Capability Computing Systems Based on Multi-Core Systems on Chip

Benoît Dupont de Dinechin

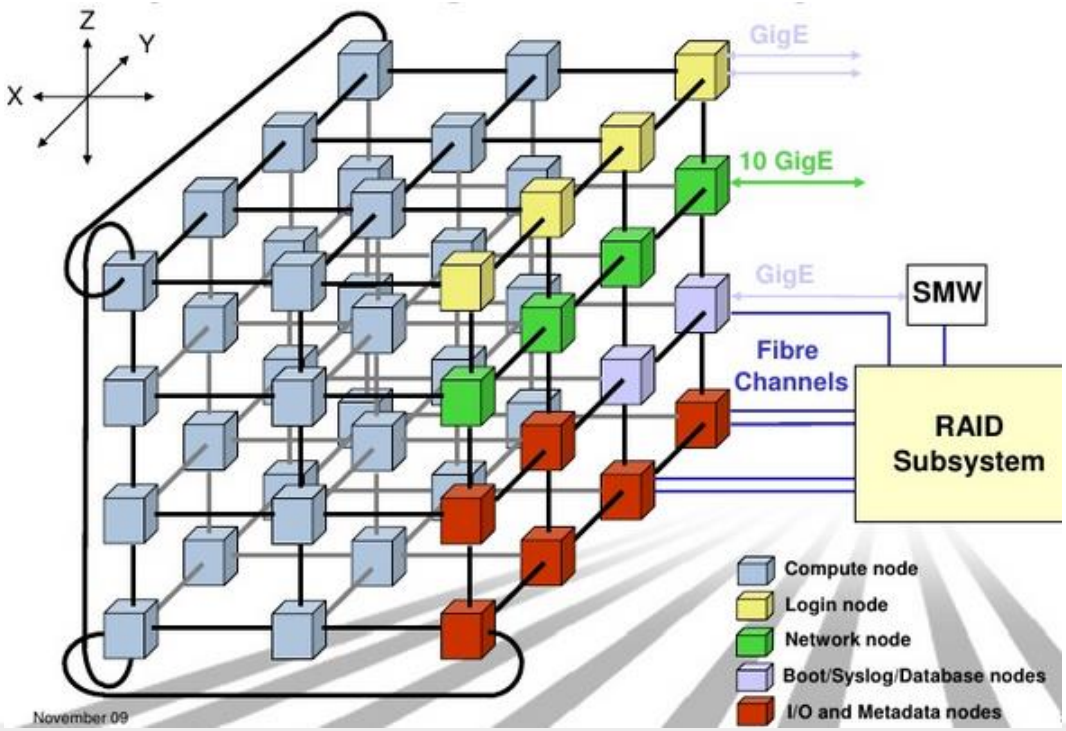
ASPROM 2014

Outline

- **Node Abstract Architectures**
- SoC-Based Projects and Systems
- Kalray MPPA Manycore Processors
- Perspectives and Conclusions

Supercomputer Clustered Architecture

- IBM BlueGene series, Cray XT series
  - Compute nodes with multiple cores and shared memory
  - I/O nodes with high-speed devices and a Linux operating system
  - Dedicated networks between the compute nodes and the I/O nodes

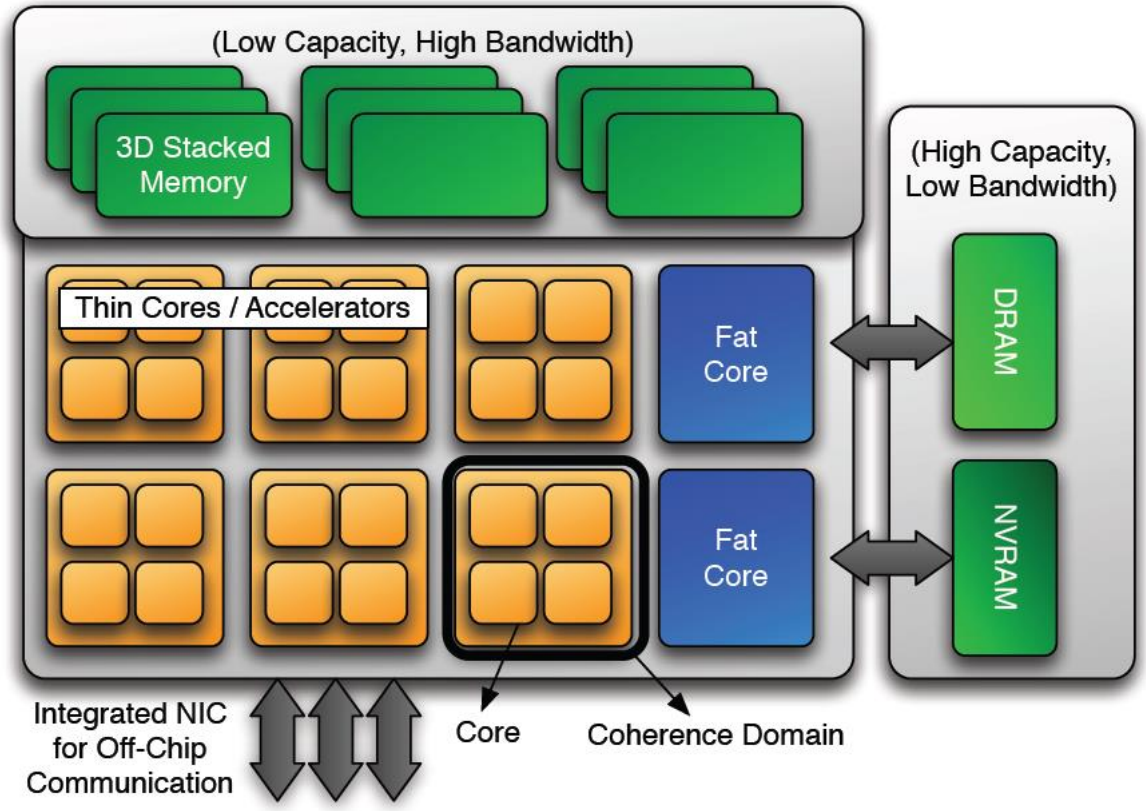


Terminology for Large-Scale Systems

- Capability computing (supercomputing)
  - Using the maximum computing power to solve a single large problem in the shortest amount of time
- Capacity computing (cloud computing)
  - Using efficient cost-effective computing power to solve a small number of somewhat large problems or a large number of small problems
- Scale-up (scale vertically)
  - Add resources to a single node in a system
    - Typically involving the addition of CPUs or memory or accelerators
- Scale-out (scale horizontally)
  - Add more nodes to a multi-node system
    - Enabled by the development of high-performance interconnects
- Proxy architecture (of compute nodes)
  - A parameterized version of an abstract machine model

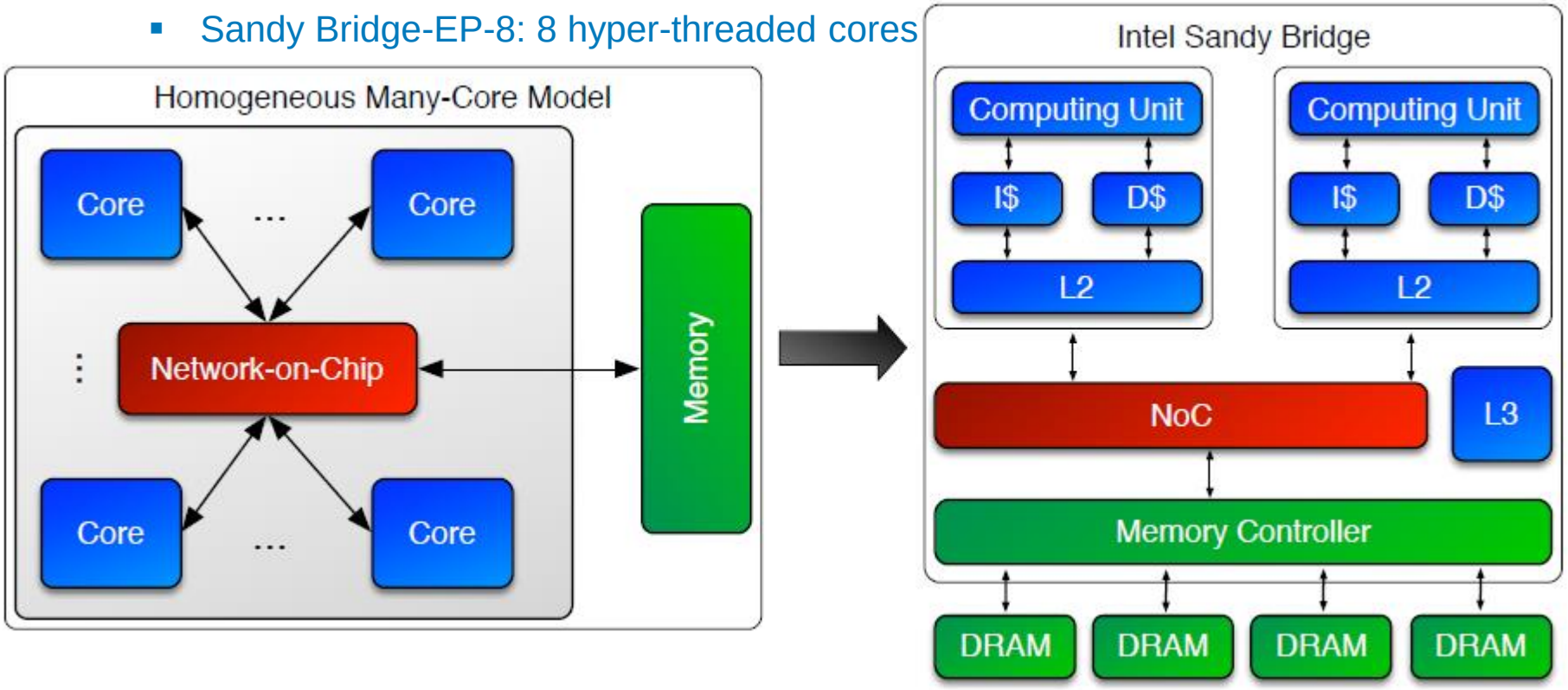
Abstract Machine Model of an Exascale Node

