Dynamically Adaptable Pipeline for Energy-Efficient Microarchitectures Under Wide Voltage Scaling

Saurabh Jain, *Member, IEEE*, Longyang Lin, *Member, IEEE*, and Massimo Alioto, *Fellow, IEEE*

***Abstract*— This paper introduces dynamically adaptable pipelines to enable microarchitecture-driven voltage scaling, adapting the microarchitecture to the most energy-efficient configuration for a time-varying throughput target or supply voltage requirement V_{DD}. Dynamic adaptation of the pipeline depth is introduced to curtail the contribution that dominates the overall energy (e.g., dynamic, clock, and leakage), as dictated by the throughput target and V_{DD}. Microarchitectural adaptation of the pipeline depth also flattens the energy dependence on V_{DD} around the minimum-energy point, thus facilitating nearly minimum-energy operation in the presence of inaccuracies in V_{DD} (e.g., discretization and non-idealities). Dynamically adaptable pipelines can be fully integrated with automated digital flows at design time and with dynamic voltage scaling schemes at run time. The proposed approach is demonstrated with a 256-point radix-4 fixed-point FFT engine on a 40-nm test chip. Measurements show energy savings up to 30% (38%) at iso-throughput (iso-voltage) in an area and the maximum performance penalty of 5% and 11%, respectively.**

***Index Terms*— Dynamic pipelining, energy efficiency, microarchitectures, wide voltage scaling.**

I. Introduction

THE energy efficiency in today's systems-on-chip is a crucial design aspect due to their operation in the power-limited regime, as occurs in a wide range of applications from servers to mobile and IoT platforms [1]. Wide voltage scaling has been extensively adopted to reduce the energy consumption over a wide range, whenever performance can be sacrificed [2]–[9]. In wide voltage scaling, a quadratic energy gain is achieved at above-threshold voltages [10], whereas diminishing returns are achieved when V_{DD} approaches the minimum-energy point (MEP), due to the increased contribution of the leakage energy [1], [10]–[13]. Due to the very different leakage-to-dynamic energy ratio across a wide voltage range, the energy-efficient operation requires different microarchitectural optimizations [14] (e.g., choice of pipeline depth).

Manuscript received July 23, 2017; revised September 29, 2017; accepted October 26, 2017. Date of publication November 20, 2017; date of current version January 25, 2018. This paper was approved by Associate Editor Vivek De. This work was supported in part by the Intel Corporation and in part by the Singaporean Ministry of Education under Grant MOE2014-T2-2-158. *(Corresponding author: Saurabh Jain.)*

The authors are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: saurabhj@u.nus.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2017.2768406

Dynamic pipeline stage unification has been previously investigated as form of microarchitectural reconfiguration at the pipeline stage level to save energy by reducing the pipeline depth at nominal voltage [15]–[20], or over a narrow range of above-threshold voltages [21]. In these simple cases, a reduced pipeline depth invariably leads to fewer registers (i.e., lower clocking energy) and less frequent hazards (i.e., lower energy per instruction) in microprocessor systems, although at the cost of degraded throughput due to shallower pipelining [15]–[21]. Among the limitations of prior art on pipeline stage-level reconfiguration, the interaction of microarchitectural reconfiguration and wide voltage scaling has not been explored due to operation at fixed or narrowly scaled voltage range, which in turn assures that the same architecture is energy optimal across such voltages, and hence makes any microarchitectural dynamic co-optimization with V_{DD} irrelevant. As second limitation, prior art explores very limited energy gains, while ignoring minimum-energy operation from a voltage scaling and a microarchitectural reconfiguration viewpoint. As third limitation, no prior art addresses the design challenges associated with joint microarchitectural and voltage scaling, as the only prior chip demonstration is limited to an intrinsically very regular architecture [17], which ignores several fundamental design issues such as automated and optimal insertion of microarchitectural reconfiguration, pipestage balancing, and design for energy optimality across a wide range of voltages.

This paper introduces dynamically adaptable pipelines to enable energy reductions through joint microarchitectural and voltage co-optimization, for voltages ranging from nominal voltage down to deep sub-threshold. As side benefit, the sensitivity of the energy to supply voltage deviations around the MEP is mitigated, thus making the pursuit of nearly minimum energy easier under inaccurate or discretized voltage generation. Techniques to insert adaptable pipeline registers are introduced in an automated fashion at design time, and microarchitectural reconfiguration is fully integrated with voltage scaling schemes at run time.

This paper is structured as follows. In Section II, the concept of dynamically adaptable pipelines is motivated and analyzed in terms of its impact on energy. Section III discusses the related design aspects, including an automated design flow for general microarchitectures and integration with conventional dynamic voltage scaling. Section IV discusses the FFT test

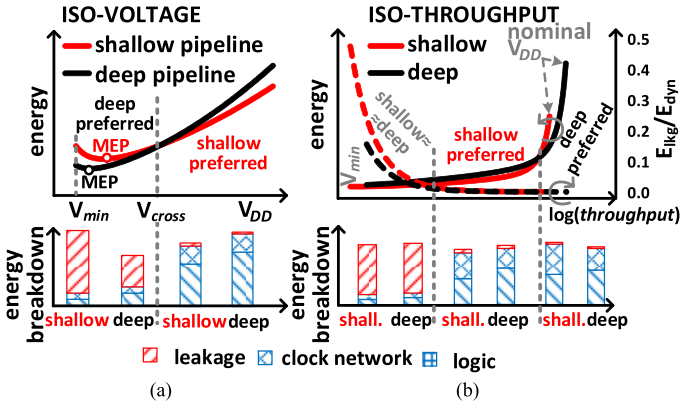


Fig. 1. (a) Qualitative trend of energy versus V_{DD} for shallow and deep pipelines, with the minimum-energy point being labeled as MEP. (b) Qualitative trend of energy versus throughput for shallow and deep pipelines. In (b), V_{DD} scaling is implicitly applied to scale the throughput. The crosspoint voltage at which the energy curves cross each other is generally different in (a) and (b).

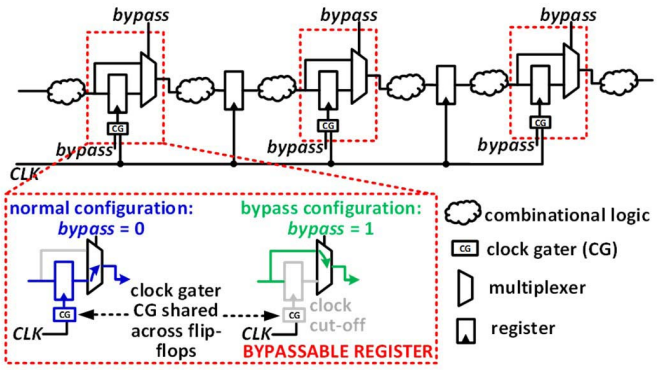


Fig. 2. Proposed dynamic adaptable pipeline.

chip demonstrating the concept, and Section V presents the measurement results. Section VI provides the conclusion.

II. Impact of Pipeline Depth on Energy Under Wide Voltage Scaling and Related Properties

The energy per cycle E_{cycle} of digital synchronous designs is due to the main contributions of the dynamic energy E_{dyn} and leakage energy E_{lkg} [14], [22]

$$E_{\text{cycle}} = E_{\text{dyn}} + E_{\text{lkg}} = \alpha_{\text{sw}} \cdot C_{\text{tot}} \cdot V_{\text{DD}}^2 + V_{\text{DD}} \cdot I_{\text{off,tot}} \cdot T_{\text{CK}} \tag{1}$$

where α_{sw} is the activity factor, C_{tot} is the switched capacitance including the clock distribution, $I_{\text{off,tot}}$ is the total leakage current drawn by the design, and T_{CK} is the minimum clock period.

