A Super-Pipelined Energy Efficient Subthreshold 240 MS/s FFT Core in 65 nm CMOS

Dongsuk Jeon, *Student Member, IEEE*, Mingoo Seok, *Student Member, IEEE*, Chaitali Chakrabarti, David Blaauw, *Senior Member, IEEE*, and Dennis Sylvester, *Fellow, IEEE*

***Abstract*—This paper proposes a design approach targeting circuits operating at extremely low supply voltages, with the goal of reducing the voltage at which energy is minimized, thereby improving the achievable energy efficiency of the circuit. The proposed methods accomplish this by minimizing the circuit's ratio of leakage to active current. The first method, *super pipelining*, increases the number of pipeline stages compared to conventional ultra low voltage (ULV) pipelining strategies, reducing the leakage/dynamic energy ratio and simultaneously improving performance and energy efficiency. Measurements of super-pipelined multipliers demonstrate 30% energy savings and 1.6× performance improvement. Since super pipelining reduces the logic depth between registers, two-phase latch based design is employed to compensate for reduced averaging effects and provide better variation tolerance. The second technique introduces a parallel-pipelined architecture that suppresses leakage energy by ensuring full utilization of functional units and reduces memory size. We apply these techniques to a 16-b 1024-pt complex-valued Fast Fourier Transform (FFT) core along with low-power first-in first-out (FIFO) design and robust clock distribution network. The FFT core is fabricated in 65 nm CMOS and consumes 15.8 nJ/FFT with a clock frequency of 30 MHz and throughput of 240 Msamples/s at $V_{dd} = 270$ mV, providing 2.4× better energy efficiency than current state-of-art and $> 10\times$ higher throughput than typical ULV designs. Measurements of 60 dies show modest frequency (energy) σ/μ spreads of 7% (2%).**

***Index Terms*—Fast Fourier Transform (FFT), subthreshold CMOS circuits, super-pipelining, ultra low voltage (ULV) design.**

I. Introduction

RECENTLY, voltage scaling has been widely applied to highly energy-constrained systems such as battery-powered sensor nodes to minimize energy consumption. Voltage scaling enables energy efficient computation by quadratic (or greater) reductions of switching and leakage power dissipation. Although voltage scaling increases gate delay and thus degrades performance, it is still advantageous for many applications with relaxed performance requirements [1], [2] and the supply voltage may be scaled down to, or below, the device threshold voltage V_{th}. However, leakage energy consumption per cycle increases due to enlarged stage delay as voltage scales and this overhead starts to exceed the switching energy savings below the optimal operating point V_{opt}, producing optimal energy consumption E_{opt}. Therefore there exists a fundamental limit for energy savings from voltage scaling in the subthreshold regime regardless of V_{th} [3]. To enhance energy efficiency beyond this point, leakage energy must be suppressed by elimination of idle gates or other techniques to boost the utilization of each gate or module in the system. Since ultra-low voltage operation incurs high process/voltage/temperature (PVT) variation [4], variation tolerance should also be considered in designing these low voltage systems. Such an energy-optimal design methodology is demonstrated on a Fast Fourier Transform (FFT) accelerator in this work.

The FFT is a key digital signal processing (DSP) algorithm and is widely used in digital communication and sensor signal processing. Aided by technology scaling, FFT accelerators have become feasible, offering higher energy efficiency than general purpose processors even for volume-constrained systems such as sensor nodes [2], [5]. We use such an FFT core as a demonstration vehicle for several circuit and architectural techniques aimed at reducing V_{opt} and E_{opt}, while achieving unusually high throughput for a subthreshold circuit. Past work in power efficient FFTs include [6], where the authors propose a cached-memory FFT architecture that processes intermediate results within cached data sets to minimize the number of main memory accesses. In [5], the authors employ voltage scaling to improve energy efficiency. They use standard cells and memories optimized for subthreshold operation and target their design at the optimal energy operating point. However, the body of prior work in this area has not investigated the key role of leakage energy in the subthreshold regime, and we show that energy efficiency can be improved beyond the conventional optimal energy operating point by suppressing leakage effectively.

This paper is an extension of [7]. It describes the use of various circuit techniques such as super-pipelining along with an architectural study focused on extending voltage scalability and enhancing performance in the design of 1024-point complex-valued FFT core. The use of super-pipelining improves performance and reduces leakage energy, but removes averaging effects of random process variability due to shorter logic depth. As a result we employ two-phase latches rather than edge-triggered registers to recapture some averaging through time borrowing. Measured results of these techniques on a multiplier show 30% energy savings concurrently with 1.6× performance improvement over a conventional unpipelined multiplier. A parallel-pipelined FFT architecture is then proposed to maximize computational element and memory

Manuscript received April 22, 2011; revised June 27, 2011; accepted August 22, 2011. Date of publication November 04, 2011; date of current version December 23, 2011. This paper was approved by Guest Editor Tanay Karnik. This work was supported by the Multiscale Systems Center, Army Research Laboratory, National Science Foundation, and National Institute of Standards and Technology.

D. Jeon, D. Blaauw, and D. Sylvester are with the University of Michigan, Ann Arbor, MI 48109-2121 USA (e-mail: djeon@umich.edu).

M. Seok is with Texas Instruments, Dallas, TX 75243 USA.

C. Chakrabarti is with Arizona State University, Tempe, AZ 85287 USA.

Digital Object Identifier 10.1109/JSSC.2011.2169311

0018-9200/$26.00 © 2011 IEEE

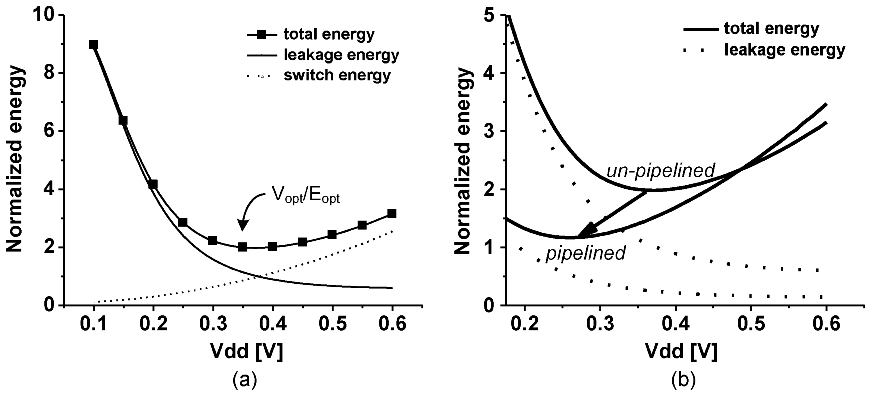


Fig. 1. (a) Simulated energy consumption of inverter chain showing energy-optimal point. (b) Energy comparison between unpipelined and pipelined inverter chains.

utilization, while reducing memory size and overall leakage energy. Such an approach is ideally suited to subthreshold design where leakage energy is significant, in contrast to traditional superthreshold design paradigm where dynamic energy and performance are the primary metrics. Combining these techniques, the parallel-pipelined FFT core with super-pipelined CEs (computational elements) is fabricated in a standard 65 nm CMOS process. The resulting design achieves 15.8 nJ/FFT operating at 30 MHz with $\mathrm{V_{dd}} = 270\ \mathrm{mV}$, enabling throughput of 240 Msamples/s.

II. Pipelining Techniques for ULV Design

Conventional pipelining in the nominal voltage regime mainly focuses on improving system performance. By dividing operations into multiple stages, stage delay is shortened and clock frequency increased, enabling higher throughput. However, pipelining incurs energy overhead in the switching energy of pipeline registers and clock distribution network, and there has been extensive research activity on optimal pipelining in power- or performance-constrained systems at nominal supply voltages. In [8], a power-aware pipeline strategy is proposed and the power–performance trade-off is demonstrated. In addition, [9] showed that there also exists an optimal pipeline depth for maximum performance in purely performance-constrained systems.

Another well-known benefit of pipelining is power reduction [10] under fixed throughput constraints. By employing more pipeline stages, the overall throughput is improved, which can then be traded off for power savings via voltage scaling. However, this approach is inappropriate for highly energy-constrained systems that usually have no performance constraints and where energy efficiency is paramount. In contrast, ultra low voltage (ULV) designs have generally used few pipeline stages with long stage delays to minimize sequential overhead from pipeline registers for low power and achieve greater tolerance to the prominent PVT variations in ULV regime via averaging effects [5], [11]. In this paper, we propose to employ aggressive pipelining for ULV designs and show that it enables substantial energy savings while addressing PVT variation tolerance by the use of latch-based design.

