Mr.Wolf: A 1 GFLOP/s Energy-Proportional Parallel Ultra Low Power SoC for IoT Edge Processing

Antonio Pullini*, Davide Rossi†, Igor Loi†, Alfio Di Mauro*, and Luca Benini*†
*Integrated Systems Laboratory, ETH Zürich, Gloriastr. 35, 8092 Zurich, Switzerland
†DEI, University of Bologna, Via Risorgimento 2, 40136 Bologna, Italy

***Abstract*—We present Mr.Wolf, a Parallel Ultra Low Power (PULP) SoC featuring a hierarchical architecture with a small (12KG) microcontroller class RISC-V core augmented with an autonomous IO subsystem for efficient data transfer from a wide set of peripherals. The small core can offload compute-intensive kernels to an 8-cores floating-point capable processing engine available on demand. The proposed SoC, implemented in a 40 nm LP CMOS technology, features a 108 μW fully retentive memory (512 kB). The IO subsystem is capable of transferring up to 1.6Gbit/s in less than 2.5mW. The 8-core compute cluster achieves a peak performance of 850 millions of 32-bit integer multiply and accumulate per second (MMAC/s), 500 millions of 32-bit floating-point multiply and accumulate per second (MFMAC/s) - 1 GFLOP/s - with an energy-efficiency up to 15 MMAC/s/mW and 9 MFMAC/s/mW. These building blocks are supported by aggressive on-chip power conversion and management, enabling energy-proportional heterogeneous computing for always-ON IoT end-nodes improving performance by several orders of magnitude with respect to traditional single core MCUs within a power envelope of 153 mW.**

I. Introduction

The majority of current ultra low power smart sensing edge devices operating for years on small batteries are limited to low-bandwidth sensors, such as temperature or pressure. While the bandwidth generated by these sensors allows to transmit raw data to the cloud where the analysis is performed, a new generation of edge application is emerging, probing the environment from data-rich sensors such as audio, video or bio-metrics, preventing data to be transmitted wirelessly for energy, aggregate sensor bandwidth, and security reasons [1]. A possible solution to solve this issue is to bring part of the data analytics close to the sensor, reducing the high-bandwidth raw data to highly compressed and informative data such as tags, classes or even simple events. However, this approach poses an extreme challenge of squeezing the computational requirements of advanced near-sensor data analysis algorithms within the mW-range power envelope of always-ON battery-powered IoT end-nodes [2].

Always-on IoT applications rely on ultra-low power MCUs featuring modest compute capability cores such as Cortex M0+ while most of this new generation of applications requires much more computational power (up to few Giga Operations Per Second - GOPS) and significant memory footprint (up to few Mbytes) in the same power envelope up to few tens of milliwatts, coupled with state retentive deep sleep modes to deal with the heavily duty-cycled behavior of several IoT applications. Moreover, the IO subsystem of current ultra-low power MCUs, relying on system DMAs placed on the same bus as the single-ported main memory and the main processor, is not suited to efficiently gather data from peripherals such as imagers and microphone arrays, or to connect to external memories such as Cypress Semiconductor's HyperRAM.

To address this challenge, we propose Mr.Wolf: a multi-GOPS fully programmable power/performance-tunable IoT-edge computing engine fabricated in 40nm LP CMOS technology. Mr.Wolf uses the flexible attributes of the RISC-V ISA to deliver a state of the art micro controller called fabric controller (FC), ease of programming, rich set of peripherals, coupled with a powerful programmable parallel processing engine for flexible multi-sensor (image, audio, biometric, inertial) data analysis and fusion. The SoC is built around an ultra-low power MCU based on a 2-pipeline stages processor optimized for low power featuring a programmable 64-to-108 μW state-retentive sleep power (for up to 512kB of system memory) and an I/O subsystem optimized for efficient and autonomous data transfers from high-bandwidth peripherals (up to 1.6 Gbit/s aggregated bandwidth in 2.5mW). The cluster is composed of 8 fully programmable processors featuring DSP extensions targeting energy-efficient digital signal processing, delivering up to 800 MMAC/s and up to 15 MMAC/s/mW, sharing 2 energy-efficient floating-point unit delivering up to 500 MFMAC/s (i.e. 1 GFLOP/s) and up to 9 MFMAC/s/mW.

