Title: Circuit Techniques for Device-Circuit Interaction toward Minimum Energy Operation

Author(s): Islam, A.K.M. Mahfuzul; Onodera, Hidetoshi

Citation: IPSJ Transactions on System LSI Design Methodology (2019), 12: 2-12

Issue Date: 2019-02-01

URL: http://hdl.handle.net/2433/243233

Type: Journal Article

Textversion: publisher

Kyoto University
Abstract: In the quest of operating a circuit at its minimum energy, both of the design-time and run-time optimizations are being employed aggressively in mobile devices today. This paper overviews some of the energy reduction techniques and highlights the role of device–circuit interaction in achieving minimum energy operation. We show that run-time energy minimization can be achieved by keeping the ratio between static energy and dynamic energy to an optimum value. Focusing on the ease-of-design, we then review several delay-based sensor circuits to monitor transistor performance, leakage-current and temperature. We also discuss activity monitors to estimate the dynamic energy. We then discuss a cell-based design approach for both of the sensing and the back-gate voltage tuning circuits for area and cost reduction. Using the sensing and the voltage tuning circuits, the system can operate at its minimum energy all the time.

Keywords: CMOS, minimum energy operation, DVFS, back-gate biasing, device–circuit interaction, on-chip monitor, on-chip sensor

1. Introduction

With the emergence of the internet of things (IoT), the design paradigm has shifted from performance-oriented design to energy-oriented design. Designers are aggressively employing low-energy techniques starting from macro-architecture to device optimization to squeeze out the last bite of energy saving. In addition to the design-time optimization, several run-time energy minimization techniques can be employed. Run-time techniques include dynamic voltage and frequency scaling (DVFS) [1], adaptive body biasing [2], power-gating [3], clock-gating [4] and so on. In this paper, we show that an additional layer of run-time energy optimization can be achieved by incorporating device–circuit interaction into the system. After the manufacture of the chip, the circuit performance is determined by the device characteristics; that is devices control the circuits. Under device–circuit interaction, the circuit also controls the devices thus forming a closed loop between device and circuit. Device–circuit interaction is a powerful tool and can be applied as not only a complementary to the many run-time techniques, but also as a new way of optimization.

The diverse nature of applications and use-case scenarios impose a need on run-time techniques to minimize energy consumption. For example, it is well known that the optimum values of supply voltage \( V_{dd} \) and threshold voltage \( V_{th} \) differ depending on the operating frequency and activity rate [5]. When a circuit operates under a low activity rate, that is the inputs are rarely altered over time, the circuit becomes leakage-dominant and as such increasing both of the \( V_{th} \) and \( V_{dd} \) is beneficial. Conversely, when the activity rate of the circuit is high, reducing both of the \( V_{th} \) and \( V_{dd} \) is beneficial. Tuning of \( V_{th} \) is realized by tuning the back-gate voltage \( V_{bg} \) of the transistors. Thus, after optimizing the design from macro-architecture down to circuit topology, an additional layer of optimization can be achieved by realizing a run-time interaction between device and circuit.

To realize a run-time interaction between device and circuit, two key circuits are required. One is sensing circuits to monitor device characteristics and system activity. The other is tuning circuits to change back-gate and supply voltages. As the sensing and tuning circuits are always-ON circuits, their own power consumptions add to the overall power consumption. Furthermore, these circuits need to be robust and tested fully for their wide usage. Ease-of-design is another challenge that needs to be addressed. In this paper, we focus on the ease-of-design and show that a cell-based design methodology that has been adopted for digital circuits for a long time can be applied to the sensing and tuning circuits. Adaptation to cell-based design methodology can reduce design time significantly and thus reduce cost. The cell-based design also enables to put the sensing circuit close to the target circuit so that better correlation between the sensor and target circuit can be achieved. To facilitate cell-based design, delay-based sensing circuits are explored in the literature which will be discussed here. We also discuss a cell-based design for an analog circuit such as a voltage tuning circuit. Finally, we propose a simple run-time minimum operation strategy that involves keeping the ratio between static and dynamic energy to an optimum value.
The remainder of the paper is organized as follows. Section 2 overviews the energy consumption mechanism of a CMOS circuit and several energy optimization techniques. We show the contours of frequency and energy on the $V_{th}$-$V_{dd}$ plane and argue that run-time tuning of $V_{th}$ and $V_{dd}$ is effective in energy reduction. In Section 3, we show that device-circuit interaction enables optimum operation of the circuits. We, then, discuss several sensing and tuning techniques focusing on ease-of-design in Section 4. Finally, we conclude our remarks in Section 5.

2. Minimum Energy Operation

In this section, we first review the mechanism of energy consumption in an LSI. We then look into the different factors that affect the energy and how optimizations from architecture down to technology are performed. We then discuss the effectiveness of run-time optimization to energy reduction and raise the need for device–circuit interaction.