- « Abstract Machine Models and Proxy Architectures for Exascale Computing » 2014 report from Sandia and Lawrence Berkeley National Laboratories



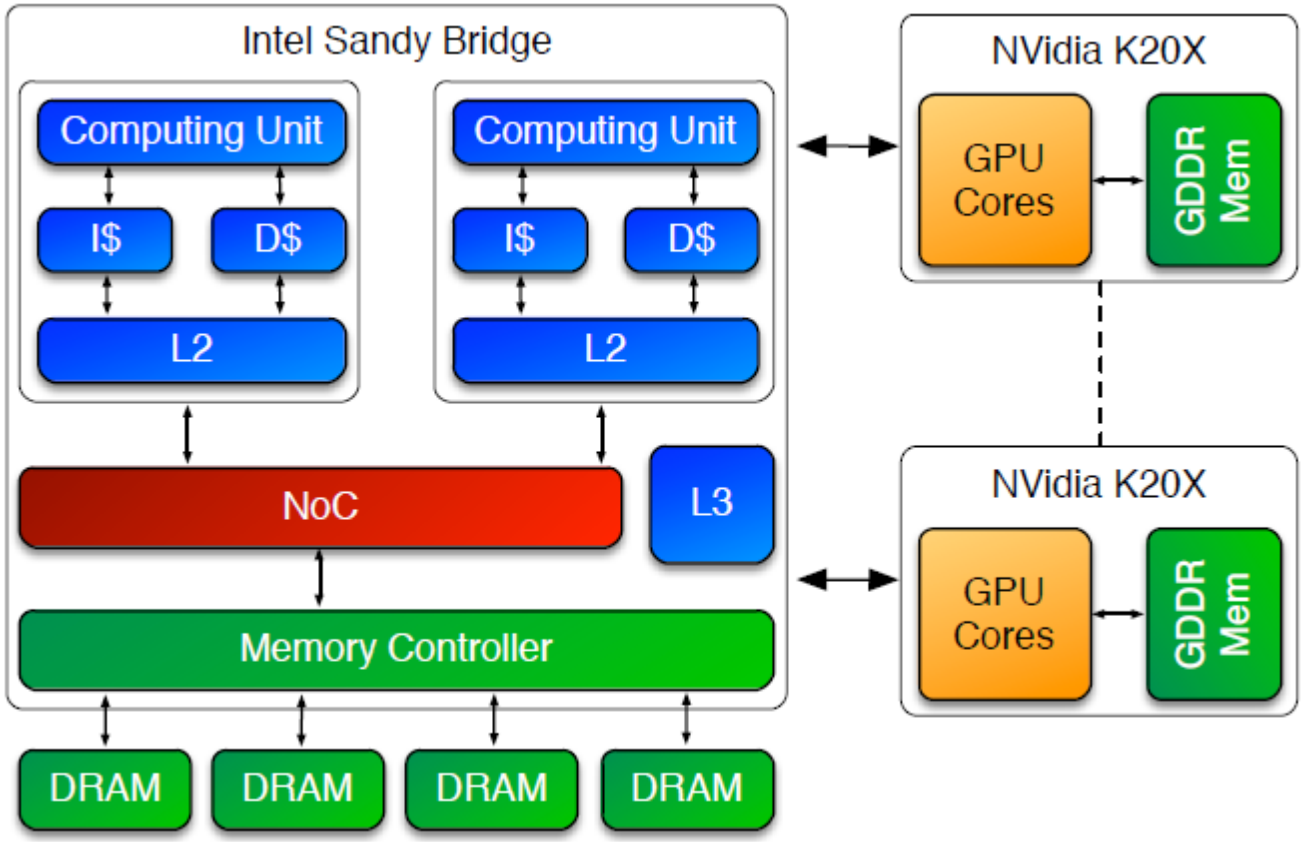
Homogeneous Multicore Model

- Each core is symmetric in performance and has the same ISA
- The cores share a single address space with full cache coherency
  - IBM BG/Q processor: 16 4-threaded cores
  - Sandy Bridge-EP-8: 8 hyper-threaded cores



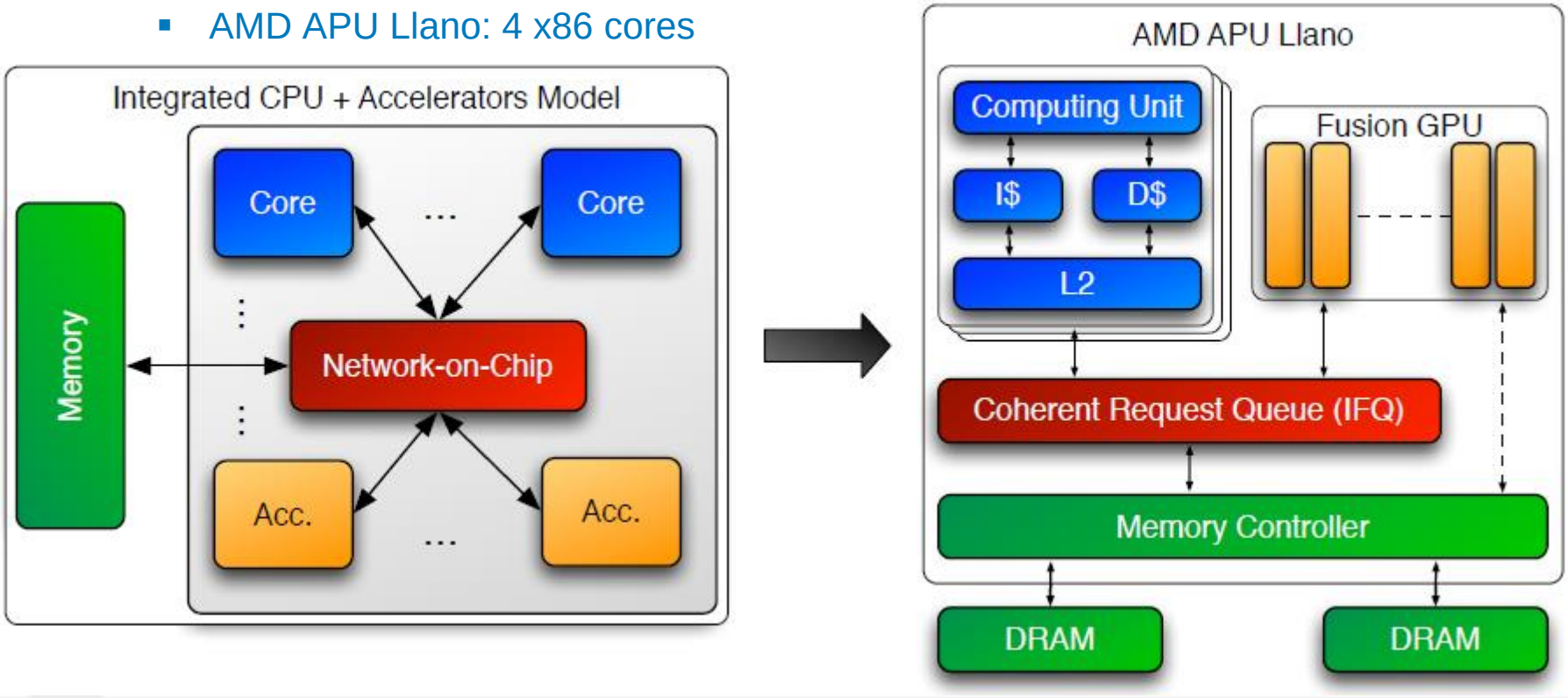
Multicore with Discrete Accelerator Model

- Homogeneous multicore CPU coupled with a series of accelerators
- Each accelerator has a local memory separate from the CPU system memory