In the above-threshold region, the leakage energy in (1) is typically a small fraction of E_{cycle}, whereas it rapidly increases in the near- and sub-threshold region due to the rapid increase in T_{CK}. As discussed later, such different $E_{\text{lkg}}/E_{\text{dyn}}$ ratio in (1) across voltages imposes very different requirements in terms of pipeline depth that cannot be satisfied by a single microarchitecture, when pursuing nearly minimum-energy operation under wide voltage scaling. In the sub-threshold region, the leakage energy becomes dominant, and it is higher in shallow microarchitectures because of its larger clock cycle T_{CK} from (1). In other words, shallow pipelines are more energy efficient at above-threshold voltages, whereas deep pipelines are more energy efficient at low voltages, when compared at iso-voltage (e.g., when V_{DD} is set by design considerations involving other blocks sharing the same voltage domain). Also, due to the lower E_{lkg}, the MEP of deep pipelines is pushed to the left of the MEP of shallow pipelines [22], [23], thus extending the voltage region under which V_{DD} scaling enables true energy reduction.

From Fig. 1(a), the adoption of a fixed microarchitecture cannot achieve the best energy efficiency at all voltages, as it would be desirable to adopt shallow pipeline at large voltages and switch to deeper pipeline at lower voltage to gain both in terms of energy efficiency (thanks to the reduced dominant leakage energy) and voltage range where voltage scaling is advantageous. As side benefit, the ability to switch to deeper pipelines at low voltages permits to mitigate the otherwise rapid energy increase when V_{DD} is pushed to the left of the MEP, as it might occur due to inaccuracies or discretization in the voltage generation (see details in Section V).

Fig. 1(b) qualitatively compares the energy trend for deep and shallow pipelines at iso-throughput, which is obtained through individual V_{DD} adjustment (as occurs when the block under design is in a separate voltage domain). At above-threshold voltages, the throughput excess in deep pipelines allows for more aggressive voltage scaling, which translates into lower energy than shallow pipelines, due to the dominance of E_{dyn}. At lower throughput targets and hence lower V_{DD} (e.g., near threshold), the throughput becomes much more sensitive to V_{DD}, hence the shallow pipeline can achieve the same throughput as the deep one through a very limited voltage increase. In this case, the operating voltage of deep and shallow pipelines at iso-throughput is almost the same, and their energy essentially differs only for the lower number of registers in shallow pipelines, which hence becomes more energy efficient than the deep one as in Fig. 1(b). Similarly, at extremely low throughput targets, the resulting voltage of deep and shallow pipelines is again nearly the same and the energy is dominated by leakage, which is essentially the same for both, thus making their energy nearly the same (i.e., both configurations are equally energy efficient). Again, it would be desirable to adjust the pipeline depth by switching from deep to shallow when scaling down the throughput target, as a single microarchitecture cannot be energy optimal across a wide throughput range.

The above considerations motivate the adoption of dynamically adaptable pipelines, where the logic depth per pipestage can be adjusted along with the voltage, to minimize the energy across a wide range of throughputs and voltages.

III. Dynamically Adaptable Pipelines for Run-Time Co-Optimization of Pipeline Depth and V_{DD} Under Wide Voltage Scaling

Fig. 2 schematizes a generic microarchitecture with dynamically adaptable pipeline, as enabled by the introduction of

bypassable registers, which can operate in bypass and normal modes depending on the value of the *bypass* signal. In the bypass mode, *bypass* is set at 1 so that the multiplexers embedded in these registers bypass the corresponding flip-flops, and the shared clock gater disables their clock to suppress their power. These multiplexers effectively merge the previous and the subsequent pipestage into a single one, thus increasing the logic depth. In the normal mode (i.e., *bypass* = 0), the multiplexer selects the flip-flop output and the clock gater enables the clock, thus leading to conventional register operation and lower logic depth. This approach does not modify the clock network, and hence maintains the same clock load in all modes.

In a dynamically adaptable microarchitecture, maximum flexibility in terms of achievable logic depth would be obtained by making all registers bypassable. However, this would be impractical since bypassed registers add a delay and energy overhead due to their embedded multiplexer. Also, this would make pipeline stage balancing more difficult for commercial electronic design automation (EDA) tools performing retiming, thus further increasing the delay overhead. In our experiments, we observed that such overhead is reasonably small when flexibility in the logic depth adjustment is limited to every other pipeline stage (i.e., the logic depth can be changed by a factor of approximately 2×, but not larger than this). In other words, a reasonable balance between microarchitectural flexibility and energy benefits is to insert bypassable registers only in odd-numbered register levels[1] as in Fig. 2, while leaving even-numbered registers as conventional.

A. Design Time: Automated Flow for Dynamically Adaptable Pipelines

From the above considerations, a conventional microarchitecture can be turned into a dynamically adaptable microarchitecture by replacing odd-numbered register levels with bypassable registers, and balancing logic in adjacent pipestages in both normal and bypass modes (i.e., for both logic depths achieved in the two modes). To be generally applicable, the register replacement process and timing optimization need to be automated, and based on commercial EDA tools. A general design flow to design balanced dynamically adaptable pipeline microarchitectures is introduced in the following (see Fig. 3), and is integrated with commercial tools through a set of scripts. As preliminary step, the logic depth required to achieve the desired cycle time (i.e., throughput) at a given feasible voltage is estimated by parsing the synthesis timing report, and dividing the overall combinational delay across the pipestages by the targeted combinational delay per pipeline stage (which in turn determines the throughput). As first step, if the original design was not able to meet the throughput, the required additional pipeline registers are inserted from the primary inputs or outputs, via simple manipulation of the gate-level netlist generated by the synthesis tool (step 1 in Fig. 3). These steps are skipped if the original design meets the throughput requirement at nominal voltage.

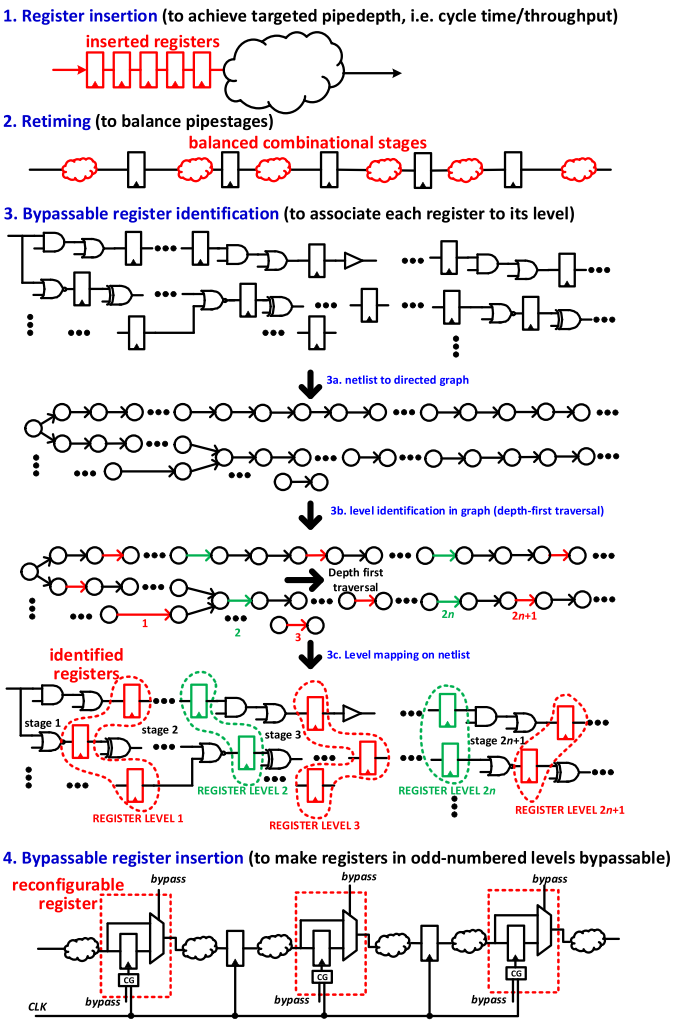


Fig. 3. Automated design flow for bypassable register insertion.