2.1 Energy Consumption in LSI

We review energy and delay models based on a delay chain as shown in Fig. 1. The delay and energy consumption of the delay path depend on the logic depth $L_{\text{depth}}$, clock frequency $f_{\text{clk}}$, and activity rate $a$. The total energy consumption of a circuit can be divided into a dynamic component and a static component as follows.

$$E_{\text{total}} = E_{\text{dynamic}} + E_{\text{static}}.$$  \hfill (1)

The dynamic energy results from the charging and discharging of the nodal capacitances during the switching and is expressed as follows.

$$E_{\text{dynamic}} = aC_{\text{load}}V_{dd}^2L_{\text{depth}}.$$  \hfill (2)

Here, $C_{\text{load}}$ is the load capacitance, $a$ is the activity rate, and $V_{dd}$ is the supply voltage. The static energy component results from the throughput leakage current from $V_{dd}$ to ground by transistor leakage-current and is expressed as follows.

$$E_{\text{static}} = I_{\text{leak}}V_{dd}T_{\text{clk}}L_{\text{depth}}.$$  \hfill (3)

Here, $T_{\text{clk}}$ is the clock period and $I_{\text{leak}}$ is the transistor leakage current which is expressed as follows.

$$I_{\text{leak}} = I_{0}e^{\frac{V_{dd}}{V_{th}}}.$$  \hfill (4)

Here, $I_0$ is the normalized leakage current which is dependent on the process. $n$ is the sub-threshold swing coefficient and $V_T$ is the thermal voltage. Note that DIBL effect is ignored here for simplicity to focus on the key mechanisms of energy consumption. Finally, the total energy can be broken down as follows.

$$E_{\text{total}} = L_{depth} \cdot V_{dd} \cdot \left( aC_{\text{load}}V_{dd} + T_{\text{clk}}I_{0}e^{\frac{V_{dd}}{V_{th}}} \right).$$  \hfill (5)

Equations (2) to (5) reveal some important insights into the effects of the device and circuit parameters on the energy consumption.

(1) Only $E_{\text{dynamic}}$ depends on the activity rate, $a$, whereas only $E_{\text{static}}$ depends on the clock period, $T_{\text{clk}}$.

(2) Temperature increase the static energy exponentially.

(3) The first term in the closed bracket of Eq. (5) linearly depends on $V_{dd}$, whereas the second term exponentially depends on $V_{th}$.

Particularly important is the fact of property (3). The linear and exponential dependence to $V_{dd}$ and $V_{th}$ results in an optimum set of values under a fixed activity rate, clock frequency and temperature. Property (3) explicitly reveals the need for run-time tuning of both of $V_{dd}$ and $V_{th}$ to achieve minimum energy operation under the variations in process ($V_{th}$), clock frequency ($f_{\text{clk}} = 1/T_{\text{clk}}$), temperature ($T$) and activity rate ($a$).

Next, we define $T_{\text{clk}}$ as the minimum delay achievable for a delay path.

$$T_{\text{clk}} = \sum_{i=1}^{L_{\text{depth}}} \tau_{i}.$$  \hfill (6)

Here, $\tau_i$ is the delay of the $i$-th stage in the path and can be modeled by Eq. (7).

$$\tau_{i} = \frac{C_{\text{load}}V_{dd}}{I_{d}}.$$  \hfill (7)

Here, $C_{\text{load}}$ is the load capacitance, $V_{dd}$ is the supply voltage, and $I_{d}$ is the transistor ON current. To model the transistor ON current, we use a modified EKV model of Eq. (8) that is continuous from weak-inversion to strong-inversion operation [6], [7].

$$I_{d} = k \cdot \frac{W}{L} \cdot \ln^n \left[ 1 + \exp \left( \frac{V_{dd} - V_{th}}{a \cdot n \cdot V_{T}} \right) \right].$$  \hfill (8)

Here, $k$ is a technology-related parameter. The validity of Eq. (8) is shown in Fig. 2. Transistor ON-current across $V_{dd}$ is shown in the figure for the three models of the $\alpha$-power law model [8], the model of Eq. (8) and the exponential model for weak-inversion [9]. The $\alpha$-power law model becomes inaccurate when $V_{dd}$ approaches $V_{th}$. Consequently, the exponential model becomes inaccurate when $V_{dd}$ crosses $V_{th}$. The model of Eq. (8) is able to represent all the regions of operations.

In this paper, we use Eqs. (2) to (8) to numerically evaluate the...
energy and delay of the model circuit of Fig. 1. The models are fitted to a commercial 65 nm process. $L_{\text{depth}} = 40$ is used. Based on the simulation results, we derive some useful insights that help us understand the conditions for minimum energy operation. We do not consider the energy loss via short-circuit current during the switching in our analysis. The short-circuit current is a small fraction (< 10%) of current to the load [10], [11]. Thus, omitting the short-circuit energy loss does not have a strong impact on the operating conditions for minimum energy operation.