II. SoC Architecture

Figure 1 provides a top-level view of the Mr.Wolf architecture. It includes two domains isolated by level-shifters and dual clock FIFOs to operate into independent voltage and frequency islands: the SoC and Cluster domains, described in the following sections.

A. SoC Domain

The SoC Domain is built around an advanced MCU controlled by a tiny (12 Kgates), 2-pipeline stages RISC-V processor optimized for power consumption [3]. The SoC includes a full set of peripherals: Quad SPI (400 Mbit/s), I2C, 4 I2S (4 x 3 Mbit/s), a parallel camera interface (400 Mbit/s) and UART, enabling parallel capture of images, sounds and vibrations, a 4 channels PWM interface, GPIOs, and a JTAG interface for debug purposes. On-chip memory is extended via a DDR HyperBus interface (800 Mbit/s). Data transfers

978-1-5386-5404-0/18/$31.00 ©2018 IEEE

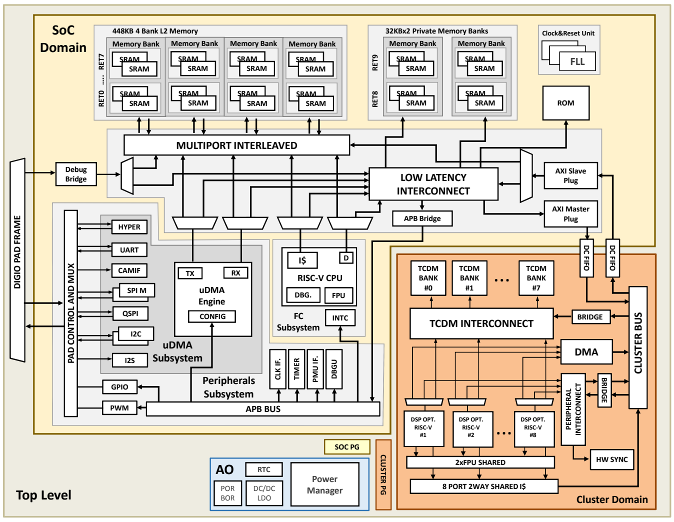


Fig. 1: Mr.Wolf Architecture

Power Mode	VDD [V]	Frequency	Power
Deep Sleep*	0.8	n.a.	72 μW
Ret. Deep Sleep*	0.8	n.a.	76.5 - 108 μW
SoC Idle	0.8 - 1.1	SoC clock gated	0.55 - 1.96 mW
SoC Active	0.8 - 1.1	32 kHz - 450 MHz	0.97 - 38 mW
Cluster Idle†	0.8 - 1.1	Cluster clock gated	1.2 - 4.6 mW
Cluster Active‡	0.8 - 1.1	32 kHz - 350 MHz	1.6 - 153 mW

*From VBAT; †SoC must be active or idle; ‡SoC must be active;

TABLE I: Mr.Wolf Power Modes.

from/to the peripherals are managed by a multi-channel I/O DMA (μDMA) to minimize the amount of interactions and the workload of the controlling core when performing IO. As opposed to traditional MCUs, where the DMA is a master on an AMBA AHB bus shared with the processor and the main memory, forming a huge performance bottleneck for IO transfers, Mr.Wolf's μDMA has dedicated connections to all IO peripherals of the SoC and a 2 dedicated 32-bit ports on the L2 memory interconnect, granting an aggregated bandwidth equal to Freq*32-bit*2 sufficient to satisfy the requirements of parallel transfers from all the peripherals (up to 1.6 Gbit/s) with a frequency of just 57 MHz, and a power of 2 mW. This architecture maximizes the transfer efficiency while avoiding the need for large buffers attached to the peripherals (16 bytes/channel are employed in Mr.Wolf), which are required to absorb possible overflows in traditional systems and to reduce the control overhead of the processors.

