Advancing Nonvolatile Computing With Nonvolatile NCFET Latches and Flip-Flops

Xueqing Li, Member, IEEE, Sumitha George, Kaisheng Ma, Wei-Yu Tsai, Ahmedullah Aziz, Student Member, IEEE, John Sampson, Member, IEEE, Sumeet Kumar Gupta, Member, IEEE, Meng-Fan Chang, Senior Member, IEEE, Yongpan Liu, Senior Member, IEEE, Suman Datta, Fellow, IEEE, and Vijaykrishnan Narayanan, Fellow, IEEE

Abstract—Nonvolatile computing has been proven to be effective in dealing with power supply outages for on-chip check-pointing in emerging energy-harvesting Internet-of-Things applications. It also plays an important role in power-gating to cut off leakage power for higher energy efficiency. However, existing on-chip state backup solutions for D flip–flop (DFF) have a bottleneck of significant energy and/or latency penalties which limit the overall energy efficiency and computing progress. Meanwhile, these solutions rely on external control that limits compatibility and increases system complexity. This paper proposes an approach to fundamentally advancing the nonvolatile computing paradigm by intrinsically nonvolatile area-efficient latches and flip–flops designs using negative capacitance FET. These desig

1. INTRODUCTION

SCHEDULED power-gating of VLSI computing systems has been widely adopted in both low-power portable devices and high-performance cloud server centers to cut off static leakage power. States of the registers and flip-flops in the pipelining logic need to be backed up to prevent loss of computation status if the supply is removed. Similarly, mandatory state backup and restore operations are required for battery-less portable devices powered by energy-harvesting techniques. This is because the ambient energy sources, such as vibration, photovoltaics, and radio, are essentially intermittent even with sophisticated design methods [1]–[8].

A. Impact of Power Supply Outages

Fig. 1(a) shows a sample of power income versus time. The power outage failures have a strong impact on the computation progress and the overall energy efficiency of IoT computing [1], [2], [4]–[8]. On the one hand, the computing system will lose its computation states if these states are not successfully backed up before the power failure. On the other hand, even if techniques of check-pointing are applied to reduce the chance of progress loss, significant overheads of energy and latency still exist for each backup and restore operation, as illustrated in Fig. 1(b). If the backup data are stored into out-of-chip nonvolatile memory, huge energy and latency are noticeable, mainly caused by the long-distance data transmission and constrained parallelism.

It is also noted that, although a larger energy storage capacitor buffer can be used to smooth the power trace, it cannot reduce the amount of energy for each backup or restore operation. A larger capacitor also results in larger leakage current and more accumulating time to respond till it reaches an appropriate voltage level for supply regulation.

B. Opportunities and Challenges for NVP

The recent success of embedding nonvolatile memory (NVM) into the same chip has incubated the concept of nonvolatile processing (NVP) with nonvolatile D flip-flop (NV-DFF) which backs up the computation states of each D flip-flop (DFF) into a local on-chip NVM [4]–[11]. Table I summarizes some existing NV-DFF designs based on ferroelectric capacitor [10]–[14], MTJ [15]–[17], ReRAM [18] etc., with their pros and cons. With a shorter data transmission distance and a wider bus than the out-of-chip backup, the energy and latency overhead could be reduced, as illustrated in Fig. 1. However, existing NVP solutions are still facing many challenges.
The first challenge is how to handle the consequence of the energy, delay and possibly high peak power of a backup or restore operation. When a clustered NVM array is used for backup, across-the-chip interconnections and the series memory access pattern still result in a long backup/restore latency, and high energy consumption. Instead, if distributed NVM cells and interface circuitry are placed close to each DFF to build a nonvolatile DFF (NV-DFF) for local parallel backup, as illustrated in Fig. 1(c), the backup/restore operation will be much faster but still consume high energy because of the duplicated interface circuit between the DFF and NVM cells. The high current one NV-DFF backup/restore operation also limits the plausible parallelism.

The second challenge is how to provide backup and restore control for distributed NV-DFFs, and how to design the backup policy, especially on when to backup [1]. To provide this control results in additional wiring with more area and energy consumption. What’s worse, the processor architecture and the software need to adapt to the control, which limits the compatibility of existing software and makes the operating system or processor design complicated. Furthermore, a backup triggered too early or too late will waste energy that could otherwise be saved, or result in a backup failure and rolling back in progress. Although architectures and even task scheduling could be optimized [1], [7], this challenge still exists because of the inherent nature of intermittency and unpredictability in ambient power sources.

### C. Proposed Solution of NCFET Nonvolatile Computing

This paper proposes an NVP solution through circuit innovations with negative capacitance FET (NCFET). NCFET is a promising emerging beyond-CMOS transistor [9], [23]–[29]. With the capability of steep switching, tunable hysteresis, and good scalability, NCFET has attracted interests in both logic and memory applications [32]–[34]. The innovation in this paper originates from the fact that an NCFET device could be treated as an NVM cell while serving as a logic transistor. The proposed nonvolatile NCFET latches and DFFs enable a new paradigm of low-power intrinsically nonvolatile operations:

- The proposed nonvolatile latches and flip-flops have only a few extra transistors added to the conventional CMOS designs, leading to a compact cell design. Unlike most existing NV-DFF designs based on the Ferroelectric capacitor in [4] and [11] and the Spin Orbit Torque in [14], no additional circuitry or a different supply voltage is needed for sensing or driving functions.

The first challenge is how to handle the consequence of the energy, delay and possibly high peak power of a backup or restore operation. When a clustered NVM array is used for backup, across-the-chip interconnections and the series memory access pattern still result in a long backup/restore latency, and high energy consumption. Instead, if distributed NVM cells and interface circuitry are placed close to each DFF to build a nonvolatile DFF (NV-DFF) for local parallel backup, as illustrated in Fig. 1(c), the backup/restore operation will be much faster but still consume high energy because of the duplicated interface circuit between the DFF and NVM cells. The high current one NV-DFF backup/restore operation also limits the plausible parallelism.