2.2 Minimum Energy Operation

2.2.1 Without Delay-constraint

When circuit delay is not a concern, the minimum energy operation can be realized by reducing the supply voltage to a point where the energy is minimum. The minimum energy point is dependent on several parameters such as activity, temperature and threshold voltage. Figure 3 shows the energy per cycle and delay of the 40-stage inverter chain against the supply voltage change. Activity rate of 0.1 is assumed. The values of energy and delay are normalized by the values at 1.0 V operation. Decreasing $V_{dd}$ from 1.0 V reduces the total energy consumption with the expense of delay increase. At around 0.3 V, the energy consumption reaches the minimum value and then increases again. Thus, lowering $V_{dd}$ is most effective to minimize energy when the delay is not constraint. However, in real applications, most designs are delay constraint and therefore energy minimization by $V_{dd}$ and $V_{th}$ scaling needs to be realized while ensuring that the delay meets the target speed.

2.2.2 With Delay-constraint

The optimum values of $V_{dd}$ and $V_{th}$ vary depending on the delay constraint. To illustrate the change in optimum values, contour plots of energy and frequency of the 40-stage inverter path is shown in Fig. 4. Contour curves are plotted on a $V_{th}$–$V_{dd}$ plane. Broken lines show frequency contour curves. Black solid lines show energy contour curves. Cross points show the minimum energy point for each clock frequency. We see that for a fixed clock frequency, there are numerous combinations of $V_{th}$ and $V_{dd}$ values that meet the target frequency. The optimum values of $V_{th}$ and $V_{dd}$ change depending on the clock frequency. For example, at 10 MHz of clock frequency, the optimum $V_{th}$ and $V_{dd}$ values are 0.42 V and 0.38 V, respectively. Then, at an increased clock frequency of 400 MHz, the optimum values become 0.28 V and 0.60 V. The optimum values also change depending on the activity rate and temperature. As activity rate changes largely in many circuits, the ability to tune $V_{th}$ and $V_{dd}$ values to their optimum values can reduce significant amount of energy. A similar locus of the minimum-energy-point is obtained by a transistor-level simulation in Ref. [12], which validates our model-based analysis.

2.3 Variability Effect on Energy

In scaled processes, process variation has become a significant issue where the design needs to be tuned to cope with all the possible cases of variation. Figure 5 shows a typical process corner window enclosed by corners labeled as “FF (Fast nMOSFET, Fast pMOSFET)”, “SF”, “SS” and “FS”. Besides process variation, supply voltage and temperature also vary and thus additional corners are also established for accounting those variations. As a result of these variations, designers need to check whether the circuit is functional even at the worst possible scenario. This results in setting $V_{dd}$ to much higher values than the optimal values. Higher $V_{dd}$ causes energy loss. The energy loss in a “worst-case design” is illustrated in Fig. 6 which shows the simulated energy consumption per clock cycle for two design cases against the clock frequency. One is the “worst-case design.” The other is the “typical-case design.” To incorporate the corner conditions in the simulation, foundry provided transistor models for a commercial 65 nm bulk process are used here instead of the model equations of Eqs. (2) to (8). In the “worst-case design”, $V_{dd}$ values have to be set such that chips from all the corners meet the timing. Thus, additional loss in dynamic power becomes as follows.
The energy values of corner chips under the “worst-case design” also find that energy loss occurs even when the chip is located at different process corners show significant energy loss compared with the “typical-case design.” We also find that energy loss occurs even when the chip is located at “TT.” The energy values of corner chips under the “worst-case design” are not the same because of the differences in the static energy. The “ Typical-case design” with the help of self-adaptation can lower the energy consumption by more than half because $V_{dd}$ can be set to its optimal values.

\[
\frac{\Delta E_{\text{dynamic}}}{E_{\text{dynamic}}} = \left(1 + \frac{\Delta V_{\text{dd}}}{V_{\text{dd}}}\right)^2 - 1. \tag{9}
\]

In Fig. 6, we find that chips at different process corners show significant energy loss compared with the “typical-case design.”

### 2.4 Run-time Energy Reduction Techniques

#### 2.4.1 Standby Energy Reduction

In many applications, a module may be in idle state for some time and then be activated again. In these scenarios, applying a clock gating technique that turns off the clock for the idle modules can save dynamic energy drastically. 66.6% of power reduction has been reported for an MPEG-4 decoder by employing clock gating [4]. After employing clock-gating, leakage power becomes the dominant source of energy consumption. To reduce the leakage power, power gating technique that disconnects the module from the supply by introducing footer or header transistors can be employed [3]. Introduction of footer or header transistors also impacts performance thus careful design is required.

