Application-Aware Quality-Energy Optimization: Mathematical Models Enabled Simultaneous Quality and Energy-Sensitive Optimal Memory Design

Yiwen Xu, Hritom Das, *Student Member, IEEE*, and Na Gong, *Member, IEEE*

Abstract—Energy efficiency is nowadays a well-known principal design goal across all layers of computing systems (e.g., sensors, mobile, cloud). With diminishing benefits from CMOS technology scaling and increasing demands from unprecedented data size, new memory hardware innovation is greatly needed to enhance energy efficiency of computing systems. Recently, quality-aware hardware design techniques have been developed from different stack layers (device/circuit/architecture/system) to enable near-threshold/sub-threshold voltage operation by trading off between quality and power efficiency. Specifically, based on the energy-quality trade-off and application requirement, memory hardware is designed for the work mode with maximum quality first (step1-design for the work mode) and then the supply voltage will be adjusted in the sleep mode with just-enough quality to achieve maximum efficiency (step2-adjust in the sleep mode). We first propose mathematical models for this two-step design method to avoid time-consuming and laborious ASIC design iterations in traditional hardware design process. However, such a two-step design method focuses more at the application quality than the energy efficiency in all cases. In addition, as the supply voltage in the second step depends on the design from the first step, the solution space of the second step may be greatly limited, far from the true minimum supply voltage. To handle these issues, we propose a new design concept, the simultaneous quality and energy-sensitive optimal design (SQEOD), in which the two objectives are considered simultaneously rather than by two separate steps. By introducing a system-wised importance weight parameter in the modeling process, our method demonstrates system-specific SQEOD mathematical models for different memory designs with various requirements on the application quality and/or energy efficiency. The results of the numerical studies on embedded memory design show that the proposed models provide a useful and fast tool to enable the optimal hardware designs.

Index Terms—Application requirement, memory design, optimization models, power efficiency, quality

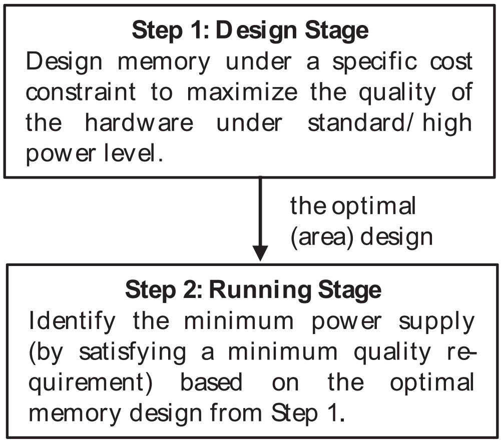
1 Introduction

With the exponential growth of semiconductor technology, computing systems are widely applied in many areas, such as cloud computing, laptop computers, cell phones, and Internet-of-the-Things (IoT) sensors. The broad application range results in a huge variation of working (active) state of computing systems. For example, leading servers are performance-driven with a long working cycle to support computation-intensive tasks (e.g., artificial intelligence algorithms and other growing cloud applications [1], [2]); mobile devices such as smart phones typically have moderate performance requirement; in the IoT areas such as actuators, wearable devices, and various sensors, computing systems are inactive most of the time, staying in low-power mode (sleep mode), and the performance can be compromised. Although the percentage of working state and application requirement vary, two challenges remain the same. First, energy efficiency enhancement is crucial for all of those computing platforms. Efficient servers and high performance computing systems are extremely important to reduce the operational cost [3]. Energy reduction is also critical in mobile electronics, due to the limited thermal budget and battery energy availability [4]. Similarly, the design of energy-efficient IoT devices is essential, as those miniaturized systems are powered from tightly-constrained batteries or harvest energy from the surrounding environment [5]. Meanwhile, embedded memory hardware, plays an increasingly important role in today's computing systems. For example, embedded memories usually dominate the system power consumption (e.g., over 50 percent of the video decoders [6], [7], [8] and over 60 percent deep learning systems [9]).

Due to the huge variation of computing applications, an emerging memory design problem is how to meet the quality requirement and meanwhile optimize the energy efficiency of the hardware. Here, the "quality" of a computing component or system represents the delivered results and it is typically quantified by application-specific metrics [10], such as peak

- Y. Xu is with the Department of Industrial and Manufacturing Engineering, North Dakota State University, ND 58108, USA. E-mail: yiwen.xu@ndsu.edu.
- H. Das is with the Department of Electrical and Computer Engineering, North Dakota State University, ND 58108, USA. E-mail: hritom.das@ndsu.edu.
- N. Gong is with the Department of Electrical and Computer Engineering, University of South Alabama, AL 36688, USA. E-mail: nagong@southalabama.edu.

*Manuscript received 9 July 2019; revised 12 Mar. 2020; accepted 21 May 2020. Date of publication 4 June 2020; date of current version 8 Dec. 2021.
(Corresponding author: Na Gong.)
Recommended for acceptance by D. Atienza.
Digital Object Identifier no. 10.1109/TSUSC.2020.2999882*



Step 1: Design Stage

Design memory under a specific cost constraint to maximize the quality of the hardware under standard/high power level.

Step 2: Running Stage

Identify the minimum power supply (by satisfying a minimum quality requirement) based on the optimal memory design from Step 1.

Fig. 1 Flow chart of the two-step method.

signal-to-noise ratio (PSNR) for images and videos, bit error rate (BER) for communication, signal-to-noise ratio (SNR) for analog interfaces, and prediction accuracy for machine learning systems. Consider the display buffer of a smart phone for example. When the power is at the standard level, one prefers a better video output quality with a higher PSNR value to maximize the user experience; but when the power level is lower than a threshold (for example, 20 percent), the power saving mode will be activated to save the power consumption at the cost of acceptably reducing some function quality. In general, the issue is how to design a memory such that (i) while a circuit is working at the standard power level, it will maximize the quality, and that (ii) as the hardware comes to the sleep/power saving mode, the supply voltage will be scaled to meet the minimum quality requirement (i.e., the "just-enough quality", as mentioned in [10]).

However, in most cases the above-mentioned two objectives – maximizing quality and maximizing energy efficiency – cannot be satisfied simultaneously and the memory designers have to make a trade-off. To achieve the first objective, the memory has to consume high-level energy, which is inconsistent with the second objective that dedicates to minimize such consumptions. Indeed, in a traditional hardware-design process, these two types of designs (i.e., the design for the working mode with maximum application quality and for the sleeping mode with maximum power efficiency) are considered separately. Since in many cases the requirement of the application quality of a hardware dominates the overall design objective, designers often aim at implementing hardware under a specific cost constraint (e.g., silicon area) to maximize the quality first, and then during the runtime process, identify the minimum power supply under this design to enable the low power situation, [3], [11], [12], [13], [14]. The minimum power supply in the second step has to satisfy a given minimum application quality requirement. We call this design method *the two-step method*. The flow chart of the two-step method is illustrated in Fig. 1. Traditionally, application-specific Integrated Circuits (ASIC) designers focus on Step 1 to customize the hardware during design time [11], [12], [15], [16], and based on the developed hardware, system designers work on the optimal voltage during the runtime [3].

A concern of the traditional two-step method is that the low-level power supply design (i.e., the second objective) is not considered when the design of the maximum application quality under the standard (or target) power supply (i.e., the first objective) is obtained in the first step. As a consequent, the solution in the second step is based on the implemented hardware of the first step, which will limit the solution space of the low power supply design. In other words, the solution obtained via this methodology is usually not the true overall optimal solution considering both design objectives simultaneously. It assumes a high priority to the quality objective in ALL design cases. This is obviously not proper for those systems that have a very low percentage of working state and that the performance can be compromised, such as various IoT devices [5].

In this paper, we first propose mathematical models for the widely-used two-step method, thereby avoiding time-consuming and laborious ASIC design iterations in traditional hardware design process [11], [12], [15], [16]. Two types of models, continuous and discrete models, will be discussed. Then, we propose a new design concept under which the two objectives will be considered simultaneously, rather than by two separate steps. Specifically, during the hardware design process in which we are trying to maximize the quality under standard power level, we need to meanwhile consider the low-level voltage setting to reduce the power consumption. We call this design concept *simultaneous quality and energy-sensitive optimal design* (SQEOD). To the best of our knowledge, this concept has never been proposed in literature. We formulate the SQEOD concept to mathematical models. The results of the numerical studies on SRAM and hybrid memory design demonstrate the effectiveness of the proposed models for different memories with a variety of performance requirement.

It should be noted that, in addition to the embedded memory which we focus on in this paper, the proposed SQEOD concept can be also applied to general computing hardware design process. The mathematical models developed in this paper can be considered as a standard design process for general hardware design in the future computing systems.

The paper is organized as follows. In Section 2, we discuss related research that consider low-voltage memory design and compare their work with our proposed work. Sections 3 and 4 introduce the models of the two-step method and the SQEOD concept. In each part, we discuss two models: (i) the memory failure rate is a continuous function (i.e., the continuous models), and (ii) both silicon area and voltage have discrete options (i.e., the discrete models). Application of hybrid memory structure without silicon area overhead is also mentioned. Section 5 includes numerical examples of the models introduced in Sections 3 and 4. Finally, Section 6 concludes the paper.