As second step (step 2 in Fig. 3), the gate-level netlist enhanced with the additional pipeline registers is retimed through any commercial synthesis tool supporting register retiming, so that pipeline stages are balanced in terms of delay. The input and output delay/load constraints are kept the same as the original design, whereas the cycle time target is updated to the above mentioned value that assures the throughput target. The retiming is performed by using the cell characterization under the above mentioned voltage, at which the throughput is specified. In this paper, the cell library was characterized from 1.1 V down to 0.3 V with a step of 0.1 V. In the example in Fig. 3, five registers have been inserted and then retimed to achieve six balanced pipeline stages. The resulting design is the conventional deep pipelined version of the design, as the successive selective replacement of registers with their bypassable counterpart will only increase the logic depth per pipeline stage.

As third step, registers need to be grouped based on the level they belong to. This step is necessary because retiming at step 2 spreads flip-flops in a flattened netlist and renames them randomly, thus completely losing the association between

[1]In general, the register bypass choices that keep pipestages balanced bypass an integer number m of consecutive registers, followed by a non-bypassed pipestage. Among these potential choices, the bypass of odd-numbered levels (i.e., $m = 1$) in shallow mode minimizes the number of bypassable registers, which in turn minimizes the area, delay, and leakage overhead associated with the multiplexers added to the bypassable flip-flops (see Fig. 2).

flip-flops and the pipeline stages they belong to (or the related signals). Accordingly, it is necessary to associate again every flip-flop to the corresponding register level, progressing from inputs (i.e., the first level of registers) to outputs (i.e., the last level of registers). To this purpose, the gate-level netlist is first parsed and converted to a directed graph (step 3a in Fig. 3), where each wire in the gate-level netlist is represented as a node in the graph, and each pair of input/output pins in every cell is represented as a graph edge. On the graph, depth-first traversal is performed to progressively identify the level of each flip-flop from the first to the last one, as shown in step 3b in Fig. 3. The above steps are simultaneously performed in all registers in linear and feedforward paths, whereas loops of registers are preliminarily identified and are separately dealt with. Indeed, the proposed approach is subject to the same limitations that are imposed by conventional (static) pipelining. After traversal, the list of flip-flops belonging to every register level is obtained.

As fourth step, the flip-flops that were found to belong to odd levels are then replaced by reconfigurable flip-flops (see Fig. 2), i.e., adding multiplexers and shared clock gaters to the gate-level netlist. After step 4, the original microarchitecture is turned into a dynamically adaptable pipeline with balanced pipeline stages. Observe that the pipeline stages were balanced with registers not being bypassed at step 2 (i.e., the microarchitecture is reconfigured for minimum logic depth), as relevant to the case when higher performance is required. To this aim, all paths going through the multiplexers of reconfigurable flip-flops have been set as false paths during synthesis, placement and routing. Otherwise, the tool would have naturally optimized the design for shallow configuration since it would have considered also the paths going through the multiplexers, thus completely skipping the balance between pipestages in the deep pipeline configuration. Once pipestages in the deep pipeline configuration are balanced, the worst-case delays of pipestages in the shallow configuration are automatically balanced among each other, although they are less critical than the deep configuration (which targets higher performance). From the resulting gate-level netlist, placement and routing are performed conventionally by using the same timing, delay and load constraints at step 2. Throughput and energy at lower voltages can be then evaluated by performing gate-level power and timing analysis, based on the cell characterization at the relevant voltages.

B. Run Time: Integration of Dynamically Adaptable Pipelines With Dynamic Voltage Scaling

The above discussed microarchitectural adaptation to the voltage and/or the throughput target at run time can be straightforwardly integrated with conventional dynamic voltage scaling schemes. Indeed, the analysis in Section II showed that the configuration with maximum logic depth [shallow microarchitecture in Fig. 1(a)] is preferred at voltages ranging from the crosspoint voltage V_{cross} to the nominal voltage, whereas the deep microarchitecture should be adopted at $V_{\text{DD}} < V_{\text{cross}}$ for a given voltage. The opposite choice is preferable when V_{DD} can be independently optimized to minimize the energy and meet a given throughput constraint [see Fig. 1(b)].

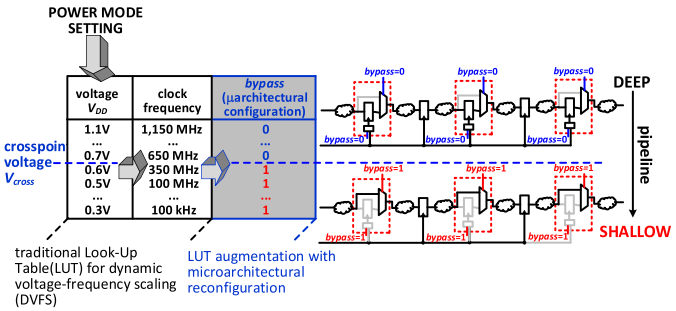


Fig. 4. Integration of microarchitectural reconfiguration in conventional dynamic voltage scaling schemes and run-time co-optimization, under a given V_{DD} [i.e., as derived from considerations in Fig. 1(a)]. The choice of deep/shallow microarchitecture is swapped when V_{DD} is optimized to achieve a given throughput [as derived from considerations in Fig. 1(b)].

According to the above considerations, the scheme in Fig. 4 can be adopted to jointly scale the voltage and reconfigure the microarchitecture while pursuing energy minimization. The voltage–frequency lookup table (LUT) that is conventionally used for dynamic voltage scaling is extended to include the appropriate value of *bypass*, and hence the microarchitecture. In this way, operation at a given voltage [see Fig. 1(a)] sets both the clock frequency and the logic depth of the microarchitecture (i.e., deep if $bypass = 0$, shallow if $bypass = 1$). Alternatively, operation at a given throughput [see Fig. 1(b)] sets the voltage and the microarchitecture.

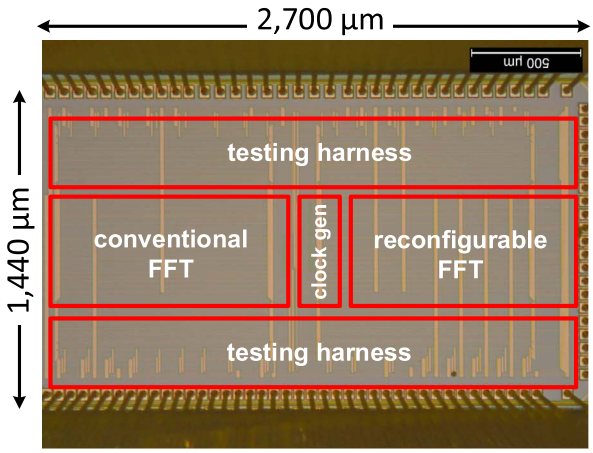
The crosspoint voltage below which the microarchitecture is dynamically changed can be determined at design time, by comparing the energy consumption of the two microarchitectural configurations across benchmarks. When the voltage regulator powering the module under design has the capability to measure power, the choice between the two microarchitectural configurations can be made for the specific workload at hand.

