A 0.4-0.9V, 2.87pJ/cycle Near-Threshold ARM Cortex-M3 CPU with In-Situ Monitoring and Adaptive-Logic Scan

Markus Hiienkari[1], Navneet Gupta[1], Jukka Teittinen[1], Jesse Simonsson[1], Matthew Turnquist[1], Jonas Eriksson[1], Risto Anttila[1], Ohto Myllynen[1,2], Hannu Rämäkkö[1], Sofia Mäkikyrö[1], Lauri Koskinen[1,2]

[1]Minima Processor,
Rantakatu 3, 90100 Oulu, Finland
first.last@minimaprocessor.com
[2]University of Turku, Department of Future Technologies

Abstract

An adaptive ARM Cortex-M3 with near-threshold to nominal voltage operation is presented. Using in-situ monitoring technology, the CPU energy is minimized across process, voltage, temperature, and applications. At the Minimum-Energy Point (MEP), the CPU achieves 2.87pJ/cycle (7.8MHz/0.378V). Absolute energy consumption is reduced by 76% by operating at the MEP. The core includes a novel configuration of industry-standard ATPG-compatible DFT architecture for adaptive logic testing and the system includes Dynamic Voltage and Frequency Scaling software. Additionally presented is an execution-frequency analysis algorithm based on software execution trace. The algorithm achieves 60% energy savings for an industry-standard speech-recognition software without any penalty in application throughput.
Keywords: Near-Threshold, DVFS, CPU, System-on-Chip

Ultra-Wide Dynamic Voltage and Frequency Scaling (DVFS) ARM Cortex-M3

Modern applications demand large performance spreads within the hardware and simultaneous conflicting low-power demands. For example, AI-based keyword spotting with always-on noise detection requires performance from tens of MHz to hundreds of MHz. For novel architectures, testability and compatibility with industry-standard testing is critical to achieve production-quality designs, time-to-market, and high yield.

The SoC architecture is shown in Fig. 1a. The ultra-wide DVFS capable Cortex-M3 CPU is supplemented with an on-chip LDO and stock memory, peripherals, and interfaces. The CPU is manufactured in 28nm CMOS technology (Fig. 1b) and designed with the process-standard gate library and industry-standard EDA. In-situ critical-path monitoring and time-borrowing flip-flops are used to reduce the large and energy-consuming timing safety margins at near threshold thereby enabling the ultra-wide DVFS. Critical-path FFs are replaced by the in-house monitor / time-borrowing FF combination with the replacement decision made by in-house EDA. A dynamically-stretched clock generated from the system clock is used to cancel time borrowing (based on [1]) every clock cycle a monitor fires (timing events). The novel critical path monitor is shown in Fig.1c. The DVFS of the processor is based on in-house controller Soft IP placed on a bus (AXI, APB etc.) and in-house DVFS SW. Therefore, the CPU can monitor the frequency of the monitors' output and the SW can be configured for energy minimization with guaranteed performance (i.e. frequency requirement) or MEP operation. Additionally, the monitor signals can be configured to directly signal the voltage control HW to raise or decrease the voltage. The direct HW control is the preferred configuration as it is the fastest DVFS method, but the control can also issue interrupts (e.g. in the case of exceeding timing event or idle cycle thresholds). Figure 2a depicts the scenario where the control directly does HW voltage adjustment and then issues an interrupt to the SW and Fig. 2b shows the measured output of the power management. The controller also can inform SW about situations where there are no timing events at all. This makes it possible for the SW to adjust the voltage down until there are some timing events occurring. This equilibrium is used in search of the minimum energy point.

The core includes a novel DFT method for adaptive-logic testing as shown in Fig. 3a. The monitor output is added to the scan path and an observation test-point flip-flop is added on the output of the OR-tree that collect the monitor signals. During scan, the monitoring can be configured in a multitude of modes, such as time borrowing on/off, capture with or without time-borrowing, preventing the timing-event signal from being reset, stuck-at and at-speed timing-event capture. Fig. 3b shows characterization of path-delay to the cell (A), forced time-borrow events generation (C), false-positives (D) and false-negatives (B). All tests are performed with industry-standard ATPG stuck-at and at-speed tests.

Measurement Results

The CPU was designed for operation from 12.5MHz/0.4V to 200MHz/0.9V in typical TT/25C conditions. The process-corner samples are functional over a wide operating frequency range up to 307MHz (@0.99V) across temperature corners. The operation was verified with Dhrystone and SHA256 with and without FreeRTOS OS. Typical samples required 0.407V on average to operate at 12.5MHz in room temperature (Fig. 4). With a fixed 12.5MHz performance requirement (i.e. where a chip designed with 200MHz/0.9V specification is operated for an application requirement of 12.5MHz), the SS and FF chips show increased energy due to higher switching- and leakage-currents. Without AVS, the energy budget is restricted by FF leakage and the maximum frequency by ON-current of SS samples. A best case of 76% improvement in energy efficiency is achieved at 0.378V/7.8MHz for Dhrystone and 0.377V/6.9MHz for SHA256. The energy is 2.87pJ/cycle at the lowest voltage of 0.37V and 3.2 pJ/cycle at the specified 12.5MHz. The AVS technology brings energy benefits for all chips in the process spread with up to 58% reduction in average power. For high duty cycles, samples on the SS side of the process spread will be less energy efficient, due to their reduced possibility for VDD decrease. Even in this case, TT and FF samples will still have energy gain.

