A 0.27V, 30MHz, 17.7nJ/transform 1024-pt complex FFT core with super-pipelining

Mingoo Seok, Dongsuk Jeon, Chaitali Chakrabarti[1], David Blaauw, Dennis Sylvester

University of Michigan, Ann Arbor

[1] Arizona State University

Recently, aggressive voltage scaling has been shown to be an important technique in achieving highly energy efficient circuits. Specifically, scaling V_{dd} to near or sub-threshold regions has been proposed for energy-constrained sensor systems to enable long lifetime and small system volume [1][2][4]. However, voltage scaling has several limitations, including significant performance degradation and heightened delay variability due to large I_d sensitivity to PVT variations in the ultra-low voltage (ULV) regime. In addition, energy efficiency degrades below a certain voltage, V_{min}, due to rapidly increasing leakage energy consumption. This paper proposes a new approach to ultra-energy efficient design that uses circuit and architectural methods to further reduce the minimum energy point, or E_{min}, while simultaneously improving performance and robustness. The approaches are demonstrated on an FFT core in 65nm CMOS.

Pipelining is a well-known method to improve performance, typically at the expense of energy consumption due to added sequential elements. However, in this paper, we make the counterintuitive observation that inserting additional pipeline latches improves both energy efficiency and performance in the ULV operating regime. Since pipelining shortens the clock period, it limits leakage energy consumed by idling gates, which reduces energy consumption and allows further voltage scaling. Simulations of inverter chains show that reducing stage depth from 65 to 11 fanout-of-four (FO4) delays yields 36% energy savings and a V_{min} reduction from

0.37V to 0.26V (Fig. 1). By applying this "super-pipelining" approach to the multipliers in an FFT core, we find that it consumes minimum energy when pipelined in 6 stages at a stage depth of 17 FO4 delay. This design approach differs radically from conventional ULV designs, which tend to use limited pipelining and typically have cycles times in the 50-200 FO4 range [1][2]. In this paper, we also show how clocking overhead can be reduced through circuit techniques for facilitating super-pipelining while process variation is addressed through the use of latch-based design. Additionally, architecture modifications are proposed to improve energy efficiency and throughput. Measurements show that the FFT core consumes only 17.7nJ per 1024-pt complex FFT while operating at 30MHz at V_{dd}=0.27V, demonstrating best reported FFT energy efficiency [2][3][4].

An important principle driving the proposed ULV design methodology is to suppress leakage energy, allowing for larger potential energy savings by enabling voltage scaling. We first address this by architectural modifications through minimizing idling modules (Fig. 2). In a traditional memory-based FFT (Fig 2, bottom right), most memory cells idle while a single butterfly unit processes data word by word over many clock cycles. These idling cells waste leakage energy, harming energy efficiency and voltage scalability. On the other hand, conventional pipeline architectures such as MDC (Multi-path Delay Commutator) have high memory utilization but low butterfly unit activity [5]. We therefore modify MDC to accept 4 inputs concurrently by proposing a new commutator configuration, enabling full utilization of both butterflies and memory elements. Additionally, we use two of the modified MDC lanes to double throughput and halve memory counts per lane, reducing leakage energy consumption from commutators. As shown in Fig. 1, these modifications improve energy efficiency and throughput by 2.8× and 6.2×, respectively, compared to a radix-4 memory-based FFT core.

Multipliers in the FFT are super-pipelined as shown in Fig 3. To successfully employ super-pipelining, sequential element overhead must be limited. Six latches share a local clock driver to reduce clock load. The drivers also use minimum-width fingers that enhance drivability at iso-input capacitance due to smaller V_{th} from inverse narrow width effects. Additionally, two latches are embedded in a mirror adder to save two transistors per latch. Latches are upsized from min-width for robustness such that they pass corners and 2 million Monte-Carlo mismatch simulations, providing an estimated 99% chip-level yield with 10k latch instances per chip at 0.2V. We implement the proposed multiplier along with an unpipelined baseline multiplier, separately from the FFT core. Measured results in Fig. 1 shows that the super-pipelined multiplier operates at 18MHz at 0.225V. It is 1.6× faster while consuming 30% less energy than an unpipelined multiplier. It operates 3.6× faster at iso-V_{dd}.

The FIFOs in the commutators contribute as much as 29% of the total FFT energy consumption in this architecture. To address this, we replace the address decoder with a cyclic address generator for reduced energy and use logic-based readout paths for improved performance, as shown in Fig. 5. Simulation results show that the proposed FIFO design consumes 12% lower energy while improving performance by 20% over a memory with MUX-based readout.

