KALRAY

Kalray MPPA®

Massively **P**arallel **P**rocessor **A**rray

Revisiting DSP Acceleration with the Kalray MPPA Manycore Processor

Benoît Dupont de Dinechin, CTO
HotChips 2015

MPPA® MANYCORE Processor Roadmap

2014	2015	2016	2017

MPPA®-1024

Volume

Samples Volume

Samples Volume

MPPA®-256

MPPA®-64

TSMC 28nm (HP)		28nm/16nm
Andey	**Bostan**	**Coolidge**
Kalray 1st generation	**Kalray 2nd generation**	**Kalray 3rd generation**
211 GFLOPS SP	845 GFLOPS SP	1056 GFLOPS SP
70 GFLOPS DP	422 GFLOPS DP	527 GFLOPS DP
32-bit addressing	64-bit addressing	64-bit addressing
400Mhz	800Mhz	1000Mhz

Production

Development

Under Specification

Accelerators and Specialized Processors

- Field-Programmable Gate Arrays (FPGA)
  - Most effective on bit-level computations
  - Require HDL programming
  - **Suitable for time-critical computing**
- Digital Signal Processors (DSP)
  - Most effective on fixed-point arithmetic
  - Require low-level programming
  - **Suitable for time-critical computing**
- Graphics Processing Units (GPU)
  - Most effective on regular computations with dense memory accesses
  - Require CUDA or OpenCL programming
  - **Unsuitable for time-critical computing**
- Intel Many Integrated Core (MIC)
  - Require multicore programming + exploitation of SIMD instructions (AVX)
  - **Unsuitable for time-critical computing**

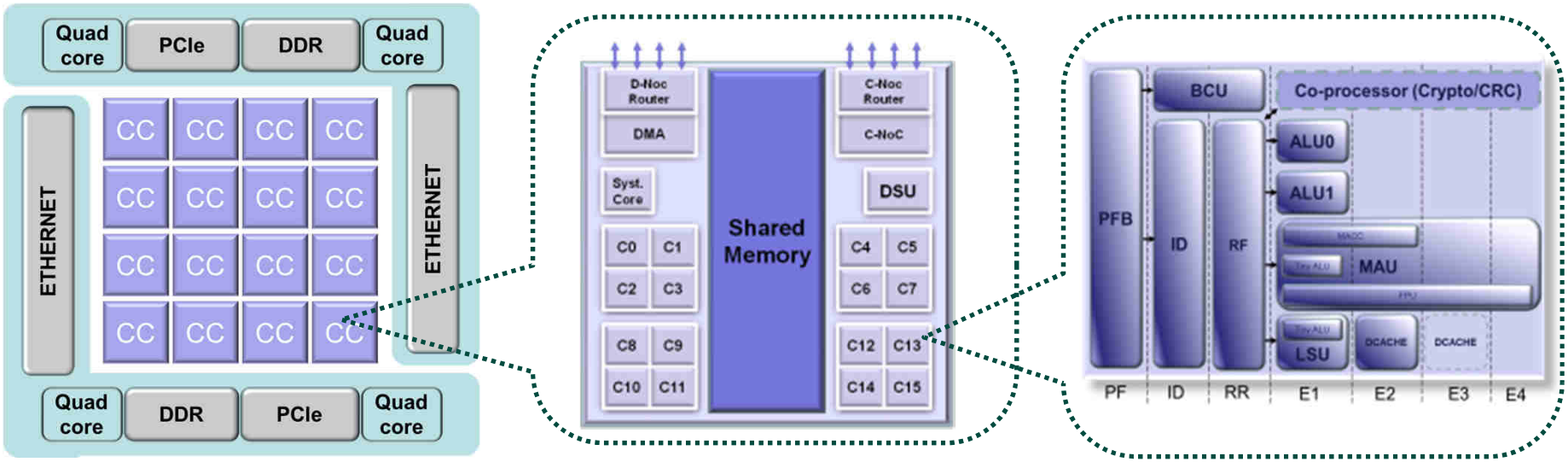
Time-Critical Computing

- Application requirements
  - Time constraints associated with information manipulation
    - Acquisition, processing, transport, storage, coordination, delivery
  - Execution time determinism, predictability, composability
  - Certification of worst-case response time by static analysis
- Execution timing issues
  - Intra-core non-deterministic timing: speculation, branch prediction
  - Inter-core competition for shared resources: caches, busses, devices
  - Ordering of task executions in critical sections
  - Cache turbulence caused by interrupt processing
- Application domains considered
  - Aerospace & defense; autonomous vehicles;
  - Financial trading; large-scale physical instrumentation;
  - industrial robotics; manufacturing equipment;

MPPA® MANYCORE Architecture Highlights

- DSP type of acceleration
  - Energy efficiency
  - Timing predictability
  - Software programmability
- CPU ease of programming
  - C/C++ GNU programming environment
  - 32-bit or 64-bit addresses, little-endian
  - Rich operating system environment
- Integrated many-core processor
  - 32 management cores on chip
  - 256 application cores on chip
  - High-performance low-latency I/O
- Scalable massively parallel computing
  - MPPA® processors tiled together through NoC extensions

MPPA®-256 Bostan Processor Architecture



Manycore Processor

- 16 compute clusters
- 2 I/O clusters each with quad-core CPUs, DDR3, 4 Ethernet 10G and 8 PCIe Gen3
- Data and control networks-on-chip
- Distributed memory architecture
- 634 GFLOPS SP for 25W @ 600Mhz

Compute Cluster

- 16 user cores + 1 system core
- NoC Tx and Rx interfaces
- Debug & Support Unit (DSU)
- 2 MB multi-banked shared memory
- 77GB/s Shared Memory BW
- 16 cores SMP System

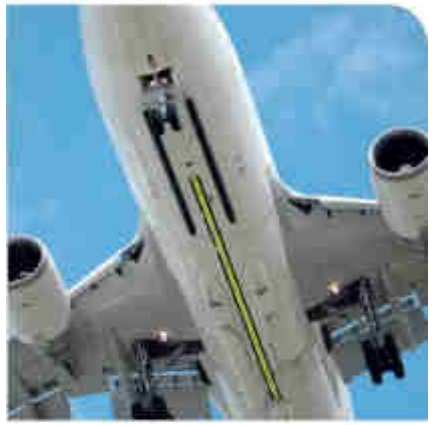
VLIW Core

- 32-bit or 64-bit addresses
- 5-issue VLIW architecture
- MMU + I&D cache (8KB+8KB)
- 32-bit/64-bit IEEE 754-2008 FMA FPU
- Tightly coupled crpto co-processor
- 2.4 GFLOPS SP per core @600Mhz

