# Kalray MPPA Many-Core Processors

Stefan Kosnac
Computer Architecture Group
University of Heidelberg, ZITI
Germany, 68131 Mannheim
Email: kosnac@stud.uni-heidelberg.de

***Abstract*—This paper has been written as part of the "Advanced Seminar Computer Engineering" at University of Heidelberg and gives a brief overview of the Many-Core Architecture developed by Kalray, focussing on the first generation processor called "Andey". It contains 256 user and 32 system cores implementing a 32-bit 5-issue VLIW architecture. The cores are divided in 16 compute clusters and 4 I/O subsystems and each got a private address space. A data and a control Network-On-Chip are used for communication and synchronization.**

**To provide a context, a short explanation of the VLIW-concept in general is given and the following Many-Core architectures are briefly presented: The Intel Many Integrated Core Architecture, the tile architecture from Tilera and a GPU. Three example applications this architecture could be used for are discussed (Seismic Wave Propagation, H.264 (AVC) Videoencoding and Monte Carlo Option Pricing). It turns out that mapping code to the MPPA architecture results in a relatively high energy efficiency compared to conventional solutions.**

## I. Introduction

There is no strict definition when a processor is called many-core but empirically the defining limit should be somewhere between 32 to 64 cores. Compared to current multi-core architectures many-cores run at lower frequencies and have a lower core complexity. Since cache coherence is hard to scale with that many cores, groups of cores use a private address space. This demands to usage of an on-chip interconnection network for communication and synchronization between the cores. Simple cores and low operating frequencies are resulting in a bad single thread performance. So why go this way? The power consumption of modern processors is getting higher and higher. At some point it is not possible anymore to prevent a processor from overheating. So reducing the power consumption is critical. The dynamic power of a chip is given by $P = \alpha C_{Load} V_{DD}^2 f$, where $\alpha$ is an activity factor between 0 and 1. When lowering the frequency of processors, the power consumption can be reduced more than linearly, because the operating voltage can be reduced at same time. By doing this and additionally building simpler cores it is possible to use several hundreds of them on a single chip. This results in a decent theoretical performance at low power consumption. The drawback of course is a much higher need for parallelism in software. Furthermore, there is the problem hardware gets more unreliable when moving to smaller processing nodes – probably due to increased leakage. Therefore, many cores could also perform redundant calculations to overcome this problem [1].

## II. MPPA Architecture

The MPPA (abbreviation for Multi-Purpose Processor Array) processors are developed by a french company called Kalray. Kalray was founded in 2008 and the headquarters are currently located in Paris. There are three different generations of MPPA processors as shown in Fig. 1. The first generation called "Andey" is a 32-bit 5-issue VLIW architecture with 256 user and 32 system cores on a single chip, manufactured in the $28\,\mathrm{nm}$ node (TSMC). The chip is running at $400\,\mathrm{MHz}$, consumes $8\,\mathrm{W}$ to $15\,\mathrm{W}$ and provides $70\,\mathrm{GFLOPS}$ double precision [5]. It is intended for low to medium volume professional applications, where energy efficiency and real time behaviour are required [2]. The second generation, called "Bostan", which became available at the end of 2015, is also manufactured in $28\,\mathrm{nm}$. It does not differ from the first generation in terms of general architecture, but it implements a 64-bit architecture and runs at $600\,\mathrm{MHz}$ to $800\,\mathrm{MHz}$. Therefore, a higher double precision floating point performance of $422\,\mathrm{GFLOPS}$ is achieved. Regarding memory, it can use more DDR memory and contains a little more on-chip memory. The third generation, called "Coolidge", is currently under specification and scheduled for release in late 2016 and 2017 [5]. It is planned to be manufactured in a smaller node size ($14/16\,\mathrm{nm}$) and will run at $1\,\mathrm{GHz}$. Furthermore, there will be three versions with a different number of cores (64, 256 and 1024). All three are 64-bit architectures.

