Regular paper

Design and implementation of energy-efficient near-threshold standard cell library for IoT applications

AbdelRahman Hesham a,*, Amin Nassar a, Hassan Mostafa a,b

a Electronics and Communications Engineering Department, Cairo University, Egypt
b University of Science and Technology, Nanotechnology and Nanoelectronics Program, Zewail City of Science and Technology, Egypt

A R T I C L E   I N F O

Keywords:
- Near-threshold CMOS
- Ultra-low energy
- IoT
- Body-biasing
- CAESAR

A B S T R A C T

A CMOS standard cells library of low-energy, minimum-area, and fitted for IoT applications is introduced in this paper. The paper uses two solutions to provide significant energy saving. The first is to design the library to be operating in the Near-Threshold Voltage (NTV) region. The second is to create layouts of cells at the minimum possible area that can be achieved for a given technology process. To partially recover the speed loss due to operating in the NTV region, the pMOS performance is boosted by a proposed body biasing technique that connects pMOS body to the ground. Furthermore, minimum energy consumption is considered at the selection of the library supply voltage and the selection of each cell transistor sizing, while keeping the library performing in the range of 1 MHz up to 20 MHz. This range is sufficient for IoT applications. Another challenge for the NTV is Performance Sensitivity to the process variations, which is analyzed, then a design solution is provided to assure timing closure with such sensitivity. The UMC 130 nm CMOS process technology was used to design and characterize the proposed library. Library timing and physical views were created to enable its usage in both synthesis and physical design tools. Library benchmark was done on three cryptography algorithms to show the benefit for IoT applications. The used algorithms are AEGIS-128, ASCON, and AEZ. The maximum achieved frequency for these cores is 14 MHz, 18 MHz, and 16 MHz, and the corresponding energy consumption is 4.25 pJ, 10.03 pJ, and 30.57 pJ, respectively.

1. Introduction

The advancement in the Internet of Things (IoT) field is widely noticed. The number of IoT applications seen in daily life is increasing with the wide adoption of such applications in significant engineering fields, like the automotive industry, home automation, and wearable devices. Many of the recent wireless communication research is either driven towards or interleaving with the deployment of mature IoT applications. In addition to this, the used devices in the IoT have a new design paradigm, which is needed to align with the general requirement for IoT.

The three major aspects considered when designing an IoT device are (1) how many computations it can do, (2) how efficient it is in energy-saving, and (3) how secure it is against the possible threats.

There is significant research done to provide computation-powerful devices that can be used in IoT applications. The proposed processors in [1–3] have shown high throughput while considering the low energy consumption constraint. The energy reduction techniques proposed in the literature are covering all design levels. Out of these techniques, circuit-level ones are the most useful for IoT devices. On top of these techniques comes voltage scaling feasibility to enable a new paradigm that can provide significant energy saving. This new paradigm considers optimizing circuits towards the Minimum Energy Point (MEP), unlike the conventional paradigm of optimizing towards the Minimum Delay Point (MDP) [1]. The MEP requires the circuit to operate in the sub-threshold region, where the supply voltage is considerably below the threshold voltage. However, the cost of the achieved energy saving is the increased delay seen at MEP [2]. That significant performance loss created a need towards a balance point between energy saving and performance loss. So, the Near-Threshold Voltage (NTV) operation is proposed. By operating at a supply value near to the transistor threshold voltage value, a significant energy saving can still be noticed while having an improved performance [2]. As per [3], the NTV region shows a 10X degradation in performance with a reduced energy saving by 10X over the sub-threshold region. Moreover, another major challenge seen in the NTV operation is the increased impact of Process, Voltage, and Temperature (PVT) variations on the performance variation. [3] showed that about 20X increase is noticed in delay variation due to PVT variations at NTV supply. The application of restrictive fixed

* Corresponding author.
E-mail addresses: a.heshamm@gmail.com (A. Hesham), amin.nassar@yahoo.com (A. Nassar), hmostafa@uwaterloo.ca (H. Mostafa).

https://doi.org/10.1016/j.aeue.2021.153907
Received 28 February 2021; Accepted 22 July 2021
Available online 31 July 2021
1434-8411/© 2021 Elsevier GmbH. All rights reserved.
timing derates can help to mitigate this variation problem, but it will end up with too pessimistic designs that are operating at lower-than-possible frequencies. A proper derating approach is needed to meet this requirement.

