Near- and Sub-V_t Pipelines Based on Wide-Pulsed-Latch Design Techniques

Wei Jin, *Student Member, IEEE*, Seongjong Kim, Weifeng He, Zhigang Mao, *Member, IEEE*, and Mingoo Seok, *Member, IEEE*

***Abstract*— This paper presents a methodology and chip demonstration to design near-/sub-threshold voltage (V_t) pipelines using pulsed latches that are clocked at very wide pulses. Pulsed-latch-based design is known for time borrowing capability but the amount of time borrowing is limited due to hold time constraint. To enable more cycle borrowing, in this paper, we aim to pad short paths to ∼1/3 cycle time using multi-V_t cell library. While delay padding using multi-V_t cells is common in super-V_t design, the small delay difference among multi-V_t cells has not allowed such extensive short path padding due to large area overhead. However, in near-/sub-V_t regime, circuits delay becomes exponentially sensitive to V_t, suggesting that high-V_t cells can significantly reduce the overhead of padding. We build a semi-automatic short path padding flow around this idea, and use it to design: 1) ISCAS benchmark circuits and 2) an 8-bit 8-tap finite impulse response (FIR) core, the latter fabricated in a 65-nm CMOS technology. The chip measurement shows that the proposed FIR core achieves 45.2% throughput (frequency), 11% energy efficiency (Energy/cycle), and 38% energy-delay-product improvements at 0.35 V over the flip-flop-pipelined baseline. The measurement results also confirm that the proposed FIR core operates with the same pulsewidth setting robustly across process, voltage, and temperature variations.**

***Index Terms*— Energy efficiency, finite impulse response (FIR), frequency improvement, multi-V_t, near-/sub-threshold, pulsed-latches, time borrowing.**

I. Introduction

NEAR-/sub-threshold voltage (V_t) circuit design is one of the promising approaches for increasing energy efficiency in digital systems. The key challenges in this approach are large performance degradation and extreme variability. Among several recent efforts to mitigate those challenges, the aggressive two-phase latch-based sequencing was used for time borrowing, which can improve performance and tolerance of the variability [1], [2]. However, two-phase sequencing has an inherently larger sequential overhead than flop-based design [3]. Additionally, the logic depth per latch stage is reduced by half after re-timing, which can exacerbate the impact of local variations.

To overcome the shortcomings above, the pulsed-latch can be an attractive option in designing pipelines for near-/sub-V_t digital circuits because of its lower sequential overhead and larger variation tolerance than the edge-triggered flip-flops (FFs) [4]–[7]. Some previous works [8]–[16], [33]–[37] researched on pulsed-latch-based circuits design. Compared with the two-phase latch-based design [30]–[32], the pulsed-latch-based design can consume smaller sequential logic area (while ignoring pulse generation and distribution overhead). Also, there is no need to perform re-timing during pipeline design. We can simply replace all the FFs with pulsed-latches to immigrate from an FF-based circuit to a pulsed-latch-based circuit. In addition, the logic depth per latch stage is reduced by half after re-timing, which can exacerbate the impact of local variations in the two-phase latch-based pipeline. Furthermore, the pulsed-latch-based design keeps the time borrowing ability.

However, the hold time constraint limits the pulsewidth. For example, one of the existing studies used narrow pulses whose width is <∼5 Fan-Out-of-4 (FO4) delays [5]. As a result, the time borrowing ability and variation tolerance of pulsed-latch based pipelines are limited. The small amount of time borrowing ability only allows roughly a 10% decrease in the clock period [5], which is smaller than the level that a two-phase latch pipeline can support [17] (half the cycle time [T_C]). To enable more amount of timing borrowing, pulsewidth allocation combined with clock skew scheduling or re-timing was proposed in previous works [8]–[16]. These techniques can decrease the clock period by an additional 20% [5]. However, the large worst case delay variability of near-/sub-V_t circuits makes it challenging to apply these techniques.

Moreover, the pulsewidth exacerbates the clock network design. It is a challenging task to distribute narrow pulses as the slew can be easily degraded when a pulse travels through a network. To guarantee the integrity of the narrow pulse shape, the design of pulsers (pulse generators) and placement of pulsed-latches should be carefully performed [5]. Coupled with wire parasitics, the chip requires strong buffers in its pulse distribution network and local pulse generators [5], [18], incurring large area and power overhead.

In this paper, we propose a methodology to design pulsed-latch-based pipelines for near-/sub-V_t operation with more cycle borrowing yet low overhead [21]. Specifically, we pursue

Manuscript received February 1, 2017; revised May 9, 2017 and June 9, 2017; accepted June 9, 2017. Date of publication July 27, 2017; date of current version August 22, 2017. This paper was approved by Associate Editor Marian Verhelst. This work was supported in part by the U.S. National Science Foundation under Grant 1453142, in part by DARPA under Grant HR0011-13-C-0003, in part by the Catalyst Foundation, in part by the China Scholarship Council under Grant 201406230166, and in part by the Hi-Tech Research and Development Program (863) of China under Grant 2012AA012702. *(Corresponding author: Wei Jin.)*

W. Jin is with the School of Microelectronics, Shanghai Jiao Tong University, Shanghai 200240, China, and also with Columbia University, New York, NY 10027 USA (e-mail: kings2005@sjtu.edu.cn).

S. Kim and M. Seok are with Columbia University, New York, NY 10027 USA (e-mail: sk3667@columbia.edu; mgseok@ee.columbia.edu).

W. He and Z. Mao are with the School of Microelectronics, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: hewf@sjtu.edu.cn; maozhigang@sjtu.edu.cn).

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.2017.2717927

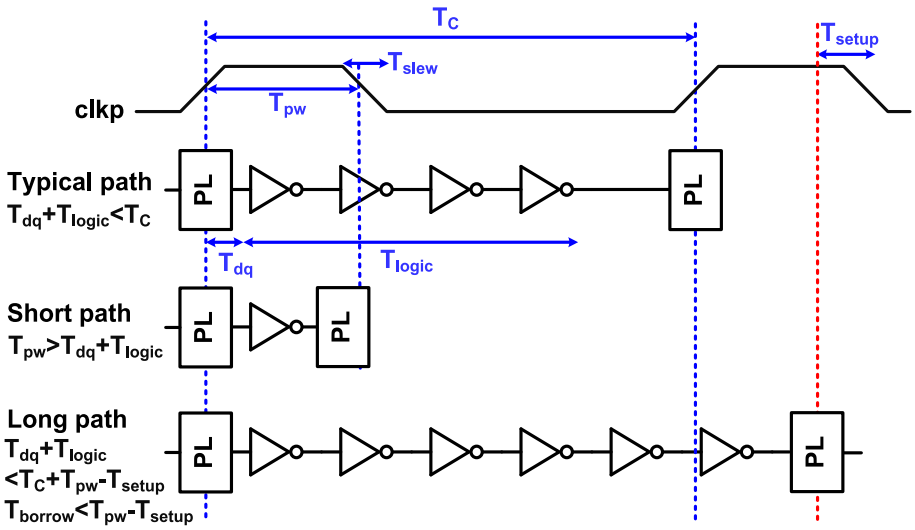


Fig. 1. Concept of pulsed-latch sequencing with the setup and hold time constraints and the case for time borrowing.

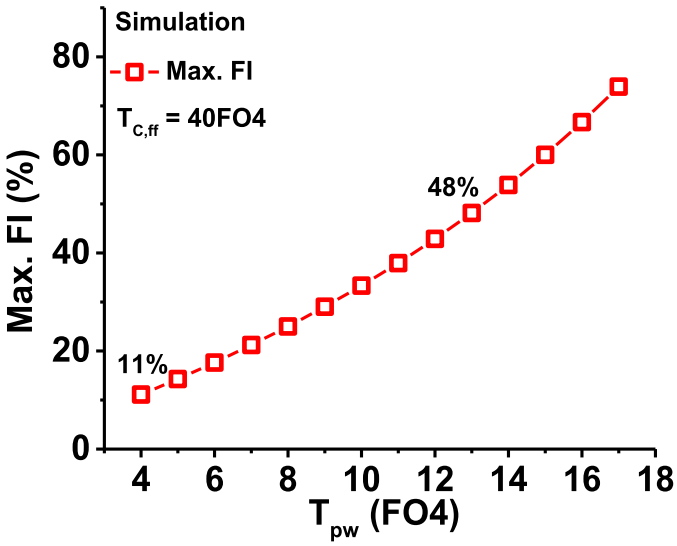


Fig. 2. Max. FI with different T_{pw} values.

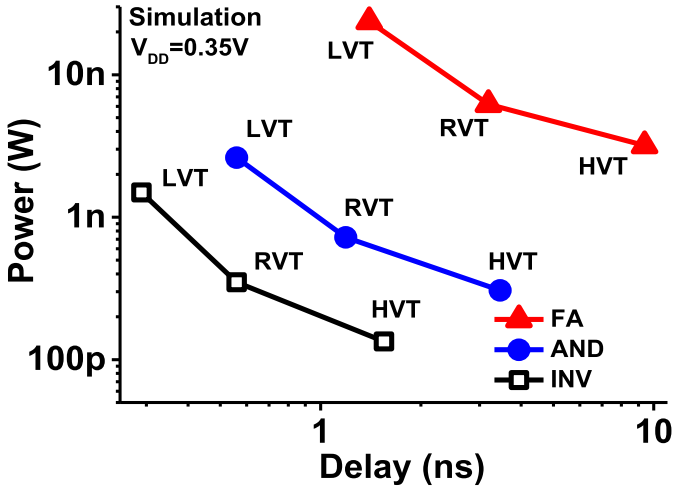


Fig. 3. Delay and power of multi-V_t cells. The high-V_t cells have significantly longer delay and less power than reg-V_t and low-V_t cells.

to enable the use of wider pulses in pipelines by padding short paths using multi-V_t cells. Padding with multi-V_t cells is not a new practice in super-V_t pipeline design. However, as the delay difference between regular-V_t and high-V_t cells is not significant in super-V_t digital circuits, it is unusual to pad short paths more than a fraction of cycle time, e.g., 5 FO4 delays. In this paper, we take advantage of the large delay difference between multiple V_t cells in near-/sub-V_t circuits, and pursue to pad short paths to ~1/3 of clock cycle time. This enables a proportionally larger amount of cycle borrowing per stage, thereby improving cycle time by ~33%. Moreover, the reduction of cycle time also saves active-leakage energy dissipation, since it equals to the product of leakage power and cycle time when circuits are in the active mode. The overhead of the proposed multi-V_t padding is 2.7 times less than that of the regular-V_t padding.

