KALRAY

Using Manycore Processors in Complex Embedded Systems

with Kalray MPPA® customer cases

Teratec Forum - June 25-26, 2013

Visit us on booth #10

AGILE PERFORMANCE

Presentation Outline

- Kalray : the company, products and markets
- The rising of Manycore architectures
  - Feartures and benefits
- Using Manycore processors
  - Concrete customer cases
- Q&A

Kalray at A Glance

- Founded in 2008 – located in Paris, Grenoble (France) & Tokyo (Japan)
- 55+ people
- Independent and unique technology : MPPA® MANYCORE processor (Multi-Purpose Processing Array) and software programming environment
- Targeting the industrial, embedded and computing intensive markets
- Large patent portfolio
- Several awards over the past years
  - Kalray ranked in “EETimes’ silicon 60 : Hot start up to watch” in 2012
  - “Best technology award” from “Les Trophées de l’Embarqué” in 2012
  - “Startup of the year” by ElectroniqueS magazine in 2013

First MPPA®-256 Chips with CMOS 28nm TSMC

Released November 2012

- High processing performance 700 GOPS – 230 GFLOPS
- Low power consumption - 5W
- High execution predictability
- Software programmable

KALRAY, a global solution

MPPA MANYCORE

Powerful, Low Power and Programmable Processors

MPPA ACCESSCORE

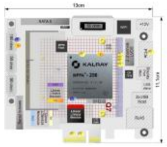
C/C++ based Software Development Kit (SDK) for massively parallel programing

MPPA DEVELOPER

Development platform
Reference Design Board

MPPA BOARDS

Reference Design board
Application specific boards
Multi-MPPA or Single-MPPA boards



Target Application Areas

INTENSIVE COMPUTING

- Finance
- Numerical Simulation
- Geophysics
- Life sciences

IMAGE & VIDEO

- Broadcast
- Medical Imaging
- Digital Cinema
- Augmented reality
- Vision

EMBEDDED SYSTEMS

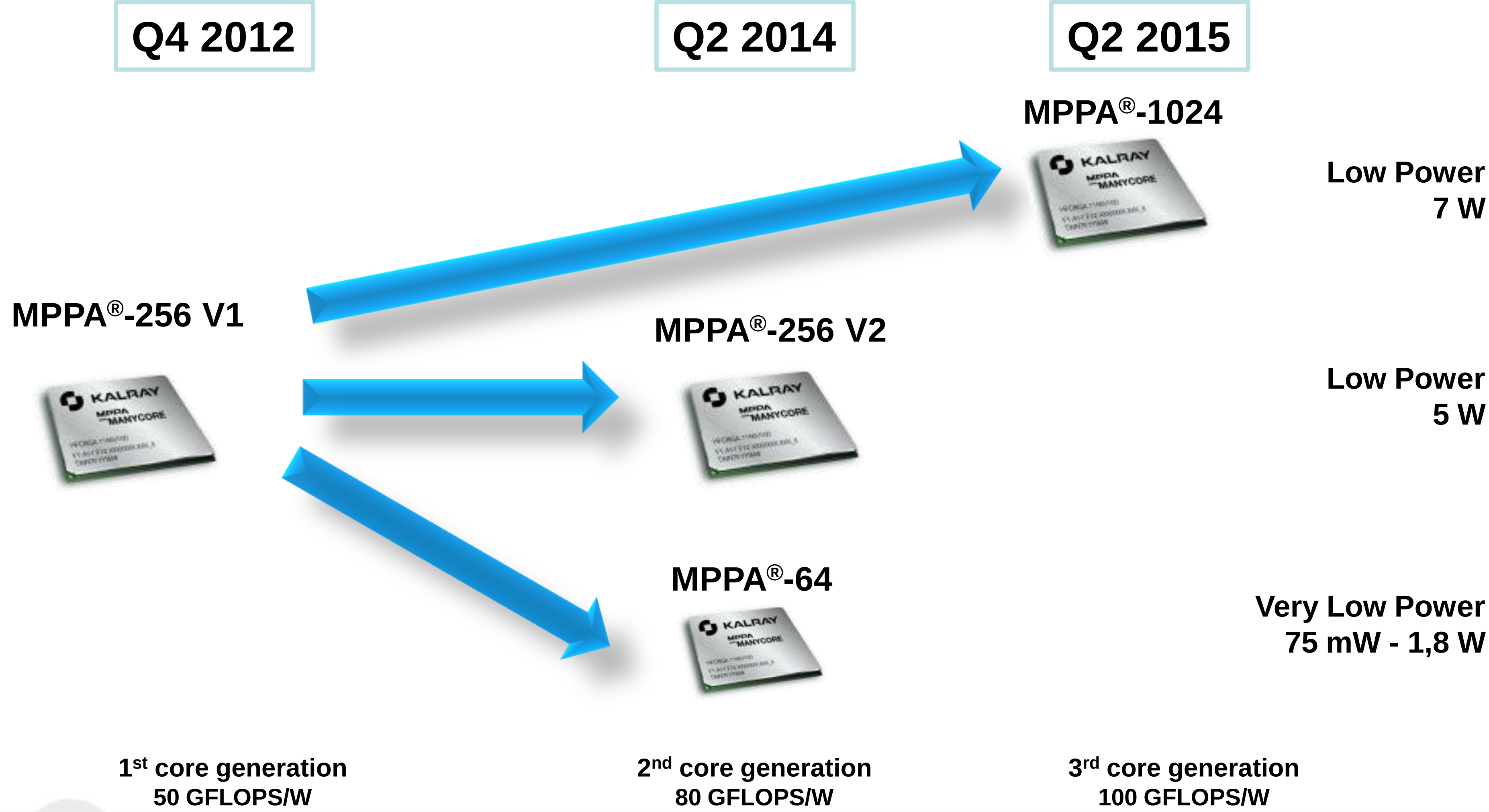
- Signal Processing
- Aerospace/Defence
- Transport
- Industrial Automation
- Video Protection

