A 1.15 TOPS/W, 16-Cores Parallel Ultra-Low Power Cluster with 2b-to-32b Fully Flexible Bit-Precision and Vector Lockstep Execution Mode

A 1.15 TOPS/W, 16-Cores Parallel Ultra-Low Power Cluster with 2b-to-32b Fully Flexible Bit-Precision and Vector Lockstep Execution Mode

Angelo Garofalo*, Gianmarco Ottavi*, Alfio di Mauro†, Francesco Conti*,
Giuseppe Tagliavini†, Luca Benini*‡, and Davide Rossi*
*Department of Electrical, Electronic and Information Engineering (DEI), University of Bologna, Italy
†Department of Computer Science and Engineering (DISI), University of Bologna, Italy
‡IIS Integrated Systems Laboratory, ETH Zurich, Switzerland

***Abstract*—IoT end-nodes require extreme performance and energy efficiency coupled with high flexibility to deal with the increasing computational requirements and variety of modern near-sensor data analytics applications. Low-Bitwidth and Mixed-Precision arithmetic is emerging as a trend to address the near-sensor analytics challenge in several fields such as linear algebra, Deep Neural Networks (DNN) inference, and on-line learning. We present Dustin, a fully programmable Multiple Instruction Multiple Data (MIMD) cluster integrating 16 RISC-V cores featuring 2b-to-32b bit-precision instruction set architecture (ISA) extensions enabling fine-grain tunable mixed-precision computation, improving performance and efficiency by 3.7 × and 1.9 × over state-of-the-art fully programmable devices. The cluster can be dynamically configured in Vector Lockstep Execution Mode (VLEM), turning off all IF stages except one, reducing power consumption by up to 38% with no performance degradation. The cluster, implemented in 65nm CMOS technology, achieves a peak performance of 58 GOPS and a peak efficiency of 1.15 TOPS/W.**

I. Introduction

Near-sensor data analytics applications increasingly require to run complex workloads, such as deep neural networks, on top of IoT end-nodes severely constrained in terms of power envelope, memory, and cost (i.e., silicon area and technology). An emerging trend to approach the complexity of this problem is to employ the simplest data representation usable for each given sub-task of a workload, using *Low-Bitwidth Mixed-Precision* arithmetic. This approach is well-established in the floating-point computation, where transprecision techniques have been demonstrated in domains such as traditional near-sensor data analytics [1] and training of neural networks [2]. In the integer domain, emerging fixed-point transprecision and mixed-precision techniques can be pushed down even more significantly to extreme low-bitwidth for applications based on linear algebra [3] and inference of deep neural networks [4]. Up to now, extreme Low-Bitwidth Mixed-Precision arithmetic has been mainly applied in specialized accelerators [2], [4] – its application to fully programmable architectures is challenged by the saturation of encoding space and the related complexity of instruction fetch and decode. Moreover, mixed-precision operations require data casting as well as packing/unpacking operations when the format is updated on the fly, triggering overheads that significantly reduce the effectiveness of this approach [5].

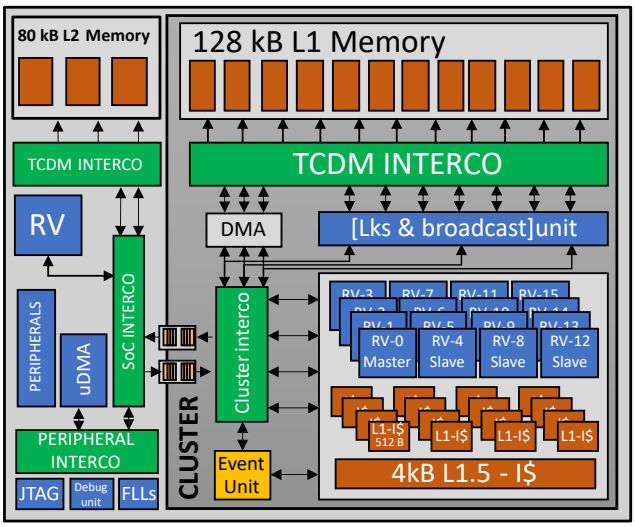


Fig. 1. Overview of the Dustin SoC Architecture.

To deal with this challenge, we present DUSTIN: a low-power IoT end-node with an accelerator cluster of 16 RISC-V cores. Each DUSTIN core is augmented with mixed 2b-to-32b single instruction multiple data (SIMD) instructions, where the format of the input operands is customized on-line through a dedicated control register.

