Energy-Efficient Near-Threshold Parallel Computing: The PULPv2 Cluster

Davide Rossi
University of Bologna

Antonio Pullini
ETH Zurich

Igor Loi
University of Bologna

Michael Gautschi, Frank Kağan Gürkaynak
ETH Zurich

Adam Teman, Jeremy Constantin, Andreas Burg
EPFL

Ivan Miro-Panades, Edith Beignè, Fabien Clermidy
CEA-LETI

Philippe Flatresse
STMicroelectronics

Luca Benini
University of Bologna and ETH Zurich

This article presents an ultra-low-power parallel computing platform and its system-on-chip (SoC) embodiment, targeting a wide range of emerging near-sensor processing tasks for Internet of Things (IoT) applications. The proposed SoC achieves 193 million operations per second (MOPS) per mW at 162 MOPS (32 bits), improving the first-generation Parallel Ultra-Low-Power (PULP) architecture by 6.4 and 3.2 times in performance and energy efficiency, respectively.

A growing number of Internet of Things (IoT) applications require flexible processing of datastreams generated by multiple sensors, such as accelerometers, low-resolution cameras, microphone arrays, and vital signs monitors.[1] These applications share the need for high performance (gigaoperations per second, or GOPS) and extreme energy efficiency (GOPS/W) in a power envelope of few mW. A promising approach to achieve vast improvements in the energy efficiency of integrated circuits is near-threshold (NT) computing.[2] Low-voltage operation has been exploited in a number of recent low-power microcontrollers based on Cortex-M processors that demonstrated high energy efficiency—up to 100 million operations per second (MOPS) per mW—within a power envelope of a few mW.[3,4] Unfortunately, peak performance is limited to a few tens of MOPS. Some approaches leverage ultra-wide supply voltage scaling to reliably meet performance requirements, at the cost of energy efficiency.[5,6] Another major challenge in NT operation is the increased sensitivity of devices to process, temperature, and voltage (PVT) variations,[1] which leads to poorly controlled performance levels and unreliable memory operation.

Parallel Ultra-Low-Power (PULP) tackles the performance and variability challenges in NT computing by leveraging two key ideas. First, it pursues thread-level parallelism

Published by the IEEE Computer Society
0272-1732/17/$33.00

with a tightly coupled multicore architecture to overcome the performance degradation at low voltage, while maintaining flexibility and high energy efficiency. Second, it leverages body biasing (BB) as an additional control knob to reduce the effect of PVT variations with significantly reduced energy efficiency losses compared to supply voltage scaling alone.[7] In addition, PULP's microarchitecture is tailored for fine-grained and wide-range operating-point management of cores and memories to ensure high energy efficiency across a wide range of operating voltages (0.32 to 1.1 V), even when application-level parallelism is insufficient to achieve their full utilization.

Davide Rossi and colleagues presented the first embodiment of the PULP platform (PULPv1) in previous work.[7] In this article, we present the second-generation PULPv2 system on a chip (SoC), which achieves much higher energy efficiency and performance. In contrast to PULPv1, implemented with the regular voltage threshold (RVT) or conventional-well flavor of the technology,[7] this design exploits low-voltage threshold (LVT) or flip-well transistors, which significantly extends the system's operating range (down to 0.32 V). On the basis of this capability, much higher operating frequency can be achieved at low voltage (below 0.7 V), providing additional headroom for energy-efficiency improvement.[8] Moreover, in contrast to bulk CMOS and fin field-effect transistor (FinFET), ultra-thin body and box (UTBB) fully depleted silicon-on-insulator (FD-SOI) technology enables aggressive forward body biasing (FBB),[9] leading to a major operating frequency boost at the cost of leakage power. The PULPv2 platform exploits this feature to selectively boost the operating frequency of processors through a fine-grained, low-area-overhead body-biasing architecture, reducing the intrinsic overhead typical for parallel computing platforms during execution of sequential portions of code in boosted mode. For example, applying 2-V FBB to the whole cluster increases leakage power by 80 times; this can be limited to 32 times if only one core is body biased.

To overcome voltage scalability issues of six-transistor static RAMs (6T SRAMs) at low voltage, which is the critical bottleneck for voltage scaling in NT designs,[3,4,7] we introduce a heterogeneous memory architecture, composed of a mixture of latch-based standard cells memories (SCMs)[10] and SRAMs. To adapt the architecture to the computational features of applications and operating conditions, we introduce a set of level-1 (L1) memory reconfiguration knobs, namely lightweight memory management units (MMUs) and reconfigurable pipeline stages (RPSs).[11] By using these reconfiguration knobs, we can tune different parameters of the L1 memory architecture—namely, the access latency, memory access scheme, and number of active memory banks—at runtime, driven by application characteristics and workload requirements to optimize performance, energy efficiency, or both during their execution on the cluster.

PULPv2: Second-Generation PULP Architecture

The platform's computational engine, shown in Figure 1, is a cluster with four cores based on a four-stage, in-order pipeline, implementing the OpenRISC ISA. GCC 4.9 and LLVM 3.7 toolchains are available, with support for OpenMP 3.0 on top of a bare-metal parallel runtime. The streamlined core microarchitecture minimizes static and dynamic power waste, featuring only approximately 30 kilo gates (KGE) per core, architectural clock gating on 98 percent of the core registers, and balanced pipeline stages. The cores do not have private data caches to avoid memory coherency overhead and significantly improve leakage and area efficiency for data memory. Instead, the architecture relies on a tightly coupled data memory (TCDM) featuring eight word-level interleaved banks that are connected to the processors through a nonblocking interconnect to minimize banking conflicts. The cores rely on 1 Kbyte of direct mapped private instruction cache that converges on a shared instruction bus.