We also devise a semi-automatic short path padding flow using commercial-grades logic-synthesis and automatic-placement-and-route (APR) tools. We then apply the flow to implement the ISCAS benchmark circuits. The results show that our methodology reduces the average area overhead of short path padding by 2.71 times as compared with the padding using single-V_t cells. Furthermore, the wide pulsewidth simplifies clock network design, which could require four times more buffers if narrow pulses had to be used.

Based on the earlier proposed techniques, we implement an 8-bit finite impulse response (FIR) core in a 65-nm general-purpose CMOS process. The measurement shows that the core operates robustly with 12 FO4 delay wide pulses (~1/3 of T_C of the core) across process, voltage, and temperature (PVT) variations. With such large cycle-borrowing capability, the core also achieves 45.2% better throughput, 11% less Energy/cycle, and 38% smaller energy-delay-product (EDP) at 0.35 V, as compared with the edge-triggered FF design. The area overhead is 15%.

This paper is organized as follows. In Section II, we describe our proposed techniques and the design of pulsed-latch pipelined 8-bit FIR core, including the design principle, short path padding method, clock network distribution, and FIR core design. Section III presents the measurement results of the test chips, including the measurements and comparisons across PVT variations. Finally, the conclusions are drawn in Section IV.

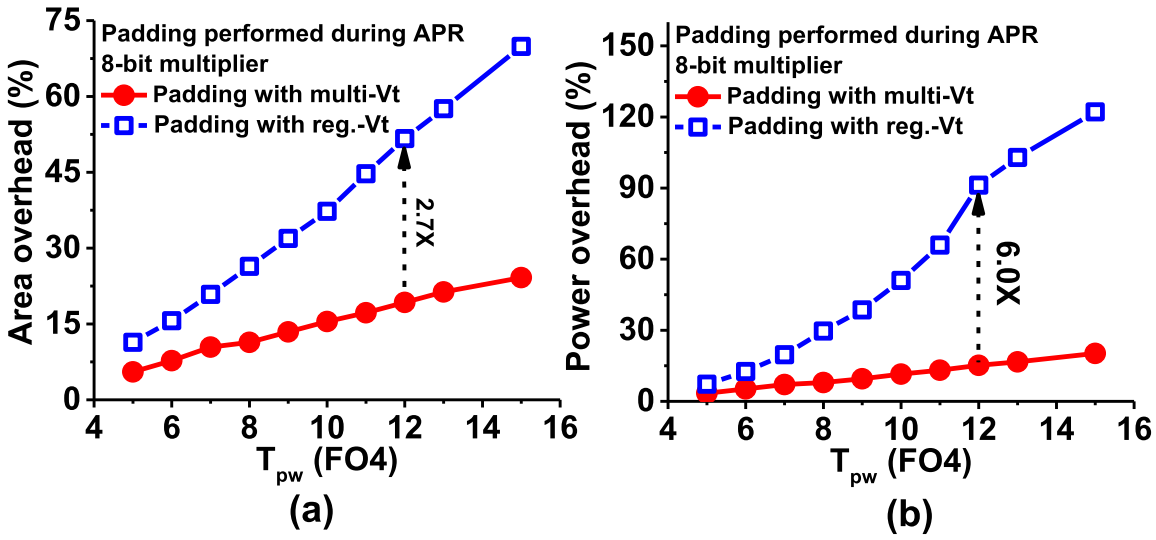


Fig. 4. (a) Area overhead and (b) power overhead of padding on an 8-bit multiplier with different T_{pw} values.

TABLE I

SHORT PATH PADDING AREA OVERHEAD COMPARISON WITH ISCAS BENCHMARKS

	gate count			area (μm^2)			area overhead (%)		area overhead reduction from single-V_t to multi-V_t (ratio)
	org	padding (multi-V_t)	padding (single-V_t)	org	padding (multi-V_t)	padding (single-V_t)	padding (multi-V_t)	padding (single-V_t)	
ISCAS'85									
c432	101	331	618	180	503.28	848.16	179.60	371.20	**2.07**
c499	258	643	902	526.68	1103.76	1542.96	109.57	192.96	**1.76**
c880	264	495	627	458.64	816.12	1095.48	77.94	138.85	**1.78**
c1908	269	515	969	493.56	909.72	1607.4	84.32	225.67	**2.68**
c2670	402	868	936	723.24	1701	2397.24	135.19	231.46	**1.71**
c3540	720	1044	1144	1257.48	1748.16	1977.84	39.02	57.29	**1.47**
c5315	938	1929	2364	1725.12	3289.68	4280.76	90.69	148.14	**1.63**
c6288	723	1505	1623	2665.44	3970.44	4330.44	48.96	62.47	**1.28**
c7552	970	1649	2128	1875.96	3068.64	4158.72	63.58	121.68	**1.91**
ISCAS'89									
s344	93	114	124	244.08	281.88	339.12	15.49	38.94	**2.51**
s510	201	202	252	353.88	371.16	428.04	4.88	20.96	**4.29**
s713	137	223	275	278.64	405.72	526.68	45.61	89.02	**1.95**
s953	364	433	510	720.72	833.76	1044.72	15.68	44.96	**2.87**
s1196	477	567	649	849.24	982.8	1136.16	15.73	33.79	**2.15**
s1238	472	597	640	845.64	1035.72	1134.72	22.48	34.18	**1.52**
s1488	508	530	579	849.24	894.6	964.8	5.34	13.61	**2.55**
s5378	1028	1103	1282	2576.88	2744.28	3040.2	6.50	17.98	**2.77**
s9234	833	868	915	2081.88	2128.32	2332.08	2.23	12.02	**5.39**
s13207	891	977	975	2665.08	2825.28	3084.48	6.01	15.74	**2.62**
s15850	2373	2957	2815	6490.44	7552.8	7765.56	16.37	19.65	**1.20**
Avg.	-	-	-	-	-	-	14.21	30.98	**2.71**

II. WIDE-PULSED-LATCH-BASED PIPELINES

A. Design Principle

Fig. 1 shows the concept of the classical pulsed-latch-based pipeline. The conventional standard latches are used in the pipeline, and all the latches in a pipeline are driven by a single-phase clock (clkp), whose high-phase time is defined as T_{pw}. In a typical logic path, the sum of the D-to-Q delay of a latch (T_{dq}) and logic delay (T_{logic}) should be smaller than T_C. If a path has a delay longer than T_C, however, the path may borrow time from the next stage. The maximum time that can be borrowed from the next stage (T_{borrow}) is constrained

TABLE II
SHORT PATH PADDING POWER OVERHEAD COMPARISON WITH ISCAS BENCHMARKS

	gate count			power (nW)			power overhead (%)		power overhead reduction from single-V_t to multi-V_t (%)
	org	padding (multi-V_t)	padding (single-V_t)	org	padding (multi-V_t)	padding (single-V_t)	padding (multi-V_t)	padding (single-V_t)	
ISCAS'85									
c432	101	331	618	33.80	60.30	72.64	78.40	114.91	**36.51**
c499	258	643	902	69.70	105.00	135.20	50.65	93.97	**43.33**
c880	264	495	627	80.21	80.27	118.70	0.07	47.99	**47.91**
c1908	269	515	969	49.43	66.72	107.00	34.98	116.47	**81.49**
c2670	402	868	936	127.10	155.10	292.20	22.03	129.90	**107.87**
c3540	720	1044	1144	145.00	141.00	178.60	-2.76	23.17	**25.93**
c5315	938	1929	2364	252.30	282.80	398.70	12.09	58.03	**45.94**
c6288	723	1505	1623	117.60	152.80	168.10	29.93	42.94	**13.01**
c7552	970	1649	2128	211.40	210.50	315.90	-0.43	49.43	**49.86**
ISCAS'89									
s344	93	114	124	5.61	5.36	6.99	-4.46	24.70	**29.16**
s510	201	202	252	2.52	2.31	2.84	-8.22	12.87	**21.09**
s713	137	223	275	8.06	8.07	10.17	0.19	26.19	**26.01**
s953	364	433	510	15.04	14.32	17.83	-4.79	18.55	**23.34**
s1196	477	567	649	23.90	24.10	27.08	0.84	13.31	**12.47**
s1238	472	597	640	23.64	23.43	25.77	-0.89	9.01	**9.90**
s1488	508	530	579	11.59	10.79	13.01	-6.90	12.25	**19.15**
s5378	1028	1103	1282	56.27	55.02	61.43	-2.22	9.17	**11.39**
s9234	833	868	915	31.84	32.12	36.74	0.88	15.39	**14.51**
s13207	891	977	975	60.77	59.76	68.76	-1.66	13.15	**14.81**
s15850	2373	2957	2815	69.23	69.00	80.77	-0.33	16.67	**17.00**
Avg.	-	-	-	-	-	-	9.87	42.40	**32.53 (4.3X)**

by T_{pw} minus the setup time of a latch (T_{setup}). This implies that wider pulsewidth gives more time to borrow. However, the maximum T_{pw} is also constrained by the delay of the shortest path in a pipeline, because the delay of any path less than T_{pw} (named short path) will cause hold time violation. This constraint can be illustrated as follows:

$$T_{\text{pw}} < T_{\text{logic}} + T_{\text{dq}}. \tag{1}$$

Although it is desirable to pad a short path longer (i.e., making its delay longer by adding delay cells) and enable more time borrowing, excessive short path padding causes large area overhead. Our experiment using an 8-bit multiplier shows that the short path padding to 12 FO4 delays, which is about 1/3 T_C of the multiplier, can cause 51.7% area overhead. Because of this severe overhead, narrow pulses (<5–6 FO4 delays [5]) are typically used in conventional pulsed-latch pipelines [5], [27], [28].