Although the above techniques improve energy efficiency and performance, we must pay attention to delay variability and overall design robustness given the ULV design point. We propose the use of 2-phase latches rather than flip-flops. Although the stage depth is drastically reduced in super-pipelined designs, time borrowing removes hard boundaries in the pipeline, re-establishing averaging of process variations along long paths that are present in unpipelined designs. Fig. 1 shows Monte Carlo simulations on latch and flip-flop pipelined multipliers

indicating that a latch pipelined multiplier can absorb delay variations, leading to higher performance yield. In addition, variability-induced hold time violations must also be avoided to ensure functionality. We identify short paths, aided by the regular structure of multipliers, and add delay elements that incur a marginal energy overhead of 2.4% per multiplier. Padded short paths were verified to satisfy hold times using 150k Monte-Carlo simulations under random process variations and corners.

The clock distribution network is uniquely designed to suppress process variation induced skew and resulting hold time violations. Conventionally, many clock buffers are used to mitigate RC mismatch. However, at low V_{dd} the mismatch in these buffers is exacerbated and contributes significant skew, while RC delay is small compared to gate delay. Therefore, we design a 3-level clock network where reduced number of large buffers and matched RC interconnect are used. The lowest and middle levels of clock network are implemented with minimum width thin interconnect while the top level uses thick metal interconnect for lower RC delay and better slew. Fig. 4 shows that the simulated worst-case RC mismatch is less than 0.15ns (0.2×FO4 at V_{dd}=0.3V).

The FFT core is fabricated in 65nm CMOS using the above circuit and architectural techniques. Measurements in Fig. 6 show that it computes 234k 16b 1024-pt complex FFT per second. The clock frequency is measured as 30MHz with V_{dd}=0.27V compared to frequencies of 10's of KHz for typical ULV designs at the same supply voltage. The FFT consumes 17.7nJ/transform, which is 4× smaller than prior work when scaled for FFT size and technology [3]. A die photograph is shown in Fig. 7.

Acknowledgement

The IC fabrication support of STMicroelectronics is gratefully acknowledged.

References

[1] G. K. Chen et al., "Millimeter-Scale Nearly-Perpetual Sensor System with Stacked Battery and Solar Cells," *IEEE International Solid-State Circuits Conference*, 2010.

[2] A. Wang et al., "A 180mV FFT Processor using Subthreshold Circuit Techniques," *IEEE International Solid-State Circuits Conference*, 2004.

[3] Y. Chen et al., "A 2.4-Gsample/s DVFS FFT Processor for MIMO OFDM Communication Systems," *IEEE Journal of Solid-State Circuits*, vol.43, no.5, May 2008.

[4] S. Sridhara et al., "Microwatt Embedded Processor Platform for Medical System-on-Chip Applications," *IEEE Symposium on VLSI Circuits*, pp.15-16, 2010.

[5] Y. Jung et al., "New Efficient FFT Algorithm and Pipeline Implementation Results for OFDM/DMT Applications," *IEEE Transactions on Consumer Electronics*, vol.49, pp.14-20, Feb. 2003.