Off-cluster level-2 (L2) memory and peripheral access is managed by a tightly coupled DMA through an AXI4-compliant interconnect. The cluster is designed for instantiation in an independent clock and power domain. It features a private clock tree that can be completely gated by an external controller and has level shifters and dual-clock first-in, first-out buffers at its boundaries. The cluster clock is generated through a small frequency-locked

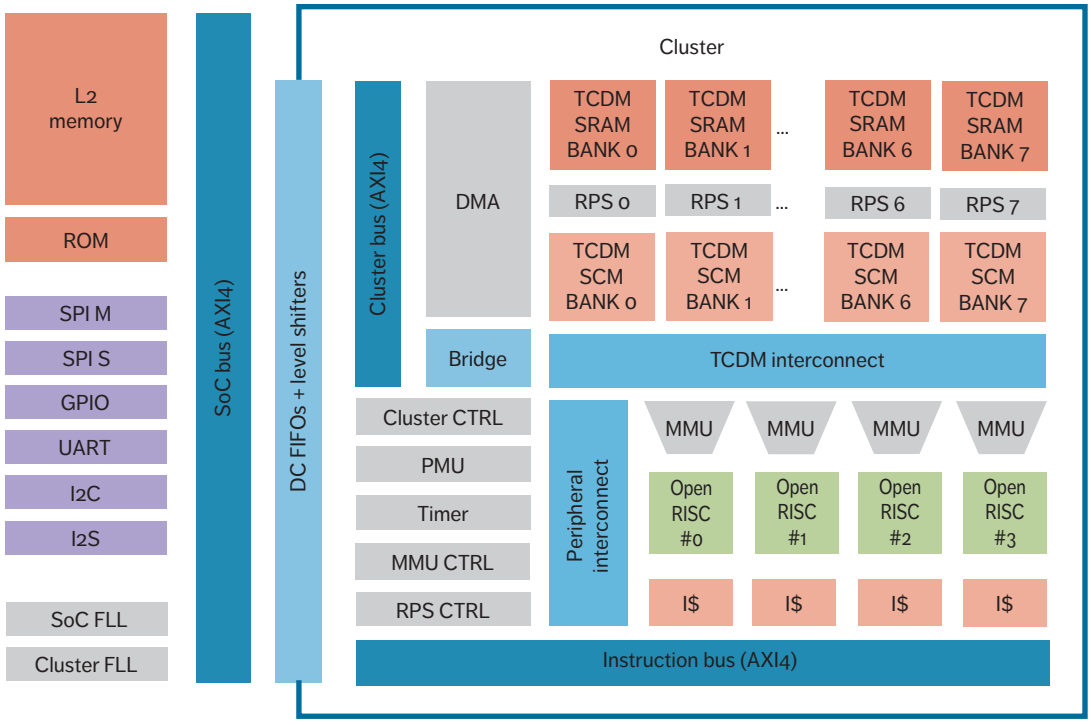


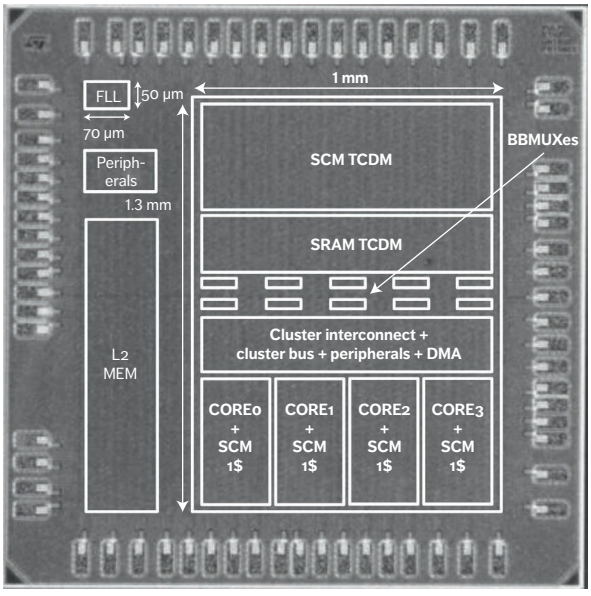
Figure 1. Parallel Ultra-Low-Power version 2 (PULPv2) system-on-chip (SoC) architecture. The SoC is split into two separate voltage and frequency domains, with one including the tightly coupled multiprocessor cluster, and one containing the level-2 (L2) memory and the peripherals.

loop, which requires only 0.3 percent of the cluster area. Several peripherals are available in the PULPv2 SoC, including Serial Peripheral Interfaces with streaming support, general-purpose I/Os, a boot-up ROM, and a Joint Test Action Group interface for debug and test purposes. Figure 2 shows the PULPv2 SoC die micrograph and summarizes its main features.

Heterogeneous Memory Architecture

On-chip memory is a major bottleneck for the energy efficiency of ultra-low-power (ULP) designs for two reasons. First, the access energy of SRAM memories often surpasses the power consumption of the datapath of ULP processors.[3,4,7] Second, the minimum reliable operating voltage for SRAMs is higher than that of logic, and SRAM timing deteriorates faster. In PULPv2, we adopted an approach based on latch-based SCMs[10] to overcome both issues. Because SCMs are constructed exclusively from standard cells, they scale with the core digital logic and continue to operate below the limit of standard 6T SRAM arrays (0.32 versus 0.45 V). This is accompanied by a tradeoff of size, as the basic SCM storage latch is much larger than a standard 6T SRAM bit cell.