Indeed, the performance improvement from cycle borrowing is proportional to T_{pw}. The cycle time of a pulsed-latch pipeline, $T_{C,\text{pulse}}$, can be derived as

$$\text{Min} \cdot T_{C,\text{pulse}} = T_{C,\text{ff}} - T_{\text{pw}} \tag{2}$$

where the cycle time of an FF pipelined baseline is $T_{C,\text{ff}}$. Hence, the maximum frequency improvement (Max. FI) over the baseline can be formulated as

$$\text{Max} \cdot \text{FI} = \frac{F_{\text{pulse}} - F_{\text{base}}}{\text{F}_{\text{base}}} = \frac{\frac{1}{T_{C,\text{ff}} - T_{\text{pw}}} - \frac{1}{T_{C,\text{ff}}}}{\frac{1}{T_{C,\text{ff}}}} = \frac{T_{\text{pw}}}{T_{C,\text{ff}} - T_{\text{pw}}} \tag{3}$$

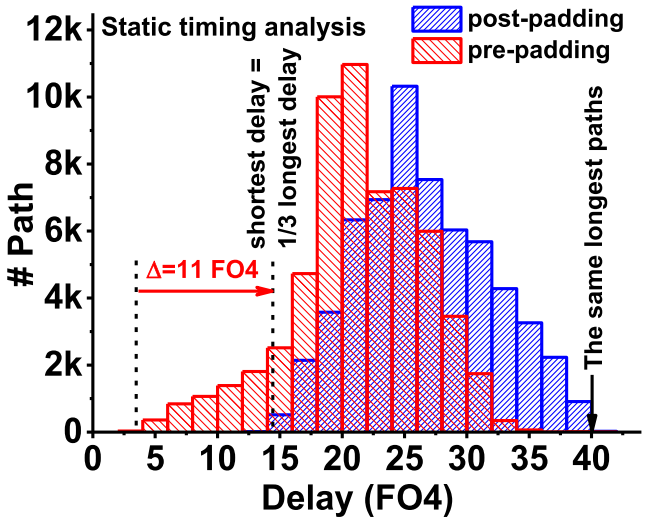


Fig. 5. Path delay distribution of an 8-bit multiplier of pre- and post-padding using static timing analysis.

where F_{base} is the frequency of the FF pipelined baseline, and F_{pulse} is the frequency of the pulsed-latch pipeline. Fig. 2 shows the maximum FI curve across different T_{pw} values when $T_{C,\text{ff}}$ is 40 FO4 delays. For a design with $T_{\text{pw}} = 1/10\ T_{C,\text{ff}}$, the maximum FI can be 11%; while for a design with $T_{\text{pw}} = 1/3\ T_{C,\text{ff}}$, the maximum FI can be up to 50%. This promises a significant amount of performance improvement by using wide pulses in pulsed-latch-based pipelines.

The above frequency improvement analysis considers only one path and if one path borrows time from the next stage, the next stage will have a more strict constraint. In practice, however, it is very rare an input exercises only critical paths across all the pipelines [29], which the pulsed-latch pipeline can take advantage of the frequency improvement in the entire pipeline.

B. Proposed Short Path Padding Technique With Multi-V_t Cells

To reduce the overhead for short path padding, we exploited exponential relationship between V_t and delay in near-/sub-V_t circuits. We simulated the delay and power of three typical logic cells, i.e., full adder, AND gate, and inverter, with different V_t values. As shown in Fig. 3, high-V_t cells have significantly longer delay and smaller power dissipation[1] than regular-V_t and low-V_t cells. This implies that we can use fewer high-V_t cells than regular-V_t and low-V_t cells for the same amount of short path padding, saving both area and power. While multi-V_t design is well known for super-V_t circuits, we find that the use of the multi-V_t library for short-path padding is very effective particularly for near- and sub-V_t circuits, since high-V_t cells are exponentially slower than regular- and low-V_t cells. With the multi-V_t library, we can have a large amount of padding at low area and power overhead, allowing the use of wider pulses. This also alleviates the overhead associated with pulse generation and distribution.

We performed an APR experiment on an 8-bit multiplier. In the experiment, we padded all short paths of the multiplier with different T_{pw} values and analyzed the area overhead. As shown in Fig. 4, at 12 FO4 delay T_{pw}, the padding area and power overhead are 51.7% and 91.2% with regular-V_t cells; while the padding area and power overhead are only 19.3% area and 15.2% with multi-V_t cells. This marks 2.7 times area and 6 times power reduction. Fig. 5 shows the path delay distribution before and after the short path padding. Before padding, the shortest path delay is originally 3 FO4 delays; however, after delay padding, the shortest path delay is extended to 14 FO4 delays (about 1/3 of the cycle time) without compromising the critical path delay.

C. Design Flow for Short Path Padding

The overhead of padding is sensitive to circuit structures. To verify the effectiveness of our proposed techniques across a range of various circuits, we devised a semi-automatic flow to perform the multi-V_t-based short path padding in near-/sub-V_t circuits. Fig. 6 shows the flowchart of the proposed flow. It starts with a logic synthesis tool (Synopsys Design Compiler) with constraints on both critical paths and short paths. The synthesis tool uses a multi-V_t library, which we characterize at $V_{\text{DD}} = 0.35$ V using Cadence Encounter Library Characterization. Then, we perform another iteration of short path padding in the APR phase using Cadence Encounter. This iteration is critical to consider the impact of parasitics on timing. Then, we perform static-timing analysis to determine whether the timing constraints are met. The short paths information and long paths information are generated using Synopsys PrimeTime. If the timing constraints are not met, we iterate the above process until they become met.

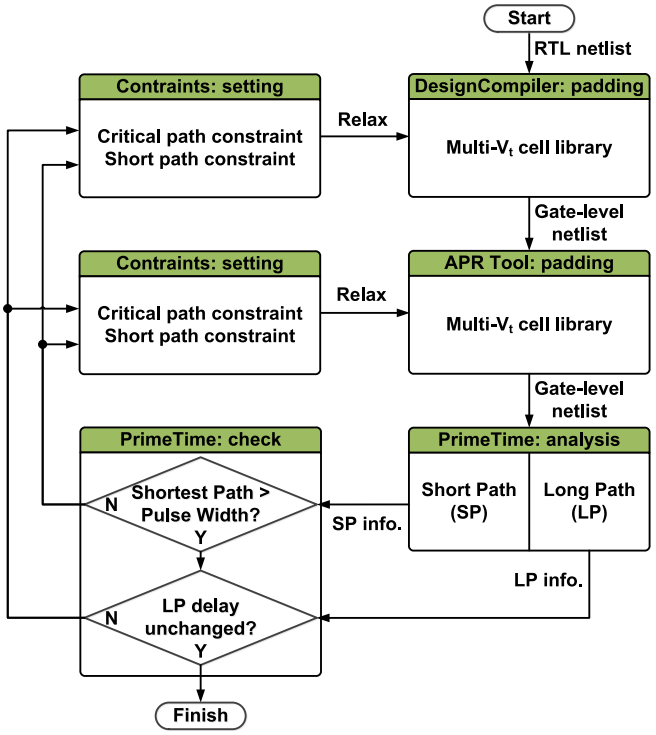


Fig. 6. Semi-automatic flowchart for the short path padding process with a multi-V_t cell library.

With this flow, we performed padding on nine circuits from the ISCAS'85 and 11 circuits from the ISCAS'89. The total 20 circuits contain both combinational and sequential logics. For each benchmark, we set the target T_{pw} to 1/3 T_C of the original circuits (org). Tables I and II summarize the area and power overhead and other metrics. Our proposed multi-V_t padding achieves 14.21% (area) and 9.87% (power) average overhead across 20 benchmark circuits while padding using single-V_t cells exhibits 30.98% (area) and 42.4% (power) average overhead. Consequentially, the average area and power overhead reduction from padding with single-V_t cells to that padding with multi-V_t cells are 2.71 times and 4.3 times, respectively.

As a summary, the use of multi-V_t library for padding enables significant reduction in padding overhead. This large degree of improvement is unique to only near-/sub-V_t circuits, since high-V_t cells become much slower relative to regular- and low-V_t cells only in near-/sub-V_t circuits.

[1]High-V_t devices consume less, since they leak less and their gate capacitance reduces due to smaller channel-to-substrate capacitance.

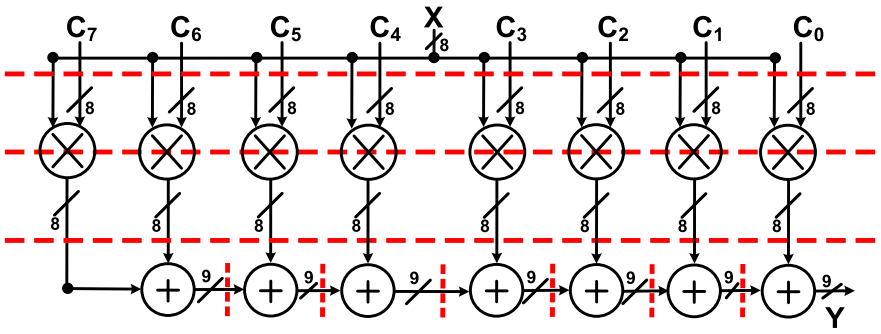


Fig. 7. 8-bit 8-tap FIR filter based on the data-broadcast architecture.