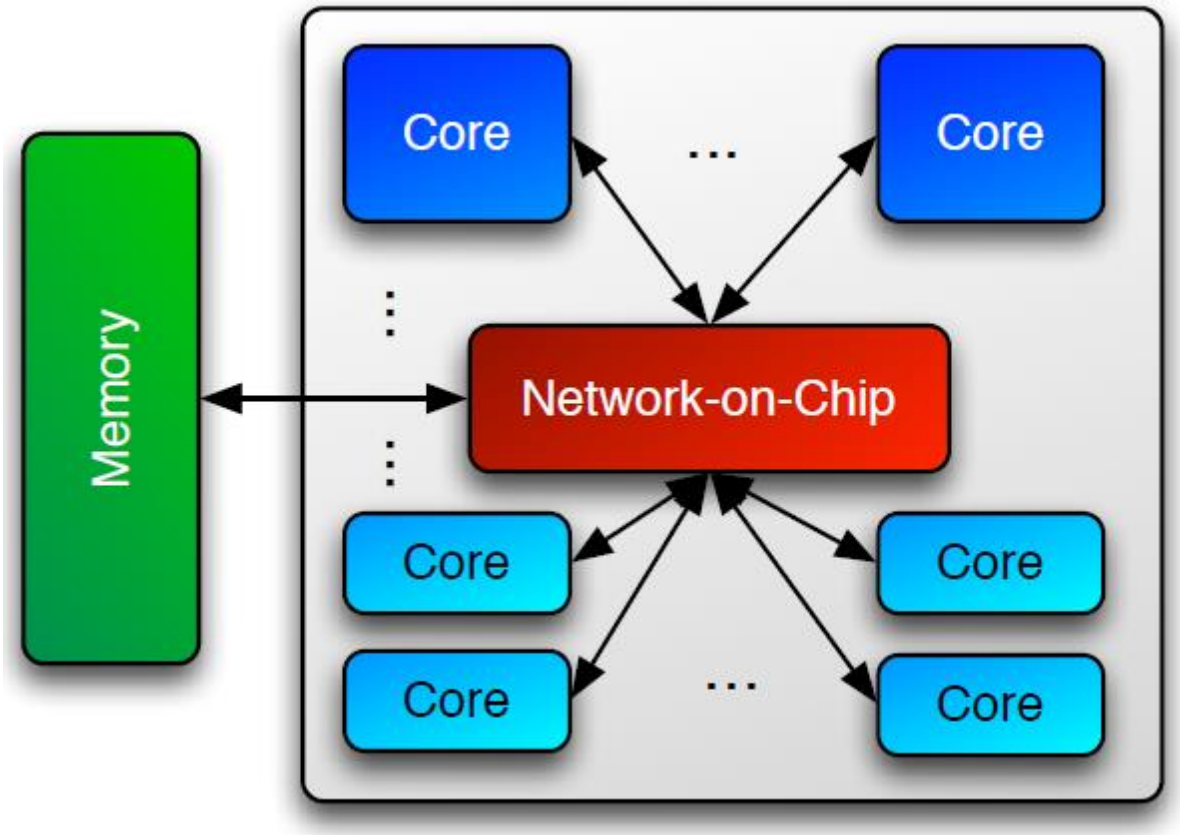
Integrated CPU and Accelerators Model

- Combine latency-optimized CPU cores with many accelerators
- Single shared memory address space, today not fully cache coherent
- Heterogenous System Architecture (HSA) aims at full cache coherence
  - AMD APU Llano: 4 x86 cores



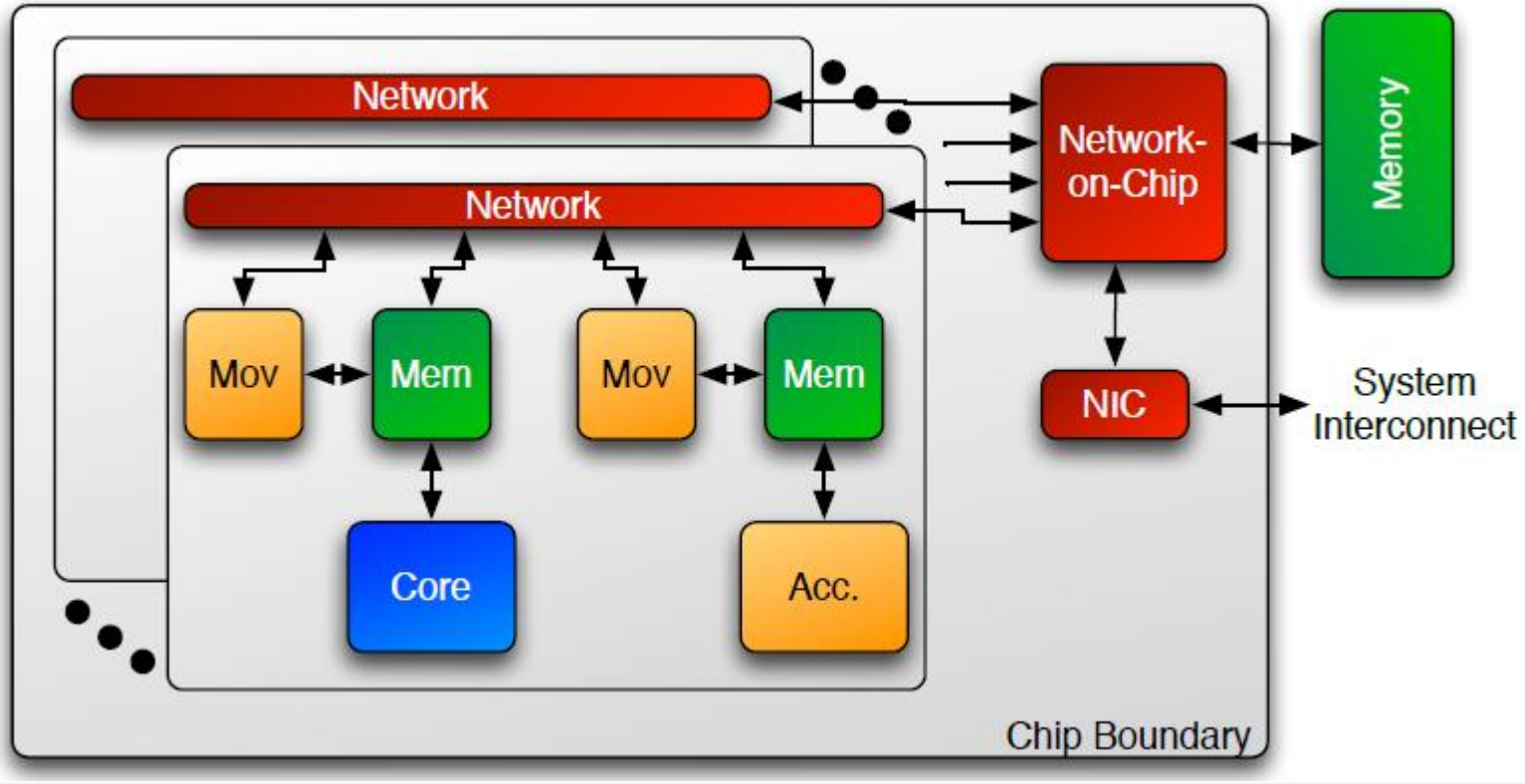
Heterogenous Multicore Model

- Potentially many different classes of processor cores with their own ISA
- Each processing core supports independent branching and threading



Homogeneous Multicore-Accelerator-Memory Model

- Internal memory blocks are integrated into the CPU cores as well as main memory that is directly connected and is available through a system's network
- At least two different kinds of accelerators: vector units and data move units