MPPA® Processor Co-Design for Avionics

- U. Saarland / Absint GMBH recommendations on VLIW core and cache micro-architecture design
  - Absint provides the aiT static timing analysis tool used to certify the flight control system of Airbus A380, Airbus A350 and Airbus A400M
  - Absint aiT tool also targets the Kalray VLIW cores
- Architecture with a focus on timing predictability
  - Core level: micro-architecture
    - ✓ Fully timing compositional core
    - ✓ LRU caches and cache bypass memory accesses
  - Cluster level: multi-banked shared memory
    - ✓ Core-private buses for memory bank access
  - Processor level: NoC with guaranteed services
    - ✓ Minimum bandwidth & maximum latency

CERTAINTY



Certification of Real Time Applications desigNed for mixed criticaliTY

THALES

UNIVERSITE JOSEPH FOURIER

ARTTIC

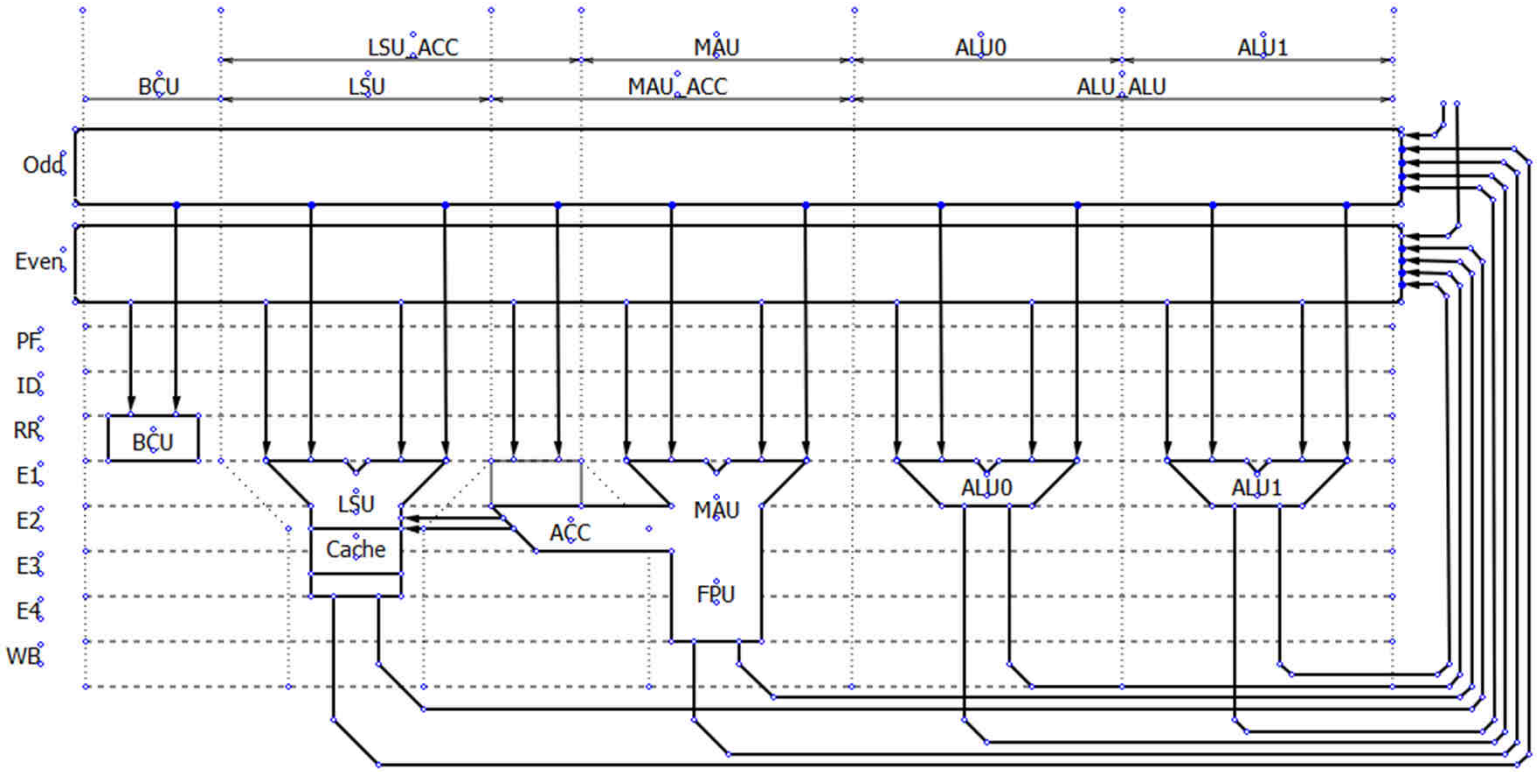
Technische Universität Braunschweig

AbsInt

UPPSALA UNIVERSITET

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

MPPA®-256 Bostan VLIW Core Data Path



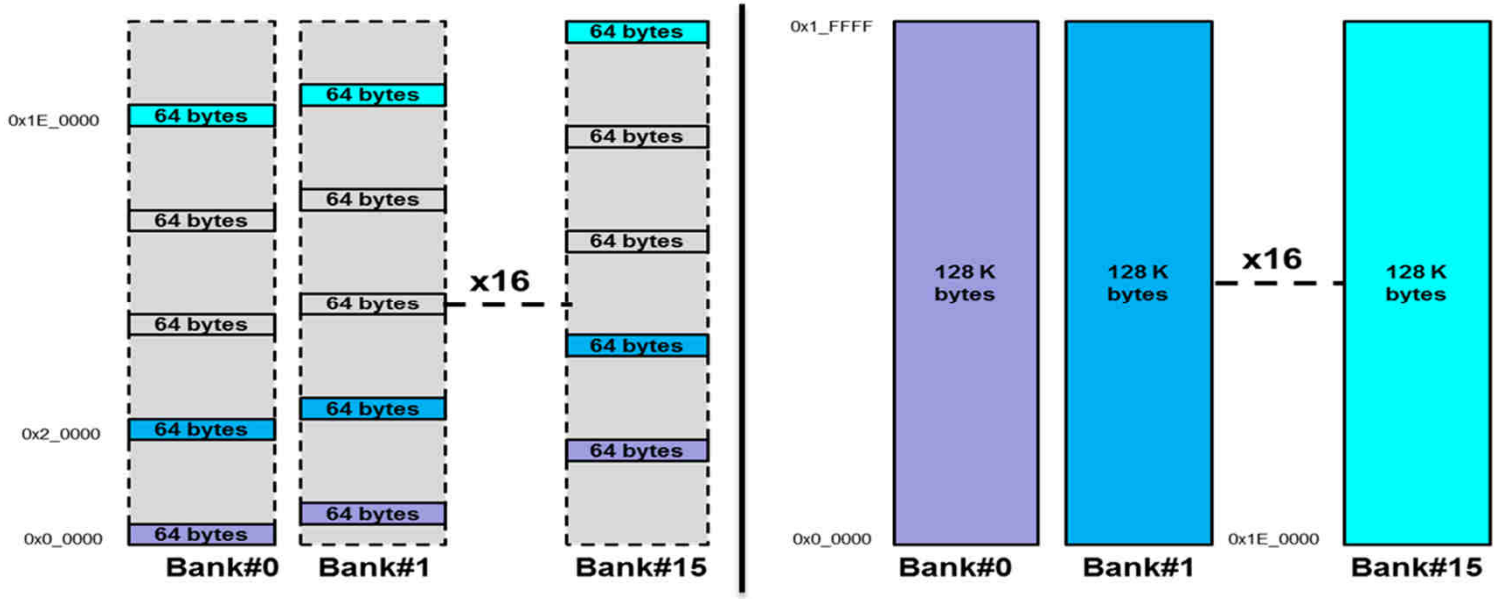
- 5-issue, polycyclic property
- Unified 10 read, 5 write 64x32-bit register file
- 64-bit data in register pairs