Fig. 5a. shows the energy for the entire operation range including best/worst process and temperature corners, only limited by the PLL [2] minimum frequency (6.25MHz) and the maximum allowed voltage (0.99V). Fig. 5b shows the leakage over the operation range and the importance of AVS in the case of high temperatures and leaky (FF) process corners. Fig. 6 shows the benefit with a fixed performance requirement. For 12.5MHz, the best-case sample operates at 0.37V and the worst-case sample at 0.45V. Without AVS, worst-case energy would be the FF leakage but with AVS the SS chip is the worst. The energy saving in this case is 18.75%. Conversely, the FF chip could be operated at 46MHz at the same voltage. For 200MHz the best-case energy saving is 34%. Table 1 shows a comparison to other near-threshold CPUs where the presented core has the best energy when normalized to efficiency.

The system also includes application-code analysis SW that defines execution frequencies based on software execution trace and user provided data on timing constraints and sets the voltage accordingly. By analyzing the code, the system can achieve optimal energy savings for the code and avoid run-to-completion energy loss. The SW 1) collects and pre-processes the application code execution trace, 2) profiles the code, and optimizes the execution frequencies based on Dual-Simplex algorithm. The analysis was tested on ARM always-on Keyword Spotting (KWS) SW [3]. By automatically increasing the KWS execution time 77.0% (while still meeting the task DL) the energy consumption decreased 59,7%, as shown in Fig. 7.

[1] M. Turnquist, M. Hiienkari, J. Mäkipää, R. Jevtic, E. Pohjalainen, T. Kallio, L. Koskinen, "Fully Integrated DC-DC Converter and a 0.4V 32-bit CPU with Timing-Error Prevention Supplied from a Prototype 1.55V Li-ion Battery" Symposium on VLSI Circuits, C320-C321, 2015

[2] Silicon Creations, "Ultra Low Area Frequency Synthesizer PLL" Internet: https://www.design-reuse.com/sip/ultra-low-area-frequency-synthesizer-pll-5nm-90nm-ip-30957/, Feb. 3. 2020

[3] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: keyword spotting on microcontrollers", arXiv: 1711.07128v3 [cs.SD], Feb. 2018.

[4] S. Paul, et. al., "An energy harvesting wireless sensor node for IoT systems featuring a near-threshold voltage IA-32 microcontroller in 14nm tri-gate CMOS", Symposium on VLSI Circuits, pp. 1-2, 2016

[5] J. Myers, et. al., "A 12.4pJ/cycle sub-threshold, 16pJ/cycle near-threshold ARM Cortex-M0+ MCU with autonomous SRPG/DVFS and temperature tracking clocks", Symposium on VLSI Circuits, C332-C333, 2018

	VLSI2015 [1]	VLSI 2016 [4]	VLSI 2018 [5]	**This work**
Technology	28nm FDSOI	14nm Tri-gate CMOS	65nm	**28nm HPC+**
CPU	LatticeMico RISC	x86 IA	ARM Cortex-M0+	**ARM Cortex-M3**
Core area	0.037mm^2	N/A	N/A	**0.04mm^2**
VDD range	0.3-0.5V	0.308-1V	0.3-0.8V	**0.4-0.9V**
Freq. range	1-77MHz*	0.5-297MHz	0.012-60MHz	**12.5-200MHz**
Eff [DMIPS/MHz]	1.14	1.6**	0.95	**1.25**
E_{CPU} [pJ/cyc]	4.9	4.64***	6.3****	**3.2 (@12.5MHz)**
E_{CPU} (norm to Eff)	1	0.67	1.54****	**0.60**
*Upper range limited by test setup, **Estimate based on Intel Atom CPU results, ***Calculated via pie chart, ****includes SRAM				