IV. Test Chip Design

The above concept of dynamically adaptable pipelines was experimentally evaluated through a test chip implementing a radix-4, 256-point, 16-bit input, fixed-point complex FFT engine with a modified multi-path delay commutator (MDC) architecture [24] (see die photograph in Fig. 5). The architecture of the FFT engine is summarized in Fig 6. The FFT engine was implemented in both a conventional design with fixed-pipeline depth and a reconfigurable design with dynamically adaptable pipeline.

The targeted throughput was required to cover and exceed the wide range required by practical applications, from low (∼1 MS/s) to high performance (∼4–5 GS/s) [26], [27]. The minimum throughput is achieved through aggressive voltage (and energy) scaling, as usual in wide voltage scaling schemes. The maximum throughput demands a clock period of approximately 1 ns, which translates into a logic depth of 33Fan out of 4 (33FO4) per pipestage[2] at a nominal voltage of 1.1 V. Design at nominal voltage for maximum throughput is adopted at step 1 in the design flow in Fig. 3. Lower voltages

[2]FO4 is the fan-out-of-4 delay, i.e., the delay of an inverter gate driving four equal inverter gates. In the adopted technology and transistor flavor, FO4 is 27.8 ps at typical process corner and nominal voltage.



	conventional FFT	reconfigurable FFT	
gate count	182k	230k	
area	0.53mm^2	0.53mm^2	
technology	40nm LP	40nm LP	
V_{min}	0.34V	0.34V	
		deep	shallow
nominal frequency @ 1.1V	657MHz	971 MHz	571 MHz
logic depth	60 *FO4*	33 *FO4*	64 *FO4*

Fig. 5. Test chip micrograph and related data.

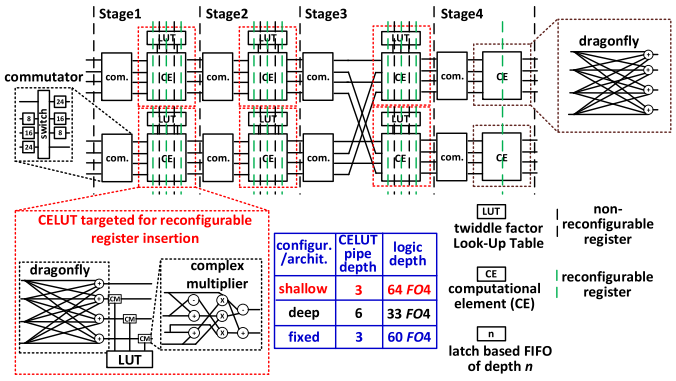


Fig. 6. Block diagram of 256-point 16-bit radix-4 fixed-point complex FFT with modified MDC architecture [24].

cannot meet the maximum throughput target since logic depth lower than 33FO4 would be required, which is not feasible for the reasons discussed below.

The critical path of the above architecture lies in the computational element (CE) of the FFT architecture in Fig. 6, which includes a fixed-point complex multiplier and three fixed-point complex adders. Accordingly, the bypassable registers were inserted in the CELUT block containing the CE and the LUT (which is totally unrelated to the voltage-frequency LUT in Fig. 4). In the dynamically adaptable pipelined design, the CELUT belonging to first, second, and third stages was pipelined with six stages in deep configuration as in steps 1 and 2 in Fig. 3, resulting into a logic depth of 33FO4 per stage. Since the last CELUT actually does not contain the LUT and the complex multiplier, it was divided in only two pipeline stages as shown in Fig. 6 (see green dashed line). As mentioned above, more aggressive eight-stage pipelining or higher would be unfeasible since it would make the CELUT logic depth 25FO4 or lower, shifting the critical path from the CELUT to the peripheral circuitry for the first-in first-out (FIFO) read-out. The latter contains feedback paths that would make reconfiguration difficult, and would limit the performance and energy gain under deep pipelining. In addition, adoption of logic depths of 25FO4 and lower would make the design very sensitive to variations, and their counteraction would require the adoption of non-trivial solutions [24] (e.g., latch clocking, aggressive hold fix buffer insertion).

Once the initial fixed-pipeline design is defined as above (step 1 in Fig. 3), bypassable registers were inserted to replace conventional registers in odd-numbered register levels as in steps 3 and 4 in Fig. 3. When bypassing registers, the logic depth in the shallow configuration turned out to be 64FO4 per stage, which is about twice the logic depth of the deep configuration as expected. The ratio of the delay of the deep and shallow microarchitecture is not exactly two because of the slightly different gate sizing discrete optimization performed by the synthesis tool, due to the presence of the additional multiplexers in bypassable registers. The conventional FFT engine was designed to achieve approximately the same logic depth as the shallow configuration of the dynamically adaptable pipelined design. The effective logic depth of the fixed-pipeline design results to 60FO4 and is 4FO4 lower than the shallow configuration of the dynamically adaptable pipelined design, as explained by the absence of the multiplexers that were instead inserted in the latter.

V. Measurement Results

Fig. 7(a) and (b) shows the energy consumption for the two configurations of the FFT engine with dynamically adaptable pipelined microarchitecture at margined clock frequency determined by the worst-case process corner, and considering a 5% and 10% voltage margin, respectively. As expected from Section II, the deep configuration is more energy optimal than the shallow configuration at higher throughput targets. In detail, this occurs at a throughput of 3.6 GS/s for a 5% voltage margin (3.8 GS/s for a 10% voltage margin) and above, which translates into a clock frequency of 450 MHz and above, since the considered architecture delivers eight FFT samples per cycle. For such throughput targets, transistors operate in the above-threshold region, and the energy gain of the deep configuration over the shallow one is up to 25% as shown in the inset of Fig. 7(a) and (b). The maximum throughput at nominal voltage is 6.3 GS/s. Conversely, the shallow configuration is energy optimal at lower throughputs ranging from 100 MS/s to 3.6 GS/s (3.8 GS/s) for a 5% (10%) voltage margin, and the energy improvement over the shallow configuration is up to 30%, as shown in the inset of Fig.7(a) and (b). Throughputs of hundreds of MS/s or more are achieved at above-threshold voltages, whereas near-threshold operation takes place for throughputs of several MS/s to a few hundreds of MS/s. At lower throughputs on the order of MS/s, the leakage energy is dominant and similar in

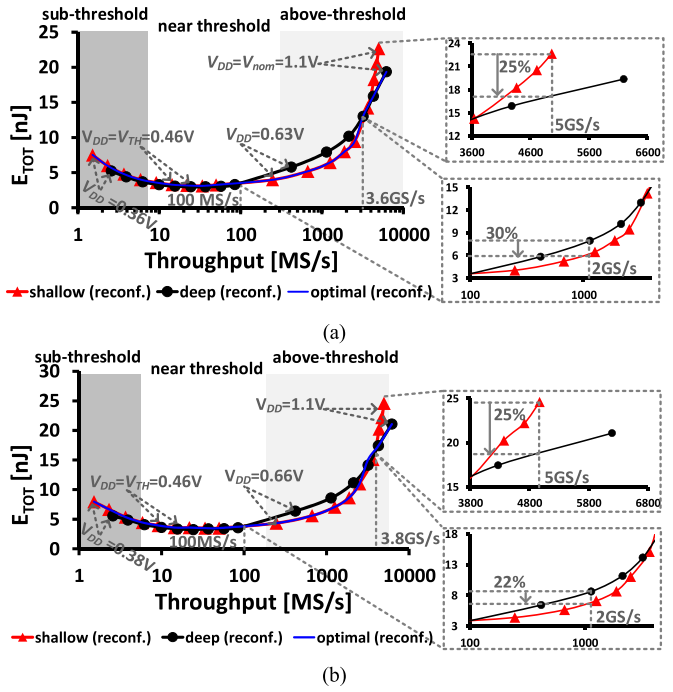


Fig. 7. Energy versus throughput in the reconfigurable microarchitecture design for the shallow/deep configurations, with frequency set by the worst-case process corner under (a) 5% supply voltage margin and (b) 10% voltage margin. The optimal configuration is chosen as the most energy-efficient microarchitectural option for each throughput target.

