19.1 Computationally Enabled Total Energy Minimization Under Performance Requirements for a Voltage-Regulated 0.38-to-0.58V Microprocessor in 65nm CMOS

Fahim ur Rahman, Rajesh Pamula, Akshat Boora, Xun Sun, Visvesh Sathe

University of Washington, Seattle, WA

Integrated circuits for ultra-low-power applications strive to minimize total system energy, while satisfying performance requirements. The supply voltage (V_{dd}) can be set to a Minimum Energy Point (MEP) [1,2], where leakage and dynamic energy are suitably balanced. However, controlling operating frequency (f_{clk}), while concurrently tracking a MEP sensitive to PVT and switching activity is not possible. Meanwhile, the traditional approach of locking to the minimum required frequency (f_{targ}), and adjusting V_{dd} to maintain timing slack precludes the possibility of minimum-energy computing. Therefore, there exists a need for a minimum-energy computing architecture that meets performance requirements.

Prior work has demonstrated MEP tracking across PVT and switching activity variation [3]. The approach relies on large capacitors for sample-and-hold operation at sub-threshold frequencies, and cannot account for the significant regulator losses (often amounting to 10%–50% of total energy) necessary for total system energy minimization. Furthermore, clock generation using a free-running oscillator is a requirement, precluding any regulation of f_{clk}.

This paper presents a digital architecture for total system energy minimization subject to performance requirements (Fig. 19.1.1). The design supports two modes of operation – *MEP-lock* and *perf-lock* – and seamless, uninterrupted execution during transitions between them. In *MEP-lock*, the design tunes V_{dd} to first search for and then track the minimum total Energy Per Cycle (tEPC) point inclusive of regulator losses. The system clock is generated by a V_{dd}-powered Critical-path Replica Oscillator (CRO), whose clock period matches the system critical path delay across V_{dd}. In doing so, the CRO automatically adjusts f_{clk} to ensure timing slack. This resulting MEP frequency (f_{MEP}) is sampled and compared against f_{targ}. If $f_{targ} > f_{MEP}$, the system is determined to be performance-constrained and seamlessly transitions to *perf-lock*, a state corresponding to phase-locked operation at f_{targ}.

Although the proposed architecture is applicable across regulator types, we demonstrate it here for a 2:1 switched-capacitor (SC) converter powering a Cortex M0 processor and an FFT accelerator (Fig. 19.1.2). The FFT clock can be selectively gated to modulate switching activity. Both *MEP-* and *perf-lock* rely on elastic clocking provided by the CRO to minimize V_{dd} margins for supply-noise and PVT variation. Phase-lock is achieved during *perf-lock*, despite a V_{dd}-powered CRO, by unifying clock and power regulation into a single loop (UniCaP) [4]. The SC converter is subsumed into the PLL and modulates the CRO phase through V_{dd} control to achieve lock without affecting timing slack. V_{dd} is regulated by the SC converter through its output impedance by modulating its switching capacitance (C_{fly}); the frequency of the converter remains fixed [5].

The total energy delivered by the supply to the regulator and the load over one SC cycle (E_{tot}) can be modeled by a single equation that depends only on V_{in}, C_{fly} and V_{dd} (Fig. 19.1.2). Combining E_{tot} with the number of execution cycles that occur within this SC cycle (f_{clk}/f_{SC}), tEPC can be computationally derived:

$$\text{tEPC} = \frac{E_{tot}}{f_{clk}/f_{SC}} = \frac{E_{tot} f_{SC}}{N_{clk} f_{REFCLK}} = \frac{C_{fly}(V_{in}-2V_{dd})V_{in} \cdot f_{SC}}{N_{clk} f_{REFCLK}}, \quad (1)$$

where N_{clk} is the number of CRO cycles that occur within one REFCLK, and f_{REFCLK} is the PLL reference clock frequency.

Transition into *MEP-lock* initially involves a rapid MEP search using sign-gradient descent, followed by fine grained MEP tracking across PVT and load variation (Fig. 19.1.3). In each search step, the system is briefly operated at the search voltage to determine the V_{dd}-dependent variables, C_{fly} and N_{clk}, needed for tEPC comparison. *MEP-ctrl* uses a DAC to set V_{ref}, relying on closed loop operation to determine C_{fly} as needed to regulate V_{dd} to V_{ref}. In this duration, the PLL frequency divider, which also counts V_{dd}-powered CRO clock cycles during each *REFCLK* cycle, provides N_{clk}. Thus, the *MEP-ctrl* module computes $E_{tot}[n]$ and $N_{clk}[n]$ for search iteration *n*. The module exploits convexity in the *tEPC* search space to perform search in four phases, each with successively finer step-size for rapid search. Finding the MEP requires knowledge of relative (not absolute) values of *tEPC* between successive search points. This observation is exploited to avoid computationally expensive division. By noting that *tEPC* is proportional to the ratio of E_{tot} and N_{clk}, the module instead compares $E_{tot}[n] \cdot N_{clk}[n+1]$ vs. $E_{tot}[n+1] \cdot N_{clk}[n]$. Once the MEP is found, tracking is performed by continuously evaluating tEPC at adjacent V_{dd} points and moving to the lowest tEPC point.

