Heterogeneous 3D Integration for a RISC-V System with STT-MRAM

Lingjun Zhu, Lennart Bamberg, Anthony Agnesina, Francky Catthoor, *Fellow, IEEE*, Dragomir Milojevic, Manu Komalan, Julien Ryckaert, Alberto Garcia-Ortiz, *Senior Member, IEEE*, and Sung Kyu Lim, *Senior Member, IEEE*

Abstract—Spin Torque Transfer Magnetic RAM (STT-MRAM) is a promising Non-Volatile Memory (NVM) technology achieving high density, low leakage power, and relatively small read/write delays. It provides a solution to improve the performance and to mitigate the leakage power consumption compared to SRAM-based processors. However, the process heterogeneity and the sophisticated back-end-of-line (BEOL) structure make it difficult to integrate the STT-MRAM in two-dimensional integrated circuits (2D ICs). In this paper, we implement a RISC-V-based processor with STT-MRAM using a heterogeneous 3D integration methodology. Compared with the SRAM-based 2D counterpart, the MRAM-based 3D IC provides up to 17.55% silicon area saving, together with either 34.74% performance gain or 13.90% energy reduction.

1 Introduction

With the development of memory-intensive applications such as machine learning, computer vision, etc., the demands of memory capacity and bandwidth grow rapidly in modern processor systems. However, the traditional SRAM-based cache systems are faced with several challenges: (1) The density of SRAM is hard to improve as the transistor scaling is approaching the physical limit [9]. (2) The system performance is limited because of the interconnection overhead between the logic modules and the large SRAM blocks in 2D ICs. (3) The sub-threshold leakage power becomes significant in designs with large SRAM blocks [4]. Therefore, it is necessary to develop alternative memory technology and physical design methodologies for the next-generation processors.

STT-MRAM is one of the most promising NVM technology to replace the SRAM. Using the Magnetic Tunnel Junction (MTJ), the STT-MRAM blocks consume negligible standby leakage power except for the peripheral circuits, which mitigate the power consumption in the SRAM-based cache system [10]. As the optimal 1T-1MTJ MRAM bit-cell has 75% smaller area compared to the conventional SRAM bit-cell, the STT-MRAM block is up to 66.7% smaller than the SRAM counterpart with the same capacity [6]. In addition, the STT-MRAM is compatible with the logic devices, so it can be integrated together with the standard cells on the same tier.

On the downside, the STT-MRAM has a unique back-end-of-line (BEOL) structure: it requires extra BEOL layers to complete routing within the memory block, due to the unique structure of the MTJ. This structure creates more obstructions in the higher routing BEOL layers, proposes challenges to traditional 2D physical design methodologies, and may increase the manufacturing costs. The existing studies have explored the benefits of STT-MRAM at the architectural level [5], [8], but BEOL structure and the impacts of STT-MRAM on placement and routing (P&R) and at the physical-design level have not been sufficiently considered yet.

Memory-on-logic (MoL) 3D is a heterogeneous 3D integration scheme which separates memory blocks and logic modules into two tiers, and then stacks the memory tier on top of the logic tier. Using Cu-Cu direct bonding technology, the BEOL layers of the two tiers are connected and form a face-to-face 3D (F2F 3D) structure. Due to the small diameter and pitch (less than $2um$) of the F2F vias [3], this 3D integration method provides a high vertical interconnection density at a low manufacturing cost. By integrating the STT-MRAM in this MoL 3D ICs, we provide more flexibility for memory designing and mitigate the negative effects of the BEOL structure of the STT-MRAM on P&R.

In this paper, we, for the first time, integrate STT-MRAM blocks into MoL 3D ICs. Using OpenPiton [1], a silicon-proven RISC-V-based System-on-Chip (SoC) as the benchmark design, we implement the 2D and 3D baseline designs with SRAM and STT-MRAM. We compare power, performance, and area (PPA) of the designs at the maximum frequency. Results show that our MRAM-based 3D IC has 17.55% smaller silicon area, and provides either 34.74% higher frequency, or 13.90% smaller energy consumption, compared with the SRAM-based 2D IC.