### A. "Andey" Architecture

Fig. 2 shows a simplified block-diagram of the MPPA chip. On the edges, there are 4 I/O subsystems and there are 16 compute clusters in the middle. Each I/O subsystem contains a quad-core for symmetric multi processing with a shared data cache of $512\,\mathrm{kB}$. It can run either Linux or a real time OS called RTEMS (Real-Time Executive for Multiprocessor Systems) [10][12]. The DDR controller can access up to $64\,\mathrm{GB}$ of external DDR3-1600 memory. The PCIe interface supports 8-lane Gen3 PCIe for a peak throughput of $16\,\mathrm{GB/s}$ in full duplex. The Ethernet links each provide a throughput up to $40\,\mathrm{Gb/s}$. The compute clusters and I/O subsystems are connected through an on-chip network (NoC) and have their own private address space. So the overall architecture of the chip is similar to what is currently used in high

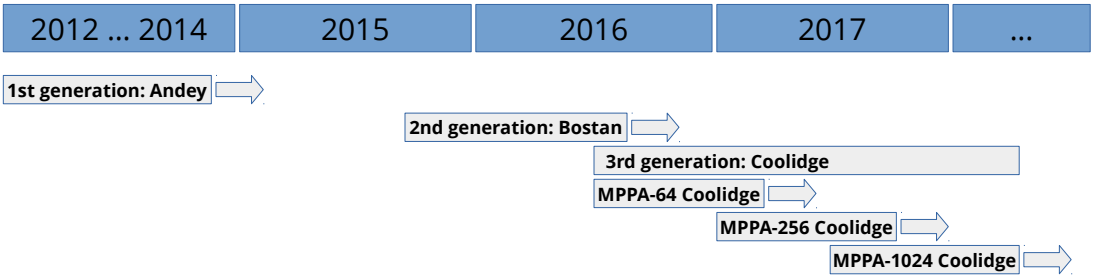


Fig. 1: MPPA processor timeline. Based on [5].

performance computing. Therefore, message passing is used to communicate between clusters. The network on chip can be extended over a PCB to other MPPA chips by the Interlaken interface. It is shown as NoCX here.

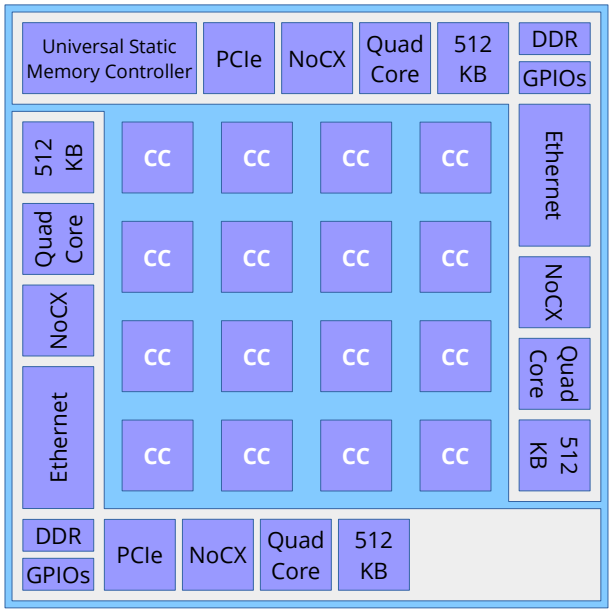


Fig. 2: MPPA "Andey" toplevel architecture. Based on [6].