#### 2.4.2 Dynamic Voltage and Frequency Scaling (DVFS)

As we can see in Fig. 3, performance and energy are in a trade-off relationship. The most effective way to save energy is to find the lowest possible performance for the circuit that satisfies a given timing constraint so that supply voltage can be reduced accordingly. For a microprocessor, the workload varies over time meaning that some time a lower performance is enough to meet the job deadline, whereas sometimes a maximum performance may be required. Thus, having a scheduler to vary the system performance according to the workload, and scale the supply voltage accordingly can save a lot of energy. The concept, called dynamic voltage and frequency scaling (DVFS) is now being widely used in many mobile devices. **Figure 7** shows a general framework of a DVFS system [13]. A software which can be the operating system (OS) runs on the hardware. The hardware here includes a microprocessor chip, DC/DC converter, and phase-locked loops (PLL) for frequency synthesis. The DC/DC converter and the PLL can be integrated onto the same microprocessor chip or they can be implemented separately onto a printed circuit board (PCB). An additional component is added in the DVFS system that is a controller to translate the required speed into the microprocessor performance. The controller can be implemented in software or in hardware. In any case, additional hardware is required to decode the requirement and send signals to the DC/DC converter and PLL. A hardware-based implementation can also be on-chip. In a traditional implementation, a look-up table (LUT) can be used to translate $f_{\text{clk}}$ into the corresponding $V_{\text{dd}}$ [14]. Without the use of sensors and run-time tuning of $V_{\text{dd}}$ and $V_{\text{n}}$, pairs of $V_{\text{dd}}$ and $f_{\text{clk}}$ values are set considering the worst-possible variability.

LUT based pairing between $V_{\text{dd}}$ and $f_{\text{clk}}$ is too pessimistic, and thus different variants of DVFS have been proposed. In one variant, $f_{\text{clk}}$ is set first, and then $V_{\text{dd}}$ is tuned such that the circuit delay meets the target speed [14]. Critical path monitors are used to tracking circuit delay in this case. As we have discussed earlier, scaling of $V_{\text{dd}}$ and $f_{\text{clk}}$ only is not enough to ensure minimum energy operation. Activity rate and temperature play important roles in the overall energy. Under the variations in activity rate and temperature, additional tuning of $V_{\text{n}}$ is necessary for the minimum energy operation. Fortunately, there is a way to tune the $V_{\text{n}}$ values by means of back-gate biasing which will be explained next.

#### 2.4.3 Back-gate Biasing

The ability to tune the $V_{\text{n}}$ values is a great tool for designers. Post-silicon tuning of $V_{\text{n}}$ by back-gate voltage has been explored a lot in the literature. Back-gate voltage tuning can be applied in many different ways. In a conventional bulk process, a well itself acts as the body of a transistor. Thus, the term “body-biasing” is being used to refer that the well voltage is tuned to forward- or reverse-bias of the source–p–n junction. Forward-biasing lowers $V_{\text{n}}$ and reverse-biasing raises $V_{\text{n}}$. As the p–n junction is directly biased here, applying forward-bias also causes an exponential increase in the junction leakage current. As junction leakage increases also, a designer has to be careful in applying body-bias for a bulk process. In processes such as fully-depleted-silicon-on-insulator (FDSOI) process [15], a well is separated from a transistor by a thin insulator layer. A well here acts as a back-gate for transistors and thus the term “back-gate biasing” is more relevant compared to the term “body-biasing”. In this case, there is no p–n junction at the source and the drain. Thus, a wider range of

---

**Fig. 6** Simulated energy per clock cycle for a “worst-case” and a “typical-case” design. Foundry provided transistor models are a 65 nm bulk process are used for the simulation. Activity rate of 0.1 is assumed.

\[
\Delta E_{\text{dynamic}} = \left(1 + \frac{\Delta V_{\text{dd}}}{V_{\text{dd}}}\right)^2 - 1.
\]

**Fig. 7** Block diagram of a widely used DVFS system.
forward and reverse back-gate biasing can be applied compared with a bulk process. The maximum values of forward and reverse bias, in this case, is limited by the p–n junctions between wells. To enable large forward back-gate biasing of ±1.3 V, a flip-well technology has been introduced [16]. A wider range of biasing gives us the increased tunability of the $V_{th}$ values.

Back-gate biasing can be applied statically or dynamically. In Ref. [2], a static value of back-gate biasing is applied to each processor chip after the manufacturing such that the chip meets the target operating frequency. For example, slower chips are applied forward back-gate biasing and faster chips are applied reverse back-gate biasing. Back-gate biasing can be applied during runtime for self-adjustment of $V_{th}$ [17]. Back-gate biasing can also be applied dynamically to reduce standby leakage power [18].

2.4.4 Combining DVFS and Back-gate Biasing

Combining the DVFS and back-gate biasing comes naturally. Tuning of $V_{th}$ compliments the DVFS system and therefore many flavors of combination are explored [19], [20], [21].

3. Device–Circuit Interaction for Energy Minimization

3.1 Device–Circuit Interaction by Self-Adaptation

Device–circuit interaction refers to the sensing and tuning of device characteristics using on-chip circuits. This realizes a closed loop where the circuit tunes the device characteristics, and in consequence, the circuit performances are changed.

Device–circuit interaction realizes a self-awareness into chip. In other words, a chip becomes aware of its performance variability and adapt itself to a desired performance by tuning device characteristics. The capability of self-adaptation to desired values can be exploited in many ways. First, the process window as explained in Section 2 can be shrunk significantly. This enables a near typical-case design without the need of large design margins. Second, depending on the circuit parameters such as activity rate, clock frequency, and supply voltage, the optimum operating point can be realized during run-time to minimize the energy consumption.