The proposed design was fabricated in a low-power 65nm CMOS process (die photograph in Fig. 19.1.7). Computational MEP tracking incurs a 0.043mm^2 area overhead that is further amortized, either in larger designs or through shared use by multiple voltage domains. The approach also presents a 0.5% energy overhead at a nominal frequency of $0.1 f_{REFCLK}$ which can be further reduced in applications requiring less frequent MEP tracking. All measurements were qualified by correct processor and FFT module operation.

Transitioning between *perf-* and *MEP-lock*, and rapidly searching and tracking the MEP is critical to tEPC minimization under performance requirements. Fig. 19.1.4 shows measured V_{dd} waveforms during mode transitions between *perf-* and *MEP-lock*. The system initially operates in *perf-lock* with V_{dd} controlled to lock the CRO to REFCLK. Transitioning into *MEP-lock* first begins with a MEP search, performed using sign-gradient descent in four phases, each performing successively finer V_{dd} adjustments (step-size). Each phase begins with $V_{dd,MEP}$ obtained from the previous search with the larger step size. After finding the MEP with the smallest step size (5mV), the system tracks the MEP, evaluating relative *tEPC* at excursions of ±5mV. Subsequently as $f_{targ} > f_{MEP}$, the system transitions back to *perf-lock*.

Figure 19.1.5 shows measured V_{dd} waveforms during *MEP-lock* mode, while tracking switching activity changes due to intermittent FFT operation. FFT turn-off reduces dynamic energy dissipation, increasing the MEP voltage. The inset shows the resulting 50mV change in the MEP being tracked by the system after recovering from the initial regulator IR surge. Fig. 19.1.6 demonstrates the quality of the computational MEP tracking approach across multiple test chips and variation in temperature and dynamic loading. The computationally derived MEP was benchmarked against the actual MEP, which was found using direct bench measurement. Direct measurement involved sweeping C_{fly} to adjust V_{dd} (Fig. 19.1.3) with 2mV resolution, recording f_{clk}, and measuring the current drawn from the test supply to determine *tEPC* at each point. The figure shows the resulting *tEPC* curves obtained across V_{dd} for the case of varying dynamic load. The computationally derived MEP was found to be within 5mV of directly measured MEP. A similar approach was used to benchmark tracking performance against temperature variation and under nominal conditions across 20 test chips. Computational MEP was within 5mV of direct measurement, equivalent to the resolution of the finest MEP search step.

Figure 19.1.7 shows a die photograph of the test chip and comparison to prior related work. This work represents a low-overhead digital architecture for minimum energy computing while providing performance guarantees. Incorporating MEP tracking within a unified clock and power framework minimizes wasteful V_{dd} guardbands, and enables seamless transition between durations of minimum energy computing and performance constrained computing. This approach is especially conducive to advanced CMOS nodes.

Acknowledgments:
A Loke, T. Zhang, C. Tokunaga, M. Khbeis for valuable inputs. Funded by SRC (TxACE Task 2712.006) and Qualcomm Technologies Inc.

References:

[1] A. Wang and A. Chandrakasan, "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology," *IEEE JSSC*, vol. 40, no. 1, pp. 310-319, 2005.
[2] B. Calhoun, et al., "Modeling and Sizing for Minimum Energy Operation in Subthreshold Circuits," *IEEE JSSC*, pp.1778-1786, 2005.
[3] Y. Ramadass and A. Chandrakasan, "Minimum Energy Tracking Loop with Embedded DC-DC Converter Delivering Voltages Down to 250mV in 65nm CMOS," *ISSCC*, pp. 64–65, 2007.
[4] F. Rahman, et al., "An All-Digital Unified Clock Frequency and Switched-Capacitor Voltage Regulator for Variation Tolerance in a Sub-Threshold ARM Cortex M0 Processor," *IEEE Symp. VLSI Circuits,* pp. 65–66, 2018.
[5] Y. Ramadass, et al., "A Fully-Integrated Switched-Capacitor Step-Down DC-DC Converter with Digital Capacitance Modulation in 45 nm CMOS," *IEEE JSSC*, vol. 45, no. 12, pp. 2557-2565, 2010.