### B. NoC - Network on Chip

The on-chip network implements a 2D torus topology between the 4 I/O cores on each I/O subsystem and the 16 compute clusters, which is shown in Fig. 3. The horizontal rings of the torus are drawn in black and the vertical ones in red. Compute clusters are not directly connected to neighbouring compute clusters. Instead, the neighbour always gets skipped or an I/O core is in between. This is a trick to avoid a long wire for connecting both borders. Although, the I/O cores on opposite sides are connected directly. These are the orange lines. To extend the network to other MPPA chips the I/O cores use their fourth link to connect to the Interlaken interface (blue lines). There are actually two networks in parallel with the same topology. One for bulk data transfer, called data NoC, and one for small messages at low latency, called control NoC. They also differ by the amount of packet buffering. Wormhole switching is used to distribute the packets. The flow control is minimal and is relying on software, to avoid deadlocks. By making the network explicitly configurable by software the programmer can predict its behaviour, which is important for real time applications [4].

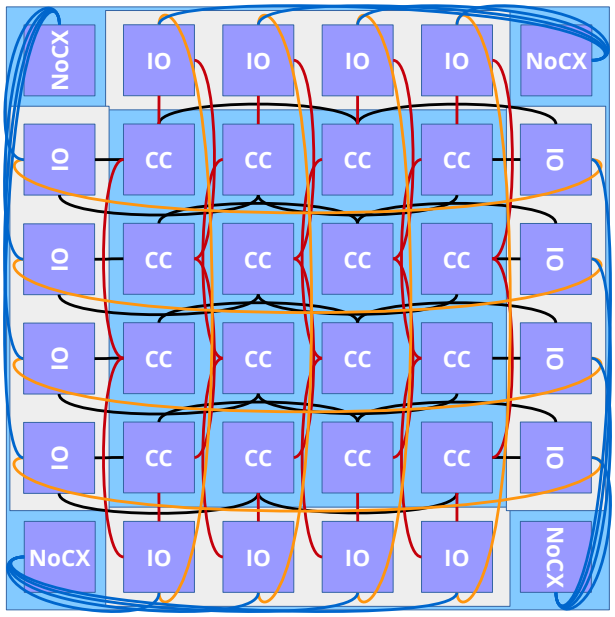


Fig. 3: Network on Chip topology. Based on [4].

### C. Compute Cluster

The compute clusters (Fig. 5) are the main processing units in the processor. Each contains 16 VLIW cores (processing elements) running user threads and one system core, which is sometimes also called resource management core. All 17 cores share the same design. As the name indicates the system core is responsible for system functions and does not run user threads. Instead it is operating the direct memory access (DMA) engine, among other things. The DMA is transferring data between the shared memory and the data NoC. It can deliver up to $3.2\,\mathrm{GB/s}$ in full duplex. The Debug and Support Unit (DSU) supports debug and diagnostic capabilities. It is connected to the outside with a JTAG chain. The DSU can deliver up to $1.6\,\mathrm{Gb/s}$ of trace data to give almost non-intrusive insight on the applications behaviour. The cores in

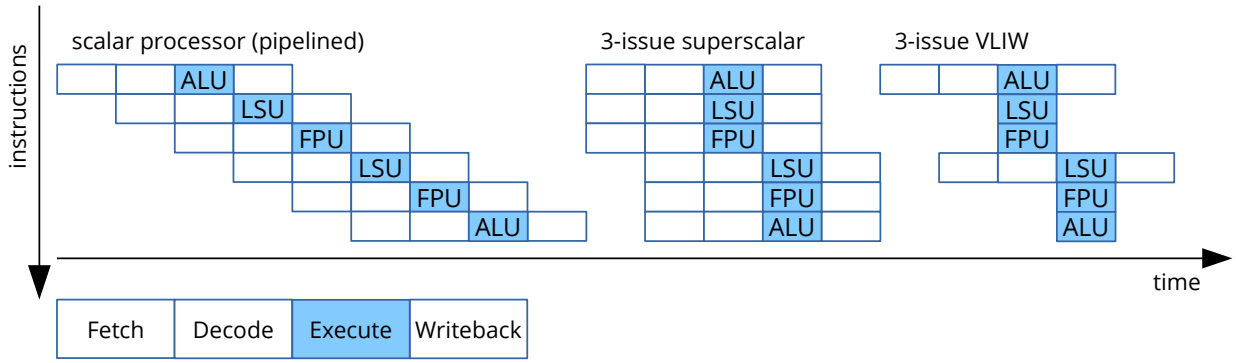


Fig. 4: Instruction issue behaviour. Based on [11].