The second challenge is how to provide backup and restore control for distributed NV-DFFs, and how to design the backup policy, especially on when to backup [1]. To provide this control results in additional wiring with more area and energy consumption. What’s worse, the processor architecture and the software need to adapt to the control, which limits the compatibility of existing software and makes the operating system or processor design complicated. Furthermore, a backup triggered too early or too late will waste energy that could otherwise be saved, or result in a backup failure and rolling back in progress. Although architectures and even task scheduling could be optimized [1], [7], this challenge still exists because of the inherent nature of intermittency and unpredictability in ambient power sources.

### TABLE I

**PERFORMANCE COMPARISONS OF RECENTLY REPORTED NV-DFF DESIGNS**

<table>
<thead>
<tr>
<th>Year</th>
<th>[10] (measured)</th>
<th>[11] (measured)</th>
<th>[19] (simulated)</th>
<th>[18] (simulated)</th>
<th>[20] (simulated)</th>
<th>This Work (Simulated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature size</td>
<td>130nm</td>
<td>130nm</td>
<td>45nm</td>
<td>180nm</td>
<td>10nm</td>
<td>10nm</td>
</tr>
<tr>
<td>Nonvolatile material</td>
<td>PZT Capacitor</td>
<td>PZT Capacitor</td>
<td>MTJ Al/TiO$_2$/Al ReRAM</td>
<td>NCFET with HfO$_2$ and PZT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retention Time</td>
<td>10 hours to 10 years; varying by design</td>
<td>-10 years</td>
<td>-10 years</td>
<td>Same as PZT capacitor in theory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endurance</td>
<td>Varying by material; possibly &gt;$10^{12}$ [21]</td>
<td>&gt;$10^{12}$ [22]</td>
<td>10$^7$-10$^9$ [22];</td>
<td>Same as PZT capacitor in theory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voltage</td>
<td>1.5V</td>
<td>1.5V</td>
<td>1.1V</td>
<td>1.8V</td>
<td>0.4V-0.8V</td>
<td>0.4V-1.0V</td>
</tr>
<tr>
<td>Area overhead</td>
<td>64%</td>
<td>49%</td>
<td>-2%</td>
<td>-</td>
<td>35%</td>
<td>25% more transistors</td>
</tr>
<tr>
<td>Backup time</td>
<td>1.64µs</td>
<td>2.22µs</td>
<td>909ps</td>
<td>10ns @2.4V</td>
<td>1.4ns @0.5V</td>
<td>1.0ns @0.8V</td>
</tr>
<tr>
<td>Restore time</td>
<td>1.25µs</td>
<td>2.2µs</td>
<td>177ps</td>
<td>1.3µs @0.4V</td>
<td>75ps @0.5V</td>
<td>56ps @0.8V</td>
</tr>
<tr>
<td>Backup energy</td>
<td>2.4pJ</td>
<td>3.44pJ in total</td>
<td>82.2Ω</td>
<td>735Ω</td>
<td>7.0Ω @0.5V</td>
<td>1.3Ω @0.8V</td>
</tr>
<tr>
<td>Restore energy</td>
<td>2.34pJ</td>
<td>-</td>
<td>735Ω</td>
<td>9.0Ω @0.5V</td>
<td>1.1Ω @0.8V</td>
<td></td>
</tr>
<tr>
<td>Additional Control</td>
<td>Needed</td>
<td>Needed</td>
<td>Needed</td>
<td>Needed</td>
<td>Needed</td>
<td>Not Needed</td>
</tr>
</tbody>
</table>

The first challenge is how to handle the consequence of the energy, delay and possibly high peak power of a backup or restore operation. When a clustered NVM array is used for backup, across-the-chip interconnections and the series memory access pattern still result in a long backup/restore latency, and high energy consumption. Instead, if distributed NVM cells and interface circuitry are placed close to each DFF to build a nonvolatile DFF (NV-DFF) for local parallel backup, as illustrated in Fig. 1(c), the backup/restore operation will be much faster but still consume high energy because of the duplicated interface circuit between the DFF and NVM cells. The high current one NV-DFF backup/restore operation also limits the plausible parallelism.

The second challenge is how to provide backup and restore control for distributed NV-DFFs, and how to design the backup policy, especially on when to backup [1]. To provide this control results in additional wiring with more area and energy consumption. What’s worse, the processor architecture and the software need to adapt to the control, which limits the compatibility of existing software and makes the operating system or processor design complicated. Furthermore, a backup triggered too early or too late will waste energy that could otherwise be saved, or result in a backup failure and rolling back in progress. Although architectures and even task scheduling could be optimized [1], [7], this challenge still exists because of the inherent nature of intermittency and unpredictability in ambient power sources.

C. Proposed Solution of NCFET Nonvolatile Computing

This paper proposes an NVP solution through circuit innovations with negative capacitance FET (NCFET). NCFET is a promising emerging beyond-CMOS transistor [9], [23]–[29]. With the capability of steep switching, tunable hysteresis, and good scalability, NCFET has attracted interests in both logic and memory applications [32]–[34]. The innovation in this paper originates from the fact that an NCFET device could be treated as an NVM cell while serving as a logic transistor. The proposed nonvolatile NCFET latches and DFFs enable a new paradigm of low-power intrinsically nonvolatile operations:

- The proposed nonvolatile latches and flip-flops have only a few extra transistors added to the conventional CMOS designs, leading to a compact cell design. Unlike most existing NV-DFF designs based on the Ferroelectric capacitor in [4] and [11] and the Spin Orbit Torque in [14], no additional circuitry or a different supply voltage is needed for sensing or driving functions.
These designs are free from external backup and restore controls because all backup and restore operations are carried out autonomously, improving compatibility and complexity of existing logic designs with drop-in replacement of latches and DFFs, as shown in Fig. 1(d). Getting rid of the need for external controls could significantly reduce the control overhead for check-pointing and power-gating.