2 Methodologies

2.1 OpenPiton System Setup

We use a single tile of the OpenPiton system as the benchmark design. Fig. 1 shows the architecture of the many-core system and the single tile. The OpenPiton architecture is highly configurable in terms of core number, cache size, etc., while the tile is an atomic unit of the many-core system. Therefore, our experimental results and conclusions on the single tile also apply to the many-core system. The single tile contains a 64-bit RISC-V-based in-order Ariane core, which supports out-of-order execution and in-order commit [11]. The core is integrated with L1, L2, L3 caches, while the network-on-chip (NoC) routers are used to establish inter-tile and inter-chip communication. Table 1 shows the detailed configuration of the

- *Lingjun Zhu, Anthony Agnesina, and Sung Kyu Lim are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. E-Mail: {lingjun, agnesina}@gatech.edu; limsk@ece.gatech.edu*
- *Lennart Bamberg and Alberto Garcia-Ortiz are with the Institute of Electrodynamics and Microelectronics (ITEM) of the University of Bremen, Germany. E-Mail: {bamberg, agarcia}@item.uni-bremen.de*
- *Francky Catthoor, Dragomir Milojevic, Manu Komalan, and Julien Ryckaert are with IMEC, Belgium. E-Mail: {Francky.Catthoor, Dragomir.Milojevic.ext, Manu.Perumkunnil, Julien.Ryckaert}@imec.be*

Manuscript submitted: 07-Mar-2020. Manuscript accepted: 06-Apr-2020. Final manuscript received: 13-Apr-2020.

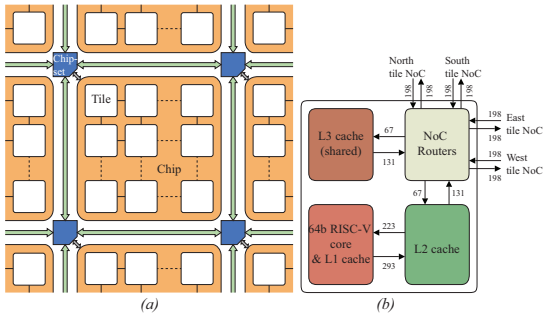

Fig. 1. Architecture of (a) the many-core OpenPiton system; (b) a single tile (adopted from [1]).

TABLE 1
Cache configurations for the single-tile OpenPiton design. (*) All the caches are single-port RAM except the dual-port L3 state array.

level	L1 inst.	L1 data	L2	L3
tech.	SRAM	SRAM	SRAM	SRAM/MRAM
size	8 kB	16 kB	16 kB	256 kB
shared	no	no	no	yes
# of ways	4	4	4	4
line size	256 bit	128 bit	128 bit	512 bit
# of banks	1	1	1	1 per tile
port	single	single	single	single*

caches. The memory blocks occupy more than 50% of the area in the SRAM-based 2D IC, and the SRAM blocks of the L3 caches are especially large in size. This motivates us to move the memory blocks to another tier in 3D ICs and to replace the L3 data array with STT-MRAM. We do not replace the L1 or L2 caches with STT-MRAM because the STT-MRAM with such a small capacity does not provide significant area or energy benefits due to the overhead of the peripheral circuitry [6].

2.2 STT-MRAM Generation and Modeling

We generate the STT-MRAM block using the IMEC STT-MRAM compiler for the 28 nm technology node [6]. We use two of these STT-MRAM blocks to replace the L3 data array inside the tile. The size of the STT-MRAM block is $224um \times 449um$, 57.85% smaller than the SRAM counterpart with the same capacity, due to its high memory density. However, the STT-MRAM occupies two extra BEOL layers (M5, M6), which causes routing problems in 2D ICs as extra BEOL layers are needed to access the pins of the STT-MRAM. The geometric properties and BEOL structure of the STT-MRAM block are reflected in the library exchange file (LEF). The timing and power information of the STT-MRAM block are stored in the Synopsys® Liberty File (LIB) generated by the memory compiler.

2.3 Physical Design and Evaluation

To implement the MoL 3D IC, we use Macro-3D, the physical design methodology proposed in [2]. In this flow, we place standard cells only on one tier (the logic tier), while the other tier (the macro tier) is full of memory macros. With the initial floorplan, we project the memory pins of the macro tier to the logic tier and complete P&R with a 3D metal stack. We assume that the F2F via size is 0.5 um, and the pitch is 1.0 um in the 3D designs. With these technology settings, the Macro-3D flow generates a 3D layout highly optimized by commercial Electrical Design Automation (EDA) tools.