- Shared ACC read port
- Unified MAU and FPU
- LSU and MAU also execute single-cycle ALU instructions

Kalray VLIW Architecture Compared to Fisher's Lx

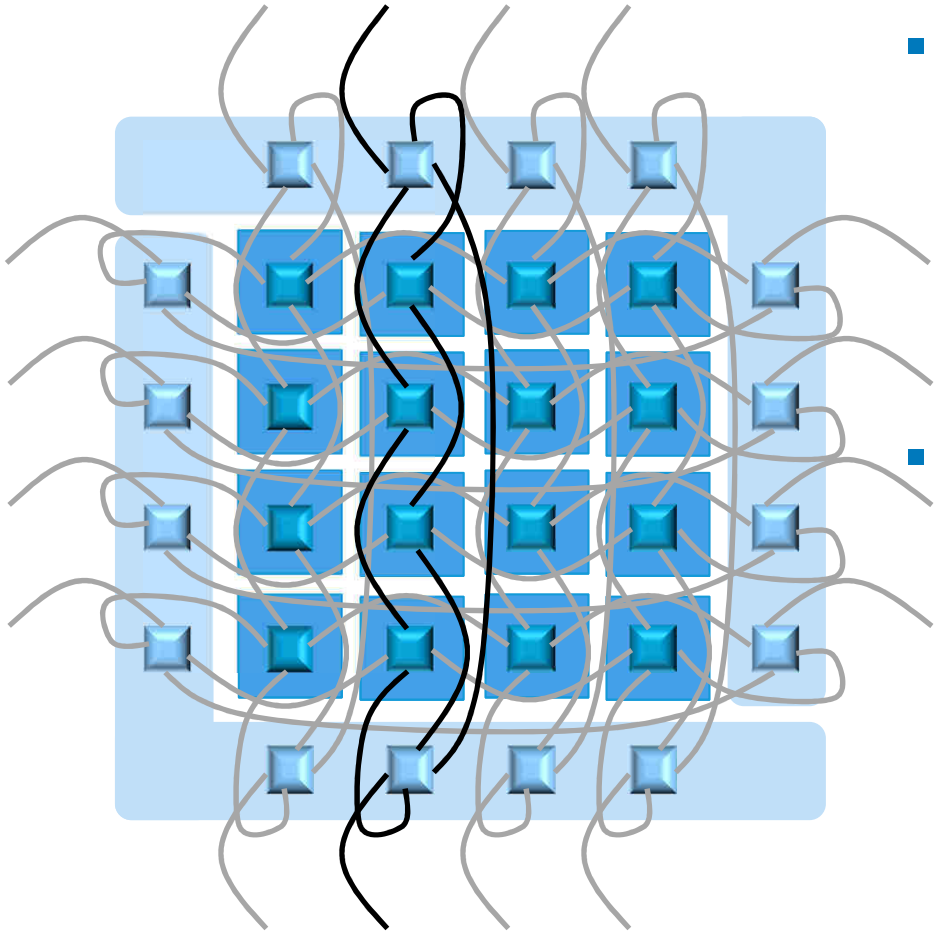
- Optimize use of the data memory bandwidth
  - Widening to 64-bit, no alignment restrictions
  - Enable large immediate values in instruction stream
  - All memory accesses may bypass the L1 data cache & write buffer
- Eliminate DLX ISA features and restrictions
  - Instructions with 3 or 4 source operands, 1 or 2 target operands
  - No aliasing between registers and special resources (LR, zero)
  - Memory addressing modes similar to those of PowerPC
  - Effective floating-point support with Fused Multiply Add
- Rework if-conversion support
  - Remove Boolean registers and SELECT instructions
  - Use CMOV and conditional load/store instructions
- Support hardware looping

MPPA®-256 Bostan Compute Cluster

- 20 bus masters
  - 16 application cores
  - 1 management core
  - NoC Tx and Rx interfaces
  - Debug support unit (DSU)
- 16-banked shared memory
  - 2MB extensible to 4MB
  - No bus interferences between cores
  - RR arbitration between bus masters
  - Interleaved or blocked address map