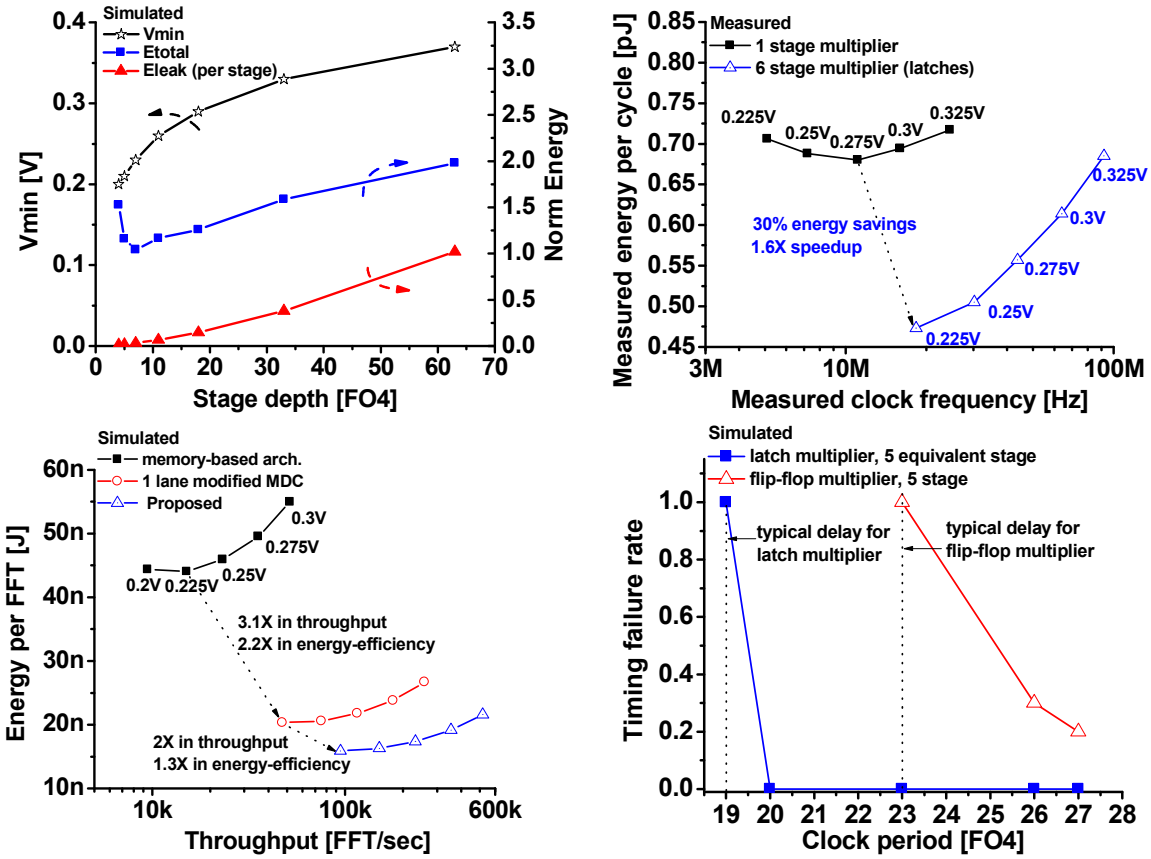


Figure 1. Simulations with 60×FO4 delay inverter chains show that pipelining can improve energy efficiency by reducing leakage energy consumption, also enabling greater voltage scalability by shifting V_{min} down (upper left). The measured energy consumption and clock frequency of an unpipelined and proposed multiplier is shown (upper right). Simulated results show that architecture modifications improve both energy efficiency and throughput (lower left). The timing failure rates across Monte-Carlo simulations with random process variations are shown for two pipelined multipliers. Latches provide lower timing failure rate along with shorter clock period by mitigating stage imbalances. (lower right)