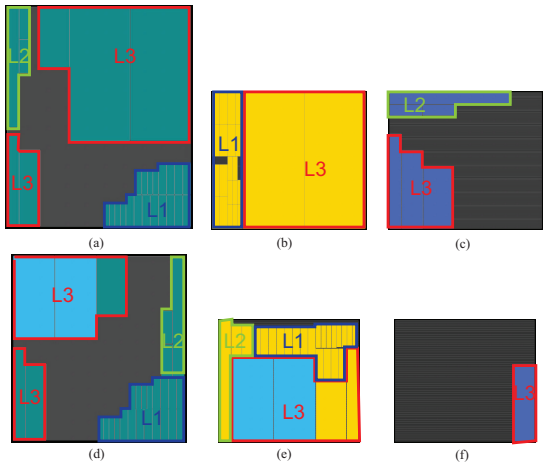

Fig. 2. Floorplan of the OpenPiton single-tile designs: (a) SRAM 2D; (b) SRAM 3D macro tier; (c) SRAM 3D logic tier; (d) MRAM 2D; (e) MRAM 3D macro tier; (f) MRAM 3D logic tier. The light blue blocks are the STT-MRAM as the L3 data array.

We compare four different types of implementation in this paper: (1) SRAM 2D: the conventional 2D IC with only SRAM blocks; (2) SRAM 3D: the MoL 3D IC with only SRAM block; (3) MRAM 2D: the 2D IC with STT-MRAM as the L3 data array and SRAM as the other caches; (4) MRAM 3D: the MoL 3D IC with STT-MRAM as the L3 data array and SRAM as the other caches. We conduct the max-performance and iso-performance experiments to evaluate the performance, power, and area (PPA) of these 2D and 3D ICs. In the max-performance experiment, we sweep target frequency for each IC and report the PPA metrics at the maximum frequency; in the iso-performance experiment, the ICs are all implemented at the same target frequency (500 MHz). The timing and power metrics are reported by Synopsys PrimeTime® at the typical corner.

3 Experimental Results

3.1 Floorplan and Area Saving

We design the 2D and 3D floorplan separatedly for each type of implemenation, as shown on Fig. 2.

For the 2D floorplan, we carefully place the memory macros to minimize the memory-to-memory distance and leave a large empty space for standard cell placement in the center. However, there are still long interconnections between memory blocks that can hardly be optimized in 2D, for example, the connections between the L3 data array and the L3 directory array for cache coherence [1]. We apply a few guidelines when designing the 3D floorplan: 1) Ensure the 3D floorplan size is around 50% of the 2D floorplan size. 2) Move as many L1 blocks and large L3 blocks as possible to the macro tier. 3) Separate memory blocks with long 2D interconnects to different tiers. With these guidelines, we maximize the benefits of 3D integration on the area saving and wirelength reduction.

The combination of STT-MRAM and 3D integration provides significant area saving. By replacing the L3 data array with the smaller STT-MRAM blocks, we are able to reduce the 2D floorplan area by 21.45%. And by implementing the design as a 3D IC, we further reduce the footprint area by 50%. Due to the timing benefits described in the next section, the P&R engine does not need to insert too many buffers in the 3D ICs to meet the timing constraints at higher frequencies. As a result, the 3D implementation with STT-MRAM has 17.55%

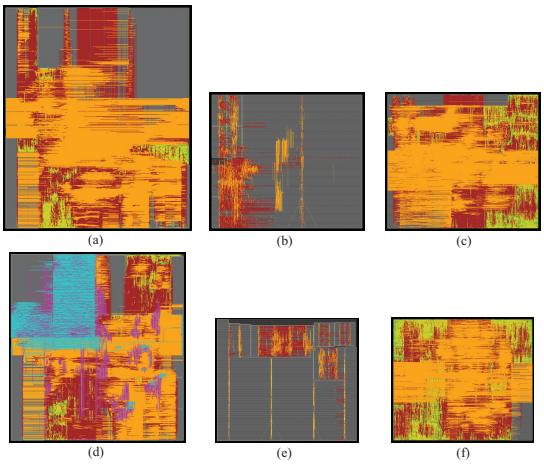

Fig. 3. Layouts of the OpenPiton single-tile designs: (a) SRAM 2D; (b) SRAM 3D macro tier; (c) SRAM 3D logic tier; (d) MRAM 2D; (e) MRAM 3D macro tier; (f) MRAM 3D logic tier.

smaller silicon area and 58.78% smaller footprint compared to the SRAM-based 2D counterpart.