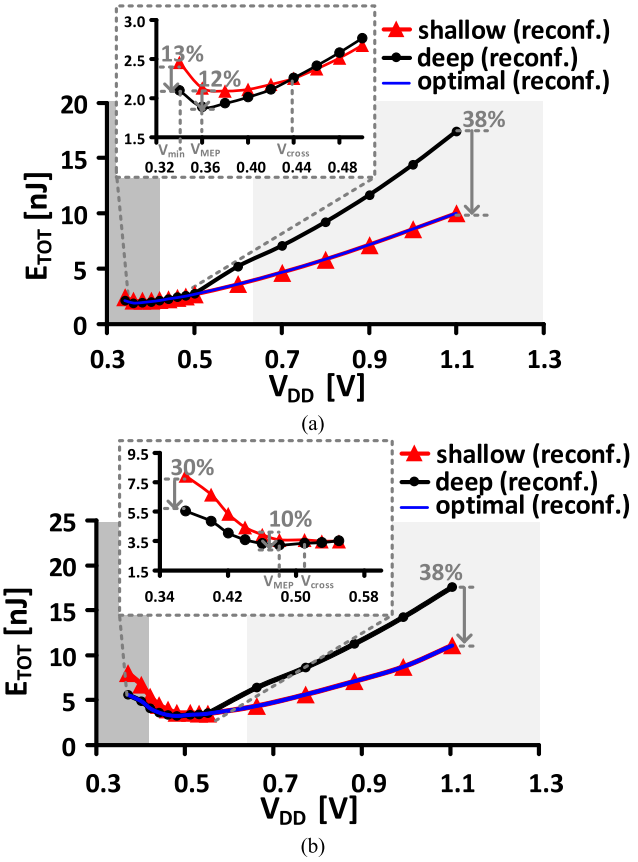


Fig. 8. Energy versus V_{DD} at (a) maximum frequency for shallow and deep configurations and (b) frequency margined for worst-case process corner and 10% V_{DD} margin.

both configurations, thus yielding similar energy at a given throughput as expected from Fig. 1(b).

Fig. 8(a) shows the energy comparison at iso-voltage for the shallow and deep configurations, as well as for the optimal configuration that corresponds to the best choice of the two for each voltage, as enabled by the run-time microarchitectural reconfiguration. From Fig. 8(a), the shallow configuration is energy optimal at V_{DD} larger than $V_{cross} = 0.44$V as qualitatively expected from Fig. 1(a), whereas the deep configuration exhibits lower energy at lower voltages. The energy gain at the minimum voltage $V_{min} = 0.34$ V, the minimum-energy voltage $V_{MEP} = 0.36$ V, and the nominal voltage $V_{nom} = 1.1$ V are 13%, 12%, and 38%, respectively. When the frequency is margined to account for the worst-case process corner and 10% voltage margin, V_{cross} increases to 0.51 V, and the energy savings at $V_{min} = 0.36$ V, $V_{MEP} = 0.48$ V, and $V_{nom} = 1.1$ V become 30%, 10%, and 38% as shown in Fig. 8(b).

The above measurements were repeated on 18 dice, and the resulting distribution of the clock frequency at 0.5, 0.6, and 0.8 V is shown in Fig. 9(a). From Fig. 9(a), the average ratio between the clock frequency of the deep and shallow configurations is 1.7–1.8×, which is close to the value of 1.9× expected from simulations (i.e., the deep configuration is almost 2× faster than the shallow one). As expectable, the clock frequency variability increases when V_{DD} approaches the threshold voltage ($V_{TH} = 0.46$ V at $V_{DD} = 1.1$ V for the adopted technology), and is, respectively, 12% and 8% for deep and shallow configuration at $V_{DD} = 0.5$ V. Also, the clock frequency variability of the deep configuration is 1.4–1.5×, the variability of the shallow one, as expected from the fact that the frequency variability is proportional to the square root of the logic depth [28]–[30].

The histogram of the voltage and the energy at the minimum-energy point is plotted in Fig. 9(b) and (c). From Fig. 9(b) and (c), the MEP of the deep (shallow) configuration lies in the 0.48–0.51 V (0.53–0.55 V) voltage range, and hence in the near-threshold region since $V_{TH} = 0.46$ V. Fig. 9(c) shows that the deep configuration at the MEP voltage offers 9% energy reduction on average and 15% in the best case across the tested dice. Comparison of Fig. 9(b) and (d) shows that V_{cross} lies in the near-threshold region and is always at the right of the MEP of the deep configuration and close to the MEP of the shallow configuration. Occurring at the right of V_{cross} (see Figs. 7 and 8), the maximum energy reduction is achieved in the above-threshold region, and at voltages that are above the minimum-energy point [Fig. 9(e)]. Overall, Fig. 9(b)–(e) shows that the MEP voltage, V_{cross}, and the voltage $V_{max,gain}$ at which the energy gain is maximum are consistent across dice.

The microarchitectural reconfigure ability comes at an energy, performance and area overhead due to the insertion of bypassable registers, compared to a conventional fixed pipeline. To evaluate such overhead, the dynamically adaptable pipelined FFT engine in its shallow configuration was compared with the fixed-pipeline version on the same test

