# Benchmarking of Standard-Cell Based Memories in the Sub-$V_T$ Domain in 65-nm CMOS Technology

Pascal Meinerzhagen, *Student Member, IEEE,* S.M. Yasser Sherazi, *Student Member, IEEE,* Andreas Burg, *Member, IEEE,* and Joachim Neves Rodrigues, *Senior Member, IEEE*

***Abstract*—In this paper, standard-cell based memories (SCMs) are proposed as an alternative to full-custom sub-$V_T$ SRAM macros for ultra-low-power systems requiring small memory blocks. The energy per memory access as well as the maximum achievable throughput in the sub-$V_T$ domain of various SCM architectures are evaluated by means of a gate-level sub-$V_T$ characterization model, building on data extracted from fully placed, routed, and back-annotated netlists. The reliable operation at the energy-minimum voltage of the various SCM architectures in a 65-nm CMOS technology considering within-die process parameter variations is demonstrated by means of Monte Carlo circuit simulation. Finally, the energy per memory access, the achievable throughput, and the area of the best SCM architecture are compared to recent sub-$V_T$ SRAM designs.**

***Index Terms*—Embedded memory, flip-flop array, latch array, low-power, sub-$V_T$ operation, reliability, process parameter variations.**

## I. Introduction

DEVICES such as hearing aids, medical implants [1], and remote sensors impose severe constraints on size and energy dissipation. Supply voltage scaling reduces both active energy dissipation and leakage power. When applied aggressively, voltage scaling leads to sub-threshold (sub-$V_T$) operation [2]. In this regime, severely degraded on/off current ratios $I_{on}/I_{off}$ and increased sensitivity to process variations are the main challenges for sub-$V_T$ circuit design [3] in 65-nm technologies and below.

As an alternative to variation-tolerant full-custom circuit design, [4]–[6] promote the design of sub-$V_T$ circuits based on conventional standard-cell libraries. In such conventional standard-cell based designs, embedded memory macros may limit the scalability of the supply voltage, and thus the minimum achievable energy per operation, as the noise margins gradually decrease with the supply voltage, which leads to write and read failures in the sub-$V_T$ regime [7].

P. Meinerzhagen and A. Burg are with the Institute of Electrical Engineering, EPFL, Lausanne, VD, 1015 Switzerland (e-mail: pascal.meinerzhagen@epfl.ch, andreas.burg@epfl.ch).

Y. Sherazi and J. Rodrigues are with the Department of Electrical and Information Technology, Lund University, Lund, 22100 Sweden (e-mail: yasser.sherazi@eit.lth.se, joachim.rodrigues@eit.lth.se).

This work was kindly supported by the Swiss National Science Foundation under the project number PP002-119057. The project was conducted at Lund University with financial support from the Swedish VINNOVA Industrial Excellence Centre (SOS) and Swedish Foundation for Strategic Research (SSF)



Manuscript received December 31, 2010; revised May 20, 2011.

The main options for embedded memories which may be operated reliably in the sub-$V_T$ domain are: 1) specially designed SRAM macros, and 2) storage arrays built from flip-flops or latches. Standard SRAM designs require non-trivial modifications to function reliably in the sub-$V_T$ regime [3], [8]–[13]. However, flip-flop and latch arrays, commonly referred to as *standard-cell based memories* (SCMs), originally intended for super-$V_T$ operation [14], and easily synthesized with standard digital design tools may directly be adopted in the sub-$V_T$ domain, where they are still fully functional.

Beside being immediately compatible with voltage scaling until deep into the sub-$V_T$ domain, SCMs bring other advantages over SRAM macros. The use of SCMs described in a hardware description language eases the portability of a design to other technologies and modifications in the memory configuration at design time. Furthermore, designs comprising SCMs can be placed automatically using the standard place-and-route tools. Consequently, SCMs may be merged with logic blocks, which may improve data locality [15] and reduce routing. Also, for reconfigurable designs targeting low power consumption, memories are preferably organized in many small blocks which can be turned on and off separately. In the context of such fine-granular memory organizations, SCMs provide more flexibility, which may result in smaller overall area, and are more adequate to reduce the overall power consumption.

*Contribution:* In this paper, the SCM architectures reported in [14] are reconsidered in the sub-$V_T$ regime. The analysis is extended to account for the energy per memory access and the maximum achievable frequency with sub-$V_T$ voltage scaling. By means of Monte Carlo circuit simulation, it is shown that SCM architectures operate reliably in the sub-$V_T$ domain even in the presence of within-die process parameter variations. Finally, the best SCM architecture is compared to full-custom sub-$V_T$ SRAM designs regarding the energy per memory access, the maximum achievable throughput, and the silicon area.

*Outline:* Sections II and III introduce the investigated SCM architectures and explain the sub-$V_T$ characterization model, respectively, before the different architectures are characterized and compared by means of this model in Section IV. Section V verifies the reliability of SCMs in the sub-$V_T$ domain, while Section VI compares SCMs to full-custom SRAM macros. Section VII concludes the paper.

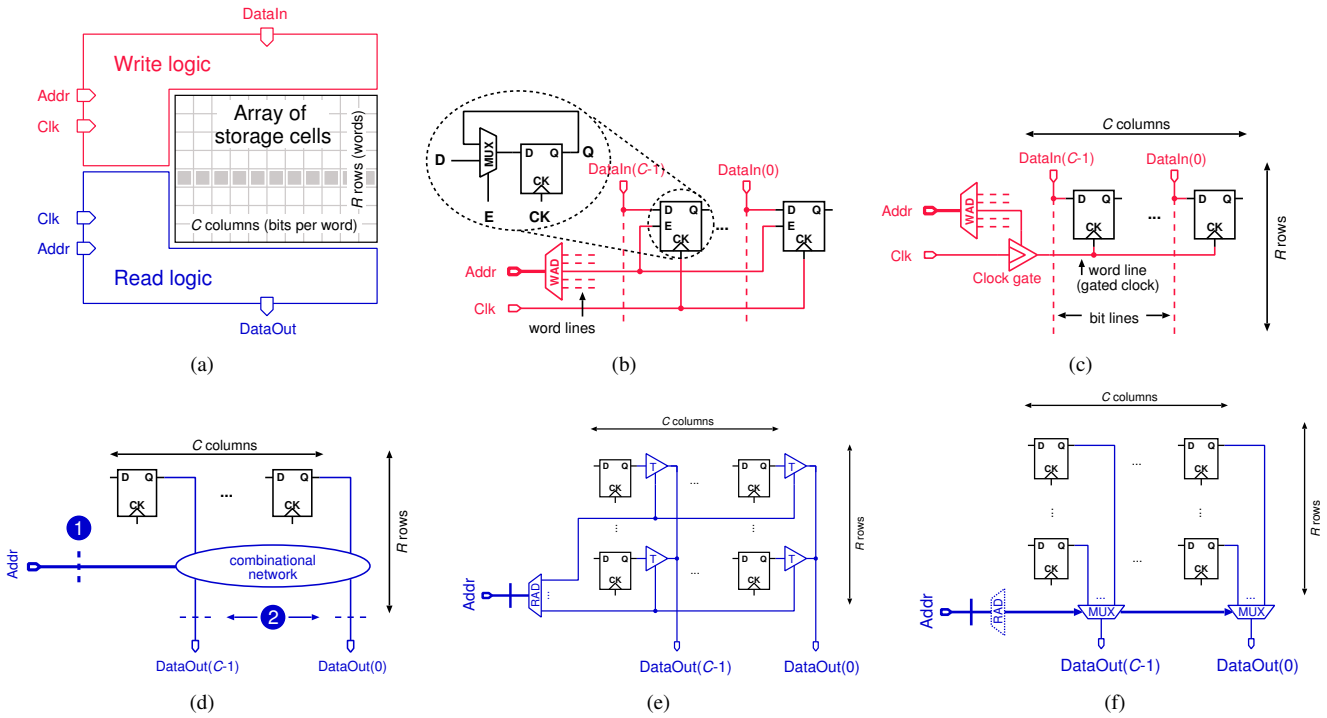



Fig. 1. Building blocks of a generic standard-cell based memory architecture (a). Write logic relying on enable flip-flops (b) and basic flip-flops in conjunction with clock-gates (c). Achieving typical one-cycle read latency (d). Read logic relying on tri-state buffers (e) and multiplexers (f).