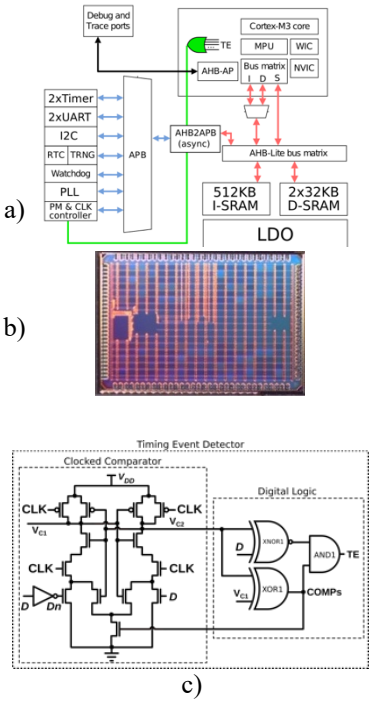

Fig. 1 a) CPU system diagram, b) Chip photograph, c) Monitor schematic

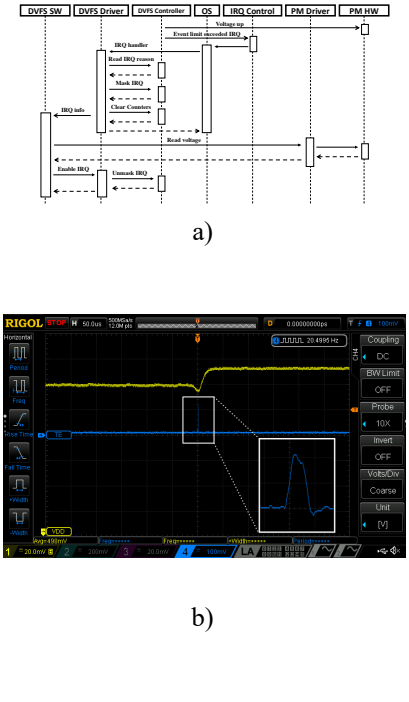

Fig. 2 a) AVS HW/SW scenario when the timing event limit is exceeded, b) AVS loop raises the LDO voltage (in yellow) due to timing event (TE, in blue)

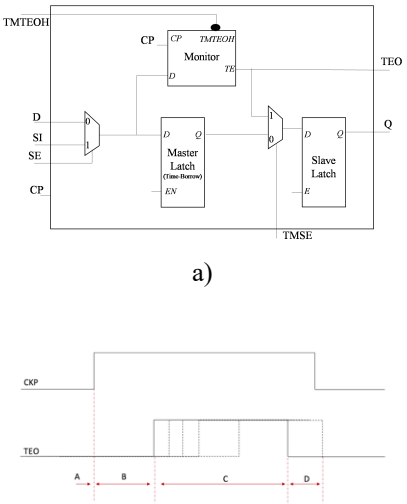

Fig. 3 a) DFT arrangement for extensive testing of the adaptive logic. b) Measurement of the event detection window

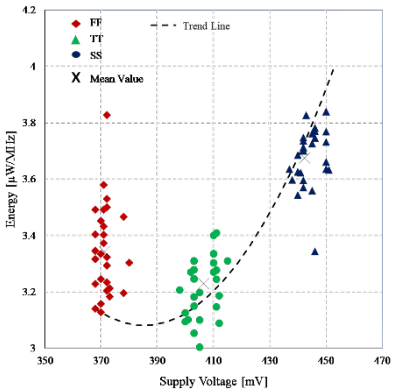

Fig. 4 Energy for all samples at 12.5MHz/20C

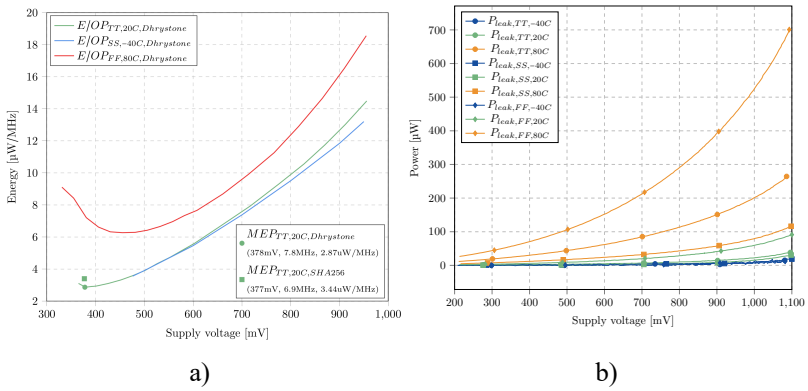

Fig. 5 Measurement of typical, BC and WC conditions, a) Energy, b) Leakage

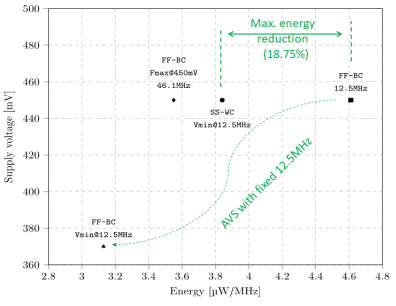

Fig. 6 Minimum energy at equal performance is the difference between FF-BC and SS-WC

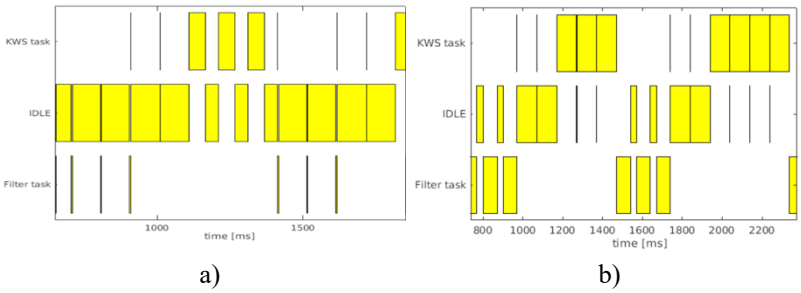

Fig. 7 By using the idle time to increase execution time (time spent shown in yellow) and lowering the operating voltage, the KWS algorithm executed with 59,7% less energy. a) before, b) after