To improve the cluster memory's energy efficiency without paying the area overhead of a fully standard-cell-based TCDM, the TCDM logical banks are implemented as heterogeneous memories, comprising a mix of 6T SRAM banks supporting body biasing on the periphery and fully body-biased SCM banks. The private instruction caches are also implemented with SCMs. We adopted a controlled SCM placement methodology to reduce the area overhead and further improve energy efficiency.[10] The resulting 128 × 32 SCM cuts used in the design feature an area of 86 µm × 160 µm. They consume 3.3 times less active energy and 2.2 times less leakage power with respect to an equivalent solution with 6T SRAM macros, at the expense of just 2.7 times area overhead (as opposed to more than 4 times for nonoptimized SCM approaches). At the cluster level, the introduction of SCMs improves energy efficiency by 38 percent compared to a fully SRAM-based implementation.



Technology	UTBB FD-SOI 28 nm
Transistors	Flip well L = 24 nm
Cluster area	1.3 mm^2
V_{DD} range (SRAMs)	0.32 – 1.15 V (0.45 – 1.15 V)
BB range	0 – 1.75 V
SRAM macros	8 x 32 Kbits (TCDM)
SCM macros	16 x 4 Kbits (TCDM) 4 x 2 x 4 Kbits (1$)
Gates	200K
Frequency range	NO BB: 40.5 – 710 MHz MAX FBB: 63.5 – 825 MHz
Power range	NO FBB: 0.56 – 85 mW MAX FBB: 6.9 – 480 mW

Figure 2. Chip micrograph and main features of the PULPv2 SoC. The figure on the left shows the floorplan of the chip. The table on the right provides quantitative information about the chip.

The 64 Kbyte L2 memory is implemented with high-density 6T SRAM banks.

Reconfigurable Pipeline Stages

Reconfigurable pipeline stages extend the TCDM interconnect to deal with the performance degradation and variability of the SRAM slice of the TCDM at low voltage (see Figure 3), in which we can identify three operating regions. Above 0.7 V, SRAMs are not a performance limiter; between 0.55 and 0.7 V, SRAM performance starts degrading; below 0.55 V, SRAM delay dominates. Each path from and toward the SRAM banks is extended with two modules: a request and a response reconfigurable pipeline block. Each reconfigurable pipeline stage can be independently activated through a set of configuration registers mapped on the peripheral interconnect. It is therefore possible to fully or partially surround each memory bank with additional pipeline resources, providing the capability to deal with global or local variations of SRAM bank timing, at the cost of up to two additional cycles of latency during load and store operations. It is important to note that when additional pipeline stages for the SRAMs are activated, compiler-managed latency-aware allocation techniques can be exploited to leverage the unpipelined SCMs as an energy-efficient, low-latency buffer for frequently accessed variables, minimizing the number of high-latency accesses to SRAM banks.

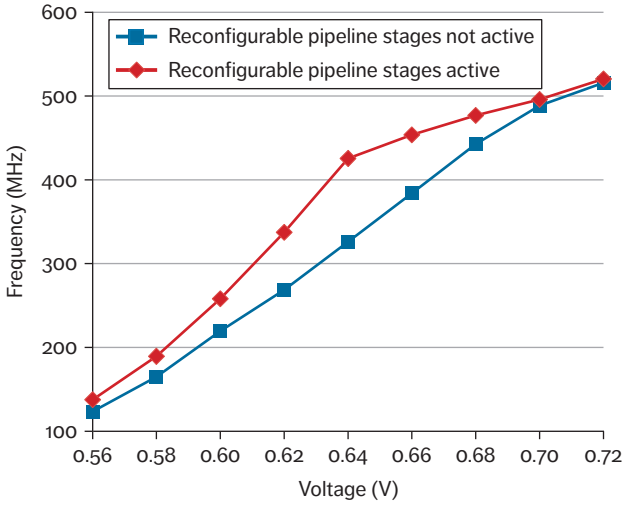


Figure 3. Impact of reconfigurable pipeline stages on the cluster's operating frequency. Reconfigurable pipeline stages can improve the cluster's operating frequency by up to 30 percent.

Architectural Support for Selective Boost

The selective boost architecture of Figure 4 provides the capability for fine-grained