A. *Super-Pipelining Technique*

In subthreshold design, due to exponentially increased stage delay the leakage energy contributes appreciably to the total energy consumption. The conventional relaxed pipelining approach with long cycle times increases leakage energy per cycle, saturating energy efficiency as shown in Fig. 1(a). However, if this leakage energy can be suppressed with low overhead, voltage can be scaled further to reduce switching energy, resulting in a new lower $\mathrm{V_{opt}}$ and $\mathrm{E_{opt}}$ as depicted in Fig. 1(b). To achieve this, we make use of *super-pipelining*, which employs significantly more pipeline stages and shorter stage delays, improving performance and, counter-intuitively, simultaneously enhancing energy efficiency. The shorter stage delay reduces leakage energy per cycle since leakage energy is the integral of leakage power over a clock period, and this savings can exceed the sequential overhead of added pipeline registers in the subthreshold regime. With the reduced leakage energy, $\mathrm{V_{opt}}$ shifts downward and allows for more voltage scaling and lower $\mathrm{E_{opt}}$ (Fig. 1(b)). While conventional deep pipelining techniques seek to improve performance or scale supply voltage at a fixed performance, our approach focuses solely on energy minimization. Specifically, it acts to move the minima of the energy consumption plot as shown in Fig. 1(b).

However, as the circuit is pipelined more deeply, sequential overheads begin to outweigh the leakage energy savings, causing energy efficiency to degrade beyond some point. In addition, as registers commonly exhibit functional failure at higher supply voltages than combinational logic [12], a larger pipeline register count may increase the lowest functional operating voltage $\mathrm{V_{min}}$, making it larger than $\mathrm{V_{opt}}$. Fig. 2 shows SPICE simulation results with 60 fanout-of-4 (FO4) delay inverter chains, indicating that the total energy per cycle continuously shrinks until the stage depth is 7 FO4, after which it increases abruptly due to sequential overheads. The optimal energy is reduced by 46%, while performance is improved by 31% at $\mathrm{V_{opt}}$ over an unpipelined inverter chain. The performance improvement is achieved through pipelining (60 to 7 FO4s per stage) but offset by increased delay due to the additional voltage scaling at $\mathrm{V_{opt}}$ (from 370 mV to 240 mV). These simulations used a switching activity of 0.2, and similar energy savings are observed for activity factors ranging from

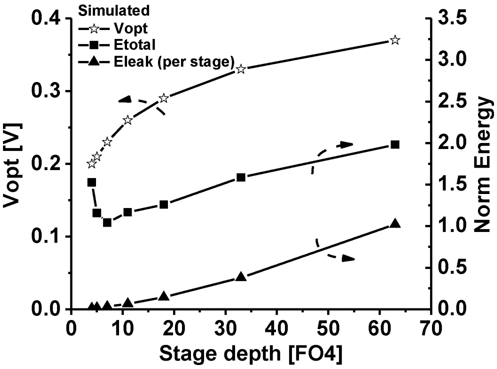


Fig. 2. Energy savings of super-pipelining from simulations with 60 FO4 delay inverter chains.

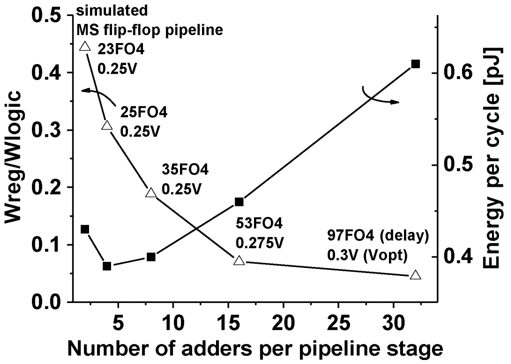


Fig. 3. Energy consumption and total register size (relative to logic) for different pipeline stage depths.

0.1 to 0.5. This insensitivity to switching activity arises since super-pipelining acts on both leakage and dynamic energy (i.e., reduces leakage energy through a shorter cycle time and improves switching energy via further voltage scaling).

B. Latch-Based Design

Although super-pipelining improves energy efficiency, the shortened stage delay worsens delay variations in the subthreshold regime, requiring more timing margin for error-free operation and limiting the benefits of super-pipelining. We seek to mitigate delay variations without significant energy overhead. While global variations can be compensated effectively with system-wide techniques such as body biasing and supply voltage scaling [13], these techniques are ineffective for local (random) variations.

To mitigate delay variability from random process variations, we suggest the use of two-phase latch-based pipelining rather than hard-edge flip-flop based pipelining. Latch-based design enables time borrowing up to nearly half a clock period [14]. This time borrowing recovers the averaging effects of delay variations by stretching the effective stage length through the soft edge of latches without performance degradation. The large time borrowing window of two-phase latch-based design is particularly advantageous in subthreshold circuits with high PVT variations compared to other soft-edge clocking approaches such as soft-edge flip-flops and pulsed latches [15], [16], which have limited time borrowing capabilities and design difficulties with reliably generating short pulses in the subthreshold regime. Although logic depth between latches is decreased by 50% in latch-based design, the logic depth between latches of the same type (transparent low and transparent low) is unchanged, providing the same averaging effect.

Since two-phase latch-based design is prone to hold time violations during clock phase overlaps, non-overlapping clocks are preferable for design robustness. However, this creates energy overhead and increases design complexity. Instead, hold time violations can be avoided by inserting delay elements in short paths during the design phase, particularly in circuits with well-defined structures such as arithmetic units, and hence we take this approach in the FFT design.

C. Super-Pipelined Baugh–Wooley Multiplier

A multiplier is a fundamental component within the FFT core and the long critical paths of an unpipelined multiplier incurs significant leakage energy. We apply super-pipelining and 2-phase latch-based design as described above to a 16-bit Baugh–Wooley multiplier to investigate the benefits of the proposed pipelining scheme. We choose a Baugh–Wooley topology due to its popularity and regular structure, making it an appropriate testbench to show the impact of super-pipelining with various logic depths.

We first explore the energy-optimal super-pipelining scheme for the multiplier. The Baugh–Wooley multiplier is pipelined at various depths with master–slave flip-flops (MSFFs) and the overall energy consumption per cycle is estimated from simulations in 65 nm CMOS. The unpipelined multiplier uses a ripple carry adder (RCA) to minimize total transistor width and energy consumption following conventional practices in ultra-low power design that seek to limit switching energy [5]. Fig. 3 shows the effects of deep pipelining on the Baugh–Wooley multiplier along with the register and logic width. Initially, the use of more pipeline stages enables a large energy savings by reducing leakage, after which the energy saving saturates for a 6-stage pipelined multiplier due to increased sequential overheads. The ratio of register widths to total logic width continuously increases as more pipeline registers are inserted, translating to larger sequential overheads including registers and clock distribution networks. The energy-optimal operating voltage reduces from 300 mV to 250 mV.

The effectiveness of two-phase latch-based pipelining is also investigated. MSFFs are replaced with two-phase latches to allow time borrowing. Based on additional timing margin obtained by latch-based design, one more pipeline stage is included to boost performance without sacrificing energy efficiency and reduce leakage energy in higher level. Fig. 4 depicts the proposed multiplier design. To avoid hold time violations, we explore possible short paths in the multiplier and insert delay elements accordingly. These paths are confirmed with extensive (150,000) random process Monte Carlo and corner simulations to assure more than 99% functional yield for 2000 paths in the multipliers. The energy overhead from this modification is 2.4% of overall multiplier energy.

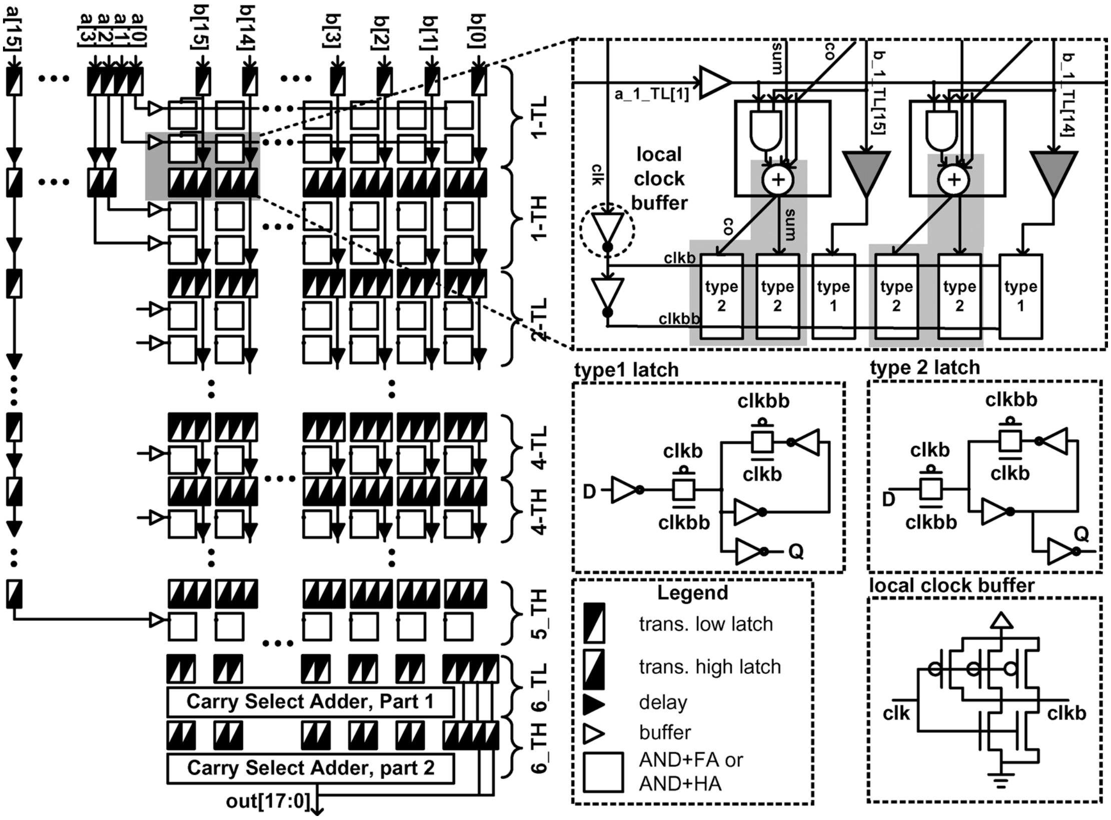


Fig. 4. Proposed super-pipelined multiplier with latches embedded in a full adder (gray box).

The 6-stage pipelined multiplier above is optimized further with various circuit techniques to improve energy efficiency. Two registers are embedded into a full adder cell to save two transistors per register and registers share local clock buffers. The local clock buffer is implemented with multiple minimum width fingers, improving drivability at constant switching power due to inverse narrow width effects. Simulation results show that a clock buffer design using multiple minimum size fingers consumes 16.8% less energy than a conventional layout with fewer fingers. In addition, the pipelined RCA for the final accumulation stage is replaced with a faster Variable Carry Select Adder, enabling 5-stage pipelining due to single-stage final accumulation and improving overall energy efficiency due to leakage energy reduction, despite a larger gate count than RCA.

We fabricated 5-stage MSFF-based and 6-stage latch-based pipelined multipliers (the optimal number of stages for each, respectively) along with an unpipelined design serving as the baseline. Energy and performance measurements with random input vectors (switching activity of ~0.46) are shown in Fig. 5. All multipliers operate down to 225 mV[1] and we do not observe any impact on functional $V_{\min}$ due to pipelining. The proposed latch-based design consumes 0.47 pJ/multiplication at 225 mV and achieves 30% energy savings with 1.6× higher performance at its own energy-optimal point V_{opt}. Simulations predicted a similar V_{opt} change from 275 mV for an unpipelined design to 200 mV for the proposed design. The energy per cycle for the proposed design at its V_{opt} was simulated to be 0.5 pJ/multiplication, which is 6.4% higher than measurements at 225 mV, showing good model-hardware correlation. Looked at differently, the proposed design achieves 18% energy savings with 3.6× better performance at a fixed operating voltage of 275 mV at the expense of 79% area increase compared to the unpipelined design (note that this area overhead will be amortized in the FFT design, as described in Section VI). In addition, the latch-based design has superior energy efficiency and performance compared to the MSFF-based implementation (Fig. 6). Although the two designs have similar σ/μ values due to global variations across the dies, the latch-based design recovers averaging effects through its time borrowing capabilities and achieves better variation tolerance, translating to 1.4× higher average performance than MSFF-based design, similar to the simulated perfor-

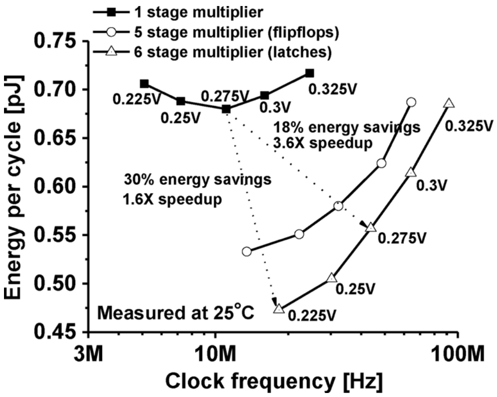


Fig. 5. Measurement results for three different multiplier designs.

[1] Multiplier operation below this value was limited by timing interface issues with the test circuitry.

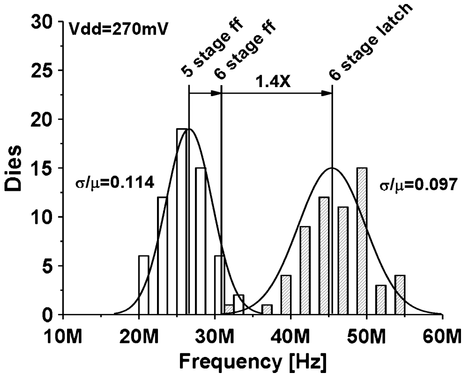


Fig. 6. Measured performance distribution of two pipelined multipliers from 60 dies.

mance improvement of 33%. Note that multiplier measurement results are for a 5-stage MSFF-based design and a 6-stage latch-based design. Expected performance improvement in moving from 5 to 6 pipeline stages in a MSFF-based design are obtained from simulation and used to scale 5-stage results for an iso-stage comparison. These measurement results confirm the benefits of super-pipelining technique and two-phase latch-based design.

III. Energy-Optimal FFT Architecture

Contrary to nominal voltage operation, leakage energy limits energy efficiency in the ultra-low voltage regime. Conventional FFT architectures mainly focus on reducing dynamic power and hardware cost while meeting a performance target, and energy efficiency can be simply calculated from the number of required computations such as complex multiplications and additions [17]. Idle cells or modules, while contributing to hardware or area overhead, are not seriously explored for improving energy efficiency for these reasons. As voltage scales, however, idle cells consume significant leakage energy per cycle while the switching energy of arithmetic units is reduced. Therefore it is critical to eliminate idle cells or modules as much as possible, enabling high utilization and improving energy efficiency.

The best-known FFT architectures are the memory-based and pipelined architectures. The memory-based architecture is one of the simplest ways of implementing the FFT algorithm. It consists of a large memory divided into several banks storing initial input data and intermediate results for the next butterfly operation, as depicted in Fig. 7(a). The computational element (CE) pulls one set of data from memory and stores results into the same memory space after processing. While the CE is processing data, the unaccessed memory cells simply store their previous values and consume leakage energy. Although the CE is fully utilized, memory utilization is very low and only 0.8% of total memory is accessed every cycle for 1024-pt radix-4 FFT. SPICE simulations indicate that 85% of overall energy is dissipated in memory at 300 mV for a radix-4 memory-based architecture, implying this architecture is not appropriate for highly energy efficient operation in subthreshold regime. In addition, it requires the use of a ping-pong type input buffer, increasing memory size, to perform successive FFT of incoming data for applications such as audio processing or OFDM [18], [19].

The pipelined architecture comprises several stages connected in series with CEs and multiple FIFOs to store and re-order intermediate values. It leads to higher dynamic power since multiple CEs switch simultaneously. However, CEs only access FIFOs of the previous stage, and these FIFOs are relatively small compared to the large memory in a memory-based architecture. This significantly reduces the average number of memory cells per CE, enabling high memory utilization and in turn reducing memory leakage energy.