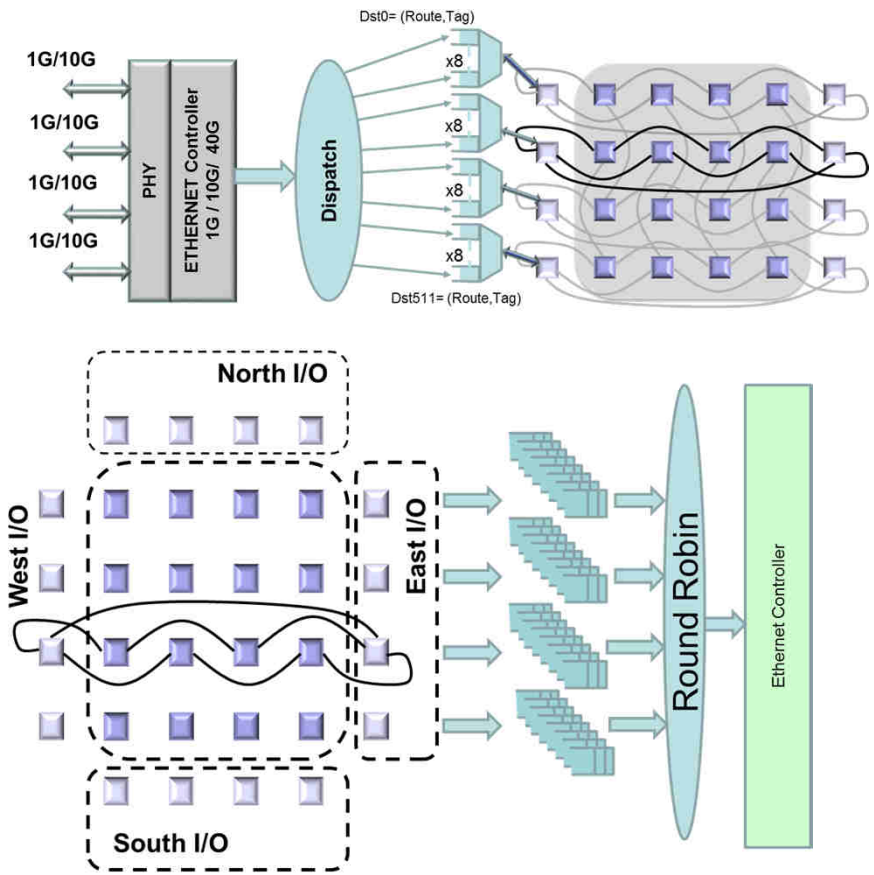


MPPA®-256 Bostan Network-on-Chip (NoC)



- Dual 2D-torus NoC
  - D-NoC: High bandwidth RDMA
  - C-NoC: Low latency mailboxes
  - 4B/cycle per link direction per NoC
  - Nx10Gb/s NoC extensions for connection to FPGA or other MPPA®
- Predictability
  - Data NoC is configured by selecting routes and injection parameters
  - Injection parameters are the (σ,ρ) or (burst, rate) of Cruz network calculus
  - Guaranteed services rely on same methods as in AFDX Ethernet

MPPA®-256 Bostan Ethernet Support



- Ethernet as the main high-performance / low-latency IO
  - Integration of Ethernet Rx and Tx to the D-NoC architecture
- Per Ethernet Rx port
  - 8 classification tables for hardware dispatch
  - Round-robin or classified cluster & core allocation
- Per Ethernet Tx port
  - 64 independent Tx FIFOs
  - Weighted round-robin between Tx FIFOs
  - Flow control between clusters and Tx FIFOs