3.2 Demonstration of Self-Adaptation

As explained in Section 2, the worst-case design is too expensive in terms of energy. An ideal case would be to have the $V_{th}$ values set at a point where the design is optimized. After the manufacture of the chip, self-adaptation can ensure the optimum $V_{th}$ values. The concept is illustrated in Fig. 8. In the left-hand side, each point in the plot corresponds to the $V_{thp}$ and $V_{thn}$ values from each chip. Note that within-die random variability is not considered here. Each point here represents the average value for a particular chip. With self-adaptation functioning, the $V_{th}$ values for all chips are tuned such that the optimum values are realized. Thus, a distribution of $V_{th}$ values shown in the right-hand side will be obtained.

A proof-of-concept for the above concept is experimentally demonstrated in Ref. [22]. Here, the target values of $V_{thp}$ and $V_{thn}$ are set to the typical values provided by the foundry. Figure 9 shows the self-adaptation scheme proposed in Ref. [22]. The chip is divided into multiple islands with a self-adaptation circuit in each of the islands. The self-adaptation circuit then tunes the $V_{th}$ values such that the target $V_{th}$ values are achieved. The scheme compares the delays of a $V_{thp}$-sensitive delay path and a $V_{thn}$-sensitive delay path to a pulse width created from the clock signal. The delays are set such that they are equal to the pulse width when the target $V_{th}$ values are achieved. Up/down counters count the error every N-cycle and generate digital representations of the required $V_{th}$ values. $N$ is a design parameter that trades tuning frequency for lower power consumption. Digital-to-analog (DAC) converters then translate the digital codes into $V_{th}$ values. The changes in the $V_{th}$ values are then reflected automatically into the delay path via back-biasing thus updating the delays. This way a feedback loop is created that dynamically adjusts the $V_{th}$ values. Here, the delays of a $V_{thp}$-sensitive delay path and a $V_{thn}$-sensitive delay path are simulated at the “TT” corner.

The clock frequency of the system is then set and adjusted so that the circuits work at the “TT” corner but fail at the “SS” corner. As only delay is considered in this case, chips at “FF” corner is functional and the $V_{th}$ values are not tuned. The target $V_{thp}$ and $V_{thn}$ ranges are set as follows.

$$V_{thp} < V_{thp,TT} \text{ and } V_{thn} < V_{thn,TT}.$$  \hspace{1cm} (10)

Figure 10 shows the measurement results of 5 corner-chips. X-axis and Y-axis shows the normalized oscillation frequencies of ring oscillators consisting of the same delay paths used for sensing. The 5 corner-chips are fabricated to represent each of the 5 process corners. The open-circle points correspond to the initial values of $V_{thp}$ and $V_{thn}$. Thus, these open-circle points refer to the chip condition before the self-adaption. Next, the self-adaptation mechanism is turned on, and the final adapted $V_{thp}$ and $V_{thn}$ values for each chip are shown by closed red circles in the figure. The gray zone in the figure here shows the $V_{thp}$–$V_{thn}$ space where circuit delay meets the requirement. In the figure, 4 chips among the 5 chips have moved from their initial locations to newer locations inside the gray acceptable zone. The chip labeled as “FF” does not move as the $V_{th}$ values are already lower than the tar-
targeting the optimum values of $V$ than the original corner window of Fig. 5. Optimizing a circuit as lack of correlation between the sensing circuits and the actual guaranteed due to the errors in sensing and tuning circuits as well
design is optimum. Let’s label these values as $V_{\text{thp, opt}}$ and $V_{\text{thn, opt}}$.
Then we design the self-adaptation circuit such that the following conditions are met.

$$V_{\text{thp, opt}} - \epsilon_1 < V_{\text{thp}} < V_{\text{thp, opt}} + \epsilon_2,$$

$$V_{\text{thn, opt}} - \epsilon_3 < V_{\text{thn}} < V_{\text{thn, opt}} + \epsilon_4.$$  \hspace{1cm} (11)

Here, $\epsilon_1$ to $\epsilon_4$ are error terms that reflect the errors in sensing and tuning. The self-adaptation mechanism guarantees that the fabricated chips will have the $V_{\text{thp, opt}}$ and $V_{\text{thn, opt}}$ values despite the initial deviation caused by process variation. It should also be noted that the exact values of $V_{\text{thp, opt}}$ and $V_{\text{thn, opt}}$ values cannot be guaranteed due to the errors in sensing and tuning circuits as well as lack of correlation between the sensing circuits and the actual circuits. From Eq. (11), a window can be drawn around the optimum values where the actual values will fit within that window. The size of the windows depends on the errors. Even if a larger margin is set for the window, the window is still much smaller than the original corner window of Fig. 5. Optimizing a circuit targeting the optimum values of $V_{\text{th}}$ and then verifying the circuit within a smaller window around the optimum point is much easier compared with that designing for the entire process corner window.