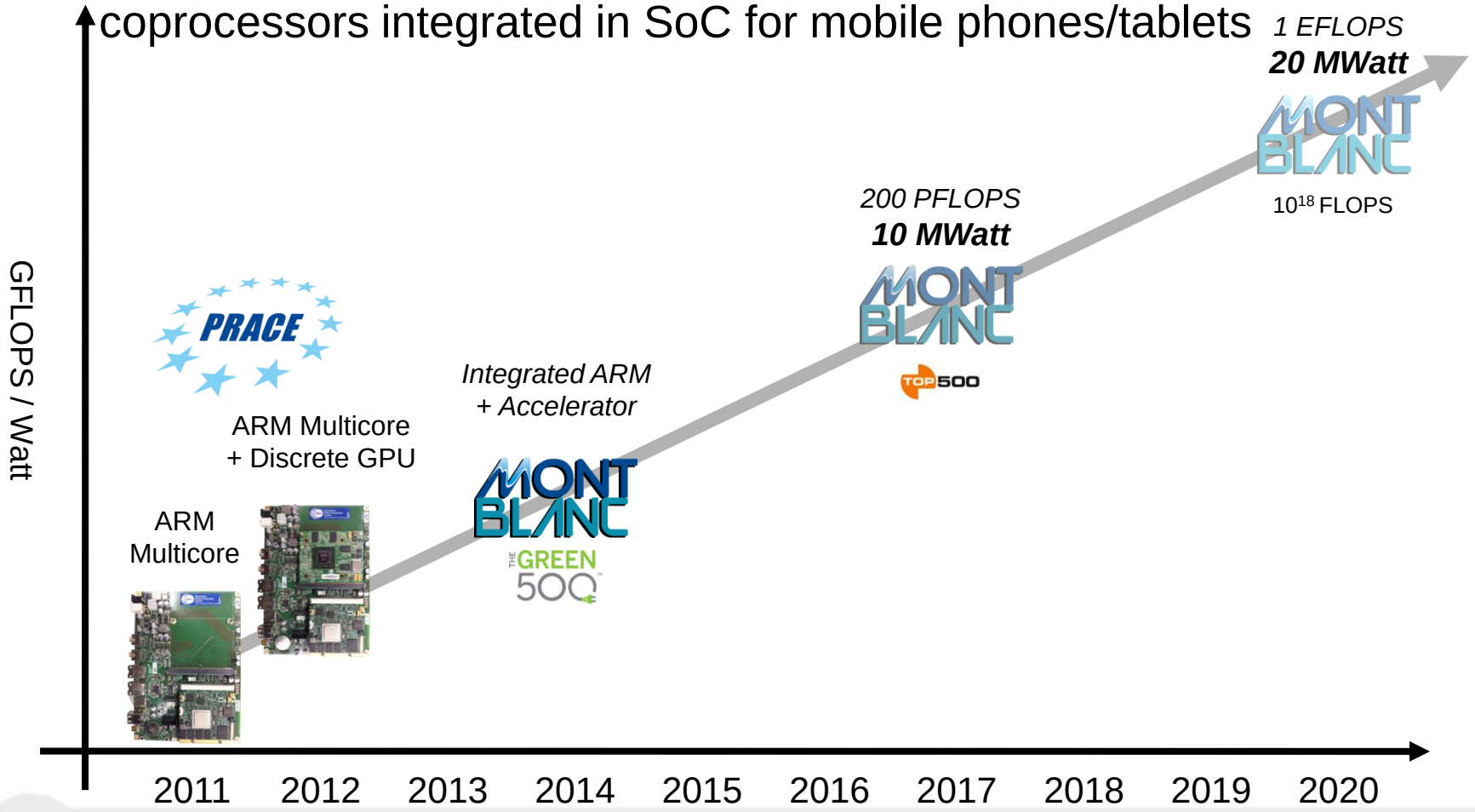


Outline

- Node Abstract Architectures
- **SoC-Based Projects and Systems**
- Kalray MPPA Manycore Processors
- Perspectives and Conclusions

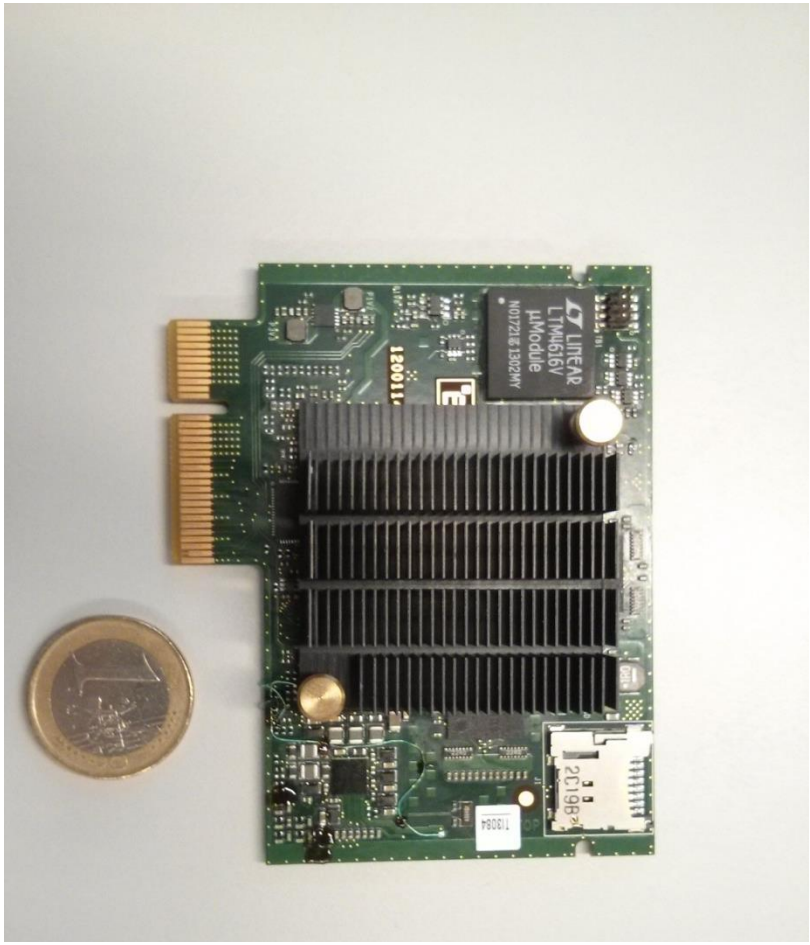
FP7 MontBlanc Project

- Energy-efficient supercomputer based on ARM processors and GPU coprocessors integrated in SoC for mobile phones/tablets



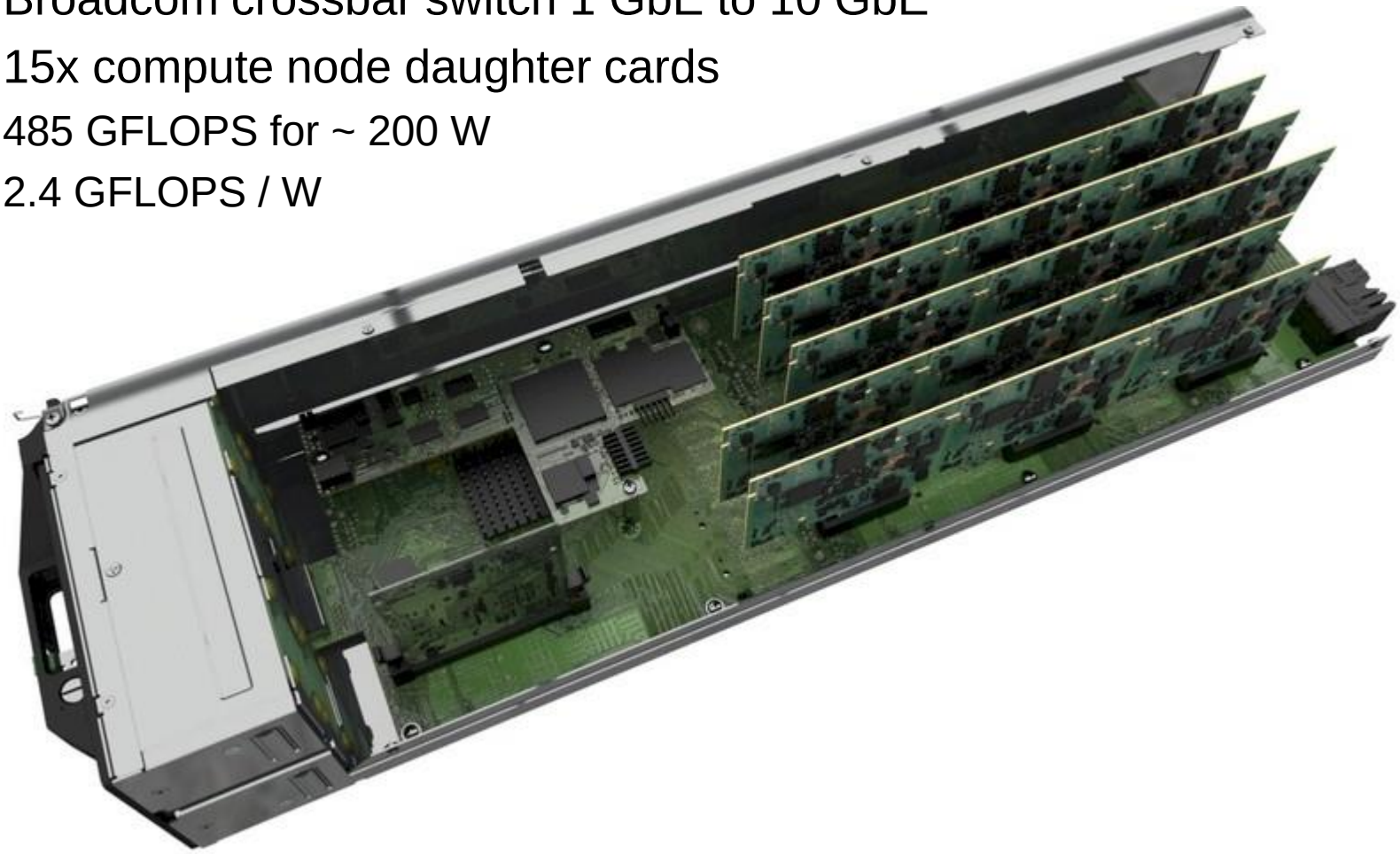
FP7 MontBlanc Project Compute Node

- Mont Blanc node is a daughter card made by Samsung
  - SoC with the CPU and GPU
  - 4 GB of memory (1.6 GHz DDR3),
  - MicroSD slot for flash storage
  - 1 Gb/s Ethernet interface
  - 3.3 by 3.2 inches daughter card
  - ~ 10 watts => 3.2 GFLOPS/W
- SoC is a dual-core Exynos 5 chip from Samsung
  - Integrated CPU and Accelerators Model
  - 6.8 GFLOPS FP64 on the dual-core Cortex-A15 CPU at 1.7 GHz
  - 25.5 GFLOPS FP64 on the quad-core ARM Mali-T604 GPU