2 Related Work

Voltage scaling is one of the most effective techniques to reduce the power consumption of memories. Studies have shown that the energy efficiency is maximized when a memory operates at a near-threshold voltage. Significant amounts of research that target low-voltage SRAM design have been reported in the literature. In this section, we briefly review some existing work related to the proposed technique. Low-voltage SRAMs can be broadly classified into two different categories including upsizing 6T bitcells and hybrid memory design.

2.1 Upsizing 6T Bitcells

Upsizing SRAM 6T is a widely used low voltage memory design technique. At low voltages, SRAM failures are mainly caused by process variations, in particular threshold voltage variations (σV_{th}), which can be expressed as [15], [17]:

$$\sigma V_{th} \ = \frac{A_{VT}}{\sqrt{WL}}, \tag{1}$$

where A_{VT} is a technology dependent constant, and W and L represent the width and length of the transistor. According to (1), σV_{th} is inversely proportional to $\sqrt{WL}$, which means that as W and L increase, the threshold voltage variations are reduced. Accordingly, upsizing the transistors of 6T can effectively reduce the memory failure at low voltages.

To upsize SRAM 6T bitcells for better power efficiency, in [15], a heterogeneous sizing scheme was presented to reduce the failure probability of conventional 6T bitcells by assigning different sizing values for different bitcells. In [18], Gong *et al.* analyzed the size-dependent SRAM failure characteristics and presented bitcell sizing techniques for area-priority and quality-priority mobile video applications. Recently, Kim *et al.* developed a sizing technique to adaptively select the bitcell sizes of SRAMs according to their sensitivities to the quality degradation in video compression while maintaining the total SRAM silicon area [19]. However, all the above designs aim to identify the sizing solutions using heuristic searching algorithms without a mathematical model, and the optimal soluton may not be achieved. In this paper, two mathematical models, continuous models and discrete models, are developed to meet different design conditions. During the optimization process, a varity of applications with different percentages of working state are also considered for both two-step process and the proposed SQEOD design process.

2.2 Hybrid Memory Design

In addition to tradition 6T bitcells, various more-than-6T have been developed to achieve low power operation, such as single-ended read-decoupled 8T bitcells [20], 10T bitcells [21], [22], and dual-feedback 13T bitcells [23]. However, those more-than-6T bitcells usually cause significant silicon area (e.g., 3× area overhead [22]). Accordingly, recent optimized memory designs often adopt hybrid SRAM bitcells for low voltage operations. For example, in [12], Chang *et al.* presented a hybrid 6T+8T SRAM to achieve video memory with quality-power optimization. The video memory reported in [13] was designed as a hybrid 8T+10T memory to store different pixel data for power savings. However, such hybrid structures increase the implementation complexity of peripheral circuitries such as memory decoders. In our earlier work [11], we discussed different cases and provide optimization solutions considering 6T upsizing design and hybrid memory design options. But all these works fix the power supply to be a constant, which is equivalent to the first step of the two-step method. In this paper, we consider the supply voltage as a decision variable and intent to obtain an optimal solution that provides a tradeoff between the quality performance and energy efficiency under a given imprance weight.

2.3 The Calculation of the Expected Mean Squared Error

Mean squared error (MSE) is a widely used quantity to measure the application quality of a memory [12], [13], [14], [15], since minimizing the MSE is equivalent to maximizing the quality. In our earlier work [11], we derived an explicate formula for the MSE.

Consider a memory chip including m-by-n bytes where each byte is composed of 8 bitcells. Let $Y_{ijk}^{(O)}$ and $Y_{ijk}^{(D)}$ denote the binary data of the $k\text{th}$ bitcell in the $i\text{th}$ row $j\text{th}$ column ($k = 0, \cdots, 7; i = 1, \cdots, m; j = 1, \cdots, n$) of the byte of the original and degraded chip, respectively. The MSE of the whole chip can be expressed as

$$E\ (MSE) = \frac{1}{mn} \sum_{i\,=\,1}^{m} \sum_{j\,=\,1}^{n} \sum_{k\,=\,0}^{7} 4^k q_{ijk}\ ,$$

where q_{ijk} is the failure rate of the ijk^{th} bitcell. Specially, if $q_{ijk} \equiv q_k$, then $E(MSE) = \sum_{k=0}^{7} 4^k q_k$. More details and the proof can be found in [11].

3 The Two-Step Method

In this section, we derive two mathematical models for the two-step method: the continuous and the discrete models. Two factors, the silicon area of a memory bitcell and the supply voltage of the chip, are considered to affect the memory failure rate. We also assume that all memory bitcells have the same power supply.

3.1 Continuous Models

Let $f(s, v)$ be the function of failure rate of a bitcell, where s and v denote the silicon area of the bitcell and the supply voltage of the memory chip. In the continuous model, we assume that $f(s, v)$ is known or can be estimated.

According to the definition of the two-step method, the first step is to find the optimal application quality design under the standard voltage supply, which can be formulated by the following mathematical model.

Step 1: finding the optimal application quality design

$$[\mathbf{2MC1}] \qquad \min_{\boldsymbol{s}} \sum_{i\,=\,1}^{m} \sum_{j\,=\,1}^{n} \sum_{k\,=\,0}^{7} 4^k f\big(s_{ijk}, v_0\big) \tag{2}$$

$$\textbf{s.t.} \qquad \sum_{i=1}^{m} \sum_{j\,=\,1}^{n} \sum_{k\,=\,0}^{7} s_{ijk} \leq s_{total} \tag{3}$$

$$s_{ijk} \geq s_{min},\ i\ =\ 1, \ldots, m; j\ =\ 1, \ldots, n; \quad k\ =\ 0, \ldots, 7. \tag{4}$$

The objective function (2) is to minimize the MSE (i.e., to maximize the application quality) under the standard supply voltage v_0. The only decision variable here is $\boldsymbol{s}\ =\ [s_{ijk}]$ $(i\ =\ 1, \ldots, m; j\ =\ 1, \ldots, n; k\ =\ 0, \ldots, 7)$, where s_{ijk} indicates the silicon area of the $ijkth$ bitcell. Constraint (3) guarantees that the sum of the silicon areas of all bitcells cannot exceed a given constant s_{total}, and constraint (4) gives the minimum allowed bitcell area (due to the implementation cost constraint) in the design.

Let s_{ijk}^* be the optimal solution to problem [2MC1]. The second step is to determine the minimum supply voltage for the low-level power case, under the given silicon area design s^*_{ijk}. The model of the second step can be formulated as follows.

Step 2: determining the optimal low-level voltage design based on the solution to [2MC1]

$$[\mathbf{2MC2}] \qquad \min_{v} \ v \tag{5}$$

$$\text{s.t.} \qquad v_{min} \le v \le v_0 \tag{6}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7} 4^k f\left(s^*_{ijk}, v\right) \le mse_{\max}. \tag{7}$$

The objective function (5) is to minimize the supply voltage of the low-level power case. Constraint (6) sets the feasible range of the voltage supply. Constraint (7) assures that the MSE of the memory under the low-level voltage design will not exceed a given threshold $mse_{\max}$, preventing the memory from getting a very low application quality due to an over-low trivial voltage setting (for example, $v = 0$). Note that there is no area-limit constraint (like constraint (3)) in [2MC1], since the silicon area-design is already fixed to be s^*_{ijk}. The final design solution of the two-step method is given by $(s^*_{ijk},\ v^*)$, where v^* is the optimal solution to problem [2MC2].

3.2 Discrete Models

In many real cases, however, the failure rate $f(s,v)$ is not known, nor can it be well-fitted. Instead, the designer may have several (discrete) options to select the area for each bitcell s_{ijk} [15] and the lower-level supply voltage v. Suppose we have d options for the silicon area design of the $ijkth$ cell in the first step and $g+1$ options (including the standard voltage v_0) for the low-level supply voltage design in the second step. Let q_{ijkbc} and s_{ijkb} be the failure rate and silicon area of the $ijkth$ bitcell, respectively, when the bitcell's area is selected to be the bth $(b = 1, \ldots, d)$ option and the supply voltage is selected to be the cth option $(c = 0, 1, \ldots, g)$. For notation convenience, we define $c = 0$ to be the option of selecting the standard supply voltage (i.e., v_0). The first step of the discrete two-step method can be formulated by the following integer linear program (ILP).

Step 1: finding the optimal application quality design

$$[\mathbf{2MD1}] \qquad \min_{\boldsymbol{x}} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7}\sum_{b=1}^{d} 4^k q_{ijkb0} x_{ijkb} \tag{8}$$

$$\text{s.t.} \qquad \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7}\sum_{b=1}^{d} s_{ijkb} x_{ijkb} \le s_{total} \tag{9}$$

$$\sum_{b=1}^{d} x_{ijkb} \ge 1,\ \forall i,j,k \tag{10}$$

$$x_{ijkb} \in \{0,1\},\ \forall i,j,k,b \tag{11}$$

where the binary decision variable $x_{ijkb} = 1$, if the b^{th} area-option is selected as the area of the ijk^{th} cell; $x_{ijkb} = 0$, otherwise.

The objective function (8) is to minimize the MSE. Constraint (9) defines the total silicon area limit, and constraint (10) assures that each bitcell will finally get exactly one option as their area selection. Note that since [2MD1] is a minimization problem, constraint (10) is equivalent to $\sum_{b=1}^{d} x_{ijkb} = 1, \forall i,j,k$. We use the inequality because usually an inequality constraint is easier to be handled than an equality constraint in an optimization problem.