A common trait of machine learning and data-analytics algorithms is data-parallelism. To boost energy efficiency on data-parallel code, the cluster can be dynamically configured into a fine-grain *Vector Lockstep Execution Mode* (VLEM), turning off all instruction fetch stages except one. This reduces power consumption by up to 38% with no performance degradation on critical data-parallel kernels while keeping multiple instruction multiple data (MIMD) flexibility for general-purpose code.

Implemented in robust and cost-effective 65 nm CMOS technology, DUSTIN achieves 15 GOPS and 303 GOPS/W on 8-bit integer arithmetic, similar to SoA fully programmable systems implemented in much more scaled technology nodes

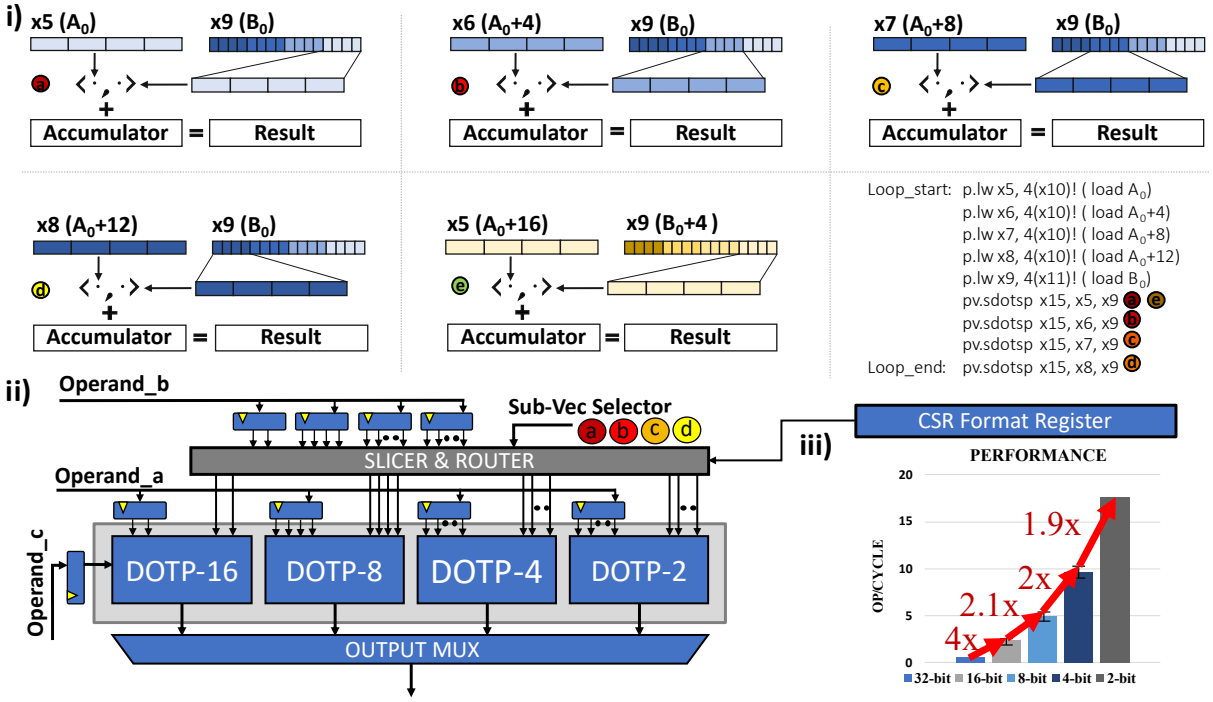


Fig. 2. i) Mixed-Precision Dot Product 8x2; ii) Dot product functional Units; iii) Performance spanning through bit-widths.

(40 nm and 22 nm) – with a further boost in performance (3.7×) and efficiency (1.9×) on Low-Bitwidth Mixed-Precision workloads, up to 58 GOPS and 1.15 TOPS/W.

II. SoC Architecture

Fig. 1 shows the architecture of DUSTIN. It is built around a tightly-coupled cluster of 16 32-bit RISC-V cores sharing a 128 kB, 32-banks Tightly-Coupled Data Memory (TCDM) through a single-cycle latency logarithmic interconnect (LIC) leveraging a request/grant protocol. The LIC implements a word-level interleaving scheme to reduce banking conflict probability (typically 5% even for highly memory-intensive applications). The cores share a 2-level latch-based instruction cache: the first level (512 B) is private, the second level (L1.5) is a 4 kB 8-banks shared cache connected to the L1s with an interconnect similar to the LIC. The L1.5 refills from a larger 80 kB L2 memory hosting resident code. A dedicated hardware block (Event Unit) assists the cores to accelerate parallel computation patterns, such as thread dispatching and barriers. Finally, the SoC includes a controlling RISC-V core, a set of standard peripherals, and 3 FLLs for frequency control.