Besides the energy and computational power considerations in IoT devices, there comes a major concern about their safety while getting adopted widely. The count of IoT devices was expected to reach about 50 billion devices by 2020 [4]. Several solutions were developed to provide the needed system security; on top of them comes Cryptography. Devices implementing cryptography algorithms can ensure the safety of data transfer in IoT networks and protect them against possible attacks. A significant research effort was made to provide system-level and architecture-level cryptography solutions. The European Network of Excellence in Cryptology (ECRYPT) held a competition to develop trusted authenticated ciphers. The competition was called CAESAR, a short for Competition for Authenticated Encryption: Security, Applicability, and Robustness [5]. In 2019, three algorithms were selected. Each of the three algorithms covers one of the major IoT concerns: lightweight, high performance, and defense-in-depth. Thus, the paper uses the three algorithms to validate and benchmark the proposed solutions.

Currently, the most used approach in digital design is the standard-cell-based one. This makes the starting point to address the IoT design challenges in digital systems is the standard cell library.

A design and implementation methodology of a low-energy and minimum-area standard cell library is proposed and implemented in this paper. The methodology was initially introduced in [6], which covers the pre-layout design challenges. By pre-layout we mean up till the synthesis step of the implementation flow. In this methodology, energy saving is achieved by operating the designed library in the NTV region at 350 mV supply, and the minimum area is achieved by minimizing each cell's layout.

This paper is completing the library design methodology by covering the placement and routing challenges, and the timing signoff challenges coming from the increased performance variation in the NTV region. The methodology is then used to implement a library in UMC 130 nm process technology. The comparison of the proposed library to literature/commercial libraries shows the gains in energy and area coming from using the proposed library. An average gain of 34.36% in cells’ area is shown in this paper while meeting the placement and routing requirements. The three algorithms from the CAESAR competition, namely: ASCON [7], AEGIS-128 [8], and AEZ [9], are used to benchmark the proposed library and to show the library’s Power, Performance, and Area (PPA) when used in a critical IoT application.

The rest of the paper’s structure comes as follows. Section 2 discusses the solutions used to achieve library PPA improvements. Section 3 describes the library design flow and how it is applied to the proposed library. Section 4 provides the library benchmark results using CAESAR algorithms. Lastly, the work done is concluded in Section 5.

2. Design of the library architecture

Standard cell libraries are widely used in digital design. All cells provided in the library come in the same height – or integer multiples of it – and in integer multiples of a certain width value [10,11]. The minimum height and width of a cell define a term called the “Unit Tile”. The unit tile specifies the minimum resolution a placement algorithm will follow to decide each cell placement. So, the design area will be viewed as rows of unit tiles or what is called “Placement Rows”. Once Unit Tile is defined for a library, a cell layout is created to include the cell devices with consideration of Design Rule Checks (DRC), pin access, and device power and performance.

2.1. Calculation of minimum cell height for the used process technology

The design of a minimum cell height is a function of process technology parameters. [6] introduced the governing equation to calculate the minimum cell height in terms of the process geometry parameters. Based on this equation, and for the used UMC 130 nm process, the minimum cell height expressed as the count of horizontal tracks is calculated to be 5T. This is the minimum number of tracks reported in the literature [6]. Prior to this, the minimum architecture in literature was 6T [12] at the same dimension process. Furthermore, the work in [13–16] developed bigger architectures of 8T–12T.

2.2. Design of cell layout

The placement of transistors inside a certain area can be minimized by utilizing Euler’s path theory [17–20]. There are many routing approaches that can be done to connect the resulting transistors’ placement. The most suitable one should minimize pins’ capacitance and allow for clear pin access when connected externally in placement and routing. With the small cell height described in the previous section, pin access becomes more challenging. Clear pin accessibility was achieved by keeping all signal pins on metal2 with at least two hit-points and aligning pins to metal2 tracks. A possible hit-point for a pin is where the pin can be connected to an external shape on the same layer or connected from upper or lower layers through a via without resulting in DRC violation. Fig. 1 shows a situation where the pin has two hit-points.