MPPA® ACCESSCORE

Software Development Kit

Standard GNU C/C++ Programming Environment

Simulators & Profilers, Debuggers & System Trace

Operating Systems & Device Drivers

Low Level Programming
DSP Style

POSIX-Level Programming
CPU Style

OpenCL Programming
GPU Style

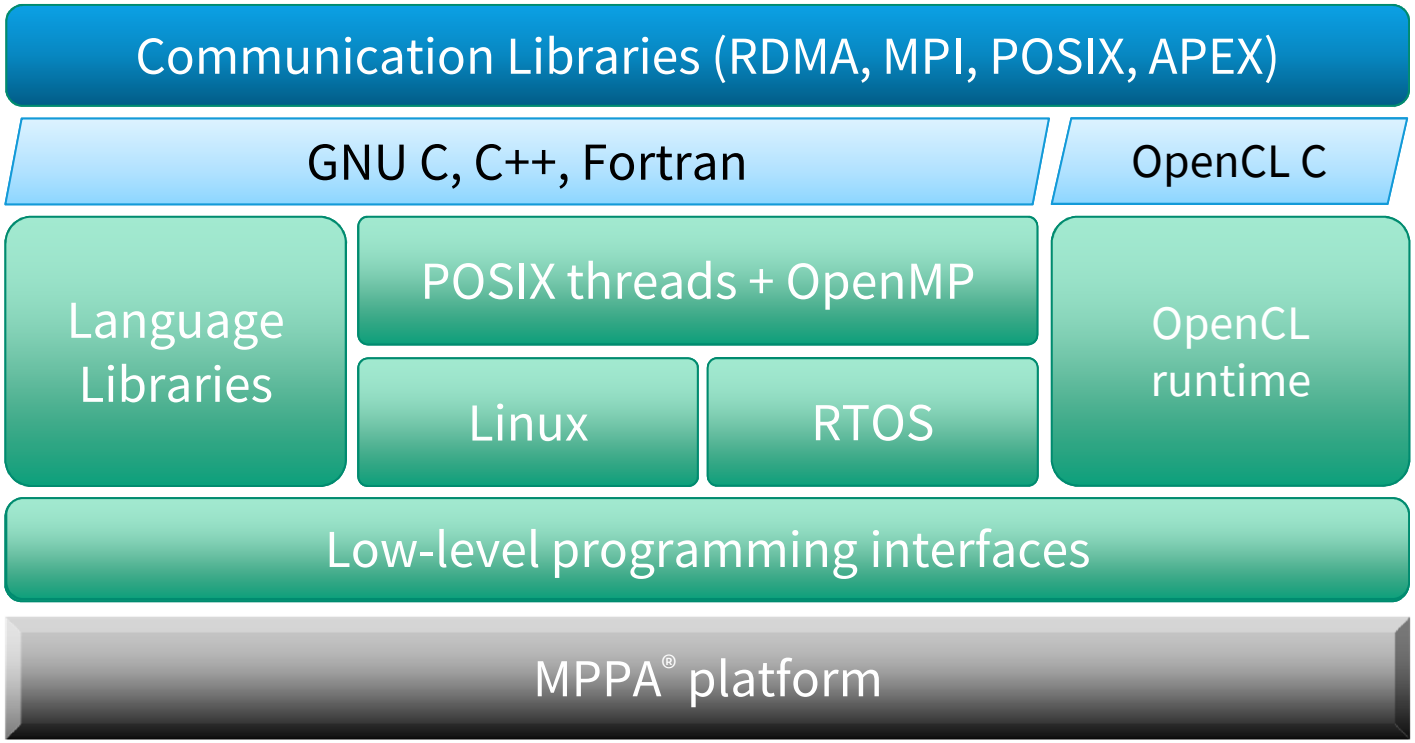
Linux

3rd party Real Time OS

eSOL
embedded intelligence

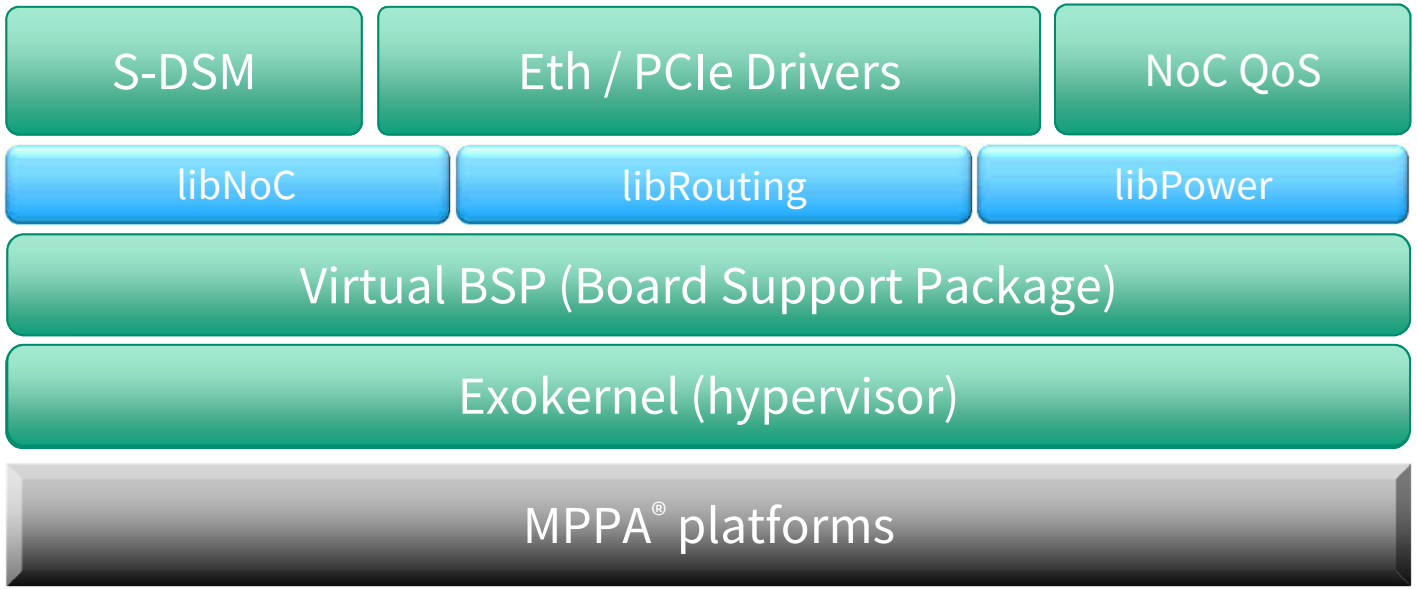
ERIKA ENTERPRISE

MPPA® Software Stack Overview



Low-Level Programming Environment

- Performance programming and 3rd party RTOS interface
  - Exokernel + Virtual-BSP + LibNoC, LibRouting, LibPower
  - Support of software DSM (Distributed Shared Memory)
  - Implement the Low-Level programming interfaces



Operating Systems & Device Drivers

- NodeOS for compute clusters
  - Provides POSIX threads, timers and run-time support for GCC OpenMP
  - The RM manages the NoC interfaces and supports security functions
  - The PEs execute application code on top of exokernel, one thread per core
- SMP Linux on I/O clusters
  - Running on a Kalray quad-core, the other quad-core manages the clusters
  - Device drivers for flash, I2C & SPI (sensors, small peripherals), GPIO
  - OpenMP, uClibc C library, gcc –fdpic
  - PREMPT-RT patch
- 3rd-party RTOS and middleware
  - eSOL eMCOS is the world's first commercially available manycore RTOS for use in embedded systems first ported to Tilera, now supported on the MPPA® processors
  - ERIKA Enterprise, the first open-source OSEK/VDX certified RTOS (automotive)

Applications of MPPA® MANYCORE Processors

- Cloud and Data Center acceleration
  - Offloading of real-time or compute intensive functions from x86 applications
  - Domains of application: video, networking, storage, OHPC, data analytics, cybersecurity
  - MPPA® Compute Clusters seen as OpenCL Compute Units or pools of DSP processors

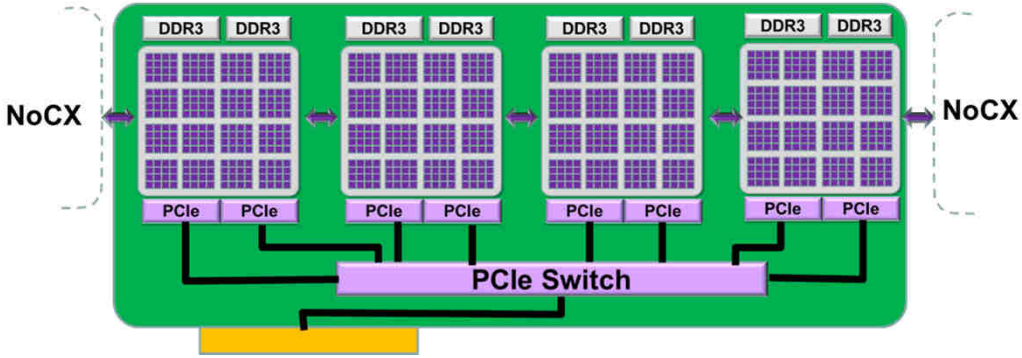
- High Performance Embedded Computing
  - Stand-alone computing enables increased integration of functions including those constrained by real-time
  - Domains of application: aerospace, automotive, transport, energy
  - MPPA® Compute Clusters seen as precision-timed multicore CPUs