These designs are also fast and energy-efficient, with similar delay and energy consumption to conventional volatile designs under a stable supply. Typical energy and latency for a backup plus restore operation are only at levels of fJ and ns, respectively. Note that the restore time also depends on the supply voltage recovery time. The operations could be even faster with kinetic coefficient improvement in the NCFET device. With such a low energy and delay overhead, the utilization of harvested energy for the purpose of computing will be significantly improved. For the same reason, check-pointing and power-gating could be carried out in fine grain as needed with significantly reduced energy and delay penalties.

In the rest of this paper, Section II introduces the NCFET background, Section III and IV describe the proposed NCFET nonvolatile latches and flip-flops, along with simulation results and performance comparisons. Section V discusses related applications and Section VI concludes this paper.

II. NCFET DEVICE BACKGROUND

This session introduces the NCFET device basics, including device structures, operating mechanisms, recent advance in the device fabrication, and finally device modeling.

A. Device Structure and Operating Theory

NCFET, also known as ferroelectric FET (FeFET), is essentially a metal-oxide-semiconductor FET with a negative capacitance material layer placed at the gate [23]. A few NCFET structures have been reported, including a MOSFET with an externally connected negative capacitor [24], [25] and a MOSFET with an integrated negative capacitor at the gate [26]. Fig. 2(a-c) illustrates some design structures.

Many ferroelectric materials, such as PbTiO, BaTiO, Pb(ZrTi)O, HiZrO, etc., have exhibited negative capacitance [35]. Strictly speaking, negative capacitance should be “negative differential capacitance”, whose charge decreases as the applied voltage increases within a certain voltage range. Fig. 3(a-b) illustrates a simplified concept for NCFET with the negative capacitor connected to the internal MOSFET gate capacitor. One interesting phenomenon that fits well with Boolean logic computation is the boosted voltage change ∆V_{MOSFET} at the internal MOSFET gate induced by an external NCFET gate voltage change ∆V_{NCFET} with |C_{FE}| > C_{MOS}, in which C_{FE} and C_{MOS} represent the ferroelectric negative capacitance and the MOSFET gate capacitance, respectively. The voltage gain at the internal MOSFET gate over the overall applied NCFET gate voltage could be approximated by G = ∆V_{MOSFET}/∆V_{NCFET} = C_{FE}/(C_{FE} + C_{MOS}) > 1. This indicates a steep switching that enables a lower supply to achieve the same ON-OFF ratio, as shown in Fig. 3(c).

More interestingly, recent reports have also theoretically predicted and experimentally confirmed that, by tuning C_{FE}, so that |C_{FE}| < C_{MOS} within some voltage range, a hysteresis loop in the I_{DS}-V_{GS} appears and even expands to the negative V_{GS} range [23]–[25], as shown in Fig. 4(a).

The mechanism behind this is the polarization P versus electric-field E hysteresis of the ferroelectric capacitor shown in Fig. 4(b). The charge Q_{FE} of the ferroelectric capacitor is a function of E and P in the way of Q_{FE} = e_0 E + P, where e_0 is the permittivity of the vacuum, and P is typically ≫ e E. The switching of P is electrically determined by E and a switching occurs when the applied voltage exceeds the coercive voltage. Before the polarization switching occurs, C_{FE} is positive. As V_{NCFET} increases from 0V to close to

---

**Fig. 2.** NCFET device structures: externally-connected ferroelectric capacitor in (a), integrated ferroelectric layer on the planar MOSFET gate in (b) and around the gate in a FinFET structure in (c), and atop a fin structure in (d).

**Fig. 3.** NCFET: (a) N-type NCFET device concept; (b) Capacitance network of an NCFET; (c) I-V curves: steep-switching and hysteresis with the baseline MOSFET curve in black.

**Fig. 4.** NCFET I-V hysteresis in (a) and P-E loop of ferroelectric capacitor in (b). The curves in (a) are for the fin-structure NCFET in Fig. 2(c).
TABLE II

<table>
<thead>
<tr>
<th>Ferroelectric material</th>
<th>Structure</th>
<th>SS (Forward) (mV/dec)</th>
<th>SS (Reverse) (mV/dec)</th>
<th>Hysteresis</th>
</tr>
</thead>
<tbody>
<tr>
<td>HfZrO</td>
<td>Planar</td>
<td>50</td>
<td>45-52 w/o and 2-4 w/ hysteresis</td>
<td>Negligible</td>
</tr>
<tr>
<td>P(VDF-TrFE)</td>
<td>Planar</td>
<td>8.5-11</td>
<td>16.50</td>
<td>Depends</td>
</tr>
<tr>
<td>BiFeO</td>
<td>Fin</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>HfZrO</td>
<td>Fin</td>
<td></td>
<td></td>
<td>Depends</td>
</tr>
</tbody>
</table>

the $I_D - V_G$ rising edge, the voltage across the ferroelectric material $V_{FE}$ increases from the initial value of 0 to close to the coercive voltage. As $V_{NCFET}$ further increases, $V_{FE}$ steps beyond the coercive voltage, and $C_{FE}$ becomes negative, leading to the unstable negative total series capacitance of $C_{FE}$ and $C_{MOS}$ that triggers a very steep switching of $I_D$ until $C_{FE}$ settles to be positive again. The outbound trip is similar. In both outbound and inbound curves away from the switching edges, a relatively flat $I_D - V_G$ may be observed if $C_{FE}$ (being positive in this region) is much smaller than $C_{MOS}$, because $\Delta V_{MOS} = \Delta V_{NCFET} \times C_{FE}/(C_{FE} + C_{MOS}) \approx 0$.

Such a hysteresis feature covering $V_G = 0$V indicates a stable remnant polarization when power is removed [35]. Actually, a zero-polarization of the ferroelectric material is not stable as the free energy is not the lowest [35]. It is also noted that the hysteresis width is a function of the dynamics of both $C_{FE}$ and $C_{MOS}$, and could be typically tuned by varying the ferroelectric layer thickness, as observed in Fig. 3(c) and Fig. 4 [23], [24].