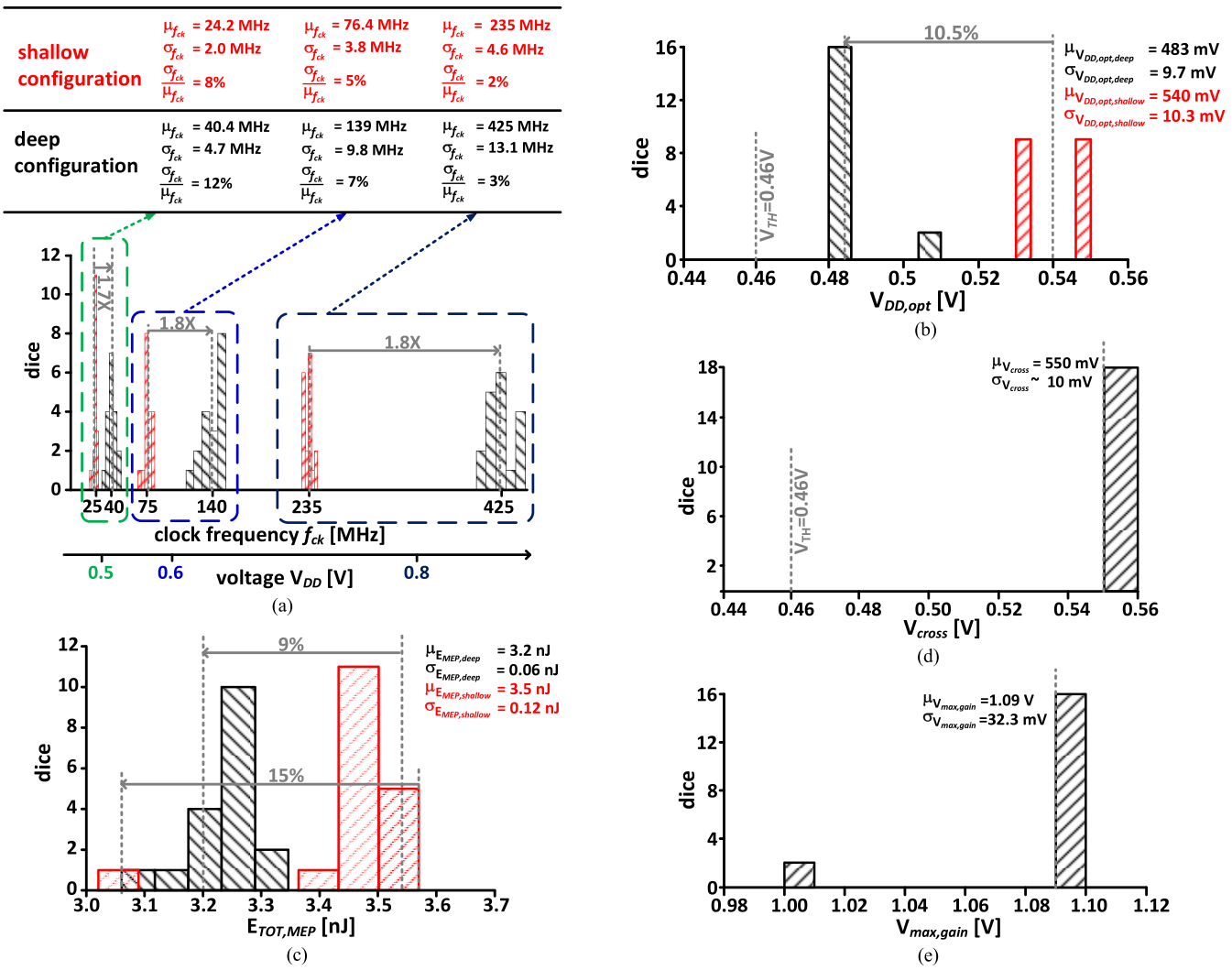


Fig. 9. (a) Histogram of maximum clock frequency f_{CK} for deep and shallow configurations at 0.5, 0.6, and 0.8 V across 18 dice. (b) Histogram of energy-optimal voltage $V_{DD,opt}$ at the minimum-energy point across 18 dice. (c) Histogram of energy at the minimum-energy point across 18 dice. (d) Histogram of V_{cross} across 18 dice. (e) Histogram of voltage $V_{max,gain}$ at which the maximum energy gain occurs across 18 dice.

chip, as they were both designed to have essentially the same logic depth and timing constraints. The clock network of both the reconfigurable and the fixed-pipeline version was designed by relying on the clock tree synthesis flow of a well-known commercial EDA tool, using the same target clock skew and slew.

Fig. 10(a) shows the energy of both versions of the FFT engine compared at iso-throughput,[3] along with the energy penalty entailed by the reconfiguration. From Fig. 10(a), the energy overhead at nominal voltage is 5.7%, whereas it is 7.9% at the throughput target of 700 MS/s (i.e., at 0.6 V). Results are consistent across the 18 considered dice, whose energy under the same conditions is plotted in the same figure and shows that the average energy of the fixed-pipeline design is 3.3 nJ, whereas the average energy of the shallow configuration is 3.6 nJ, leading to an average energy overhead of 9.1% across dice [see Fig.10(b)]. The energy overhead due to reconfiguration reaches the maximum value of 10.5% at $V_{min} = 0.34$ V, and its overall value averaged across voltages and dice is 6%.

Fig. 10(c) shows the throughput versus V_{DD} and the penalty due to reconfiguration, which is mainly due to the extra multiplexers in the bypassable registers. The performance penalty can be as high as 10.8% at a voltage of 0.6 V, and it decreases to 6.7% when averaged across the 18 considered dice at the same voltage [see Fig. 10(d)]. Overall, the performance penalty averaged across voltages and dice is 6.7%. As shown in Table I, the area overhead due to reconfiguration is 3.1%, when fairly comparing the dynamically adaptable design with the fixed deep pipeline version (which was not implemented in the test chip, but was placed and routed with the same timing constraint as the dynamically adaptable pipeline design in its deep configuration).

Fig. 10(e) summarizes the measured energy distribution for the reconfigurable and the fixed microarchitecture. The clocking energy was measured by disabling the FFT logic,

[3]Simulations show 5% maximum throughput degradation compared to the fixed deep pipeline, due to the additional delay of the multiplexers inserted in bypassable registers.

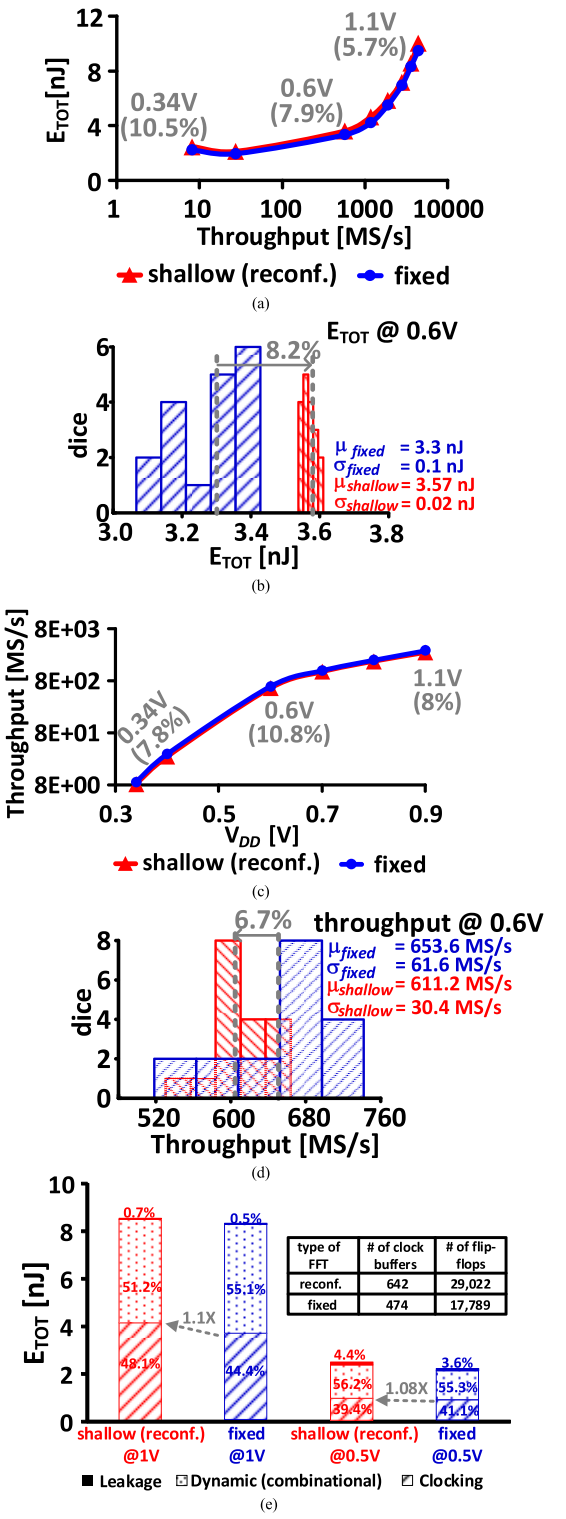


Fig. 10. (a) Energy versus throughput and percentage energy overhead due to reconfiguration at three-pipestage compared to conventional fixed three-pipestage microarchitecture (as indicated in parentheses). (b) Energy distribution across 18 dice at 0.6 V. (c) Maximum throughput versus V_{DD} and percentage performance penalty over conventional fixed microarchitecture due to reconfiguration at iso-voltage (as indicated in parentheses). (d) Throughput distribution across 18 dice at 0.6 V. (e) Energy distribution per cycle compared to conventional fixed microarchitecture.

while still running the clock. The clocking energy contributes to 48.1% (44.4%) of the total energy budget in the reconfigurable (fixed) microarchitecture. The clocking energy overhead