Kalray TURBOCARD Family

Kalray TURBOCARD2 (2014)

- 4 Andey MPPA®-256 processors
- 0.83 TFLOPS SP / 0.28 TFLOPS DP
- 8x DDR3L @1266 MT/s => 81 GB/s

- First engineering samples: Q4-14
- Volume Production: Q1-15



Kalray TURBOCARD3 (2015)

- 4 Bostan MPPA®-256 processors
- 2.5 TFLOPS SP / 1.25 TFLOPS DP
- 8x DDR3 @ 2133 MT/s => 136 GB/s

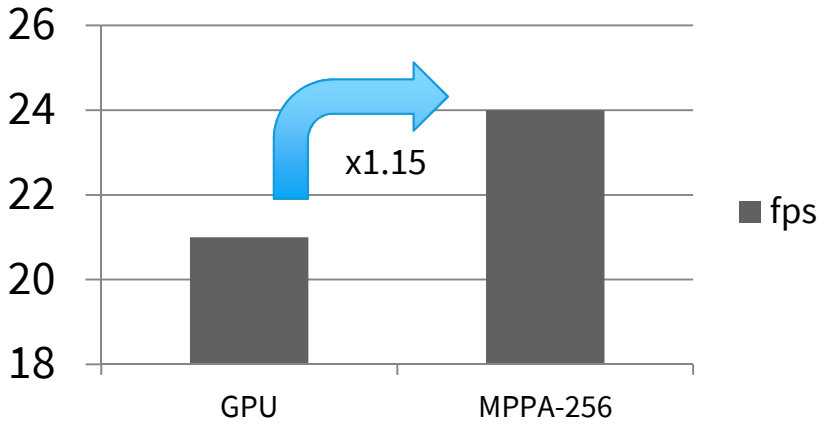
- First engineering samples: Q3-15
- Volume Production: Q4-15

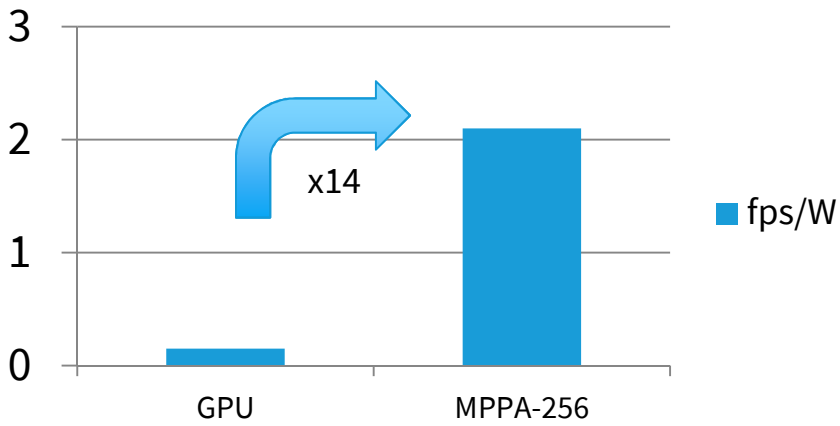
Computer Vision

Object Recognition and Tracking on VGA images

- 5 image processing blocks + 2 categorizers running concurrently
  - Initial customer implementation running on GPU (140 W)
- Application ported on MPPA using OpenCL
  - Better performance than GPU and **10x lower power consumption**