B. Recent NCFET Fabrication and Modeling Progress

The recent NCFET device as a beyond-CMOS steep-slope device apparently got an impetus for low-voltage Boolean operations, like Tunnel FETs [30], [31]. To make NCFET compatible with conventional CMOS logic, some works try to prevent the $I-V$ hysteresis [25]-[28], while others make use of it to build a new device different from CMOS [24]. Table II summarizes some recent NCFET devices showing the advantage of subthreshold swing (SS) that is substantially lower than the fundamental limit of 60mV/decade in CMOS in the room temperature. While a demonstration of integrated circuits is not yet available, these reports have shown the scalability of NCFETs below 100 nm.

NCFET modeling needs to consider the dynamic mechanism as discussed above. There have been some works on time-domain-based modeling supporting SPICE transient simulations [36], [37], although device variations, temperature dependence, etc. are not yet supported. This paper adopts the ferroelectric capacitor model in [20], [32], and [36] that has been experimentally calibrated, and then attach it to a PTM CMOS FinFET to build NCFET models. The ferroelectric layer thickness is set to 8 nm for proper hysteresis as shown in Fig. 4. The optimization of the ferroelectric layer thickness, will be further discussed in Section V. The switching time has been calibrated based on experimental results by setting the kinetic coefficient $\rho$ of the ferroelectric material to be 0.25 [36].

III. PROPOSED NCFET NONVOLATILE LATCHES

This section presents the proposed two categories of NCFET nonvolatile latch designs in Fig. 5: NV-Latch-N in Fig. 5(a) with a differential-driving input pair D/DN, and NV-Latch-NC in Fig. 5(c) with only D as the data input. In Fig. 5(c), the two bottom NMOS transistors have a lower $V_{TH}$ for higher yield (see Section V). Their baseline designs are shown in Fig. 5(b) and Fig. 5(d), respectively. As to be revealed later in this section, they behave differently in area, energy per switching, delay, etc. When the power supply is steady, their logical transfer function is the same as shown in Fig. 5(e).

A. NCFET NV-Latch-N: Functionality and Operation Theory

NV-Latch-N has two extra N-type NCFETs over the CMOS counterpart design. Both latches are designed with sufficient input driving strength so as to overwrite the state by the input pair D/DN directly, when the clock signal $CLK$ is high. When $CLK$ is high, it connects the input differential signals $D$ and $DN$ to the output ports of $Q$ and $QN$, respectively.

Fig. 6 illustrates a steady state of NV-Latch-N with $CLK$ being high. Let us assume $Q$ is logically ‘1’ (representing
VDD in voltage) and QN is logically ‘0’ (representing GND in voltage). Controlled by Q or QN, M3 and M8 are turned off, and M4 and M7 are turned on. For NCFET M5, its gate voltage is high, and the drain connecting to QN is low, its source is discharged to GND by M7. As a result, M5 is turned on and stays at the positive polarization state with low drain-source resistance. For NCFET M6, it stays OFF at the negative polarization state with high drain-source resistance. It is noted that, a low voltage of GND at QN will turn off M8, and set M6 as negatively polarized (high drain-source resistance) state, M6 source will be charged to Q, which is VDD, by M6. As a result, the drain-source voltage of M6 becomes ~VDD. After that, M6 will remain negative in its polarization even if its source node is slowly discharged by M8.

Fig. 7 shows the NV-Latch-N transient waveforms of the sample and hold operations with a steady supply. The last P plot indicates “polarization” of the two NCFETs in the NV-Latch-N scheme (Blue: M5 polarization; Purple: M6 polarization).

of D and CLK, the duration of the total time of Q to stay at the new state, i.e. ‘1’ from 60ns to 67ns, is too short for the polarization switching to finish. As a result, none of them in this 7ns period of time succeeds in switching the polarization state, until after 67ns when Q stays at ‘1’ for more than 0.5nS.

Such a phenomenon of a polarization switching speed being lower than the input data-rate actually reveals the possibility of having a pseudo-floating ‘0’ of Q at the latch output, with high resistance to VDD (with turned-off PMOS) and also medium-to-high resistance to GND (due to partially turned-off NCFET in series with a turned-on NMOS). Note that a pseudo-floating ‘1’ does not exist because a Q output of ‘1’ is always securely connected to VDD by the PMOS.

The pseudo-floating ‘0’ has three major effects. First, functionality will stay correct with proper noise shielding. This is because a short period of ns in time of being pseudo-floating will not have the state corrupted if external coupling noise is properly isolated. For example, most standard embedded DRAM cells have a retention time of a few microseconds [39] at the internal floating MOSFET gate. After the polarization switching finishes in the order of ns or sub-ns, this pseudo-floating ‘0’ becomes a steady ‘0’ connecting to GND.

Second, endurance will improve. Although existing research has not yet found out a fundamental bottleneck in improving the endurance or aging effects of ferroelectric materials [38], existing ferroelectric materials degrade faster in terms of number of full-swing switching cycles than CMOS transistors. The way above of slower polarization switching helps to reduce the number of full-swing switching activities, which improves the endurance of NCFETs.

Third, delay and energy performance may also improve. First, considering the conventional volatile CMOS latch in Fig. 5(b), in order to overwrite the previous ‘0’ state of Q to be ‘1’, the new input datum of ‘1’ has to “fight” with the pulling down transistor in the latch. Such a racing condition causes significant power consumption until the switching completes. On the contrary, in NV-Latch-N, considering the pseudo-floating ‘0’ Q due to the higher resistance from Q to GND through the not-fully-ON NCFET, overwriting this ‘0’ to be ‘1’ has less short current or racing conditions, which improves the speed and energy performance. It is also noted that, using a high threshold NMOS transistor could also
Fig. 9. Transient waveforms of backup and restore operations in NV-Latch-N.

Fig. 10. Proposed NV-Latch-P in (a) and NV-Latch-NP in (b).

mitigate the racing “fight” but could not provide equally low pull-down resistance after the polarization switching finishes.

Fig. 9 shows the transient waveforms with backup and restore operations due to power failures. At around 180nS and 2,175ns, backup of ‘0’ and ‘1’ is carried out, respectively; at around 2,100ns and 4,080ns, restore of ‘0’ and ‘1’ is observed, respectively. The simulation is accompanied with close to a 2us power-down window to mimic really field scenarios.