3.3 Best-Case Design

The self-adaption capability enables a new way of designing the circuits that can be named as “best-case design”. What it means is that we can target the values of $V_{\text{thp}}$ and $V_{\text{thn}}$ where our design is optimum. As noted that the exact values of $V_{\text{thp, opt}}$ and $V_{\text{thn, opt}}$ can be realized. To demonstrate the importance of run-time $V_{\text{th}}$ scaling along with $V_{\text{dd}}$, we plot the energy consumption per clock cycle against clock frequency for two cases in Fig. 12. In the first case, $V_{\text{th}}$ is fixed at a value of 0.35 V during the manufacture. Clearly, run-time $V_{\text{th}}$ scaling reduces energy at a wide frequency operation. The difference is particularly significant at low-frequency operation where static energy becomes dominant.

Computing the optimum values for $V_{\text{dd}}$ and $V_{\text{bb}}$ is not straightforward as these parameters depend on activity rate, clock frequency, temperature, and initial values of $V_{\text{dd}}$ and $V_{\text{th}}$. Even if we can model the relationship between the parameters, we need calibrations to correlate the model to the actual hardware. Instead, we can take a simpler approach by observing an interesting property of energy minimum operation. That is under a fixed activity rate, the ratio between the static energy $E_{\text{static}}$ and the dynamic energy $E_{\text{dynamic}}$ is almost constant across the $f_{\text{clk}}$ values [25]. Using the analytical models, we calculate the ratio of $E_{\text{static}}$ to $E_{\text{dynamic}}$ for each $f_{\text{clk}}$ and then plot the ratio values against the $f_{\text{clk}}$ values. The plot is shown in Fig. 13. Energy ratio values are shown for two different activity rates of 0.001 and 0.01 in the figure. For both of the activity rates, the energy ratio where the energy is minimum is around 0.2. Therefore, we can tune the $V_{\text{dd}}$ and $V_{\text{bb}}$ values such that an energy ratio of 0.2 is achieved by estimating $E_{\text{static}}$ and $E_{\text{dynamic}}$. The interesting feature is that the total energy is quite flat between the ratio values of 0.1 and 0.5. This shows that we do not need to precisely target a fixed ratio. Rather, we just need to ensure that $E_{\text{static}}$ and $E_{\text{dynamic}}$ is balanced with a ratio somewhere around 0.3.

As $E_{\text{static}}$ is proportional to transistor leakage current and $E_{\text{dynamic}}$ is proportional to circuit activity rate, monitoring the leakage current and activity rate gives the information required
to calculate the energy ratio. Thus, the following two sensors are necessary here.

(1) Leakage current sensor, and
(2) activity rate sensor.

Calibration is required to correlate the model which can be performed during the test time of the chip. The calibrated information can be implemented in the software.

After calculating the energy ratio value, the next step is to tune the $V_{dd}$ and $V_{th}$ values such that the target energy ratio is realized while the circuit delay meeting the required clock frequency. As there are many possible combinations of $V_{dd}$ and $V_{th}$ values, an efficient algorithm is required. For example, when the proportion of $E_{dynamic}$ is higher than the optimum value, $V_{dd}$ should be scaled down. Please note that $V_{th}$ should also be scaled down to meet the target delay. In Refs. [26], [27], an algorithm is proposed to scale the $V_{dd}$ and $V_{th}$ values at run-time to achieve the minimum energy operation.

4. Circuits for Device–Circuit Interaction

4.1 Delay-based Sensing

4.1.1 $V_{th}$-sensitive Sensor

For self-adaption of $V_{th}$ values to their optimum values, so that “best-case design” can be possible, $V_{th}$ sensors are required. Although there is no universal definition of $V_{th}$, an arbitrary definition can be used as a reference. For example, the $V_{gs}$ value that gives a fixed $I_d$ value is often used to define the $V_{th}$ value. Conversely, we can track the $V_{th}$ value by observing the change of $I_d$ value if the $V_{gs}$ can be set as a function of $V_{th}$. Then the delay change resulting from the $I_d$ change can be measured and digital operation can be performed after. Figure 14 shows two delay paths consisting of inverter gates where either the nMOSFET $V_{gs}$ or the pMOSFET $|V_{gs}|$ voltage becomes a function of the corresponding $V_{th}$ values. The $V_{th}$-sensitive gate–source voltage is realized using pass-transistors as shown in Fig. 14 [28], [29].

As will be shown next, driving the load with the transistor leakage current also gives us a delay that is exponentially related to $V_{th}$ value change. However, driving the load using leakage current requires careful design because leakage currents through the pull-up and the pull-down paths get involved also. The topology of Fig. 14 gives us a compact design that is minimal and fulfills the purpose.