a cluster share 2 MB local memory, which handles cache misses on the first level caches and buffers main memory. Additionally, the shared memory provides the main mechanism for implementing intercore communication in a compute cluster because the cluster is not cache coherent. The first level caches are inclusive with regard to the shared memory. It supports ECC on 64-bit words and is physically implemented in 128 kB banks. It can be accessed through 12 ports. Eight ports are arbitrated between pairs of processing elements, one is reserved for the system core, one for the DSU and two for the DMA. On each compute cluster runs an operating system called NodeOS. It implements an asymmetric multi-processing architecture that can use the asymmetry between the system core and the processing elements. According to [10], about 500 kB of the shared memory is always in use by the NodeOS, thus reducing the available shared memory significantly. The NodeOS is conforming to the POSIX API. Therefore, it is possible to use OpenMP on the compute clusters.

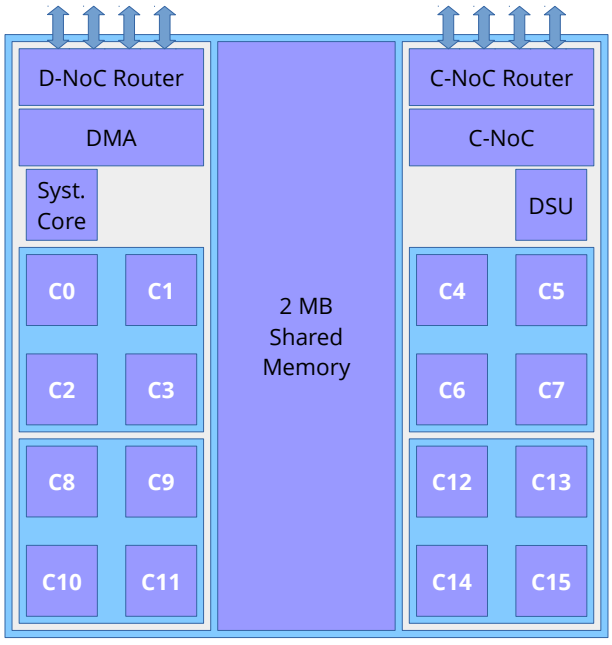


Fig. 5: Compute cluster. Based on [2].

### *D. VLIW-Core - Very Long Instruction Word*

Very Long Instruction Word is one among several ways to design an Instruction Set Architecture (ISA). It is based on bundling instructions together for simultaneous execution. Fig. 4 is a simplified view of the steps conducted by processors of different architectures when executing the instructions of an instruction stream. Instructions are issued from top to bottom in this picture. The fields here are “Fetch” for fetching new instructions from the instruction cache, “Decode” for the instruction decode unit and “Writeback” for writing back results to the register file. The “Execute”-stage can be for example a floating point unit, an arithmetic logic unit or a load and store unit. A scalar pipelined processor contains one instruction pipeline to exploit instruction level parallelism (ILP). So the maximum instructions-per-cycle ratio (IPC) is 1, i.e. every cycle an instruction is fetched with no no-operation (NOP) instructions in between. A 3-issue superscalar processor has three seperate pipelines for different execution units that can work in parallel. Furthermore, there is additional hardware to dynamically (during runtime) decide which instructions can be executed in parallel. This should not be confused with simultaneous multithreading, which is also known by Intel’s name “hyper-threading”. Simultaneous multithreading fetches instructions from several threads, whereas a superscalar processors is build to improve single-thread performance. With ideal conditions an $m$-issue superscalar processor can have an IPC of $m$. A 3-issue VLIW processor on the other hand uses longer instruction words, named bundles, which contain three instructions that can be executed in parallel by the different execution units. In this case, the decision which instructions can be executed in parallel needs to be made by a (good) compiler beforehand. This saves hardware and power to dynamically schedule instructions. An “ideal” instruction stream can reach an IPC of $m$ on a $m$-issue VLIW processor. In real cases the degree of ILP is between 2 to 5, i.e. there is no use in building $>$5-issue processors.