FP7 MontBlanc Project Carrier Blade

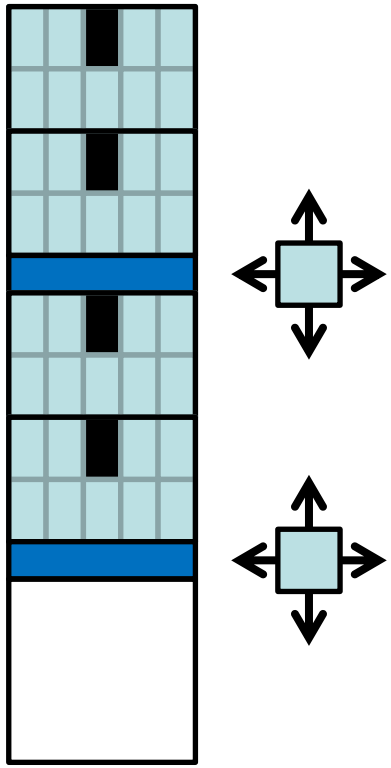
- Broadcom crossbar switch 1 GbE to 10 GbE
- 15x compute node daughter cards
- 485 GFLOPS for ~ 200 W
- 2.4 GFLOPS / W



FP7 MontBlanc Project System

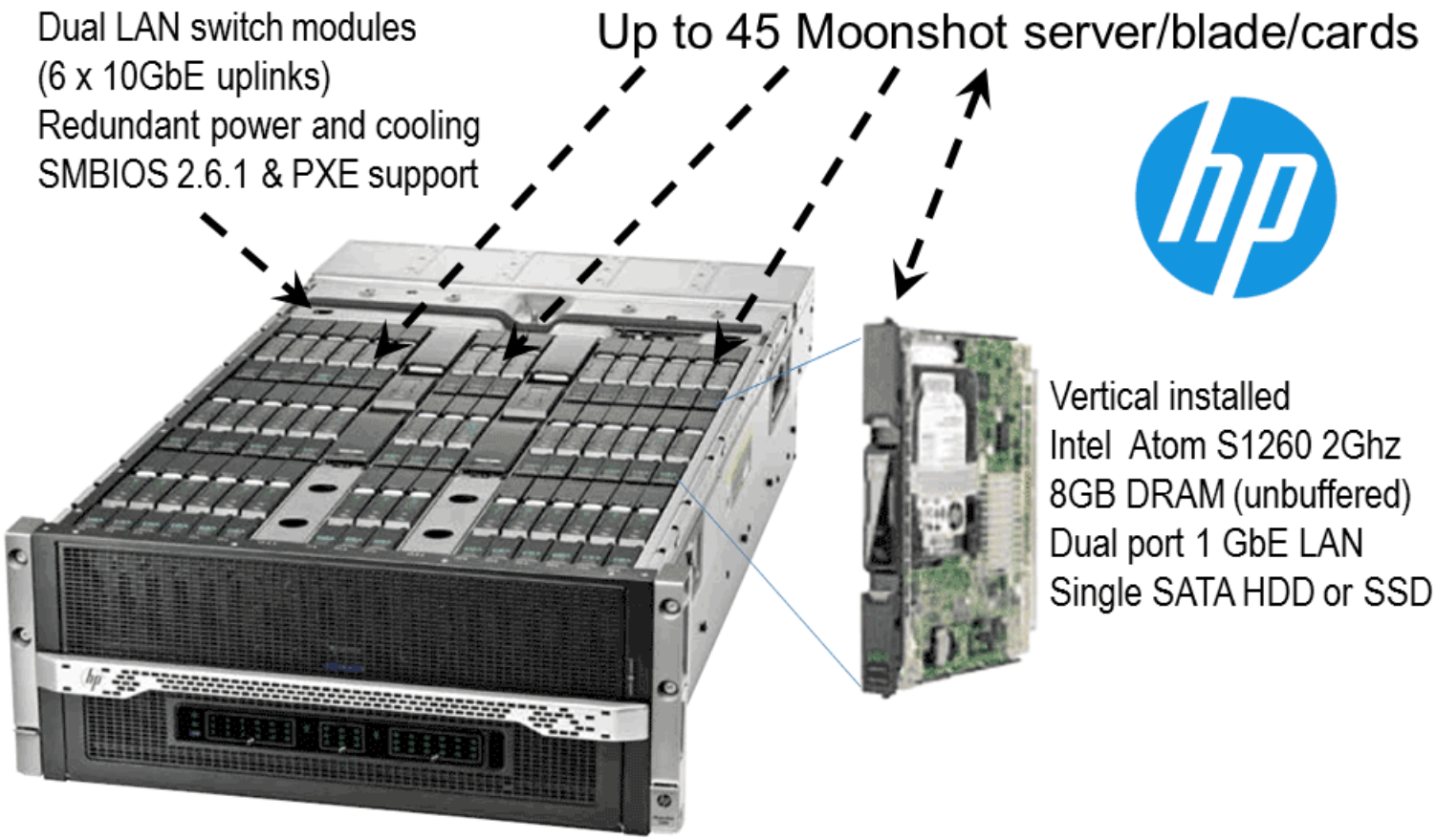
- Bull B505 7U blade chassis
  - 9x Carrier blade
    - 135 x Compute cards
  - 4.3 TFLOPS for ~ 2 KWatt
  - 2.2 GFLOPS/W

- Rack of 4 x blade cabinets
  - 36 blades
  - 540 compute cards
  - 2x 36-port 10GbE switch
  - 8-port 40GbE uplink
  - 17.2 TFLOPS for ~8.2 KWatt
  - 2.1 GFLOPS/W



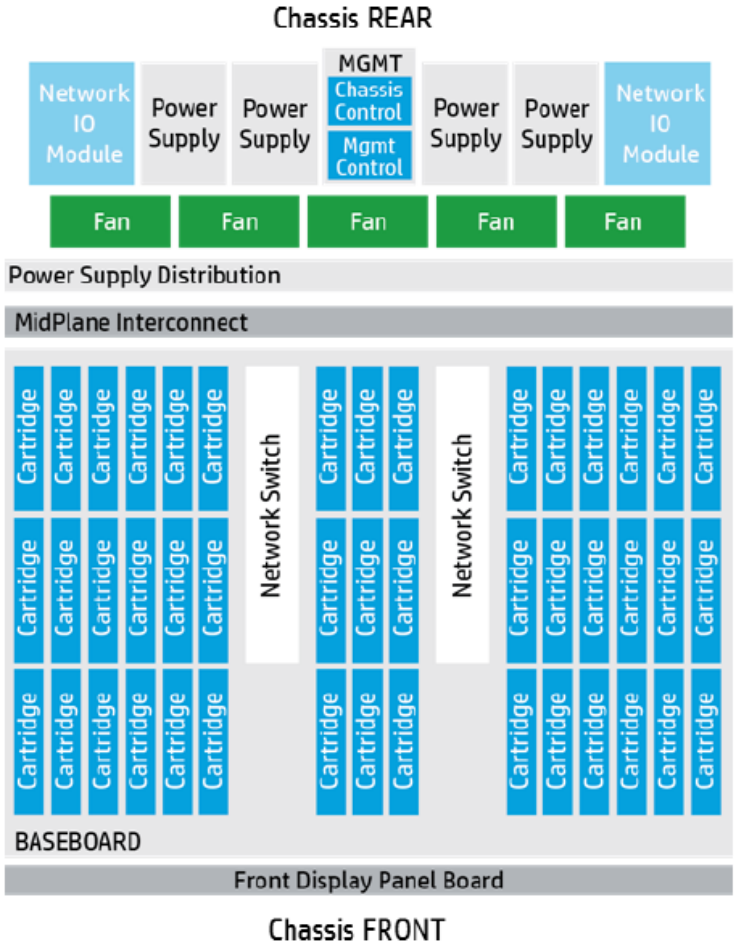
HP Moonshot 1500 System

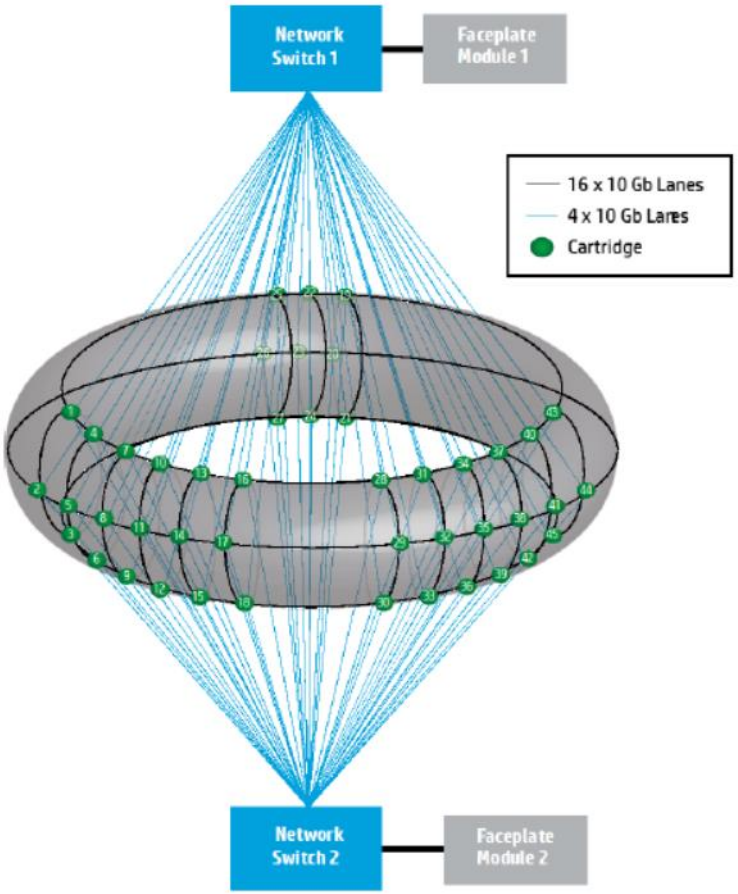
- The world's first software defined servers