Let x^*_{ijkb} be the optimal solution to problem [2MD1]. Then the second step is to select the minimum supply voltage for the low-level power case among totally $g+1$ voltage-options, under the given silicon area design x^*_{ijkb}. Let v_c be the value of the low-level supply voltage if it is selected to be the c^{th} option $(c = 0, 1, \ldots, g)$. Then the second step can be formulated as the following ILP.

Step 2: determining the optimal voltage design based on the solution to [2MD1]

$$[\mathbf{2MD2}] \qquad \min_{\boldsymbol{y}} \sum_{c=0}^{g} v_c y_c \tag{12}$$

$$\text{s.t.} \qquad \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7}\sum_{c=0}^{g} 4^k q_{ijkbc} x^*_{ijkb} y_c \le mse_{max} \tag{13}$$

$$\sum_{c=0}^{g} y_c \ge 1 \tag{14}$$

$$y_c \in \{0,1\}, c = 0, 1, \ldots, g, \tag{15}$$

where the binary decision variable $y_c = 1$, if the c^{th} option is selected as the low-level supply voltage of the memory; $y_c = 0$, otherwise.

The objective function (12) is to minimize the voltage supply, and constraint (13) is to guarantee that the MSE of the memory under the low-level power design will not exceed a given threshold. Constraint (14) assures that only one voltage of the low-level power supply can be selected for the whole memory.

4 The SQEOD

While the models in the two-step method are relatively easy to solve (as the silicon area of the hardware in the second step is fixed to be the optimal solution to the first-step problem), a drawback of it is that it enforces the performance-quality objective to dominate the energy-efficiency objective.

Indeed, if the system is known to have a low percentage of working state, a simple way to overcome the drawback is to reverse the two-step method: solving the second-step problem before solving the first-step problem. However, this methodology has two serious flaws. First, since the hardware silicon area in the second step is unknown (which is given by the first-step problem in the original two-step method), one has to consider it as another decision variable. But in this case the "two-step" loses its meaning, as within only one step the solutions of both area-design and the low-level voltage design are obtained, which completely neglects the performance-quality objective. Second, this simple procedure cannot differentiate two low-duty cycle systems with varied percentages of working states. Consider two smart devices for example. In the first device 10 percent is consider as the low-level power supply threshold under which the device will automatically turn to the power-save mode, while in the second device it is 25 percent. Compared with the second device, the first one has a greater emphasis on the significance of application quality as 90 percent of the time the

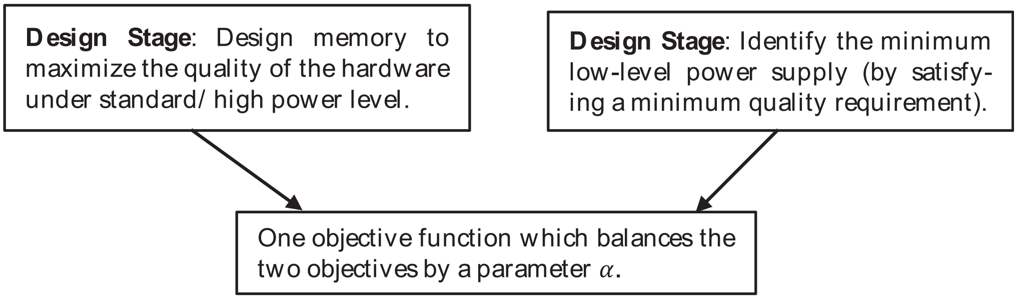

Design Stage: Design memory to maximize the quality of the hardware under standard/ high power level.
Design Stage: Identify the minimum low-level power supply (by satisfying a minimum quality requirement).
One objective function which balances the two objectives by a parameter α.

Fig. 2. Flow chart of SQEOD design.

device is set at a high application quality. But if we use the above-mentioned reversed procedure, these two devices will inevitably be handled by the same way and we will get exactly the same solutions.

To overcome these issues, we propose the concept of the SQEOD in which the objectives of maximizing the application quality and maximizing the energy efficiency are considered simultaneously and hence we avoid the issue that one objective dominates the other. We use a parameter, α, to indicate the relative significance of the two objectives. The setting of α will be discussed in Section 4.1.2 and 4.2.2. The flow chart of the SQEOD concept is shown in Fig. 2. One can see the difference between the two-step method and the SQEOD by comparing Figs. 1 and 2. Similar to the two-step method, we discuss two types of models: the continuous and the discrete SQEODs in this section.

4.1 The Continuous SQEOD Models

In the continuous SQEOD model we assume that the function of failure rate of a bitcell, $f(s, v)$ is known or can be estimated.

4.1.1 The Mathematical Model

The continuous SQEOD model can be formulated as the follows.

$$[\mathbf{SMC}] \qquad \min_{\boldsymbol{s},\, v} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7} 4^k f\big(s_{ijk}, v_0\big) + \alpha v^2 \tag{16}$$

$$\text{s.t.} \qquad \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7} s_{ijk} \leq s_{total} \tag{17}$$

$$s_{ijk} \geq s_{min},\ i = 1,\ldots,m; j = 1,\ldots,n; \\ k = 0,\ldots,7 \tag{18}$$

$$v_{min} \leq v \leq v_0 \tag{19}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7} 4^k f\big(s_{ijk}, v\big) \leq mse_{\max}, \tag{20}$$

The objective function (16) is a combination of the two design objectives: (i) to minimize the MSE at the standard supply voltage v_0, by $\sum_i^m \sum_j^n \sum_k^7 4^k f(s_{ijk}, v_0)$, and (ii) to minimize the supply voltage when the power is at a low level, by αv^2. Since in general it is very difficult to find the accurate relationship of the power consumption (E_p) versus the bitcell voltage (v), we approximate this relationship by a quadratic function

$$E_p = \mu v^2,$$

where μ is assumed to be constant for a given circuit system and it is loading capacitor dependent. More discussions can be found in [24]. The meaning of the constraints (17), (18), (19), (20) are similar to that of constraints (3), (4), (6), (7) in the continuous two-step models [2MC1] and [2MC2]. But note that in (20) s_{ijk} is still a decision variable, while in (7) it is fixed to be constant s^*_{ijk}.

It is worth mentioning that the complete objective function should be

$$\min_{\boldsymbol{s},\, v} \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7} 4^k f\big(s_{ijk}, v_0\big) + \tilde{\alpha}\mu v^2,$$

where $\tilde{\alpha}$ is the original parameter for setting the weight of the two objectives. To simplify it, we cancel the constraint coefficient $\frac{1}{mn}$ and set $\alpha = mn\tilde{\alpha}\mu$. As m, n, μ are constant, tuning α is equivalent to tuning $\tilde{\alpha}$.

4.1.2 Setting the Weight Parameter α

In model [SMC], parameters such as v_0, s_{total}, s_{min}, v_{min}, and mse_{max} are usually engineering specifications determined by the target application. But one needs to set the weight parameter α before running the model.

Given the two design objectives (i) to minimize MSE when the power level is standard, and (ii) to minimize supply voltage when the power level is low, clearly, the larger the α is, the more significant the second design objective will be in the objective function of [SMC]. However, the scale of the solutions to the two design objectives can vary greatly. A reasonable way to set the parameter α is to first normalize the scale of the two objectives, and then determine the significance of them in (16) by a system threshold. To this end, we first solve the following subproblem

$$[\mathbf{SPC1}] \qquad z^*_{1.1C} = \min_{\boldsymbol{s}} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{7} 4^k f\big(s_{ijk}, v_0\big) \tag{21}$$

$$\text{s.t.} \qquad \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k\,=\,0}^{7} s_{ijk} \leq s_{total} \tag{22}$$

$$s_{ijk} \geq s_{min},\ i = 1,\ldots,m; j = 1,\ldots,n; \\ k = 0,\ldots,7, \tag{23}$$

where the objective function is only to minimize the overall MSE, assuming that the supply voltage is standard.

Then, we solve the second subproblem to get the overall minimum voltage supply:

$$[\mathbf{SPC2}] \qquad z^*_{1.2C} = \min_{\boldsymbol{s},\, v} \ v \tag{24}$$

$$\text{s.t.} \qquad \sum_{i\,=\,1}^{m}\sum_{j\,=\,1}^{n}\sum_{k\,=\,0}^{7} s_{ijk} \le s_{total} \tag{25}$$

$$s_{ijk} \ge s_{min},\ i \ = \ 1,\ldots,m; j \ = \ 1,\ldots,n; \quad k \ = \ 0,\ldots,7 \tag{26}$$

$$v_{min} \le v \le v_0(27) \tag{27}$$

$$\sum_{i\,=\,1}^{m}\sum_{j\,=\,1}^{n}\sum_{k\,=\,0}^{7} 4^k f\big(s_{ijk}, v\big) \le mse_{max}. \tag{28}$$

Clearly, solving (24) is equivalent to solving $\min \alpha v^2$. Note that problem [SPC2] is different from [2MC2], because the silicon area s_{ijk} is a decision variable in the former but fixed in the latter.