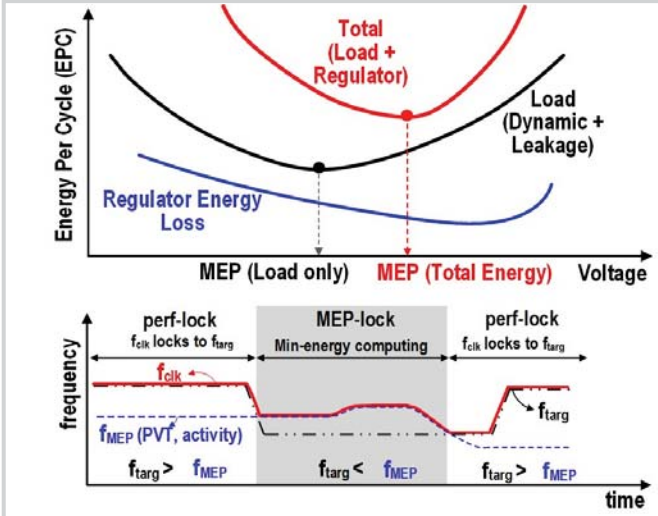


Figure 19.1.1: (Top) The proposed system minimizes total energy dissipation inclusive of VR losses. (Bottom) System-level operation: *perf-lock* mode when $f_{targ}>f_{MEP}$ and seamless transition to *MEP-lock* when $f_{targ}<f_{MEP}$.

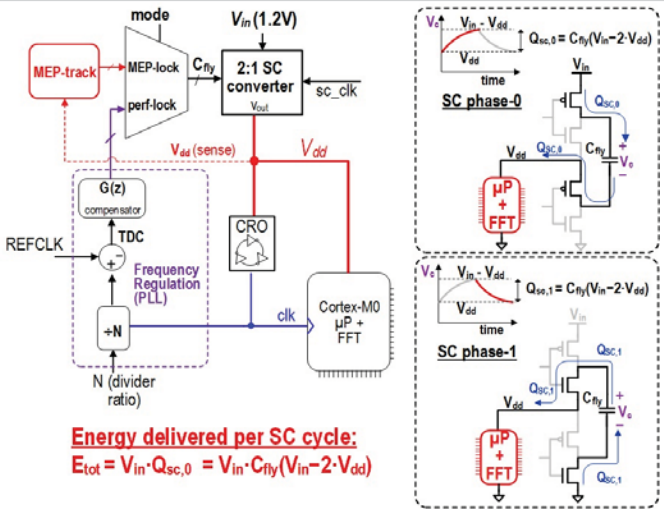


Figure 19.1.2: (Left) Performance-constrained MEP tracking architecture: V_{dd} is controlled by tuning the flying capacitance (C_{fly}). (Right) Variables V_{in}, V_{dd}, and C_{fly} determine E_{tot}.

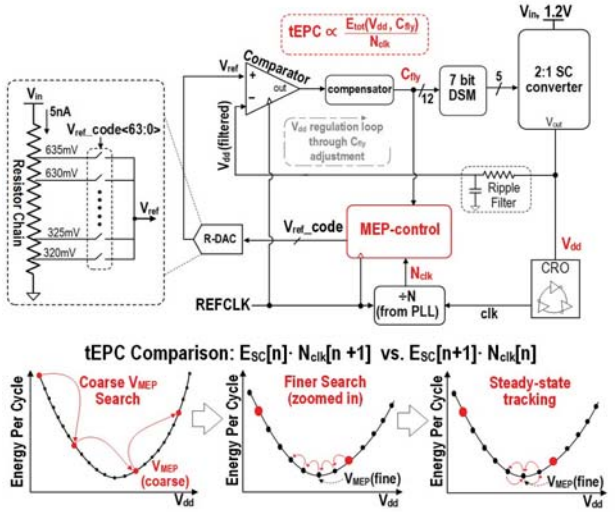


Figure 19.1.3: (Top) Circuit architecture: Closed-loop operation at given search-V_{dd} yields key variables, C_{fly} and N_{clk} for *tEPC* evaluation. (Bottom) MEP search methodology using successively finer search-steps.

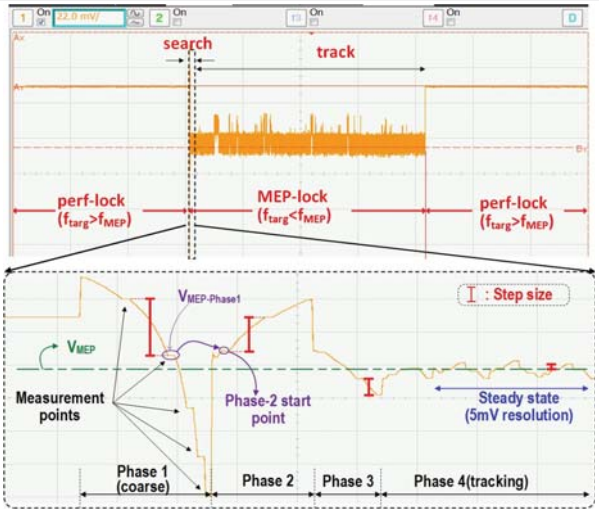


Figure 19.1.4: Measured V_{dd} waveforms during transition between *perf-lock* and *MEP-lock* modes. (Inset) Phases of MEP search, and corresponding V_{dd} transition.