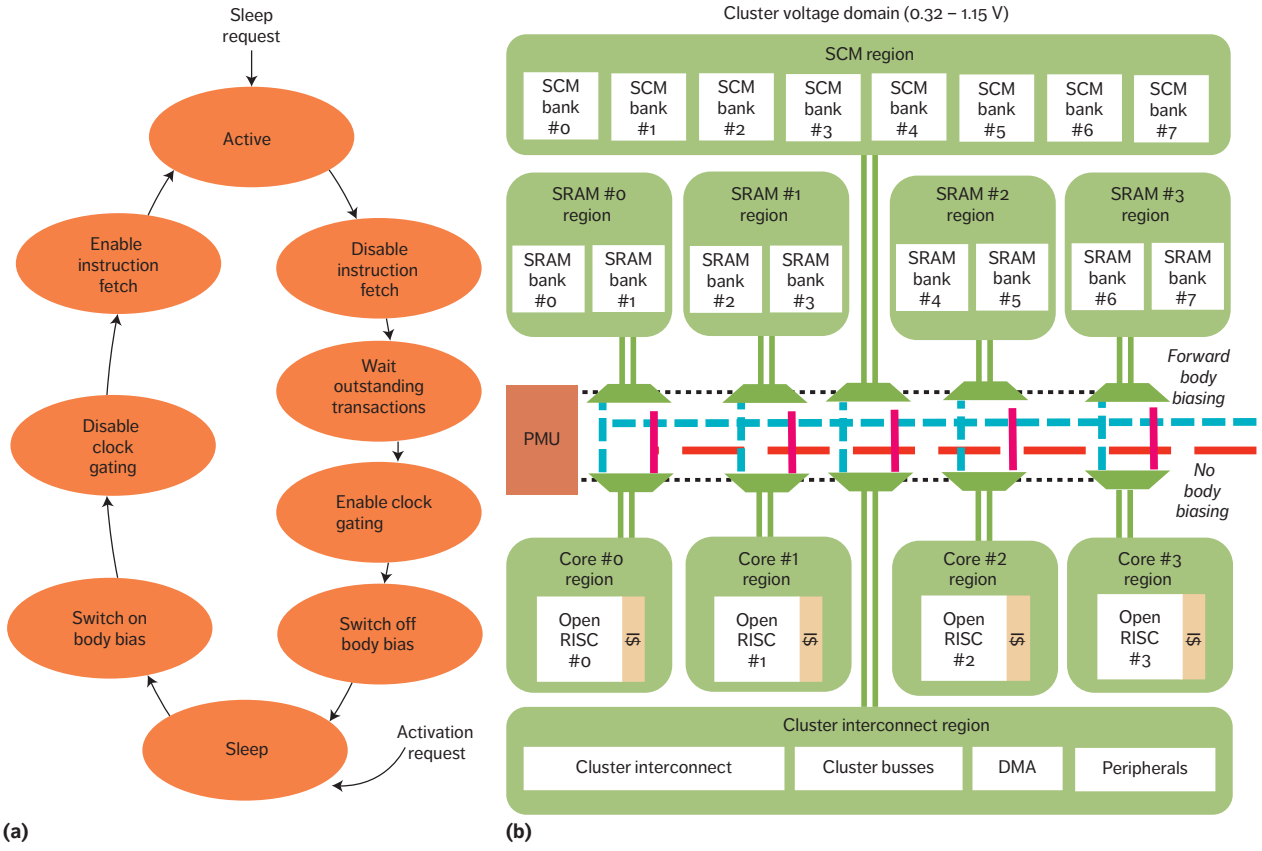


Figure 4. Power management architecture. (a) Cluster partitioning in clock gating and body bias regions. (b) Power management unit (PMU) finite state machine description.

ultra-dynamic BB management of architectural entities to reduce leakage overheads typical of parallel computing platforms during boosted execution, for example, during sequential portions of applications. Ten BB multiplexers (BBMUX), controlled by a memory-mapped power management unit (PMU), enable dynamic configuration of the BB voltage of the regions, featuring approximately 30-ns transitions between the FBB and no BB modes. This approach is one order of magnitude faster than typical voltage supply transition times (that is, tens of microseconds), and it has minimal overhead in terms of area (the spacing required to isolate the wells between the body bias regions is only 3.5 µm). The sleep and wake-up procedure is managed by the PMU (see Figure 4), which handles the control signals to gate the clocks of idle regions and related BBMUXes. Power management of idle resources improves energy efficiency during boost execution of sequential code by up to 60 percent in both dynamic and leakage power dominated regions. Control of power management knobs is fully integrated in the OpenMP runtime, hence completely transparent from the programmer's viewpoint.

Lightweight MMUs

A set of lightweight MMUs, connected between the cores and the interconnect (see Figure 1), provides the capability to dynamically reconfigure the address space according to application needs and workload. This feature enables fine-grained division of the memory space into private and shared data to maximize the TCDM's energy efficiency and performance. To reduce banking conflicts, shared memory is interleaved across memory banks, whereas the private data is not interleaved, and each section of private data can be located in a separate physical memory bank. Moreover, in cases where the application's memory requirements are smaller than the actual physical memory space, the