A. Bit-Scalable Precision Processor

The proposed processor extends RI5CY, a 32-bit 4-pipeline stages core featuring DSP extensions such as 16-bit and 8-bit SIMD dot product fully supported by a GCC 7.1.1 toolchain [6]. The key efficiency-boosting enhancement is a new mixed-precision SIMD dot product execution unit, shown in Fig. 2. It includes 4 multiplexed sub-units implementing 16b down to 2b dot products (DOTP). To enable any SIMD mixed-precision computation, a slicer-and-router unit selects the correct bits in the source registers and forwards them to the DOTP unit featuring the higher precision between the two operands after optional bit manipulation. A dedicated circuit gates the clock of the input registers of the unused SIMD units. With no timing overhead and an increase in area smaller than 10% with respect to RI5CY, the proposed power-aware design allows the extended core to run in the same power envelope as the original one, safeguarding its general-purpose computing efficiency. To encode the new mixed-precision SIMD instructions, we define a *virtual instruction*: the opcode (e.g., dotp) is decoded in the ID stage, the precision of its operands (e.g., 4x8) is specified by a control and status register (CSR), written by the processor before issuing a portion of code containing virtual SIMD instructions. This approach is essential to address the saturation problem of the RISC-V encoding space, as it avoids to explicitly encode all the 500 combinations of mixed-precision operands.

B. Vector Lockstep Execution Mode

The second key efficiency enhancement is at the cluster level: we support a new Vector Lockstep Execution Mode (VLEM), where all cores execute the same instructions cycle-by-cycle. In VLEM, only the master core's L1 cache and IF stage are active, forwarding instructions to the ID stages of all cores (Fig. 3). The related activity reduction by clock gating saves up to 38% total power. To enter in VLEM, all cores have to i) synchronize on a barrier, ii) write to a memory-mapped register. Banking conflicts on TCDM are solved by delaying the grant signal assertion for the time required to serve all requests. To avoid systematic conflicts (e.g., when all cores access the same address in memory – a

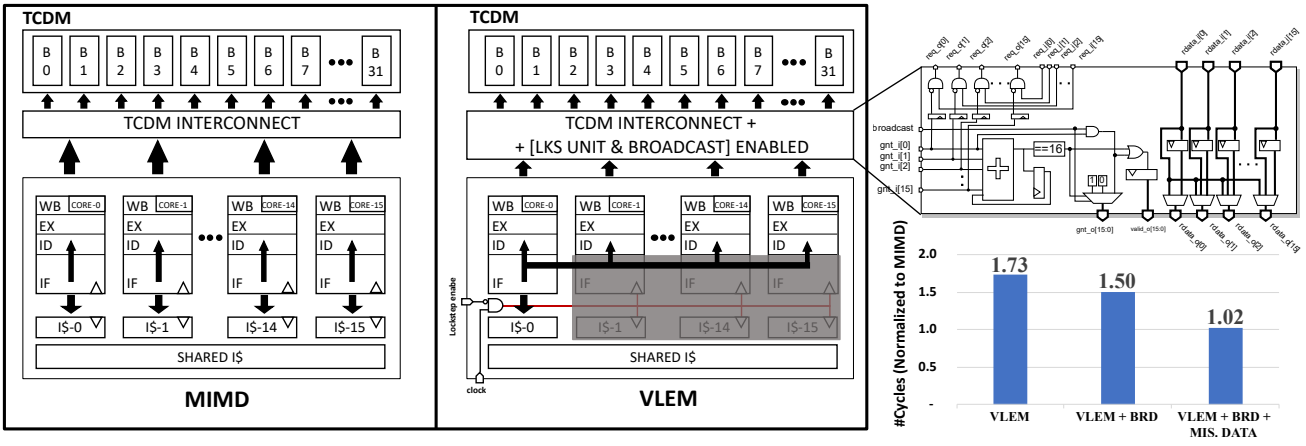
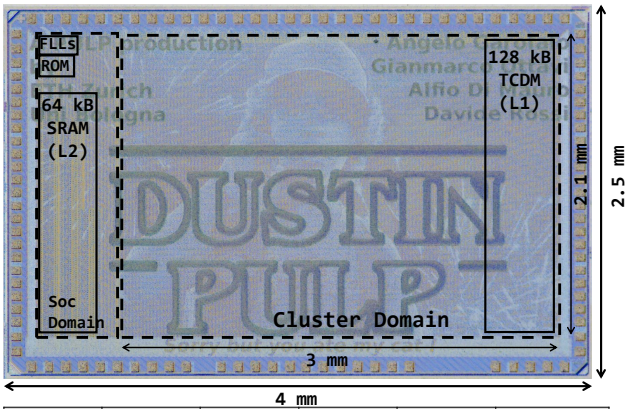