Advancing deeper into the hierachy one gets to the architecture of a VLIW core (Fig. 6). The MPPA cores are

composed of five different execution units. Two arithmetic & logic units (ALU), one multiply-accumulate & floating-point unit (MAU), one load/store unit (LSU) and one branch & control unit (BCU). Actually, there are more execution units: The LSU & MAU each contain an additional tiny ALU. However, as only five can execute at the same time, we have a 5-issue VLIW core here. The first stage of the instruction pipeline contains a prefetch buffer (PFB) which can hold up to three 128-bit entries. It decouples accessing instructions from loading them from the instruction cache. The instruction and data cache are both 2-way set associative and have a size of 8 kB. But they differ in terms of cache line size. The instruction cache has got 64 B lines, the data cache 32 B lines. The second stage does the instruction decoding. Next is a register read stage. The register file contains 64 registers, which are 32-bit wide, and supports 11 reads and 4 writes per cycle. Registers can be paired for 64-bit operations. The next four stages are the execution stages for the actual execution. Since scheduling of instructions is done by the compiler the pipeline is short compared to superscalar architectures. Therefore, the BCU has just a two cycle penalty for mispredicted conditional branches [4]. The ALUs can be paired for 64-bit integer operations. The MAU can handle integer and floating-point data. Different addressing modes are supported by the LSU. Not shown here is the write back stage.

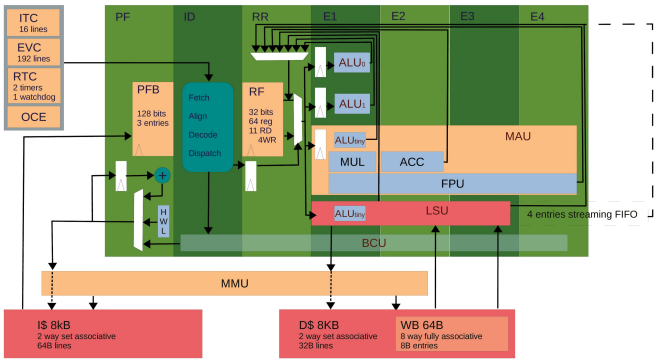


Fig. 6: VLIW core architecture [2].

## III. Programming & Software

There are two programming models available at the moment. Firstly, a cyclostatic dataflow (CSDF) language based on C syntax called ΣC. And secondly, since SMP Linux is running on the I/O subsystems and the NodeOS provides a POSIX API on the compute clusters, it is possible to use POSIX threads and OpenMP on the compute clusters. The CSDF model is a specialization of Kahn process networks and has the property that synchronization is implicitly expressed by the specified data dependencies. For further information on CSDF see [2]. Kalray provides support for the Eclipse IDE and tools for simulation, power measurement, etc. to simplify code development and testing. When mapping code to the MPPA architecture, one needs to be aware of the hardware level to effectively use the NoC and on-chip memory. Obviously, existing code needs to be compiled for the VLIW ISA, but since the MPPA architecture requires code changes a recompilation is necessary anyway.

## IV. Other Many-Core Processors

Beside the Kalray MPPA Processors there are other Many-Core processors available on the market. This section shortly presents some of them to give a comparison in terms of performance, power efficiency, architecture and coding effort.