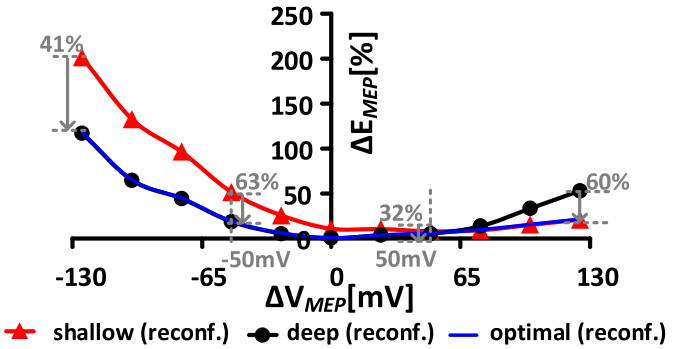


Fig. 11. Percentage energy increase ΔE_{MEP} with respect to the MEP energy in the optimal configuration of the reconfigurable microarchitecture versus voltage deviation ΔV_{MEP} with respect to the MEP voltage. The adoption of the optimal configuration extends the flat energy region toward the left (right), compared to the single shallow (deep) configuration.

TABLE I

AREA OVERHEAD DUE TO RECONFIGURATIONS AND AREA BREAKDOWN BY TYPE OF CELL (SEQUENTIAL CELLS = FLIP-FLOPS + FIFO LATCHES)

FFT	# of sequential cells	# of combinational cells	area (mm^2)
fixed (shallow)	32k	150k	0.270
fixed (deep)*	43.3k	179k	0.384
reconf.	43.3k	188k	0.396
overhead (reconf. to fixed(deep))	0%	5%	3.1%

* evaluated from P&R

was measured to be 10% (8%) at 1 V (0.5 V), and is mainly contributed by the extra clock buffers in the clock network of reconfigurable microarchitecture.

Microarchitectural reconfiguration brings the additional benefit of reducing the energy sensitivity to V_{DD}, when operating around the MEP. This is shown in Fig. 11, which plots the percentage energy increase ΔE_{MEP} with respect to the MEP when the voltage deviates from the correct MEP voltage by ΔV_{MEP}, which represents the effect of inaccuracies and discretization in the voltage generation circuitry. Due to the exponential leakage energy increase at the left of the MEP [22], relatively small voltage deviations lead to a substantial energy increase. However, when V_{DD} is lower than the MEP voltage, the optimal configuration in the dynamically adaptable design is the deep one and mitigates the energy degradation by 63% (41%), compared to the fixed shallow configuration when ΔV_{MEP} is -50 mV (-125 mV) as shown in Fig. 11. When V_{DD} is greater than the MEP voltage, the optimal configuration is the shallow one and offers 32% (60%) lower energy than the fixed deep configuration when ΔV_{MEP} is 50 mV (125 mV).

For completeness, Table II shows that the FFT engine with the proposed microarchitectural reconfiguration achieves the lowest energy per FFT computation of 1.88 nJ, when the energy-optimal configuration is adopted at its MEP voltage $V_{MEP} = 0.36$ V, compared to the state of the art [24]–[27]. For the sake of fairness, the energy consumption of the FFT test chip in this paper is evaluated at its true (non-margined) maximum frequency, as was assumed in [24]–[27].

TABLE II

COMPARISON OF RECONFIGURABLE FFT (THIS WORK) WITH STATE-OF-THE-ART FFT ENGINES AT MEP

	[24]	[25]	[26]	[27]	this work
technology	65nm	65nm	180nm	90nm	40nm
size	1024	128-2048	128-1024	256	256
word width	16 bit	12 bit	16 bit	10 bit	16 bit
area	8.5mm^2	1.37mm^2	5.5mm^2	5.1mm^2	0.53mm^2
design point	CV 0.27V, 30 MHz,240MS/s	CV 0.43V, 10 MHz,80MS/s	RV 0.35V, 10 KHz,NA	CV 0.85V, 300 MHz,2.4GS/s	CV 0.36V, 2.5 MHz,20MS/s
energy/FFT*	15.8 nJ	6.2 nJ	30 nJ	12.8 nJ	1.88 nJ
normalized energy/FFT**	2.43 nJ	5.6 nJ	6.7 nJ	10.5 nJ	1.88 nJ***

* all measured at true maximum (non-margined) frequency

** energy normalized as in [26]: $E_{norm}=\text{energy}*\frac{256}{\text{FFTsize}*\frac{\text{tech}}{40\text{nm}}*0.66*\frac{WL}{16}+0.33*\frac{WL^2}{16}}$

*** energy at margined frequency is 2.6nJ

To make the comparison fair and independent of the technology and the specific FFT parameters (e.g., number of points, wordlength), the energy was normalized according to the popular figure of merit in [24] and the adopted FFT parameters (256-point, 16-bit input, and 32-bit output bitwidth). The proposed dynamically adaptable pipelined FFT design achieves an energy improvement of 1.3–5.6× compared to [24]–[27].

VI. CONCLUSION

In this paper, dynamically adaptable pipelines have been introduced to reduce energy via joint microarchitectural and voltage co-optimization, for voltages ranging from nominal voltage down to V_{min}. Based on an FFT engine test chip in 40 nm, energy benefits of up to 38% (30%) have been demonstrated in voltage (throughput)-constrained designs. The dynamic pipeline adaptation is achieved at a moderate energy, area, and performance overhead of 6%, 3.1%, and 6.7%, respectively.

As further benefit, dynamically adaptable pipelines extend the voltage region with relatively flat energy, thus relaxing the accuracy requirement in the V_{DD} generation. Full integration of dynamically adaptable pipelines in automated design flows and run-time microarchitectural voltage co-adjustment have been demonstrated and applied to the design of the FFT test chip.

Starting from the above results and findings, future work will explore architectures with complex control flow (e.g., out-of-order microprocessors).

ACKNOWLEDGMENT

The authors would like to thank TSMC for sponsoring the chip fabrication. They would also like to thank G. Ponnusamy and T. Q. Kien for their assistance during chip testing.

REFERENCES

[1] M. Alioto, Ed., *Enabling the Internet of Things*. Cham, Switzerland: Springer 2017.

[2] T. D. Burd, T. A. Pering, A. J. Stratakos, and R. W. Brodersen, "A dynamic voltage scaled microprocessor system," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2000, pp. 294–295.

[3] S. Jain *et al.*, "A 280 mV-to-1.2 V wide-operating-range IA-32 processor in 32 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2012, pp. 66–68.

[4] W. Wang and P. Mishra, "System-wide leakage-aware energy minimization using dynamic voltage scaling and cache reconfiguration in multitasking systems," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 5, pp. 902–910, May 2012.

[5] A. P. Chandrakasan *et al.*, "Technologies for ultradynamic voltage scaling," *Proc. IEEE*, vol. 98, no. 2, pp. 191–214, Feb. 2010.

[6] M. Seok, D. Jeon, C. Chakrabati, D. Blaauw, and D. Sylvester, "Extending energy-saving voltage scaling in ultra low voltage integrated circuit designs," in *Proc. IEEE Int. Conf. IC Design Technol.*, Austin, TX, USA, May/Jun. 2012, pp. 1–4.

[7] D. Jacquet *et al.*, "A 3 GHz dual core processor ARM cortex TM −A9 in 28 nm UTBB FD-SOI CMOS with ultra-wide voltage range and energy efficiency optimization," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 812–826, Apr. 2014.

[8] F. Abouzeid, S. Clerc, B. Pelloux-Prayer, F. Argoud, and P. Roche, "28 nm CMOS, energy efficient and variability tolerant, 350 mV-to-1.0 V, 10 MHz/700 MHz, 252 bits frame error-decoder," in *Proc. ESSCIRC*, Bordeaux, France, Sep. 2012, pp. 153–156.