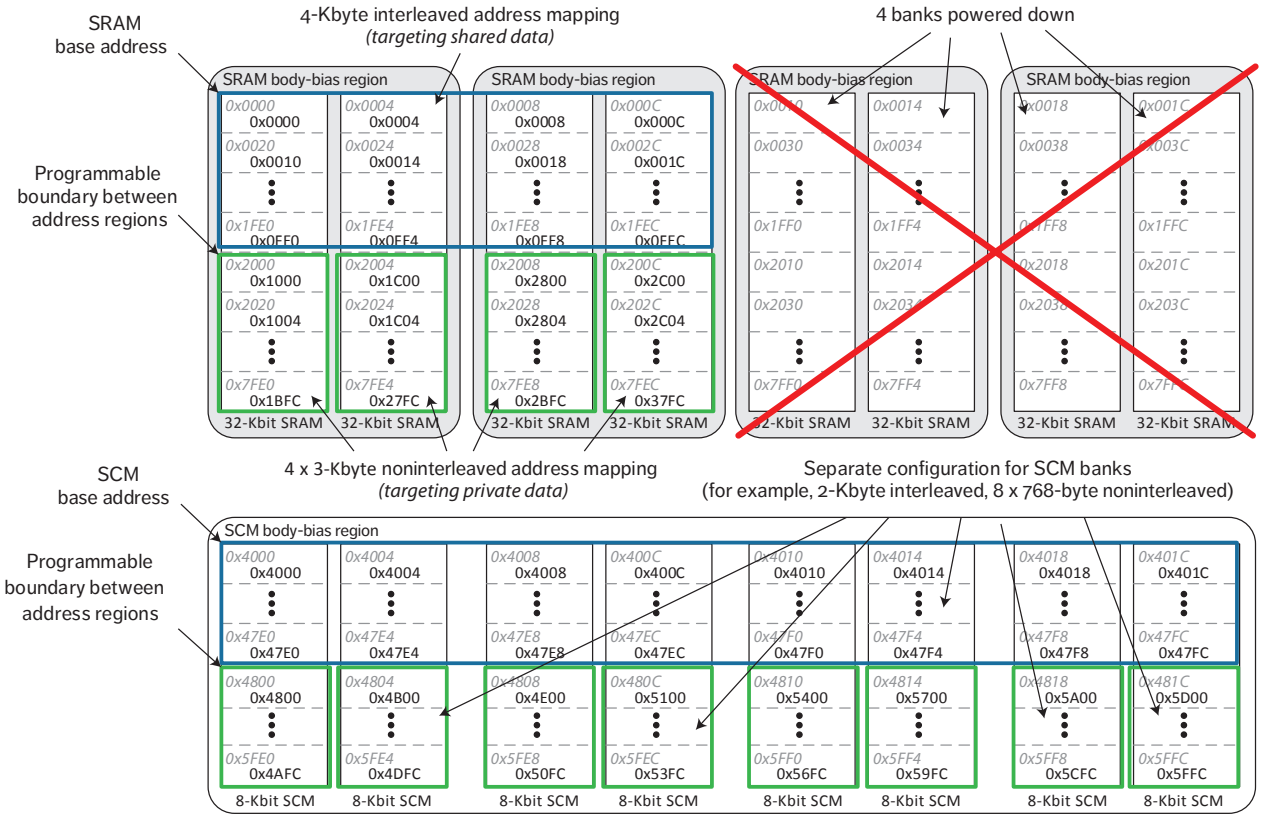


Figure 5. Functionality of the memory management units (MMUs) for tightly coupled data memory (TCDM) address remapping. The address space is reconfigured for partial shutdown of static RAM (SRAM) banks and to manage bank-interleaved buffers shared among cores and private per-core buffers.

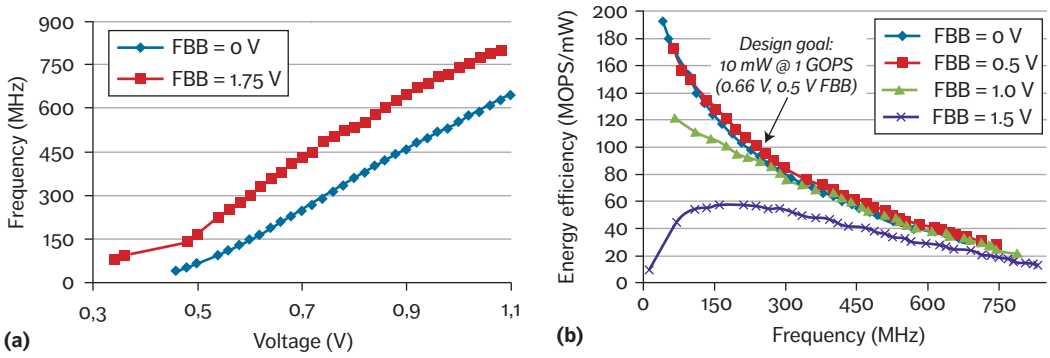
shared and private data can be organized, such that not all physical memory banks are used, and the unused banks can then be powered down. Figure 5 gives an example of a possible configuration. In this case, the application utilizes four cores and requires 4 Kbytes of shared data memory with 3 Kbytes of private data memory per core. In order to efficiently meet the requirements, the shared memory space is interleaved across the first 1-Kbyte segment of four of the eight SRAM banks, whereas the private memory is allocated to each core as noninterleaved portions of the remaining memory of the four individual banks. Because the entire SRAM space required in this configuration is only 16 Kbytes and the addressing is organized such that it is entirely mapped to four of the eight physical SRAM banks, the remaining four banks are powered down for energy savings. This separation of memory spaces can be independently applied to the SRAM banks and to the SCMs, as shown in Figure 5, with the SCMs programmed to have 2 Kbytes of interleaved memory for shared data and eight separate noninterleaved private address spaces.

Evaluation of PULPv2 SoC

This section provides a quantitative evaluation of the PULPv2 SoC, describing the chip measurements and an extensive exploration of the SoC architecture.

Chip Measurements

Figure 6 shows the maximum frequency measured on the silicon prototype running a typical high-utilization workload (matrix multiplication), the energy efficiency, and the related power consumption. At the best energy-efficiency point of 193 MOPS/mW (0.46 V, 0-V FBB), the cluster's measured power is 840 μW at 40.5 MHz (162 MOPS), whereas within a 10-mW power envelope, the cluster achieves the design goal of 100 MOPS/mW at



Power (mW)	0.5 V	0.6 V	0.7 V	0.8 V	0.9 V	1.0 V	1.1 V
Multicore power (mW)	1.5	4.8	10.6	20.2	33.1	50.8	68.6
Single-core power (mW)	0.75	2.3	5.2	9.8	16.3	24.7	32.3
Leakage power (mW)	0.19	0.37	0.62	0.91	1.5	2.3	3.5

(c)

Figure 6. Performance evaluation of the PULPv2 SoC. (a) Maximum operating frequency with reconfigurable pipeline stages off. (b) Energy efficiency in the multicore configuration. (c) Power consumption with FBB = 0 V silicon measurements with matrix multiplication running on the cluster. (FBB: forward body biasing.)

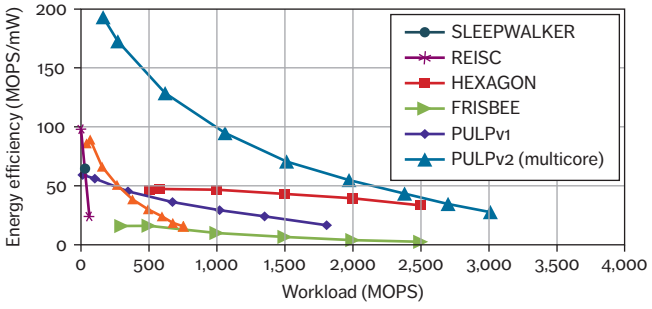


Figure 7. Energy-efficiency comparison with recent ultra-low-power (ULP) microcontrollers and energy-efficient digital signal processors (DSPs).