## II. Standard-Cell Based Memory Architectures

The remainder of this paper assumes SCMs with a separate read and write port, a word access scheme, and a read and write latency of one cycle, which are typical requirements for memories distributed within dedicated datapaths. As shown in Fig. 1(a), any such SCM accomodates the following building blocks: 1) a write logic, 2) a read logic, and 3) an array of storage cells. Different ways to implement the write and read logic are presented in Sections II-A and II-B, respectively, assuming flip-flops as storage cells. The use of latches instead of flip-flops as storage cells is discussed in Section II-C.

### A. Write Logic

Consider an array of $R \times C$ flip-flops, where $R$ and $C$ denote the number of rows (words) and the number of columns (bits per word), respectively. Assuming a word-access scheme and a write latency of one cycle, the write logic needs to select one out of $R$ words, according to the given write address, and update the content of the corresponding flip-flops on the next active clock edge. Accordingly, the *write address decoder* (WAD) produces one-hot encoded row select signals, which select one row of the flip-flop array. Next, the flip-flops in the selected row need to update their state according to the data to be written. One option is to use flip-flops with enable feature or with a corresponding logic, as shown in Fig. 1(b). A second option is to use basic flip-flops in conjunction with clock-gates, as shown in Fig. 1(c), which generate a separate clock signal for each row so that only the currently selected row receives a clock pulse to sample the provided data, while all other rows receive a silenced clock, thereby keeping their current state.

### B. Read Logic

As shown in Fig. 1(d), the read logic may be purely combinational or contain sequential elements, which leads to a read latency. Assuming a word access scheme, one out of $R$ words needs to be routed to the data output, according to the read address. The typical one-cycle latency is obtained by inserting flip-flops either at the read address input, see case (1) in Fig. 1(d), or at the data output, see case (2) in Fig. 1(d). The former and latter case require $\text{ceil}(\log_2(R))$ and $C$ additional flip-flops, impose gentle and hard read address setup-time requirements, and cause considerable and negligible output delays, respectively. The task of routing one out of $R$ words to the output is accomplished using either tri-state buffers or multiplexers.

*1) Tri-state buffer based read logic:* This approach asks for a *read address decoder* (RAD) to produce one-hot encoded row select signals, and $R \cdot C$ tri-state buffers, i.e., exactly one per storage cell, as shown in Fig. 1(e). Notice that it is generally difficult to buffer tri-state buses [16], which might be necessary to maintain reasonable slew rates if these buses are routed over long distances.

*2) Multiplexer based read logic:* $C$ parallel $R$-to-1 multiplexers are required to route an entire word to the output, as shown in Fig. 1(f). The R-to-1 multiplexer may be implemented in many ways. Binary selection tree multiplexers do not require one-hot encoded row select signals and can therefore save the RAD. However, some glitches or activity on unselected data inputs can propagate all the way to the input of the last stage, giving rise to unnecessary power consumption. A better approach is to use a glitch-free RAD to mask (AND operation) unselected data at the leaf-level of an OR-tree to

realize the multiplexer functionality.

### C. Array of Storage Cells

Instead of flip-flops, latches can be used as storage cells, while the previous discussions on the write and read logic remain valid. However, setup-time requirements on the write port become considerably more stringent when using latches. The reason for this is that when sticking to a single-edge-triggered one-phase clocking discipline and a duty cycle of 50%, the WAD together with the clock-gates in the latch-based design can use only the first half of a clock period to generate one clock pulse and $R-1$ silenced clocks, which will make the latches in one out of $R$ rows transparent and keep the latches in all other rows non-transparent, during the second half of the clock period. The latches, which receive a clock pulse, store the applied input data on the next active clock edge.

Furthermore, if the currently transparent latches are also selected by the output multiplexers, the SCM becomes transparent from its data input to its data output, and combinational loops through external logic can arise. To avoid this problem, a restriction on the choice of read and write addresses needs to be imposed. If such a restriction is not desired, latches which are non-transparent during the second half of the clock period needs to be inferred at either the SCM's data input or output, or alternatively, registers needs to be inserted into any path feeding the SCM's data output back to the data input.

## III. Sub-$V_T$ Modeling

To exhaustively compare energy dissipation and critical path delay of the various SCM architectures, a gate-level sub-$V_T$ characterization flow is applied. The sub-$V_T$ characterization model is briefly described in Section III-A, and its accuracy is discussed in Section III-B.

### A. Sub-$V_T$ Characterization Model

The total energy dissipation $E_T$ of static CMOS circuits operated in the sub-$V_T$ regime is modelled as

$$E_T = \underbrace{\alpha C_{tot} {V_{DD}}^2}_{E_{dyn}} + \underbrace{I_{leak} V_{DD} T_{clk}}_{E_{leak}} + \underbrace{I_{peak} t_{sc} V_{DD}}_{E_{sc}}, \quad (1)$$

where $E_{dyn}$, $E_{leak}$, and $E_{sc}$ are the average energy dissipation due to switching activity, the energy dissipation resulting form integrating the leakage power over one clock cycle $T_{clk}$, and the energy dissipation due to short circuit currents, respectively. The energy dissipation $E_{sc}$ has been shown to be negligible in the sub-$V_T$ regime [17]. The switching current causing the energy dissipation $E_{dyn}$ results from sub-threshold currents [18], i.e., from the drain currents of MOS transistors whose gate-to-source voltage $V_{GS}$ is equal to or lower than the threshold voltage $V_T$ ($V_{GS} \leq V_T$). Whenever the sub-threshold current is not used to switch a circuit node, it contributes to $E_{leak}$ together with all other types of leakage currents.

For a given clock period $T_{clk}$, (1) may be rewritten as

$$E_T = \mu_e C_{inv} k_{cap} {V_{DD}}^2 + k_{leak} I_0 V_{DD} T_{clk}, \quad (2)$$

where $I_0$ and $C_{inv}$ are the average leakage current and the input capacitance of a single inverter, respectively. Furthermore, $k_{leak}$ and $k_{cap}$ are the average leakage and the capacitance of the circuit, respectively, both normalized to a single inverter. Moreover, $\mu_e$ is the circuit's average switching activity.

In the sub-$V_T$ domain, it is beneficial to operate at the maximum achievable frequency to reach minimum energy dissipation per operation. In the following, (2) is therefore rewritten for the case where the clock period $T_{clk}$ is equal to the critical path delay ($T_{clk}$ denotes the critical path delay in the remainder of this section). The critical path delay itself may be written as

$$T_{clk} = k_{crit} T_{sw\_inv}, \quad (3)$$

where $k_{crit}$ is the critical path delay of the circuit normalized to the inverter delay $T_{sw\_inv}$. In [17], the delay $T_{sw\_inv}$ of an inverter operating in the sub-$V_T$ regime is given by

$$T_{sw\_inv} = \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}, \quad (4)$$

where $n$ and $U_t$ denote the slope factor and the thermal voltage, respectively. By introducing (4) into (3), the the critical path delay is now given by

$$T_{clk} = k_{crit} \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}, \quad (5)$$

and the reciprocal of (5) defines the maximum frequency at which the circuit may be operated for a given supply voltage $V_{DD}$.

Finally, the total energy dissipation $E_T$ assuming operation at the maximum frequency is found by introducing (5) into (2), which yields

$$E_T = C_{inv} {V_{DD}}^2 \left[ \mu_e k_{cap} + k_{crit} k_{leak} e^{-V_{DD}/(nU_t)} \right]. \quad (6)$$

The key parameters which this sub-$V_T$ characterization model relies on are extracted from fully placed, routed, and back-annotated netlists and gate-level power simulations. For the architectural analysis presented in the following section, (6) has been used. For more details, the reader is referred to [19].

### B. Accuracy of Sub-$V_T$ Model

In [19], the accuracy of the sub-$V_T$ characterization model is verified by comparison with HSPICE transient simulations. It is found that the sub-$V_T$ model predicts the energy dissipation with less than 3.8 % error for all considered ISCAS85 benchmark circuits.

Furthermore, accuracy of the model is validated by measurements in [6] and [20]. It is shown that the measured energy is in the near vicinity of the simulated energy dissipation. The mean of the absolute modelling error is calculated as 5.2 %, with a standard deviation of 6.6 %. Moreover, it is also shown that the predicted maximum frequency at a given $V_{DD}$ matches well with the measured maximum frequency of the implemented ASIC.

