with Kalray Smart Cards

Intelligent Processors have the capability to process a huge flow of data in parallel and to apply heterogeneous type of data processing on such a flow. They are the perfect candidates to address the need of **DPU**.
And they can do more…

3

AUTOMOTIVE USE CASE

Performance, aggregation of functions and safety are the keys for upcoming intelligent vehicles.

KALRAY

The Challenges of Intelligent Systems: Example in the Automotive Industry

The automotive industry is currently facing 3 major challenges. The first one is to address the performance needs required by the demanding compute nature of ADAS (Advanced Driving Assistant System) and AD (Autonomous Driving). Secondly, the need to simplify the overall system architecture by reducing the number of other ECU's (Electronic Component Unit), other words, break the "yet-another ECU for another function" approach. Finally, safety, for obvious reasons.

3.1 Need for High Performance

Car manufacturers estimate that a fully autonomous vehicle (Level 5) would require computing power greater than 200 TOPS (Tera Operations Per Second) for the inference of artificial intelligence in perception functions, and 10s of TFLOPS (Tera Floating-Point Operations Per Second) for numerical calculations in trajectory planning functions.

Even if it takes time to see fully autonomous vehicles on the road, partially autonomous vehicles (Level 3 to Level 4) still require a huge amount of computing performance for a fraction of the power a standard CPU consumes. Significant advances are therefore needed to improve energy efficiency, both in electronic technology and in computer architecture. For the latter, the main path lies in the exploitation of massive parallelism.

3.2 Function Aggregation is Key

Obviously, running such a complex system as an autonomous vehicle requires combining a multitude of different types of computing functions within one single electronic component while ensuring that each of these functions are isolated from each other.

This is an important feature of manycore processors. Kalray's MPPA® (Massively Parallel Processor Array) latest generation of manycore, is a single component providing 80 cores (CPU-type) coupled with 80 accelerators (AI type/ Math-type), grouped into clusters. Each cluster (a group of 16 cores with 16 acceleration co-processors) can run independently thanks to a communication fabric ensuring isolation.

The number of powerful compute units, combined with the capability of parallelization of execution (core, cluster and chip level) provides

the performance that can scale from ADAS to AD vehicle (from Level 2 to 5). The spatial isolation is ensured by hardware design. It provides freedom-from-interference and removes the need for a software hypervisor to reach ISO 26262 ASIL-B level.

With Kalray's processors, system architects can now consider flexibility and scalability not only at ECU level but at component level.

3.3 Safety, a Critical Priority

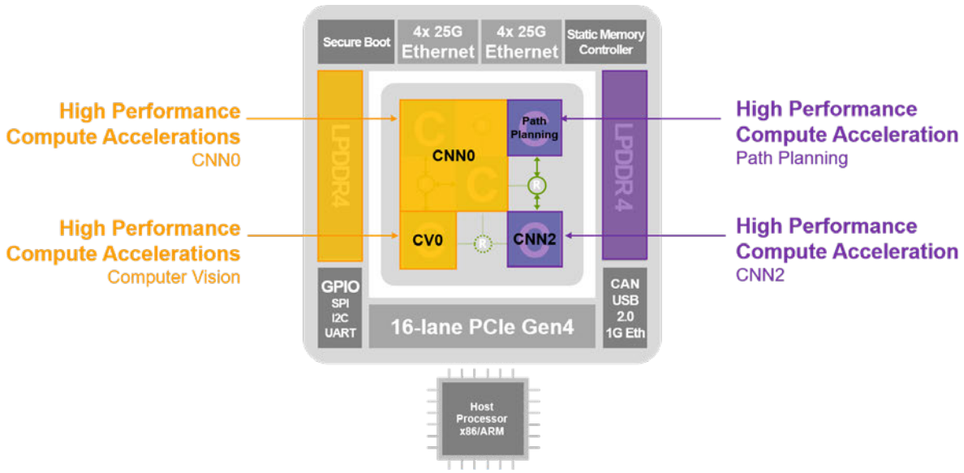
Intelligent systems are expected to not only deliver high computing power at very low power consumption, but also to bring a high level of safety.

Specifically, it is about ensuring the continuity of service (availability), responding in the allotted time to avoid catastrophic failures (operational reliability), and resisting deliberate intrusions (digital security). As we will detail, the real-time requirement and operational safety aspects constitute the major challenges with which designers of parallel computers are confronted.

A system is qualified as real time if it guarantees to respond to an input stimulus within a limited time specified in advance. In a software-controlled system, response time depends not only on the intrinsic performance of the cores, but also on interference that can delay the execution of the task that processes the stimulus. In particular, this concerns the time sharing of a core between several tasks, as well as conflicts between cores for access to common hardware resources (communiction bus, memory elements). In addition, modern cores are equipped with features designed to increase performance by exploiting run history, such as branch predictors and cache memories.