512 kB of L2 memory are available on the SoC, together with a ROM storing the primary boot-code. The L2 memory layout of Mr.Wolf is arranged as 4 112 Kbytes word-level interleaved banks (448 Kbytes overall) to minimize conflicts during parallel accesses through the master ports of the L2 memory interconnect (i.e. the μDMA, the processor and the Cluster domain), plus 2 banks of 32 Kbytes that can be used privately by the FC (e.g. program, stack, private data) without incurring banking conflicts. From a performance viewpoint, this memory organization allows transparent sharing of the L2, increasing by 4x the system memory bandwidth (i.e. Freq*32-bit*4) to satisfy the significant requirements of all the master resources (i.e. FC, Cluster's processors, Cluster's DMA and μDMA). The SoC has an APB subsystem including pad GPIO and multiplexing control, clock and power control, timer, *μDMA* configuration port and PWM controller. Finally, the connection with the accelerator is done thought 2 asymmetric AXI plugs with a 64-bit width for accelerator to memory communication and 32-bit for memory to accelerator communication. Mr.Wolf's memory hierarchy is organized as a single name space: every single core in the chip can access all memory locations.

B. Cluster Domain

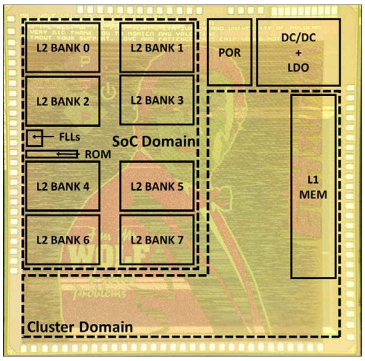
The cluster, residing on a dedicated voltage and frequency domain, is turned on and adjusted to the right voltage and frequency when applications running on the FC offload highly intensive computation kernels. It contains 8 RISC-V cores supporting the RVC32IM instruction set, plus extensions targeting energy-efficient digital signal processing such as hardware loops, load/store with pre/post increment, multiply and accumulate (MAC) vectorial instructions (RVC32IMFX)[3]. Two floating-point units (FPU) are shared among the 8 processors of the cluster, implementing common floating point operations including FMAC, a key operation for near sensor tasks such as filtering and neural networks. This shared FPU approach reduces by 4x the area overhead for latency bound and infrequent FP operations, while leading to a maximum performance degradation up to 10%.

The cluster is served by a multi-banked L1 memory, enabling shared-memory parallel programming models such as OpenMP. The L1 memory can serve all memory requests in parallel with single-cycle access latency and low average contention rate (<10% even on the most data intensive kernels). The cluster program cache is also shared to maximize efficiency in fetching data-parallel code. Fast event management, parallel thread dispatching, and synchronization are supported by a dedicated hardware block (HW Sync), enabling very fine-grained parallelism and hence high energy efficiency in parallel workloads. The HW Sync block also controls the top-level clock gating of every single core in the cluster. A core waiting for an event (attached to a synchronization barrier or general event) is instantly brought into a fully clock gated state, zeroing its dynamic power consumption and resumes the execution after the event in 2 clock cycles.

III. Power Management

To maximize power efficiency and to minimize the number of external components, the SoC contains an internal DC/DC converter that can be directly connected to an external battery[1]. It can deliver voltages in the range of 0.8V to 1.1V when the circuit is active with an efficiency of 70% for very low loads and up to 95% for medium and high loads. When the circuit is in sleep mode this regulator is turned off and a low-dropout (LDO) regulator is used to power the real-time clock fed by a 32kHz crystal oscillator, which controls programmed

[1] Provided by Dolphin Integration https://www.dolphin-integration.com.