The idea of using the N-type NCFET to build NV-Latch-N could be extended straightforward to two other nonvolatile latch designs, namely NV-Latch-P with P-type NCFETs and NV-Latch-NP with both N-type and P-type NCFETs. Their circuit topologies are shown in Fig. 10. The operation theory is similar to that of NV-Latch-N.

In terms of the physical layout design, the overhead of the extra NCFETs could be mitigated by properly sharing the drain and source active region without the need for adding additional contacts for the drain and source terminals of the NCFETs.

B. NV-Latch-NC: Operation Theory and Analysis

The second type of external-control-free nonvolatile NCFET latch, namely NV-Latch-NC is shown in Fig. 5(c). Compared with the baseline design of 10-transistor CMOS volatile latch in Fig. 5(d), the proposed NV-Latch-NC has 6 more transistors, including the two NCFETs, two extra NMOS transistors in parallel with the NCFETs, and a CMOS switch that is set to be always turned-on by GND and VDD. Both NV-Latch-NC and the CMOS latch in Fig. 5(d) could cut off the inverter latching loop during the sampling phase.

Fig. 11 explains the operation theory. During the sampling phase (high CLK), M3 and M5b form an inverter that provides a fast settling output of QN and no pseudo-floating status. The always-on CMOS switch consisting of M9 and M10 delivers QN to drive the second inverter to provide Q. The feedback loop of the switch is turned off to prevent racing short current during this sampling phase, and will be turned on to form a stable closed loop during the hold phase (low CLK), as shown in Fig. 11(b). The restore operation is similar to NV-Latch-N, by sensing a different Q or QN to ground resistance and continuing to settle down to VDD or GND in the feedback loop.

Fig. 12 shows the transient simulation waveforms of the proposed NV-Latch-NC. Along with the normal sampling and
Fig. 13. Energy versus delay comparison between CMOS and NCFET latches: (a) x1 fan-out driving strength; (b) x4 fan-out driving strength (only CMOS transistors are increased to x4). VDD ranges from 1.0V to 0.5V from left to right.

The proposed NCFET NV-Latch-NC has about 64% and 39% extra energy-delay product (EDP) over the CMOS latch. If there is no need for a complete polarization switching (in cases with a high data rate), it consumes less energy for each switching event.

Fig. 14 shows the energy and time for backup and restore operations. As supply voltage increases, backup and restore time decreases, and backup and restore energy increases. Being an intrinsic nonvolatile latch, simulation results in Fig. 14 also show that the proposed NCFET latches are very fast and energy-efficient for backup and restore operations. For example, under 0.8V VDD, the backup time is lower than 1.0ns, and the intrinsic restore time is lower than 75ns for both NV-Latch-N and NV-Latch-NC. It is noted that, in Fig. 14, the intrinsic restore time is obtained with an assumption of very fast supply voltage recovery in a few ps. In real scenarios the supply voltage recovers slower (usually much larger than ns) and the capability of fast intrinsic restore operation guarantees that the overall restore operation closely follow the supply recovery. In Fig. 14, the restore energy is obtained based on 0.8μs supply voltage ramp-up time to mimic real scenarios. Meanwhile, the backup and restore energy is lower than 3.0fJ for the entire operating range of 0.5V to 1.0V. This enables extremely low-energy backup and restore operations of the entire processor being carried out autonomously and in parallel without worrying about the peak current.

IV. PROPOSED NCFET NONVOLATILE DFFS

A nonvolatile DFF could be directly built by replacing the master latch or the slave latch with the proposed NCFET nonvolatile latches. For a positive-edge-triggered DFF, making the slave latch to be nonvolatile simplifies clocking during restore, as the slave latch is isolated from the input data when CLK is low (i.e. in hold/restore phase). In a pipelined design, the output Q/QN will be sent to the following stages, and setting the master slave latch of the subsequent-stage DFFs.

Fig. 15(a) shows an example with the slave latch implemented with an NV-Latch-N. This structure is simple and has the minimum overhead in terms of number of additional number of transistors. However, the driving strength of the first latch should be sufficient to make sure a new state could be set up in time for the slave latch. Also, the possible bleeding current from Dthrough the input switch affects the timing
Fig. 15. NCFET NV-DFF with the slave latch replaced by an NCFET NV-Latch. (a) Master and slave latches connected with switches; (b) Master and slave latches connected with clocked inverters for isolation.

behavior. Fig. 15(b) gives another example that uses clocked inverters as a buffer to drive the slave latch and to isolate the possible bleeding current between master and slave stages. The main shortcoming of the designs in Fig. 15 is the conflict between the differential input with an existing different latched state that causes extra energy and time to settle down.

In contrast, based on NV-Latch-NC, Fig. 16(a) shows another NV-DFF design, namely NV-DFF-NC. Fig. 16(b) shows the corresponding widely used CMOS volatile DFF scheme. In both designs, clocked inverters instead of CMOS switches are used to drive master and slave latches to provide better kick-back isolation between the two stages and the driver of \( D \). It is noted that the proposed NCFET NV-DFF has only 6 more transistors than the CMOS DFF, with a total of 30 transistors including a local clock driver.

Fig. 17 shows the transient simulation waveforms of the proposed NV-DFF-NC. With a steady \( V_{DD} \), the DFF carries out the positive-edge-triggered sample-and-hold operations. The intrinsic non-volatility of NV-Latch-NC is inherited. As shown in Fig. 17, during power failures, the polarization of the NCFETs maintains stable. When power supply recovers, the restore operations automatically restore the output \( Q \) and \( QN \) to its previous state before the power failure. In Fig. 17, the backup and restore operations for bit ‘0’ and bit ‘1’ are shown around 4-6\( \mu s \) and 9-11\( \mu s \), respectively.