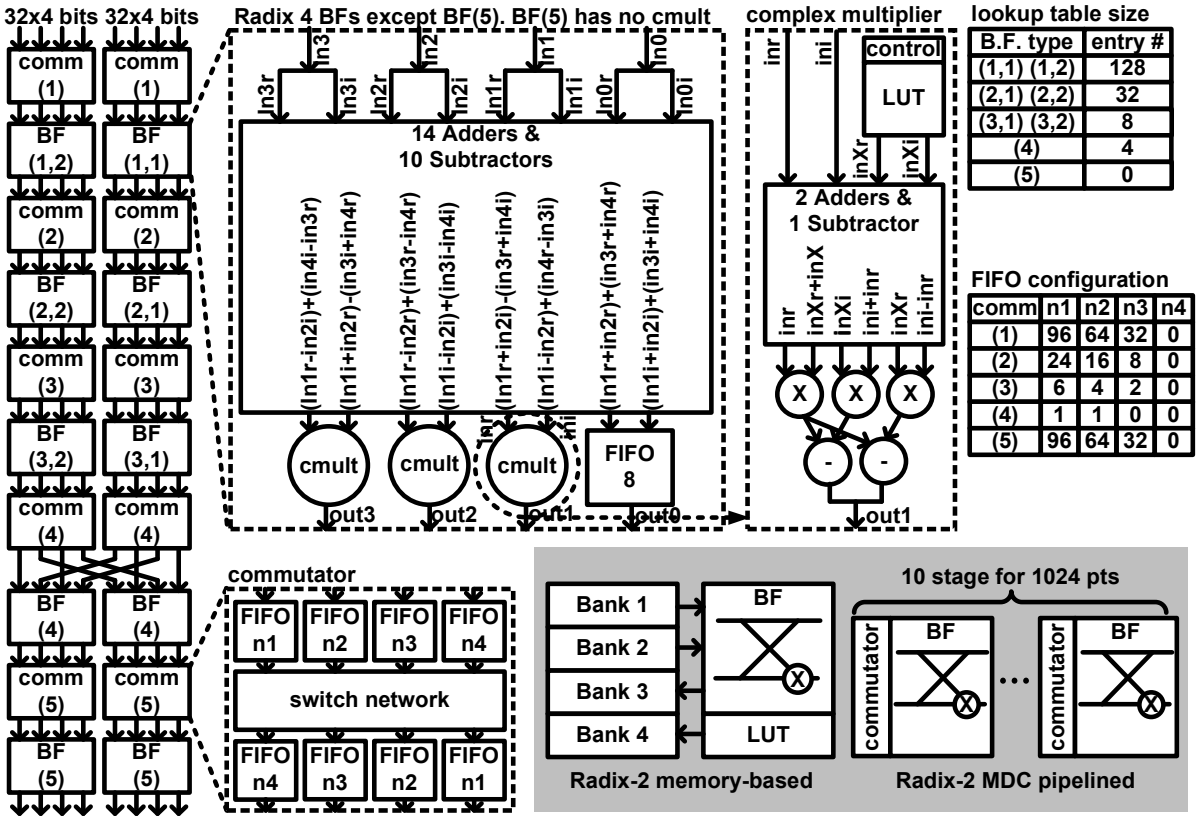


Figure 2. A pipelined, 8×32b input, radix-4, 2-lane, 1024-pt, complex FFT architecture is proposed. Each butterfly except BF(5) has 3 complex multipliers, 32b-8w FIFO and 24 adders/subtractors. BF(5) has no complex multipliers. A basic memory-based architecture and MDC pipelined architecture are shown for references.