Technology	CMOS 40nm LP
Chip area	10 mm^2
VDD range	0.8V - 1.1V
Memory Transistors	576 Kbytes
Logic Transistors	1.8 Mgates
Frequency Range	32 kHz – 450 MHz
Power Range	72 µW – 153 mW

Fig. 2: Chip Micrograph and Main Features.

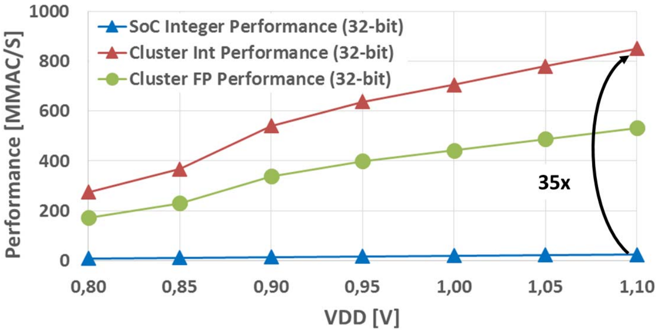


Fig. 3: Mr.Wolf Performance when executing an integer matrix multiplication on the SoC CPU and on the Cluster, and a floating-point matrix multiplication on the Cluster.

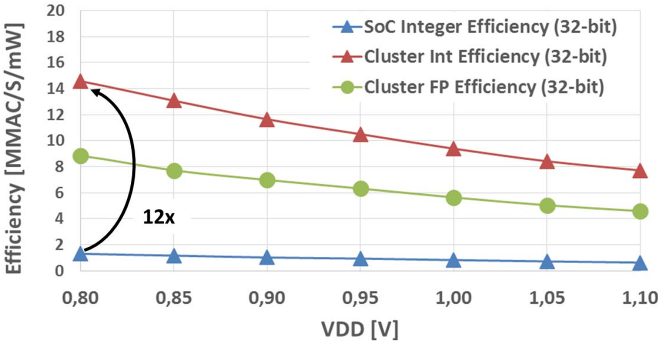


Fig. 4: Mr.Wolf Energy Efficiency when executing an integer matrix multiplication on the SoC CPU and on the Cluster, and a floating-point matrix multiplication on the Cluster.

wake-up and, optionally, part of the L2 memory allowing retention of application state for fast wake-up. 8 physical memory banks can be independently power gated, allowing to implement an incremental state-retentive mechanism for the L2 memory, where the leakage power of the memory arrays is paid only for the memory cuts that are actually being used by the applications. When in deep sleep the current consumption is reduced to 72 μW (from VBAT) assuming the RTC is active and no data retention, and 108 μW assuming full L2 retention. The two main domains have their own separate clocks, generated by two frequency-locked loops placed on the SoC domain. Special attention has been paid to the time needed to turn on and turn off the cluster. The typical turn-around time from FC idle to cluster active is between 300 μs allowing for agile power state transitions. Table I shows the power modes of Mr.Wolf, together with maximum frequency and power consumption.

IV. Measurements

Figure 2 shows a die photograph together with its main features. The SoC was implemented in 40nm CMOS LP technology, the die size is 10 mm^2, integrating 1.8 million of equivalent logic gates and 576 Kbytes of memory.