Fig. 18 compares the backup and restore performance metrics with existing NV-DFF designs based on ferroelectric capacitor and NCFETs in [20]. This prior NCFET NV-DFF in [20] used a more complex backup and restore circuit scheme and needs multiple steps to carry out the operations. Meanwhile, the NCFET NV-DFF restore operation in [20] may cause static leakage current if the NCFET polarization state is positive (low-resistance state). Such a circuit design in [20] causes tens of times higher energy overheads. When compared with the NV-DFF based on stand-alone ferroelectric capacitor, more significant amount of energy could be saved for backup and restore operations. This is mainly because of the removal of the complex driving and sensing schemes for the ferroelectric capacitor. Thanks to the deeply embedded logic-in-memory operation in the proposed simple and effective circuit structure, the proposed design enables a new low-energy operation paradigm.

It is also noted that the proposed design has lower backup and restore speed than [20]. This is because of the adoption of low-power (high \( V_{TH} \)) CMOS transistors to achieve low leakage current for low-power IoT applications. For example, the prior NCFET NV-DFF has around 0.12\( \mu W \) static leakage power when operating at 0.8V supply as reported
in [20], while the proposed design has less than 0.2nW at 0.8V. Luckily, the backup and restore speed of the proposed NV- DFF is sufficiently fast for most scenarios, as charging and discharging the supply network on the chip following a power failure and recovery usually takes much longer time than a few nanoseconds. As will be discussed in Section V, the adoption of high \( V_{TH} \) MOSFETs in the main signal routes also improves the reliability during backup and restore operations.

Fig. 19 compares the energy versus delay between NV-DFF-NC and CMOS volatile DFF with \( x1 \) and \( x4 \) driving strength. The NCFET latch NV-DFF-NCx4 has around 35% energy-delay-product overhead. Note that the energy here includes the backup energy consumed after \( Q/\bar{Q} \) settles down. For applications with a higher bit-rate or dynamic frequency scaling (DFS) in which the clock cycle is shorter than the polarization switching time, the energy overhead could be much smaller. More importantly, for many IoT applications in which the processor is normally idle, most energy is consumed by MTJ and ReRAM. Furthermore, the great widening time window for write operations due to the structure also improves the reliability during backup and restore operations.

Table I also compares the proposed NV-DFF-NC with other reported NV-DFF designs at different technology nodes. Considering the lack of the same baseline of technology node and operating voltage, etc., Table I is not an apple-to-apple comparison. Nevertheless, there are fundamental advantages in the proposed NCFET NV-DFF over existing designs based on MTJ and ReRAM. This is because MTJ and ReRAM are two-terminal devices and the change of their memory state requires a static current or voltage across them for a certain period of time. Such static current, especially considering the widened time window for write operations due to the impact of device variations, significant amount of energy will be consumed by MTJ and ReRAM. Furthermore, the great scalability, low-voltage operation, high ON-OFF ratio, and more importantly, the unique external-control-free feature, highlight the advantage of the proposed promising NCFET NV-DFF designs.

For the proposed NV-DFF-NC in Fig. 16(a), the setup and hold time will be almost identical to that of the CMOS volatile DFF in Fig. 16(b). This property stems from a master latch of the same structure and size being used in both designs that features a clocked inverter to provide isolation.

V. FURTHER DISCUSSIONS ON RELATED APPLICATIONS

This section continues to discuss a few important perspectives relating to the design and applications.

A. Device Optimization

Given an NCFET structure and the ferroelectric material, the tunable NCFET design parameter is the ferroelectric layer thickness \( T_{FE} \). During device optimizations, the first concern is the retention time. For NCFET memory devices, it depends on the energy barrier between the two polarization states. Increasing \( T_{FE} \) helps increasing the coercive voltage that is required to change the polarization (see Fig. 4). However, it results in a larger minimum required supply voltage, which indicates more energy consumption each time the polarization is switched. Another device optimization considering yield will be discussed in the next sub-section.

There are also other design considerations, such as circuit area and fabrication cost, operating voltage range, endurance, etc. which will be of significance when the model becomes available. For the endurance consideration, the proposed NCFET NV-DFF may not be suitable for high-frequency applications. This makes improving device endurance useful.

B. Design Variation and Yield Analysis

It is noted that the restore functionality of the NCFET latches and DFFs depends on the difference of the sensed resistance from \( Q/\bar{Q} \) to GND or VDD, which indicates that the sensitivity of the sensed resistance may be critical for yield. For simplicity, only the sensed resistance to GND, i.e. \( R_{Q2GND} \), is analyzed here.

Taking NV-Latch-N for example, the sensed resistance is the sum of the series NCFET drain-source resistance \( R_{NCFET} \) and the NMOS drain-source resistance \( R_{CMOS} \). When storing a different latch bit information, the key difference in the initial sensed resistance, without considering process non-idealities, varies in \( R_{NCFET} \). As a result, the ratio of \( R_{Q2GND} \) between the two branches in the latch is

\[
\Gamma = \frac{R_{Q2GND}\, '/V}{R_{Q2GND}\, '/0} = \frac{(R_{NCFET}\, '/V + R_{CMOS}\, '/V)/(R_{NCFET}\, '/0 + R_{CMOS}\, '/0)}{R_{CMOS}\, '/V/(R_{NCFET}\, '/0 + R_{CMOS}\, '/0)}. \tag{1}
\]

A smaller \( \Gamma \) will lead to more stable restore operation and is more noise-resistant. A small \( R_{CMOS} \) or a large \( R_{NCFET} \) is thus helpful. Considering the orders of difference in the ON-OFF resistance of \( R_{NCFET} \), the approximation in (1) is rather safe. For this purpose, \( T_{FE} \) is set to be 8 nm so as to provide large ON-OFF state resistance while enabling low-voltage operation. Since the degradation of \( \Gamma \) can be easily caused by the variation of the NMOS initial resistance, \( R_{CMOS}\, '/0 \) and \( R_{CMOS}\, '/V \) analysis should be carried out. This is especially important for designs of NV-Latch-NC and NV-DFF-NC, because another parallel branch (see MSb and M6b in Fig. 11) is affecting \( \Gamma \), too.
All CMOS transistors have a higher $V_{TH}$ range, the design is reliable with results showing that, within the 0.5V to 1.0V supply voltage range, the time equal to $0.8V_{DD}$.