Obtaining reliable bounds on the response time of a task has thus become difficult on modern multicore processors, but remains achievable when a core and its memory hierarchy have been designed not to exhibit any temporal anomaly.

Performance and Heterogeneous Multi-Processing on MPPA® DPU Processor: Run Multiple High Performance Accelerations Simultaneously



The role of safety mechanisms is to detect and then correct errors or prevent their propagation. These mechanisms, such as two-core lockstep execution, are present in automotive components that host critical functions.

For critical avionics applications, the components are also monitored by an external device called a safety net. The need to contain errors, locate the scope of safety mechanisms and observe operation from the outside in a non-intrusive manner, are currently only fully met by specialized processors limited to a small number of cores.

Although demonstrating the advantages of manycore architectures in terms of performance and energy efficiency by exploiting massive parallelism, GPGPU processors have intrinsic limitations which prevent their generalization in intelligent systems and in particular their difficulty in guaranteeing maximum response times. Kalray's Manycore processors do not suffer from these limitations.

Massively Parallel Processor Array

THE MPPA® DPU PROCESSOR ARCHITECTURE

KALRAY

The MPPA® DPU Processor Architecture

4.1 A Massively Parallel Architecture

The overall architecture of the Kalray third generation of MPPA® DPU processor aka Coolidge™ is based on a "Massively Parallel Processor Array" architecture, which is characterized by the association of computing clusters connected to each other, to the external memory and to the I/O interfaces via two independent interconnects.

The MPPA® DPU processor has 2 global on-chip interconnects, suited to different types of data transfers. The first interconnect is an AXI Fabric bus grid, for use read/write access from cores to memories and peripherals connected by PCIe. The second interconnect by an RDMA NoC (Network-on-Chip), that supports data transfers to or from the Ethernet network interfaces and connect all clusters together.

The robust partitioning necessary for safe operation of the processor is carried out at the granularity of the computing cluster and is based on the configuration of memory management units (MMUs), memory protection units (MPUs), and on the deactivation or not of network on chip links.

4.2 An Innovative Cluster Partition

The Coolidge™ processor cluster is partitioned between a secure area and an user application area. The secure area includes a core (RM core) dedicated to security and safety functions, associated with an isolated memory bank.

The user application zone brings together 16 processing cores (PE cores) and a data movement engine (DMA) connected to a 4MB local memory called SMEM and composed of 16 banks.

The user application zone brings together 16 processing cores (PE cores) and a data movement engine (DMA) connected to a 4MB local memory called SMEM and composed of 16 banks.

The 16 processing cores operates in two modes:

- **An SMP (Symetric Multi-Processing) mode** intended for high performance applications, where the PE cores behave like a multicore processor CPU. In this mode, the SMEM memory is partitioned between SPM (Scratch Pad Memory) working memories and a level 2 cache, while consistency is enabled between the level 1 data caches of the cores. Adjacent memory accesses are distributed over all banks, with interleaved addressing similar to that of local memory in an streaming multiprocessor in a GPGPU processor.
- **An AMP ("Assymetric Multi-Processing") mode** intended for real-time applications, where the PE cores behave like sixteen independent single-core CPUs. In this mode, all SMEM memory is configured as SPM, while core level 1 data cache consistency is disabled. Adjacent memory addresses run through an entire bank before moving on to the next, allowing software to locate code and data for a job on a particular bank and associate a single core with it.

4.3 Kalray's VLIW Core

The cores used by the Coolidge™ processor all implement the same architecture, of VLIW ("Very Long Instruction Word") type. VLIW architecture is used on embedded processors for signal and predictability, as well as increased

resilience to Meltdown and Spectre security attacks. On a VLIW core, the exploitable parallelism between the instructions is detected by the compiler then explained in the binary code by marking the packets of instructions that can be executed in parallel. This allows cores with precise temporal behavior, more compact for a given processing capacity, which also allows the integration of a greater number of cores.

Although already performing very well in numerical commuting, the Coolidge™ MPPA® DPU processor has been significantly enhanced to increase its performance in neural network inference applications.

4.4 Tightly Coupled Accelerators

The solution adopted is to tightly couple each core to a coprocessor specializing in performing mixed precision matrix operations.

The data is transferred in blocks of 32 bytes between the memory and the registers of the coprocessor, according to the flow of the program executed by the core. When this data is processed by the coprocessor, it is interpreted as arrays of four rows and a varying number of columns depending on the size of the elements: integers between 8 and 64 bits, 16 or 32 bit floating point number.

Operating on two-dimensional data allows the coprocessor to achieve high computational intensity, up to sixteen dot products between vectors of eight elements and sixteen accumulations per cycle.

4.5 Very High-Speed Interfaces

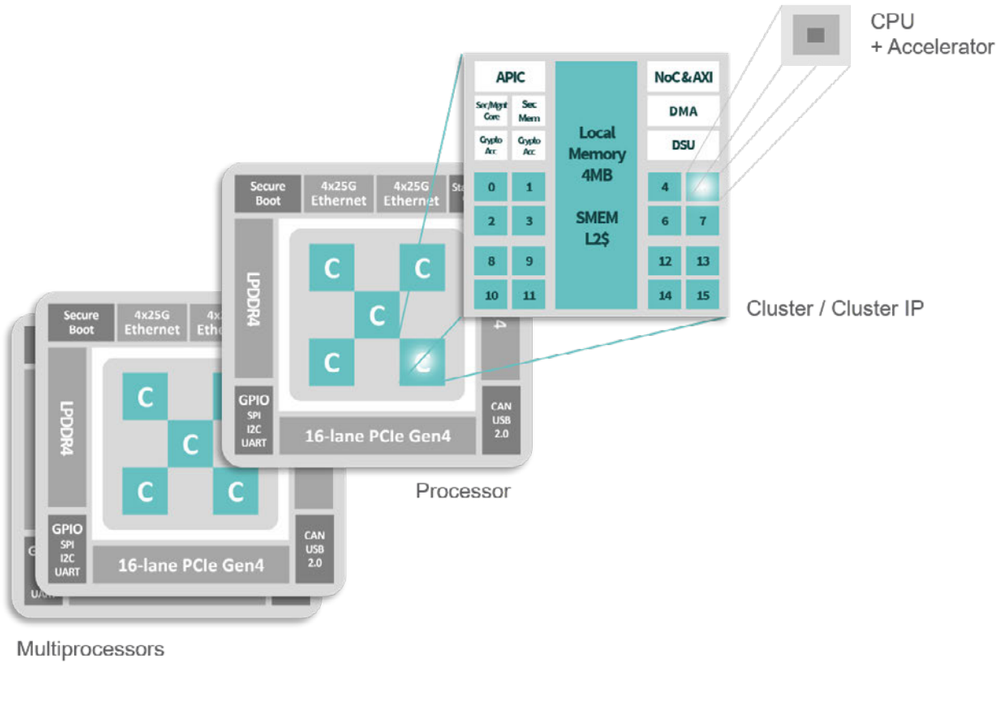
Data centers are undergoing a very important revolution with rapidly evolving network architectures and technologies. This is due to the explosion of data to be processed, the exponential growth of the number of machines to be managed to support this growth, an ever-increasing need for acceleration and some technological breakthrough (e.g. NVMe-over-Fabric to support of Solid-State Drive/SDD or flash memory versus Hard Drives/HDD and disaggregation).

Kalray's solution offers both extremely high computing power capable of processing considerable data volumes while minimizing energy consumption; as well as on-the-fly heterogeneous processing capabilities. The main product is a standard PCIe card, highly configurable for composability needs, based on Kalray's MPPA® DPU processor implementing leading edge interfaces, including PCIe GEN4 x16 and 2x100G Ethernet. As an example, coupled with the latest PCIe GEN4 AMD processor, MPPA® DPU can be used as an accelerator delivering a full duplex bandwidth of up to 256 Gbit/s.

4.6 Scalability: From Cluster and Cluster IP to Multi-processors

One of the key benefits of Massively Parallel Processor Array architecture used on manycore processors is the ability to scale from 1 to N clusters. This was indeed one of the key triggers of the invention of manycore processors i.e. integrate hundreds of cores into a single piece of silicon with a high capability to scale and as a deliverance from the limitations of multicore processors as presented earlier in the paper.

At chip level, scalability still exist i.e. you can add multiple MPPA® DPU processors to increase performances of your system. Either within a monolithic implementation or multi-chip implementation, cluster scalability capability is unique to the MPPA® DPU architecture.



Kalray MPPA® DPU Manycore Processor: From Cluster/Cluster IP to Multi-Processors

In conclusion, taking full advantage of Kalray's patented MPPA® DPU (Massively Parallel Processor Array) architecture and 16nm FinFet technology, the MPPA® DPU Coolidge™ processor is a scalable 80-core processor designed for intelligent systems.

It offers a unique alternative to conventional approaches such as GPU, ASIC and FPGA, bringing unique value to multiple applications from Data Centers, to Edge or Embedded systems.