Figure 3 shows the SoC performance measured on the silicon prototype running a typical high-utilization workload (matrix multiplication), while Figure 4 shows the related energy efficiency. The first two curves (blue and red) show the FC and cluster performance when executing an integer matrix multiplication. It is possible to note that similarly to other low power MCUs [4][5], based on tiny processors optimized for low-power control tasks, the FC can achieve a peak performance of 25 MMAC/s at 450 MHz, 1.1V, and a peak efficiency of 1.5 MMAC/s/mW @ 150 MHz, 0.8V. The differentiating factor and the power of Mr.Wolf stands on the possibility to power on the parallel processing cluster and offload compute intensive tasks with significant performance and efficiency. Thanks to the instruction set extensions (i.e. hardware loops, load and store with automatic index increment and MAC unit), the optimized pipeline of the 8 RI5CY processors, and to the efficient memory sharing through the L1 memory, the cluster can execute 2.5 MAC/cycle on 8 cores. This execution efficiency leads to the peak performance of 850 MMAC/s at 350 MHz, 1.1V and a peak energy efficiency of 15 MMAC/s/mW at 110 MHz, 0.8V, improving DSP performance and energy efficiency of the FC by 35x and 12x, respectively. Both performance and energy efficiency increase almost linearly when exploiting the single instruction multiple data (SIMD) capability of the processors, reaching 1700 MMAC/s and 30 MMAC/s/mW for a 16-bit matrix multiplication, and 3400 MMAC/s and 60 MMAC/s/mW for a 8-bit matrix multiplication. The third curve (green) shows the performance and efficiency of the cluster when executing a FP matrix multiplication, expressed as MFMAC/s and MFMAC/s/mW, respectively. It is interesting to note that even if the 2 FP units are shared, and despite the 2 pipeline stages of the FMAC units (required to reach the target frequency) the architectural efficiency is 1.57 FMAC/cycle, leading to a peak performance of 500 MFMAC/s - 1 GFLOP/s - and a peak energy efficiency of 9 MFMAC/s/mW.

Figure 5 shows in the lower 2 curves the measurements of power consumption of the SoC subsystem for different I/O input bandwidths with and without 50 millions of operations per second (MOP/s) of load on the FC. The system can sustain 800 MBit/s operating at as low as 28.5 MHz and consuming 1.21 mW when only I/O to memory transfer is involved and work at 62.5 MHz and consume 2.7 mW when executing a kernel on the CPU at 50 MOP/s in parallel to the I/O transfer. The other curves in the graph show an estimate of the consumption of more traditional system with CPU, memory and DMA sharing the same system bus. When DMA is used the performances

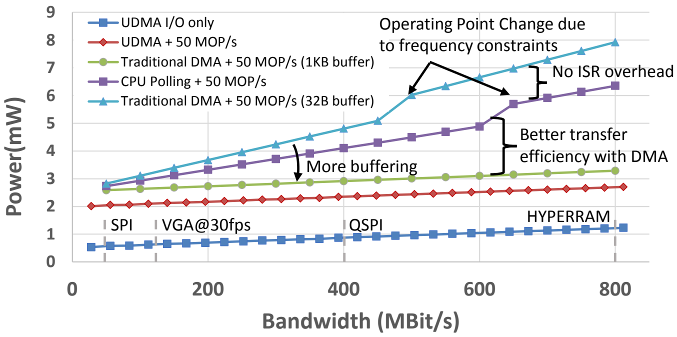
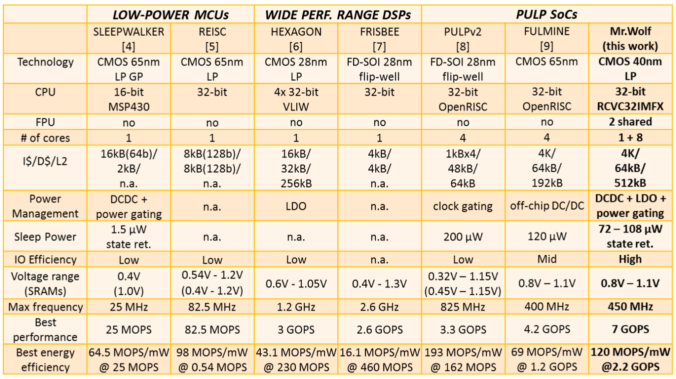


Fig. 5: Mr.Wolf IO Subsystem Efficiency.