HP Moonshot 1500 Chassis Interconnects

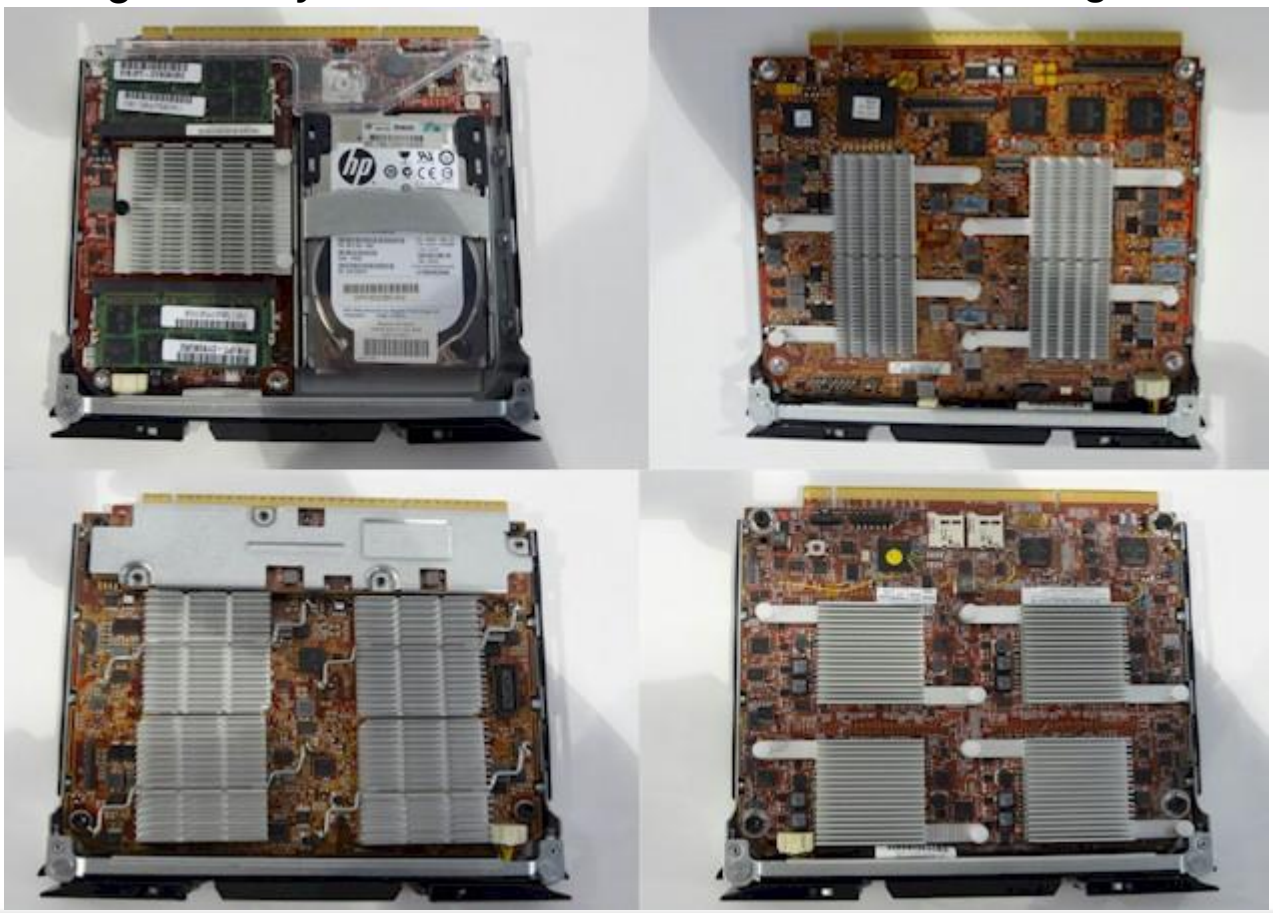
- Up to 2x 4x 10Gb Ethernet and 4x 4x 10Gb lanes per cartridge





HP Moonshot 1500 Cartridges end of 2013

- Clockwise: the m300 using Intel's Avoton, the m700 using AMD's Kyoto, the m800 using TI's KeyStone-II, and an unnamed card using Calxeda's Midway



HP Proliant m800 Server Cartridge

- SoC
  - Texas Instruments KeyStone-II based 66AK2H
  - Quad Cortex-A15 cores 1.0 GHz => 8 GFLOPS FP64
  - Eight C66x DSP cores 1.25 GHz => 60 GFLOPS FP64
- Cartridge
  - Four TI KeyStone-II SoCs
  - Dual 1 GB Ethernet (radial fabric)
  - Four SODIMM slots DDR3-1600 => 51.2 GB/s
  - 32 GB (4x8 GB) DDR capacity
  - 272 GFLOPS for ~85 W => 3.2 GFLOPS/W
- Chassis
  - 45x m800 Server Cartridges
  - 12.24 TFLOPS for ~4KW => 3.06 GFLOPS/W

Outline

- Node Abstract Architectures
- SoC-Based Projects and Systems
- **Kalray MPPA Manycore Processors**
- Perspectives and Conclusions

Kalray MPPA®-256 Andey Processor

256 VLIW processing engine cores + 32 VLIW resource management cores

Shipping since January 2013

- High processing performance
  700 GOPS – 230 GFLOPS SP
- Low power consumption
  5 – 15W at 200 – 400MHz
- High execution predictability
- High-level programming models
- PCI Gen3, Ethernet 10G, NoCX

MPPA® Processor Roadmap

2014	2015	2016	2017	2018

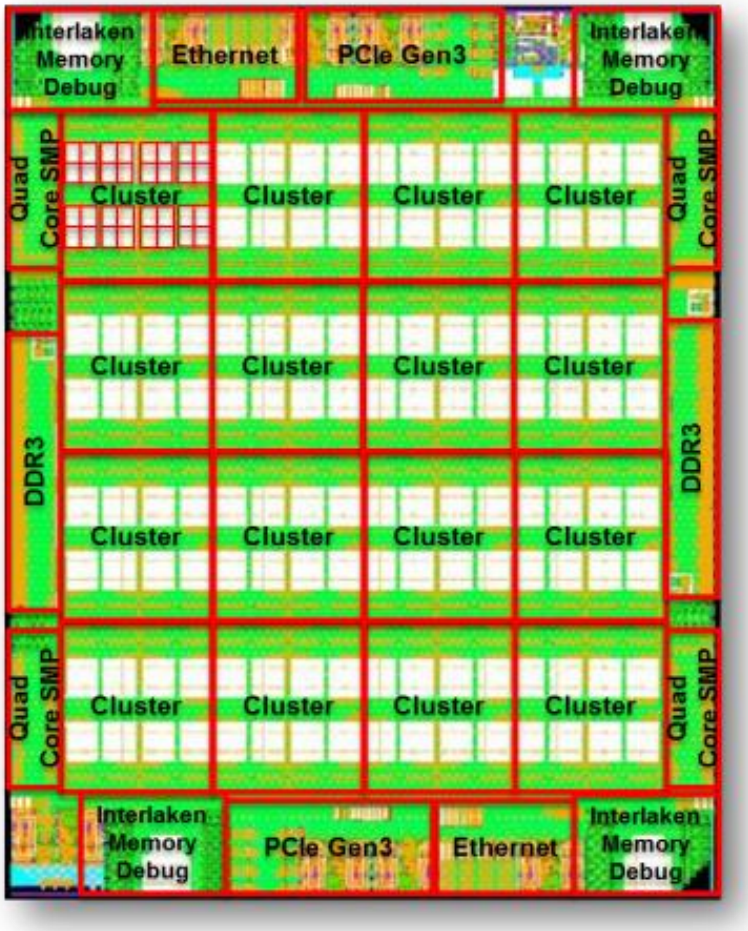
MPPA®-1024

MPPA®-256

MPPA®-64

28nm		20/16nm
Andey	**Bostan**	**Coolidge**
Kalray 1st generation*	**Kalray 2nd generation***	**Kalray 3rd generation***
32-bit addresses	32-bit / 64-bit addresses	64-bit addresses
230GFLOPS SP	700 GFLOPS SP	
25 GFLOPS/W	50 GFLOPS/W	75 GFLOPS/W
75 GOPS/W	80 GOPS/W	115 GOPS/W