1 GOPS (0.66 V, 0.5-V FBB). Although a large BB range does not improve energy efficiency across the voltage range in nominal conditions (typical process, 25°C), it is very useful to compensate process and temperature variations, especially for devices operating at low voltage subject to strong thermal inversion.[7,12]

Table 1 shows a comparison with recent energy-efficient digital signal processors (DSPs), and Figure 7 highlights the energy efficiency with respect to the computational workload. By exploiting the energy boost provided by the heterogeneous memory architecture and an implementation methodology carefully tuned for energy efficiency in NT,[7] the proposed cluster's peak energy efficiency surpasses existing ULP microcontrollers by more than two times in terms of energy efficiency while achieving 40 times better peak performance, and it exceeds existing wide-voltage-range DSPs by more than 4.5 times in terms of energy efficiency, achieving 1.13 times better peak performance. When operating in the single-core configuration, the cluster still surpasses other reported peak energy efficiency figures by 1.3 to 7.6 times, showing the effectiveness of the dynamic and leakage power management of idle blocks exploited in the design. With respect to the first-generation PULP platform, the best energy operating point of the v2 cluster provides 6.4 times improvement in performance and 3.2 times improvement in energy efficiency.

Beyond the described architectural and design methodology enhancements,[7] the choice of the high-performance, flip-well, UTBB FD-SOI technology with respect to the low-leakage, conventional-well flavor greatly increases the best energy efficiency and the operating frequency of the best energy point, at the cost of approximately 10 times higher leakage power. Hence, the flip-well flavor is more suitable for mostly active applications and

Table 1. Comparison with recent ULP microcontrollers and wide-voltage-range DSPs.

	Low-power MCUs		Wide-voltage-range DSPs		PULP SoCs	
Parameter	**Sleepwalker[3]**	**REISC[4]**	**HEXAGON[5]**	**FRISBEE[6]**	**PULPv1**	**PULPv2**
Technology	CMOS 65-nm LP GP	CMOS 65-nm LP	CMOS 28-nm LP	FD-SOI 28-nm flip-well	FD-SOI 28-nm conventional-well	FD-SOI 28-nm flip-well
Data format	16-bit	32-bit	4x 32-bit VLIW	32-bit	32-bit	32-bit
No. of cores	1	1	1	1	4	4
I$/D$/L2	16 Kbytes (64-bit)/ 2 Kbytes/N/A	8 Kbytes (128-bit)/ 8 Kbytes (128-bit)/N/A	16 Kbytes/ 32 Kbytes/ 256 Kbytes	4 Kbytes/ 4 Kbytes/N/A	1 Kbyte x4/ 16 Kbytes/ 16 Kbytes	1 Kbytes x4/ 48 Kbytes/ 64 Kbytes
Voltage range (SRAMs)	0.4 V (1.0 V)	0.54 – 1.2 V (0.4 – 1.2 V)	0.6 – 1.05 V	0.4 – 1.3 V	0.44 – 1.2 V (0.54 – 1.2 V)	0.32 – 1.15 V (0.45 – 1.15 V)
Max frequency	25 MHz	82.5 MHz	1.2 GHz	2.6 GHz	475 MHz	825 MHz
Best power density	15.5 µW/MHz	10.2 µW/MHz	58 µW/MHz	62 µW/MHz	65 µW/MHz	20.7 µW/MHz
Peak performance	25 MOPS	82.5 MOPS	3 GOPS	2.6 GOPS	1.8 GOPS	3.3 GOPS
Peak energy efficiency	64.5 MOPS/ mW @ 25 MOPS	98 MOPS/mW @ 0.54MOPS	43.1 MOPS/ mW @ 230 MOPS	16.1 MOPS/ mW @ 460 MOPS	60 MOPS/mW @ 25.6 MOPS	193 MOPS/ mW @ 162MOPS

for switchable computing engines that can be power gated and activated on demand. On the other hand, the conventional-well flavor is suitable for always-on applications, requiring low power consumption and targeting workloads in the range of few MHz.

Architectural Evaluation of PULPv2 SoC

Table 2 shows an evaluation of a set of representative benchmarks on the PULPv2 SoC, featuring widely different characteristics in terms of parallelism, computational density, and code locality. The instructions per cycle (IPC) of the applications is greater than 0.85 when running on a single core and greater than 0.97 for computationally intensive kernels, thanks to the highly optimized microprocessor microarchitecture that stalls only in the case of read-after-write hazards during load operations. The IPC slightly degrades in applications with complex control flow, due to the increase of instruction cache stalls. When moving to multicore execution, most of the intrinsically parallel, computationally intensive algorithms still perform optimally on the SoC, providing almost ideal speedup (that is, four times), especially when the MMUs are enabled to reduce contention on the TCDM. Stalls due to instruction cache misses, caused by calls to the functions of the OpenMP runtime, affect some of these applications (DCT, convolution, SOR), slightly reducing the IPC with respect to single-core execution (approximately 0.8). On the other hand, LU and Dijkstra feature

Table 2. Analysis of parallel and sequential signal-processing applications on the PULPv2 SoC.