The most straightforward pipelined architecture is MDC (Multi-path Delay Commutator), shown in Fig. 7(b) [20]. Each stage corresponds to one segment of the signal flow graph and the CE performs butterfly computation in the same way as in the memory-based architecture. The commutator consists of switching network and FIFOs, and re-orders the data flow for CE of next stage. However, this architecture suffers from CE utilization as low as 25% for R4MDC (radix-4 MDC) architecture since it accepts only one input per cycle, then waits for several cycles to build a full set of data required for a butterfly operation. This drawback becomes worse in higher-radix FFT algorithms since more data is required for a butterfly operation. In addition, for an N-point FFT this architecture requires $3(\log_4 \mathrm{N} - 1)$ complex multipliers and $5\mathrm{N}/2 - 4$ memory cells, making the hardware costs higher than other pipelined architectures.

A widely used pipelined architecture is $\mathrm{R2^2SDF}$ (Single-path Delay Feedback) [17]. This approach introduces a feedback path composed of local memory beside each CE to temporarily store intermediate values rather than a large memory in the commutator module of the R4MDC architecture. $\mathrm{R2^2SDF}$ requires $\log_4 \mathrm{N} - 1$ complex multipliers and only $\mathrm{N} - 1$ memory cells, achieving 75% utilization of complex multipliers.

A. Modified R4MDC Architecture

The conventional R4MDC architecture accepts only one input per cycle while processing 4 input data concurrently per cycle, and thus the CEs performing radix-4 butterfly operations are activated partially to match throughput with input data rate. This incurs significant leakage energy overhead due to low CE and memory utilization. However, if four inputs from a single channel are obtained in one cycle, full CE utilization can be achieved, which also minimizes memory leakage energy since the number of cycles required to perform one FFT is reduced.

This modified R4MDC architecture with 4 inputs per cycle is depicted in Fig. 8. In altering the original R4MDC architecture to achieve full utilization, we made several changes:

- Altered data scheduling within a commutator
- Different configurations for each FIFO
- Process four input data per cycle

While the original R4MDC architecture employs large input buffers to convert serial input to a parallel data stream for the first CE, the modified architecture allows the input re-ordering buffer to be half as large since four input data are fed into the FFT core per cycle. Therefore data scheduling within a commutator had to be changed from the original R4MDC architecture, which results in smaller FIFO configurations as shown in Fig. 8. The switching network remains the same as the original

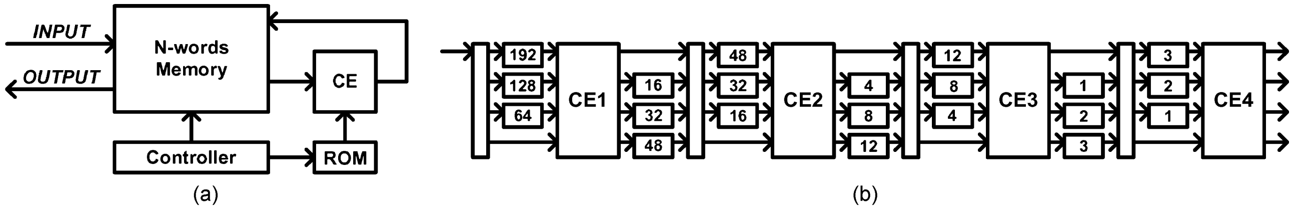


Fig. 7. (a) Memory-based architecture and (b) R4MDC pipelined architecture.

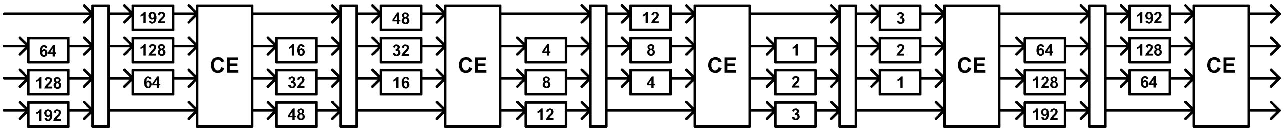
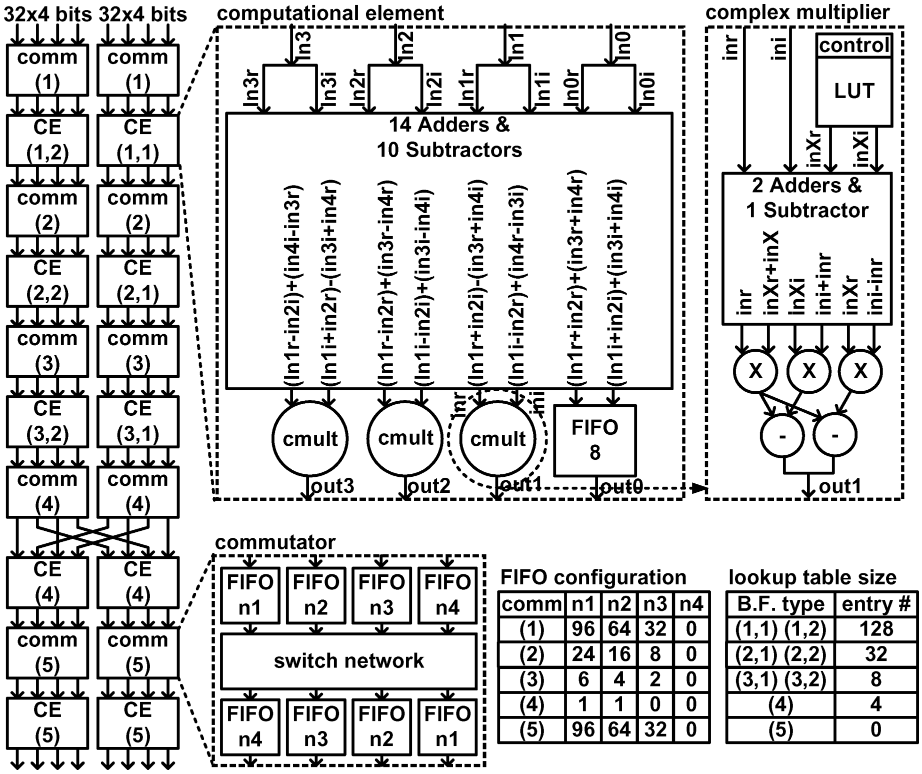


Fig. 8. Modified R4MDC architecture for 1024-pt FFT.



FIFO configuration

comm	n1	n2	n3	n4
(1)	96	64	32	0
(2)	24	16	8	0
(3)	6	4	2	0
(4)	1	1	0	0
(5)	96	64	32	0

lookup table size

B.F. type	entry #
(1,1) (1,2)	128
(2,1) (2,2)	32
(3,1) (3,2)	8
(4)	4
(5)	0

Fig. 9. Proposed FFT core architecture with two processing lanes.

R4MDC and a look-up table for twiddle factors is embedded in each CE along with a controller.

In addition to full CE utilization, this architecture requires fewer memory cells for commutators. Specifically, modified R4MDC contains $7N/4-4$ memory cells compared to $5N/2-4$ in R4MDC, reducing memory size by 30% for a 1024-pt FFT. Although a CE accesses memory every cycle, implying full memory utilization, this does not reflect the actual number of activated versus idle cells. Instead the average number of memory cells per complex multiplier can be used as an alternate metric in energy-constrained applications. For a 1024-pt FFT, the modified R4MDC requires 149 cells per complex multiplier while $\mathrm{R2^2SDF}$ needs 255.75 cells (72% more). Taken together, modified R4MDC achieves 4× higher throughput and more energy efficient operation than $\mathrm{R2^2SDF}$ with smaller hardware cost. Simulation results indicate that modified R4MDC and $\mathrm{R2^2SDF}$ consume 52% and 68% of their total energy in memory at 250 mV, respectively, with modified R4MDC consuming 43.2% less energy with 2.6× better performance than $\mathrm{R2^2SDF}$ at their respective energy-optimal points.

B. Parallel-Pipelined Architecture

The improvements above significantly increase memory utilization and suppress leakage energy. However, even the modified R4MDC architecture consumes 52% of its total energy in memory, indicating that further room for improvement exists in memory utilization. This section shows that FFT architecture parallelization reduces energy consumption per operation and improves performance simultaneously.

