Soft N-Modular Redundancy

Eric P. Kim, *Student, IEEE,* and Naresh R. Shanbhag, *Fellow, IEEE*

Abstract—Achieving robustness and energy-efficiency in nanoscale CMOS process technologies is made challenging due to the presence of process, temperature and voltage variations. Traditional fault-tolerance techniques such as N-modular redundancy (NMR) employ deterministic error-detection and correction, e.g., majority voter, and tend to be power hungry. This paper proposes *soft NMR* that non-trivially extends NMR by consciously exploiting error statistics caused by nanoscale artifacts in order to design robust and energy-efficient systems. In contrast to conventional NMR, soft NMR employs Bayesian detection techniques in the voter. Soft voter algorithms are obtained through optimization of appropriate application-aware cost functions. Analysis indicates that, on average, soft NMR outperforms conventional NMR. Furthermore, unlike NMR, in many cases, soft NMR is able to generate a correct output even when all N replicas are subject to errors. This increase in robustness is then traded-off through voltage scaling to achieve energy efficiency. The design of a discrete cosine transform (DCT) image coder is employed to demonstrate the benefits of the proposed technique. Simulations in a commercial $45nm$, $1.2V$, CMOS process show that soft NMR provides up to $10\times$ improvement in robustness, and 35% power savings over conventional NMR.

Index Terms—Low-power design, signal processing systems, redundant design.

1 Introduction

Modern nanoscale CMOS exhibit a number of artifacts such as process, temperature and voltage variations, leakage, and soft errors due to particle hits. It is expected that such non-idealities will only increase with rapid scaling of CMOS technology [1], and dominate the behavior of post-silicon device fabrics. Nanoscale non-idealities make it hard to design reliable computing systems [2]. Worst-case designs address the robustness issue with a severe power penalty. Nominal-case designs, though energy-efficient, suffer from reliability problems. Thus, energy-efficiency and reliability need to be addressed jointly.

Recently, error-resiliency has emerged as an attractive approach [3], [4], [5], [6] towards achieving robust and energy-efficient operation in nanoscale CMOS. Error-resiliency permits circuit errors to occur, and compensate for these at either the circuit, architecture or system levels. Communication-inspired error-resiliency techniques such as algorithmic noise-tolerance (ANT) [4], and stochastic sensor-network-on-chip (SSNOC) [5], treat application-specific nanoscale circuits and architectures as noisy communication channels, and employ statistical signal processing techniques in order to enhance robustness with minimal hardware overhead. ANT has been successfully applied to various media processing kernels such as filtering, motion estimation [7], and Viterbi decoding [8], while SSNOC-based CDMA PN-code acquisition system [5] has similarly demonstrated orders-of-magnitude enhancement in robustness along with significant energy-savings. Such techniques are able to achieve say a 90% system reliability with very high component error probabilities up to 40%, which corresponds to orders-of-magnitude enhancement in robustness over conventional systems. Energy-savings ranging from 20%-to-60% is achieved simultaneously. These techniques are termed as being *effective*, as they achieve robustness and energy-efficiency simultaneously. Memory-specific techniques have also been developed [6]. However, a key drawback of ANT and SSNOC-like approaches is that these are *application-specific*, i.e., error compensation assumes the knowledge of the algorithm being implemented. RAZOR [3] overcomes this limitation by focusing on error compensation at the logic and microarchitectural levels, but its effectiveness is limited to component error probabilities of 1.62% [3]. N-modular redundancy (NMR) [9], [10], [11], [12] (see Fig. 1(a)) is a commonly employed fault-tolerance technique with general applicability. In NMR, the same computation is executed in N processing elements (PEs), and the outputs are majority voted upon to select the correct one. However, its $N\times$ complexity and power overhead restricts its applicability to cost-insensitive critical applications such as those in the military, medical and high-end servers. Similarly, a number of other fault-tolerance techniques have been proposed in the past such as checkpointing [13], and coding techniques [14], [15]. Checkpointing is a technique that takes a snapshot of the entire execution state. In case of an error detection, the system rolls back to the most current checkpoint and re-executes. However, complex systems that have many processes sharing data can make checkpointing a non-trivial task. Also the storage requirement can be very large. On recovery, there is a significant time and energy overhead due to re-execution. Coding techniques make use of redundant bits that are organized

• *E.P. Kim and N.R. Shanbhag are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA e-mail: [epkim2, shanbhag]@illinois.edu.*