[9] S. Hsu *et al.*, "A 280 mV-to-1.1 V 256b reconfigurable SIMD vector permutation engine with 2-dimensional shuffle in 22 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2012, pp. 178–180.

[10] S. Hanson *et al.*, "Ultralow-voltage, minimum-energy CMOS," *IBM J. Res. Develop.*, vol. 50, no. 4.5, pp. 469–490, Jul./Sep. 2006.

[11] S. Hanson, B. Zhai, D. Blaauw, D. Sylvester, A. Bryant, and X. Wang, "Energy optimality and variability in subthreshold design," in *Proc. ISLPED*, vol. 6. Tegernsee, Germany, Oct. 2006, pp. 363–365.

[12] W. Zhao, Y. Ha, and M. Alioto, "Novel self-body-biasing and statistical design for near-threshold circuits with ultra energy-efficient AES as case study," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 8, pp. 1390–1401, Aug. 2015.

[13] Y. Zhang *et al.*, "8.8 iRazor: 3-transistor current-based error detection and correction in an ARM cortex-R4 processor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Jan./Feb. 2016, pp. 160–162.

[14] S. Jain and M. Alioto, "A closed-form energy model for VLSI circuits under wide voltage scaling," in *Proc. ICECS*, Monte Carlo, France, 2016, pp. 548–551.

[15] H. Shimada, H. Ando, and T. Shimada, "Pipeline stage unification: A low-energy consumption technique for future mobile processors," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 2003, pp. 326–329.

[16] A. Efthymiou and J. D. Garside, "Adaptive pipeline depth control for processor power-management," in *Proc. IEEE Int. Conf. Comput. Design, VLSI Comput. Process.*, Sep. 2002, pp. 454–457.

[17] S. Chellappa, C. Ramamurthy, V. Vashishtha, and L. T. Clark, "Advanced encryption system with dynamic pipeline reconfiguration for minimum energy operation," in *Proc. 16th Int. Symp. Quality Electron. Design (ISQED)*, Santa Clara, CA, USA, Mar. 2015, pp. 201–206.

[18] S. Vijayalakshmi, A. Anpalagan, I. Woungang, and D. P. Kothari, "Power management in multi-core processors using automatic dynamic pipeline stage unification," in *Proc. Int. Symp. Perform. Eval. Comput. Telecommun. Syst. (SPECTS)*, Toronto, ON, Canada, Jul. 2013, pp. 120–127.

[19] H. M. Jacobson, "Improved clock-gating through transparent pipelining," in *Proc. Int. Symp. Low Power Electron. Design*, Newport Beach, CA, USA, Aug. 2004, pp. 26–31.

[20] S. Manne, A. Klauser, and D. Grunwald, "Pipeline gating: Speculation control for energy reduction," in *Proc. 25th Annu. Int. Symp. Comput. Archit.*, Barcelona, Spain, Jul. 1998, pp. 132–141.

[21] H. Shimada, H. Ando, and T. Shimada, "A hybrid power reduction scheme using pipeline stage unification and dynamic voltage scaling," in *Proc. IEEE COOL Chips*, Apr. 2006, pp. 201–214.

[22] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 1, pp. 3–29, Jan. 2012.

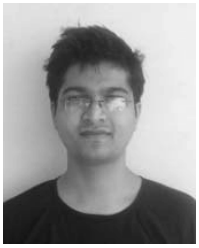
[23] B. H. Calhoun and A. P. Chandrakasan, "Characterizing and modeling minimum energy operation for subthreshold circuits," in *Proc. Int. Symp. Low Power Electron. Design*, Newport Beach, CA, USA, 2004, pp. 90–95.

[24] D. Jeon, M. Seok, C. Chakrabarti, D. Blaauw, and D. Sylvester, "A super-pipelined energy efficient subthreshold 240 MS/s FFT core in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 23–34, Jan. 2012.

[25] C.-H. Yang, T.-H. Yu, and D. Markovic, "Power and area minimization of reconfigurable FFT processors: A 3GPP-LTE example," *IEEE J. Solid-State Circuits*, vol. 47, no. 3, pp. 757–768, Mar. 2012.

[26] A. Wang and A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, Jan. 2005.

[27] Y. Chen, Y. W. Lin, Y. C. Tsao, and C. Y. Lee, "A 2.4-Gsample/s DVFS FFT processor for MIMO OFDM communication systems," *IEEE J. Solid-State Circuits*, vol. 43, no. 5, pp. 1260–1273, May 2008.
[28] M. Alioto, G. Palumbo, and M. Pennisi, "Understanding the effect of process variations on the delay of static and domino logic," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 5, pp. 697–710, May 2010.
[29] M. Eisele, J. Berthold, D. Schmitt-Landsiedel, and R. Mahnkopf, "The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 5, no. 4, pp. 360–368, Dec. 1997.
[30] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, Feb. 2002.



Saurabh Jain (M'17) received the B.Tech. and M.Tech degrees from IIT Kanpur, Kanpur, India, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

His current research interest includes the development of reconfigurable architectures for widely voltage-scalable memory and logic.

Longyang Lin (M'16) received the dual bachelor's degree from Shenzhen University, Shenzhen, China, and Umeå University, Umeå, Sweden, in 2011, and the master's degree from Lund University, Lund, Sweden, in 2013. He is currently pursuing the Ph.D. degree with the National University of Singapore, Singapore.

His current research interests include the design of ultra-low power circuits and systems.

Massimo Alioto (M'01–SM'07–F'16) was an Associate Professor at the University of Siena, Siena, Italy. He was a Visiting Professor at the Intel Labs–CRL, Hillsboro, OR, USA the University of Michigan, Ann Arbor, MI, USA, the BWRC–University of California, Berkeley, CA, USA, and the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. He is currently an Associate Professor with the National University of Singapore, Singapore, where he is also the Director of the Integrated Circuits and Embedded Systems area. He has authored or co-authored more than 240 publications in journals and conference proceedings. One of them is the second most downloaded TRANSACTIONS ON CIRCUITS AND SYSTEMS I (TCAS-I) paper in 2013. He has co-authored three books, *Enabling the Internet of Things—from Circuits to Systems* (Springer, 2017), *Flip-Flop Design in Nanometer CMOS—from High Speed to Low Energy* (Springer, 2015), and *Model and Design of Bipolar and MOS Current-Mode Logic: CML, ECL, and SCL Digital Circuits* (Springer, 2005). His current research interests include ultra-low power very large scale integration (VLSI) circuits, self-powered and wireless sensor nodes, near-threshold circuits for green computing, widely energy-quality scalable VLSI circuits, circuit techniques for emerging technologies and hardware-level security, and accelerators for embedded machine learning.

Dr. Alioto was a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2009 to 2010, for which he has also been a member of the Board of Governors since 2015; he was the Chair of the VLSI Systems and Applications Technical Committee from 2010 to 2012. In the last five years, he has given 50+ invited talks in top universities and leading semiconductor companies. He currently serves as an Associate Editor-in-Chief for the IEEE TRANSACTIONS ON VLSI SYSTEMS, and served as a Guest Editor for various journal special issues. He also serves or has served as an Associate Editor for a number of journals such as the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS, the *ACM Transactions on Design Automation of Electronic Systems*, and the IEEE TCAS-I). He was the Technical Program Chair of SOCC, ICECS, VARI, NEWCAS, ICM, and PRIME and the Track Chair in a number of conferences such as ICCD, ISCAS, ICECS, VLSI-SoC, APCCAS, and ICM.