To normalize $z^*_{1.1C}$ and $z^*_{1.2C}$, let

$$\alpha_{50} : = \frac{z^*_{1.1C}}{\big(z^*_{1.2C}\big)^2}. \tag{29}$$

We take advantage of α_{50} to indicate the importance of the two objectives in model [SMC]. If they are equivalently important, then we can simply set $\alpha \ = \alpha_{50}$ in (16). If we have a quantitative system threshold $\eta\%$, in terms of the low-level power indication or idle cycles percentages for example, then we can set

$$\alpha \ = \frac{\eta}{100 - \eta} \ \alpha_{50}. \tag{30}$$

The greater the α is, the more sensitive model [SMC] is to the second objective, and hence the more important the second objective will be. For example, the typical low-level power supply threshold of today's smart phone is 20 percent, which roughly means that under 80 percent of the usage time the device is required to provide a high-quality performance, while 20 percent of the usage time it is during the power-save mode. Thus, we can set $\alpha \ = (20\%/80\%) \quad \alpha_{50} = \ 0.25\alpha_{50}$, indicating that after the normalization the first design objective (i.e., to minimize the MSE when the power level is standard) is roughly four times as important as the second design objective (i.e., to minimize the supply voltage when the power level is low). For IoT sensors with very low duty cycles (e.g., 99-99.9 percent idle time) [25], [26], the α varies from $(99\%/1\%)\ \alpha_{50} = \ 99\alpha_{50}$ to $999\alpha_{50}$, indicating an extremely high requirement for power saving (i.e., the second design objective).

Let the optimal solution to [SMC] be $(\boldsymbol{s}^*_{SMC}, v^*_{SMC})$. For any $\alpha \ge 0$, we must have

$$\sum_{i\,=\,1}^{m}\sum_{j\,=\,1}^{n}\sum_{k\,=\,0}^{7} 4^k f\Big(s^*_{ijk}, v_0\Big) \ge z^*_{1.1C}, \ \text{and}\,(v^*)^2 \ge \big(z^*_{1.2C}\big)^2,$$

and hence the optimal value of [SMC] has a lower bound $z^*_{1.1C} + \alpha z^*_{1.2C}$. Moreover, problems [2MC1] and [2MC2] in the two-step method are two special cases of problem [SMC]: by setting $\alpha \ = \ 0$ and $\alpha \to +\infty$, respectively.

Algorithm 1 shows our procedures to solve the optimal solution to the continuous SQEOD. The reason we solve problem [SPC2] first rather than [SPC1], is that if there is no feasible solution to [SPC2], then the problem [SMC] will definitely have no solution. In this case, to make the problem feasible, one may consider to increase s_{total} and/or mse_{max}.

Algorithm 1. For Solving Continuous SQEOD Models

```
Input parameters: $v_0$, $s_{total}$, $s_{min}$, $v_{min}$, $mse_{max}$ and $\eta$.
Output: optimal design $(\boldsymbol{s}^*_{SMC}, v^*_{SMC})$.
1 Solve [SPC2] and get $z^*_{1.2C}$. If there is no solution, then the
    problem is not feasible and STOP.
2 Solve [SPC1] and get $z^*_{1.1C}$.
3 Calculate $\alpha_{50}$ and $\alpha$ by (29) and (30).
4 Solve [SMC], and RETURN the optimal solution
    $(\boldsymbol{s}^*_{SMC}, v^*_{SMC})$. STOP.
```

4.2 The Discrete SQEOD Models

4.2.1 The Mathematical Model

In case that the failure rate $f(s, v)$ is not known or cannot be well-fitted, we need to consider discrete SQEOD models. Using the notations defined in Section 3.2, the discrete SQEOD can be formulated as follows.

$$\min_{\boldsymbol{x},\boldsymbol{y}} \sum_{i\,=\,1}^{m}\sum_{j\,=\,1}^{n}\sum_{k\,=\,0}^{7}\sum_{b\,=\,1}^{d} 4^k q_{ijkb0} x_{ijkb} + \alpha \sum_{c\,=\,0}^{g} v_c^2 y_c \tag{31}$$

$$\text{s.t.} \ \sum_{i\,=\,1}^{m}\sum_{j\,=\,1}^{n}\sum_{k\,=\,0}^{7}\sum_{b\,=\,1}^{d} s_{ijkb} x_{ijkb} \le s_{total} \tag{32}$$

$$\sum_{i\,=\,1}^{m}\sum_{j\,=\,1}^{n}\sum_{k\,=\,0}^{7}\sum_{b\,=\,1}^{d}\sum_{c\,=\,1}^{g} 4^k q_{ijkbc} x_{ijkb} y_c \le mse_{max} \tag{33}$$

$$\sum_{b\,=\,1}^{d} x_{ijkb} \ge 1, \ \forall i, j, k \tag{34}$$

$$\sum_{c\,=\,1}^{g} y_c \ge 1 \tag{35}$$

$$x_{ijkb}, y_c \in \{0, 1\}, \ \forall i, j, k, b, c, \tag{36}$$

where the decision variable $x_{ijkb} = \ 1$, if the bth silicon area-option is selected as the area of the $ijkth$ cell; $x_{ijkb} = \ 0$, otherwise. Decision variable $y_c = \ 1$, if the cth voltage-option is selected as the low-level voltage; $y_c = \ 0$, otherwise.

In the objective function (31), the first part, $\sum_{i\,=\,1}^{m} \sum_{j\,=\,1}^{n} \sum_{k\,=\,0}^{7} \sum_{b\,=\,1}^{d} 4^k q_{ijkb0} x_{ijkb}$, represents the MSE of the whole memory at the standard voltage v_0. The second part, $\alpha \sum_{c\,=\,0}^{g} v_c^2 y_c$, is the approximate power consumption with the weight parameter α. The interpretations of constraints (32)–(35) are similar to constraints (9), (13), (10), and (14), respectively, in models [2MD1] and [2MD2]. Note that (13) is a linear constraint (since $x_{ijkb} = x^*_{ijkb}$ is fixed), but (33) is not (for x_{ijkb} is also a decision variable). To linearize constraint (33), we define a new binary decision variable w_{ijkbc} to equivalently indicate $x_{ijkb} y_c$, by the following additional constraints

$$\begin{cases} \frac{x_{ijkb}+y_c-1}{2} \leq w_{ijkbc} \leq \frac{x_{ijkb}+y_c}{2}, \ \forall i,j,k,b,c. \\ w_{ijkbc} \in \{0,1\}, \ \forall i,j,k,b,c \end{cases} \tag{37}$$

Consequently, the problem can be reformulated as the following ILP.

$$[\mathbf{SMD}] \ \min_{\boldsymbol{x},\boldsymbol{y}} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=0}^{7} \sum_{b=1}^{d} 4^k q_{ijkb0} x_{ijkb} + \alpha \sum_{c=0}^{g} v_c^2 y_c, \tag{38}$$

s.t. Constraints (32), (34) – (36), (37)

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=0}^{7} \sum_{b=1}^{d} \sum_{c=1}^{g} 4^k q_{ijkbc} w_{ijkbc} \leq mse_{max}. \tag{39}$$

4.2.2 Setting the Weight Parameter α

To determine the problem-based weight parameter α, similar to Section 4.1.2, we propose the following two subproblems.

$$[\mathbf{SPD1}] \qquad z^*_{1.1D} = \min_{\boldsymbol{x},\boldsymbol{y}} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=0}^{7} \sum_{b=1}^{d} 4^k q_{ijkb0} x_{ijkb} \tag{40}$$

$$\text{s.t.} \qquad \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=0}^{7} \sum_{b=1}^{d} s_{ijkb} x_{ijkb} \leq s_{total} \tag{41}$$

$$\sum_{b=1}^{d} x_{ijkb} \geq 1, \ \forall i,j,k \tag{42}$$

$$x_{ijkb}, \in \{0,1\}, \ \forall i,j,k,b, \tag{43}$$

and

$$[\mathbf{SPD2}] \ z^*_{1.2D} = \min_{\boldsymbol{x},\boldsymbol{y}} \sum_{c=0}^{g} v_c \ y_c \tag{44}$$

$$\text{s.t.} \ \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=0}^{7} \sum_{b=1}^{d} s_{ijkb} x_{ijkb} \leq s_{total} \tag{45}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=0}^{7} \sum_{b=1}^{d} \sum_{c=1}^{g} 4^k q_{ijkbc} w_{ijkbc} \leq mse_{max} \tag{46}$$

$$\sum_{b=1}^{d} x_{ijkb} \geq 1, \ \forall i,j,k(47) \tag{47}$$

$$\sum_{c=1}^{g} y_c \geq 1(48) \tag{48}$$

$$\frac{x_{ijkb}+y_c-1}{2} \leq w_{ijkbc} \leq \frac{x_{ijkb}+y_c}{2}, \forall i,j,k,b,c \tag{49}$$

$$x_{ijkb}, y_c, w_{ijkbc} \in \{0,1\}, \ \forall i,j,k,b,c. \tag{50}$$

Let

$$\alpha_{50} = \frac{z^*_{1.1D}}{\left(z^*_{1.2D}\right)^2}, \tag{51}$$

and we set

$$\alpha \ = \frac{\eta}{100-\eta} \ \alpha_{50}. \tag{52}$$

The whole procedures to solve the optimal solution to the discrete SQEOD is given in Algorithm 2.