Fig. 3. Overview of the cluster architecture to operate in VLE mode and comparison with the classic MIMD mode. The chart (bottom right) shows the optimizations to reduce the VLE execution overhead: First we introduce the broadcasting feature (+ BRD), then we operate the misalignment of the data (+ MIS. DATA).



Technology	Chip Area	Total SRAM	VDD range	Freq. Range	Power Env.
CMOS 65nm	10mm²	208 kB	0.8 – 1.2V	60–205 MHz	156 mW
Int. Perf. (8b)	**Int. Eff. (8b)**	**Int. Perf. (4b)**	**Int. Eff. (4b)**	**Int. Perf. (2b)**	**Int. Eff. (2b)**
15 GOPS	303 GOPS/W	30 GOPS	562 GOPS/W	58 GOPS	1.15 TOPS/W

Fig. 4. Chip micrograph and specifications.

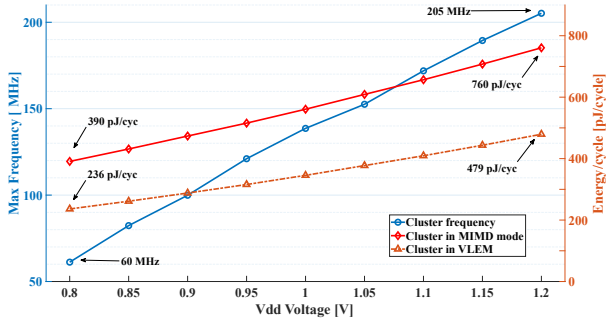


Fig. 5. Voltage Sweep vs. Max Freq. vs. Energy/Cycle.

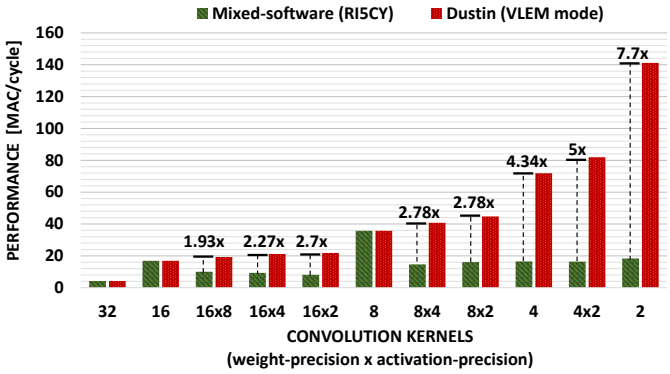


Fig. 6. The chart compares the execution of mixed-precision convolution kernels running on the baseline 16 cores cluster with the RI5CY core (software mixed-precision kernels) and on Dustin's cluster in VLEM mode (featuring the Mixed-precision ISA extensions).

common pattern in linear algebra kernels), the VLEM unit is enhanced with a broadcast control, activated when all cores access the same memory location. Together with proper data organization, broadcast can entirely eliminate the overheads introduced by banking conflicts, as shown in Fig. 3, and can reduce the number of memory accesses up to 66%. After the execution of a kernel in lockstep, the cores exit VLEM by writing into a memory-mapped register. The increase in area of the slave cores (gating and isolation) is negligible (<3%) compared to the baseline as well as the design cost of the entire lockstep unit, which impacts for less than 1% on the total cluster area.

III. Measurements

Figure 4 shows a die photograph of DUSTIN, together with its main features. The SoC is implemented in 65 nm CMOS technology with a die size of 10 mm^2. Figure 5 reports the maximum operating frequency and the energy per cycle of the cluster over the 0.8V to 1.2V voltage range. The measurements are carried out on the silicon prototype, running a typical high-utilization deep neural network workload, the matrix-multiplication (matmul), with 8-bit precision operands. Linearly increasing with the voltage, we can reach the highest operating frequency of 205 MHz at 1.2V.