### A. Intel MIC

Intel is offering its Many Integrated Core (MIC) architecture with the Xeon Phi coprocessors – often called with their codenames Knights Corner or Knights Landing for the second generation. There are currently three families of Xeon Phi Coprocessors: 3100, 5100, 7100. The 7100 family provides the highest performance. There are 61 cores running at 1.238 GHz which deliver around 1.2 TFLOPS of double precision floating point performance [7]. The processors have a Thermal Design Power (TDP) of 300 W. This is not exactly the consumed energy of the processor, but shows the direction. Another important feature is that the Intel Xeon Phi series uses the IA-32 instruction set architecture. So transferring code from a Xeon processor to a Xeon Phi coprocessor involves adding some directives and therefore is little work.

### B. Tilera

This company originates from a MIT research project and was founded in 2004. Their current processor architecture (TILE-Gx72) is very similar to the Kalray processor. It features 72 RISC cores running at 1.2 GHz with a typical power consumption of 65 W [8]. The processor is a 64-bit architecture. It uses a 2D-mesh network topology and has 18 MB of coherent L3 cache. This is surprising because one would expect no cache coherence in many-core processors. The MPPA processors don't have cache coherence and use a 2D-torus network topology. Unfortunatly, there is no information on the floating point performance available.

### C. GPUs

GPUs can be considered as many-core processors as well. To give a feeling how they compare to the previously presented architectures, the properties of the Tesla K40 are exemplarily stated in the following [9]: 2880 cores are running at frequencies between 745 MHz and 875 MHz achieving a double precision floating point performance of 1430 GFLOPS to 1660 GFLOPS. The whole PCB consumes a power of 235 W. Mapping code to a GPU is a bigger task.

## V. Application Area & Energy Efficiency

In contrast to the raw floating point performance in the previous section, this section will look at real applications: Seismic Wave Propagation, H.264 (AVC) Encoding and Monte Carlo Option Pricing.

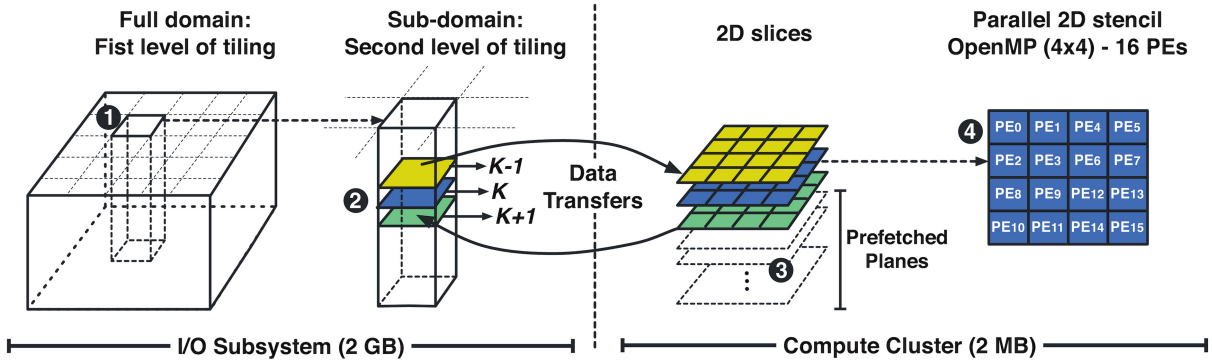


Fig. 7: Tiling scheme to exploit the MPPA memory architecture [10].

### A. Seismic Wave Propagation

Seismic wave propagation refers to simulating sound waves within the earth mantle. This is used to calculate damage invoked by future earthquakes or for oil and gas exploration. The equations for an isotropic medium are shown below. $v$ is the velocity of the grid points, $\sigma$ is the stress tensor, $F$ is an external force and $\lambda$ and $\mu$ are some parameters. These equations result in a stencil code.

$$\rho \frac{\partial v_i}{\partial t} = \frac{\partial \sigma_{ij}}{\partial x_j} + F_i \quad (1)$$

$$\frac{\partial \sigma_{ij}}{\partial t} = \lambda \delta_{ij} \frac{\partial v_k}{\partial x_k} + \mu \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \quad (2)$$