Algorithm 2. For Solving Discrete SQEOD Models

Input parameters: $v_0, \ s_{total}, \ s_{min}, \ v_{min}, \ mse_{max}$ and η.
Output: optimal design selections $(\boldsymbol{x}^*_{SMD}, \boldsymbol{y}^*_{SMD})$.

```
1 Solve [SPD2] and get $z^*_{1.2D}$. If there is no solution, then the
  problem is not feasible and STOP.
2 Solve [SPD1] and get $z^*_{1.1D}$.
3 Calculate $\alpha_{50}$ and $\alpha$ by (51) and (52).
4 Solve [SMD], and RETURN the optimal solution (i.e., the
  option selections for the silicon area design and the low-level
  voltage design) $(\boldsymbol{x}^*_{SMD}, \boldsymbol{y}^*_{SMD})$. STOP.
```

4.3 Hybrid Memory Design Without Area Overhead

As technology continually advances, hybrid memory structures have drawn increasing attentions [12], [13]. As discussed in Section 2.2, a great advantage of using hybrid structure is that it can provide a more flexible and intelligent energy-quality-cost tradeoff compared with traditional single-technology structure. We assume that (i) area overhead does not exist in the hybrid structure. In other words, the total silicon area of a memory equals the sum of silicon areas of all bitcells; and that (ii) all bitcells have the same supply voltage options.

One can use Algorithm 2 to handle the optimal design of such hybrid structures, by including all area-options in models [SMD], [SPD1], and [SPD2]. For example, consider two memory structures, where the first has d_1 options for s_{ijk} and the second structure has d_2 options. Then we can define $d \ = d_1 \ + d_2$, where the first d_1 options stand for structure one and the following d_2 options stand for structure two. A numerical example of hybrid memory design is included in Section 5.2.

5 Numerical Studies

In the following we investigate two numerical studies on the continuous and discrete models for the proposed video memory design using two-step method and SQEOD models. In our numerical studies, we assume that all bit cells have the same design, which is the situations for almost all practical memory designs. We can hence remove the i, j subscripts in the proposed models.

5.1 Numerical Study 1: Continuous Model

We study a 6T SRAM video memory storage design, where the standard power supply $v_0 \ = \ 0.75$V and $v_{min} \ = \ 0.5$V. The data used in this numerical study is shown in Table 1. In the fifth column of the table, a normalized "area ratio" is calculated so that we can conveniently set $s_{min} \ = \ 1$, where the real area of C61, 0.685 μm^2, is considered as the basis of the normalization. For example, the area ratio of C62 is 1.053, calculated by $0.721/0.685 \ = \ 1.053$. The failure rates in the table are obtained by HSPICE Monte Carlo simulations using a 45

TABLE 1
45 nm 6T Data for Fitting, Numerical Study 1

Type	Height (μm)	Width (μm)	Area (μm^2)	Area ratio	Voltage \ Failure rates										
					0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
C61	0.45	1.523	0.685	1.000	0.3436	0.2885	0.2527	0.2243	0.19790	0.172400	0.15360	0.13750	0.12740	0.11785	0.11140
C62	0.45	1.603	0.721	1.053	0.2771	0.2355	0.2019	0.1703	0.14110	0.110000	0.08350	0.05980	0.04380	0.02521	0.01025
C63	0.45	1.698	0.764	1.115	0.2311	0.1929	0.1597	0.1269	0.09570	0.066500	0.04390	0.02380	0.00720	0.00125	5e-5
C64	0.45	1.758	0.791	1.154	0.2107	0.1737	0.1409	0.1087	0.07970	0.052200	0.02960	0.01240	0.00220	1.25e-4	n/a
C65	0.45	1.848	0.831	1.213	0.1822	0.1478	0.1163	0.0858	0.05819	0.034230	0.01598	0.00433	0.00047	n/a	n/a
C66	0.45	1.938	0.872	1.273	0.1611	0.1276	0.0967	0.0678	0.04239	0.022225	0.00762	0.00104	2.5e-5	n/a	n/a
C67	0.45	2.008	0.903	1.319	0.1474	0.1150	0.0847	0.0569	0.03337	0.015450	0.00457	0.00039	n/a	n/a	n/a
C68	0.45	2.087	0.939	1.371	0.1295	0.0982	0.0697	0.0450	0.02459	0.009545	0.00197	0.00006	n/a	n/a	n/a
C69	0.45	2.148	0.966	1.411	0.1179	0.0873	0.0607	0.0379	0.01939	0.006740	0.00106	0.00005	n/a	n/a	n/a
C610	0.45	2.218	0.998	1.456	0.1085	0.0791	0.0537	0.0318	0.01488	0.004415	0.00043	n/a	n/a	n/a	n/a
C611	0.45	2.288	1.029	1.502	0.0961	0.0683	0.0447	0.0254	0.01106	0.002640	0.00019	n/a	n/a	n/a	n/a
C612	0.45	2.358	1.061	1.548	0.0867	0.0602	0.0380	0.0201	0.00773	0.001470	0.00006	n/a	n/a	n/a	n/a
C613	0.45	2.438	1.097	1.601	0.0760	0.0520	0.0318	0.0157	0.00514	0.000790	0.00003	n/a	n/a	n/a	n/a
C614	0.45	2.518	1.133	1.654	0.0690	0.0453	0.0263	0.0122	0.00375	0.000466	1e-6	n/a	n/a	n/a	n/a
C615	0.45	2.588	1.164	1.700	0.0607	0.0395	0.0218	0.0092	0.00229	0.000300	n/a	n/a	n/a	n/a	n/a
C616	0.45	2.668	1.200	1.752	0.0544	0.0338	0.0181	0.0073	0.00172	0.000146	n/a	n/a	n/a	n/a	n/a
C617	0.45	2.738	1.232	1.798	0.0475	0.0284	0.0141	0.0052	0.00097	0.000060	n/a	n/a	n/a	n/a	n/a
C618	0.45	2.828	1.272	1.857	0.0419	0.0242	0.0120	0.0040	0.00066	0.000020	n/a	n/a	n/a	n/a	n/a
C619	0.45	2.898	1.304	1.903	0.0369	0.0206	0.0097	0.0029	0.00036	0.000010	n/a	n/a	n/a	n/a	n/a
C620	0.45	2.978	1.340	1.956	0.0325	0.0175	0.0078	0.0024	0.00027	0.000003	n/a	n/a	n/a	n/a	n/a
C621	0.45	3.058	1.376	2.008	0.0288	0.0152	0.0065	0.0018	0.00015	0.000002	n/a	n/a	n/a	n/a	n/a

nm CMOS technology [27], where "n/a" indicates that after one million runs no memory failure can be found. Fig. 3 shows the relationship between the failure rate and the voltage for memories C61, C611 and C621. One can see that the failure rate of a memory is monotone decreasing with respect to voltage and area (by comparing the failure rate of C61, C611 and C621 under the same voltage).

Borrowing the idea from [28], we fit the failure rate function $f(s, v)$ using

$$f(s,v)=\exp(\beta_1 s+\beta_2 v^3+\beta_3 v^2+\beta_4 v+\beta_5 v^3 s+\beta_6 v^2 s+\beta_7 v s+\beta_8), \tag{53}$$

where $\beta_1, \ldots,\ \beta_8$ are the fitting parameters and the data of area ratio is used for the s. Matlab curve fitting toolbox is used for the fitting, and the regression results are shown in Table 2. We use the mean of the fitting parameters as their values in the f. Fig. 4 is plotted to compare the predicted failure rates and the real ones. The x-axis of Fig. 4 includes 11 intervals, each interval representing for one voltage: 0.5V, 0.55V , ..., 1V. In each interval, the points are the real (red) and predicted (blue) failure rates of memory types C61–C621 (if available) under the corresponding voltage.

Then, we use the $f(s, v)$ fitted by (53) as the failure rate in the models and run Algorithm 1 to solve the continuous SQEOD models using MOSEK solver, under $s_{total} =\ 9.0,\ 9.5, 10.0,$ and $10.5,$ and $\alpha\ = \alpha_{50}\ /4,\ 7\alpha_{50}/8,$ and $9\alpha_{50},$ with parameters $v_0 =\ 0.75\text{V},\ s_{min} =\ 1,\ v_{min} =\ 0.5\text{V},$ and $ms\ e_{max} =\ 129.742$. The reason that we set $ms\ e_{max} = 129.742$ is to set PSNR to be 27 dB. PSNR is an quantity to describe the quality of a video [16], calculated by

$$PSNR\ =\ 20\log_{10}255 - 10\log_{10}(MSE).$$

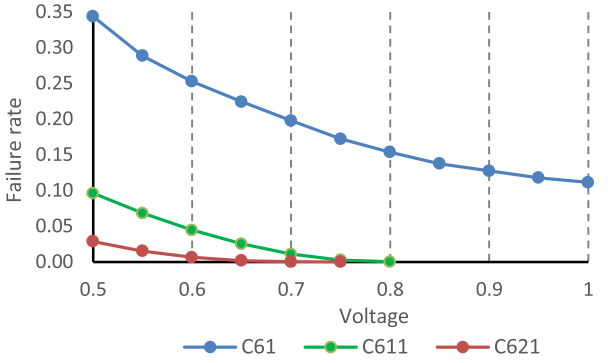


Fig. 3. Comparison of failure rates, Numerical Study 1.

The larger PSNR is, the higher quality a video has. We consider a 27-dB PSNR as an acceptable video quality.

The solutions to [SPC1] and [SPC2] are shown in Table 3, where s_7–s_0 refer to the 8 bitcells to store a pixel: s_7 standing for silicon area of the most-significant-bit (MSB) and s_0 for the least-significant-bit (LSB). It can be seen from Table 3 that at $s_{total} =\ 9.0$ in [SPC2] there is no feasible solution (denoted as "n/a"). This is because the minimum expected MSE under this case is 201.264, which is still greater than the mse_{max}. The solutions to [SMC] are included in Table 4, where second column is the optimal value of the SQEOD objective function and the last column incldes the expected MSE under the standard supply voltage $\ v_0$. We also get the solutions to the two-step method (i.e., problems [2MC1] and [2MC2]) using MOSEK solver, and show them in Table 5.