Figure 6 shows the performance of heavily quantized and mixed-precision convolutional kernels on the proposed cluster. On kernels where the activations are the only sub-byte precision operands, the performance benefits of the mixed-precision hardware extension are marginal due to the unpacking of data

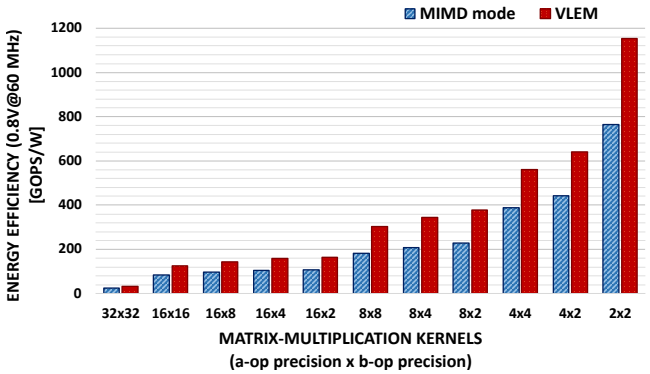


Fig. 7. Comparison in terms of Energy Efficiency of Dustin configured in MIMD and VLE mode, running Mixed-precision Matrix Multiplication kernels.

executed in a less arithmetic intensive portion of the kernel. In all other configurations, the mixed-precision instruction set extensions provide a significant advantage ranging from 2× to 7.7× improvements with respect to a baseline cluster.

To highlight the energy savings of the VLEM mode on regular computing kernels, we measure energy consumption with the cluster running the matrix-multiplication in two modes: the classic MIMD mode and the VLEM mode, enabled via software. Fig. 6 shows the related efficiency. The execution of linear kernels in VLEM mode achieves 1.5× better energy efficiency and no performance overhead with respect to the default MIMD execution.

Figure 8 shows a comparison with the SoA. Compared to similar fully programmable IoT end-nodes [7], [8], [9], [10], the proposed SoC delivers similar performance and energy efficiency on 8-bit format, despite the less scaled technology node used for implementation. This is achieved thanks to the larger parallelism of the cluster and the VLEM mode saving up to 38% of overall power consumption. The proposed work is the only one featuring support for fully flexible bit-scalable precision from 2b to 32b, improving performance and efficiency by 3.7x and 1.9x over the state-of-the-art (SoA) for heavily quantized and mixed-precision workloads, delivering a peak performance of 58 GOPS and a peak efficiency of 1.15 TOPS/W.

IV. Conclusion

We presented DUSTIN, a fully programmable Multiple Instruction Multiple Data (MIMD) cluster integrating 16 RISC-V cores featuring 2b-to-32b bit-precision instruction set architecture (ISA) extensions enabling fine-grain tunable mixed-precision computation. The cluster can be dynamically configured into a Vector Lockstep Execution Mode (VLEM), turning off all instruction fetch stages and L1 I$ except one, thereby reducing power consumption by up to 38% with no performance degradation. The cluster, implemented in 65nm CMOS technology, achieves a peak performance of 58 GOPS and a peak efficiency of 1.15 TOPS/W – competitive with IoT end-nodes using much more scaled and expensive technology nodes.

	SleepRunner [7]	SamurAI [8]	Mr.Wolf [9]	VEGA [10]	Dustin (this work)
Technology	CMOS 28nm FD-SOI	CMOS 28nm FD-SOI	CMOS 40nm LP	CMOS 22nm FD-SOI	CMOS 65nm
Die Area	0.68 mm²	4.5 mm²	10 mm²	12 mm²	10 mm²
Applications	IoT GP	IoT GP + DNN	IoT GP + DNN	IoT GP + DNN	IoT GP + DNN + QNNs
CPU/ISA	CM0DS Thumb-2 subset	1x RI5CY RVC32IMFXpulp	9 x RI5CY RVC32IMFXpulp	9 x RI5CY RVC32IMFXpulp	16 x MPIC CORES (RISC-V)
Int Precision (bits)	32	8, 16, 32	8, 16, 32	8, 16, 32	2, 4, 8, 16, 32 (plus Mixed-Precision)
Supply Voltage	0.4 - 0.8 V	0.45 - 0.9 V	0.8 - 1.1 V	0.5 - 0.8 V	0.8 - 1.2 V
Max Frequency	80 MHz	350 MHz	450 MHz	450 MHz	205 MHz
Power Envelope	320 µW	96 mW	153 mW	49.4 mW	156 mW
[1]Best Integer Performance	31 MOPS (32b)	1.5 GOPS (8b)[2]	12.1 GOPS (8b)	15.6 GOPS (8b)	**15 GOPS (8b) 30 GOPS (4b) 58 GOPS (2b)**
[1]Best Integer Efficiency	97 MOPS/mW @ 18.6 MOPS (32b)	230 GOPS/W @110 MOPS (8b)[2]	190 GOPS/W @ 3.8 GOPS (8b)	614 GOPS/W @ 7.6 GOPS (8b)	**303 GOPS/W @4.4 GOPS (8b) 570 GOPS/W @8.8 GOPS (4b) 1152 GOPS/W @17.3 GOPS (2b)**