In [10] a MPPA platform with 2 GB memory was used. To fit the problem size to this memory, a cube with 180 grid point in every dimension was simulated. As mentioned before the 2 MB of shared memory in the compute clusters are caches for the main memory. The data needs to be tiled and transferred from DDR to shared memory and vice versa. The whole 3D domain is allocated in main memory and divided in sub-domains corresponding to the number of compute clusters (Fig. 7, point 1). These zones have a two grid point overlap due to the fourth order stencil code. They are still to large to fit into the shared memory of each cluster. Therefore, each sub-domain is divided into 2D slices (point 2). To overlap computation and slice transfer into shared memory an "explicit prefetching mechanism" (point 3) is applied. On the compute clusters OpenMP is used for parallelization. This implementation has been compared to reference implementations on a Xeon E5 (CPU) and a Quadro K4000 (GPU). The results of 500 time steps are shown in Fig. 8. The MPPA processor may be the slowest to calculate the solution but it consumes less energy in doing so. For more information on how scalable this application is see [10].

### B. H.264 (AVC) Encoding

H.264, Advanced Video Coding (AVC), is currently the standard format for compression of video, especially high definition video. Although, software and hardware implementations of the sucessor standard High Efficiency Video Coding (HEVC) are under heavy development these days. In [2]

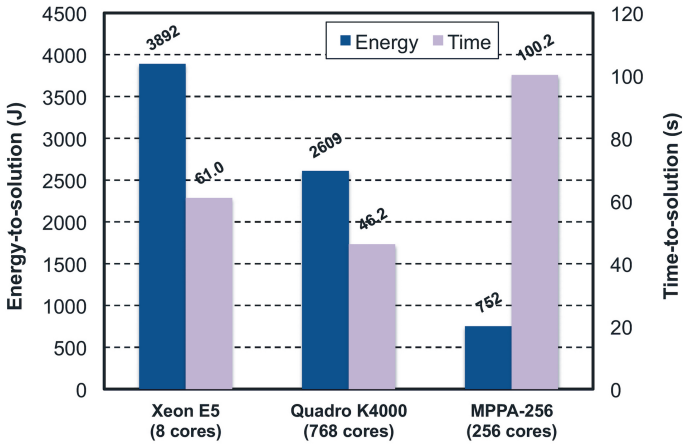


Fig. 8: Comparison of the different systems [10].

a H.264 Encoder has been implemented and compared to the x264 library, which is an open source implementation of a H.264 encoder and technically on the same level as commercial encoders. The x264 library is executed on an Intel Core i7-3820. Using structural similarity (SSIM) and peak signal to noise (PSNR) for objective quality measurement and taking the file size into account, the ΣC implementation shows better results. Also, the energy efficiency of the MPPA processor is much higher. But the encoding performance in frames per second (fps) is roughly the same (Tab. I).

| Processor | Performance | Energy efficiency |
|---|---|---|
| Intel Core i7-3820 | 49 fps | 2.60 W/fps |
| Kalray MPPA1-256 | 52 fps | 0.14 W/fps |

TABLE I: H.264 Encoder performance results [2].

The better video quality achieved on the MPPA processor is due to the fact that many motion vectors could be calculated in parallel. So much more information is available when deciding what is the best way to encode frames.

### C. Monte Carlo Option Pricing

This application is important for the financial sector. Options are financial objects which allow the holder to buy or sell

something to a given price at a certain time. There are different types of options. European options are exercised at a certain time, whereas American options can be exercised at any time before a given deadline [3]. For simple European options the Black-Scholes-equation (3) describes the price of the option $V(S,t)$ as a function of stock price and time. $r$, the so called risk-free interest rate, and $\sigma$, the volatility of the stock, are parameters. There are thousands of options that need to be priced at a given time. So there is a high demand for parallelism. This fits perfectly to the MPPA architecture and of course HPC in general.