MPPA® DPU MAIN FEATURES

80-Cores Architecture

Core

64-bit/32-bit architecture

From 600MHz to 1.2 GHz

6-issue VLIW

Instruction data cache with MMU

IEEE 754-2008 Floating Point Unit (FPU)

Up to 256-bits per cycle Load/Store

Co-Processor (One per Core)

Acceleration of INT8, INT16 or FP16 accuracy

Power & Cooling

16 Application Cores + 1 Management/Security Core

4 MB of Memory / L2 Cache – 600GB/s Low Latency / High Speed

Configurable cluster/chip cache coherency & deterministic modes

System-On-Chip

5 clusters (total of 80 Application Cores + 5 Management Cores)

Up to 1.15 TFLOPs (SP) / 384 GFLOPs (DP)

Up to 3 TFLOPs (16 bits) / 25 TOPs (8bits) for Deep Learning

PCIe Gen4 Interface

16-lane PCIe GEN4 Endpoint (EP) or Root Complex (RC)

Bifurcation up to 8 downstream ports in RC mode

SR-IOV up to 8 Physical Functions / 248 Virtual Functions

Address translation and protection

Support for Hot Plug

Support for NVMe and VIRTIO emulation

LPDDR4/DDR4 Interface

64-bit DDR4/LPDDR4-3200 channels with sideband/ inline ECC

Up to two ranks per DDR4 Channel

2 DDR channels (up to 32GB) with channel interleaving

2x100GbE Ethernet Interface

8x1/8x10/8x25/2x40/4x50/2x100 GbE

RDMA over Converged Ethernet (RoCE v1 and v2)

Support for PTP/IEEE 1588v2

Priority Flow Control (PFC), IEEE 802.1Qbb

In-line packet classification / smart load balancing

Security

Secure Boot

True Random Number Generators (TRNG)

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

Parallel Trace Interface

Safety & Predictability

Mix criticality support

Lockable critical configuration

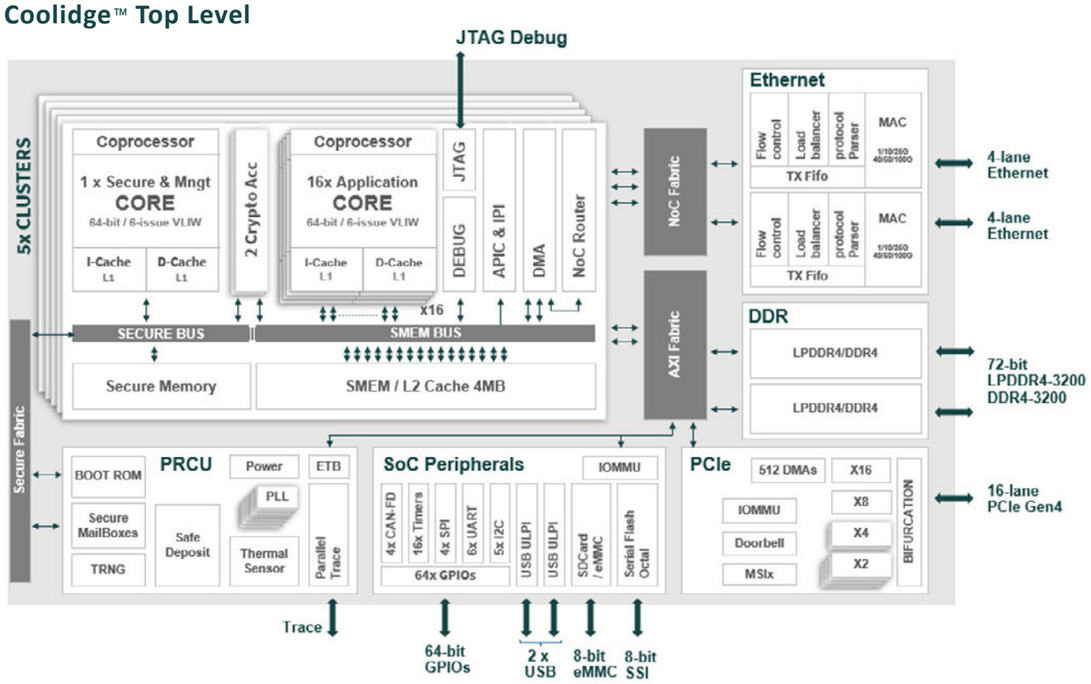
Capability to bank memory and caches for non-interference & time-predictable execution

L1 Cache coherency enabling/disabling

Kalray software development kit, relying on standard tools and open standards provides non proprietary environment to developers to enable seamless port of existing applications or easy development of new ones. Using, C/C++, OpenCL-C, using on GCC or LLVM, using GDB and Eclipse UI, all developers will find a familiar environment to start with.