Comparisons of the solutions to SQEOD (under $\alpha\ = \alpha_{50}\ /4, 7\alpha_{50}/8,\ 9\alpha_{50})$ and the two-step method are given in Fig. 6. One can see that under $\alpha\ = \alpha_{50}\ /4$ and $7\alpha_{50}/8$, the SQEOD MSEs at the standard voltage v_0 and the low-level power supply v^* are almost the same with those of the two-step method. By contrast, under $\alpha\ =\ 9\alpha_{50}$, the SQEOD gets to a solution that has a better low-level power supply v^* but a worse MSE under standard power supply v_0. This reflects the idea of SQEOD in which the weight parameter α controls the importance of the two objectives: (i) to minimize MSE when the power level is standard, and (ii) to minimize supply voltage when the power level is low. The larger α is, the more important the second objective will be in the SQEOD. And that is also the reason when $\alpha\ =\ 9\alpha_{50}$ the SQEOD solution

TABLE 2
Statistics of the Fitting, Numerical Study 1

Parameter	Mean	95% confidence interval	Parameter	Mean	95% confidence interval
β_1	47.72	(34.95, 60.50)	β_5	−307.50	(−361.90, −253.10)
β_2	308.70	(253.90, 363.50)	β_6	488.70	(386.90, 590.40)
β_3	−490.00	(−592.80, −387.20)	β_7	−267.80	(−330.50, −205.00)
β_4	265.50	(201.90, 329.20)	β_8	−47.56	(−60.56, −34.55)
R-square: 99.79%				Adjusted R-square: 99.78%	

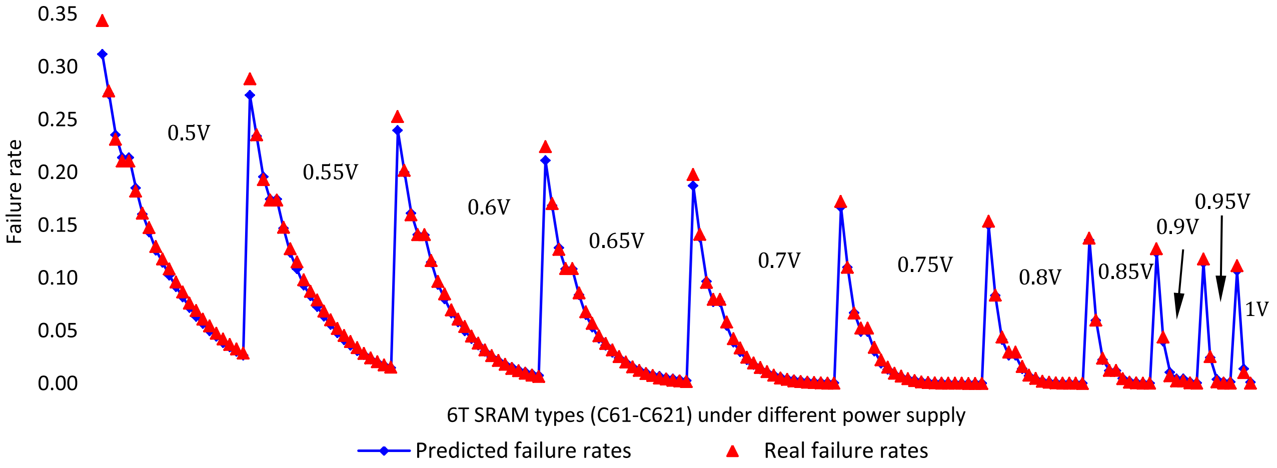


Fig. 4. Predicted failure rates versus real failure rates, Numerical Study 1.

TABLE 3
Solutions to [SPC1] and [SPC2], Numerical Study 1

				[SPC1]					
s_{total}	Obj. value (z^*_{SPC1})	s^*_7	s^*_6	s^*_5	s^*_4	s^*_3	s^*_2	s^*_1	s^*_0
9.0	201.264	1.507	1.333	1.159	1.000	1.000	1.000	1.000	1.000
9.5	83.244	1.636	1.462	1.288	1.114	1.000	1.000	1.000	1.000
10.0	38.873	1.748	1.574	1.400	1.226	1.052	1.000	1.000	1.000
10.5	19.457	1.848	1.674	1.500	1.326	1.152	1.000	1.000	1.000
				[SPC2]					
s_{total}	Obj. value (z^*_{SPC2})	s_7	s_6	s_5	s_4	s_3	s_2	s_1	s_0
9.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
9.5	0.725	1.683	1.478	1.272	1.067	1.000	1.000	1.000	1.000
10.0	0.686	1.894	1.631	1.369	1.106	1.000	1.000	1.000	1.000
10.5	0.652	2.101	1.784	1.466	1.149	1.000	1.000	1.000	1.000

performs better in v^* but worse in MSE under v_0, compared with the other two cases.

5.2 Numerical Study 2: A Discrete Model

We study a hybrid video memory storage design with optional 6T and 8T SRAM. To enable near threshold voltage operation, we assume the power supply options are 0.36V (v_{min}), 0.38V, ..., 0.48V, 0.5V (v_0). Other parameter settings are $s_{min} = 1$, $s_{total} = 8.3, 8.4, \ldots, 8.7$, and $ms\ e_{max} = 51.651$ (equivalent to $PSNR = 31$dB). The data used in this numerical study is shown in Table 6. Compared with 8T SRAM, in general, 6T SRAM are smaller in size but with relatively higher failure rates.

We apply the discrete SQEOD model [SMD] (associated with [SPD1] and [SPD2]), where there are $d = 6 + 4 = 10$ options for the area design (i.e., C81–C86 for 8T SRAM and C61–C64 for 6T SRAM) and $g + 1 = 8$ options for the low-level voltage design (i.e., 0.36V, 0.38V, ..., 0.5V).

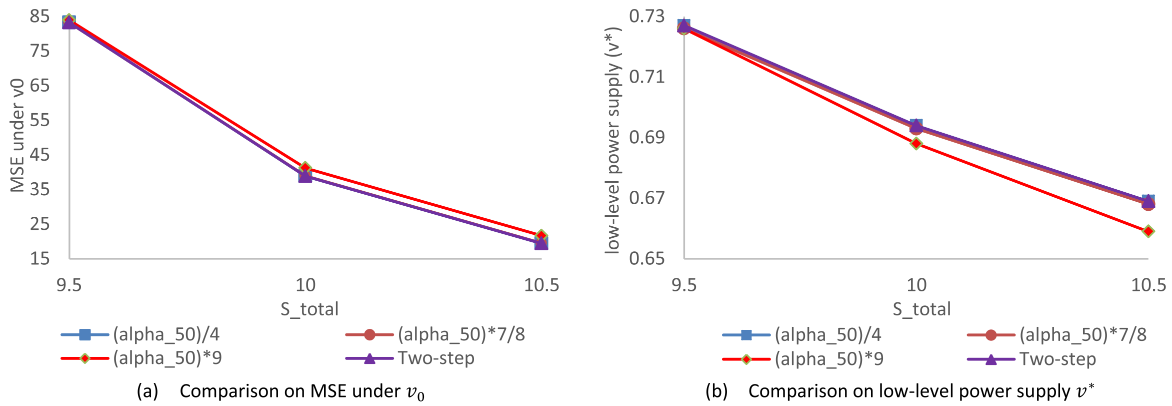
We first use Algorithm 2 to solve the [SMD] for the SQEOD design using Gurobi solver (version 7.0.2). The solutions to [SPD1] and [SPD2] are shown in Table 7. Based on these solutions, we can calculate the α_{50}, and in Table 8 we show two cases of the α settings for the SQEOD [SMD] problem: $\alpha = \alpha_{50}/4$ and $9\alpha_{50}$. Fig. 6 includes comparisons of solutions to SQEOD (under $\alpha = \alpha_{50}/4$ and $9\alpha_{50}$) and the two-step method (see Table 9 for details). The solutions to SQEOD under $\alpha = \alpha_{50}/4$ are the same with those to the two-step method. But under $\alpha = 9\alpha_{50}$ they are different. Specifically, at $s_{total} = 8.6$, compared with the two-step method, the solution to SQEOD can save 8.7% of power supply (equivalent to

TABLE 4
Solutions to [SMC] Under $\alpha = \alpha_{50}/4,\ 7\alpha_{50}/8$ and $9\alpha_{50}$, Numerical Study 1