The results are summarized in Fig. 21, with $V_{TH}$ variation at different supply voltages for both NV-Latch-NC and NV-DFF-NC.

The difference in $R_{CMOS,0'}$ and $R_{CMOS,1'}$ mainly comes from device size mismatch and threshold voltage $V_{TH}$ variation $\Delta V_{TH}$. By manually adding an opposite in-series gate driving voltage to the gate, as shown in Fig. 20, the impact of $\Delta V_{TH}$ can be quantified through a series of simulations. The results are summarized in Fig. 21, with $V_{DD}$ ramp-up time equal to 0.8$\mu$s to mimic typical real scenarios. The results show that, within the 0.5V to 1.0V supply voltage range, the design is reliable with $\Delta V_{TH}$ no more than 30mV. Note that this result provides a fairly large margin for design, as all the MOSFETs in Fig. 20 are having an unfavorable direction of $V_{TH}$ variation if the NCFETs on the left and right branches store a negative and positive polarization state, respectively.

It is also interesting to find out, that the variation impact is independent on the supply voltage within the given range of 0.5V to 1.0V. This is because of the relatively long rising time for the supply voltage to recover, and the fact that the initial restore trend is almost equal for scenarios with different $V_{DD}$.

To provide a small $\Gamma$, the latch is designed in a way that all CMOS transistors have a higher $V_{TH}$ than the bottom two transistors connecting to NCFETs, i.e. M7 and M8 in Fig. 11 and Fig. 20. By doing this, the following goals could be achieved:

(i) The resistance of M7 and M5b, or that of M8 and M6b, plays a less significant role than that of NCFETs, as M7 is in series with the NCFET, and M5b is in parallel with the NCFET. This will help build the correct rising trend of Q and QN when $V_{DD}$ starts to recover.

(ii) Static leakage current of the latch will not increase. This can be guaranteed by a proper NCFET design with a high OFF-state resistance. Given a certain ferroelectric material and transistor structure, this OFF-state resistance could be tuned by varying $T_{FE}$ and the width of NCFET.

In this paper, the device variation of NCFET is not considered due to lack of model. Luckily, the large inherent ON-OFF resistance ratio will help to reduce the impact of NCFET variation.

C. Related Applications

The proposed NCFET NV-DFF could be strongly complementary to existing power-gating approaches in both low and high-performance systems. In aggressive, high-speed systems using fine-grained, low-latency power-gating techniques [40], the ability to power-gate stateful units up to and including entire processor cores within a handful of cycles would both expand the scope of what can be power-gated and simplify design constraints. While the impact of no need for backup and restore control for power-gating is still not yet fully explored, it is promising to open up new possibilities for further energy savings and architecture optimizations due to the reduced control complexity.

On the energy-harvesting end of the spectrum, one apparent benefit is the reduced backup and restore energy consumption and latency, which improves the utilization of harvested energy for the purpose computation. The intrinsic non-volatility ensures no missing backup without the need for backup control, leading to prevention of roll-back operations in the computation progress. While there are already works on nonvolatile processor optimizations [41]–[43], further architecture-level optimizations would be useful to capture the intrinsic non-volatility of NCFET flip-flops. Meanwhile, it has been shown that the recovery time in NVPs after a power emergency are sometimes dominated by the recovery of analog components, such as ensuring PLL stability [44], [45]. While this limits the impact of the rapid recovery time that NC-DFFs have in such systems, their rapid, completely distributed and low energy backup properties may allow power gating of data path and other digital components fast enough to divert energy during shorter or less severe power emergencies in order to preserve analog functionality, using the NCFET NV-DFF cycle-latency (at NVP frequencies) power gating potential to shave microseconds off recovery times.

VI. Conclusion

This paper has proposed a set of intrinsically nonvolatile latches and DFFs by harnessing the built-in non-volatility of negative capacitance field-effect transistors (NCFETs) with novel circuitry. Simulation results have shown fJ-level energy and ns-level intrinsic latency for a backup plus restore operation. With such low-energy and low-latency backup and restore operations, these fully external-control-free latches and DFFs enable a new nonvolatile computing paradigm for future IoT applications and power-gating applications.
ACKNOWLEDGMENT

The authors would like to thank Prof. Sayeef Salahuddin and Prof. Asif Khan from University of California, Berkeley, Dr. Francky Catthoor from IMEC, Prof. Sharon Hu from University of Notre Dame, and Prof. Peter Asbeck from University of California, San Diego, for useful discussions and suggestions.

REFERENCES


[19] M. H. Lee et al., “Prospects for ferroelectric HfZO FETs with experimentally C = 0.98 nF, SS = 42 mV/dec, SSrev = 28 mV/dec, switch-off <0.2 V, and hysteresis-free strategies,” in IEDM Tech. Dig., Dec. 2015, pp. 22.5.1–22.5.4.


Xueqing Li (M’13) received the B.S. degree and Ph.D. degree in electronics engineering from Tsinghua University, Beijing, China. Since 2013, he has been a Post-Doctoral Research Associate with the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA, USA. He has authored or co-authored over 50 journal and conference papers and is a reviewer of over ten IEEE/ACM journals. His interests include high-performance CMOS data converters and high-efficiency circuits and systems, including energy harvesting, low-voltage digital circuits, nonvolatile memory and circuits, and nonvolatile processing architectures using emerging devices, such as tunnel FETs, ferroelectric negative-capacitance FETs, and phase-transition FETs. He was a recipient of Best Paper Awards at the HPCA’15 and the ASP-DAC’17.

Sumitha George was born in India. She received the B.Tech. degree in electronics and communication from Kerala, India, and the M.Tech. degree from IIT Delhi. She joined the Systems and Technology Group, IBM, in 2007. Her main area of focus at IBM was microprocessor physical design. She is currently a Graduate Student of Computer Science and Engineering with The Pennsylvania State University. Her current research interests include computer architecture and circuit design.