TELECOM / NETWORKING

- Packet Switching
- Network Optimisation
- Security Services
- Software Defined Radio
- Software Defined Network

MPPA MANYCORE Roadmap

Architecture scalability for high performances and low power



Q4 2012

Q2 2014

Q2 2015

MPPA®-1024

Low Power
7 W

MPPA®-256 V1

MPPA®-256 V2

Low Power
5 W

MPPA®-64

Very Low Power
75 mW - 1,8 W

1st core generation
50 GFLOPS/W

2nd core generation
80 GFLOPS/W

3rd core generation
100 GFLOPS/W

Kalray Software Development Kit

MPPA ACCESSCORE – MPPA ACCESSLIB

Linux Today

Microsoft Windows Q4 2013

Standard C/C++ Programming Environment

Simulators & Profilers, Debuggers & System Trace

Operating Systems & Device Drivers

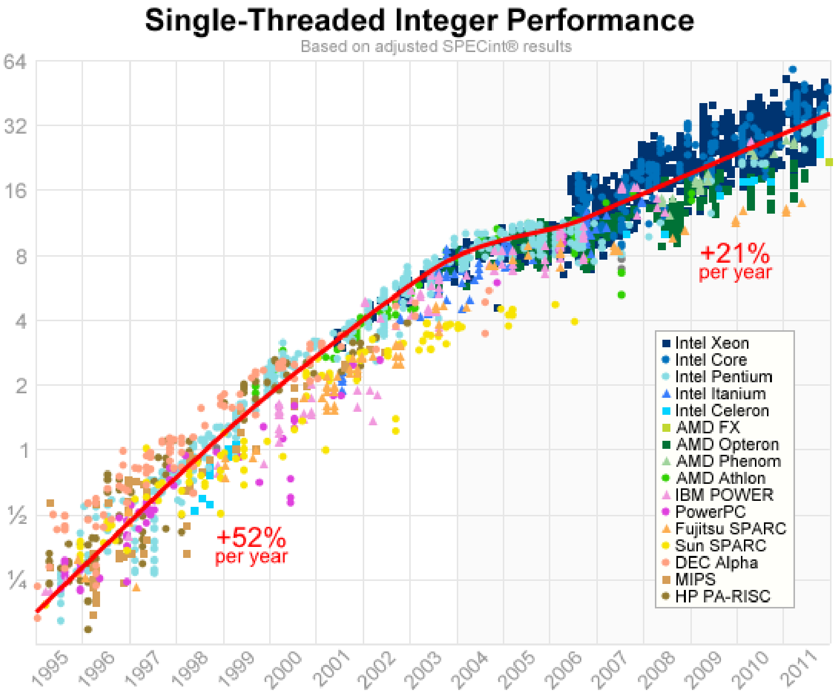
Dataflow Programming
FPGA Style

POSIX-Level Programming
DSP Style

Streaming Programming
GPU Style

The rising of Manycore architectures

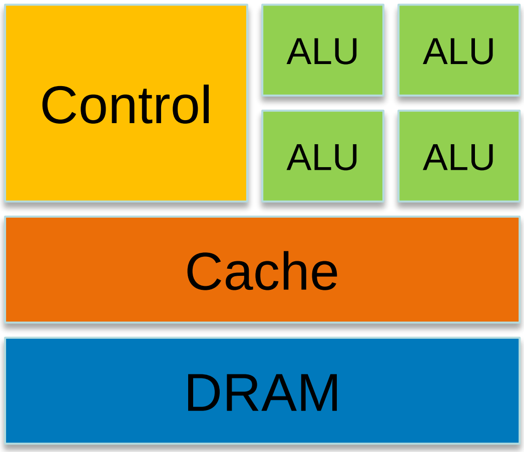
Single-Threaded Integer Performance



- +52% increase yoy from 1996 to 2004
- +21% from 2007 to 2011
- Current figures suggest a +10% increase yoy since 2011 for most CPUs

No more progress in time scale, so enlarge in space with more cores

Multicore CPUs vs GPUs vs Manycore

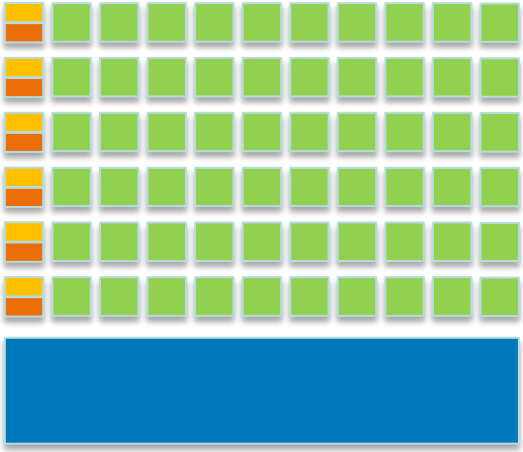


CPU

- CPU are optimized for sequential code performance
- Sophisiticated Control Logic to execute several instruction at the same time
- Very large cache to reduce access time to instruction and data of complex applications
- Clock frequency limits reached long ago
- 50 to 150 Watt

CONSTRAINED BY

- Power consumption , Complexity, Scalability

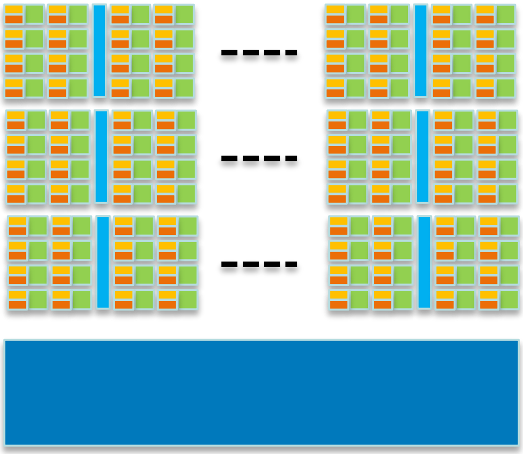


GPU

- Originates from the video game industry
- Very simple control units and huge number of floating point units working in parallel
- SIMD processing model
- Smaller cache than CPUs
- 20 to 300 Watt

CONSTRAINED BY

- Power consumption , Programming model, Communication overhead



Manycore

- Rich CPU-like Control units + FPU
- Cluster of processors share a local memory
- Clusters communicate through a high speed low latency network on chip
- MIMD paradigm (Multiple instructions Multiple Data) rather than MIMD
- 5 to 10 Watt

CONSTRAINED BY

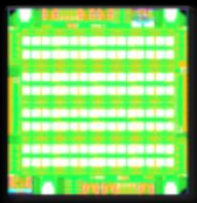
- Disruptive technology adoption curve

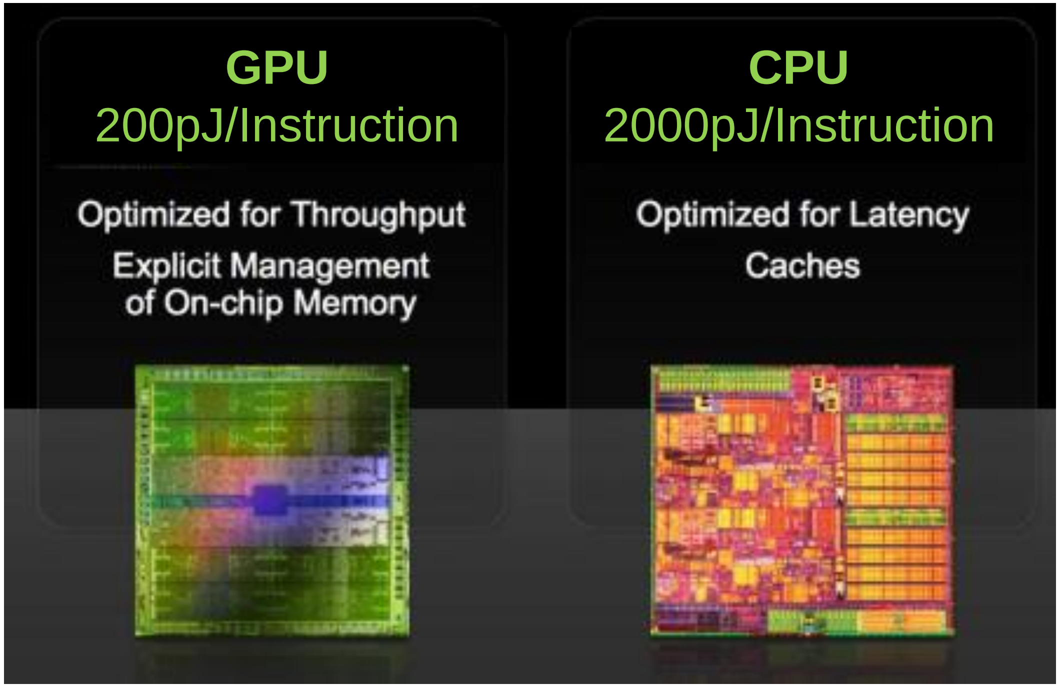
MPPA® Technology Compared to GPU & CPU

MPPA-256

20pJ/Instruction

Optimized for GFlops/Watt and $
Distributed Memory model
Low Latency





GPU

200pJ/Instruction

Optimized for Throughput

Explicit Management
of On-chip Memory

CPU

2000pJ/Instruction

Optimized for Latency

Caches

Source: Bill Dally, “To ExaScale and Beyond” - NVidia

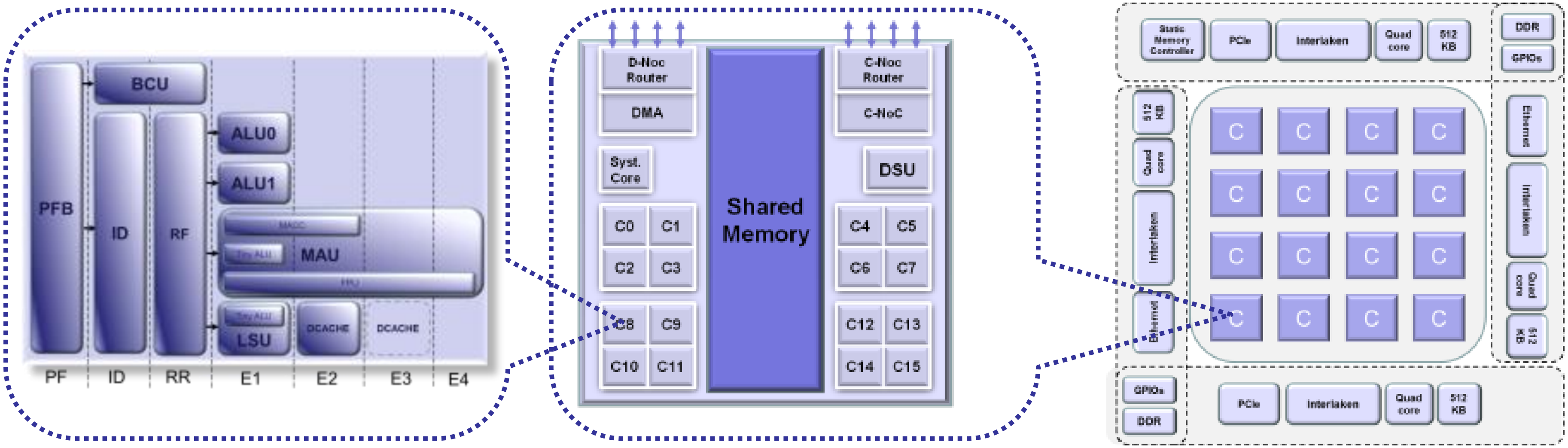
230 GFlops 5 W	3000 GFlops 300 W	700 GFlops 130 W
~50	~10	~5

MPPA®-256 Processor Hierarchical Architecture

VLIW Core

Compute Cluster

Manycore Processor



Instruction Level Parallelism

Thread Level Parallelism

Process Level Parallelism

KALRAY

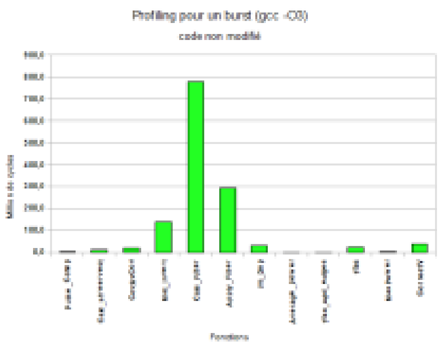
Kalray's MPPA customer cases

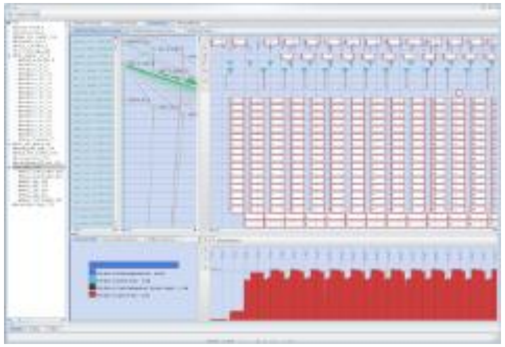
Intensive Computing Driving Factors

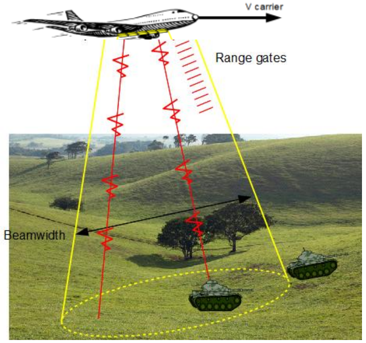
- Computing efficiency (GFlops/Watt)
  - Energy consumption becomes an absolute barrier whether in high-end embedded sytems or data-centers
- Hardware efficiency (GFlops/$)
  - Remove unnecessary hardware overhead
- Bandwidth (MB/s)
  - Bring data in and out fast and avoid bus bottleneck

Signal Processing Example

- Radar applications: STAP, …
- Beam forming : Sonar, Echography
- Software Defined Radio (SDR)
- Dedicated libraries (FFT, FTFR, …)







Well suited for massively parallel architectures

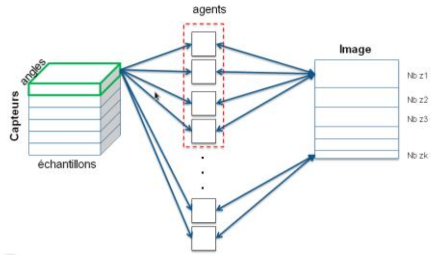
MPPA as an alternative to DSP+FPGA or CPU+GPU platforms

Customer case #1

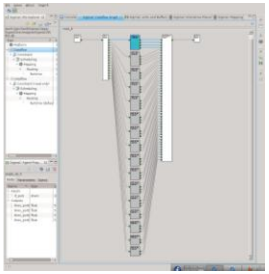
- Healthcare industry sector
- Electronic device performing real time, very intensive signal and image processing tasks
- Current status
  - Uses conventional CPU+GPU
- Problems identified
  - Excessive energy footprint : won’t fit conventional electric setup at customer site, too bulky, too noisy.
  - GPUs hard to program and almost complete lack of tools for debugging/profiling/optimization work.
  - Totally inadequate for a forthcoming portableversion of the device

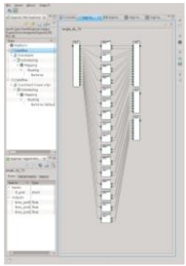
Kalray’s 3-step Approach

1. Analyze
   - Profile legacy code using MPPA ACCESSCORE simulation tools

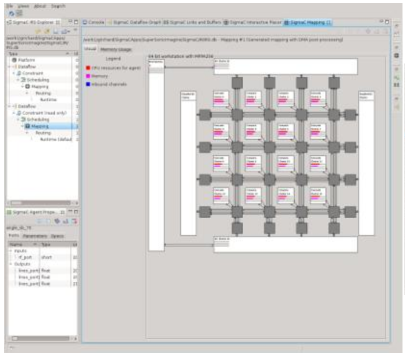


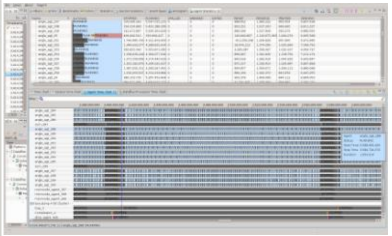
2. Parallelize
   - Use MPPA Dataflow to express parallelism at a high level



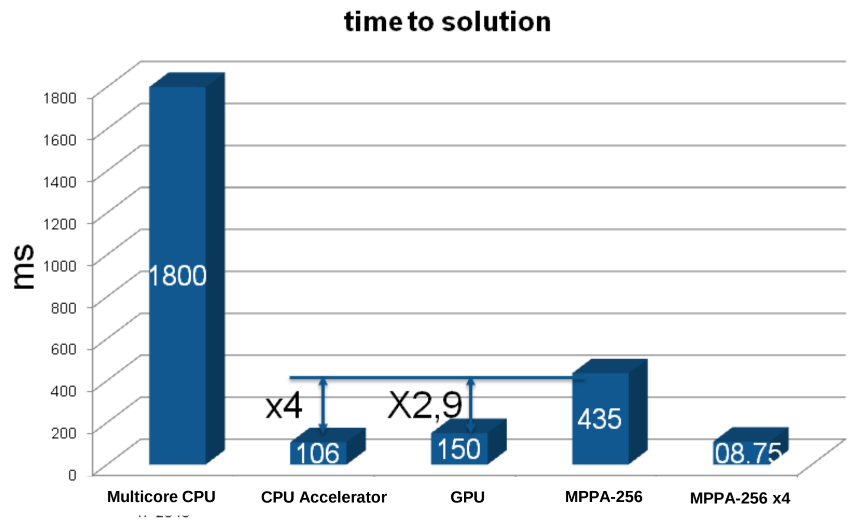


3. Optimize
   - Automatically map application on the 256 cores of MPPA processor
   - Run the code, trace and fine tune

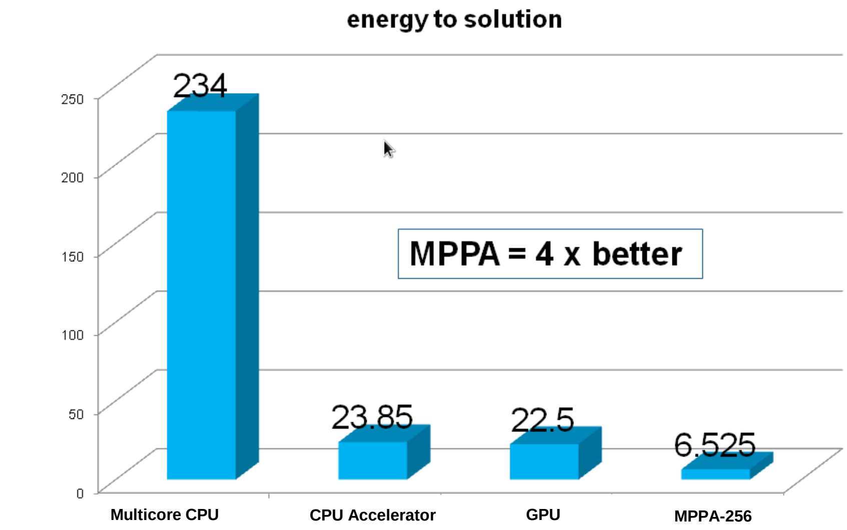




Customer Benefits



PWR	130W	225W	150W	15W	60W



PWR	130W	225W	150W	15W

- Hardware benefits
  - GFlops/W is 4x better
  - No need for host CPU (better GFlops/$)
  - Extensible processing array without PCI bus
  - MPPA v2 triples the GFlops/W

- Software benefits
  - Porting effort < one month of work
  - Powerful and unique optimization tools
  - Unified toolset Host + MPPA
  - No need to change the code when extending the processing array

Q&A

www.kalray.eu - info@kalray.eu

Visit us at Teratec Booth #10