3.2 Max-Performance Experiments and Timing Analysis

We implement the 2D and 3D ICs with Cadence® Innovus®. The P&R runtime for each implementation is around 16 hours on a 32-core machine, with fewer than 10k design rule violations.

Fig. 3 shows the layouts of the 2D and 3D ICs after P&R. The routing layer numbers are different for the 2D and 3D ICs: 1) for the 2D IC with SRAM only, we use 6 BEOL layers for routing; 2) while for the 2D IC with STT-MRAM, 8 BEOL layers are necessary to allow the signal nets to access the pins of the STT-MRAM; 3) for the 3D ICs, we use 6 layers for the logic tier and 6 layers for the macro tier. The power delivery network (PDN) is not implemented in these ICs because we focus on signal routing in this work. The 3D pin density is similar to the pin density in the 2D designs, and F2F via number (lower than 5000) in these 3D designs are much smaller than in previous studies [7], because the macro tier only has a few memory pins and the F2F vias are mainly used to access these pins in MoL 3D. However, these 3D interconnections are optimized to overcome the routing problems in 2D.

Table 2 shows the PPA metrics of the ICs in the max-performance analysis. First, if we compare the 2D and 3D ICs with SRAM only, the 3D IC has 14.91% frequency improvement. This is mainly contributed by the wirelength saving (-12.88%) and wire capacitance reduction (-7.86%) with vertical interconnection in the 3D IC. On the other hand, in the 2D IC with STT-MRAM, the wirelength reduction is also negligible, because the BEOL structure of the STT-MRAM blocks signal routing. In contrast, in the 3D IC with STT-MRAM, the obstructions of STT-MRAM are removed from the logic tier, and the small size of the STT-MRAM allows more memory blocks, including all the L2 caches, to move to the macro tier. Therefore, the maximum operating frequency of the MRAM-based 3D IC is even higher (+17.26%) than the SRAM-based 3D IC.

Fig. 4 shows the critical paths of the 2D and 3D ICs. In the 2D IC with SRAM only, the critical path is from one register to one of the L1 cache blocks. This suggests the register-to-memory paths tend to become the timing bottleneck in 2D ICs. We notice that the long interconnections are hard to avoid in 2D. If the memory blocks are clustered together, the memory-to-memory connections will be shortened, but the register-to-memory distance will be enlarged. Similarly, if the memory blocks are spread out, there will be enough space to place registers around the memory blocks and to reduce the register-to-memory distance, but the memory-to-memory paths tend to become critical. As a result, in 2D ICs, designers have to optimize the floorplan carefully and balance these constraints.

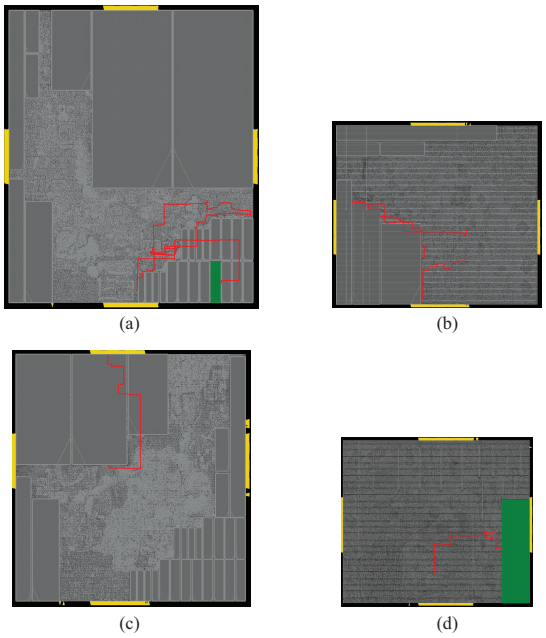

Fig. 4. Critical paths in the OpenPiton single-tile design. The cells and memory blocks on the critical paths are highlighted with the green color. (a) SRAM 2D; (b) SRAM 3D logic tier; (c) MRAM 2D; (d) MRAM 3D logic tier.

However, in the 3D IC with SRAM, the register-to-memory path is no longer critical, but the register-to-register path is still affected by the obstructions of the remaining SRAM blocks on the logic tier. In the 2D IC with STT-MRAM, we clearly observe that the register-to-output path is affected by the BEOL structure of the STT-MRAM blocks.

In the 3D IC with STT-MRAM, the benefits of 3D integration and STT-MRAM are combined, leading to 21.65% wirelength saving and 34.74% performance improvement. The F2F via number is not large, but the vertical interconnections are fully optimized by the router in the 3D space. Because most of the memory blocks are moved to the macro tier, the placement and routing blockages on the logic tier are minimized. In addition, the small footprint size helps reduce the distance from the I/O pins and the registers. As a result, all the register-to-memory, register-to-register, and register-to-I/O paths are improved.

Although the 3D IC with STT-MRAM has 33.14% higher total power compared to the SRAM-based 2D baseline, it does not mean the design is less energy-efficient. Since it runs at 34.74% higher frequency, the energy-delay product is actually 1.19% lower than the 2D IC with SRAM. Hence, the 3D IC with STT-MRAM can execute a task 1/3 faster while consuming less energy. The energy benefits will be further exploited in the iso-performance experiment.

3.3 Iso-Performance Experiments and Power analysis

Table 3 shows the iso-performance analysis results. The 3D IC with STT-MRAM provides 13.90% total power saving and energy efficiency improvement, compared to the SRAM-based 2D IC, which is mainly contributed by the switching and logic power reduction, due to smaller wirelength and buffer number.

TABLE 2
Max-performance analysis of the PPA metrics in the 2D and 3D ICs.

flow	SRAM 2D	SRAM 3D	Δ	MRAM 2D	Δ	MRAM 3D	Δ
BEOL structure	6M	M6TM6B	-	8M	-	M6BM6T	-
clk. freq. (MHz)	519.00	596.40	14.9%	600.18	15.6%	699.32	34.7%
std. cell #	198587	199746	0.6%	201136	1.3%	202287	1.9%
footprint (mm^2)	1.20	0.60	-50.0%	0.94	-21.4%	0.49	-58.8%
silicon area (mm^2)	1.20	1.20	0.1%	0.94	-21.4%	0.99	-17.6%
wirelength (m)	6.28	5.47	-12.9%	6.27	-0.2%	4.92	-21.7%
F2F via #	0	4596	-	0	-	4708	-
worst slack (ns)	0.01	0.00	-	0.00	-	0.01	-
tot. power (mW)	123.1	140.4	14.1%	149.8	21.7%	163.9	33.1%
tot. cap (pF)	1288.7	1242.7	-3.6%	1341.1	4.1%	1183.5	-8.2%
pin cap (pF)	397.7	421.7	6.0%	428.6	7.8%	448.6	12.8%
wire cap (pF)	891.0	821.1	-7.9%	912.5	2.4%	734.8	-17.5%
edp (pJ)	237.2	235.4	-0.7%	249.6	5.2%	234.4	-1.2%

TABLE 3
Iso-performance comparisons of the PPA metrics in the 2D and 3D ICs.

flow	SRAM 2D	SRAM 3D	Δ	MRAM 2D	Δ	MRAM 3D	Δ
BEOL structure	6M	M6TM6B	-	8M	-	M6BM6T	-
clk. freq. (MHz)	500.00	500.00	0.0%	500.00	0.0%	500.00	0.0%
std. cell #	206534	205494	-0.5%	201096	-2.6%	199972	-3.2%
wirelength (m)	7.26	6.39	-12.0%	5.94	-18.1%	4.88	-32.8%
worst slack (ns)	0.06	0.19	-	0.21	-	0.29	-
tot. power (mW)	133.8	127.9	-4.4%	122.6	-8.4%	115.2	-13.9%
switching power (mW)	68.0	63.1	-7.2%	57.1	-15.9%	50.1	-26.3%
internal power (mW)	63.4	62.7	-1.2%	63.9	0.7%	63.4	-0.1%
leakage power (mW)	2.3	2.1	-8.6%	1.6	-32.2%	1.7	-27.8%
logic power (mW)	66.7	62.1	-6.9%	55.6	-16.6%	50.5	-24.3%
register power (mW)	30.6	29.8	-2.6%	28.8	-5.9%	29.0	-5.2%
clock power (mW)	9.9	9.8	-0.9%	9.5	-3.7%	9.5	-3.6%
macro power (mW)	26.0	25.7	-1.2%	28.0	7.7%	25.7	-1.2%
edp (pJ)	267.6	255.8	-4.4%	245.2	-8.4%	230.4	-13.9%