To meet routability requirements with the shrunked area, the internal routing between transistors required three metal layers. It is worth mentioning that metal3 is used in few cases and should not affect its usability for signal routing.

2.3. Comparison of the library achieved area against foundry commercial library

The standard cells’ area achieved with the minimum area technique, while keeping clear pin access, is provided in Table 1. For each cell, the table shows the corresponding area in the foundry library. From the table, it is shown that the proposed library is achieving 34.36% area shrinkage. It is worth mentioning that the area reduction is achievable regardless of the region of operation.

![Fig. 1. Accessibility of cell pin.](image-url)
2.4. Transistor sizing calculation

Transistor sizing is an important factor in determining the speed and power behavior of the proposed library. [6] defined a maximum limit for transistor sizing as a function of process geometry parameters. For the selected 5T architecture and in UMC 130 nm, this limit is calculated to be 1.04 μm, which allows for transistor sizing around the smallest pMOS and nMOS widths.

Based on the analysis done in [6], the sizing of PMOS (Wp) needs to be >350 nm to avoid the peak of the threshold voltage and achieve higher performance, but it cannot be largely increased to avoid input capacitance increase, which will hurt overall performance. For NMOS (Wn), it needs to be at the minimum value to benefit from the Inverse-Narrow-Width Effect (INWE).

2.5. pMOS performance boost by a proposed body biasing

In [6], a body biasing technique was proposed to minimize the difference between the NMOS and PMOS currents. This minimization helps to reduce the cell delay without changing PMOS transistor sizing. This assures that the gains achieved by the transistor sizing described in the previous section are not opposed. The technique relies on the pMOS forward body biasing to provide significant performance gain for the pull-up network while keeping the pull-down one at the same power consumption. The challenge for such techniques is that the body biasing from the conventional biasing is the uncertainty of performance sensitivity to PVT variations. However, the proposed body biasing helps to reduce the increased sensitivity coming from operating in the NTV region, which will be shown later in this paper.

2.6. Library architecture testing and supply voltage selection

The minimum Wp of 160 nm is used to design the inverter INV_1X, while connecting the p-substrate terminal to the ground. The simulation is done on a Fanout-Of-4 (FO4) testbench for the inverter cell. Figs. 2 and 3 show that the proposed body biasing provides better performance with a slight increase in power in the sub-threshold and the near-threshold regions compared to the conventional body biasing. This is not the case when the supply voltage increases, as power starts to grow faster than the delay reduction. The supply voltage is selected to be 350 mV, where the performance gain is achieved at the cost of power increase while achieving the same PDP as of the conventional body biasing. At the 350 mV supply, INV_1X is showing 276ps delay reduction. So, it can operate at 4.2 GHz rather than 3.6 GHz, which is an improvement of 16.6%. The selection of this supply in the middle of the flat range in the PDP curve will help with less sensitivity for supply variations while maintaining the same PDP value.

2.7. Library characterisation

To characterize the timing and power of each cell, a FO4 circuit is created. Both High-To-Low and Low-To-High delays were measured. If the cell is a multi-input one, the delay is calculated for each input while fixing the rest. Similarly, power consumption was calculated for each cell. Also, to show the benefit of selecting 350 mV as supply voltage, the library was characterized at two other supply voltages: 300 mV and 400 mV. Table 2 shows the average cell delay and average power consumption for each cell in the FO4 circuit operating at the different voltage supplies. Then, the Power-Delay-Product was calculated to indicate the overall energy consumption. The characterization was done at the tthp35v25c and using the nominal parasitics.

From the table, it can be shown that increasing the supply voltage to 400 mV can achieve reduced delay by 0.46X but at the cost of 2.71X increase in power consumption. On the other hand, reducing the supply voltage to 300 mV can reduce the power consumption by 0.56X, but at the cost of increased delay by 3.31X. PDP indicates that the maximum energy saving can be achieved by operating the library at 350 mV compared to 400 mV and 300 mV by 25% and 79%, respectively.