Parameter	MatrixMul	Convolution	Sparse	DCT	SOR	LU	Dijkstra
Single core							
No. of clock cycles	60,964	131,937	14,847	29,252	106,586	74,132	17,568
No. of instructions executed	60,821	130,766	14,735	28,431	93,348	70,913	14,964
Core stalls (%)*	0.01	0.02	0.44	0.07	0.45	0.02	11.22
I$ stalls (%)*	0.22	0.87	0.32	2.74	11.97	4.32	3.60
IPC	1.00	0.99	0.99	0.97	0.88	0.96	0.85
Energy efficiency (MOPS/mW)§	32.4	32.2	32.3	31.6	28.5	31.1	27.7
Energy efficiency (power-managed) (MOPS/mW)§	51.9	51.5	51.6	50.5	45.5	49.7	44.3
Multicore							
No. of clock cycles	16,369	38,182	3,980	8,027	31,027	58,110	74,410
No. of instructions executed	61,003	139,492	14,933	29,857	101,587	166,905	134,497
Core stalls (%)*	0.10	0.09	0.80	1.57	0.22	0.31	3.05
I$ stalls (%)*	2.19	3.86	2.74	12.7	16.1	21.2	23.0
TCDM stalls (%)*	5.38	6.59	0.38	4.61	0.05	0.03	0.46
IPC†	0.92	0.89	0.96	0.81	0.84	0.78	0.73
TCDM accesses per cycle	1.28	1.38	0.88	1.15	0.10	0.02	0.29
SCM TCDM accesses (%)‡	12.1	60.9	2.28	46.8	77.5	88.1	83.4
Slaves sleep (%)*	1.44	1.01	4.35	3.05	7.83	16.0	46.6
Speedup vs. single core	3.72	3.46	3.73	3.64	3.44	1.28	0.24
Energy efficiency (MOPS/mW)§	92.9	85.6	92.6	88.5	75.2	30.5	5.0

Table 2. Analysis of parallel and sequential signal-processing applications on the PULPv2 SoC.

Parameter	MatrixMul	Convolution	Sparse	DCT	SOR	LU	Dijkstra
Reconfigurable pipeline stages on							
No. of cycles increase (%)	35.5	28.6	23.8	16.7	0.40	0.83	14.9
Normalized execution time (%)§	+0.74	−4.37	−7.92	−13.2	−25.0	−24.7	−14.5
Normalized energy efficiency (%)§	−14.2	−11.4	−9.5	−6.7	−0.2	−0.3	−6.0
Normalized energy efficiency with V_{DD} scaling (%)‖	−14.6	−8.7	−4.6	+1.6	+15.3	+15.3	+3.1
Memory management unit on							
IPC†	0.97	0.95	0.96	0.82	0.84	0.78	0.72
Speedup vs. single core	3.92	3.68	3.74	3.98	3.44	1.28	0.24
Normalized energy efficiency (%)§	+2.1	+2.7	+0.1	+3.7	+0.1	+0.0	+0.7
Normalized energy efficiency with V_{DD} scaling (%)‖	+4.5	+5.0	+2.5	+6.1	+2.2	+0.8	+0.8

** Percentage of time is calculated with respect to the number of clock cycles. † IPC is calculated considering the number of active cycles for slave cores. ‡ SCM TCDM accesses are calculated as a percentage of the overall TCDM accesses. § 0.65 V, 0.5 FBB, 250 MHz (with RPS off) is assumed as nominal operating point, which provides peak performance of 1 GOPS @ 10 mW, and peak energy efficiency of 100 MOPS/mW in multicore configuration. || The V_{DD} that allows achievement of the same performance (GOPS) as the nominal operating point is assumed.*

low parallelism (see the percentage of idle slaves in Table 2), especially for the small datasets involved in deeply embedded applications (that is, a few Kbytes of data). Moreover, the complex parallelization patterns required to extract some parallelism from these algorithms trigger several calls to the OpenMP runtime. This significantly increases the overall number of instructions executed (by 2.3 times for LU and 8.9 times for Dijkstra) and the stalls due to instruction cache miss (by 3.8 times for LU and 27 times for Dijkstra), further reducing the IPC and speedup with respect to execution on a single core.

In this scenario, power management of idle cores and selective boost of active cores provides an effective way to execute applications featuring pathological parallelization bottlenecks, improving energy efficiency by up to 1.6 times with respect to a single-core execution on a fully active cluster. Moreover, leveraging reconfigurable pipeline stages provides an additional knob to optimize applications, especially those dominated by irregular, control-dominated programs (hence, small TCDM bandwidth), providing up to 25 percent reduction of the execution time. For this class of applications, reconfigurable pipeline stages also lead to an improvement of energy efficiency of up to 15 percent, if combined with voltage scaling.

We have presented PULPv2, an energy-efficient parallel accelerator for near-sensor processing applications. The SoC demonstrates outstanding performance and energy efficiency with respect to existing architectures, achieving the milestone of 1 GOPS within 10 mW for a fully programmable 32-bit architecture. The active and idle power management circuits and the heterogeneous memory hierarchy coupled with reconfiguration knobs provide an effective way to maintain high energy efficiency for our tightly coupled multicore cluster deep into the NT region. Future work will focus on hardware-assisted synchronization mechanisms to reduce parallelization overheads and on the optimization of the instruction cache architecture to improve the execution of the parallel runtime that often reduces the code locality typical of computing-intensive applications. ■

Acknowledgments

This work is supported by the European FP7 ERC Advanced project MULTITHERMAN (g.a. 291125) and by the Swiss National Science Foundation (SNF) project (no. 162524) "MicroLearn: Micropower Deep Learning." We thank STMicroelectronics for chip fabrication.