4.1.2 Leakage Current Sensor

As explained earlier, driving the load with transistor OFF current gives us a simple leakage current monitor. Figure 15 shows a RO (ring oscillator) structure where inverter gates are designed such that the discharge of the load capacitance of each stage is performed by nMOSFET OFF current [30]. During the charging, the pMOSFETs turn ON and charges the load capacitances. Multiple stack transistors are used for the pull-up path so that the pull-up leakage current does not interfere with the discharging of the loads.

4.1.3 Temperature Sensor

Leakage current sensor of Fig. 15 can also be used as temperature sensors as leakage current is sensitive to both $V_{th}$ and temperature $T$. Because of the variation in $V_{th}$ values, post-fabrication calibration is required to obtain absolute temperature values. Reference [30] has shown that a one-point calibration for delay paths of Fig. 15 realizes a temperature inaccuracy within $\pm 1.5$ °C. For energy and thermal management purposes, we surmise that such accuracy is enough. One-point calibration is preferred against two-point calibration to reduce test cost significantly. The calibrated values can be incorporated into the software.

4.1.4 Multi-sensing Using Reconfigurable Architecture

Delay-based sensing enables us to design a reconfigurable architecture to monitor multiple parameters by configuring the delay path accordingly [30], [31]. Figure 16 shows a reconfigurable architecture where the topology of each inverter can be config-
energy is required. As dynamic energy is proportional to circuit activity rate, we can estimate the dynamic energy by calculating the activity rate. A digital dynamic power meter (DDPM) has been used that computes a rolling average of signal activity over a fixed number of clock cycles [37]. The accuracy of the power estimation here depends on careful selection of signals, such that they correspond to the activity of structures that have high power consumption. Instead of monitoring key logic signals, a clock activity adder (CAA) for switching power estimation is proposed [38]. The approach of the CAA takes advantage of the fact that switching power is highly correlated to register clock activity. Similarly, hardware-event sensors such as memory-access counters and instruction-execution counters can be used for dynamic energy estimation [26].

4.4 Back-gate Voltage Regulator
To realize the target $V_{th}$ values, on-chip circuits to tune the MOSFET back-gate potential are required. Voltage tuning circuits are analog circuits and therefore conventionally they are designed as a hard macro. They are then placed separately from the target logic circuit. Separate treatment as a hard macro causes area overhead. This area overhead can be minimized if the circuit can be placed and routed with the same flow that is used to place and route the logic circuits. Another common treatment for the analog circuit is to dedicate a separate supply voltage for the circuit. The supply voltage is often fixed to ensure correct operation. However, this treatment causes large area and consume resources which could have been used for other optimizations. Therefore, there is a need to establish a cell-based design flow for the back-gate voltage tuning circuit that operates under a common supply voltage. The circuit also needs to operate under a wide supply voltage range to realize run-time energy minimization. Although there are several back-gate voltage tuning circuits reported in the literature, we review a particular circuit technique proposed in Refs. [39], [40] which meet all the above criteria.

Figure 18 shows the concept of $V_{th}$ generator that operates under the single supply voltage used for the logic. Compact design enables dividing the circuit into several islands where each island has its own $V_{th}$ tuning circuits. Optimization using the fine-grain tuning is a different topic which we do not address in this paper. We focus that the circuit operates under the same supply voltage and must be compact.

4.4.1 Cell-based Design
Figure 19 shows the layout of an AES cipher and decipher circuit with integrated back-gate bias generators. To apply a cell-based design flow, a $V_{th}$ tuning circuit is first divided into several components. A manual layout has been performed for each of the components. All the cells share the same height with other

---

**Figure 16** Reconfigurable ring oscillator (RO) based circuit topology for multi-sensing.

**Figure 17** A reconfigurable inverter cell topology for $V_{thp}$ and $V_{thn}$ monitoring. C is a control signal. (a) Reconfigurable topology. (b) $V_{thp}$-sensitive configuration, and (c) $V_{thn}$-sensitive configuration.

---

4.4.1 Cell-based Design

Figure 19 shows the layout of an AES cipher and decipher circuit with integrated back-gate bias generators. To apply a cell-based design flow, a $V_{th}$ tuning circuit is first divided into several components. A manual layout has been performed for each of the components. All the cells share the same height with other
standard cells. Then, timing and physical characterizations are performed for automatic place and route. The design is then described by a hardware description language (HDL) such as Verilog. To make sure that the cells are not removed by the optimizer during automatic place and route, “don’t touch” properties are used where necessary. Furthermore, relative positions are defined for the place and route to ensure that the interconnect does not get too long. We can see in the figure that the cells highlighted by purple color and labeled as “BBG” are grouped together. These cells construct the $V_{bb}$ tuning circuit. Thus using the state-of-the-art cell-based design flow, the circuit could successfully be implemented. The chip was then fabricated in a 65 nm bulk process.

4.4.2 Measurement Results
A circuit with 6-bit of resolution is fabricated in a 65 nm bulk process. Peak-to-peak differential non-linearity (DNL) over a wide supply range of 0.5 V to 1.2 V are measured to be ±10 mV. Using a back-gate biasing coefficient of 0.1, the non-linearity of 10 mV corresponds to only 1 mV of $V_{bb}$ value change. Thus, designing with a cell-based flow is possible for $V_{bb}$ tuning circuits with acceptable linearity and wide voltage operation.