SQEOD [SMC], under $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{50}/4$													
s_{total}	Obj. value	s_7^*	s_6^*	s_5^*	s_4^*	s_3^*	s_2^*	s_1^*	s_0^*	v^*	$\boldsymbol{\alpha}_{50}$	$\boldsymbol{\alpha}$	$E(MSE)$ under v_0
9.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
9.5	104.145	1.637	1.462	1.287	1.113	1.000	1.000	1.000	1.000	0.727	158.371	39.593	83.245
10.0	48.816	1.751	1.575	1.399	1.224	1.05	1.000	1.000	1.000	0.694	82.604	20.651	38.878
10.5	24.576	1.853	1.675	1.499	1.324	1.149	1.000	1.000	1.000	0.669	45.770	11.442	19.460
SQEOD [SMC], under $\boldsymbol{\alpha} = 7\boldsymbol{\alpha}_{50}/8$													
s_{total}	Obj. value	s_7^*	s_6^*	s_5^*	s_4^*	s_3^*	s_2^*	s_1^*	s_0^*	v^*	$\boldsymbol{\alpha}_{50}$	$\boldsymbol{\alpha}$	$E(MSE)$ under v_0
9.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
9.5	156.383	1.64	1.463	1.286	1.110	1.000	1.000	1.000	1.000	0.726	158.371	138.575	83.263
10.0	73.635	1.759	1.577	1.398	1.221	1.044	1.000	1.000	1.000	0.693	82.604	72.278	38.922
10.5	37.346	1.863	1.678	1.497	1.319	1.143	1.000	1.000	1.000	0.668	45.770	40.049	19.495
SQEOD [SMC], under $\boldsymbol{\alpha} = 9\boldsymbol{\alpha}_{50}$													
s_{total}	Obj. value	s_7^*	s_6^*	s_5^*	s_4^*	s_3^*	s_2^*	s_1^*	s_0^*	v^*	$\boldsymbol{\alpha}_{50}$	$\boldsymbol{\alpha}$	$E(MSE)$ under v_0
9.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
9.5	834.447	1.661	1.469	1.279	1.091	1.000	1.000	1.000	1.000	0.726	158.371	1425.343	83.923
10.0	388.835	1.825	1.600	1.388	1.187	1.000	1.000	1.000	1.000	0.688	82.604	743.434	41.219
10.5	200.717	1.964	1.708	1.479	1.271	1.078	1.000	1.000	1.000	0.659	45.770	411.930	21.655

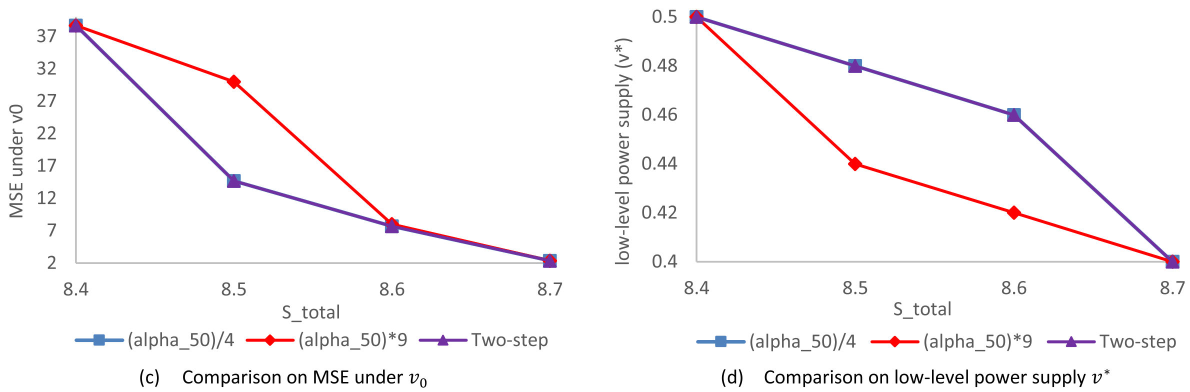
TABLE 5
Solutions to the Continuous Two-Step Method, Numerical Study 1

Two-step method, [2MC1] & [2MC2]									
s_{total}	s_7^*	s_6^*	s_5^*	s_4^*	s_3^*	s_2^*	s_1^*	s_0^*	v^*	$E(MSE)$ under v_0
9.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
9.5	1.636	1.462	1.288	1.114	1.000	1.000	1.000	1.000	0.727	83.244
10.0	1.748	1.574	1.4	1.226	1.052	1.000	1.000	1.000	0.694	38.873
10.5	1.848	1.674	1.500	1.326	1.152	1.000	1.000	1.000	0.669	19.457



(a) Comparison on MSE under v_0 (b) Comparison on low-level power supply v^*

Fig. 5. Comparisions of solutions to SQEOD (under $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{50}/4, 7\boldsymbol{\alpha}_{50}/8,\ 9\boldsymbol{\alpha}_{50}$) and the two-step method, Numerical Study 1.



(c) Comparison on MSE under v_0 (d) Comparison on low-level power supply v^*

Fig. 6. Comparisions of solutions to SQEOD (under $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{50}/4$ and $9\boldsymbol{\alpha}_{50}$) and the two-step method, Numerical Study 2.

TABLE 6
45 nm 8T and 6T Data for Numerical Study 2

Type	Height (μm)	Width (μm)	Area (μm^2)	Area ratio	Voltage \ Failure rates							
					0.36	0.38	0.40	0.42	0.44	0.46	0.48	0.50
C81	0.45	1.669	0.685	1.000	0.0185	0.0153	0.0121	0.0083	0.0050	0.0035	0.0018	0.00082
C82	0.45	1.681	0.721	1.053	0.0137	0.0105	0.0076	0.0048	0.0033	0.0020	0.0009	0.00031
C83	0.45	1.700	0.764	1.115	0.0097	0.0073	0.0043	0.0032	0.0020	0.0011	0.0003	0.00009
C84	0.45	1.720	0.791	1.154	0.0067	0.0044	0.003	0.0021	0.0012	0.0004	0.0002	0.00006
C85	0.45	1.740	0.831	1.213	0.0046	0.0028	0.002	0.0012	0.0005	0.0002	0.00006	0.00002
C86	0.45	1.760	0.872	1.273	0.00303	0.00192	0.00122	0.0007	0.0003	0.00007	0.00003	0.00001
C61	0.45	1.523	0.685	1.000	0.7253	0.6551	0.5897	0.5234	0.4669	0.4190	0.3773	0.3436
C62	0.45	1.563	0.703	1.026	0.6757	0.6056	0.5341	0.4678	0.4129	0.3696	0.3352	0.3074
C63	0.45	1.603	0.721	1.053	0.6318	0.5539	0.4803	0.4195	0.3720	0.3318	0.3010	0.2771
C64	0.45	1.6425	0.739	1.079	0.5818	0.5041	0.4342	0.3778	0.3344	0.2992	0.2730	0.2521

TABLE 7
Solutions to [SPD1] and [SPD2], Numerical Study 2

[SPD1]									
s_{total}	Obj. value (z^*_{SPD1})	s^*_7	s^*_6	s^*_5	s^*_4	s^*_3	s^*_2	s^*_1	s^*_0
8.3	126.445	C82	C81	C81	C61	C61	C61	C61	C61
8.4	38.693	C82	C81	C81	C81	C61	C61	C61	C61
8.5	14.666	C82	C82	C81	C81	C81	C61	C61	C61
8.6	7.666	C83	C81	C81	C81	C81	C81	C61	C61
8.7	2.365	C86	C84	C83	C82	C81	C81	C61	C61
[SPD2]									
s_{total}	Obj. value (z^*_{SPD2})	s_7	s_6	s_5	s_4	s_3	s_2	s_1	s_0
8.3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
8.4	0.50	C82	C81	C81	C81	C61	C61	C61	C61
8.5	0.44	C86	C85	C82	C81	C61	C61	C61	C61
8.6	0.42	C85	C83	C83	C83	C81	C61	C61	C61
8.7	0.40	C86	C84	C83	C82	C82	C61	C61	C61

TABLE 8
Solutions to [SMD] Under $\alpha = \alpha_{50}/4$ and $9\alpha_{50}$, Numerical Study 2

SQEOD [SMD], under $\alpha = \alpha_{50}/4$													
s_{total}	Obj. value	s^*_7	s^*_6	s^*_5	s^*_4	s^*_3	s^*_2	s^*_1	s^*_0	v^*	α_{50}	α	E(MSE) under v_0
8.3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
8.4	48.366	C82	C81	C81	C81	C61	C61	C61	C61	0.50	154.772	38.693	38.693
8.5	19.029	C82	C82	C81	C81	C81	C61	C61	C61	0.48	75.754	18.939	14.666
8.6	9.965	C83	C81	C81	C81	C81	C81	C61	C61	0.46	43.458	10.865	7.666
8.7	2.956	C85	C85	C83	C82	C81	C81	C61	C61	0.40	14.781	3.695	2.365
SQEOD [SMD], under $\alpha = 9\alpha_{50}$													
s_{total}	Obj. value	s^*_7	s^*_6	s^*_5	s^*_4	s^*_3	s^*_2	s^*_1	s^*_0	v^*	α_{50}	α	E(MSE) under v_0
8.3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
8.4	386.930	C82	C81	C81	C81	C61	C61	C61	C61	0.50	154.772	1392.948	38.693
8.5	161.973	C86	C85	C82	C81	C61	C61	C61	C61	0.44	75.754132	681.787	29.979
8.6	76.974	C85	C85	C83	C81	C81	C61	C61	C61	0.42	43.45805	391.122	7.980
8.7	23.650	C86	C84	C83	C82	C81	C81	C61	C61	0.40	14.78125	133.031	2.365

TABLE 9
Solutions to the Discrete Two-Step Method, Numerical Study 2

Two-step method, [2MD1] & [2MD2]										
s_{total}	s^*_7	s^*_6	s^*_5	s^*_4	s^*_3	s^*_2	s^*_1	s^*_0	v^*	E(MSE) under v_0
8.3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
8.4	C82	C81	C81	C81	C61	C61	C61	C61	0.50	38.693
8.5	C82	C82	C81	C81	C81	C61	C61	C61	0.48	14.666
8.6	C83	C81	C81	C81	C81	C81	C61	C61	0.46	7.666
8.7	C86	C84	C83	C82	C81	C81	C61	C61	0.40	2.365

saving approximately 16.6 percent of power consumption [24]) but causes 3.93 percent of extra MSE.

6 Conclusion

Quality-aware hardware design techniques have been recently developed by adding quality as a novel dimension to traditional design space to enable energy efficiency enhancement. The main contributions of this paper are

a) developing mathematical models for the traditional two-step method, which enables the search for the exact system optimal solutions; and

b) proposing the SQEOD concept with its mathematical models, which generalizes the traditional two-step method: when the weight parameter $\alpha \rightarrow 0^+$, then the SQDOD tends to be the two-step method.

Given two objectives, (i) to minimize the MSE at the standard supply voltage and (ii) to minimize the supply voltage when the power is at a low level, the main difference between the two-step method and the SQEOD is that the former always solve objective (i) first, and then solve objective (ii) based on the solution to objective (i). This may limit the solution to the low-level power supply. By contrast, in the SQEOD, the two objectives are solved simultaneously and the importance of them, i.e., the tradeoff, is controlled by the weight parameter α.

Acknowledgments

This work was supported by National Science Foundation under Grant CCF-1855706.

References

[1] J. Panneerselvam, L. Liu, and N. Antonopoulos, "An approach to optimise resource provision with energy-awareness in datacentres by combating task heterogeneity," *IEEE Trans. Emerg. Topics Comput.*, early access, Jan. 16, 2018, doi: 10.1109/TETC.2018.2794328.

[2] P. Arroba, J. M. Moya, J. L. Ayala, and R. Buyya, "DVFS-aware consolidation for energy-efficient clouds," in *Proc. Int. Conf. Parallel Archit. Compilation*, 2015, pp. 494–495.

[3] S. Wang, Z. Qian, J. Yuan, and I. You, "A DVFS based energy-efficient tasks scheduling in a data center," *IEEE Access*, pp. 13090–13102, 2017.

[4] F. C. Fernandes, E. Faramarzi, X. Li, Z. Ma, and X. Ducloux, "Mobile display power reduction for video using standardized metadata," *IEEE Trans. Mobile Comput.*, vol. 18, no. 1, pp. 165–178, Jan. 2019.

[5] Z. Abbas and W. Yoon, "A survey on energy conserving mechanisms for the Internet of Things: Wireless networking aspects," *Sensors*, vol. 15, no. 10, pp. 24818–24847, 2015.

[6] T.-M. Liu, T.-A. Lin, S.-Z. Wang, W.-P. Lee, J.-Y. Yang, and K.-C. Hou, "A 125 uW, fully scalable MPEG-2 and H.264/AVC video decoder for mobile applications," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 161–169, Jan. 2007.

[7] M. Cho, J. Schlessman, W. Wolf, and S. Mukhopadhyay, "Reconfigurable SRAM architecture with spatial voltage scaling for low power mobile multimedia applications," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 19, no. 1, pp. 161–165, Jan. 2011.

[8] F. Sampaio, M. Shafique, B. Zatt, S. Bampi, and J. Henkel, "Energy-efficient architecture for advanced video memory," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2014, pp. 132–139.

[9] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, "DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," *ACM Sigplan Notices*, vol. 49, no. 4, pp. 269–284, 2014.

[10] M. Alioto, "Energy-quality scalable adaptive VLSI circuits and systems beyond approximate computing," in *Proc. Des. Autom. Test Europe Conf. Exhib.*, 2017, pp. 127–132.

[11] Y. Xu, H. Das, Y. Gong, and N. Gong, "On mathematical models of optimal video memory design," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 256–266, Jan. 2020.

[12] I. Chang, D. Mohapatra, and K. Roy, "A priority-based 6T/8T hybrid SRAM architecture for aggressive voltage scaling in video applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 2, pp. 101–112, Feb. 2011.

[13] N. Gong, S. Jiang, A. Challapalli, S. Fernandes, and R. Sridhar, "Ultra-low voltage split-data-aware embedded SRAM for mobile video applications," *IEEE Trans. Circuits Syst. II, Express Briefs*, vol. 59, no. 12, pp. 883–887, Dec. 2012.

[14] A. Kazimirsky, A. Teman, N. Edri, and A. Fish, "A 0.65-V, 500-MHz integrated dynamic and static RAM for error tolerant applications," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 25, no. 9, pp. 2411–2418, Sep. 2017.

[15] J. Kwon, I. Lee, and J. Park, "Heterogeneous SRAM cell sizing for low power H.264 applications," *IEEE Trans. Circuits Syst. I, Regular Papers*, vol. 99, no. 2, pp. 1–10, Feb. 2012.

[16] J. Edstrom, D. Chen, Y. Gong, J. Wang, and N. Gong, "Data-pattern enabled self-recovery low-power storage system for big video data," *IEEE Trans. Big Data*, vol. 5, no. 1, pp. 95–105, Mar. 2019.

[17] J. Edstrom, Y. Gong, D. Chen, J. Wang, and N. Gong, "Data-driven intelligent efficient synaptic storage for deep learning," *IEEE Trans. Circuits Syst. II, Express Briefs*, vol. 64, no. 12, pp. 1412–1416, Dec. 2017.

[18] N. Gong, S. A. P. Pourbakhsh, X. C. Chen, X. Wang, D. Chen, and J. Wang, "SPIDER: Sizing-priority-based application-driven memory for mobile video applications," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 25, no. 9, pp. 2625–2634, Sep. 2017.

[19] H. Kim, I. J. Chang, and H.-J. Lee, "Optimal selection of SRAM bit-cell size for power reduction in video compression," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 3, pp. 431–443, Sep. 2018.

[20] K. Kushida *et al.*, "A 0.7 V single-supply SRAM with 0.495 cell in 65 nm technology utilizing self-write-back sense amplifier and cascaded bit line scheme," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1192–1198, Apr. 2009.

[21] H. Noguchi *et al.*, "A 10T non-precharge two-port SRAM for 74% power reduction in video processing," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, 2007, pp. 107–112.

[22] F. Abouzeid *et al.*, "Scalable 0.35 V to 1.2 V SRAM bitcell design from 65 nm CMOS to 28 nm FDSOI," *IEEE J. Solid-State Circuits*, vol. 49, no. 7, pp. 1499–150, Jul. 2014.

[23] L. Atias, A. Teman, R. Giterman, P. Meinerzhagen, and A. Fish, "A low-voltage radiation-hardened 13T SRAM bitcell for ultralow power space applications," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 24, no. 8, pp. 2622–2632, Aug. 2016.

[24] B. Zhai, R. G. Dreslinksk, D. Blaauw, T. Mudge, and D. Sylvester, "Energy efficient near-threshold chip multi-processing," in *Proc. Int. Symp. Low Power Electron. Des.*, 2007, pp. 32–37.

[25] E. Morgin, M. Maman, R. Guizzetti, and A. Duda, "Comparison of the device lifetime in wireless networks for the Internet of Things," *IEEE Access*, vol. 5, pp. 7097–7114, 2017.

[26] "New smart energy technology boosts battery life for IoT devices," Accessed: Dec. 12, 2018. [Online]. Available: https://www.maximintegrated.com/content/dam/files/design/technical-documents/white-papers/new-smart-energy-technology-boosts-battery-life-for-iot-devices.pdf

[27] "FreePDK," Accessed: Dec. 1, 2017. [Online]. Available: https://www.eda.ncsu.edu/wiki/FreePDK

[28] X. Li, "Maximum-information storage system: Concept, implementation and application," in *Proc. Int. Conf. Comput.-Aided Design*, 2010, pp. 39–46.

[29] M. Alioto, "Energy-quality scalable adaptive VLSI circuits and systems beyond approximate computing," in *Proc. Des. Autom. Test Europe Conf. Exhib.*, 2017, pp. 127–132.

Yiwen Xu received the PhD degree in systems and industrial engineering from the University of Arizona. He is currently an assistant professor with the Department of Industrial and Manufacturing Engineering at North Dakota State University, Fargo, ND. His research interests include applied operations research (especially probabilistic network optimization and applied integer programming) and reliability engineering.

Hritom Das (Student Member, IEEE) received the BS degree in electrical and electronic engineering from American International University-Bangladesh, Dhaka, Bangladesh, in 2012, and the MS degree in electronic engineering from Kyungpook National University, Daegu, South Korea, in 2015. He is currently working toward the PhD degree in electrical and computer engineering at North Dakota State University, Fargo, ND, USA. From 2016 to 2017, he was a faculty member (lecturer) with Uttara University, Uttara, Bangladesh. His research interests include the low power circuit design, testing, and machine learning implementation in traditional electronics.

Na Gong (Member, IEEE) received the BE degree in electrical engineering, the ME degree in microelectronics from Hebei University, Hebei, China, and the PhD degree in computer science and engineering from the State University of New York, Buffalo, in 2004, 2007, and 2013, respectively. Currently, she is an associate professor of electrical and computer engineering with the University of South Alabama, Mobile, AL, USA. Her research interests include power-efficient computing circuits and systems, video memory optimization, and neuromorphic hardware.