References

1. F. Bonomi et al., "Fog Computing and Its Role in the Internet of Things," *Proc. 1st MCC Workshop Mobile Cloud Computing*, 2012, pp. 13–16.
2. R.G. Dreslinski et al., "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits," *Proc. IEEE*, vol. 98, no. 2, 2010, pp. 253–266.
3. D. Bol et al., "A 25MHz 7μW/MHz Ultra-Low-Voltage Microcontroller SoC in 65nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes," *Proc. IEEE Int'l Solid-State Circuits Conf.*, 2012, doi:10.1109/ISSCC.2012.6177104.
4. N. Ickes et al., "A 10 pJ/Cycle Ultra-Low-Voltage 32-Bit Microprocessor System-on-Chip," *Proc. ESSCIRC*, 2011, doi:10.1109/ESSCIRC.2011.6044889.
5. M. Saint-Laurent et al., "A 28 nm DSP Powered by an On-Chip LDO for High-Performance and Energy-Efficient Mobile Applications," *Proc. IEEE Int'l Solid-State Circuits Conf.*, 2014, doi:10.1109/ISSCC.2014.6757388.
6. R. Wilson et al., "A 460MHz at 397mV, 2.6GHz at 1.3V, 32b VLIW DSP, Embedding FMAX Tracking," *Proc. IEEE Int'l Solid-State Circuits Conf.*, 2014, doi:10.1109/ISSCC.2014.6757509.
7. D. Rossi et al., "A 60 GOPS/W, −1.8V to 0.9 V Body Bias ULP Cluster in 28 nm UTBB FD-SOI Technology," *J. Solid-State Electronics*, vol. 117, 2016, pp. 170–184.
8. D. Rossi et al., "193 MOPS/mW @ 162 MOPS, 0.32V to 1.15V Voltage Range Multi-Core Accelerator for Energy Efficient Parallel and Sequential Digital Processing," *Proc. IEEE Symp. Low-Power and High-Speed Chips* (Cool Chips 19), 2016, doi:10.1109/CoolChips.2016.7503670.
9. P. Flatresse et al., "Ultra-Wide Body-Bias Range LDPC Decoder in 28nm UTBB FDSOI Technology," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC), 2013, doi:10.1109/ISSCC.2013.6487798.
10. A. Teman et al., "Power, Area, and Performance Optimization of Standard Cell Memory Arrays through Controlled Placement," *ACM Trans. Design Automation of Electronic Systems (TOADES)*, vol. 21, no. 4, 2016, article 59.
11. M.R. Kakoee, I. Loi, and L. Benini, "Variation-Tolerant Architecture for Ultra Low Power Shared-L1 Processor Clusters," *IEEE Trans. Circuits and Systems II*, vol. 59, no. 12, 2012, pp. 927–931.
12. S. Clerc et al., "A 0.33V/−40°C Process/Temperature Closed-Loop Compensation SoC Embedding All-Digital Clock Multiplier and DC-DC Converter Exploiting FDSOI 28 nm Back-Gate Biasing," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC), 2015, doi:10.1109/ISSCC.2015.7062970.

Davide Rossi is an assistant professor in the Energy Efficient Embedded Systems Laboratory at the University of Bologna. His research interests include ultra-low-power multicore SoC design and applications. Rossi received a PhD in electronics engineering from the University of Bologna. Contact him at davide.rossi@unibo.it.

Antonio Pullini is a PhD student at ETH Zurich. His research interests include ultra-low-power system-on-chip design with a special focus on peripheral subsystems. Pullini received an MsC in electronics engineering from the University of Bologna. Contact him at pullinia@iis.ee.ethz.ch.

Igor Loi is an assistant professor in the Energy Efficient Embedded Systems Laboratory at the University of Bologna. His research interests include ultra-low-power multicore systems and memory systems evolution. Loi received a PhD in electronics engineering from the University of Bologna. Contact him at igor.loi@unibo.it.

Michael Gautschi is a PhD student at ETH Zurich. His current research interests include low-power digital circuits, processor architectures, and mobile communication. Gautschi received a PhD in electronics engineering from ETH Zurich. Contact him at gautschi@iis.ee.ethz.ch.

Frank Kağan Gürkaynak is a senior scientist at ETH Zurich. His research interests include digital VLSI design. Gürkaynak received a PhD in electrical and computer engineering from ETH Zurich. Contact him at kgf@ee.ethz.ch.

Adam Teman is a tenure-track senior lecturer in the Emerging Nanoscaled Integrated Circuits and Systems (EnICS) labs at Bar-Ilan University, Israel. His research interests include embedded memory design for energy-efficient systems. Teman received a PhD in electronics engineering from EPFL, where he completed the work for this article. He is a member of IEEE. Contact him at adam.teman@biu.ac.il.

Jeremy Constantin is a PhD student in the Telecommunications Circuits Laboratory at EPFL. His research interests include instruction-set architectures and microarchitectural techniques for low-power design. Constantin received a PhD in electronics engineering from EPFL. Contact him at jeremy.constantin@epfl.ch.

Andreas Burg is a professor at EPFL, where he leads the Telecommunications Circuits Laboratory. His research interests include low-power VLSI signal processing. Burg received a Dr. sc. techn. in electronics engineering from ETHZ. Contact him at andreas.burg@epfl.ch.

Ivan Miro-Panades is a research engineer at CEA-LETI, Grenoble. His research interests include multiprocessor architectures and power optimization on advanced CMOS technology nodes. Miro-Panades received a PhD in computer science from the University of Pierre & Marie Curie. Contact him at ivan.miro-panades@cea.fr.

Edith Beignè is the head of the low-power design team within the digital design laboratory at CEA-LETI. Her research interests include fine-grained power control and local voltage and frequency scaling innovative features. Beignè received degrees from Grenoble Polytechnical Institute, France, and her higher degree research (research director) in 2014. Contact her at edith.beigne@cea.fr.

Fabien Clermidy is the head of the digital architecture and design laboratory at CEA-LETI. His research interests include multicore architectures and design with a focus on emerging technologies and memories. Clermidy received a PhD in engineering science from the Grenoble Institute of Technology. Contact him at fabien.clermidy@cea.fr.

Philippe Flatresse is a design architect at STMicroelectronics Central R&D. His research interests include low-power and high-performance digital design techniques in both bulk and SoI technologies. Flatresse received a PhD in microelectronics from Grenoble Institute of Technology. Contact him at philippe.flatresse@st.com.

Luca Benini is a full professor of electronics at the University of Bologna and ETH Zurich. His research interests include energy-efficient system design and multicore SoC design. Benini received a PhD in electrical engineering from Stanford University. Contact him at luca.benini@unibo.it.