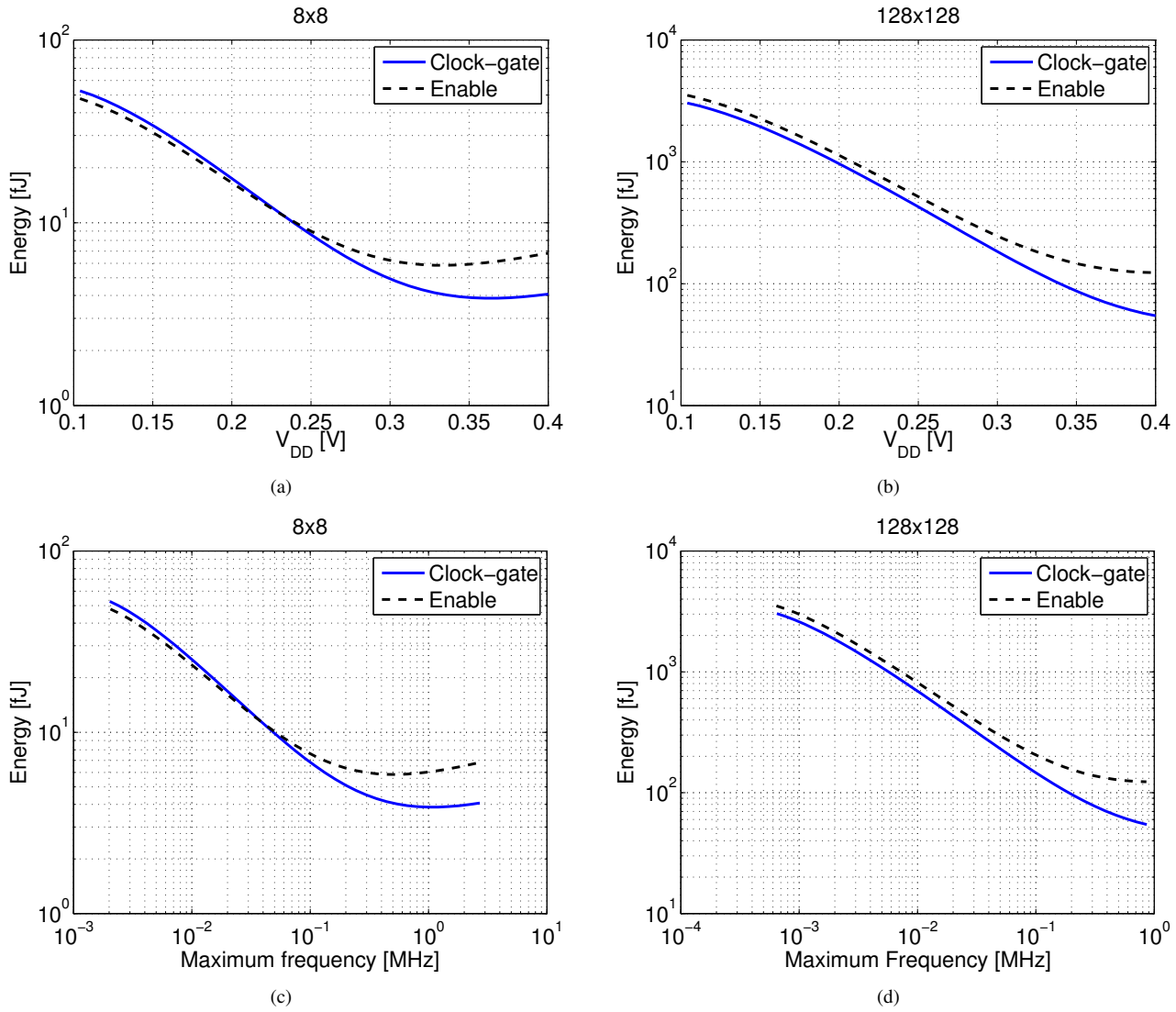



Fig. 2. Energy versus $V_{DD}$ for different write logic implementations, namely *enable flip-flops* and *basic flip-flops in conjunction with clock-gates*, assuming a multiplexer based read logic, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$. Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (c) and (d).

## IV. SCM Architecture Evaluation

After the presentation of different architectural choices for SCMs and the sub-$V_T$ characterization model, we now aim at identifying the SCM architecture that performs best in terms of energy, but also in terms of throughput, and silicon area. All SCMs are mapped to a 65-nm CMOS technology with low-power (LP) high threshold-voltage (HVT) transistors ($V_T$ is above 450 mV) and the results are based on fully synthesized, placed, and routed netlists with back-annotated layout parasitics. The average switching activity $\mu_e$ is obtained using voltage change dumps (VCDs) for 1000 write and read cycles. All inputs of the SCMs are driven by buffers of standard driving strength; highly capacitive nets such as the bit lines are buffered inside the SCMs. For the comparisons between SCMs of different sizes $R \times C$, energy figures are reported as *energy per written bit* and *energy per read bit*, commonly referred to as *energy per accessed bit*. In Sections IV-A and IV-B the different implementations of the write and read ports are compared and in Section IV-C flip-flop arrays are compared with latch arrays.

### A. Comparison of Write Logic Implementations

In order to compare different write logic implementations, we choose a multiplexer-based read logic and flip-flops as storage cells. We consider two memory configurations ($R = 8$, $C = 8$ and $R = 128$, $C = 128$) which are expected to have a smaller and to full-custom sub-$V_T$ SRAM designs comparable area cost, respectively.

Fig. 2(a) and Fig. 2(b) show the energy per written bit as a function of the supply voltage $V_{DD}$ for the small and the larger memory configuration, respectively. In both cases, the write logic relying on clock-gates in addition to basic flip-flops exhibits lower energy per written bit than the architecture that employs flip-flops with enable, for the range around the energy-minimum supply voltage. In the sub-$V_T$ regime, there are two main reason for this behavior: First, the architecture

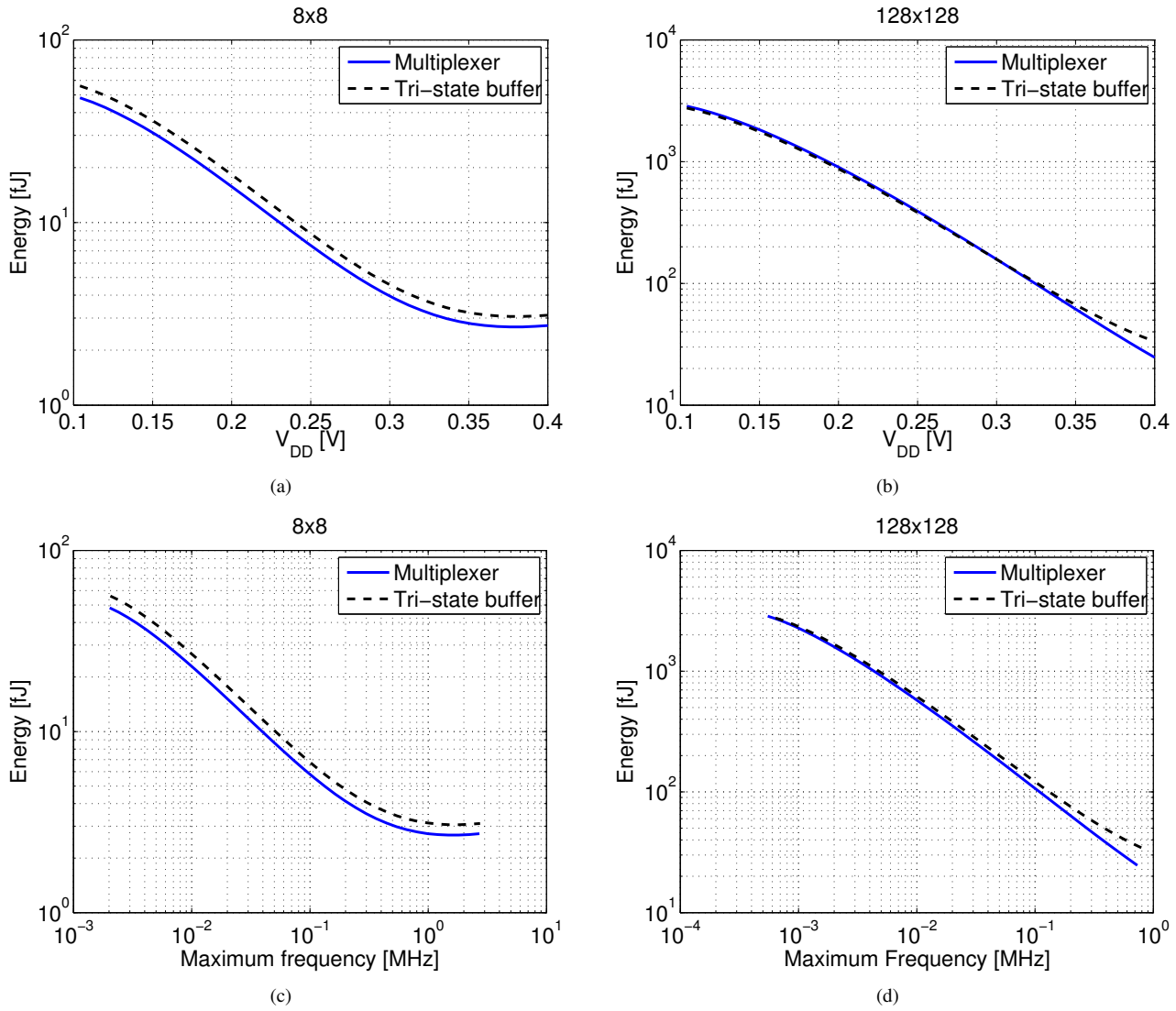



Fig. 3. Energy versus $V_{DD}$ for different read logic implementations, namely *tri-state buffers* and *multiplexers*, assuming a clock-gate based write logic and latches as storage cells, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$. Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (c) and (d).

based on clock-gates dissipates less active energy than the architecture based on enable flip-flops, as the latter distributes the clock signal to each storage cell, while the former silences the clock signal of all, but the selected row. The second reason is more visible for the larger storage array whose energy dissipation is dominated by leakage. This leakage is larger for the case of the more complex storage cells that require additional circuitry to realize the enable for each cell in a standard-cell based implementation.