are deeply affected by the buffer size. The two extremes are shown in the graph, a small buffer of 32B, and an hypothetical system with 1KB buffer dedicated to each single peripheral. The proposed solution shows an improvement of up to 4.5x in power consumption when operating at high bandwidth (800 MBit/s) compared to a traditional DMA solution with a small 32 B/peripheral buffer. Moreover, even when very big buffers are employed in traditional architectures, the μDMA is 1.2x more efficient, thanks to the higher frequency that traditional solutions are required to use to compensate for contention with the processor on the system memory, thanks to the smaller buffers enabled by the tightly coupled integration with the L2 memory, and by the fact that in the traditional systems all the interconnect is active during I/O transfers. In the figure CPU polling is presented as a reference and to highlight that a DMA based solution the buffering resources at the peripheral has to be big enough to hide the overhead of interrupt service routine and DMA programming. In a traditional system the buffering has to be allocated at design time and cannot be dynamically allocated as in our solution.

Figure 6 shows a comparison with recent energy-efficient digital signal processors (DSPs)[4][5][6][7][8][9]. Exploiting the heterogeneous architecture which couples the IO efficiency and state-retentive sleepwalking capabilities of the SoC domain with the powerful and energy-efficient 8-processors cluster, Mr.Wolf represent a significant advance in the state of the art of IoT processors for processing of high-bandwidth near-sensor data streams. Mr.Wolf overcomes the performance of all existing energy-efficient processors (by more than 1.6x) with significant energy efficiency (120 MOPS/mW) surpassed only by PULPv2 due to a better technology used for implementation (28nm FD-SOI vs. CMOS 40nm LP), but lacking internal power management circuits (i.e. DC/DC, LDO, power gating). Although the tiny microcontrollers such as Sleepwalker [4] feature a better deep sleep power, Mr.Wolf has the capability to store in a full retentive way up to 512 Kbytes of data (instead of few Kbytes). The hierarchical and energy-proportional architecture of Mr.Wolf allows to periodically wake-up the SoC only to efficiently transfer sensor data to L2 with the μDMA, accumulate data on the state-retentive L2 memory (enabling retention only on used banks to minimize sleep power), and activating the cluster when enough data has been acquired for energy-efficient (floating point) digital signal processing, paving the way for always-on data analytics of high-bandwidth sensors data at the edge of the Internet of Things.



	LOW-POWER MCUs		***WIDE PERF. RANGE DSPs***		***PULP SoCs***		
	SLEEPWALKER [4]	REISC [5]	HEXAGON [6]	FRISBEE [7]	PULPv2 [8]	FULMINE [9]	**Mr.Wolf (this work)**
Technology	CMOS 65nm LP GP	CMOS 65nm LP	CMOS 28nm LP	FD-SOI 28nm flip-well	FD-SOI 28nm flip-well	CMOS 65nm	**CMOS 40nm LP**
CPU	16-bit MSP430	32-bit	4x 32-bit VLIW	32-bit	32-bit OpenRISC	32-bit OpenRISC	**32-bit RCVC32IMFX**
FPU	no	no	no	no	no	no	**2 shared**
# of cores	1	1	1	1	4	4	**1 + 8**
I$/D$/L2	16kB(64b)/ 2kB/ n.a.	8kB(128b)/ 8kB(128b)/ n.a.	16kB/ 32kB/ 256kB	4kB/ 4kB/ n.a.	1kBx4/ 48kB/ 64kB	4K/ 64kB/ 192kB	**4K/ 64kB/ 512kB**
Power Management	DCDC + power gating	n.a.	LDO	n.a.	clock gating	off-chip DC/DC	**DCDC + LDO + power gating**
Sleep Power	1.5 µW state ret.	n.a.	n.a.	n.a.	200 µW	120 µW	**72 – 108 µW state ret.**
IO Efficiency	Low	Low	Low	n.a.	Low	Mid	**High**
Voltage range (SRAMs)	0.4V (1.0V)	0.54V - 1.2V (0.4V - 1.2V)	0.6V - 1.05V	0.4V - 1.3V	0.32V – 1.15V (0.45V – 1.15V)	0.8V – 1.1V	**0.8V – 1.1V**
Max frequency	25 MHz	82.5 MHz	1.2 GHz	2.6 GHz	825 MHz	400 MHz	**450 MHz**
Best performance	25 MOPS	82.5 MOPS	3 GOPS	2.6 GOPS	3.3 GOPS	4.2 GOPS	**7 GOPS**
Best energy efficiency	64.5 MOPS/mW @ 25 MOPS	98 MOPS/mW @ 0.54 MOPS	43.1 MOPS/mW @ 230 MOPS	16.1 MOPS/mW @ 460 MOPS	193 MOPS/mW @ 162 MOPS	69 MOPS/mW @ 1.2 GOPS	**120 MOPS/mW @2.2 GOPS**