The small size of the STT-MRAM helps reduce the wirelength in 2D and 3D. The results show that the 3D IC with STT-MRAM has the smallest wirelength and wire capacitance. As a result, the switching power consumption caused by the parasitic wire capacitance is reduced by 20.58% and 26.29%, compared to SRAM-2D and SRAM-3D, respectively.

In addition, due to the timing benefits, as discussed in the previous section, the 3D IC with STT-MRAM can easily meet the timing constraints at 500MHz, without inserting too many buffers. According to the Table 3, the 3D IC with STT-MRAM has the largest worst slack (0.29 ns), which means it has a huge margin in timing to perform power optimization. Therefore, the 3D IC with STT-MRAM provides a 24.29% logic power reduction compared to the 2D IC with SRAM. Moreover, the leakage power reduction is also significant in the designs with STT-MRAM (larger than 25%), which mitigates the energy drawback of SRAM-based systems. In conclusion, the 3D integration of STT-MRAM also provides remarkable power benefits when power optimization is the first priority of the design objectives.

4 Summary

In this paper, we analyze the benefits of heterogeneous 3D ICs with STT-MRAM technology. Results show that the MRAM-based 3D ICs provide 17.55% silicon area saving, and either up to 34.74% performance gain or 13.90% power reduction, compared to the 2D baseline design using SRAM blocks only. It not only overcomes the drawbacks of 2D integration but also combines the benefits of STT-MRAM and 3D integration with no requirement of additional BEOL layers.

References

[1] J. Balkind et al. Openpiton: An open source manycore research framework. In *ACM SIGARCH Comput. Archit. News*, volume 44, pages 217–232. ACM, 2016.

[2] L. Bamberg et al. Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs. In *2020 Des., Auto. & Test in Euro. Conf. & Exhibition (DATE)*. IEEE, 2020.

[3] E. Beyne et al. Scalable, sub 2μm Pitch, Cu/SiCN to Cu/SiCN Hybrid Wafer-to-Wafer Bonding Technology. In *2017 IEEE Int. Electron Devices Meeting (IEDM)*, pages 32–4. IEEE, 2017.

[4] A. Calimera et al. Design Techniques and Architectures for Low-Leakage SRAMs. *IEEE Trans. Circuits Syst. I, Reg. Papers*, 59(9):1992–2007, 2012.

[5] M. Komalan et al. Feasibility exploration of NVM based I-cache through MSHR enhancements. In *2014 Des., Auto. & Test in Euro. Conf. & Exhibition (DATE)*, pages 1–6. IEEE, 2014.

[6] M. Komalan et al. Cross-Layer Design and Analysis of a Low Power, High Density STT-MRAM for Embedded Systems. In *2017 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, pages 1–4. IEEE, 2017.

[7] B. W. Ku et al. Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs. In *Proc. of the 2018 Int. Symp. on Phys. Des.*, pages 90–97. ACM, 2018.

[8] E. Kültürsay et al. Evaluating STT-RAM as an energy-efficient main memory alternative. In *2013 IEEE Int. Symp. on Perf. Analysis of Syst. and Software (ISPASS)*, pages 256–267. IEEE, 2013.

[9] A. Makosiej et al. CMOS SRAM Scaling Limits under Optimum Stability Constraints. In *2013 IEEE Int. Symp. Circuits and Syst. (ISCAS2013)*, pages 1460–1463. IEEE, 2013.

[10] S. P. Park et al. Future Cache Design using STT MRAMs for Improved Energy Efficiency: Devices, Circuits and Architecture. In *Proc. 49th Des. Auto. Conf.*, pages 492–497, 2012.

[11] F. Zaruba et al. The Cost of Application-Class Processing: Energy and Performance Analysis of a Linux-Ready 1.7-GHz 64-Bit RISC-V Core in 22-nm FDSOI Technology. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, 27(11):2629–2640, 2019.