2.8. Analysis and handling of performance variation

In digital circuits, the body effect is canceled by tying the n-well terminal to the supply and tying the p-substrate terminal to the ground. However, it can be observed that the body biasing helps to reduce the increased sensitivity coming from operating in the NTV region, which will be shown later in this paper.

<table>
<thead>
<tr>
<th>Cell</th>
<th>Proposed lib area (μm²)</th>
<th>Ref Lib Area (μm²)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AN2_1X</td>
<td>4.16</td>
<td>6.4</td>
<td>35.00</td>
</tr>
<tr>
<td>AO2_1X</td>
<td>5.44</td>
<td>8.96</td>
<td>39.29</td>
</tr>
<tr>
<td>DINV</td>
<td>11.36</td>
<td>23.36</td>
<td>51.97</td>
</tr>
<tr>
<td>D_LATCH_L</td>
<td>1.12</td>
<td>2.84</td>
<td>39.69</td>
</tr>
<tr>
<td>FILLER1X</td>
<td>1.04</td>
<td>1.28</td>
<td>18.75</td>
</tr>
<tr>
<td>FILLER2X</td>
<td>2.08</td>
<td>2.56</td>
<td>18.75</td>
</tr>
<tr>
<td>INV_1X</td>
<td>2.08</td>
<td>3.84</td>
<td>45.83</td>
</tr>
<tr>
<td>MX2_1X</td>
<td>9.6</td>
<td>11.52</td>
<td>16.67</td>
</tr>
<tr>
<td>ND2_1X</td>
<td>3.12</td>
<td>5.12</td>
<td>39.06</td>
</tr>
<tr>
<td>ND3_1X</td>
<td>4.16</td>
<td>7.68</td>
<td>45.83</td>
</tr>
<tr>
<td>NR2_1X</td>
<td>3.12</td>
<td>5.12</td>
<td>39.06</td>
</tr>
<tr>
<td>NR3_1X</td>
<td>4.16</td>
<td>7.68</td>
<td>45.83</td>
</tr>
<tr>
<td>OINV_1X</td>
<td>5.44</td>
<td>7.68</td>
<td>39.06</td>
</tr>
<tr>
<td>OR2_1X</td>
<td>1.12</td>
<td>6.4</td>
<td>35.00</td>
</tr>
<tr>
<td>TAP</td>
<td>2.08</td>
<td>2.56</td>
<td>18.75</td>
</tr>
<tr>
<td>TIEHI</td>
<td>2.08</td>
<td>3.84</td>
<td>45.83</td>
</tr>
<tr>
<td>TIELO</td>
<td>2.08</td>
<td>3.84</td>
<td>45.83</td>
</tr>
<tr>
<td>XOR2_1X</td>
<td>2.08</td>
<td>14.08</td>
<td>31.82</td>
</tr>
<tr>
<td>XOR2_1X</td>
<td>11.56</td>
<td>14.08</td>
<td>31.82</td>
</tr>
<tr>
<td>Average</td>
<td>–</td>
<td>–</td>
<td>34.36</td>
</tr>
</tbody>
</table>

Fig. 2. INV_1X power and delay against the supply voltage.

Fig. 3. INV_1X PDP against the supply voltage.
With this conventional biasing, threshold voltage dependence on the body voltage is negligible. Once the body connections are altered from this conventional connection, the threshold voltage starts to show increased sensitivity to PVT variations, and hence increasing performance sensitivity. The situation becomes more challenging knowing that the NTV operation itself is also increasing performance sensitivity to PVT variations. This was discussed in [1–3].

The PVT variations can be categorized into two types: systematic variations, known as global variations, and random variations, known as local variations. The conventional design flow has different approaches to handle both variation types [21]. For the global variations, the timing closure needs to be achieved at different PVT corners. This makes the design meeting timing requirements at any point in the PVT space. Based on this, the proposed library needs to be characterized for the most timing-critical PVT corners. For the local variations, their impact on delay is modeled by applying an On-Chip Variation (OCV) derate for each cell. In super-threshold operation, OCV can be represented as a fixed value applied to all cells at all PVT corners. With the increased delay variation, applying a fixed OCV derate will end up with over-design pessimism. So, the Advanced OCV (AOCV) approach is introduced. It models the variation for each cell at each corner while including the impact of logic path length on reducing delay variability. Another timing design methodology is proposed in [22] to address the delay variation while reducing the computational effort to model it.