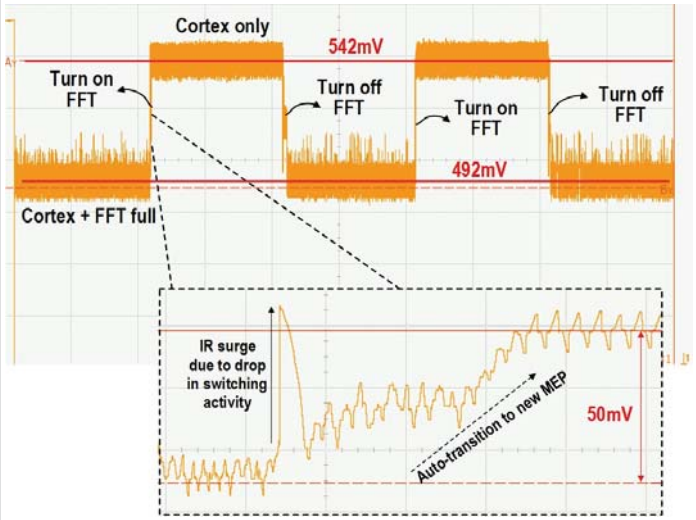
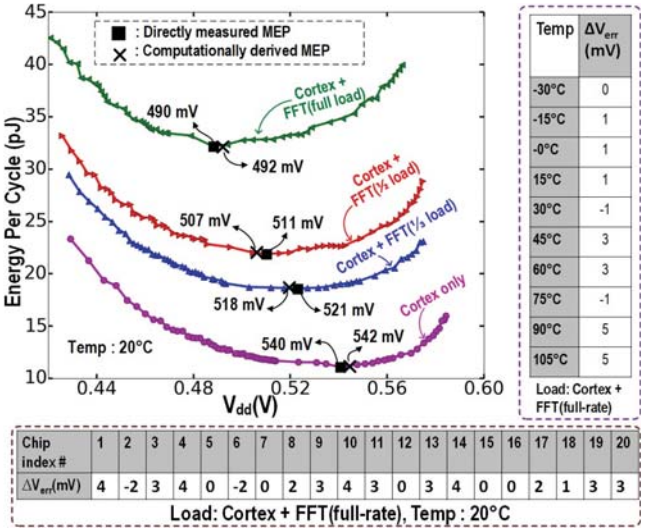


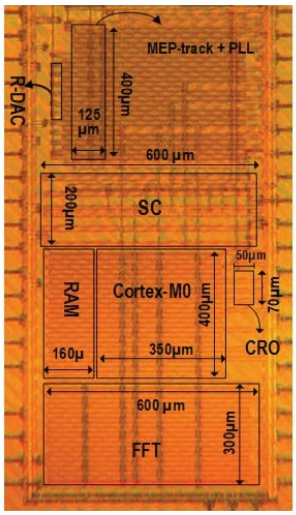
Figure 19.1.5: Measured V_{dd} waveform under MEP-lock during run-time changes in switching activity enabled by FFT on-off operation.



Temp	ΔVerr (mV)
-30°C	0
-15°C	1
-0°C	1
15°C	1
30°C	-1
45°C	3
60°C	3
75°C	-1
90°C	5
105°C	5

Chip index #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ΔVerr(mV)	4	-2	3	4	0	-2	0	2	3	4	3	0	3	4	0	0	2	1	3	3

Figure 19.1.6: Computational MEP-tracking accuracy vs. direct MEP measurement across different workloads, temperature and 20 test-chips.



	This work	ISSCC '07 [3]
Process	65nm LP CMOS	65nm CMOS
Operating voltage	0.38-0.58	0.25-0.7V
Regulator type	Switched-Capacitor	Buck converter
Load	Cortex + FFT	FIR filter
EPC read-out	Computational	Mixed signal
Temp. tracking	Yes	Yes
Load variation tracking	Yes	Yes
Total energy minimization	Yes	No*
Freq. regulation	Yes	No
Perf.-constrained min-energy	Yes	No
MEP-tracking accuracy	≤ 5mV	Not reported
% Energy reduction (load variation only)	18%	Not reported

* Regulator loss not accounted for. However the method relies on nearly constant converter efficiency across V_{dd}

Figure 19.1.7: (Left) Die-photo of the test-chip and (Right) comparison with related work.

978-1-5386-8531-0/19/$31.00