For systems that require a constrained memory bandwidth, the energy dissipation at a given frequency may also be of interest. Fig. 2(c) and Fig. 2(d) show the energy per written bit as a function of the maximum achievable operating frequency of the corresponding SCM. The frequency range on the x-axis is obtained by sweeping $V_{DD}$ from 0.1 V to 0.4 V. It can be seen that both architectures have the same maximum operating frequencies, as the critical path is in the read logic through the output multiplexers.

With respect to area, the results in [14] show that the clock-gate architecture yields smaller SCMs than the enable architecture if only $C \geq 4$. This statement is true for many different CMOS technologies and standard-cell libraries.

In summary, the clock-gate architecture exhibits lower energy, equal throughput, and smaller area compared to the enable architecture and is therefore generally preferred.

### *B. Comparison of Read Logic Implementations*

In order to compare different read logic implementations, we choose the clock-gate based write logic and a latch-based storage array for again a small and a larger SCM configuration. Fig. 3(a) and Fig. 3(b) show that the multiplexer based read logic with RAD has a small advantage over the tri-state buffer based read logic in terms of energy per read bit, at least around the energy-minimum supply voltage. Fig. 3(c) and Fig. 3(d) show that there is no significant difference between the two read logic implementations as far as the maximum achievable

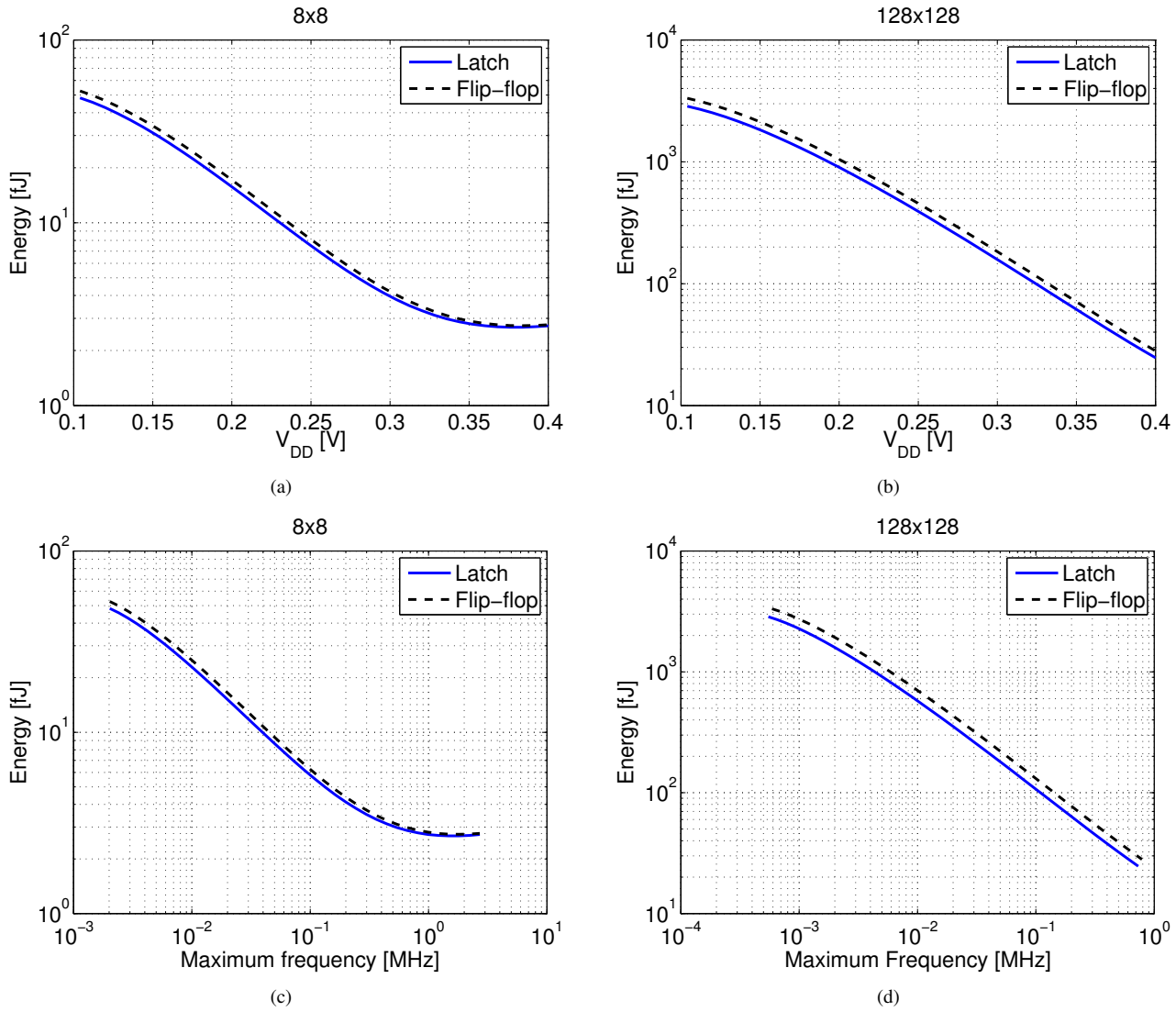



Fig. 4. Energy versus $V_{DD}$ for different storage cell implementations, namely *latches* and *flip-flops*, assuming a clock-gate based write logic and a multiplexer based read logic, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$. Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (c) and (d).

operating frequency is concerned. Indeed, the delay of the tri-state buffer is quite long and comparable to the delay through the entire multiplexer as all $R$ tri-state buffers in one column are connected to the same net, which consequently has a high capacitance.

In summary, multiplexer based SCMs have a small energy and an area advantage [14], compared to the tri-state buffer approach and are therefore preferred.

### C. Comparison of Storage Cell Implementations

In order to compare different storage cell implementations, the best write and read logic implementations and again a small and a larger SCM block are considered. Fig. 4(a) and Fig. 4(b) show that latch arrays have less energy per accessed bit than flip-flop arrays, due to smaller leakage currents drained in each storage cell and due to lower active energy of the latch implementation. However, the energy savings of using latches instead of flip-flops are only small: a latch has around 2/3 the leakage of a flip-flop in the considered standard-cell library, but only around 2/3 of all cells in an SCM are storage cells, which accounts for the approximately 22 % energy reduction visible from Fig. 4(d).

Fig. 4(c) and Fig. 4(d) show that there is no significant difference in terms of maximum frequency. In fact, the storage cells are not in the critical path, since the critical path of any SCM is through the RAD and the tri-state buffers or the multiplexers. However, flip-flops as sotrage cells allow for shorter write address setup-times than latches, as described in Sec. II-C.

Latch arrays have only slightly smaller area than flip-flop arrays [14]. Table I shows the standard-cell area $A_{SC}$ and the area $A_{P\&R}$ of fully placed and routed latch and flip-flop arrays for different configurations $R \times C$, the clock-gate based write logic, and the multiplexer based read logic. Notice that $A_{P\&R} = A_{SC}/0.75$, as the SCMs have been successfully placed and routed with a typical initial floorplan utilization of 75 %. An

TABLE I
STANDARD-CELL AREA $A_{SC}$ AND AREA $A_{P\&R}$ OF FULLY PLACED AND ROUTED LATCH AND FLIP-FLOP ARRAYS FOR DIFFERENT CONFIGURATIONS $R \times C$, CLOCK-GATE BASED WRITE LOGIC, AND MULTIPLEXER BASED READ LOGIC.

| | | Latch array | | Flip-flop array | |
|---|---|---|---|---|---|
| $R$ | $C$ | $A_{SC}$ [$\mu m^2$] | $A_{P\&R}$ [$\mu m^2$] | $A_{SC}$ [$\mu m^2$] | $A_{P\&R}$ [$\mu m^2$] |
| 8 | 8 | 738 | 984 | 811 | 1.1k |
| 8 | 32 | 2.5k | 3.3k | 2.8k | 3.7k |
| 8 | 128 | 9.5k | 12.7k | 10.6k | 14.1k |
| 32 | 8 | 2.9k | 3.8k | 3.1k | 4.2k |
| 32 | 32 | 9.9k | 13.2k | 10.9k | 14.6k |
| 32 | 128 | 37.9k | 50.6k | 42.1k | 56.2k |
| 128 | 8 | 11.2k | 15.0k | 12.3k | 16.4k |
| 128 | 32 | 39.4k | 52.5k | 43.7k | 58.3k |
| 128 | 128 | 152.2k | 202.9k | 169.0k | 225.4k |

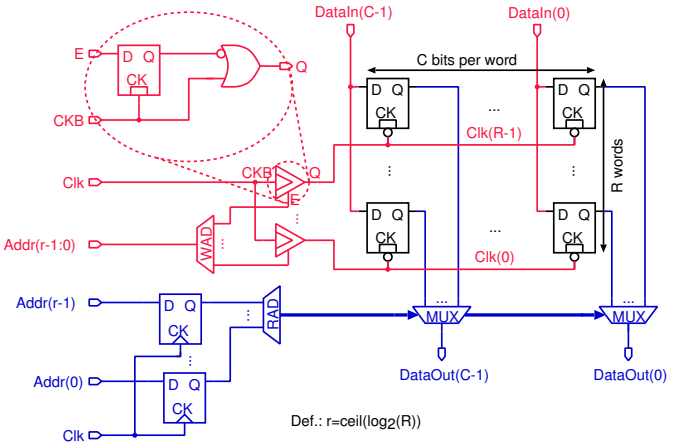



Fig. 5. Schematic of latch based SCM with clock-gates for the write logic and multiplexers for the read logic.

approximation of the area $A(R,C)$ for an arbitrary memory configuration $R \times C$ can be found according to

$$A(R,C) = \beta_1 + \beta_2 R + \beta_3 C + \beta_4 RC + \beta_5 \text{ceil}(\log_2(R)) + \beta_6 \text{ceil}(\log_2(C)). \quad (7)$$

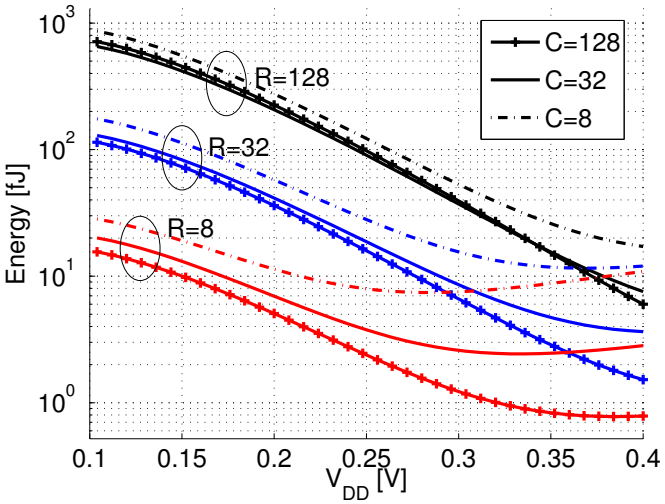
The coefficients $\beta_1 \ldots \beta_6$ are obtained through a least squares fit to a set of reference configurations in the technology under consideration such as the ones provided in Table I.

To summarize, latch arrays have slightly less energy per accessed bit, achieve the same frequency, and are smaller compared to flip-flop arrays.

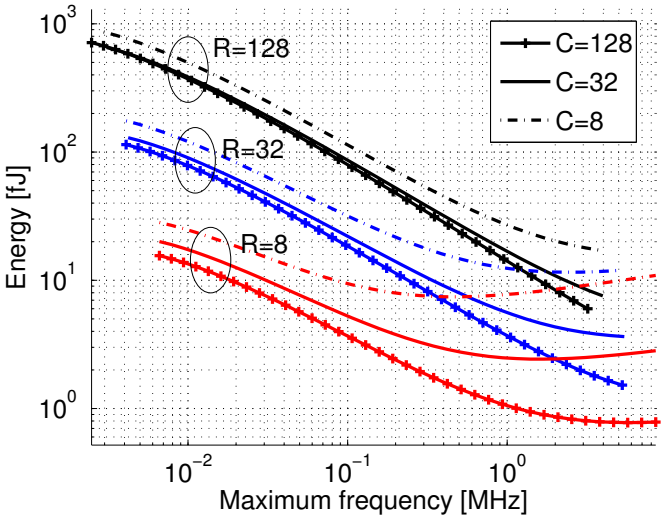
### D. Best Practice Implementation

Fig. 5 shows the schematic of the best SCM architecture. This architecture uses latches without enable feature as storage cells, clock-gates for the write logic, and multiplexers for the read logic.

With respect to the energy efficiency, we note that a significant switching activity is required to find an energy-minimum, which occurs only for the smallest memory configurations. However, for the large memory configurations, the overall switching activity is very low and the energy dissipation is clearly dominated by the integration of the leakage power over the access time, which decreases with increasing $V_{DD}$ if always operating at maximum speed. Consequently, the energy-minimum supply voltage within the sub-$V_T$ domain approaches the threshold voltage $V_T$ when increasing the memory size.




(a)




(b)

Fig. 6. Energy versus $V_{DD}$ (a) and energy versus frequency (b) for the *latch multiplexer clock-gate* architecture for different memory configurations.

For different memory configurations with the same storage capacity ($R \cdot C = \text{const.}$), we observe from Fig. 6(a) and Fig. 6(b) that the energy-efficiency improves for a larger number of columns $C$ and a smaller number of rows $R$. The reason for this behavior is that the maximum operating frequency increases as $R$ decreases which again reduces the contribution of the energy consumed due to leakage power in each access cycle.

## V. Reliability Analysis

Besides the desire to operate at the energy-minimum, one of the limiting factors with respect to voltage scaling in the sub-$V_T$ domain is the reliability of the circuit. Reliability issues arise mainly from within-die process variations and

are aggravated in deep submicron technologies. Consequently, ensuring robust operation in the sub-$V_T$ regime has been one of the most important concerns in the design of full-custom sub-$V_T$ storage arrays.

Compared to full-custom designs, SCMs are compiled from conventional combinational CMOS logic gates, such as NAND, NOR, or AOI gates, and from sequential elements, i.e., latches and/or flip-flops. The reliability issue therefore corresponds to the discussion down to which supply voltage a given standard-cell library can operate reliably. This point limits in the same way the operation of the combinational and sequential logic and of the embedded SCMs for a given process corner.

To determine the range of reliable operation of the SCMs, we distinguish between the combinational and the sequential cells in the library, used to construct the storage array. Previous work shows that when gradually scaling down the supply voltage, the sequential cells fail earlier then the combinational CMOS logic gates [5], provided that the combinational logic is built without transmission gates. Therefore, the focus is on the analysis of the sequential elements in the following.

The peripherals of SCM storage arrays, i.e., the read and write logic, are built from combinational CMOS gates and are thus less sensitive to process variation than the array of storage cells itself. Also, delay variations in SCM peripherals induced by process variation are unproblematic due to the used single-edge-triggered one-phase clocking discipline where path delays do not necessarily need to be matched. Compared to SCM peripherals, the peripherals of SRAM arrays are more sensitive to process variation: delay variations may cause the sense amplifiers to be triggered at the wrong time, and mismatch in the sense amplifiers can further compromise reliability, especially at very low supply voltages.

### A. Sensitivity of SCMs to Variations

Reliability issues in both sequential standard-cells and in dedicated SRAM storage cells essentially arise from mismatch between carefully sized transistors due to *within-die* process variations [21]. In a conventional 6T-SRAM cell, such mismatch manifests itself in three types of failures: a) read failures, b) write failures, and c) hold failures. The read failures result from the direct access of the read bit line to the storage node which is not present in a standard latch design such as the one shown in Fig. 7, where the output is isolated from the internal node with a separate driver. The write failures in a 6T-SRAM cell are caused by the inability to flip storage nodes that suffer from an unusually strong keeper. The standard-cell latch avoids this issue by turning off the feedback path during write operation. The only remaining issue are hold failures which occurr in the non-transparent phase of a latch during which the circuit behavior essentially resembles that of a basic 6T-SRAM cell. Hence, a conventional standard-cell latch may be viewed as a very conservative SRAM cell design [8] where the reliability is determined by the risk of experiencing hold failures.

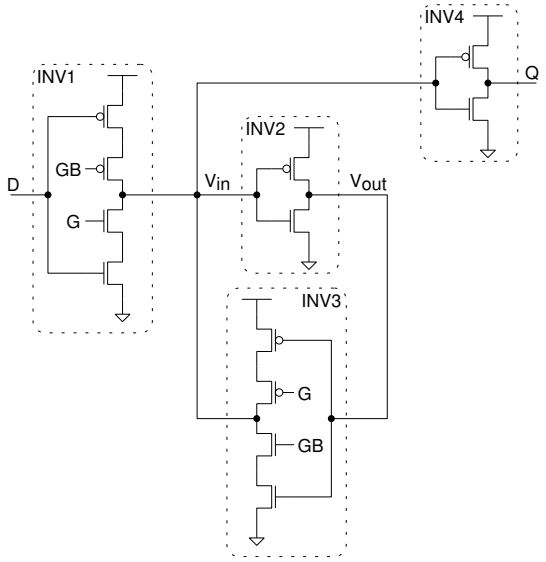



Fig. 7. Simplified schematic of the latch used in the best SCM architecture.

### B. Hold Failure Analysis

Fig. 7 shows a simplified schematic of the latch which was chosen by the logic synthesizer from a commercial standard-cell library in order to minimize leakage and area of the latch arrays described in this paper. The development of new libraries with special latch topologies is beyond the scope of this paper.

A latch needs to be able to hold data in the non-transparent phase. In this phase, INV2 and INV3 in Fig. 7 act as a cross-coupled inverter pair. The stability of the state of this pair is usually defined by the *static noise margin* (SNM) that is required to hold data in the presence of voltage noise on the storage nodes [22]. This SNM is extracted as the side of the largest embedded square for the butterfly curves shown in Fig. 8 for different supply voltages in the sub-$V_T$ domain. For each butterfly curve, there is an SNM associated with the top-left and the bottom-right eye, referred to as *SNM high* and *SNM low*. The probability distribution functions on the right-hand side of Fig. 8 are always for the minimum of *SNM high* and *SNM low*. The butterfly curves and the corresponding minimum SNM distributions are obtained from 1000-point Monte Carlo circuit simulation assuming within-die process parameter variations for the typical process corner at a temperature of 25 °C. All common parameters of the BSIM4 transistor simulation models are subject to variation according to statistical distributions provided by the foundry.

The distributions in Fig. 8 show that the SNM values decrease with the supply voltage. As can be seen in Fig. 8(a), there is a clear separation between the voltage transfer characteristic (VTC) of inverter INV2 and the inverse VTC of inverter INV3 corresponding to a comfortable SNM for a supply voltage of 400 mV, which also corresponds to the energy optimum supply voltage for most SCM architectures and sizes. Fig. 8(b) and Fig. 8(c) show that there is still a separation between the VTCs even at lower supply voltages, indicating that operation is still possible, but the SNMs are small and reliability clearly starts to become critical at 250 mV,

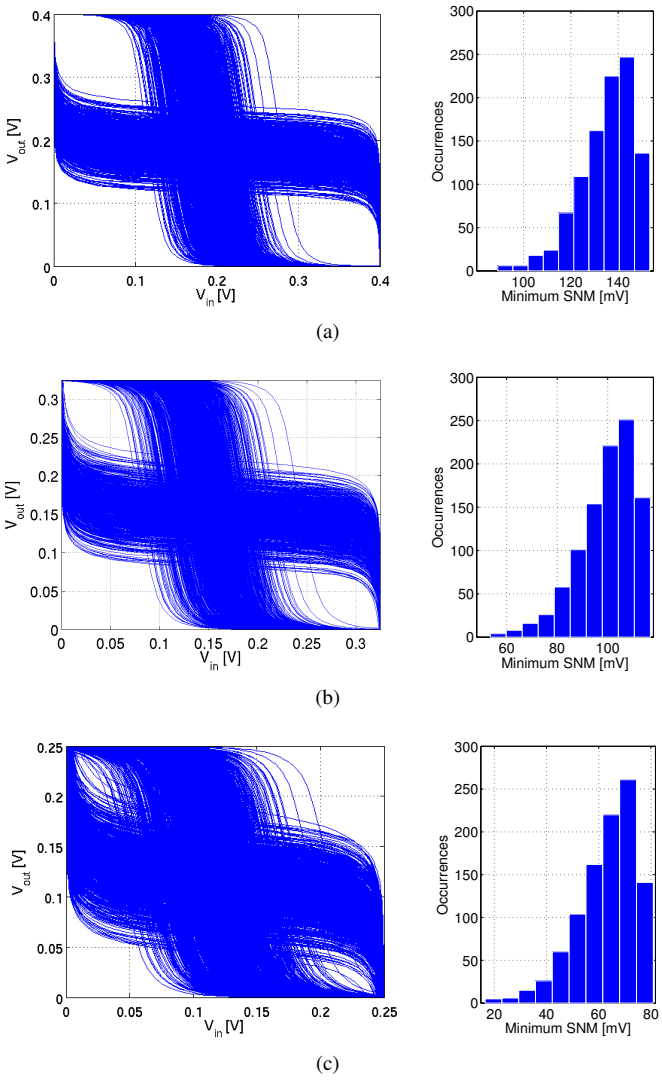



Fig. 8. Butterfly curves (left) and distribution of minimum hold SNM (right) of the latch used in the best SCM architecture for (a) $V_{DD} = 400\,mV$, (b) $V_{DD} = 325\,mV$, and (c) $V_{DD} = 250\,mV$.

limiting the range of operation.

## VI. Comparison with Sub-$V_T$ SRAM Designs

In this section, the performance and cost of sub-$V_T$ SCMs is compared to a selection of sub-$V_T$ SRAM designs in literature [8]–[11], [13]. Section VI-A gives an overview of recent sub-$V_T$ memory implementations including this work. Section VI-B compares the energy and throughput of the smallest SCM architecture with a prominent sub-$V_T$ SRAM design, while Section VI-C compares their area.

### A. Overview

Table II presents a selection of recently published sub-$V_T$ memories. $V_{DDmin}$ is defined as the minimum supply voltage, which guarantees reliable write, hold, and read operations. Unless otherwise stated, the maximum operating frequency $f_{max}$ is given for $V_{DD} = V_{DDmin}$. The reported energy includes both active energy for a read operation and the leakage energy

TABLE II
Comparison of sub-$V_T$ memories.

| Publication | [8] | [9] | [10] | [11] | [13] | This |
|---|---|---|---|---|---|---|
| Capacity [kbit] | 256 | 256 | 64 | 8 | 480 | 32 |
| Tech. [nm] | 65 | 65 | 65 | 90 | 130 | 65 |
| Basis of results | ASIC measurements | | | | | Post-layout |
| $V_{DDmin}$ [mV] | 380[a] | 350[c] | 300 | 160 | 200 | 300 |
| $f_{max}$ [kHz] | 475 (0.4 V) | 25 | 20 (0.25 V) | 200 | 120 | 1 000 (0.4 V) |
| Energy [fJ/bit] | 65.6 (0.4 V) | 884.4 | 86.0[d] (0.4 V) | 750[e] | 4.2 | 32.7 (0.4 V) |
| Area [$\mu m^2$/bit] | 2.9[b] | 4.0[b] | 7.0[b] | 19.5 | 12.8 | 12.5 |

[a]One redundant row and column per 32-kbit block are assumed to guarantee reliable operation at this supply voltage.
[b]Area estimated from die photograph.
[c]Plus 50 mV for boosting of word line drivers.
[d]Estimation extracted from a graph.
[e]Includes the energy dissipation of the package.

of the memory array during the access time. Furthermore, the total energy value is normalized by the width of the data IO bus, thereby reporting the total energy per read bit. Unless otherwise stated, the energy is given for $f_{max}$ at $V_{DDmin}$.

All sub-$V_T$ SRAM designs [8]–[10] realized in a 65-nm CMOS technology have $V_{DDmin} \geq 300\,mV$. Monte Carlo simulations indicate that SCMs mapped to the same technology should operate reliably at least down to the same minimum supply voltage. Two SRAM designs [11], [13] fabricated in older technologies are less sensitive to process parameter variations and are reported to have an even lower $V_{DDmin}$, i.e., 160 mV and 200 mV, respectively.

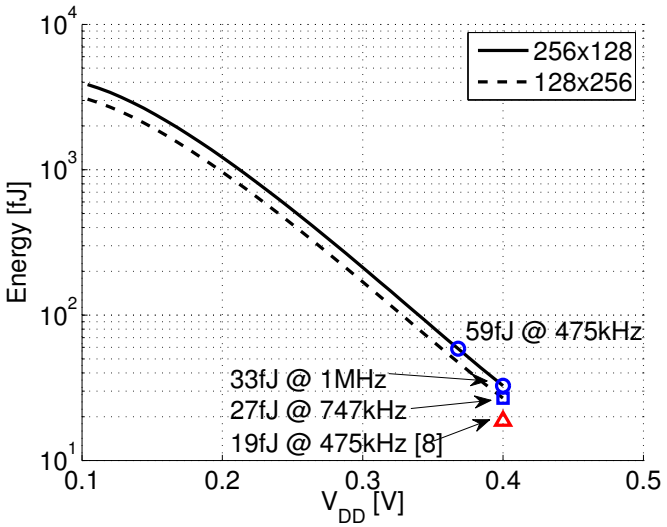
At the same technology node and supply voltage $V_{DD}$, SCMs are faster than SRAM designs, which bares the potential to lower energy dissipation per memory access if 1) speed is traded against energy, or 2) early task completion is honored by power gating. Obviously, older technologies exhibit lower leakage currents which may lead to lower energy per memory access.

With respect to area, the use of robust latches, available from conventional standard-cell libraries, instead of 8T or 10T SRAM cells, is clearly paid for by a larger area per bit for SCMs, in the same technology.

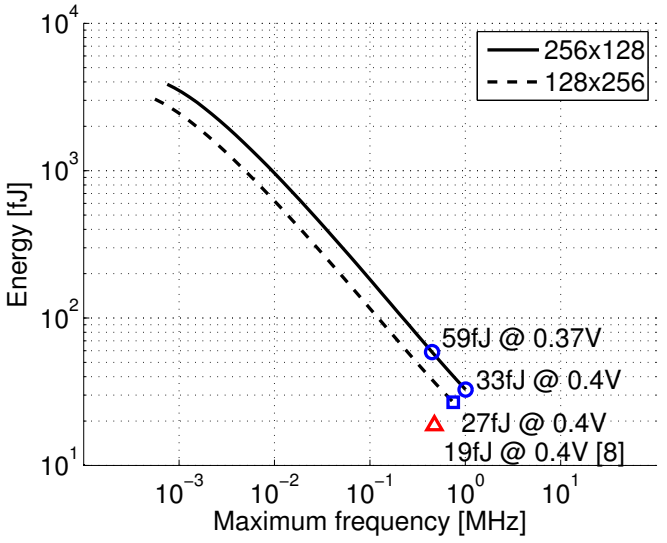
### B. Energy and Throughput

A well-cited 256-kbit 10T sub-$V_T$ SRAM [8] in 65-nm CMOS has 8 32-kbit blocks ($R = 256$, $C = 128$), which are served by a single 128-bit data IO bus. The leakage energy of this SRAM macro is divided by 8 to compare one block with the proposed 32-kbit SCM block, while the active energy is taken as is, since only one block is accessed at a time. At 400 mV, the SRAM macro is reported to be operational at $f_{max} = 475\,kHz$, and a single 32-kbit block dissipates 19 fJ per accessed bit, as indicated by the triangle in Fig. 9.

For comparison, Fig. 9(a), and Fig. 9(b), show the energy per accessed bit of the smallest SCM architecture as a function of $V_{DD}$ and $f_{max}$, respectively. Considering an SCM block with $R = 256$ and $C = 128$, $f_{max} = 475\,kHz$ is already achieved at $V_{DD} = 370\,mV$ and the energy per accessed bit for this




(a)




(b)

Fig. 9. Energy versus $V_{DD}$ (a) and energy versus frequency (b) for the *latch multiplexer clock-gate* architecture for $R = 256$, $C = 128$ and for $R = 128$, $C = 256$. The red triangle corresponds to [8].

operating point is 59 fJ, which is more than for the full-custom SRAM macro. However, when operated at the same supply voltage ($V_{DD} = 400\,\text{mV}$), the SCM is able to operate at $f_{max} = 1\,\text{MHz}$, with an energy dissipation of 33 fJ per accessed bit, which is only $1.7\times$ higher compared to the full-custom design. The energy savnings compared to the initial operating point are achieved due to a higher possible clock frequency combined with power gating after earlier completion of a task.

Changing the SCM configuration to $R = 128$ and $C = 256$ while keeping a constant storage capacity $R \cdot C$, the energy per accessed bit of the SCM is further reduced. As shown by the square marker in Fig. 9, this new SCM configuration is able to run at 747 kHz for $V_{DD} = 400\,\text{mV}$, and dissipates 27 fJ per read bit in this operating point, which is only $1.4\times$ higher than for the full-custom design. This change in the SCM configuration results in lower energy and doubled memory bandwidth at the price of a higher routing congestion during system integration.

### C. Area

The bitcell of SCMs (flip-flop or latch) is clearly larger than the SRAM bitcell. However, SRAM macrocells have an overhead to accommodate the peripheral circuitry, i.e., precharge circuitry and sense amplifiers [23]. For SRAM macrocells with small storage capacity, this area overhead may be significant. Hence, SCMs may outperform SRAM macrocells in terms of area for small storage capacities, but become bigger for large storage capacities. In [14], it is shown that the border up to which SCMs are still smaller than SRAM macrocells depends on the *number of words* and the *number of bits per word*, and may be as large as 1 kbit. However, [14] considers only circuit implementations for super-$V_T$ operation, i.e., SRAM macros based on the 6T bitcell and SCMs synthesized with a given timing constraint. When considering circuit implementations specifically optimized for sub-$V_T$ operation, SRAM macrocells become significantly larger due to the need for 8 T [9] or 10 T [8] bitcells and the additional assist circuits required for reliable sub-$V_T$ operation. As opposed to this, SCMs may be synthesized with relaxed timing constraints (and still reach 1 MHz in the current study) as speed is not of major concern for typical ultra-low-power applications and may therefore have a reduced area cost compared to super-$V_T$ implementations.

In the present case, considering a storage capacity of 32 kbit, the SCM is 4.3 times larger than a corresponding SRAM block [8]. For some applications, this area increase may be acceptable for the benefit of lower energy per memory access and higher throughput.

## VII. Conclusions

For standard-cell based ultra-low-power designs which need to operate in the sub-$V_T$ regime, standard-cell based memories (SCMs) are an interesting alternative to full-custom SRAM macros which must be specifically optimized to guarantee reliable operation. The main advantages of SCMs are the reduced design effort, reliable operation for the same voltage range as the associated logic, high speed (when compared to corresponding full-custom macros), and reasonably good energy efficiency for maximum-speed operation. The drawbacks are the area penalty (for storage arrays larger than a few kbit) and a loss in energy efficiency compared to full-custom designs when operating at the same clock frequency.

Energy-efficient SCM design is driven by the fact that most of the energy is consumed due to leakage while active energy plays only a minor role, especially for large configurations. A design based on latches using clock-gates for the write logic and glitch-free multiplexers for the read logic achieves the best energy efficiency and has the smallest silicon area. For the same maximum throughput but smaller write address setup-times, the latches may be replaced by flip-flops.

## References

[1] R. Sarpeshkar, "Ultra low power electronics for medicine," in *Proc. International Workshop on Wearable and Implantable Body Sensor Networks*, April 2006, pp. 1 pp.–37.

[2] J.-J. Kim and K. Roy, "Double gate-MOSFET subthreshold circuit for ultra low power applications," *in IEEE Trans. on Electron Devices*, vol. 51, no. 9, pp. 1468–1474, Sept. 2004.

[3] M. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 65nm SRAM achieving voltage scalability from 0.25-1.2V and performance scalability from 20kHz-200MHz," in *Proc. IEEE ESSCIRC*, sept. 2008, pp. 282–285.
[4] B. Calhoun, A. Wang, and A. Chandrakasan, "Device sizing for minimum energy operation in subthreshold circuits," in *Proc. IEEE Custom Integrated Circuits Conference*, oct. 2004, pp. 95–98.
[5] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *in IEEE J. of Solid-State Circuits*, vol. 40, no. 9, pp. 1778–1786, sept. 2005.
[6] J. Rodrigues, O. C. Akgun, and V. Owall, "A <1 pJ Sub-VT cardiac event detector in 65 nm LL-HVT CMOS," *Proc. VLSI-SoC*, June. 2010.
[7] J. Chen, L. Clark, and T.-H. Chen, "An ultra-low-power memory with a subthreshold power supply voltage," *in IEEE J. of Solid-State Circuits*, vol. 41, no. 10, pp. 2344–2353, oct. 2006.
[8] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *in IEEE J. of Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, march 2007.
[9] N. Verma and A. Chandrakasan, "A 65nm 8T sub-Vt SRAM employing sense-amplifier redundancy," in *Proc. IEEE ISSCC*, feb. 2007, pp. 328–606.
[10] M. E. Sinangil, N. Verma, and A. P. Chandrakasan, "A reconfigurable 8T ultra-dynamic voltage scalable (U-DVS) SRAM in 65 nm CMOS," *in IEEE J. of Solid-State Circuits*, vol. 44, no. 11, pp. 3163–3173, nov. 2009.
[11] S.-C. Luo and L.-Y. Chiou, "A sub-200-mV voltage-scalable SRAM with tolerance of access failure by self-activated bitline sensing," *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 57, no. 6, pp. 440–445, june 2010.
[12] M.-F. Chang, J.-J. Wu, K.-T. Chen, Y.-C. Chen, Y.-H. Chen, R. Lee, H.-J. Liao, and H. Yamauchi, "A differential data-aware power-supplied ($D^2$AP) 8T SRAM cell with expanded write/read stabilities for lower VDDmin applications," *IEEE J. of Solid-State Circuits*, vol. 45, no. 6, pp. 1234–1245, june 2010.
[13] T.-H. Kim, J. Liu, J. Keane, and C. Kim, "A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme," in *Proc. IEEE ISSCC*, feb. 2007, pp. 330–606.
[14] P. Meinerzhagen, C. Roth, and A. Burg, "Towards generic low-power area-efficient standard cell based memory architectures," in *Proc. IEEE International Midwest Symposium on Circuits and Systems*, aug. 2010, pp. 129–132.
[15] C. Roth, P. Meinerzhagen, C. Studer, and A. Burg, "A 15.8 pJ/bit/iter quasi-cyclic LDPC decoder for IEEE 802.11n in 90 nm CMOS," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Nov. 2010.
[16] J. Lillis and C.-K. Cheng, "Timing optimization for multisource nets: characterization and optimal repeater insertion," *in IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, no. 3, pp. 322–331, Mar 1999.
[17] E. Vittoz, *Low-Power Electronics Design*. CRC Press, 2004, ch. 16.
[18] H. Soeleman, K. Roy, and B. Paul, "Robust subthreshold logic for ultra-low power operation," *in IEEE Trans. VLSI Systems*, vol. 9, no. 1, pp. 90–99, Feb 2001.
[19] O. C. Akgun and Y. Leblebici, "Energy efficiency comparison of asynchronous and synchronous circuits operating in the sub-threshold regime," *in J. of Low Power Electronics*, vol. 4, OCT 2008.
[20] O. Akgun, J. Rodrigues, Y. Leblebici, and V. Owall, "High-level energy estimation in the sub-VT domain: Simulation and measurement of a cardiac event detector," *in IEEE Trans. on Biomedical Circuits and Systems*, Accepted.
[21] A. Agarwal, B. Paul, S. Mukhopadhyay, and K. Roy, "Process variation in embedded memories: failure analysis and variation aware architecture," *IEEE J. of Solid-State Circuits*, vol. 40, no. 9, pp. 1804–1814, sept. 2005.
[22] B. Calhoun and A. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *in IEEE J. of Solid-State Circuits*, vol. 41, no. 7, pp. 1673–1679, july 2006.
[23] K.-S. Yeo and K. Roy, *Low-Voltage, Low-Power VLSI Subsystems*. McGraw-Hill, 2005.

**Pascal Meinerzhagen** (S'10) was born in Bern, Switzerland, in 1984. He received his B.Sc. and M.Sc. degrees in Electrical and Electronics Engineering from the Swiss Federal Institute of Technology in Lausanne (EPFL) in 2006 and 2008, respectively. In 2008, he also received a joint certificate in Micro- and Nanotechnologies for Integrated Systems from the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland, the Grenoble Institute of Technology (INPG), France, and the Politecnico di Torino, Italy.

In 2008, Mr. Meinerzhagen was a visiting researcher in Chancellor Steve Kang's group at the University of California, Merced, USA, where he worked on the development of a 12-bit low-power SAR A/D Converter for a Neurochip. From 2009 to 2010, Mr. Meinerzhagen was a PhD student in the Integrated Systems Laboratory (IIS) at the Swiss Federal Institute of Technology in Zurich (ETHZ), Switzerland. Currently, he is finishing his PhD dissertation in the Telecommunications Circuits Laboratory (TCL) at EPFL.

Mr. Meinerzhagen's current research interests include the design of high-density memory structures for fault-tolerant VLSI systems in deep-submicron CMOS technologies and memory arrays assembled from standard-cells for ultra-low-power sub-VT VLSI systems. Mr. Meinerzhagen received a nomination for the "2010 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS) student paper contest".

**S.M. Yasser Sherazi** (S'09) received the Bachelors degree in computer engineering from COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2005, and the Masters degree in system-on-chip from Linköping University, Linköping, Sweden, in 2008.

He is currently perusing the Ph.D. degree in digital ASIC from the EIT Department, Lund University, Lund, Sweden. After his B.Sc. he worked as a Research Associate in CIIT, Islamabad, Pakistan for two years. He then received a HEC scholarship for his Masters Studies in Sweden. After completing his M.Sc. from Sweden, he returned to Pakistan and worked as a Lecturer with CIIT Islamabad for a semester. He is funded by Swedish Foundation for Strategic Research (SSF) as a PhD student in Lund University, Sweden.

He is currently working on ultra-low power wireless devices project with the EIT Department, Lund University. His main responsibility in the project is to design ultra-low power base band digital circuits. Mr. Sherazi was a recipient of a bronze medal for his performance during his Bachelors studies.

**Andreas Burg** (S'97-M'05) was born in Munich, Germany, in 1975. He received his Dipl.-Ing. degree in 2000 from the Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland. He then joined the Integrated Systems Laboratory of ETH Zurich, from where he graduated with the Dr. sc. techn. degree in 2006.

In 1998, he worked at Siemens Semiconductors, San Jose, CA. During his doctoral studies, he was an intern with Bell Labs Wireless Research for a total of one year. From 2006 to 2007, he held positions as postdoctoral researcher at the Integrated Systems Laboratory and at the Communication Theory Group of the ETH Zurich. In 2007 he co-founded Celestrius, an ETH-spinoff in the field of MIMO wireless communication, where he was responsible for the ASIC development as Director for VLSI. In January 2009, he joined ETH Zurich as SNF Assistant Professor and as head of the Signal Processing Circuits and Systems group at the Integrated Systems Laboratory. Since January 2011, he is a Tenure Track Assistant Professor at the Ecole Polytechnique Federale de Lausanne (EPFL) where he is leading the Telecommunications Circuits Laboratory in the School of Engineering.

In 2000, Mr. Burg received the "Willi Studer Award" and the ETH Medal for his diploma and his diploma thesis, respectively. Mr. Burg was also awarded an ETH Medal for his Ph.D. dissertation in 2006. In 2008, he received a 4-years grant from the Swiss National Science Foundation (SNF) for an SNF Assistant Professorship. In his professional career, Mr. Burg was involved in the development of more than 25 ASICs. He is a member of the IEEE and of the European Association for Signal Processing (EURASIP).

**Joachim Neves Rodrigues** (S'00-M'05-SM'11) holds currently an assistant professorship at the Department of Electrical and Information Technology at Lund University, Lund, Sweden. He received his degree in electrical engineering and computer science from the University of Applied Sciences, Kaiserslautern, Germany, and the Ph.D. degree from the Department of Electroscience, Lund University, in 2000 and 2005, respectively.

From 2005 to 2008 he acted as ASIC process lead in the digital ASIC department at Ericsson Mobile Platforms, Lund, Sweden, and he re-joined his current department in 2008. He has contributed to 20 ASICs both in industry and academia.

His main research interest is modelling and implementation of digital and mixed-mode microelectronics, architectures for high performance ultra-low power design which may be operated with an aggressively scaled supply voltage, with a focus on biomedical circuits and systems. He is a TC member of the biomedical circuits and systems society since 2010, and board member of the Swedish SSC chapter.