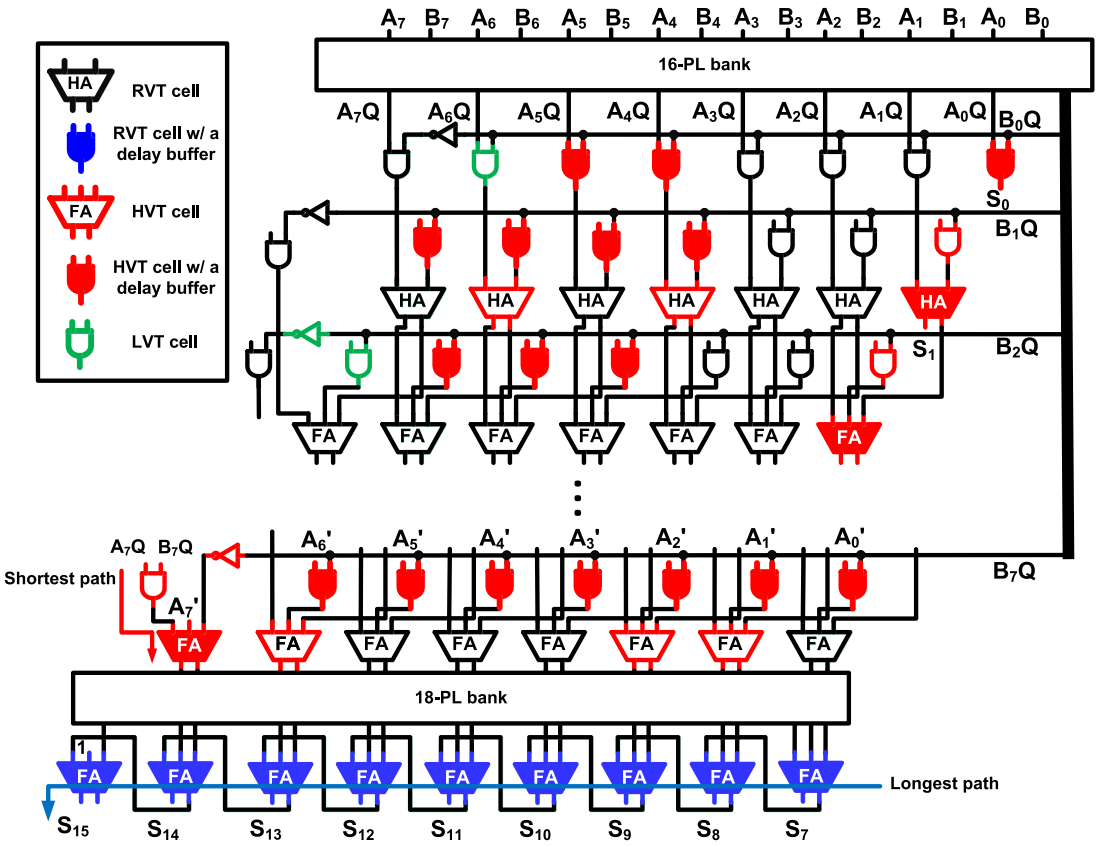


Fig. 8. Two-stage 8-bit multiplier with the proposed multi-V_t short path padding scheme.

D. Wide-Pulsed-Latch-Based FIR Prototype

We designed an 8-bit 8-tap data-broadcast architecture FIR filter based on pulsed-latch pipelines (Fig. 7). The dashed lines show the location of pulsed-latches. The multipliers, based on Baugh–Wooley architecture, were padded with high-, low-, and regular-V_t logic and delay cells (Fig. 8). The critical path delay is not changed. Each multiplier has two pipeline stages, employing total 34 pulsed-latches. We also designed ripple carry adders using the same padding technique. T_C of the FIR core is 40 FO4 delays and the target T_{pw} is set to 1/3 T_C. After the delay padding, the shortest path delay is ∼14 FO4 delays.

We also designed the pulse generator and pulse distribution network. Fig. 9 shows the distribution of the clock network and the schematics of the shared pulse generator. We implemented only one pulse generator for all the latches (∼400) in the FIR core. Also, we designed 1-level pulse distribution tree based on the merged buffer scheme [19], [20] for low skew and low power. The configurable pulse generator can supply a pulsewidth from 5 FO4 delays to 20 FO4 delays for the FIR chip.

Such a simple clock design is feasible, because the wide pulse can alleviate the slew requirement in pulse distribution. We set the high phase time of pulse to no less than 5 FO4 delays for meeting latch setup time and allowing some amount of timing borrowing. If pulses need to be narrow, the slew time should be short enough to ensure sufficient time of pulses being high. For example, for the pulsewidth <7 FO4 delays, we need to reduce the slew time down to ∼1 FO4 delay. This requires a very strong buffer to compensate wire resistance, or requires to embed local pulse generators each of which is shared by only nearby pulsed-

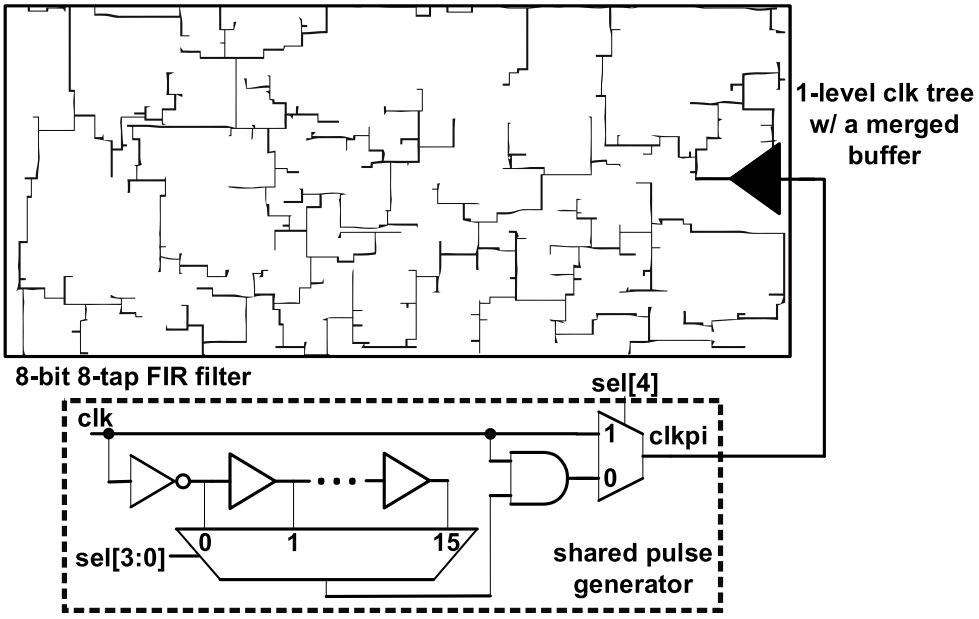


Fig. 9. Simple pulse distribution network using the 1-level clock tree with minimize skew and low power consumption and the shared pulse generator.

TABLE III
MEASUREMENTS SUMMARY ACROSS PVT VARIATIONS

Variation	Functional T_{pw} (FO4)	F_{CLK} Improvement	Energy Saving
Voltage (0.33-0.4V)	6~12	24% to 45%	7.3% to 12%
Process (11 chips)	7~12	17.4% to 56%	9.1% to 24.2%
Temperature (0-80°C)	9~14	16.7% to 51.6%	5.4% to 14%

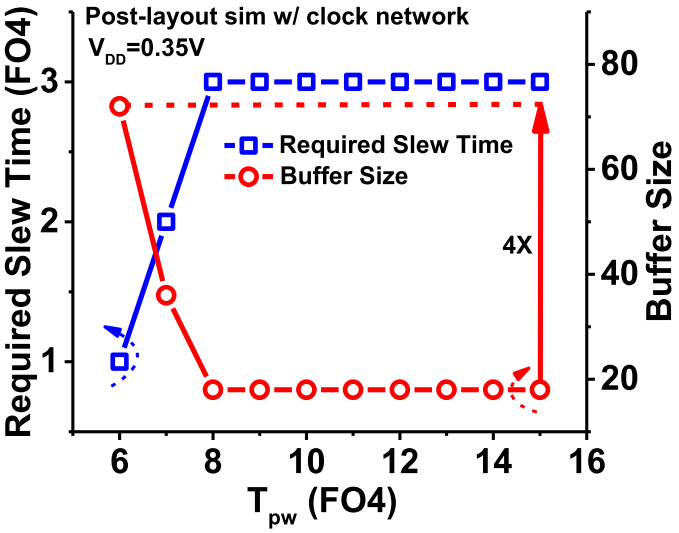


Fig. 10. Relationship between T_{pw}, the required slew time, and the buffer size for pulse distribution.

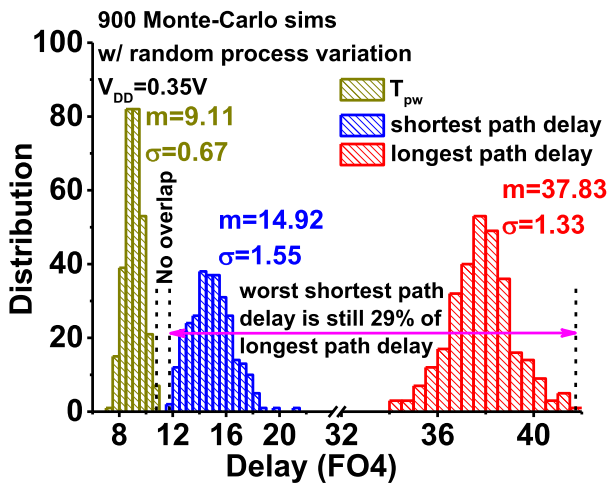
latches. In our design, the wide pulsewidth (~12 FO4 delays) relaxes the slew constraint, e.g., to 3 FO4 delays[2], and simplifies the clock design.

We quantitatively analyze the required buffer size to distribute different widths of pulses at 0.35 V using post-layout clock network. As shown in Fig. 10, it is difficult to reliably distribute pulses whose T_{pw} is <6 FO4 delays. This is because the wire parasitics are too large to achieve the required slew time (0.5 FO4 delay assuming 5 FO4 delays of high-phase duration). We also need 72×, 36×, and 18× buffers to distribute pulse whose T_{pw} is 6, 7, and 8 FO4 delays, respectively. For T_{pw} > 8 FO4 delays, however, we can keep using 18 times buffer strength for the slew constraint (<3 FO4 delays). The use of narrow pulses could require up to four times more buffering in distribution network.

Monte Carlo simulations confirmed that the proposed design is robust across random and systematic process variations. As shown in Fig. 11, the SPICE simulations with the *RC*-extracted netlists of the longest path, the shortest path, and the clock tree confirm that the maximum of T_{pw} is always less than the minimum of the shortest path delay across random process variations. In addition, the minimum of the shortest path delay is still larger than 1/4.5 of the maximum of the longest delay, confirming that the proposed technique can offer a good amount of time borrowing.

In the experiments with both random and systematic process variations effects, as shown in Fig. 12, the ratio of the minimum of the shortest path delay to the maximum of the longest path delay is in the range of 0.22 to 0.35, again confirming our proposed techniques can offer a good amount of time borrowing. This ratio indicates that with the good ability of

[2]Excessive relaxation of slew can degrade T_{dq} and T_{setup} of latches.



Path	Avg. delay (FO4)	Std. delay (FO4)
Critical path	37.83	1.33
Path 1 (2nd longest path)	37.6	2.56
Path 2	36.45	2.63
Path 3	35.87	2.29
Path 4	36.76	2.29
Path 5	36.81	2.56
Path 6	33.03	2.16
Path 7	35.12	2.28
Path 8	35.87	2.29
Path 9	35.95	2.29
Path 10	31.89	1.88

(b)

Fig. 11. Monte Carlo simulation results across random process variations of (a) longest path, shortest path, and clock tree and (b) near-critical paths.

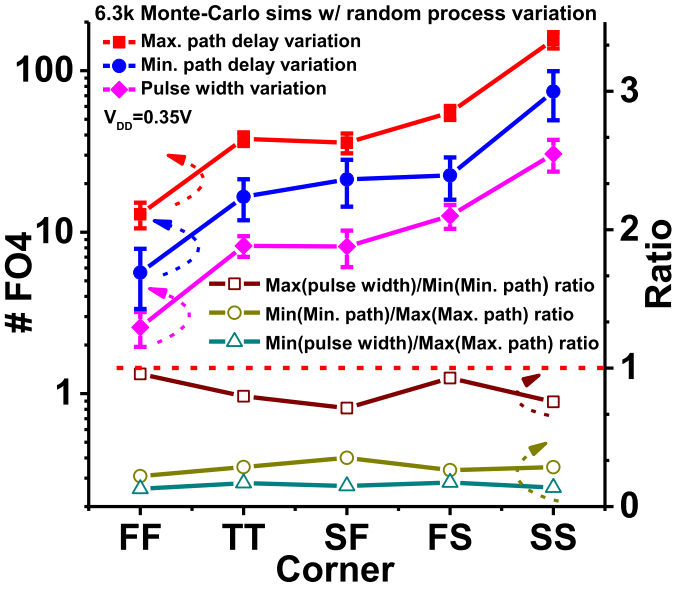


Fig. 12. Monte Carlo simulation results show the robustness (no hold time violations) of our design across five different corners.

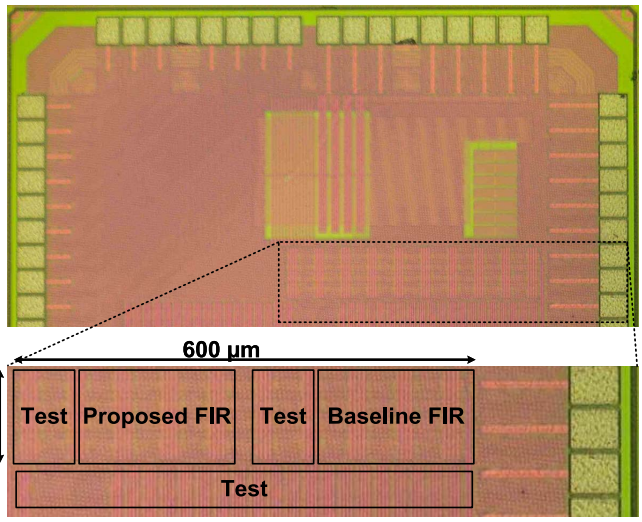


Fig. 13. Test-chip die photograph.

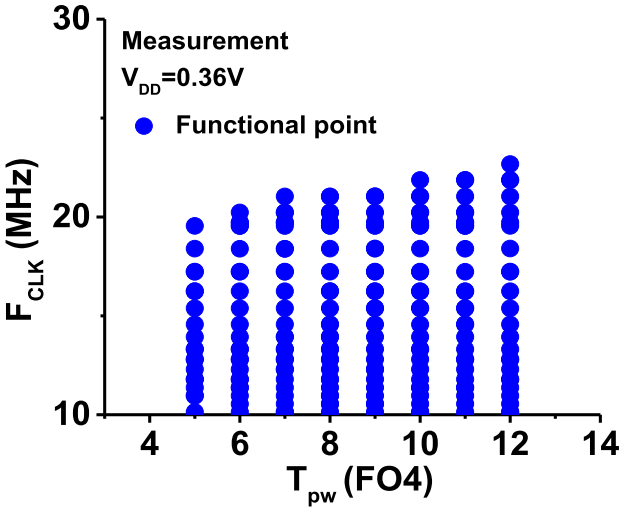


Fig. 14. Functional F_{CLK} with different T_{pw} values.

time borrowing, our proposed techniques allow a 22% to 35% reduction in clock period. Also, the ratio of the maximum of T_{pw} to the minimum of the shortest path delay is in the range of 0.71 to 0.96, which shows the maximum of T_{pw} is always less than the minimum of the shortest path delay across five different corners, and thus there is no hold time violation. These simulations show the robustness of the proposed FIR design with the shortest path delay of 14 FO4 delays, which contains a margin of 1 ∼ 2 FO4 delays.

III. Test-Chip and Measurement Results

A. Test-Chip Organization

Test chips for the FIR cores have been fabricated in a 65-nm general-purpose CMOS process. Fig. 13 shows the die photograph of the test chip. Each chip contains both the proposed and the FF-based baseline designs. The total area of the chip is 0.12 × 0.6 mm^2, the core area of proposed

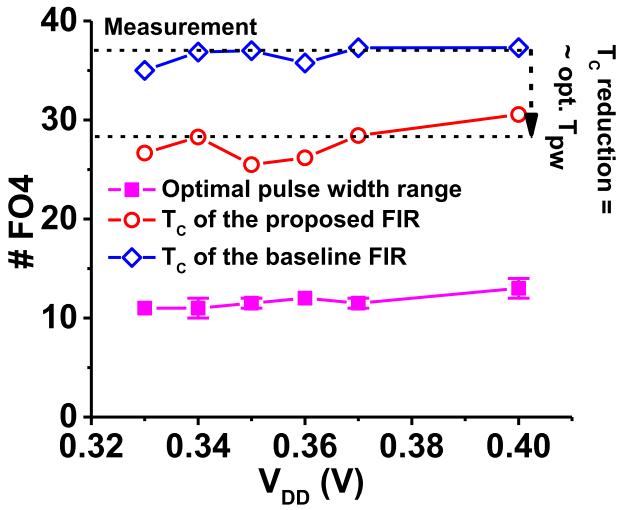


Fig. 15. Optimal T_{pw}, and T_C of the baseline and proposed FIR across V_{DD} values. The T_C improvement of the proposed FIR over the baseline is roughly the same with the optimal T_{pw}.

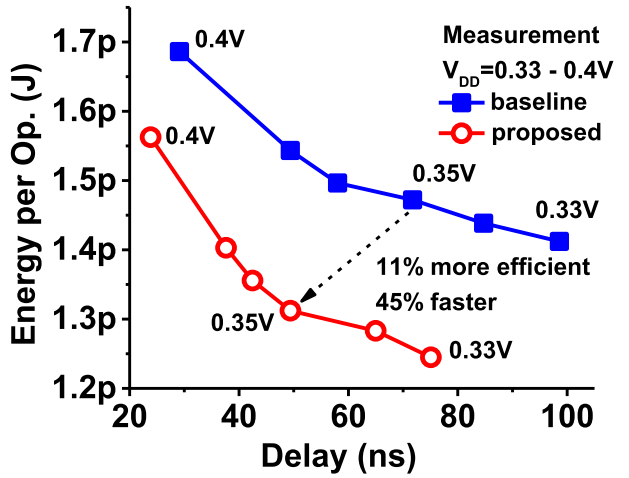


Fig. 16. Delay and energy of the baseline and proposed FIR across V_{DD} values.

FIR is 0.0105 mm^2, and the core area of baseline FIR is 0.0091 mm^2.

B. Performance Measurements

We measure the functionality and the maximum F_{CLK} across T_{pw} values at room temperature. As shown in Fig. 14, the FIR is functional at 5 to 12 FO4 delays of T_{pw}. If T_{pw} is too small, the FIR core fails due to unreliable pulse distribution. If T_{pw} is too large, it also fails because of the hold time violation. The maximum F_{CLK} improves approximately linearly with T_{pw} thanks to the greater amount of time borrowing.

Fig. 15 shows T_C of the baseline and the proposed FIR and T_{pw} at the performance-optimal point. The optimal T_{pw} is found to be between 10 and 14 FO4 delays at V_{DD} = 0.33 to 0.4 V, which is roughly 1/3 of T_C. The T_C reduction from the baseline FIR to the proposed FIR is approximately equal to the optimal T_{pw} (1/3 T_C), which is consistent to (2).

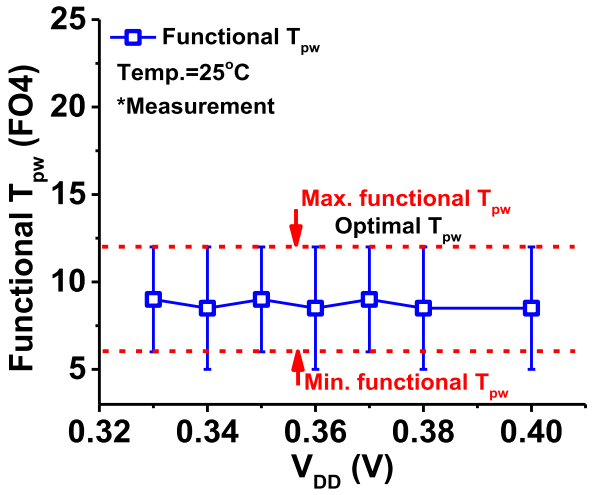


Fig. 17. Functional T_{pw} of the proposed FIR across V_{DD} values.

We also measure the delay and Energy/cycle across V_{DD} values. As shown in Fig. 16, the proposed FIR core achieves 20.2-MHz, 1.31-pJ/cycle, and 65-ns · pJ EDP at T_{pw} = 12 FO4 delays and V_{DD} = 0.35 V. The baseline FIR core achieves 13.9-MHz, 1.47-pJ/cycle, and 105-ns · pJ EDP. F_{CLK} is 45.2% higher than that of the baseline, and Energy/cycle and EDP are 11% and 38% less than those of the baseline core.

C. Measurement Across PVT Variations

As the pulse generator is on the same die with the FIR core, it can track the FIR core across PVT variations. This allows us to keep using the same T_{pw} setting across PVT variations. As shown in Fig. 17, across V_{DD} = 0.33–0.4 V, we find that T_{pw} that makes the proposed FIR core functional is from 6 to 12 FO4 delays. Fig. 18 shows that F_{CLK} improvement at the optimal T_{pw} (12 FO4 delays) is 24%–45% and energy savings are 7.3%–12%.

Across process variations in 11 chips, the functional T_{pw} is measured to be 7 FO4 to 12 FO4 delays (Fig. 19). We also measure the F_{CLK} and Energy/cycle of the proposed cores configured with the single T_{pw} setting and the baseline cores using FFs. As shown in Fig. 20, F_{CLK} improvement is 17.4%–56% and energy savings are 9.1%–24.2%. The mean (m) and standard deviation (σ) of F_{CLK} improvement are 30% and 13%, respectively. The mean and standard deviation of energy savings are 16% and 5%, respectively.

We also measured the functional T_{pw} of a typical chip across temperature variations. As shown in Fig. 21, the functional T_{pw} across 0 °C–80 °C lies from 9 FO4 to 14 FO4 delays. When operating at 0 °C–80 °C, as shown in Fig. 22, the proposed FIR core achieves 16.7% to 51.6% higher F_{CLK} and 5.4% to 14% less Energy/cycle over the baseline.

Table III summarizes the measurement results across PVT variations. Across PVT variations, the core can operate with the largest common T_{pw}: 12 FO4 delays. The maximum F_{CLK} improvement at this T_{pw} is ~50% and the maximum energy savings are 24.2%, as compared to the baseline design.

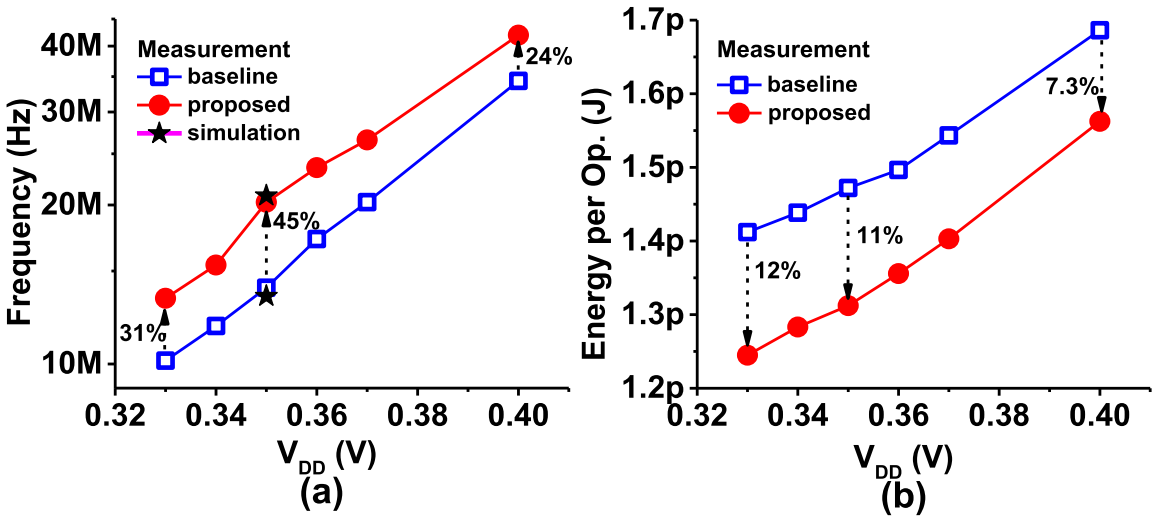


Fig. 18. (a) Frequency improvement and (b) energy savings of the proposed FIR over the baseline across V_{DD} values.

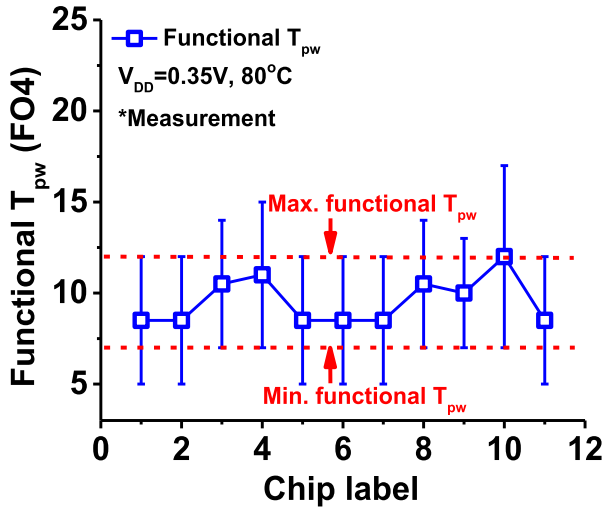


Fig. 19. Functional T_{pw} of the proposed FIR across 11 chips.

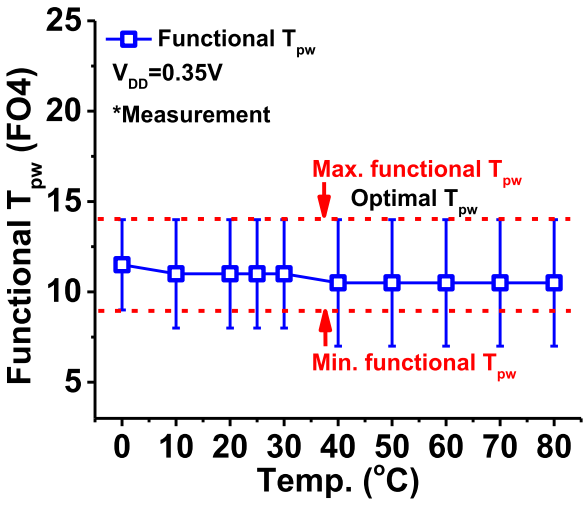


Fig. 21. Functional T_{pw} of the proposed FIR across temperatures.

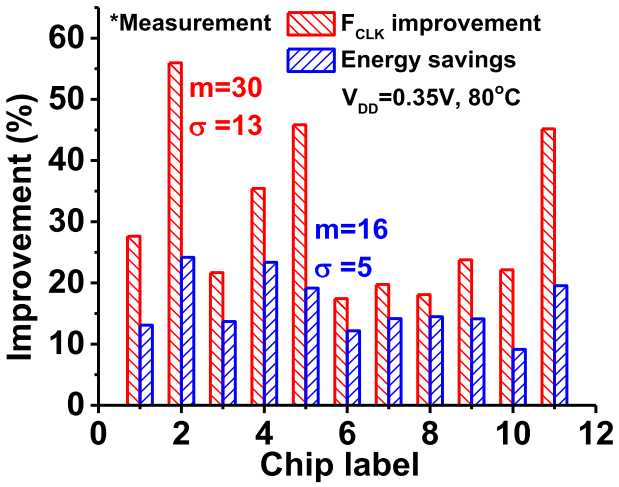


Fig. 20. F_{CLK} and energy efficiency improvement of the proposed FIR over the baseline across 11 chips.