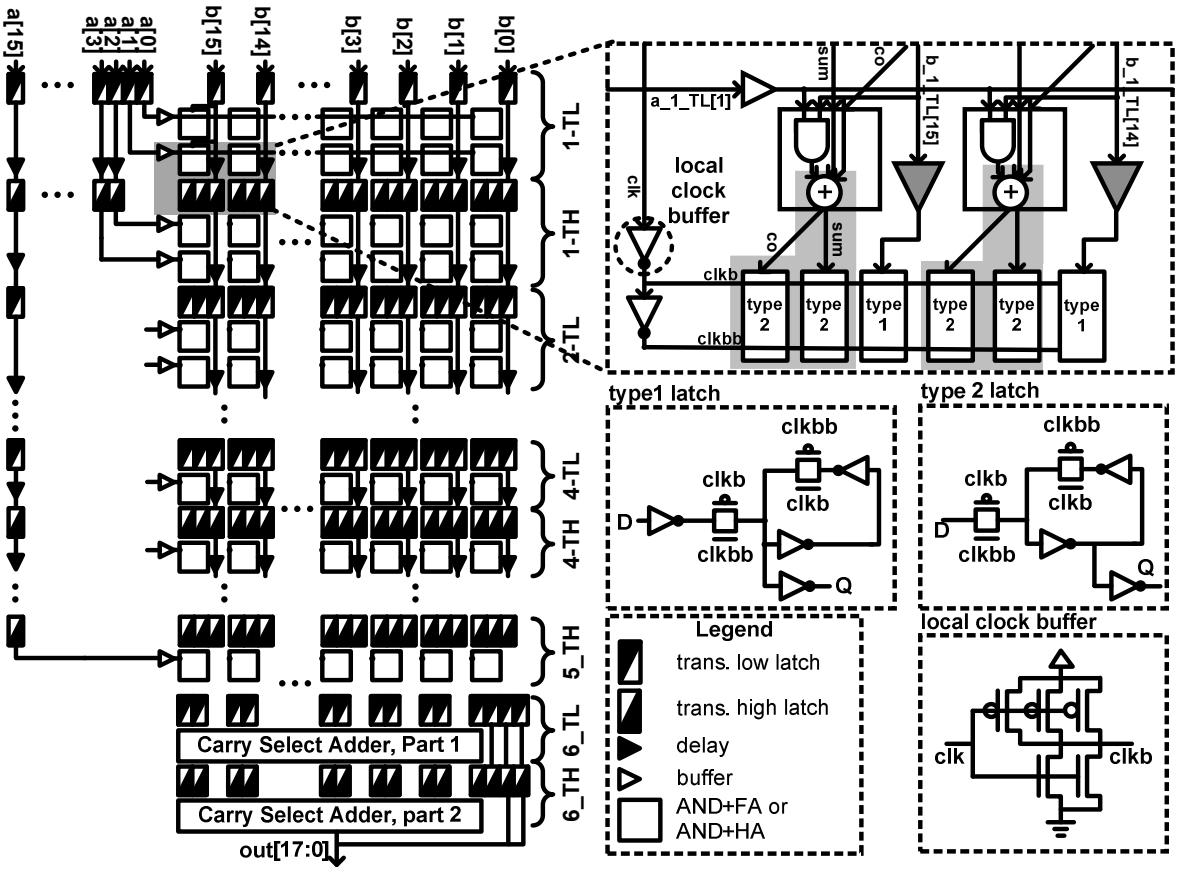


Figure 3. A 16b Baugh-Wooley multiplier is super-pipelined with 2-phase latches. It uses 12 banks of latches, equivalent to 6 flip-flop pipeline stages. Variable length (5-4-3-2-2) carry select adder is used for accumulation. For super-pipelining, latches are energy-optimized by sharing a clock buffer and embedding in FA cells.

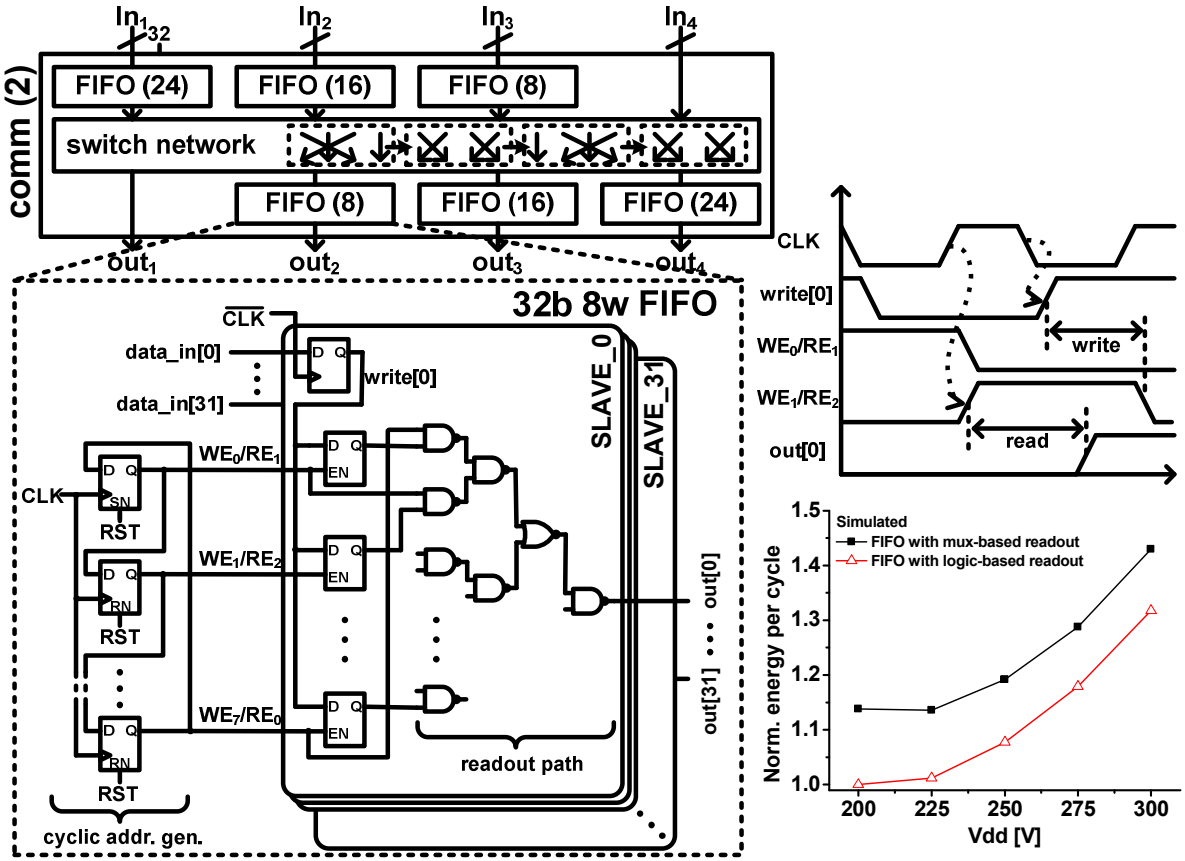
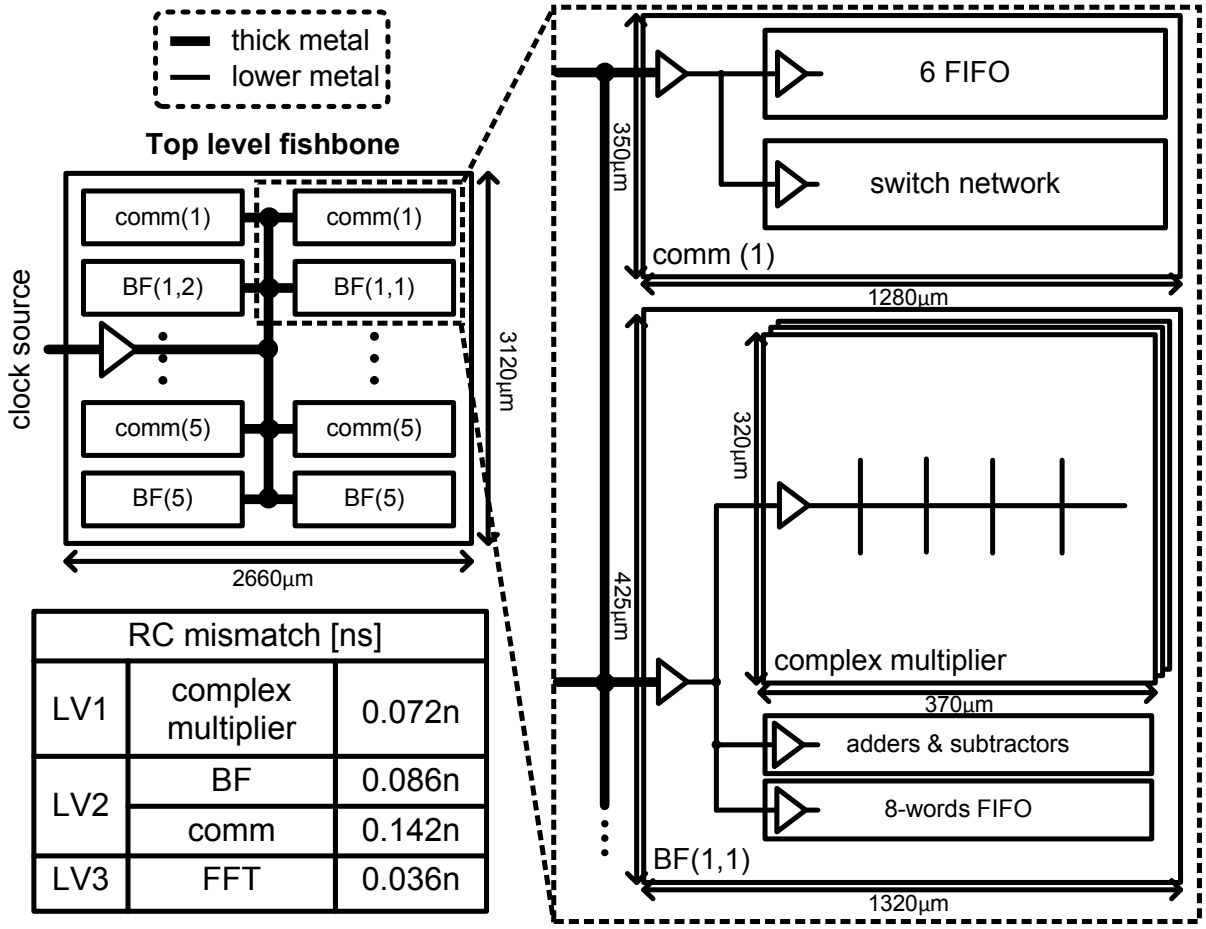
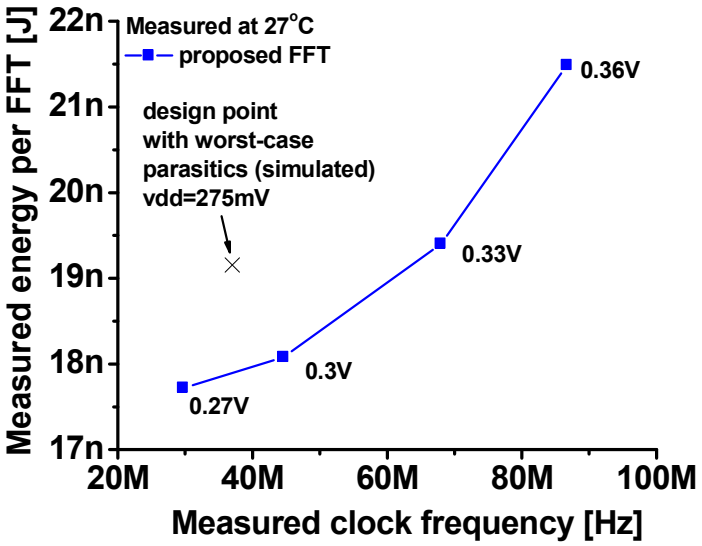


Figure 4. A commutator consists of a switch network and 4-6 FIFOs. The switch network changes configuration for transferring proper data. FIFOs are optimized with a cyclic address pointer and logic-readout path. Positive-edge read and negative-edge write operations for preventing hold time violation are described (upper right). Simulations of proposed FIFO show 12% lower energy compared to MUX-based readout memory (lower right).



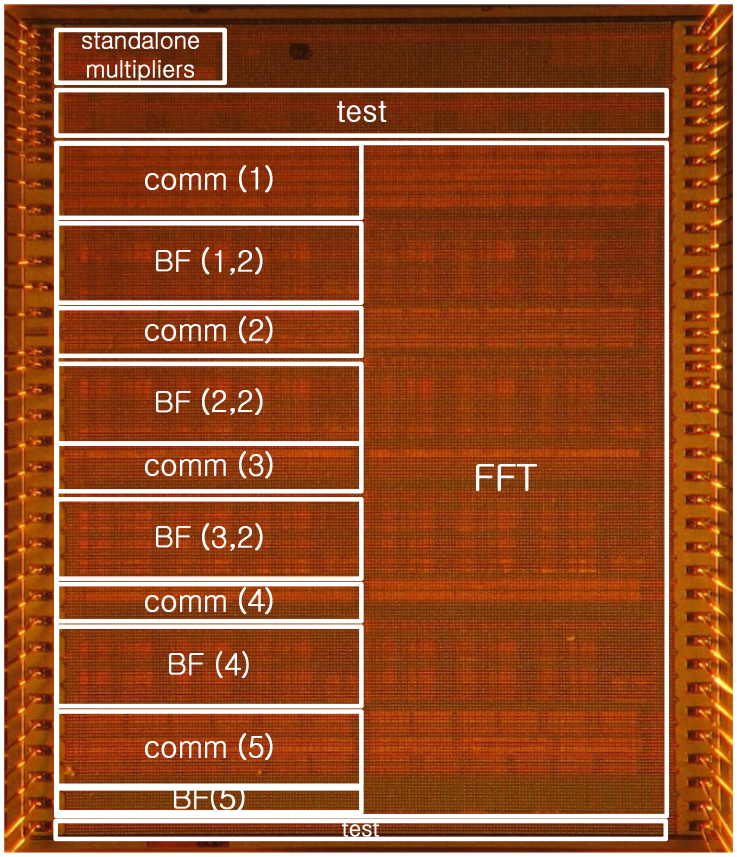
RC mismatch [ns]		
LV1	complex multiplier	0.072n
LV2	BF	0.086n
	comm	0.142n
LV3	FFT	0.036n

Figure 5. The clock network is designed with limited buffers and matched interconnect to address key ULV skew sources. Blocks up to ~0.1mm^2 use a single large buffer to eliminate the contribution of PVT-induced gate delay variation on skew. A fishbone clock network is designed for the top-level (LV3) in thick M6 and M7 to reduce RC mismatch and improve slew rates.



	Proposed	[2]	[3]	[4]
Technology	65nm	180nm	90nm	130nm
Low power config	1024-pt CV 0.27V, 30MHz	1024-pt RV 0.35V, 10KHz	256-pt CV 0.85V, 300MHz	256-pt RV 0.5V, 7KHz
Energy/FFT	17.7 nJ	155 nJ	12.8 nJ	100 nJ
Norm. energy/FFT	**17.7 nJ**	**111.9 nJ**	**71.0 nJ**	**400 nJ**

Figure 6. Top plot shows the measured energy consumption and clock frequency of the proposed FFT core along with a simulated design point. Bottom comparison table shows that the proposed design improves energy efficiency by 4.2× over prior work. For normalizing energy consumption to technology and word width, we apply the scaling formula in [3]. We assume that complex FFT operations consume 2× the energy of real FFT operations. Additionally, energy consumption is conservatively increased by 4× from 256 point FFT to 1024 point FFT.