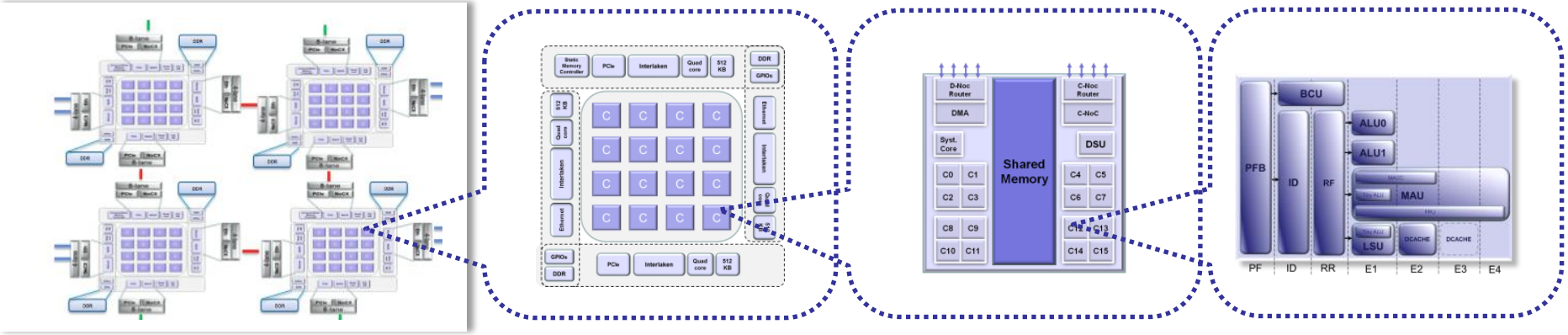
MPPA®-256 Processor I/O Interfaces



- DDR3 Memory interfaces
- PCIe Gen3 interface
- 1G/10G/40G Ethernet interfaces
- SPI/I2C/UART interfaces
- Universal Static Memory Controller (NAND/NOR/SRAM)
- GPIOs with Direct NoC Access (DNA) mode
- NoC extension through Interlaken interface (NoCX)

Quad MPPA®-256 TURBOCARD Architecture

1024 Kalray application cores and 8 DDR3-1600 under 80W



TURBOCARD Board

- 4 MPPA-256 processors
- 8 DDR3 SODIM
- PCIe Gen3 16x to host
- NoCX between MPPA

Manycore Processor

- 16 compute clusters
- 4 I/O subsystems with quad-core SMP and DDR memory
- 2 Networks-on-Chip

Compute Cluster

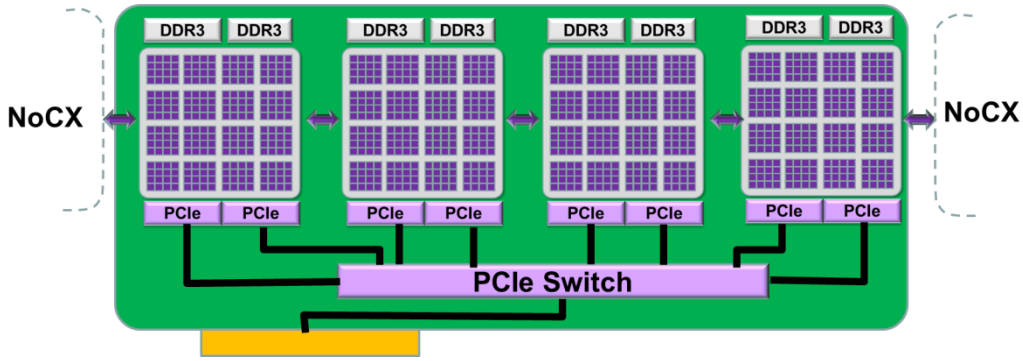
- 16 PE cores + 1 RM core
- NoC Tx and Rx interfaces
- Debug Support Unit (DSU)
- 2 MB of shared memory

VLIW Core

- 5-issue VLIW architecture
- Timing predictability
- 32-bit/64-bit IEEE 754 FPU
- MMU for rich OS support

Kalray TURBOCARD Family

1024 Kalray application cores and 8 DDR3 under 80W



Kalray TURBOCARD2

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

- First engineering samples: Q2-15
- Volume Production: Q3-15

Kalray TURBOCARD1

- 4 Andey MPPA®-256 processors
- 0.83 TFLOPS SP / 0.28 TFLOPS DP
- 8x DDR3L @1233 MT/s => 79 GB/s

- First engineering samples: Q3-14
- Volume Production: Q4-14

TURBOCARD for HEVC Encoder

From SD to UHD and beyond

Resolution	# Cores # MPPA®-256	MPPA Andey	MPPA Bostan
SD	48 ¼ MPPA®-256	1/20 TURBOCARD Less than 6W	1/40 TURBOCARD
1080p (HD) – 30fps	256 1 MPPA®-256	¼ TURBOCARD Less than 8W	1/8 TURBOCARD
4K (UHD) – 30fps	1024 4 MPPA®-256	1 TURBOCARD Less than 50W	½ TURBOCARD Less than 60W
4K (UHD) – 60fps	2048 8 MPPA®-256	2 TURBOCARD Less than 90W	1 TURBOCARD Less than 60W

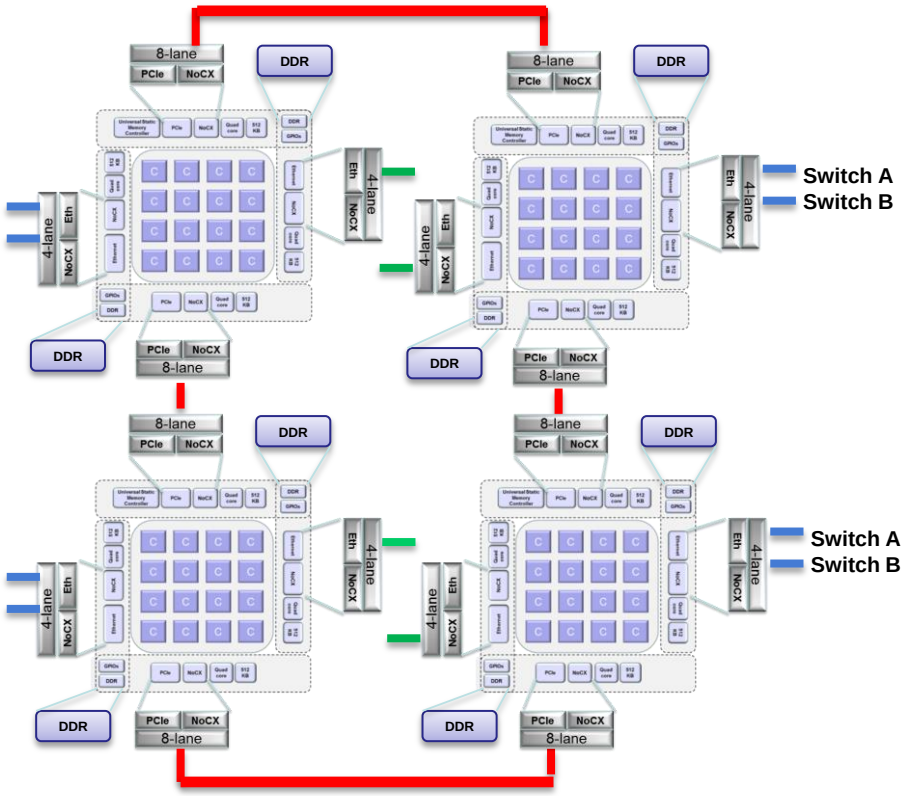
Competition	Resolution	CPU Load	Power (W)
CPU 2.66GHz 8-Core (16 thread)	1080p (HD) – 30 fps	130%	100

Outline

- Node Abstract Architectures
- SoC-Based Projects and Systems
- Kalray MPPA Manycore Processors
- **Perspectives and Conclusions**