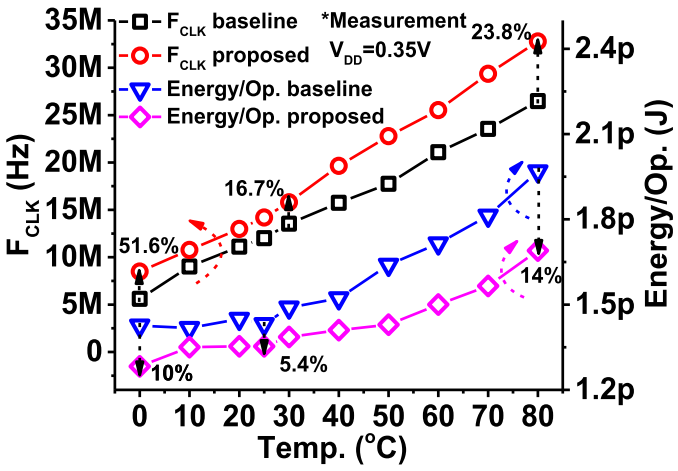


Fig. 22. F_{CLK} and efficiency improvement of the proposed FIR over the baseline across temperatures.

D. Comparisons

The core area and power breakdown of the baseline and proposed FIR are shown in Fig. 23. The core area of the baseline FIR is 9110 μm^2, with a logic area of 6374 μm^2 and FFs area of 2736 μm^2. The core area of the proposed FIR is 10 469 μm^2, with a logic area of 6374 μm^2, latches area of 2016 μm^2, and delay buffers area of 2081 μm^2. The delay buffers take 19.9% of area and 7% of power of the pro-

TABLE IV
MEASUREMENTS SUMMARY AND COMPARISONS

	Proposed	Baseline	Comparison
Technology	**65nm**	**65nm**	**-**
V_{th}	**std., high, low**	**std.**	**-**
V_{DD} (V)	**0.35**	**0.35**	**-**
F_{CLK} (MHz)	**20.2**	**13.9**	**+45.2%**
Energy/Op. (pJ)	**1.31**	**1.47**	**-11%**
EDP (ns·pJ)	**65**	**105**	**-38%**
Area (mm^2)	**0.0105**	**0.0091**	**+15%**

TABLE V
COMPARISONS WITH THE STATE-OF-THE-ART FIR CORES

	[22] VLSI'07	[23] JSSC'10	[24] JSSC'10	[25] TCAS-II'12	[7] JSSC'14	[26] TVLSI'15	This Work
FIR Type	8-tap, 8-bit	14-tap, 8-bit	8-tap, 8-bit	30-tap, 8-bit	16-tap, 8-bit	14-tap, 8-bit	8-tap, 8-bit
Technology	0.13 µm	0.13 µm	90 nm	0.13 µm	65 nm LP	0.18 µm	65 nm G
PVT-Tolerance/ Technique	Yes/ Body bias	No/-	Yes/ Asynchronous	No/-	Yes/ EDAC	No/-	Yes/ Time borrowing
V_{DD} (V)	0.2	0.27	0.29	0.35	1.2	0.31	0.35
F_{CLK} (Hz)	12k	20M	148k	29k	1.008G	100k	20.2M
Power (nW)	114	310,000	743	32	20e6	32.7	26,462
Energy consumption per operation (pJ)	1.19	1.11	0.63	0.0368	1.24	0.0234	0.16375
Energy-FoM* (fJ)	4.638	4.343	5.112	0.143	19.380	0.048	2.559
Throughput (S/s)	12k	20M	148k	29k	1.008G	100k	20.2M
Energy-FoM/Throughput (ns•pJ/S)	386	0.22	35	4.91	0.02	0.48	0.13
Area** (mm^2)	0.385	0.095	N/A	0.0145	0.1855	0.00691	0.0105

* Energy-FoM(fJ) = Power(nW)/F_{CLK}(MHz)/# of taps/input bit length/coefficient bit length·(65nm/Technology)^2
** Area = area·(65nm/Technology)^2

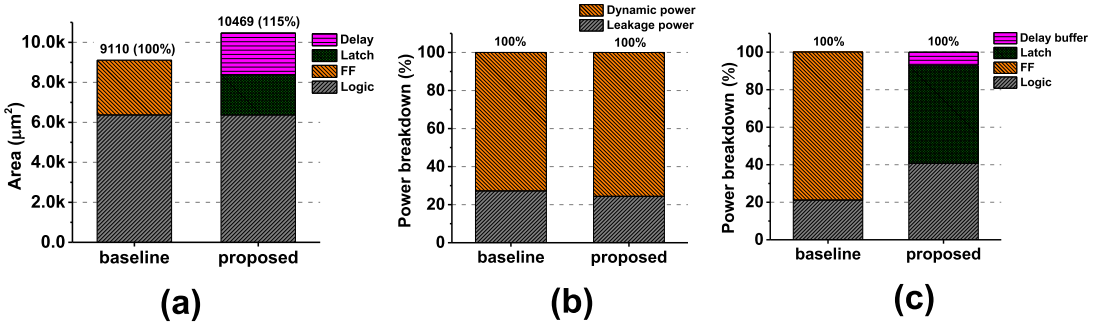


Fig. 23. (a) Core area, (b) leakage and dynamic power, and (c) power breakdown of the baseline and proposed FIR.

posed FIR. The latches in the proposed FIR are 26% smaller than the FFs in the baseline. Combined together, the area overhead is 15%, as compared with the baseline. Table IV summarizes the comparisons of the proposed and the baseline FIR cores. At the design point (V_{DD} = 0.35 V), the F_{CLK}, Energy/op., and EDP improvements are 45.2%, 11%, and 38%, respectively.

As shown in Figs. 2 and 4, higher T_{pw} can enable higher performance, but it can also cause higher energy and area consumption. The F_{CLK} improvement will be limited as we increase T_{pw} because of two reasons. The first is that we cannot gain much beyond a certain point ($T_{pw,limit}$), since the last pipeline stage would not have enough time to compute as it gives too much time to the previous stages. The second reason is that when T_{pw} is greater than half cycle time, the total maximum time can be borrowed is still equal to or less than half cycle time. Hence, when T_{pw} is greater than $T_{pw,limit}$ or half cycle time, we can receive no performance improvement yet only energy penalty. To elaborate on this, we performed the simulation to characterize the power-delay-product (PDP) of our proposed pulsed-latch-based FIR core across different T_{pw} values. We summarize the

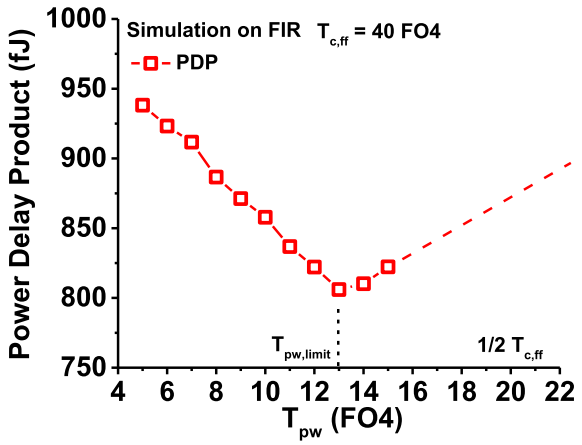


Fig. 24. PDP of the FIR core as a function of $T_{\rm pw}$ becomes optimal at $T_{\rm pw} = \sim 1/3\ T_C$.

simulation results in Fig. 24. It shows the PDP is optimized at $T_{\rm pw} = \sim 1/3\ T_C$.

Table V shows the comparison results of our FIR core and the state-of-the-art FIR cores [7], [22]–[26]. The proposed design achieves among the best energy-figure of merit (FoM) (normalized to the process node) [26] behind only by the works presented in [25] and [26]. The works in [25] and [26] employ no variation tolerant techniques, and achieve about two to three orders of magnitude lower throughput than our proposed design. We also compare the FIR cores in the energy and throughput tradeoff (i.e., Energy-FoM/Throughput), which shows that our design also achieves among the best tradeoff behind only by the work presented in [7]. Note that [7] uses a low-power process technology and also is designed for super-V_t operation.

IV. Conclusion

This paper presents a methodology to design pulsed-latch pipelines in near-/sub-V_t circuits that can use very wide pulses. Such wide pulses can severely increase area overhead due to excessive short path padding, and thus is not considered a common design choice in super-V_t digital pipeline. In this paper, we propose a multi-V_t-based padding technique to scale the overhead, which becomes significantly more effective in near-/sub-V_t pipelines. Experiments with the ISCAS benchmark circuits show that our technique can consistently reduce overhead by ∼two times. The measurement of the FIR prototypes based on the proposed technique demonstrates 45.2% $F_{\rm CLK}$, 11% Energy/cycle, and 38% EDP improvement over the baseline using edge-triggered FFs. The proposed core can also operate at the single $T_{\rm pw}$ setting (12 FO4 delays) robustly across PVT variations. The area overhead is 15%.

References

[1] 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 *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2011, pp. 342–344.

[2] M. Fojtik *et al.*, "Bubble Razor: An architecture-independent approach to timing-error detection and correction," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Apr. 2012, pp. 488–490.

[3] S. Kim and M. Seok, "Analysis and optimization of in-situ error detection techniques in ultra-low-voltage pipeline," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Des.*, Sep. 2014, pp. 291–294.

[4] M. Fojtik *et al.*, "A millimeter-scale energy-autonomous sensor system with stacked battery and solar cells," *IEEE J. Solid-State Circuits*, vol. 48, no. 3, pp. 801–813, Mar. 2013.

[5] Y. Shin and S. Paik, "Pulsed-latch circuits: A new dimension in ASIC design," *IEEE Design Test Comput.*, vol. 28, no. 6, pp. 50–57, Nov./Dec. 2011.

[6] S. Das *et al.*, "RazorII: *In situ* error detection and correction for PVT and SER tolerance," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, Jan. 2009.

[7] P. N. Whatmough, S. Das, and D. M. Bull, "A low-power 1-GHz razor FIR accelerator with time-borrow tracking pipeline and approximate error correction in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 84–94, Jan. 2014.

[8] H. Lee, S. Paik, and Y. Shin, "Pulse width allocation with clock skew scheduling for optimizing pulsed latch-based sequential circuits," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2008, pp. 224–229.

[9] H.-M. Chou, H. Yu, and S.-C. Chang, "Useful-skew clock optimization for multi-power mode designs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2011, pp. 647–650.

[10] S. Paik, L. E. Yu, and Y. Shin, "Statistical time borrowing for pulsed-latch circuit designs," in *Proc. IEEE Asia South Pacific Design Autom. Conf.*, Sep. 2010, pp. 675–680.

[11] C.-L. Chang, I. H.-R. Jiang, Y.-M. Yang, E. Y.-W. Tsai, and A. S.-H. Chen, "Novel pulsed-latch replacement based on time borrowing and spiral clustering," in *Proc. ACM Int. Symp. Phys. Design*, 2012, pp. 121–128.

[12] H.-T. Lin, Y.-L. Chuang, and T.-Y. Ho, "Pulsed-latch-based clock tree migration for dynamic power reduction," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design*, Aug. 2011, pp. 39–44.

[13] Y. L. Chuang, H. T. Lin, T. Y. Ho, Y. W. Chang, and D. Marculescu, "PRICE: Power reduction by placement and clock-network co-synthesis for pulsed-latch designs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2011, pp. 85–90.

[14] Y. L. Chuang, S. Kim, Y. Shin, and Y. W. Chang, "Pulsed-latch aware placement for timing-integrity optimization," in *Proc. ACM/IEEE Design Autom. Conf.*, Dec. 2010, pp. 280–285.

[15] K. Tanimura and N. D. Dutt, "LRCG: Latch-based random clock-gating for preventing power analysis side-channel attacks," in *Proc. IEEE/ACM/IFIP Int. Conf. Hardware/Softw. Codesign Syst. Synth.*, Apr. 2012, pp. 453–462.

[16] S. Lee, S. Paik, and Y. Shin, "Retiming and time borrowing: Optimizing high-performance pulsed-latch-based circuits," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Sep. 2009, pp. 375–380.

[17] M. Seok, D. Jeon, C. Chakrabarti, D. Blaauw, and D. Sylvester, "Pipeline strategy for improving optimal energy efficiency in ultra-low voltage design," in *Proc. ACM/IEEE Design Autom. Conf.*, Aug. 2011, pp. 990–995.

[18] H. Partovi, R. Burd, U. Salim, F. Weber, L. DiGregorio, and D. Draper, "Flow-through latch and edge-triggered flip-flop hybrid elements," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 1996, pp. 138–139.

[19] M. Seok, G. Chen, S. Hanson, M. Wieckowski, D. Blaauw, and D. Sylvester, "CAS-FEST 2010: Mitigating variability in near-threshold computing," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 1, no. 1, pp. 42–49, Mar. 2011.

[20] M. Seok, D. Blaauw, and D. Sylvester, "Clock network design for ultra-low power applications," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design*, Sep. 2010, pp. 271–276.

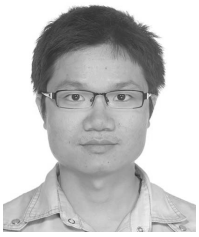
[21] W. Jin, S. Kim, W. He, Z. Mao, and M. Seok, "A 0.35 V 1.3 pJ/cycle 20 MHz 8-bit 8-tap FIR core based on wide-pulsed-latch pipelines," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Sep. 2016, pp. 129–132.

[22] M. E. Hwang, A. Raychowdhury, K. Kim, and K. Roy, "A 85 mV 40 nW process-tolerant subthreshold 8×8 FIR filter in 130 nm technology," in *Proc. IEEE Symp. VLSI Circuits*, Sep. 2007, pp. 154–155.

[23] W.-H. Ma, J. C. Kao, V. S. Sathe, and M. C. Papaefthymiou, "187 MHz subthreshold-supply charge-recovery FIR," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 793–803, Apr. 2010.

[24] I. J. Chang, S. P. Park, and K. Roy, "Exploring asynchronous design techniques for process-tolerant and energy-efficient subthreshold operation," *IEEE J. Solid-State Circuits*, vol. 45, no. 2, pp. 401–410, Feb. 2010.

[25] A. Klinefelter, Y. Zhang, B. Otis, and B. H. Calhoun, "A programmable 34 nW/channel sub-threshold signal band power extractor on a body sensor node SoC," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 937–941, Dec. 2012.
[26] M.-Z. Li *et al.*, "Energy optimized subthreshold VLSI logic family with unbalanced pull-up/down network and inverse narrow-width techniques," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 12, pp. 3119–3123, Dec. 2015.
[27] K. A. Bowman *et al.*, "Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 49–63, Jan. 2009.
[28] K. A. Bowman *et al.*, "A 45 nm resilient microprocessor core for dynamic variation tolerance," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 194–208, Jan. 2011.
[29] M. R. Choudhury *et al.*, "Masking timing errors on speed-paths in logic circuits," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, 2009, pp. 87–92.
[30] S. Kim and M. Seok, "R-processor: 0.4V resilient processor with a voltage-scalable and low-overhead *in-situ* error detection and correction technique in 65 nm CMOS," in *Proc. IEEE Symp. VLSI Circuits*, Sep. 2014, pp. 1–2.
[31] S. Kim and M. Seok, "Variation-tolerant, ultra-low-voltage microprocessor with a low-overhead, within-a-cycle *in-situ* timing-error detection and correction technique," *IEEE J. Solid-State Circuits*, vol. 50, no. 6, pp. 1478–1490, Jun. 2015.
[32] S. Kim, J. P. Cerqueira, and M. Seok, "A 450 mV timing-margin-free waveform sorter based on body swapping error correction," in *Proc. IEEE Symp. VLSI Circuits*, Sep. 2016, pp. 264–265.
[33] E. Consoli, M. Alioto, G. Palumbo, and J. Rabaey, "Conditional push-pull pulsed latches with 726fJ·ps energy-delay product in 65 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf.*, May 2012, pp. 482–484.
[34] E. Consoli, G. Palumbo, J. M. Rabaey, and M. Alioto, "Novel class of energy-efficient very high-speed conditional push-pull pulsed latches," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 7, pp. 1593–1605, Jul. 2014.
[35] E. J. Fluhr *et al.*, "The 12-core POWER8 processor with 7.6 Tb/s IO bandwidth, integrated voltage regulation, and resonant clocking," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 10–23, Jan. 2015.
[36] W. M. Elsharkasy, A. Khajeh, A. M. Eltawil, and F. J. Kurdahi, "Reliability enhancement of low-power sequential circuits using reconfigurable pulsed latches," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 7, pp. 1803–1814, Jul. 2017.
[37] B. D. Yang, "Low-power and area-efficient shift register using pulsed latches," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 6, pp. 1564–1571, Jun. 2015.



Wei Jin (S'16) received the B.S. and M.S. degrees in microelectronics, and circuits and systems from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2011, respectively, where he is currently pursuing the Ph.D. degree in circuits and systems.

He was a visiting Ph.D. Student in electrical engineering at Columbia University, New York City, NY, USA, from 2014 to 2016. His current research interests include variation-, voltage-adaptive circuits design, ultra-low power VLSI circuits, and system design.

Seongjong Kim received the B.S. degree in electrical engineering from Hanyang University, Seoul, South Korea, in 2010, and the M.S. degree from the University of Michigan, Ann Arbor, MI, USA, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering at Columbia University, New York City, NY, USA.

His current research interests include low-power and variation tolerant VLSI circuit and system design. He is a Student Fellow of the Catalyst Foundation, Valhalla, NY, USA.

Weifeng He received the B.S., M.S., and Ph.D. degrees in microelectronics and solid state electronics from the Harbin Institute of Technology, Harbin, China, in 1999, 2001, and 2005, respectively.

He is currently an Associate Professor with the Department of Micro–Nano Electronics, Shanghai Jiao Tong University, Shanghai, China. His current research interests include VLSI architecture for video signal processing, reconfigurable processor architecture, and low power circuit design.

Zhigang Mao (M'08) received the B.S. degree from Tsinghua University, China in 1986, and the Ph.D. degree from the University of Rennes 1, Rennes, France, in 1992.

From 1992 to 2006, he was with the Microelectronics Center, Harbin Institute of Technology, Harbin, China. In 2006, he joined the Department of Micro–Nano Electronics, Shanghai Jiao Tong University, Shanghai, China, where he is currently a Professor. His current research interests include DSP architecture design, video processor design, and reconfigurable processor architecture.

Mingoo Seok (S'05–M'11) received the B.S. (*summa cum laude*) degree from Seoul National University, Seoul, South Korea, in 2005, and the M.S. and Ph.D. degrees from the University of Michigan, Ann Arbor, MI, USA, in 2007 and 2011, respectively, all in electrical engineering.

He was a member of Technical Staff with Texas Instruments Inc., Dallas, TX, USA, in 2011. Since 2012, he has been an Assistant Professor with the Department of Electrical Engineering, Columbia University, New York City, NY, USA. His current research interests include variation, voltage, aging, thermal-adaptive circuits and architecture, ultra-low-power system on chip design for emerging embedded systems, machine-learning VLSI architecture and circuits, and non-conventional hardware design.

Dr. Seok received the 1999 Distinguished Undergraduate Scholarship and the 2005 Doctoral Fellowship from the Korea Foundation for Advanced Studies, and the 2008 Rackham Pre-Doctoral Fellowship from the University of Michigan. He also received the 2009 AMD/CICC Scholarship Award for picowatt voltage reference work, and the 2009 DAC/ISSCC Design Contest for the 35pW sensor platform design. He was a recipient of the 2015 NSF CAREER Award. He has been serving as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I since 2013 and the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS since 2015.