To analyze the impact of both types on the delay of the proposed library, Monte Carlo simulations are conducted on the INV_1X FO4 circuit. The results are shown in Figs. 4, 5 and 6. From the graphs, the highest delay mean value is seen at the SS315v40c corner. While the lowest delay mean value is seen at FF385v125c. So, in addition to TT350v25c, the three corners are used to characterize the library to cover the impact of the global variations. Also from the graphs, the proposed body biasing is proved to provide...
Fig. 5. Delay mean and sigma values over TT corners for proposed vs. conventional body biasing.

Fig. 6. Delay mean and sigma values over FF corners for proposed vs. conventional body biasing.
Then, the simulated derate value needed to address local process variations, calculated as delay Sigma/Mean, are shown in Figs. 7, 8 and 9.

From the graph, it is clear that a fixed derate value cannot be applied to all corners, as the range of derate variation is too wide, ranging from 0.29 up to 1.02 for 3Sigma. This confirms the need for AOCV modeling for the proposed library.

For our proposed library, the AOCV is modeled by creating Liberty Variation Format (LVF) beside the .lib timing files. The LVF is the most accurate method of specifying the OCV [23].

The delay variations described previously are describing process local variation. Voltage and temperature local variations are more correlated and are minor when compared to process local variation. For example, as the local voltage variation are coming from the IR drop seen in the power grid, it can be minimized by designing the power grid to achieve a maximum IR drop of 2% of the supply.

2.9. Placement and routing of the library

To have the library ready for placement and routing tools, some physical cells needed to be added to the library. These cells are (1) Tap...
cell, which is responsible for providing the body connection needed for our body-biasing technique, (2) Tie cells, needed to connect input pins to constant values, and (3) Filler cells, to preserve the continuity of supply/ground connections and continuity of base layers. Fig. 10 shows the tap cell connecting both n-well and p-substrate to the ground. With the placement of tap cells as in Fig. 11, there is no need to add the body connections inside each cell and reduce the cell area.

Then, a set of abstract views describing the library were created. A Liberty file (.lib) was generated to include all the cell’s timing and power views. The file contains look-up tables providing each cell’s delay and power based on its driver and load. To have this table covering all the possible situations, the driver transition time is assumed to be in the range between 0.1X of the minimum output transition seen in the library and 10X of the maximum output transition seen. The lower limit of 0.1X is unlikely to happen as long as the design uses cells from the same library. The capacitive load range was selected between 0.1 of the minimum input capacitance for all cells and 10 of the maximum capacitance load. To provide abstract information of cell pins and internal metal locations, a LEF (Library Exchange Format) file is generated. The file is needed for routing cell pins and avoiding internal metal shapes (short). In addition to these two files, a technology file describing cell architecture was created. The file describes the used metal layers pitches and defines the library’s unit tile. This is in addition to describing each metal layer DRC rules and via definitions. With all these files, the FO4 test design is taken to Placement and Routing using ICC (IC Compiler). A placement and routing flow is developed to properly use the library and minimize the power while keeping the timing performance as required. Fig. 12, Fig. 13, and Fig. 14 show the complete placed and routed design, the signal and clock routes connecting all cells, and the placement of cells, respectively.

2.10. Literature comparison

A reference library is designed to operate in the NTV at the 130 nm process. The library has a 6T architecture, and a supply of 400 mV. To have a fair comparison, the proposed library is re-characterized to operate at 400 mV. The corner used for this comparison is TT400v25c. As shown in Table 3, the proposed library is achieving a less delay by about 1.5X and a less PDP by about 106X for the 3 basic cells. This shows that the improvements are achieved by the usage of the proposed design methodology.

3. Cell design flow

The cell design flow used is described in [6]. The cells available in this library are INV, AND, OR, MUX, NAND, NOR, XOR, XNOR, D-Flipflop, and latch. This guarantees the coverage of the basic combinational and sequential functions. Literature was explored for logic families that are used for IoT applications. Adiabatic logic is discussed in [24,25] which shows improved security over CMOS logic. The static CMOS is selected for the library proposed in this paper to guarantee Rail-to-Rail swing at the low NTV supply and assures the compatibility with conventional design flow. Also, to include the variation models, Synopsys Siliconsmart tool is used to create the library LVF files.