Fig. 9 shows the modified R4MDC architecture with two processing lanes in parallel. Eight in-order input data are fed into the FFT core every cycle and each processing lane processes

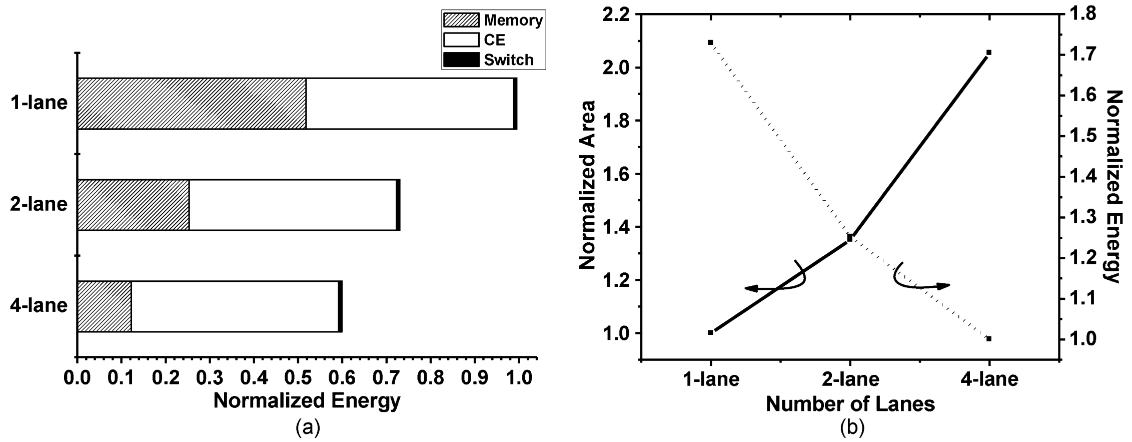


Fig. 10. (a) Energy breakdown and (b) energy–area trade-off of parallel-pipelined architectures.

data within its own data set until the second stage from the last, after which intermediate results are exchanged followed by independent processing in each lane. The proposed design in Fig. 9 requires $4N/7-8$ memory cells, similar to the 1-lane version. Thus, the average number of memory cells per multiplier is reduced by more than 50%, translating directly to memory leakage reduction. The proposed design improves performance by 2× through parallelism while consuming only 35% of its total energy in memory, indicating a greater degree of voltage scalability and potential E_{opt} improvements.

Area–time–energy optimization has been applied in a FIR filter [21] previously to improve energy efficiency with fixed throughput. The proposed FFT architecture differs from this work in that it focuses solely on achieving minimum energy consumption with emphasis on the role of leakage current in the subthreshold regime. Since ultra-low voltage (subthreshold) FFT architectures consume most of their energy in memory as leakage, the enhanced memory utilization from the proposed architecture provides significant energy savings.

The FFT core can be further parallelized to achieve continued benefits. However, hardware costs increase while the benefit from parallelization saturates as the CE active energy starts to dominate. Fig. 10(a) shows energy breakdowns of modified R4MDC architectures with different degrees of parallelism. With the CE already fully utilized, its energy consumption remains unchanged while memory energy decreases with additional parallelization. The energy–area trade-off shown in Fig. 10(b) clearly indicates that energy reductions from aggressive parallelism saturate while incurring large area overhead. Although the minimum energy will be achieved when entire signal flow of FFT algorithm is implemented without any memory modules, increased numbers of multipliers and other modules results in higher PVT-induced variability and lower maximum operating frequency, translating to more leakage energy per cycle. In addition, higher parallelism incurs wire overheads including the switching network and input/output interfaces. Based on this and available silicon area constraints, we implement the 2-lane version.

Simulation results show that the proposed design with 2 lanes consumes 27.5% less energy with 2× higher performance than a 1-lane version, translating to 2.4× better energy efficiency and 5.2× higher throughput than the conventional $\mathrm{R2^2SDF}$ architecture.

IV. Robust Subthreshold FIFO Design

The first-in first-out (FIFO) is a key module in commutators, and the proposed architecture employs a large number of these modules. Simulations of the 2-lane modified R4MDC architecture show that FIFOs consume up to 29% of total energy, making it necessary to explore energy efficient FIFO design.

The most straightforward FIFO implementation is a shift register-based FIFO [22]. For an N delay FIFO, N registers are connected in series and all of them switch every cycle, sending data to next stage. Although this is very robust due to its negative hold time property, registers toggling every cycle have significant switching power overhead, making them less attractive for low power designs. Another possible candidate for a low-power FIFO is an SRAM-based design [23], which uses conventional SRAM for storing data and reads data from each word successively, writing subsequent data into the same address. For nominal voltage operation, a simple 6T SRAM can be used for high density with lower switching power than shift registers while retaining sufficient speed to match the data rate. However, a 6T SRAM is not suitable for the subthreshold regime due to its susceptibility to process variability and resulting small read/write margins, and variants such as 10T [24] are preferred for robust operation. The maximum operating frequency of robust subthreshold SRAM designs is typically below 1 MHz [25], [26], making it impossible to meet performance requirements of our target FFT design.

To mitigate these performance and variation issues, we propose the use of a latch-based memory. Although latch-based memory as described in [5] increases read/write margins and offers potential performance improvements, additional address generator or decoder is necessary for read/write address signal generation. In addition, MUX-based readout paths are slow and suffer from leakage energy overhead due to the correspondingly long cycle time. For improved energy efficiency, we propose a latch-based memory with dedicated address generator and logic-based readout path. The entire FIFO architecture is described in Fig. 11.

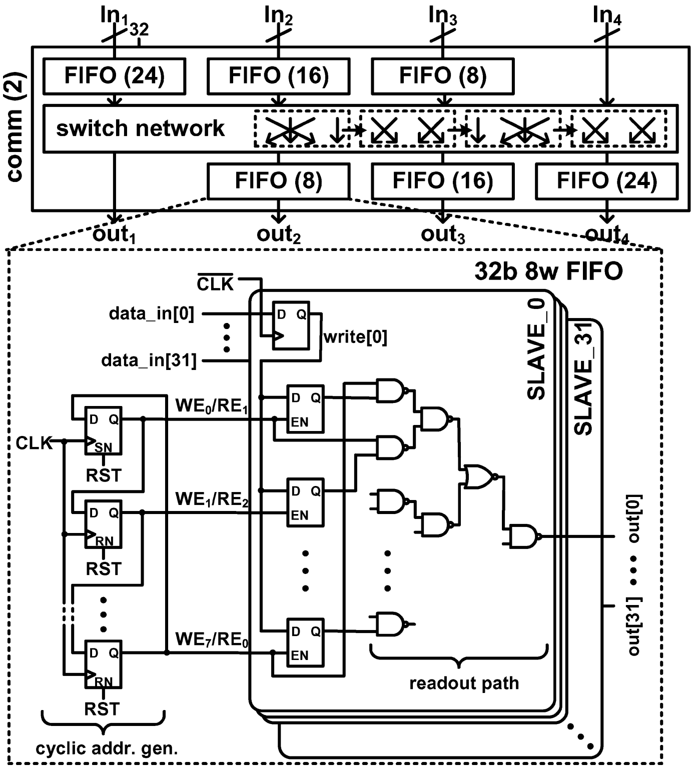


Fig. 11. Proposed 8-word FIFO design with commutator architecture.

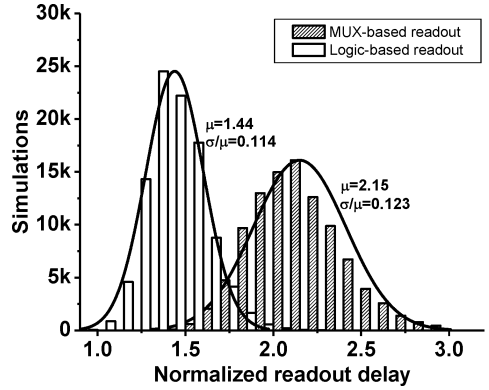


Fig. 12. Readout delay distributions from 100 k Monte-Carlo SPICE simulations.

The proposed design contains a cyclic address generator consisting of a single chain of registers, producing both write and read enable signals. While one of the latches is enabled for write operation based on the enable signal from address generator, the data stored in the next latch is read through the NAND gate using the same signal. This cyclical access pattern reduces energy by sharing one address generator for read and write operation rather than using a separate decoder for write operation as in conventional SRAM designs. The logic-based readout path operates faster than the MUX-based option, allowing the FIFO to match the performance of the super-pipelined CEs. Fig. 12 shows Monte-Carlo simulation results of readout delays of MUX-based and logic-based readout scheme. The logic-based design is 33% faster on average than the MUX-based one, indicating that the proposed design effectively reduces leakage energy by shortening stage delay while improving throughput. Simulation results further confirm that the 32-word FIFO with proposed logic-based design consumes 12% lower energy with 20% higher performance than the MUX-based design.

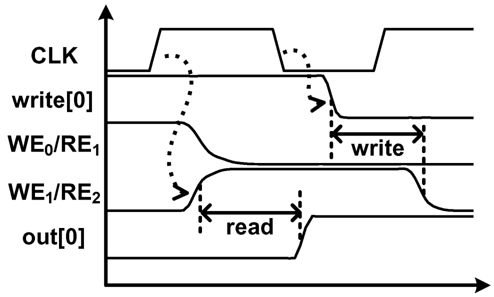


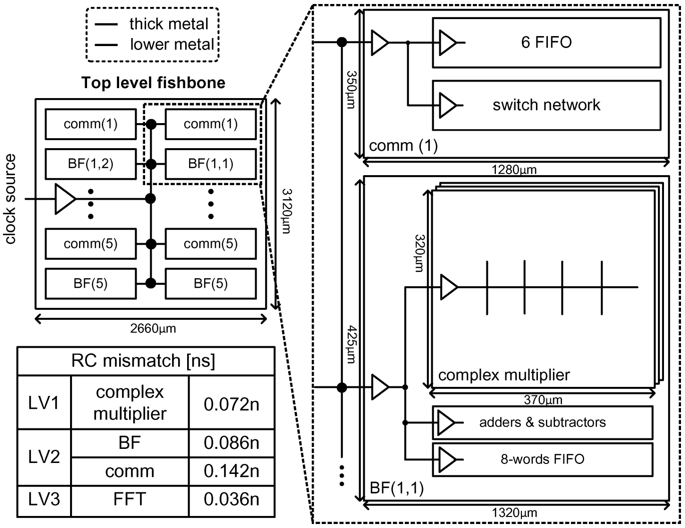
Fig. 13. Positive-edge read and negative-edge write scheme.

For robust read operation, hold time issues must be avoided between the write enable and write data signals. If write data changes while write enable of previous cycle is still asserted, the stored data of previous latch will be corrupted. To avoid this hold time issue, we introduce a negative-edge write scheme. As described in Fig. 13, the write data signal remains unchanged for a half cycle after the previous write enable signal is disabled to provide large write operation timing margin and guaranteeing timing violation-free write operation. Since the readout path is critical for this FIFO module, this increased timing margin for write operation does not impact overall performance. We implemented 8-word and 32-word FIFOs, and larger FIFOs are obtained by connecting them in series.

V. Clock Distribution Network

Clock tree distribution networks of large integrated circuits are generally composed of several levels with local clock buffers to suppress RC delay of long wires. In the nominal voltage regime, clock buffer gate delay is small compared to the RC delay of global clock networks and inserting a large number of clock buffers is used to reduce clock skew. In the subthreshold regime, however, exponentially increased gate delay dominates while RC delay no longer contributes appreciably to clock path delay mismatches. In contrast, buffer mismatch significantly impacts clock distribution delays and adding more buffers leads to higher clock skew. Mismatch can be effectively suppressed by employing only a small number of large buffers, since they are robust to random process variations [4], [27].

To suppress clock skew incurred by clock buffer mismatch, we design a 3-level clock network with a small number of relatively large buffers. RC delay mismatch is also minimized by matching wire lengths of each level. Fig. 14 shows the complete clock distribution network. The lowest and middle level networks, which have relatively small RC delays, are implemented with minimum width lower level metallization for low power consumption while the top level is designed in a fish-bone network using top thick metal layers to minimize RC delay. Worst-case RC mismatch from simulation is less than 150 ps or 0.14 FO4 delays at $\mathrm{V_{dd}} = 270\ \mathrm{mV}$. Fig. 15 also indicates that simulated 2σ clock skew due to buffer mismatch is 0.68× FO4, which is 2% of the clock cycle at 270 mV.



RC mismatch [ns]		
LV1	complex multiplier	0.072n
LV2	BF	0.086n
	comm	0.142n
LV3	FFT	0.036n

Fig. 14. RC-matched 3-level clock distribution network along with maximum RC mismatch values.

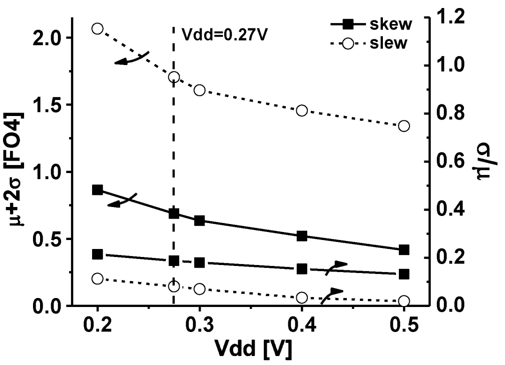


Fig. 15. Simulated clock skew and slew rates in proposed clock network design.

VI. Measurement Results

The FFT core is designed using the described circuit and architectural techniques. Fig. 16 provides measured energy and performance results for the proposed FFT core. The core consumes 15.8 nJ/FFT at a measured maximum clock frequency of 30 MHz at 270 mV, yielding 240 Msamples/s. This throughput is 10 ~ 100× higher than typical ULV designs [2], [5]. At 600 mV the proposed design consumes 35.0 nJ/FFT at a clock frequency of 290 MHz; this energy efficiency is a 2× improvement over the high performance design in [28] at the same throughput. Fig. 17 reports the measured performance and energy efficiency as a function of temperature, demonstrating functionality across a wide temperature range, which is often a challenge in subthreshold designs. Also, note that higher operating frequencies due to super-pipelining may cause temperature increases, pushing leakage energy up exponentially. However, since the proposed FFT core consumes only 3.7 mW at 270 mV and 30 MHz, power density is very low (0.043 $\mathrm{W/cm^2}$) and there are no self-heating effects that could compromise energy efficiency at low $\mathrm{V_{dd}}$.

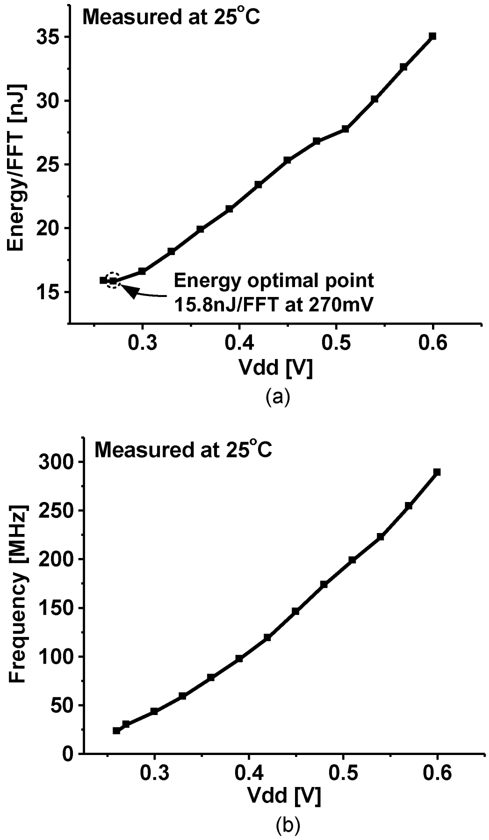


Fig. 16. Measured (a) energy consumption and (b) performance of the FFT core.

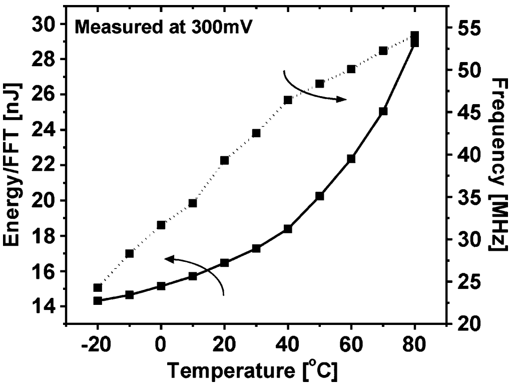


Fig. 17. Measured performance and energy consumption as a function of temperature.

The average energy consumption and clock frequency at $\mathrm{V_{dd}} = 300\ \mathrm{mV}$ are 17.1 nJ/FFT and 41 MHz, respectively, as measured across 60 dies (Fig. 18). This figure also shows modest frequency and energy spreads of only 7% and 2%, respectively, in terms of σ/μ. Table I shows performance characteristics from recent publications along with the proposed FFT design normalized to technology, FFT size, and bit width using the expression[2] in [28]. The proposed design shows 2.4× better energy efficiency than the previous state-of-the-art. Fig. 19 shows the die photograph of the fabricated FFT core with core area of 8.3 $\mathrm{mm^2}$ (2.66 × 3.12 mm) in 65 nm CMOS technology. The pipelined multipliers take 14.8% of the area for the FFT core and the overhead from super-pipelining is 6.5%.

[2] Normalized energy per FFT $=(\mathrm{Energy\ per\ FFT}/((\mathrm{Technology}/65\,\mathrm{nm})*((2/3)(\mathrm{wordlength}/16)+(1/3)(\mathrm{wordlength}/16)^2)))*(1024/\mathrm{FFT\ size})$.

TABLE I
CHARACTERISTICS OF PUBLISHED FFT CORES AND THE PROPOSED DESIGN

	Proposed	[5]	[28]	[29]
Technology	65nm	180nm	90nm	65nm
FFT mode	1024-point complex-valued	128~1024-point real-valued	256-point complex-valued	128-2048pt complex-valued
word width	16 bit	16 bit	10 bit	16 bit
Vdd	0.27~1.0 V	0.18~0.9 V	0.625~1.0 V	0.5~1.0 V
area	2.71×3.15 mm^2	2.6×2.1 mm^2	2.26×2.26 mm^2	1.375mm^2
design point	1024-point CV 0.27V, 30MHz 240MS/s	1024-point RV 0.35V, 10KHz N/A	256-point CV 0.85V, 300MHz 2.4GS/s	1024-point CV 0.43V, 10MHz 80MS/s
Energy/FFT	15.8 nJ	155 nJ	12.8 nJ	37.3nJ
Normalized Energy/FFT	**15.8 nJ**	**111.9 nJ**	**71 nJ**	**37.3nJ**

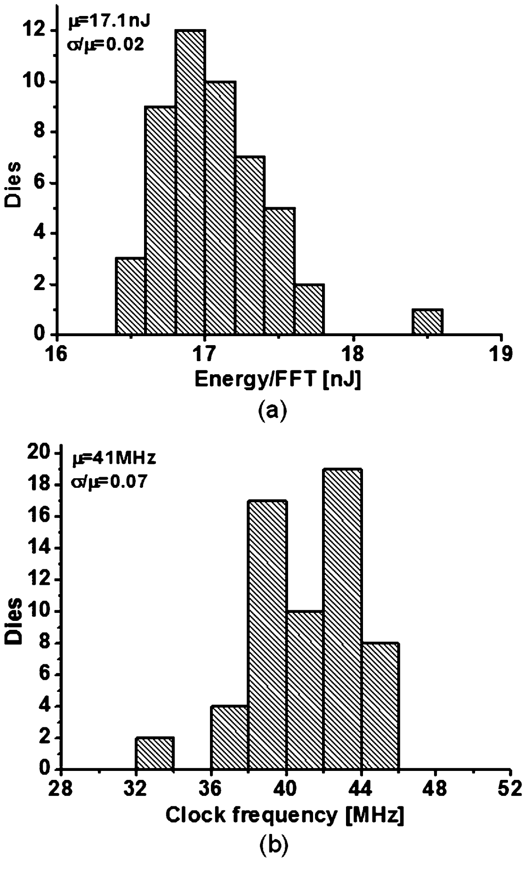


Fig. 18. Measured energy (a) and performance (b) distributions at $\mathrm{V_{dd}} = 300\ \mathrm{mV}$.

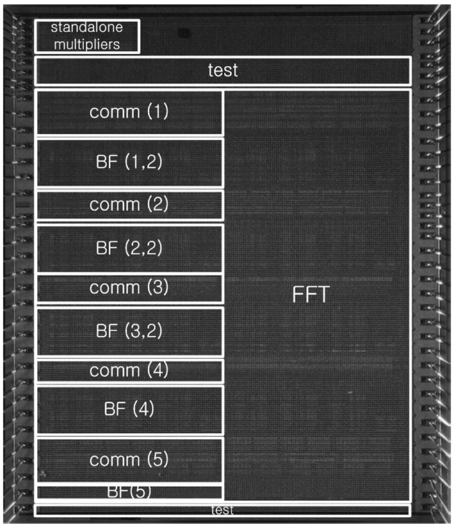


Fig. 19. Die photograph of the FFT core in 65 nm CMOS.

VII. CONCLUSIONS

This paper proposes circuit and architecture techniques to enhance energy efficiency in the subthreshold regime, with application to an FFT module. In contrast to the use of modest pipelining in past subthreshold designs, we advocate the use of super-pipelining to suppress leakage energy consumption while simultaneously improving performance. The latch-based pipelining scheme recovers delay averaging effects by introducing a longer effective stage length through time borrowing to effectively mitigate variations. A parallel-pipelined architecture maximizes CE and memory utilization, enabling greater voltage scaling range and further enhancing throughput. In addition, an energy efficient and robust FIFO design and variation-tolerant clock distribution network are employed. The proposed FFT design is fabricated in 65 nm CMOS and measurements indicate that it successfully operates at an optimal operating point of 270 mV at a clock frequency of 30 MHz. Its energy efficiency is 2.4× higher than previous state-of-art FFT designs, while performance is more than 10× higher than past ULV designs, demonstrating the feasibility of very low voltage design for moderate to high performance systems.

ACKNOWLEDGMENT

IC fabrication support of STMicroelectronics is gratefully acknowledged.

REFERENCES

[1] G. Chen, M. Fojtik, D. Kim, D. Fick, J. Park, M. Seok, M.-T. Chen, Z. Foo, D. Sylvester, and D. Blaauw, "Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2010, pp. 288–289.

[2] S. R. Sridhara, M. DiRenzo, S. Lingam, S.-J. Lee, R. Blazquez, J. Maxey, S. Ghanem, Y.-H. Lee, R. Abdallah, P. Singh, and M. Goel, "Microwatt embedded processor platform for medical system-on-chip applications," *IEEE J. Solid-State Circuits*, vol. 46, no. 4, pp. 721–730, Apr. 2011.

[3] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. Design Automation Conf.*, May 2005, pp. 868–873.

[4] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2005, pp. 20–25.

[5] 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.

[6] B. M. Baas, "A low-power, high-performance, 1024-point FFT processor," *IEEE J. Solid-State Circuits*, vol. 34, no. 3, pp. 380–387, Mar. 1999.

[7] M. Seok, D. Jeon, C. Chakrabarti, D. Blaauw, and D. Sylvester, "A 0.27 V 30 MHz 17.7 nJ/transform 1024-pt complex FFT core with super-pipelining," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2011, pp. 342–343.

[8] V. Srinivasan, D. Brooks, M. Gschwind, P. Bose, V. Zyuban, P. N. Strenski, and P. G. Emma, "Optimizing pipelines for power and performance," in *Proc. Int. Symp. Microarchitecture*, Nov. 2002, pp. 333–344.

[9] M. S. Hrishikesh, N. P. Jouppi, K. I. Farkas, D. Burger, S. W. Keckler, and P. Shivakumar, "The optimal logic depth per pipeline stage is 6 to 8 FO4 inverter delays," in *Proc. Int. Symp. Computer Architecture*, May 2002, pp. 14–24.

[10] A. Chandrakasan and R. Brodersen*, Low-Power CMOS Design*. New York: Wiley-IEEE Press, 1998.

[11] M. Seok, S. Hanson, Y.-S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "The phoenix processor: A 30 pW platform for sensor applications," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2008, pp. 188–189.

[12] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778–1786, Sep. 2005.

[13] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw, "Performance and variability optimization strategies in a sub-200 mV, 3.5 pJ/inst, 11 nW subthreshold processor," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2007, pp. 152–153.

[14] D. Harris*, Skew-Tolerant Circuit Design*. Burlington: Morgan Kaufmann, 2000.

[15] M. Wieckowski, Y. M. Park, C. Tokunaga, D. W. Kim, Z. Foo, D. Sylvester, and D. Blaauw, "Timing yield enhancement through soft edge flip-flop based design," in *Proc. IEEE Custom Integrated Circuits Conf.*, Sep. 2008, pp. 543–546.