Type	1024pts complex
Vdd	270mV
Clock Freq.	30MHz
Energy / FFT	17.7nJ
Throughput	234k FFT/sec
Area	2.66x3.12mm^2
Technology	65nm CMOS

Figure 7. Die photo of the FFT core implemented in 65nm CMOS with a summary table.

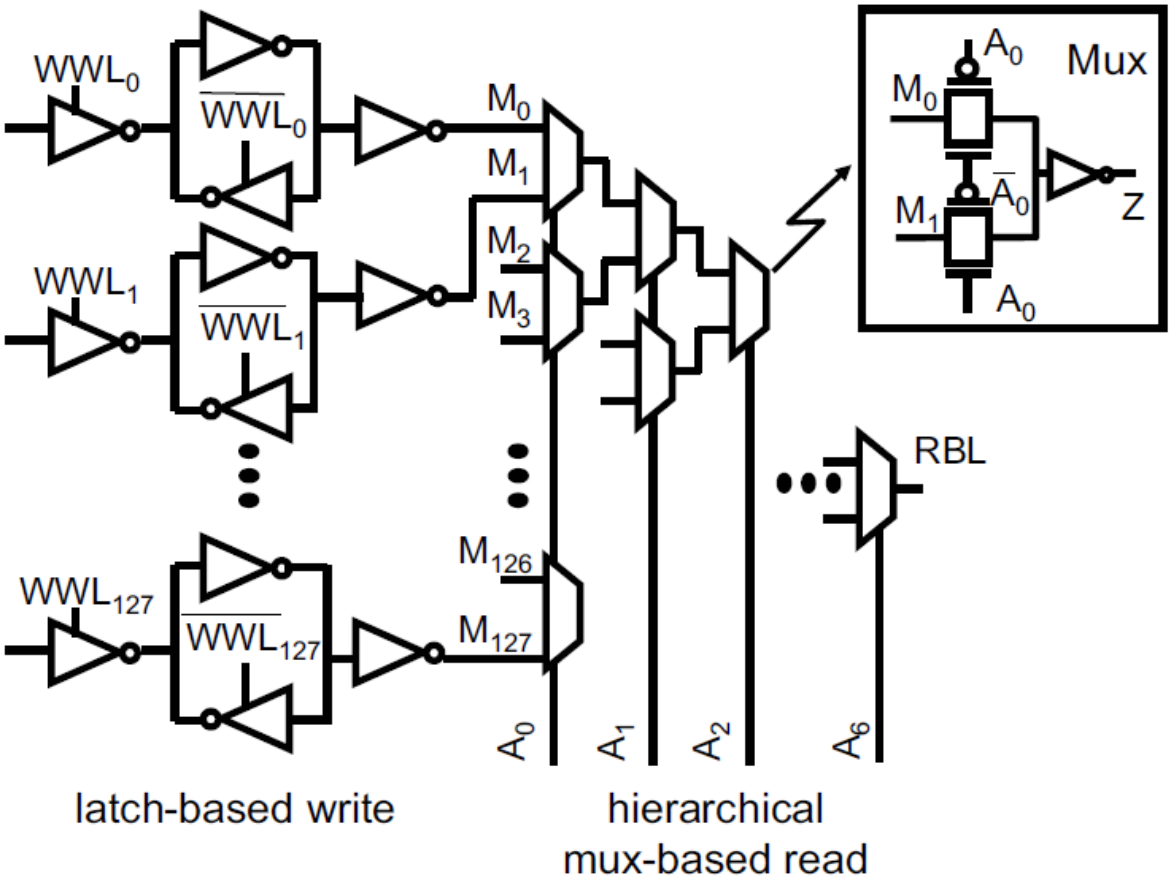


Figure S1. A memory with MUX-based readout is shown, copied from [2]. To be used as FIFO, it requires an address decoder and a cyclic counter for write operations, which induces energy overhead. Decoders also increase write delay. In addition, MUX-based readout path is slower than logic-based path in simulations, which limits read performance.