MPPA® DPU COOLIDGE™ BLOCK DIAGRAM

Coolidge™ Top Level



Coolidge™ is composed of 5 clusters, each with 16 user cores plus 1 management/safety core.

Need more performance?

Just connect several MPPA® DPU processors together to reach the level of performance you need.

MPPA® DPU, a Kalray patented technology.

MPPA® DPU KEY BENEFITS

High Performance Computing

AI Acceleration

Very High-Speed Interfaces
PCIe Gen4x16,
2x100G Ethernet

Real-time Data Processing

Standard Programming
C/C++/OpenCL™,
Linux, POSIX, RTOS

Power Efficiency

Heterogenous Multi-Processing

Security/Safety
Determinism, Freedom from Interference, Secure Boot

CONCLUSION

The speed, the complexity and the quantity of data to analyze is increasing so much that a new generation of processor is needed to face this explosion, both on Cloud and at the Edge.

Kalray Intelligent Processors have been designed with that purpose in mind. Based on its unique and patented MPPA® manycore architecture, Kalray MPPA® DPU has the ability to perform many massively parallel intensive processing tasks, including AI, while delivering low latency and energy efficiency as well as meeting security needs. In addition, unlike GPGPU and FPGA, MPPA® DPU processors are programmed using standard langages and OS.

Kalray processors offer a unique solution to build your next generation of intelligent systems.

THE AUTHORS

Benoît Dupont de Dinechin

Chief Technology Officer (CTO), Kalray

Benoît is the Kalray VLIW core main architect, and co-architect of the Kalray Multi-Purpose Processing Array (MPPA®). He is also a direct contributor to several components of the AccessCore®software development environment. Before joining Kalray, Benoît was in charge of Research and Development for the STMicroelectronics Software, Tools, Services Division. He was promoted to STMicroelectronics Fellow in 2008. Prior to STMicroelectronics, Benoît worked at the Cray Research park (Minnesota, USA), where he developed the software pipeliner of the Cray T3E production compilers.

Benoît earned an Engineering Degree in Radar and Telecommunications from the Ecole Nationale Supérieure de l'Aéronautique et de l'Espace (Toulouse, France), and a Doctoral Degree in computer systems from the University Pierre et Marie Curie (Paris) under the direction of Prof. P. Feautrier. He completed his post-doctoral studies at the McGill University (Montreal, Canada) at the ACAPS Laboratory led by Prof. G. R. Gao. Benoît has published over 50 conference papers, journal articles and book chapters, and holds 10 hardware patents.

Loic Hamon

VP Corporate Development and Strategic Marketing, Kalray

Before his time at Kalray, Loïc spent 10 years at Inside Secure. The last position he occupied was Executive Vice President of Corporate Development and communication.

Prior to that, he was the vice president of Inside Secure's NFC business unit. Loïc also served as Director of Strategic Marketing for the wireless business unit of Texas Instruments after holding several strategic and operational marketing positions within TI.

Loïc has earned a Master's Degree in Marketing Intelligence at the HEC School of Management in Paris. He has also been awarded a Master's Degree in Electrical Engineering from ESIGELEC in Rouen and a postgraduate degree in Microelectronics from Paris XI University.

ABOUT KALRAY

Kalray (Euronext Growth Paris - FR0010722819 - ALKAL) is a fabless semiconductor company, a leading provider of a new class of processors, specialized in Intelligent Data Processing from Cloud to Edge. Kalray's team have created and developed its leading edge technology and products to help its clients maximize the market possibilities presented by a world dominated by massive, disparate and pervasive data.

Thanks to Kalray's patented manycore architecture, Kalray's MPPA® Intelligent Data Processors are natively capable of managing multiple workloads with no bottlenecks to enable smarter, more efficient and energy-wise data-intensive applications. Kalray's offering includes processors, acceleration cards with associated software environment and appliances, allowing its customers to design the best solutions in fast growing sectors such as modern data centers, 5G, AI and Edge Computing, autonomous vehicles and others.

Founded in 2008 as a spin-off of CEA French lab, with investors such as Alliance Venture (Renault-Nissan-Mitsubishi), Safran, NXP Semiconductors, CEA and Bpifrance, Kalray is dedicated through technology, expertise and passion to offer more: More for a smart world, more for the planet, more for customers and developers.

Kalrayinc.com
contact@kalrayinc.com


KALRAY
THE POWER OF MORE
Intelligent Data Processing,
From Cloud to Edge
Kalrayinc.com