- GPU = 256 SIMT Cores / 142W
- MPPA-256 = 256 MIMD cores / 12W

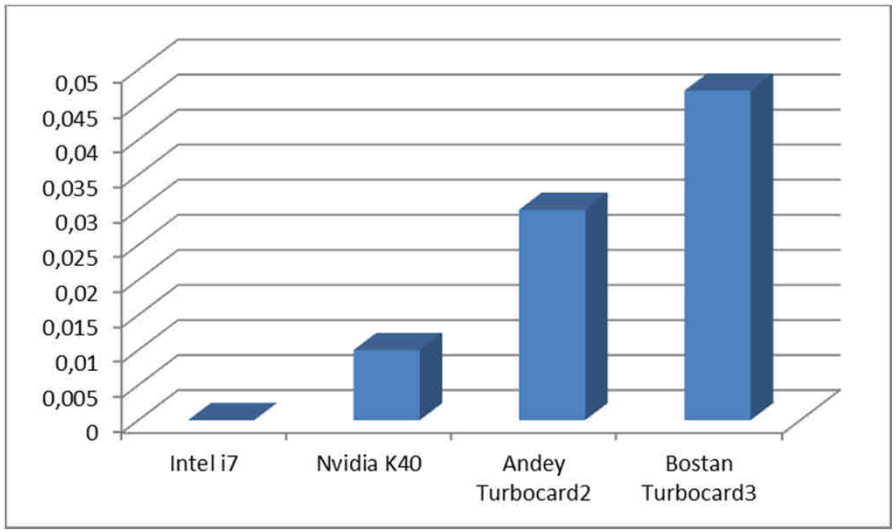


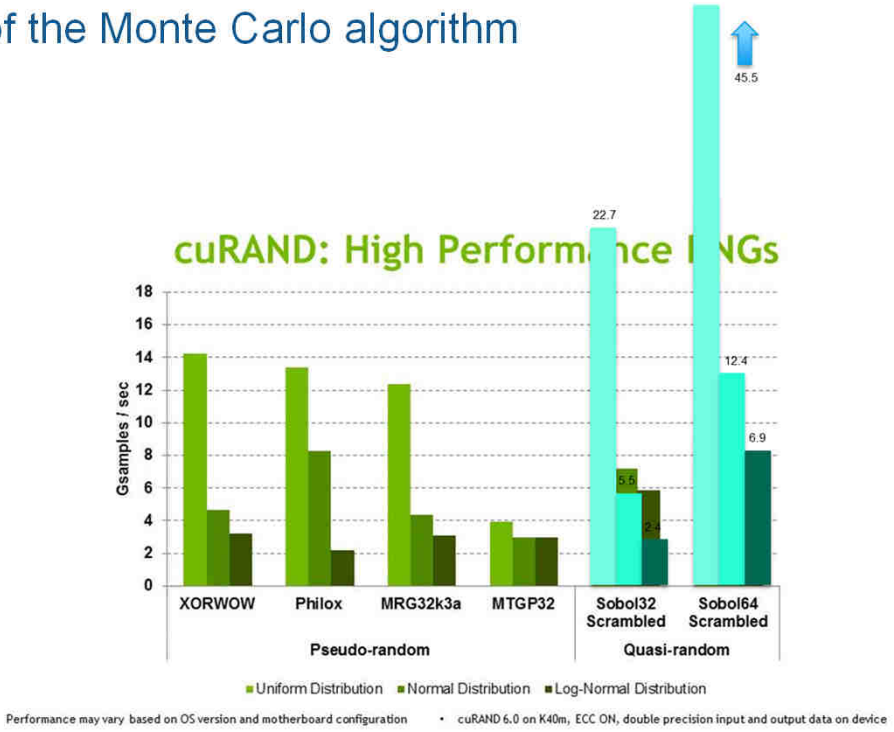


Finance: Monte Carlo Asian Options Pricing benchmark

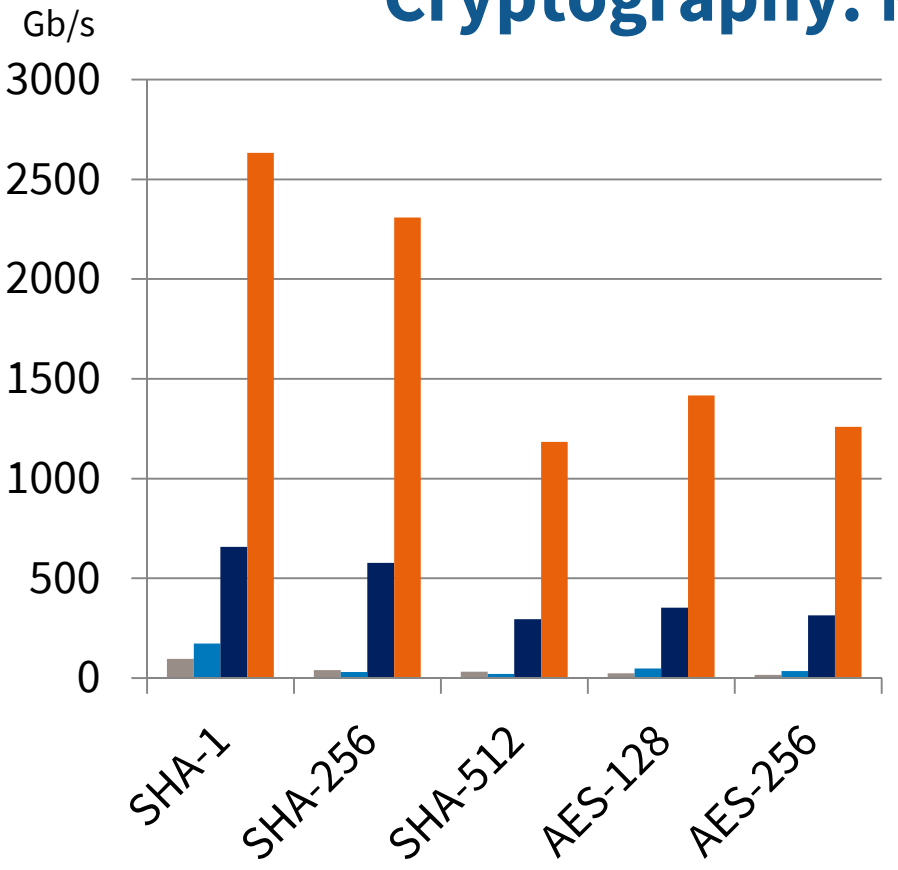
- Asian Option performance (GSamples/s):
  - Mersenne Twister RNG+ Ziggurat Log Normal Distribution+ Option calculation
  - RNG calculation in DP is a big part of the Monte Carlo algorithm

Gsamples/s/W(DP)





Cryptography: MPPA Performances

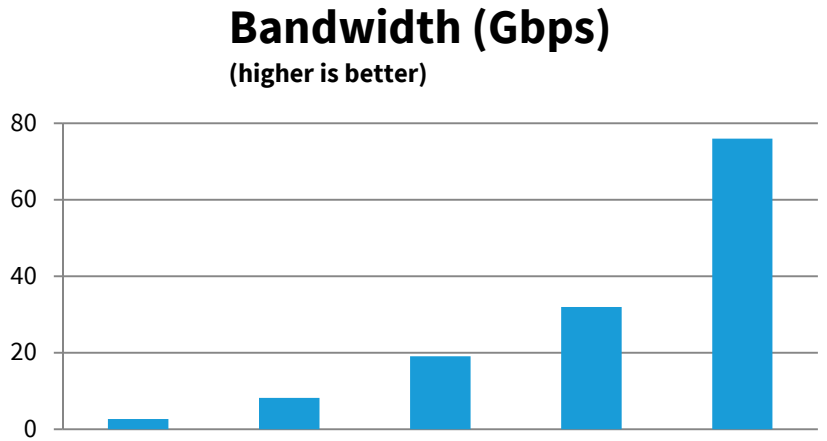


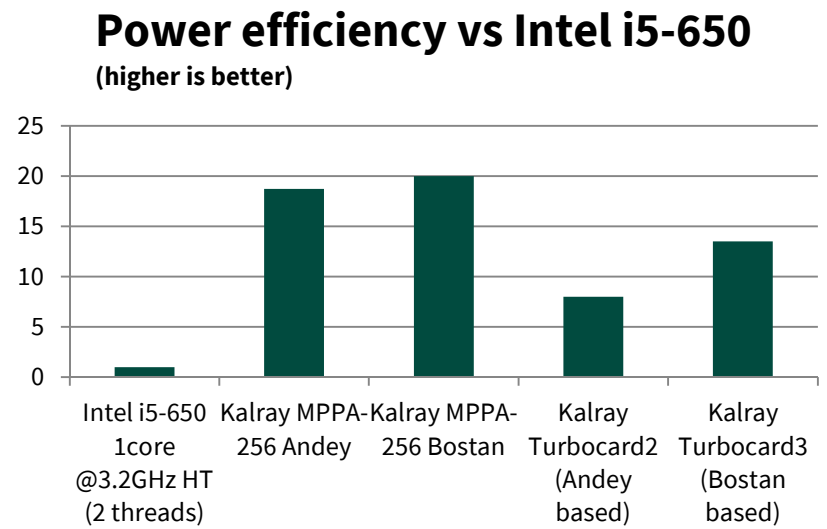
- Intel Xeon Haswell, i7 4770, http://intel.ly/1LHWCCo
- Kalray Andey, 256 cores, 400Mhz
- Kalray Bostan, 256 cores +128 CryptoCores, 800Mhz
- TurboCard 3, 1024 cores +512 CryptoCores, 800Mhz