[1] OPs = 1 8-bit (or 4-bit or 2-bit) MAC on MatMul benchmark unless differently specified.
[2] Execution on SW programmable Core.

Fig. 8. Comparison with SoA solutions.

V. Acknowledgments

This work was supported in part by the EU Horizon 2020 projects WiPLASH (g.a. 863337) and AI4DI (g.a. 826060).

References

[1] G. Tagliavini, S. Mach, D. Rossi, A. Marongiu, and L. Benini, "A transprecision floating-point platform for ultra-low power computing," in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2018, pp. 1051–1056.

[2] S. Kang, D. Han, J. Lee, D. Im, S. Kim, S. Kim, and H. Yoo, "7.4 GANPU: A 135TFLOPS/W Multi-DNN Training Processor for GANs with Speculative Dual-Sparsity Exploitation," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 140–142.

[3] A. Stojanov, T. M. Smith, D. Alistarh, and M. Püschel, "Fast Quantized Arithmetic on x86: Trading Compute for Data Movement," in *2018 IEEE International Workshop on Signal Processing Systems (SiPS)*, 2018, pp. 349–354.

[4] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H. Yoo, "UNPU: A 50.6TOPS/W unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 218–220.

[5] N. Bruschi, A. Garofalo, F. Conti, G. Tagliavini, and D. Rossi, "Enabling mixed-precision quantized neural networks in extreme-edge devices," in *Proceedings of the 17th ACM International Conference on Computing Frontiers*, 2020, pp. 217–220.

[6] M. Gautschi, P. D. Schiavone, A. Traber, I. Loi, A. Pullini, D. Rossi, E. Flamand, F. K. Gürkaynak, and L. Benini, "Near-Threshold RISC-V Core With DSP Extensions for Scalable IoT Endpoint Devices," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 10, pp. 2700–2713, 2017.

[7] D. Bol, M. Schramme, L. Moreau, P. Xu, R. Dekimpe, R. Saeidi, T. Haine, C. Frenkel, and D. Flandre, "SleepRunner: A 28-nm FDSOI ULP Cortex-M0 MCU With ULL SRAM and UFBR PVT Compensation for 2.6-3.6-μW/DMIPS 40-80-MHz Active Mode and 131-nW/kB Fully Retentive Deep-Sleep Mode," *IEEE Journal of Solid-State Circuits*, pp. 1–1, 2021.

[8] I. Miro-Panades, B. Tain, J. F. Christmann, D. Coriat, R. Lemaire, C. Jany, B. Martineau, F. Chaix, A. Quelen, E. Pluchart, J. P. Noel, R. Boumchedda, A. Makosiej, M. Montoya, S. Bacles-Min, D. Briand, J. M. Philippe, A. Valentian, F. Heitzmann, E. Beigne, and F. Clermidy, "SamurAI: A 1.7MOPS-36GOPS Adaptive Versatile IoT Node with 15,000× Peak-to-Idle Power Reduction, 207ns Wake-Up Time and 1.3TOPS/W ML Efficiency," in *2020 IEEE Symposium on VLSI Circuits*, 2020, pp. 1–2.

[9] A. Pullini, D. Rossi, I. Loi, G. Tagliavini, and L. Benini, "Mr.Wolf: An Energy-Precision Scalable Parallel Ultra Low Power SoC for IoT Edge Processing," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 7, pp. 1970–1981, 2019.

[10] D. Rossi, F. Conti, M. Eggiman, S. Mach, A. D. Mauro, M. Guermandi, G. Tagliavini, A. Pullini, I. Loi, J. Chen, E. Flamand, and L. Benini, "4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7μW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode," in *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 64, 2021, pp. 60–62.