Fig. 6: Comparison with state of the art efficient processors. Performance are normalized to RVC32IM operations.

V. Conclusion

We presented Mr.Wolf, an SoC for edge IoT applications coupling a state of the art micro-controller (MCU) featuring an advanced IO subsystem for efficient data acquisition from high-bandwidth sensors, with an 8-cores floating-point capable computing cluster. The proposed SoC, implemented in a commercial 40nm technology, features a 108 μW fully retentive memory (512kB), an efficient IO subsystem capable to transfer up to 1.6Gbit/s in less than 2.5mW, and an 8-core compute cluster achieving a peak performance of 850 MMAC/s and 500 MFMAC/s (1 GFLOP) and an energy efficiency up to 15 MMAC/s/mW (and 9 MFMAC/s/mW). We demonstrated that Mr.Wolf SoC allows to perform heavy duty floating-point digital signal processing within a power envelope typical of microcontroller architectures.

VI. Acknowledgments

This work was supported in part by the Swiss National Science Foundation under Grant 162524 (MicroLearn: Micropower Deep Learning). We thank Dolphin Integration for providing the Retention Alternating Regulator (RAR).

References

[1] F. Bonomi *et al.*, "Fog computing and its role in the internet of things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, ser. MCC '12. ACM, 2012, pp. 13–16.

[2] D. Rossi *et al.*, "Pulp: A parallel ultra low power platform for next generation iot applications," in *2015 IEEE Hot Chips 27 Symposium (HCS)*, Aug 2015, pp. 1–39.

[3] P. D. Schiavone *et al.*, "Slow and steady wins the race? a comparison of ultra-low-power risc-v cores for internet-of-things applications," in *PATMOS*, Sept 2017, pp. 1–8.

[4] D. Bol *et al.*, "A 25mhz 7 μw/mhz ultra-low-voltage microcontroller soc in 65nm lp/gp cmos for low-carbon wireless sensor nodes," in *2012 IEEE International Solid-State Circuits Conference*, pp. 490–492.

[5] N. Ickes *et al.*, "A 10 pj/cycle ultra-low-voltage 32-bit microprocessor system-on-chip," in *2011 Proceedings of the ESSCIRC*, pp. 159–162.

[6] R. Wilson *et al.*, "A 460mhz at 397mv, 2.6ghz at 1.3v, 32b vliw dsp, embedding fmax tracking," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 452–453.

[7] M. Saint-Laurent *et al.*, "A 28 nm dsp powered by an on-chip ldo for high-performance and energy-efficient mobile applications," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 81–91, Jan 2015.

[8] D. Rossi *et al.*, "Energy-efficient near-threshold parallel computing: The pulpv2 cluster," *IEEE Micro*, vol. 37, no. 5, pp. 20–31, September 2017.

[9] F. Conti *et al.*, "An iot endpoint system-on-chip for secure and energy-efficient near-sensor analytics," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 9, pp. 2481–2494, Sept 2017.