4.5 DC/DC Converter
DC/DC converter generates required $V_{dd}$ to a digital circuit core from an external voltage source. There is a power loss in converting voltage levels and as such the DC/DC converter energy efficiency significantly contributes to the system energy efficiency. DC/DC converter energy loss can be comparable to the digital circuit energy when operating at near- or sub-threshold region. DC/DC converter can be an obstacle in achieving the minimum energy operation [41].

The following three key types of DC/DC converters exist [42]. (a) Linear regulator (LR), (b) switched-capacitor regulator (SC), and (c) switching regulator (SR). LRs are best at stabilizing the output voltage but has a larger loss. SCs use a capacitive network to output a predetermined output voltage. SCs provide better efficiency than LRs at the expense of higher ripple voltage. SRs provide the flexibility in continuous output voltage and better efficiency over a wide output range but requires an inductor. For a run-time minimum energy optimization, SRs are better suited. The control of an SR whose output varies largely is a challenge. Especially, when the load current is small, reducing the switching loss in an SR DC/DC converter is a challenge. There are several attempts in integrating the DC/DC converter into the system energy minimization loop to account for the DC/DC converter energy loss [41], [42]. Although DC/DC converters have a long history, new design methodology and control mechanism are still being researched targeting ultra-low power applications.

5. Conclusion
In this paper, we have reviewed the energy consumption of digital circuits and its dependence on the circuit parameters such as activity rate and clock frequency. Energy and delay are in a trade-off relationship and as such, the most effective way to minimize energy consumption is to operate the circuit at the slowest speed that satisfies the performance requirement and then setting the supply voltage accordingly. However, under the significant presence of process, voltage, temperature and workload variations, having constant threshold devices is not optimum. Therefore, the need for device–circuit interaction arises. With the aid of device monitoring by sensors and device tuning via back-gate biasing, we have shown that run-time minimum energy operation can be achieved at all conditions. Focusing on the ease-of-design to reduce cost, we have reviewed delay-based sensors to monitor performances of p-type and n-type transistors. Having many sensors in the chip can be nice, but the challenge arises in how to use them. We, therefore, have discussed a simplified approach to realize minimum energy operation by observing the fact that the ratio between static energy and dynamic energy is almost constant near the minimum energy operation. Leakage-current and activity sensors can be used to compute the energy ratio. Then, we need to scale the $V_{dd}$ and $V_{bb}$ values accordingly. The guidelines to minimum energy operation and the circuit design techniques discussed in the paper should help designers in choosing the right approach toward minimum energy operation.

Acknowledgments This work has been supported in part by KAKENHI 16H01713 and 25280014.

References
Kuroda, T.: Optimization and Control of


Markovic, D. and Stojanovic, V.: Methods for True Energy-

Kamae, N., Islam, A.K.M.M., Tsuchiya, A., Ishihara, T. and Onodera,

IPSJ Transactions on System LSI Design Methodology Vol.12 2–12 (Feb. 2019)


A.K.M. Mahfuzul Islam received his B.E. degree in electrical and electronics engineering in 2009, his M.E. degree in communications and computer engineering in 2011, and his Ph.D. degree in informatics in 2014, all from Kyoto University, Kyoto, Japan. He was a Research Fellow of Japan Society for the Promotion of Science from 2013 to 2015. In 2015, he joined the Institute of Industrial Science, University of Tokyo, Japan, as a Research Associate. Since 2018, he has been a Junior Associate Professor of the Department of Electrical Engineering at Kyoto University. His research interests include characterization and modeling of transistor noise, low-power CMOS sensor circuits, and low-power design methodology. He received the best paper award at ICMTS 2017, student design award at A-SSCC 2013 and was nominated for best paper at CICC 2018. He is a member of IEEE.

Hidetoshi Onodera received his B.S., M.S. and Ph.D. degrees in Electronic Engineering, all from Kyoto University, Kyoto, Japan. He joined the Department of Electronics, Kyoto University, in 1983. He is currently a Professor in the Department of Communications and Computer Engineering, Graduate School of Informatics, Kyoto University. His research interests include design technologies for Digital, Analog, and RF LSIs, with particular emphasis on low-power design, design for manufacturability, and design for dependability. He served as a Program Chair and a General Chair of ICCAD and ASP-DAC. He was a Chairman of SSCS Kansai Chapter, IEEE CASS Kansai Chapter, IEEE Kansai Section, IPSJ SIG-SLDM (System LSI Design Methodology), and IEICE Technical Group on VLSI Design Technologies. He served as an Editor-in-Chief of IEICE Transactions on Electronics and IPSJ Transactions on System LSI Design methodology. He is an IEEE Fellow, an IEICE Fellow, and a Member of Science Council of Japan.

(Invited by Editor-in-Chief: Nozomu Tagawa)