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV - rS\frac{\partial V}{\partial S} \tag{3}$$

Tab. II shows the results for a performance analysis of monte carlo option pricing. Since this problem is highly parallel we get the best execution performance on the GPU. The MPPA processor is slower by more than a factor of two, but the energy consumed is more than five times lower. The CPU is way off and delivers the worst performance.

| Processor | Time [s] | Performance | Energy [J] |
|---|---|---|---|
| Intel Core i7-3820 | 13.86 | 0.17 | 1802.2 |
| NVIDIA Tesla C2075 | 2.37 | 1.00 | 531.7 |
| Kalray MPPA1-256 | 5.75 | 0.41 | 86.3 |

TABLE II: Option Pricing performance results [2].

### D. Further Application Areas

There are many other application areas the MPPA processors could be used in. Kalray offers a PCIe card for data centers to accelerate encryption. Furthermore, they could be used as powerful RAID cards for storage. There is also a HEVC video encoding solution implemented by Kalray, capable of encoding a full HD stream at 30 fps, i.e. in real time. Only mentioned briefly in this work, the properties of the processor make it possible to use it in real time applications.

## VI. Conclusion

The Kalray MPPA processor "Andey" (first generation) contains 256 user and 32 system cores running at $400\,\mathrm{MHz}$. It has a 32-bit 5-issue VLIW core architecture and uses an on-chip network for communication between the compute clusters of 16+1 cores. In terms of energy efficiency the presented applications showed excellent results. In my opinion, the MPPA processors are performing well in their intended application area, which are "low to medium volume professional applications" [2]. Regarding energy efficiency one could think of using the processors on a bigger scale on hundreds of nodes. But this would result in a huge number of cores demanding extreme parallelism to compensate for the low single thread performance. The strengths of these processors are embedded applications where a high compute power is needed at a low power budget and/or time predictability is important.

## References


[1] A. Vajda: *"Programming Many-Core Chips"*, 2011, Springer, chapter 2

[2] B. D. de Dinechin et al.: *"A Clustered Manycore Processor Architecture for Embedded and Accelerated Applications"*, 2013, IEEE Conference Publications, pp. 1-6

[3] V. Cvetanoska, T. Stojanovski: *"Using high performance computing and Monte Carlo simulation for pricing american options"*, CIIT Conference, April 2012, Bitola Macedonia

[4] D. Kanter and L. Gwennap: *"Kalray Clusters Calculate Quickly"*, 2015, Linley Group (article)

[5] B. D. de Dinechin: *"Next-Generation Accelerated Computing"*, 2012, Kalray (presentation), p. 25, http://www.anciens-amis-cnrs.com/pdf/Dupont_de_Dinechin.pdf [13.12.2015]

[6] Product Brief: *"MPPA1-256 Andey generation"*, http://www.kalrayinc.com/kalray/downloads/ [04.12.2015]

[7] Product Brief: *"The Intel Xeon Phi Product Family"*, http://www.intel.com/content/www/us/en/high-performance-computing/high-performance-xeon-phi-coprocessor-brief.html [14.12.2015]

[8] Product Brief: *"TILE-Gx72 Processor"*, http://www.tilera.com/files/drim_TILE-Gx8072_PB041-04_WEB_7683.pdf [20.12.2015]

[9] Board Specification: *"TESLA K40 GPU ACCELERATOR"*, http://international.download.nvidia.com/tesla/pdf/tesla-k40-passive-board-spec.pdf [14.12.2015]

[10] M. Castro et. al: *"Energy Efficient Seismic Wave Propagation Simulation on a Low-power Manycore Processor"*, 2014, IEEE 26th International Symposium on Computer Architecture and High Performance Computing, pp. 57-64

[11] K. Hwang: *"Advanced Computer Architecture: Parallelism, Scalability, Programmability"*, 1993, McGraw-Hill Publishing, chapter 4

[12] RTEMS Real Time Operating System, https://www.rtems.org/