Kaisheng Ma received bachelor’s degree (Hons.) in electrical engineering from Hangzhou Electronic University, Zhejiang, and the master’s degree from the Institute of Microelectronic, following Dr. D. Yu and Dr. X. Cao, Peking University, Peking, China. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, The Pennsylvania State University, following Dr. D. Yu and Dr. X. Cao, Peking University, Peking, China. His current research interests include nonvolatile processor architecture, machine learning, big data, neural networks, and neuromorphic computing.

Wei-Yu Tsai received the B.S. degree in electrical and control engineering from National Chiao Tung University, Taiwan, in 2009, and the M.S. degree in electrical engineering from National Tsing Hua University, Taiwan, in 2011. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, The Pennsylvania State University, USA. His current research interests include nonBoolean computing, such as neural networks and coupled oscillators, and the full-custom integrated circuit designs in CMOS and emerging devices such as III-V Heterojunction Tunnel FET.

Ahmedullah Aziz (S’10) received the B.Sc. degree from the Bangladesh University of Engineering & Technology in 2013, and the M.S. degree from The Pennsylvania State University in 2016, where he is currently pursuing the Ph.D. degree with the Department of Electrical Engineering. He is with the Integrated Circuits and Devices Lab as a Graduate Researcher. Previously, he was an Engineer with the Samsung Research and Development Institute. He was affiliated to the Tizen project in Solution Lab. His current research focus is on spin based memory and emerging steep slope device technology.

John Sampson (M’04) received the Ph.D. degree from the University of California at San Diego, San Diego, CA, USA. He is currently an Assistant Professor with the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA, USA. His current research interests include energy-efficient computing, architectural adaptations to exploit emerging technologies, and mitigating the impact of dark silicon.

Sunneet Kumar Gupta (M’12) received the B.Tech. degree in electrical engineering from IIT Delhi, New Delhi, India, in 2006, and the M.S. and Ph.D. degrees in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2008 and 2012, respectively. He is currently a Monkowski Assistant Professor of Electrical Engineering with The Pennsylvania State University, University Park, PA, USA. His current research interests include low-power variation aware very-large-scale integration circuit design, nanoelectronics and spintronics, device-circuit co-design, and nanoscale device modeling and simulations.

Meng-Fan Chang (M’05–SM’14) received the M.S. degree from the University of California at San Diego, La Jolla, CA, USA, and the Ph.D. degree from National Chiao Tung University, Hsinchu, Taiwan. He is currently a Full Professor with National Tsing Hua University, Hsinchu, Taiwan. He has was in industry over 10 years. From 1996 to 1997, he designed memory compilers in Mentor Graphics, NJ, USA. From 1997 to 2001, he designed embedded SRAMs and Flash in Design Service Division with Taiwan Semiconductor Manufacturing Company Ltd., Hsinchu. From 2001 to 2006, he was a co-founder and a Director of IPLib Company, Taiwan, where he developed embedded SRAM and ROM compilers, flash macros, and flat-cell ROM products. His current research interests include circuit designs for volatile and nonvolatile memory, ultra-low-voltage systems, 3-D-memory, circuit-device interactions, memristor logics for neuromorphic computing, and computing-in-memory. He is the corresponding author of numerous ISSCC, Symposium VLSI Circuits, EMD and DAC papers. He is an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS, the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, and IEICE ELECTRONICS. He has been serving on the technical program committees for ISSCC, JEDC, A-SSCC, ISCAS, VLSI-DAT, and numerous international conferences. He has been a Distinguished Lecturer of the IEEE Circuits and Systems Society. He has been serving as an Associate Executive Director of the Taiwan’s National Program of Intelligent Electronics from 2011 to 2016. He received the Academia Sinica (Taiwan) Junior Research Investigators Award in 2012, the Ts’I-Yu Wu Memorial Award of the National Science Council (NSC-Taiwan) in 2011. He also received numerous awards from the Taiwan’s National Chip Implementation Center, NTHU, MXIC Golden Silicon Awards, and ITRI.
Yongpan Liu (M’07–SM’15) is currently an Associate Professor with Tsinghua University. His current research interests include power design, emerging circuits and systems, and design automation. He has authored or co-authored over 100 papers and led over ten chip design projects for sensing applications, including the first nonvolatile processor (THU1010N) and has received Design Contest Awards from ISLPED2012, ISLPED2013 and best paper award HPCA2015 and served as TPC in DAC, ASPDAC, and ASSCC.

Suman Datta (F’13) was a Professor of Electrical Engineering with The Pennsylvania State University, University Park, PA, USA, from 2007 to 2011. He is currently the Chang Family Chair Professor of Engineering Innovation with the University of Notre Dame, Notre Dame, IN, USA. His current research interests include super-steep slope transistors, phase transition-based transistors, ferroelectric gated transistors, 3-D cross-point memories, coupled dynamical systems, and unsupervised learning systems.

Vijaykrishnan Narayanan (F’11) received the B.S. degree in computer science and engineering from the University of Madras, Chennai, India, in 1993, and the Ph.D. degree in computer science and engineering from the University of South Florida, Tampa, FL, USA, in 1998. He is currently a Professor of Computer Science and Engineering and Electrical Engineering with The Pennsylvania State University (Penn State), University Park, PA, USA. His current research interests include power-aware and reliable systems, embedded systems, nanoscale devices, and interactions with system architectures, reconfigurable systems, computer architectures, network-on-chips, and domain-specific computing.

Prof. Narayanan has received several awards, including the Penn State Engineering Society Outstanding Research Award in 2006, the IEEE CAS VLSI Transactions Best Paper Award in 2002, the Penn State CSE Faculty Teaching Award in 2002, the Association for Computing Machinery (ACM) Special Interest Group on Design Automation Outstanding New Faculty Award in 2000, the Upsilon Pi Epsilon Award for Academic Excellence in 1997, the IEEE Computer Society Richard E. Merwin Award in 1996, and the University of Madras First Rank in Computer Science and Engineering in 1993. He has received several certificates of appreciation for outstanding service from ACM and the IEEE Computer Society. He is the Editor-in-Chief of the IEEE TRANSACTIONS ON COMPUTER AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS.