Maximum-Information Storage System: Concept, Implementation and Application

Xin Li
Electrical & Computer Engineering Department, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
xinli@ece.cmu.edu

ABSTRACT

The aggressive technology scaling has made it increasingly difficult to design high-performance, high-density SRAM circuits. In this paper, we propose a new SRAM design methodology that is referred to as maximum-information storage system (MISS). Unlike most traditional SRAM circuits that are designed for maximum cell density, MISS aims to maximize the information density (i.e., the number of information bits per unit area). Towards this goal, an information model is derived to quantitatively measure the information bits stored in a given SRAM system. In addition, a convex optimization framework is developed to optimize SRAM cells to achieve maximum information storage. Our design example in a commercial 65nm CMOS process demonstrates that MISS achieves more than 3.5× area reduction over the traditional SRAM design, while storing the same amount of information. Furthermore, two real-life signal processing examples show that given the same area constraint, MISS can increase signal-to-noise ratio by more than 30 dB compared to the traditional SRAM system.

1. INTRODUCTION

On-chip embedded storage device (i.e., SRAM) is a critical component that plays an important role in defining the overall system performance of today's large-scale integrated circuits [1]. An SRAM bit cell is typically designed with minimum-size transistors in order to minimize silicon area. However, these small transistors make SRAM extremely sensitive to large-scale process variations (e.g., random dopant fluctuations) posed by nanoscale IC technology [2]-[4]. For this reason, SRAM design has been identified as one of the major bottlenecks for future IC technology scaling.

To address this technical challenge, a large number of statistical analysis and optimization algorithms have been proposed to facilitate robust SRAM design at advanced technology nodes [5]-[14]. The key idea is to accurately predict both symmetric and random variations for SRAM circuits so that design margins can be minimized to improve performance and/or reduce area. While these existing techniques rely on different statistical algorithms, all of them share the same design goal:

- **High cell robustness**: Each SRAM cell is designed with nearly zero failure probability in order to guarantee high parametric yield for the entire SRAM system.
- **High cell density**: Subject to the aforementioned robustness constraint, each SRAM cell is designed with minimum area in order to maximize cell density.

These two objectives have been considered as the "golden standard" and have never been changed during the past several decades. However, as the non-idealities at nanoscale technology pose enormous challenges for SRAM design, they also suggest an immediate need to re-think this fundamental design strategy in order to meet today's manufacturing reality.

In this paper, we propose a completely new design methodology, referred to as Maximum-Information Storage System (MISS), for SRAM circuits. The key idea is not to maximize the traditional cell density that is measured by the number of SRAM cells per unit area. Instead, we propose to maximize the *information density* (i.e., the number of information bits per unit area). Note that these two density metrics are equivalent, if and only if all SRAM cells have zero failure probability. In this case, one SRAM cell stores one bit of information. However, as each SRAM cell can possibly fail with nanoscale manufacturing technology, the proposed information density is substantially different from the traditional cell density. It offers a radically new paradigm for optimal SRAM design.

The proposed information density measures the amount of information stored in a unit-area SRAM system. Maximum-information storage cannot be achieved by simply maximizing cell robustness. Note that a reduced failure rate of SRAM cell always comes with an area penalty, e.g., by increasing transistor size or by adding extra redundancy. If SRAM cells are over-designed to achieve nearly zero failure probability, only few bit cells (hence, only few information bits) can be stored within a unit area. It, in turn, fails to offer maximum-information storage. In many *application-specific* cases, zero failure probability is not required. As will be demonstrated by the signal processing examples in Section 6, a number of "unimportant" SRAM cells can fail to work and they have negligible impact on the final signal-to-noise ratio.

On the other hand, continually increasing the number of SRAM cells within a unit area does not lead to maximum-information storage either. If an SRAM cell is designed with small-size transistors, it results in a high failure probability. In this case, the information stored in SRAM is not maximized, as a failed bit cell cannot store any information. These observations imply an important fact that the traditional SRAM design strategy for maximum cell robustness/density may not be optimal. The challenging issue here is how to optimally design the proposed MISS system (e.g., determine the silicon area and failure probability for each SRAM cell) to achieve *maximum information density*.

Towards this goal, a number of new CAD algorithms and design methodologies are developed in this paper to facilitate optimal MISS design. First, an analytical model is derived from information theory [21], [23] to quantitatively measure the information bits stored in a given SRAM system where each bit cell is subject to a given failure probability. Such an information model enables us to quickly compare different SRAM designs and assess their optimality based on information density.

Second, a convex optimization framework is developed to optimize the SRAM system to achieve maximum-information storage. Based on this optimal design methodology, the performance improvement offered by MISS is compared to the traditional SRAM design. While this paper does not focus on chip

978-1-4244-8192-7/10/$26.00 ©2010 IEEE

tape-out and several detailed design issues are not explicitly discussed, our quantitative analyses in Section 5 show extremely promising results. The proposed MISS system achieves more than 3.5× area reduction over the traditional SRAM design, while storing the same amount of information.

Third, to fully demonstrate the efficacy of MISS in real-life applications, two signal processing examples are extensively studied. Our experimental results in Section 6 demonstrate that given the same area constraint, the proposed MISS system is able to increase signal-to-noise ratio by more than 30 dB, compared to the traditional SRAM design. It, in turn, provides strong evidence to support the bold move from the traditional design with maximum cell robustness/density to the proposed design with maximum information density.

The remainder of this paper is organizing as follows. In Section 2, we briefly summarize the background of information theory, and then derive the information model for the proposed MISS system in Section 3. The detailed design methodology for MISS is discussed in Section 4, and a 65nm design example is shown in Section 5. Two signal processing applications are further studied in Section 6 to demonstrate the efficacy of MISS. Finally, we conclude in Section 7.

2. BACKGROUND

In this section, we briefly summarize the basic background on information theory. The concepts and theorems introduced here will be further used to derive the information model for SRAM circuits in Section 3.

Definition 1: The *differential entropy* $H(x)$ of a continuous random variable x with probability density function $p(x)$ is defined as [21]:

$$H(x) = -\int_{-\infty}^{+\infty} p(x) \cdot \log_2 p(x) \cdot dx \cdot \tag{1}$$

The differential entropy $H(x)$ depends on the distribution $p(x)$. It measures the uncertainty (or equivalently, information) that the random variable x carries. If the entropy $H(x)$ is large, the random variable x is highly uncertain and a large amount of information can be obtained by knowing its value.

Based on Definition 1, it is easy to prove the following theorem for differential entropy.

Theorem 1: If the continuous probability density function $p(x)$ is scaled by a factor of k, the differential entropy will be increased by $\log_2|k|$ [21]:

$$H(k \cdot x) = H(x) + \log_2 |k| \cdot \tag{2}$$

In other words, the differential entropy $H(x)$ increases, if the variance of $p(x)$ increases. This result is consistent with our intuition. Namely, a random variable with large variance is highly uncertain and, hence, it has large entropy.

In addition to variance, the differential entropy $H(x)$ also depends on the distribution of x. Even if two distributions $p(x)$ and $p(y)$ have the same variance, their entropy values $H(x)$ and $H(y)$ can be different. The following theorem proves that given a fixed variance, Gaussian distribution has the largest entropy value.

Theorem 2: Among all continuous probability density functions where the mean is μ and the variance is σ^2, the Gaussian distribution $x \sim N(u, \sigma^2)$ has the largest entropy value [23]:

$$H(x) = \frac{1}{2} \cdot \log_2 \left(2\pi \cdot e \cdot \sigma^2\right) \cdot \tag{3}$$

The differential entropy $H(x)$ measures the information carried by a single random variable x. To study an information storage system, we must simultaneously consider two random variables: (1) the original data x, and (2) the stored (probably distorted) data y. In this case, we need to introduce the concept of conditional differential entropy.

Definition 2: If x and y are two continuous random variables with joint probability density function $p(x, y)$ and conditional probability density function $p(x \mid y)$, the *conditional differential entropy* is defined as [21]:

$$H(x \mid y) = -\int_{-\infty}^{+\infty} p(x, y) \cdot \log_2 p(x \mid y) \cdot dx \cdot dy \cdot \tag{4}$$

The conditional entropy $H(x \mid y)$ measures the uncertainty of x conditioned on y. Namely, it tells us how much extra information is carried by x, if y is already known.

Given Definition 2, it is straightforward to prove the following theorem.

Theorem 3: Given two continuous random variables x and y, the following properties hold for conditional differential entropy [21]:

$$H(x \mid y) \le H(x) \tag{5}$$

$$H(x + y \mid y) = H(x \mid y) \tag{6}$$

where the equality in (5) holds when x and y are independent.

Intuitively, Eq. (5) simply means that knowing y helps to reduce the uncertainty of x, if x and y are correlated. On the other hand, Eq. (6) implies that once y is known, adding it to x does not introduce any extra uncertainty.

Based on the definition of differential entropy and conditional differential entropy, we are now ready to define the mutual information between two random variables.

Definition 3: If x and y are two continuous random variables, the *mutual information* between x and y is defined as [21]:

$$I(x, y) = H(x) - H(x \mid y) \cdot \tag{7}$$

Studying (7), we would notice that the mutual information $I(x, y)$ is equal to the difference between $H(x)$ (i.e., the information of x) and $H(x \mid y)$ (i.e., the extra information carried by x given a known y). In other words, $I(x, y)$ measures the information of x that we can learn from y. It is the key mathematical tool that we will use to design our proposed MISS system. Intuitively, if x represents the original data and y stands for the distorted data stored in an SRAM, the goal of optimal MISS design is to maximize the mutual information $I(x, y)$ so that we can extract as much information as possible for x by knowing y.

3. INFORMATION MODEL

In this section, we derive an analytical model to quantitatively measure the information stored in an SRAM system where each bit cell is subject to a given failure probability. Such an information model allows us to quickly compare different SRAM designs based on their information density, and it will be incorporated into our convex optimization framework in Section 4 to optimally design the proposed MISS system.

3.1 Mathematical Formulation

We consider an SRAM system that is particularly designed for the data cache of signal processing applications. In this case, the SRAM cells are used to store numerical data for signal processing algorithms.

Without loss of generality, we assume that a signal x is real-valued and it is within the interval [0, 1]. Any real-valued signal can be mapped to this interval after appropriate shifting and scaling. We represent the signal x by a set of binary digits $\{x_n;\ n = 1,2,\ldots\}$ where $x_n \in \{0, 1\}$:

$$x = \sum_{n=1}^{+\infty} 2^{-n} \cdot x_n \,. \tag{8}$$

In general, since x is real-valued, an infinite number of binary digits are required to exactly represent x. The upper bound of the summation in (8) is $n = +\infty$.

In practice, we always use a finite number of (say, N) digits $\{x^Q_n;\ n = 1,2,\ldots,N\}$, where $x^Q_n \in \{0, 1\}$, to approximate x. Such an approximation is referred to as quantization in digital signal processing [18]:

$$x^Q = \sum_{n=1}^{N} 2^{-n} \cdot x^Q_n \tag{9}$$

$$x = x^Q + \varepsilon \tag{10}$$

where x^Q represents the quantized signal, ε denotes the quantization noise, and the superscript "Q" stands for "quantization". In (9), each digit x^Q_n is associated with a unique weight 2^{-n}. The first digit x^Q_1 is referred to as the most significant bit (MSB), as it corresponds to the largest weight 2^{-1}. On the other hand, the N-th digit x^Q_N is referred to the least significant bit (LSB), as it corresponds to the smallest weight 2^{-N}.

To quantitatively model the quantization noise ε, we assume that the signal x can take any value within the interval [0, 1] with equal probability. Namely, the signal x is uniformly distributed over [0, 1]. Given this assumption, it can be easily verified that each digit x^Q_n can be either 0 or 1 with equal probability [18]:

$$p\left(x^Q_n\right) = \begin{cases} 0.5 & \left(x^Q_n = 0\right) \\ 0.5 & \left(x^Q_n = 1\right) \end{cases} \quad (n = 1,2,\cdots,N) \tag{11}$$

and the quantization noise ε follows the statistics [18]:

$$\mu(\varepsilon) = 0 \tag{12}$$

$$var(\varepsilon) = \frac{1}{12} \cdot 4^{-N} \tag{13}$$

where $\mu(\bullet)$ and $var(\bullet)$ represent the mean and variance of a random variable, respectively.

While the quantization process introduces the noise term ε in (10) and, hence, causes information loss, it is not the only noise source in today's SRAM system. When the binary digits $\{x^Q_n;\ n = 1,2,\ldots,N\}$ are stored in SRAM, the stored value $\{y^Q_n;\ n = 1,2,\ldots,N\}$ can be different from $\{x^Q_n;\ n = 1,2,\ldots,N\}$, since each SRAM cell can possibly fail. In other words, the stored signal:

$$y^Q = \sum_{n=1}^{N} 2^{-n} \cdot y^Q_n \tag{14}$$

can be different from the actual quantized signal x^Q. For this reason, we need to further consider the random cell failure and model the corresponding information loss.

Towards this goal, we propose the symmetric failure model shown in Figure 1. In this model, x^Q_n denotes the n-th digit of the quantized signal x^Q. On the other hand, y^Q_n represents the stored digit in SRAM that corresponds to x^Q_n. Figure 1 shows the transition probability from x^Q_n to y^Q_n, which carries a two-fold meaning. First, the probability of getting different x^Q_n and y^Q_n (i.e., the failure rate) is equal to α_n. Second, the failure model in Figure 1 is symmetric, as the conditional probabilities $p(y^Q_n = 0 \mid x^Q_n = 1)$ and $p(y^Q_n = 1 \mid x^Q_n = 0)$ are identical. Such a symmetric failure model assumes that the failure probability is independent of the binary value stored in the SRAM cell. It is a valid assumption, if the failure event is caused by random variations and the SRAM cell has a symmetric topology (e.g., the traditional 6-T SRAM cell). Combing (11) and Figure 1, we can derive the following joint probability mass function for x^Q_n and y^Q_n:

$$p\left(x^Q_n, y^Q_n\right) = \begin{cases} 0.5 \cdot (1-\alpha_n) & \left(x^Q_n = 0, y^Q_n = 0\right) \\ 0.5 \cdot \alpha_n & \left(x^Q_n = 0, y^Q_n = 1\right) \\ 0.5 \cdot \alpha_n & \left(x^Q_n = 1, y^Q_n = 0\right) \\ 0.5 \cdot (1-\alpha_n) & \left(x^Q_n = 1, y^Q_n = 1\right) \end{cases} \quad (n = 1,2,\cdots,N) \,. \tag{15}$$

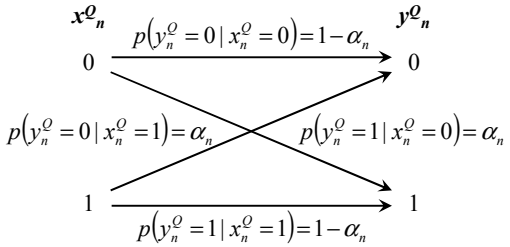


Figure 1. Symmetric failure model for SRAM cell where the failure probability is α_n.

The random failure of SRAM cell can be conceptually considered as an additive noise for the quantized signal x^Q:

$$x^Q = y^Q + \delta^Q \tag{16}$$

where

$$\delta^Q = \sum_{n=1}^{N} 2^{-n} \cdot \delta^Q_n \tag{17}$$

stands for the "equivalent" noise caused by random cell failure. Substituting (9), (14) and (17) into (16) yields:

$$\delta^Q_n = x^Q_n - y^Q_n \quad (n = 1,2,\cdots,N) \,. \tag{18}$$

In (18), each digit δ^Q_n represents the error between the original data x^Q_n and the stored data y^Q_n. Based on the joint probability mass function $p(x^Q_n, y^Q_n)$ in (15), it is easy to derive the probability mass function for δ^Q_n:

$$p\left(\delta^Q_n\right) = \begin{cases} 0.5 \cdot \alpha_n & \left(\delta^Q_n = -1\right) \\ 1-\alpha_n & \left(\delta^Q_n = 0\right) \\ 0.5 \cdot \alpha_n & \left(\delta^Q_n = 1\right) \end{cases} \quad (n = 1,2,\cdots,N) \,. \tag{19}$$

and the following statistics:

$$\mu\left(\delta^Q_n\right) = 0 \tag{20}$$

$$var\left(\delta^Q_n\right) = \alpha_n \,. \tag{21}$$

Substituting (20)-(21) into (17), we have:

$$\mu\left(\delta^Q\right) = 0 \tag{22}$$

$$var\left(\delta^Q\right) = \sum_{n=1}^{N} 4^{-n} \cdot \alpha_n \,. \tag{23}$$

Finally, we combine (10) and (16), yielding:

$$x = y^Q + \varepsilon + \delta^Q \,. \tag{24}$$

Eq. (24) implies an important fact that when approximating a real-valued signal x by N digits and store them in an SRAM, we introduce two additive noise terms. The first term ε corresponds to quantization noise, and the second term δ^Q is associated with random cell failure. Eq. (12)-(13) and (22)-(23) specify the statistics (i.e., mean and variance) for these two noise sources, respectively. The aforementioned noise model is summarized in Figure 2.

Studying Figure 2, we would have two important observations. First, both noise terms ε and δ^Q lead to information loss. To maximize the stored information, different strategies should be

applied to minimize different noise sources. For example, we should increase the number of digits (i.e., N) and, hence, the number of SRAM cells to reduce the quantization noise ε. On the other hand, to minimize the noise δ^Q caused by random cell failure, large transistors should be used and/or extra redundancy should be added. All these approaches result in increased area. The open question here is how to optimally explore the trade-off between total noise and silicon area.

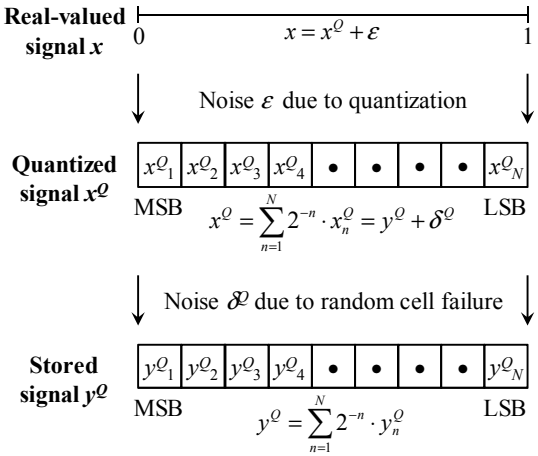


Figure 2. Two additive noise terms, ε due to quantization noise and δ^Q due to random cell failure, are modeled when approximating a real-valued signal x.

Second, when the quantized signal $\{x^Q_n;\ n = 1,2,\ldots,N\}$ is stored as N digits $\{y^Q_n;\ n = 1,2,\ldots,N\}$, different digits correspond to different levels of importance. For instance, the MSB y^Q_1 is much more important than the LSB y^Q_N, since it is assigned to a much larger weight. If the SRAM cell y^Q_1 fails, it leads to a large distortion (i.e., large information loss) of the signal x^Q. From this point of view, MSB should be better protected to achieve a smaller failure rate than LSB. The open question here is how to determine the optimal failure rate for each digit y^Q_n so that the overall information density is maximized.

Motivated by these observations, we will derive an information model in the next sub-section that allows us to quantitatively measure the information density and quickly explore the design trade-offs for an SRAM system. Such a model will be further used in Section 4 to maximize the information density for the proposed MISS system.

3.2 Information Modeling

As shown in Figure 2, a real-valued signal x is represented as y^Q, after it is quantized and stored in an SRAM. Ideally, we want x and y^Q to be exactly identical (i.e., no information loss). However, such an ideal case can never be achieved due to quantization error and random cell failure. To quantitatively model the "difference" between x and y^Q, we adopt the concept of mutual information that is defined in Section 2. In our case, the mutual information $I(x, y^Q)$ measures the information of x that we can learn from y^Q. We should maximize $I(x, y^Q)$ (or equivalently, minimize the information loss) as much as possible so that y^Q accurately approximates x.

Before moving forward, it is important to mention that the variable y^Q in (14) is discrete. In theory, the mutual information of a discrete random variable cannot be calculated by (7), as Eq. (7) is only applicable to continuous random variables. In most practical applications, however, the total number of digits (i.e., N) is large and, hence, y^Q can be approximately treated as a continuous variable. This assumption is adopted in our paper and we will use (7) to derive an approximate model for the mutual information $I(x, y^Q)$ in this sub-section.

Based on (7), we represent $I(x, y^Q)$ as the difference between two entropy metrics:

$$I(x, y^Q) = H(x) - H(x \mid y^Q). \tag{25}$$

Studying (25), one would notice that the first term $H(x)$ is the differential entropy of x. It is independent of the SRAM system that we design. Hence, our goal here is to model the second term $H(x \mid y^Q)$ and study how $H(x \mid y^Q)$ is related to the two noise terms in (24), i.e., ε due to quantization noise and δ^Q due to random cell failure. The information model we develop will represent $H(x \mid y^Q)$ as a function of N (i.e., the total number of digits) and $\{\alpha_n;\ n = 1,2,\ldots,N\}$ (i.e., the cell failure probabilities).

Towards this goal, we first apply (6) and (24) to simply the conditional differential entropy $H(x \mid y^Q)$:

$$H(x \mid y^Q) = H(y^Q + \varepsilon + \delta^Q \mid y^Q) = H(\varepsilon + \delta^Q \mid y^Q). \tag{26}$$

In (26), $H(\varepsilon + \delta^Q \mid y^Q)$ is the conditional differential entropy for $\varepsilon + \delta^Q$. Calculating the conditional entropy of the sum of two random variables is extremely difficult [15]. Hence, instead of finding the exact value of $H(\varepsilon + \delta^Q \mid y^Q)$, we aim to determine its upper bound and, equivalently, the lower bound of the mutual information $I(x, y^Q)$ in (25). Such a lower/upper bound technique is facilitated by the following two approximations: (1) independence approximation, and (2) Gaussian approximation.

1) Independence approximation: Based on (5), we have:

$$H(\varepsilon + \delta^Q \mid y^Q) \le H(\varepsilon + \delta^Q) \tag{27}$$

where the equality holds when $\varepsilon + \delta^Q$ and y^Q are independent. In general, even though the quantization noise ε and the stored signal y^Q are independent, the noise δ^Q due to random cell failure and y^Q are correlated. Such a correlation can be intuitively explained by the fact that if the n-th digit y^Q_n is 0, the corresponding noise δ^Q_n can only be 0 or 1 (but not –1), according to the noise definition in (18). In other words, since x^Q_n can be either 0 or 1, $\delta^Q_n = x^Q_n - y^Q_n$ can never be –1 given $y^Q_n = 0$. It, in turn, demonstrates that y^Q_n and δ^Q_n are correlated. Namely, the value of δ^Q_n depends on the value of y^Q_n. Hence, $H(\varepsilon + \delta^Q)$ in (27) is an upper bound to approximate $H(\varepsilon + \delta^Q \mid y^Q)$, where the equality cannot be reached in our application.

2) Gaussian approximation: We further apply Theorem 2 to derive an upper bound for $H(\varepsilon + \delta^Q)$. To this end, we define a new random variable ξ with Gaussian distribution. The mean and variance of ξ are identical to those of $\varepsilon + \delta^Q$:

$$\mu(\xi) = \mu(\varepsilon + \delta^Q) \tag{28}$$

$$var(\xi) = var(\varepsilon + \delta^Q) = var(\varepsilon) + var(\delta^Q) \tag{29}$$

where Eq. (29) utilizes the property that ε and δ^Q are independent. The mutual independence between ε and δ^Q is a valid assumption, as these two noise terms come from completely different sources. The new random variable ξ is the optimal Gaussian distribution that approximates $\varepsilon + \delta^Q$. In this case, Theorem 2 guarantees:

$$H(\varepsilon + \delta^Q) \le H(\xi). \tag{30}$$

Namely, we use the differential entropy $H(\xi)$ as an upper bound to approximate $H(\varepsilon + \delta^Q)$. Combing (3), (13), (23), (29) and (30), we have:

$$H(\varepsilon + \delta^Q) \le \frac{1}{2} \cdot \log_2\left[2\pi \cdot e \cdot \left(\frac{1}{12} 4^{-N} + \sum_{n=1}^{N} 4^{-n} \cdot \alpha_n\right)\right]. \tag{31}$$

Finally, substituting (26)-(27) and (31) into (25) yields:

$$I(x, y^Q) \geq H(x) - \frac{1}{2} \cdot \log_2 \left[2\pi \cdot e \cdot \left(\frac{1}{12} 4^{-N} + \sum_{n=1}^{N} 4^{-n} \cdot \alpha_n \right) \right]. \quad (32)$$

Eq. (32) is the information model that we aim to derive. It gives the lower bound of the mutual information between x and y^Q. To design the proposed MISS system, we should maximize the mutual information $I(x, y^Q)$ and, hence, maximize the lower bound in (32):

$$\max_{N,\alpha_1,\cdots,\alpha_N} \quad H(x) - \frac{1}{2} \cdot \log_2 \left[2\pi \cdot e \cdot \left(\frac{1}{12} 4^{-N} + \sum_{n=1}^{N} 4^{-n} \cdot \alpha_n \right) \right] \quad (33)$$

where N (i.e., the total number of digits) and $\{\alpha_n;\ n = 1,2,\ldots,N\}$ (i.e., the cell failure probabilities) are the design parameters that should be determined.

Note that the differential entropy $H(x)$ in (33) is independent of N and $\{\alpha_n;\ n = 1,2,\ldots,N\}$. In addition, the function $\log_2(\bullet)$ monotonically increases. Hence, the optimization in (33) is equivalent to:

$$\min_{N,\alpha_1,\cdots,\alpha_N} \quad \frac{1}{12} 4^{-N} + \sum_{n=1}^{N} 4^{-n} \cdot \alpha_n \,. \quad (34)$$

Eq. (34) implies an important fact that *maximum-information storage can be achieved by minimizing the total noise power*, i.e., $var(\varepsilon) + var(\delta^Q)$. This conclusion is consistent with our intuition. Namely, maximum signal-to-noise ratio can be achieved, if noise power is minimized. In practice, the optimization in (34) must be solved subject to a given area constraint so that the information density (instead of the total information) is maximized. These implementation details will be discussed in the next section.

4. MISS DESIGN

Given the information model derived in the previous section, we need to further capture the relation between the silicon area and the design parameters, i.e., N and $\{\alpha_n;\ n = 1,2,\ldots,N\}$, so that the optimization in (34) can be solved subject to a given area constraint. Assume that N SRAM cells are used to store the N digits $\{y^Q_n;\ n = 1,2,\ldots,N\}$ and the silicon area of these memory cells is denoted as $\{s_n;\ n = 1,2,\ldots,N\}$. The total silicon area s_{Total} is simply the summation of $\{s_n;\ n = 1,2,\ldots,N\}$:

$$s_{Total} = \sum_{n=1}^{N} s_n \,. \quad (35)$$

On the other hand, there are several design options that we can consider to explore the trade-offs between the cell area $\{s_n;\ n = 1,2,\ldots,N\}$ and the failure probabilities $\{\alpha_n;\ n = 1,2,\ldots,N\}$. For example, the failure probability α_n can be reduced, if the transistor size is increased and/or extra redundancy is added. Due to the page limit of this paper, we will not consider all these possible implementation options. Instead, we will focus on the transistor sizing approach for the rest of this paper.

To quantitatively model the relation between s_n and α_n, we consider a commercial 6-T SRAM cell designed in a 65nm CMOS process. Taking the initial design provided by the foundry, we proportionally scale the widths of all six transistors to vary the cell area s_n. In addition, we apply Monte Carlo analysis (i.e., importance sampling [6]-[8], [10], [13]) to estimate the failure rate α_n with the consideration of process variations. Figure 3 plots the failure rate α_n as a function of the cell area s_n. In Figure 3, s_n is normalized where $s_n = 5$ represents the initial foundry design and $s_n = 1$ denotes the minimum-size design with the transistors at their minimum feature size.

Studying Figure 3, we would notice that the failure rate α_n exponentially decreases, as the cell area s_n increases. The logarithm of the failure probability is almost a linear function of the cell area. Motivated by this observation, we use the following model to approximate the relation between α_n and s_n:

$$\alpha_n = k \cdot \exp(-\beta \cdot s_n) \quad (36)$$

where the model coefficients $k = 1.5\times10^{-4}$ and $\beta = 0.73$ are determined by least-squares fitting [24] from the data points in Figure 3. The fitted model accurately matches the Monte Carlo analysis result, as shown in Figure 3.

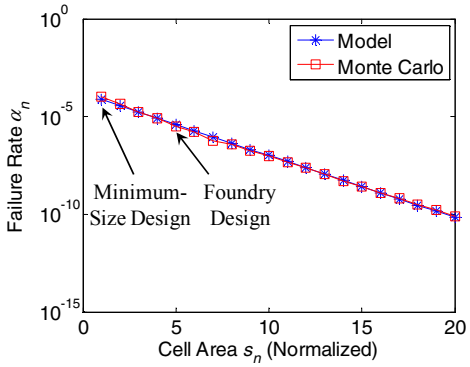


Figure 3. The SRAM cell failure rate α_n exponentially decreases as the cell area s_n increases.

Combining (34)-(36), we can formulate the following constrained optimization problem:

$$\begin{aligned} \min_{N,s_1,\cdots,s_N} \quad & \frac{1}{12} 4^{-N} + k \cdot \sum_{n=1}^{N} 4^{-n} \cdot \exp(-\beta \cdot s_n) \\ \text{S.T.} \quad & \sum_{n=1}^{N} s_n = s_{Total} \\ & s_n \geq s_{Min} \qquad (n = 1,2,\cdots,N) \end{aligned} \quad (37)$$

where s_{Total} and s_{Min} denote the given specifications of the total silicon area and the minimum cell area, respectively. Taking Figure 3 as an example, s_{Min} (normalized) is equal to 1, as is determined by the minimum feature size of the manufacturing process. In (37), N (i.e., the total number of digits) and $\{s_n;\ n = 1,2,\ldots,N\}$ (i.e., the cell area) are the design parameters that should be determined. Compared to (34), the cell failure probabilities $\{\alpha_n;\ n = 1,2,\ldots,N\}$ are replaced by $\{s_n;\ n = 1,2,\ldots,N\}$ in (37), since there is a one-to-one mapping between α_n and s_n as shown in (36).

As the variable N in (37) is an integer, the resulting nonlinear optimization is an integer programming problem. It cannot be directly solved by an efficient and robust algorithm. However, for any fixed value of N, the cost function in (37) is a convex exponential function of $\{s_n;\ n = 1,2,\ldots,N\}$ and the constraints are simply linear functions of $\{s_n;\ n = 1,2,\ldots,N\}$. Hence, if N is fixed, Eq. (37) is a convex optimization problem for the variables $\{s_n;\ n = 1,2,\ldots,N\}$. It can be easily solved by many convex programming algorithms (e.g., interior point method [19]). For this reason, instead of solving (37) directly, we propose the following hierarchical search approach.

Algorithm 1: Hierarchical Optimization for MISS System

1. Start from the cell failure rate model in (36), the total silicon area s_{Total}, and the minimum cell area s_{Min}.
2. Calculate N_{Max} as the largest integer that is no greater than s_{Total}/s_{Min}. Set $N = 1$.
3. Given a fixed value of N, solve the convex optimization problem in (37) to determine $\{s_n;\ n = 1,2,\ldots,N\}$ and calculate

the optimal cost function value c_N.

4. $N = N + 1$. If $N \le N_{Max}$, go to step 3. Otherwise, go to Step 5.
5. Find the minimum cost function value from the set $\{c_N;\ N = 1,2,\ldots,N_{Max}\}$. Determine the corresponding N and $\{s_n;\ n = 1,2,\ldots,N\}$ as the optimal solution of the MISS system.

Algorithm 1 searches the optimal N and $\{s_n;\ n = 1,2,\ldots,N\}$ for (37) via two hierarchical loops. In the inner loop, we solve a convex optimization problem to determine $\{s_n;\ n = 1,2,\ldots,N\}$ for a given N. On the other hand, N is varied from 1 to N_{Max} during the top-level iterations to search its optimal value. In Algorithm 1, N_{Max} is determined by the ratio between s_{Total} and s_{Min}. Namely, the maximum number of SRAM cells cannot be greater than s_{Total}/s_{Min}.

The optimization in (37) minimizes the total noise power and, hence, maximizes the mutual information in (32) for a given area constraint. Once it is solved by Algorithm 1, we find the optimal N and $\{s_n;\ n = 1,2,\ldots,N\}$ to store the real-valued signal in (8) with maximum information density. In most practical applications, multiple real-valued signals are involved for signal processing, and each signal should be stored in N SRAM cells where the optimal cell area is $\{s_n;\ n = 1,2,\ldots,N\}$. The efficacy of the proposed MISS system over the traditional SRAM design will be demonstrated by a 65nm design example in Section 5 and two real-life signal processing applications in Section 6.

5. DESIGN EXAMPLE

In this section, we demonstrate the efficacy of the proposed MISS system and highlight its difference over the traditional SRAM design. A commercial 6-T SRAM cell in a 65nm CMOS process is used as the test case. The failure rate of this SRAM cell is shown in Figure 3.

Two different SRAM design methodologies are implemented for testing and comparison purpose. First, an initial cell design is provided by the foundry. Its normalized cell area is 5 and the corresponding failure rate is 3.9×10^{-6}, as shown in Figure 3. If N digits are used to represent a real-valued signal, all these N SRAM cells have the same silicon area $\{s_n = 5;\ n = 1,2,\ldots,N\}$ and failure rate $\{\alpha_n = 3.9\times10^{-6};\ n = 1,2,\ldots,N\}$. The total silicon area is simply $5\times N$. The aforementioned setup is used to create the traditional SRAM design.

Second, we take the initial foundry design and proportionally scale the widths of all transistors, resulting in the cell failure rate shown in Figure 3. In this example, the normalized minimum cell area is $s_{Min} = 1$. Next, we apply Algorithm 1 to generate the optimal MISS design for a given area constraint s_{Total}. The convex optimization in Algorithm 1 is solved by CVX [17]. It takes less than 1 minute on a desktop to finish the top-level iterations of Algorithm 1. As s_{Total} varies from 5 to 100, a number of optimal MISS designs are created.

In this paper, we use signal-to-noise ratio (SNR) as a criterion to quantitatively compare the aforementioned SRAM designs. Given the real-valued signal x that is uniformly distributed over [0, 1], its energy can be measured by the variance [18]:

$$var(x) = \frac{1}{12}. \tag{38}$$

On the other hand, the total noise power, including the noise ε due to quantization and the noise δ^Q due to random cell failure, can be calculated based on (13), (23) and (29):

$$var(\varepsilon) + var(\delta^Q) = \frac{1}{12}4^{-N} + \sum_{n=1}^{N} 4^{-n}\cdot\alpha_n. \tag{39}$$

Combining (38) and (39), we have:

$$SNR = -\log_{10}\left(4^{-N} + 12\cdot\sum_{n=1}^{N} 4^{-n}\cdot\alpha_n\right). \tag{40}$$

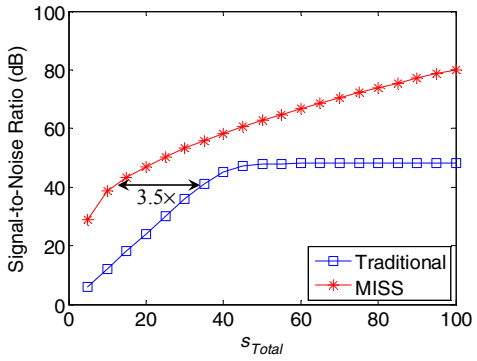


Figure 4. The proposed MISS system achieves significantly improved signal-to-noise ratio over the traditional SRAM design.

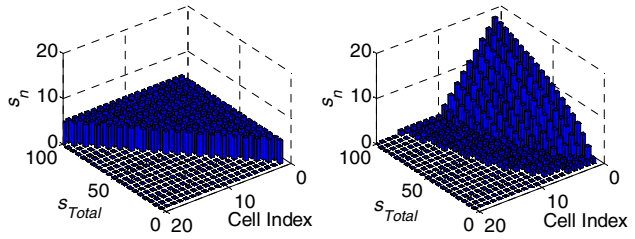


Figure 5. The number of SRAM cells and the corresponding silicon area of these memory cells are different for (a) the traditional SRAM design and (b) the proposed MISS system.

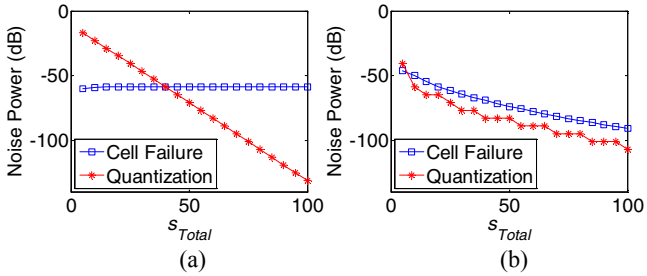


Figure 6. Noise power varies for (a) the traditional SRAM design and (b) the proposed MISS system, as s_{Total} varies from 5 to 100.

Figure 4 shows the SNR values as the normalized total area s_{Total} varies from 5 to 100. Note that the proposed MISS system achieves significantly improved SNR over the traditional SRAM design. As labeled in Figure 4, to reach the same SNR (or equivalently, to store the same amount of information), the proposed MISS system offers more than 3.5× area reduction compared to the traditional SRAM design.

To intuitively understand the advantage offered by MISS, Figure 5 plots the number of SRAM cells and the corresponding silicon area of these memory cells for both the traditional SRAM design and the proposed MISS system. Comparing Figure 5(a) and Figure 5(b), we would have two important observations. First, when s_{Total} is small, the traditional SRAM design uses a small number of digits to store a real-valued signal x. In the extreme case, if s_{Total} is equal to 5, only one digit is used to represent x,

resulting in large quantization noise. On the other hand, five digits are used by the MISS system given s_{Total} = 5. In this case, even though the corresponding five SRAM cells must take minimum size, they offer significantly improved SNR, as shown in Figure 4.

Second, when s_{Total} is large, a lot of digits are adopted by the traditional SRAM design where the corresponding SRAM cells have identical silicon area. For example, when s_{Total} is equal to 100, there are 20 SRAM cells in total where the normalized silicon area of each cell is 5. The proposed MISS system, however, only uses 16 SRAM cells in this case. In addition, these 16 memory cells are optimally sized so that the MSB cell is large (i.e., has low failure rate) and the LSB cell is small (i.e., has high failure rate). As a result, the SNR (and hence, the information density) is maximized.

Figure 6 further plots the noise power as a function of s_{Total} for both the traditional SRAM design and the proposed MISS system. Studying Figure 6(a) for the traditional SRAM design, the quantization noise dominates, when s_{Total} is small and only few digits are used to represent a real-valued signal. As s_{Total} increases, quantization noise decreases and eventually random cell failure becomes the dominant noise source. On the other hand, the proposed MISS system carefully balances these two noise sources and the noise power continuously decreases with increased s_{Total}, as Algorithm 1 is applied to optimally size all SRAM cells.

Finally, it is important to mention that while cell redundancy and error-correcting code (ECC) are typically applied to a practical SRAM system, they are not considered in this paper. Both of these techniques help to reduce cell failure rate with increased silicon area. They can be incorporated into the proposed MISS system as alternative implementation options. Due to the page limit of this paper, the detailed discussion on these topics is not presented here.

6. APPLICATIONS

In this section, we study two signal processing examples for the traditional SRAM design and the proposed MISS system. Our objective is to demonstrate the impact of MISS on real-life applications.

6.1 Image Processing

(a) (b) (c)

Figure 7. Image processing example: (a) original image, (b) image stored by the traditional SRAM design, and (c) image stored by the proposed MISS system.

Shown in Figure 7(a) is a benchmark example that has been widely used to test various image processing algorithms [22]. The image size is 512×512. It is in BMP format where each pixel is a numerical value. For testing and comparison, two different SRAM systems are designed to store this BMP image.

First, we create a traditional SRAM design using the setup described in Section 5. In this study, we set the total silicon area s_{Total} = 10 for each pixel. Since the area of one traditional SRAM cell is 5, we only have two digits to represent one pixel of the image. It, in turn, results in large quantization noise, as shown in Figure 7(b).

Second, we create an optimal MISS design with the same area constraint (i.e., s_{Total} = 10) for each pixel. In this case, Algorithm 1 optimally decides to use eight digits to represent one pixel. Even though each of these eight SRAM cells has small area (and hence, large failure rate) compared to the traditional SRAM design, MISS minimizes the total noise power (in particular, the quantization noise in this case). Therefore, the resulting image has significantly improved signal-to-noise ratio, as shown in Figure 7(c).

Finally, it is worth mentioning that the proposed MISS system is completely different from the traditional image compression method. While image compression is an alternative approach to increase the information density for data storage, it requires extra encoding and decoding steps and, hence, increases the latency for read and write operations. For this reason, image compression has never been used for on-chip data cache where access time is of great importance.

6.2 Neural Signal Processing

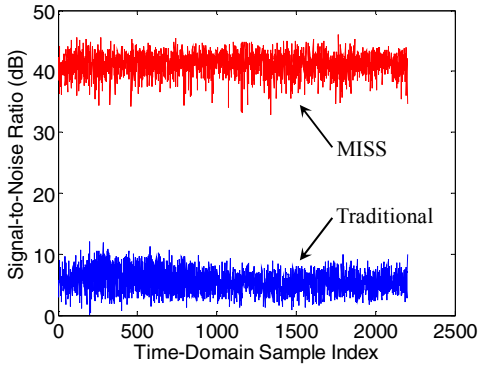
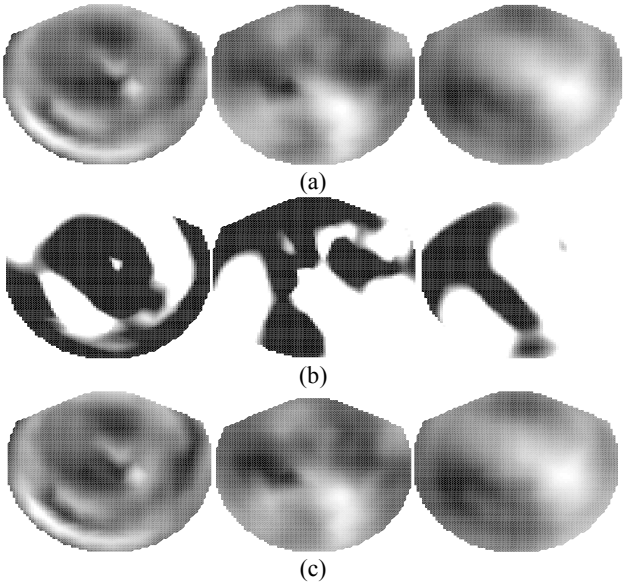


Figure 8. Signal-to-noise ratio achieved by the traditional SRAM design and the proposed MISS system for neural signal processing.



(a) (b) (c)

Figure 9. Neural signal processing example: (a) original MEG image, (b) MEG image stored by the traditional SRAM design, and (c) MEG image stored by the proposed MISS system.

In this sub-section, we consider a neural signal processing

example where magnetoencephalography (MEG) is recorded for a human subject. MEG measures the magnetic field generated by human brain [16]. In our experiment, MEG is sampled at 1 kHz for 2.2 seconds. In other words, one MEG image is sampled every one millisecond and 2200 MEG images are collected in total.

The aforementioned MEG images are stored in two different SRAM systems: the traditional SRAM design and the proposed MISS system. Both SRAMs are designed using the same setup as described in Section 6.1 (i.e., s_{Total} = 10 for each pixel). In addition, a signal space separation (SSS) algorithm [16] is applied to these MEG data to perform spatial filtering.

Figure 8 shows the signal-to-noise ratio (SNR) for both SRAM systems. Note that MISS achieves more than 30 dB improvement in SNR over the traditional SRAM design. In addition, Figure 9 shows the SSS results at a particular sampling time for (a) the original MEG image, (b) the MEG image stored by the traditional SRAM design, and (c) the MEG image stored by MISS. In each case, three different images are displayed. Roughly speaking, these three images correspond to the magnetic field at three different directions in the 3-D space. Studying Figure 9, we would notice that MISS achieves significantly reduced distortion compared to the traditional SRAM design.

7. CONCLUSIONS

In this paper, we propose a new SRAM design methodology that is referred to as maximum-information storage system (MISS). MISS aims to maximize the information density (instead of cell density) for SRAM. It offers an optimal SRAM design (e.g., maximum signal-to-noise ratio) for a number of application-specific cases such as signal processing. To optimally design the proposed MISS system, an information model is derived to quantitatively measure the information bits stored in an SRAM. In addition, a convex optimization framework is developed to determine the optimal transistor sizing to achieve maximum information density. As is demonstrated by our 65nm SRAM design example, MISS can reduce silicon area by 3.5× compared to the traditional SRAM circuit. When applied to signal processing applications, MISS achieves more than 30 dB improvement in signal-to-noise ratio over the traditional SRAM system. Based on these promising results, MISS is expected to offer a radically new design paradigm for next-generation SRAM circuits.

8. ACKNOWLEDGEMENTS

The author acknowledges the support of the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity. This work is also supported in part by the National Science Foundation.

9. REFERENCES

[1] B. Calhoun, Y. Cao, X. Li, K. Mai, L. Pileggi, R. Rutenbar and K. Shepard, "Digital circuit design challenges and opportunities in the era of nanoscale CMOS," *Proceedings of The IEEE*, vol. 96, no. 2, pp. 343-365, Feb. 2008.

[2] T. Mizuno, J. Okamura and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *IEEE Trans. on Electron Devices*, vol. 41, no. 11, pp. 2216-2221, Nov. 1994.

[3] A. Bhavanagarwala, X. Tang and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE JSSC*, vol. 36, no. 4, pp. 658-665, Apr. 2001.

[4] S. Mukhopadhyay, K. Kim, H. Mahmoodi and K. Roy, "Design of a process variation tolerant self-repairing SRAM for yield enhancement in nanoscaled CMOS," *IEEE JSSC*, vol. 42, no. 6, pp. 1370-1382, Jun. 2007.

[5] S. Mukhopadhyay, H. Mahmoodi and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," *IEEE Trans. CAD*, vol. 24, no. 12, pp. 1859-1880, Dec. 2005.

[6] R. Kanj, R. Joshi and S. Nassif, "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events," *IEEE DAC*, pp. 69-72, 2006.

[7] A. Singhee and R. Rutenbar, "From finance to flip flops: a study of fast quasi-Monte Carlo methods from computational finance applied to statistical circuit analysis," *IEEE ISQED*, pp. 685-692, 2007.

[8] A. Singhee and R. Rutenbar, "Statistical blockade: a novel method for very fast Monte Carlo simulation of rare circuit events, and its application," *IEEE DATE*, pp. 16-20, 2007.

[9] C. Gu and J. Roychowdhury, "An efficient, fully nonlinear, variability-aware non-Monte-Carlo yield estimation procedure with applications to SRAM cells and ring oscillators," *IEEE ASPDAC*, pp 754-761, 2008.

[10] L. Dolecek, M. Qazi, D. Shah and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," *IEEE ICCAD*, pp. 322-329, 2008.

[11] J. Wang, S. Yaldiz, X. Li and L. Pileggi, "SRAM parametric failure analysis," *IEEE DAC*, pp. 496-501, 2009.

[12] R. Kanj, R. Joshi, C. Adams, J. Warnock and S. Nassif, "An elegant hardware-corroborated statistical repair and test methodology for conquering aging effects," *IEEE ICCAD*, pp. 497-504, 2009.

[13] J. Jaffari and M. Anis, "Adaptive sampling for efficient failure probability analysis of SRAM cells," *IEEE ICCAD*, pp. 623-630, 2009.

[14] A. Bansal, R. Singh, R. Kanj, S. Mukhopadhyay, J. Lee, E. Acar, A. Singhee, K. Kim, C. Chuang, S. Nassif, F. Heng and K. Das, "Yield estimation of SRAM circuits using virtual SRAM fab," *IEEE ICCAD*, pp. 631-636, 2009.

[15] M. Madiman and A. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Trans. Information Theory*, vol. 53, no. 7, pp. 2317-2329, Jul. 2007.

[16] S. Taulu, J. Simola and M. Kajola, "Applications of the signal space separation method," *IEEE Trans. Signal Processing*, vol. 53, no. 9, pp. 3359-3372, Sep. 2005.

[17] M. Grant and S. Boyd, *Matlab Software for Disciplined Convex Programming* (http://stanford.edu/~boyd/cvx), Jun. 2009.

[18] A. Oppenheim, R. Schafer and J. Buck, *Discrete-Time Signal Processing*, 1999.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[20] G. Fishman, *A First Course in Monte Carlo*, Duxbury Press, 2005.

[21] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley Interscience, 2006.

[22] R. Gonzalez and R. Woods, *Digital Image Processing*, Prentice Hall, 2007.

[23] C. Bishop, *Pattern Recognition and Machine Learning*, Prentice Hall, 2007.

[24] W. Press, S. Teukolsky, W. Vetterling and B. Flannery, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, 2007.