Quad MPPA®-256 Moonshot Cartridge

1024 Kalray cores and 8 DDR3 under 85W

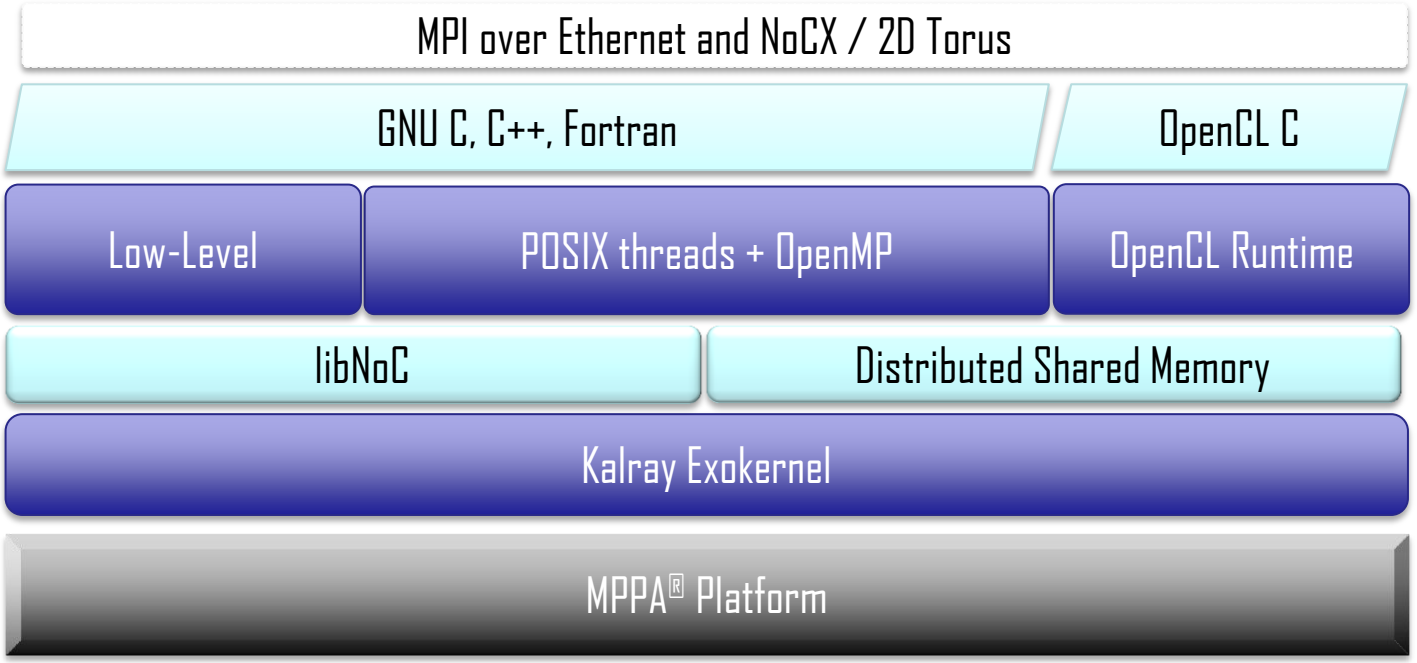


- Cartridge memory
  - 2x DDR3 per MPPA®-256
- Radial fabric
  - 8x 10G Ethernet
  - Switch A: 1 per MPPA
  - Switch B: 1 per MPPA
- NoCX fabric
  - 4x 10Gbps per link
  - 4 pairwise links
- Torus fabric
  - 4x 10Gbps per link
  - 4 neighbor links

MPPA® Bostan-Based Moonshot (2015)

- Bostan Cartridge
  - 4 x MPPA-256 Bostan processors with 4x DDR3 @ 2133MT/s
  - 2496 GFLOPS SP, 1248 GFLOPS DP @ 600MHz
  - 136 GB/s R + W DDR bandwidth
  - 64 GB DDR capacity
  - 14,7 GFLOPS/W
- HEVC Application
  - 1 Anders cartridge (Intel Haswell-based) per 8 Bostan cartridges
  - Offload by using ROCE (RDMA over Converged Ethernet)
- RTM Application
  - Only Bostan cartridges in Moonshot chassis
  - Storage servers directly addressed by the MPPA
  - 56.16 TFLOPS for 4KW => 14 GFLOPS/W

MPPA® Software Environment



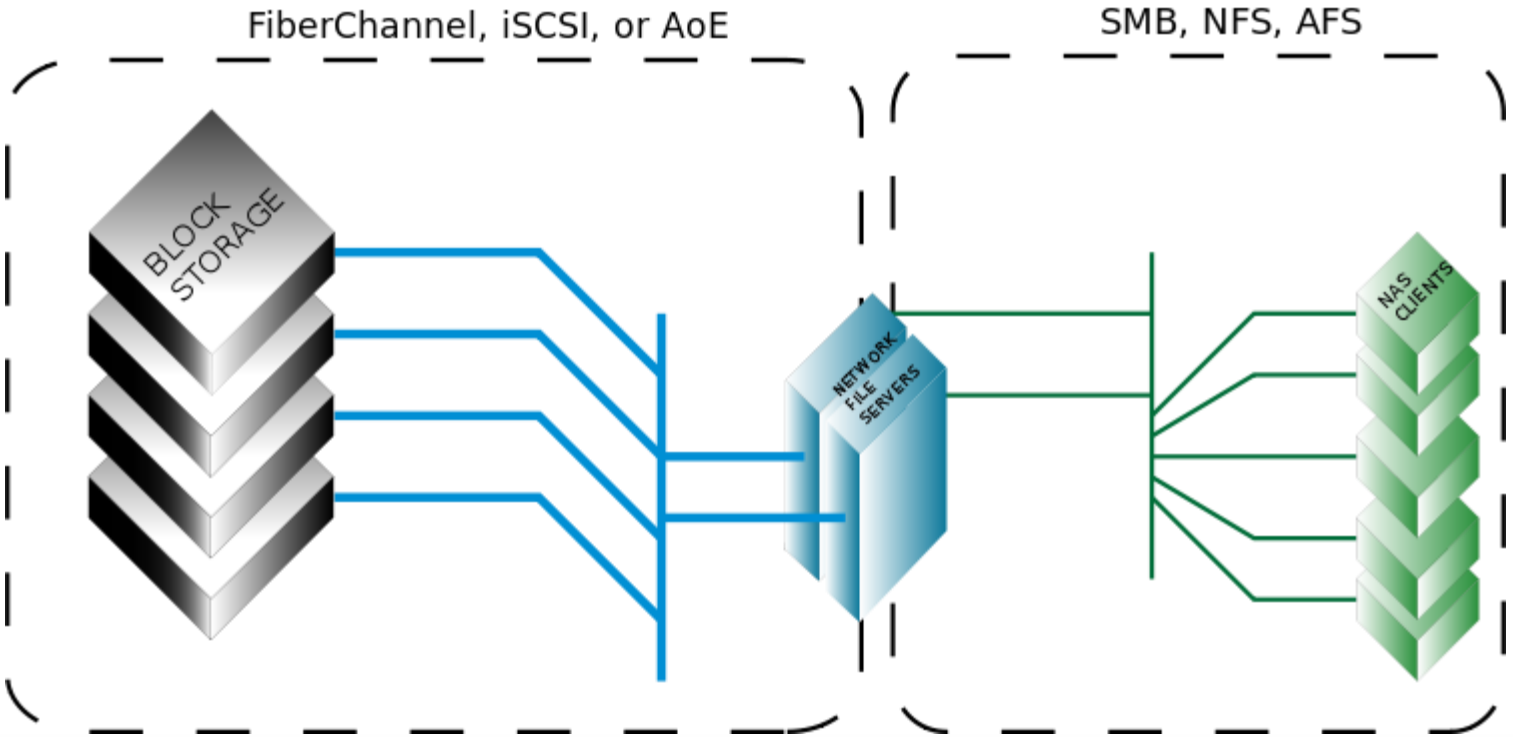
KALRAY

Network Processor Competition Landscape

	Packet throughput	PCIe I/F Throughput	ETH #Ports	#GOPs/ DDR3	Ethernet Features	Cypto
MPPA®-256 Bostan	**2x 40Gbps 120Mpps**	**GEN3 2 x 8-lane**	**2 x40G or 8 x 10G**	**1TOPs 2x 64-bit DDR3**	PTP v2 QoS	TRNG 80 Gbps Encryption
Tilera Gx36		8-lane GEN2	4 x10G	130 GOPs 2x 64-bitDDR3	PTP v2 Priority flow control datacenter	TRNG 40Gbps encryption Zip/unzip
Tilera Gx72	120 Mpps	2x 8-lane 2x 4-lane GEN2	8x10G	260 GOPs 4 x 64-bitDDR3 W/ ECC	PTP v2 Priority flow control datacenter	TRNG 40Gbps encryption Zip/unzip
CAVIUM Octeon II (1936 FCBGA)		1 x 8-lane 1 x 4-lane GEN2	Up to 4x10G	96GOPs 4x 64-bit DDR3 w/ ECC		

Network Storage for HPC and Big Data

- A storage area network (SAN) is a dedicated network that provides access to consolidated, block level data storage
- Network-attached storage (NAS) is file-level computer data storage server
- Network storage nodes are adopting network-oriented manycore processors



Summary and Conclusions

- SoC-based large-scale computing systems
  - First SoC-based supercomputer was IBM BG/Q, whose node implemented the **Homogeneous Multicore Model**
  - Current node architecture model for SoC-based computing is **Integrated CPU and Accelerators Model** (AMD Kaveri, NVIDIA Tegra)
  - Difference between a demonstration (FP7 MontBlanc) and a deployable system (HP Moonshot) is access to high-speed interconnect and storage
  - HP Moonshot is a capacity system with potential for a capability system
- Uses of MPPA® processors in SoC-based systems
  - The MPPA® processor is a SoC that implements the more advanced **Homogeneous Multicore-Accelerator-Memory Model**
  - System energy efficiency is higher, but programming models are more intrusive with mix of CPU nodes and of MPPA® processor nodes
  - Significant potential of MPPA® processors in large-scale systems are at the network storage (PCIe Gen4) and interface (Ethernet 25G) hubs

Thank you

Headquarters (Paris area)

86 rue de Paris,
91 400 Orsay
France
Tel: +33 (0)1 69 29 08 16
email: info@kalray.eu

Grenoble office

445 rue Lavoisier,
38 330 Montbonnot
France
Tel: +33 (0)4 76 18 09 18
email: info@kalray.eu

Japan office

CVML, 3-22-1, Toranomon,
Minato-ku, Tokyo 105-0001,
Japan
Tel: +81 80-4660-2122
email: japan@kalray.eu

USA office

4962 El Camino Real
Los Altos, CA
USA
Tel: +1 408-475-0550
email: usa@kalray.eu

MPPA® Processors for Next Generation Computing

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 Predicatbility

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

Ease of Use

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