- **From Andey to Bostan**
  - Added 128 CryptoCores
  - Up to x 19 factor in SHA
  - Up to x 6+ in AES
- Kalray Cryptography solution can be implemented on different boards:
  - K-ONIC 80 at Networking termination level
  - TurboCard3 as server acceleration board

Compression: GZIP DEFLATE

- Web servers and storage applications make heavy use of lossless data compression
- DEFLATE is the most widely used algorithm (gzip and zlib)
- Standard zlib is available on MPPA with high performance
- Compression can be mixed with other compute-intensive tasks
  - Erasure coding
  - Cryptography

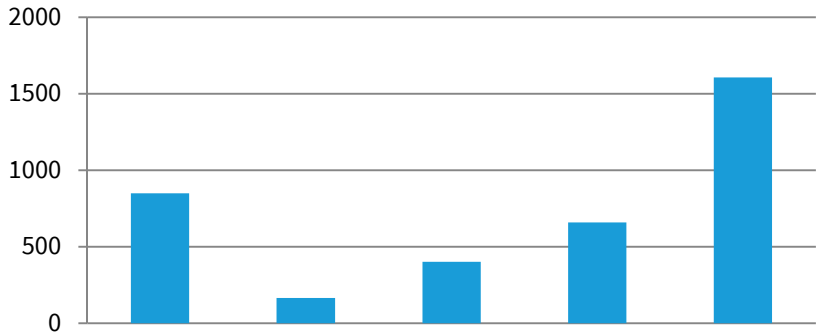




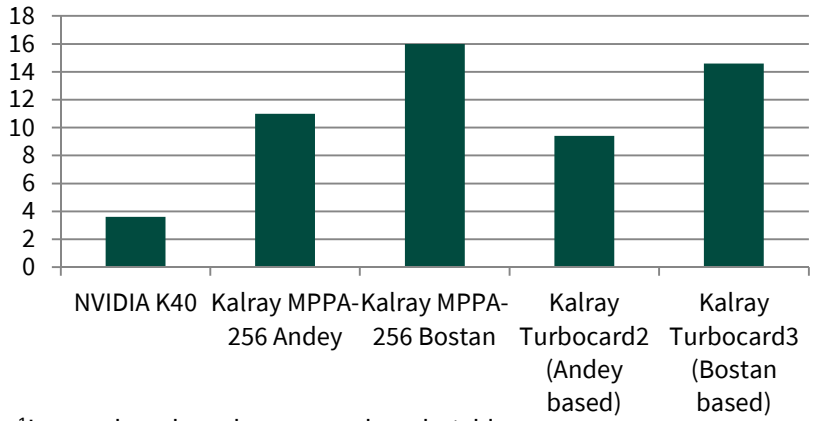
Machine Learning: Convolutional Neural Networks (CNN)

- Kalray's unique MPPA® architecture (multi-banked internal memory) is well suited for Deep Learning.
- MPPA® processor delivers a higher performance per watt ratio than industry deployed solutions
- MPPA® processor accelerates both the training phase and run the network once trained

convnet-benchmark[1] (GFLOPS)
(higher is better)



Power efficiency (GFLOPS/W)
(higher is better)



[1]Layers benchmark, averaged, no batching
https://github.com/soumith/convnet-benchmarks

MPPA® Supercomputing on a Chip

Energy Efficiency

- 5-issue 32-bit / 64-bit VLIW core
- Optimum instruction pipelining
- Shallow memory hierarchy
- Software cache coherence

Performance Scalability

- Linear scaling with number of cores
- Network on chip extension (NoCX)
- 2x DDR controllers per processor
- 24x 10 Gb/s lanes per processor

Execution Predictability

- Fully timing compositional cores
- Multi-banked parallel local memory
- Core-private busses to local memory
- Network on chip guaranteed services
- Configurable DDR address mapping

Ease of Use

- Standard GCC C/C++/Fortran
- Command-line and Eclipse IDE
- Full featured debug environment
- System trace based on LTTNG
- Linux + RTOS operating systems

MPPA®-256 Bostan Improvements

Indicator	MPPA®-256 Andey	MPPA®-256 Bostan	Improvement
# of cores	256	256	=
Max frequency (overdrive)	400MHz (N/A)	600MHz 800MHz	x1.5 x2
SP GLFOPS @400MHz @600MHz @800MHz (overdrive)	 211 (N/A) (N/A)	 423 634 845	 x2 x3 x4
DP GFLOPS @400MHz @600MHz @800MHz (overdrive)	 70 (N/A) (N/A)	 315 420	 x4,5 x6
TOPS (overdrive)	0.7 (N/A)	1.05 1.4	x1.5 x2
DDR3 MT/s	1266	2133	x 1.7
Power @400MHz @600MHz @ 800MHz (overdrive)	 12W (N/A) (N/A)	 11W 16W 24W	 X0.9

Company Overview

- Develops manycore processors since 2008
- ‘Supercomputing on a chip’™ architecture for high computing performance, low energy consumption and precision-timed execution
- Offer includes MPPA® (Massively Parallel Processing Array) single-chip processors, acceleration boards and software development tools
- Industry-recognized R&D teams of 55 engineers (HW & SW)
- 2014 EETimes Silicon 60: hot start-ups to watch
- Offices in France, US and Japan

Thank you

KALRAY S.A.
Paris - France

86 rue de Paris,
91 400 Orsay
France

Tel: +33 (0) 184 00 00 45
email: info@kalray.eu

KALRAY S.A.
Grenoble - France

445 rue Lavoisier,
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