4. Library benchmarking

In this section, comparison results are provided to show the power and performance gains of the proposed library. Table 4 compares Fmax of the proposed library against foundry commercial library and shows

![Fig. 11. Tap cell placement.](image1)

![Fig. 12. All-Cells FO4 design: Physical Layout.](image2)
the frequency loss ratio between them at the typical corner. Then, the impact of the global PVT variations on power and frequency is shown in Table 5 for the AEZ design implemented with the proposed library. As discussed previously, the worst performance corner is confirmed to be SS0p315vm40c, which had an about 16X reduction in performance over the typical one. On the other hand, the worst power consumption corner is FF0p385v125c, with an about 10X increase compared to the typical one.

Then, the implementation of three cryptographic cores is done completely till layout generation. In the final portfolio of the CAESAR competition [26], ASCON was selected as the primary choice for lightweight authenticated encryption, and AEGIS-128 was selected as the primary choice for high-performance authenticated encryption. AEZ is another algorithm that was one of the finalists in the same competition. The used RTL implementations are the ones used for CAESAR candidates benchmarking [27]. The library’s post-layout assessment was done by taking the three cores through the full digital design flow. Table 6 is showing the results of this assessment.

5. Conclusion

The paper’s designed Near-Threshold standard cell library shows significant energy saving when used in essential applications in IoT. Frequency reduction is the penalty of the achieved energy saving. The paper has provided an integrated design solution to find an optimal Power-Performance-Area operating point based on two major components. The first component is a technology-dependent methodology for minimum standard cell layout architecture design. The second component is utilizing INWE and a novel body biasing technique that boosts performance in NTV and reduces performance sensitivity to PVT variations. Three of the latest and best cryptography cores are used to benchmark the proposed library. Post-layout results are showing a significant gain in energy saving. The quality of improvement can be measured as the ratio of energy improvement and frequency reduction. AEGIS-128 achieves a ratio of 2.5, AEZ achieves a ratio of 4.1, and ASCON achieves a ratio of 1.7. The achieved frequencies are sufficient for IoT applications and can be highly accepted given the significant achieved energy saving.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is an extension of the published paper [6] in NILES 2020 Conference. Also, this work was funded partially by ONE Lab at Zewail City of Science and Technology, Egypt and at Cairo University, Egypt, NTRA, Egypt, Information Technology Industry Development Agency, Egypt, ASRT, Egypt, and NSERC, Canada.
Table 6
Post-layout comparison of power and energy at typical corner.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>AEGIS-128 @ 14 MHz Proposed</th>
<th>Lib2</th>
<th>AEGIS @ 16 MHz Proposed</th>
<th>Lib2</th>
<th>ASCON @ 18 MHz Proposed</th>
<th>Lib2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Power (mW)</td>
<td>0.022</td>
<td>0.3179</td>
<td>0.165</td>
<td>3.060</td>
<td>0.042</td>
<td>0.862</td>
</tr>
<tr>
<td>Switching Power (mW)</td>
<td>0.036</td>
<td>0.2788</td>
<td>0.172</td>
<td>1.770</td>
<td>0.084</td>
<td>0.924</td>
</tr>
<tr>
<td>Leakage Power (mW)</td>
<td>0.002</td>
<td>0.0001</td>
<td>0.091</td>
<td>0.001</td>
<td>0.015</td>
<td>0.001</td>
</tr>
<tr>
<td>Total Power (mW)</td>
<td>0.059</td>
<td>0.5968</td>
<td>0.428</td>
<td>4.831</td>
<td>0.140</td>
<td>1.787</td>
</tr>
<tr>
<td>Internal Energy (pJ)</td>
<td>4.25</td>
<td>42.63</td>
<td>30.57</td>
<td>345.08</td>
<td>10.03</td>
<td>127.64</td>
</tr>
</tbody>
</table>

Ratio of Energy Gain

12.73 10.04 11.29

*Foundry Commercial Library.

References