[16] H. Ando, Y. Yoshida, A. Inoue, I. Sugiyama, T. Asakawa, K. Morita, T. Muta, T. Motokurumada, S. Okada, H. Yamashita, Y. Satsukawa, A. Konmoto, R. Yamashita, and H. Sugiyama, "A 1.3-GHz fifth-generation SPARC64 microprocessor," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1896–1905, Nov. 2003.

[17] S. He and M. Torkelson, "A new approach to pipeline FFT processor," in *Proc. Int. Parallel Processing Symp.*, Apr. 1996, pp. 766–770.

[18] W. Tang and L. Wang, "Cooperative OFDM for energy efficient wireless sensor networks," in *Proc. IEEE Workshop on Signal Processing Systems*, Oct. 2008, pp. 77–82.

[19] D. Zhao, H. Ma, and L. Liu, "Event classification for living environment surveillance using audio sensor networks," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Jul. 2010, pp. 528–533.

[20] E. E. Swartzlander, W. K. W. Young, and S. J. Joseph, "A radix 4 delay commutator for fast Fourier transform processor implementation," *IEEE J. Solid-State Circuits*, vol. 19, no. 5, pp. 702–709, Oct. 1984.

[21] T. Gemmeke, M. Gansen, H. J. Stockmanns, and T. G. Noll, "Design optimization of low-power high-performance DSP building blocks," *IEEE J. Solid-State Circuits*, vol. 39, no. 7, pp. 1131–1139, Jul. 2004.

[22] S. Yoshizawa, K. Nishi, and Y. Miyanaga, "Reconfigurable two-dimensional pipeline FFT processor in OFDM cognitive radio systems," in *Proc. IEEE Int. Symp. Circuits and Systems*, May 2008, pp. 1248–1251.

[23] C.-C. Wang, J.-M. Huang, and H.-C. Cheng, "A 2 K/8 K mode small-area FFT processor for OFDM demodulation of DVB-T receivers," *IEEE Trans. Consumer Electronics*, vol. 51, no. 1, pp. 28–32, Feb. 2005.

[24] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, Mar. 2007.

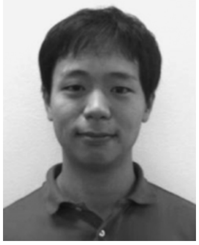
[25] C.-H. Lo and S.-Y. Huang, "P-P-N based 10T SRAM cell for low-leakage and resilient subthreshold operation," *IEEE J. Solid-State Circuits*, vol. 46, no. 3, pp. 520–529, Mar. 2011.

[26] M.-F. Chang, S.-W. Chang, P.-W. Chou, and W.-C. Wu, "A 130 mV SRAM with expanded write and read margins for subthreshold applications," *IEEE J. Solid-State Circuits*, vol. 46, no. 2, pp. 520–529, Feb. 2011.

[27] M. Seok, D. Blaauw, and D. Sylvester, "Clock network design for ultra-low power applications," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2010, pp. 271–276.

[28] 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.

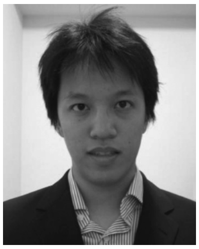
[29] C.-H. Yang, T.-H. Yu, and D. Markovic, "A 5.8 mW 3GPP-LTE compliant 8 × 8 MIMO sphere decoder chip with soft-outputs," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2010, pp. 209–210.



Dongsuk Jeon (S'10) received a B.S degree in electrical engineering from the Seoul National University, South Korea, in 2009. He is currently pursuing a Ph.D. in electrical engineering at the University of Michigan, Ann Arbor.

His research interests include energy efficient signal processing, subthreshold circuit design and error-resilient systems.

Mr. Jeon is the recipient of the Samsung Scholarship for graduate student.



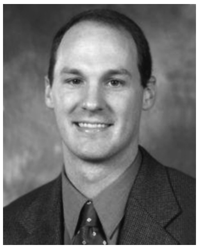
Mingoo Seok (S'05) received the Bachelor degree (summa cum laude) in electrical engineering from Seoul National University, South Korea, in 2005, and the Master degree and PhD degree from University of Michigan in 2007 and 2011, respectively, all in electrical engineering. He is currently a member of technical staff in the Systems and Applications R&D Center of Texas Instruments, Dallas Texas. He will join the EE department of Columbia University in Nov 2011 as an assistant professor.

He has published more than 25 journal and conference papers in his field of research, which includes low power circuit and system design methodologies for improving energy efficiency, performance, and variability both in digital and analog domains. Dr. Seok received 1998–2001 Excelency Fellowship from Seoul National University, 1999 Distinguished Undergraduate Scholarship from the Korea Foundation for Advanced Studies, 2005 Doctoral Fellowship from the same organization, and 2008 Rackham Pre-Doctoral Fellowship from University of Michigan, Ann Arbor. He also won 2009 AMD/CICC Student Scholarship Award for picowatt voltage reference work and 2009 DAC/ISSCC Student Design Contest for the 35 pW sensor platform design (also known as Phoenix Processor). He holds one pending US patent and several invention disclosures.

Chaitali Chakrabarti is a Professor of Electrical Engineering at Arizona State University (ASU), USA. Her research interests are in the areas of low power embedded systems and VLSI architectures for signal processing and communications.

Chakrabarti received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology (IIT), Kharagpur, India, in 1984, and the Ph.D. degree in Electrical Engineering from the University of Maryland, College Park, in 1990. She has received Best Paper Awards at SAMOS'07, MICRO'08 and SiPS'10. She is an Associate Editor of the IEEE TRANSACTIONS ON VLSI SYSTEMS (2007-) and the Journal of VLSI Signal Processing Systems (1999-). She has served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1999–2005), Technical Committee Chair of the Design and Implementation of Signal Processing Systems, IEEE Signal Processing Society, 2006–2007, and a Guest Editor of a two part special issue on Signal Processing on Platforms with Multiple Cores, Nov 2009 and March 2010.

David Blaauw (M'94–SM'07) received his B.S. in Physics and Computer Science from Duke University in 1986, and his Ph.D. in Computer Science from the University of Illinois, Urbana, in 1991. Until August 2001, he worked for Motorola, Inc. in Austin, TX, where he was the manager of the High Performance Design Technology group. Since August 2001, he has been on the faculty at the University of Michigan where he is a Professor. He has published over 350 papers and hold 40 patents. His work has focused on VLSI design with particular emphasis on ultra low power and high performance design. He was the Technical Program Chair and General Chair for the International Symposium on Low Power Electronic and Design. He was also the Technical Program Co-Chair of the ACM/IEEE Design Automation Conference and a member of the ISSCC Technical Program Committee.



Dennis Sylvester (S'95–M'00–SM'04–F'11) received a Ph.D. in electrical engineering from the University of California, Berkeley where his dissertation was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the UC-Berkeley EECS department.

He is a Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor and Director of the Michigan Integrated Circuits Laboratory (MICL), a group of ten faculty and 60+ graduate students. He previously held research staff positions in the Advanced Technology Group of Synopsys, Mountain View, CA, Hewlett-Packard Laboratories in Palo Alto, CA, and a visiting professorship in Electrical and Computer Engineering at the National University of Singapore. He has published over 300 articles along with one book and several book chapters. His research interests include the design of millimeter-scale computing systems and energy efficient near-threshold computing for a range of applications. He holds 7 US patents. He also serves as a consultant and technical advisory board member for electronic design automation and semiconductor firms in these areas. He co-founded Ambiq Micro, a fabless semiconductor company developing ultra-low power mixed-signal solutions for compact wireless devices.

Dr. Sylvester received an NSF CAREER award, the Beatrice Winner Award at ISSCC, an IBM Faculty Award, an SRC Inventor Recognition Award, and eight best paper awards and nominations. He is the recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He has served on the technical program committee of major design automation and circuit design conferences, the executive committee of the ACM/IEEE Design Automation Conference, and the steering committee of the ACM/IEEE International Symposium on Physical Design. He is currently an Associate Editor for IEEE TRANSACTIONS ON CAD and previously served as Associate Editor for IEEE TRANSACTIONS ON VLSI SYSTEMS. He is a member of ACM and Eta Kappa Nu.