Digital Object Indentifier 10.1109/TC.2010.253

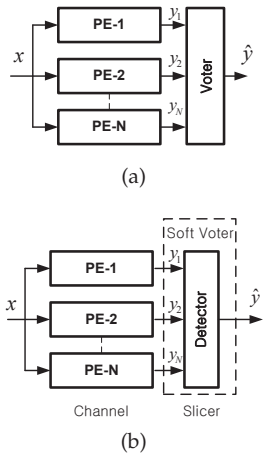


Fig. 1. Block diagram of: (a) NMR, and (b) soft NMR.

in a systematic way to enable error detection and correction. Simple techniques such as parity codes do not have sufficient coverage, while complex codes such as cyclic codes are expensive to implement. Each of these techniques are effective in enhancing robustness but at a significant energy-cost. Therefore, there is a need for effective error-resiliency techniques, i.e., those that provide energy-efficiency and robustness inherent in communications-inspired error-resiliency techniques [16], while exhibiting the generality of NMR. In this paper, we propose to employ error and signal statistics of the underlying circuit fabric and architecture in order to achieve this goal. We proposed *soft NMR* (see Fig. 1(b)) [17] as an effective error-resiliency technique with general applicability. Soft NMR incorporates communication-inspired techniques into NMR in order to improve its effectiveness, while preserving its generality. Structurally, soft NMR differs from NMR in that it incorporates a *soft voter*, which is composed of a *detector*. Thus, soft NMR views computation in the PEs as a noisy communication channel, and employs the detector as the slicer. Soft NMR enhances the robustness of NMR, which is then traded-off with energy in order to achieve energy-efficient operation. We show that soft NMR provides between $2\times$-to-$10\times$ improvement in robustness along with 13%-to-35% savings in power over NMR, for a DCT-based image compression kernel implemented in a commercial $45nm$, $1.2V$, CMOS process. It must be noted that though a number of NMR voting strategies exist, none exploit error statistics to enhance robustness, or trade-off robustness to achieve energy-efficiency.

In this paper, we describe past work in communication-inspired design techniques in Section 2. Next, the proposed soft NMR technique is introduced in Section 3. Statistical analysis of soft NMR, NMR and ANT is described in Section 4, which shows that soft NMR will always outperform NMR. Application of soft NMR, NMR and ANT to a discrete cosine transform (DCT)-based image coder is demonstrated in section 5. Finally Section 6 concludes the paper with future research directions.

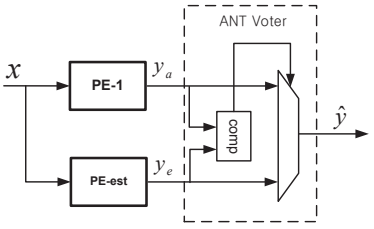


Fig. 2. An ANT based system.

2 Communication-Inspired Error-Resiliency Techniques

Communication-inspired techniques such as ANT [16] employ statistical signal processing techniques such as estimation and detection to compensate for errors in hardware, and exploit the statistical nature of system-level performance metrics. These techniques attempt to meet metrics such as signal-to-noise ratio (SNR) or bit error-rate (BER) specification of the application, instead of the somewhat arbitrary notion of numerical correctness employed conventionally. This view enables communication-inspired techniques to simultaneously achieve robustness and energy-efficiency.

2.1 Algorithmic noise-tolerance (ANT)

Algorithmic noise-tolerance in Fig. 2 has some similarities to dual-MR (DMR). It has a *main* PE and an *estimator* PE. The main PE is permitted to make errors, but not the estimator PE. Unlike DMR, the estimator PE in ANT, is a low-complexity (typically 5%-to-20% of the main PE complexity) computational block generating a statistical estimate of the correct main PE output, i.e.,

$$y_a = y_o + \eta \quad (1)$$

$$y_e = y_o + e \quad (2)$$

where y_a is the actual main PE output, y_o is the error-free main PE output, η is the hardware error, y_e is the estimator PE output, and e is the estimation error. Note: the estimator PE has estimation error e because it is a simpler than the main PE. ANT exploits the difference in the statistics of η and e. The final/corrected output of an ANT-based system $\hat{y}$ is obtained via the following decision rule:

$$\hat{y} = \begin{cases} y_a, & \text{if } |y_a - y_e| < \tau \\ y_e, & \text{otherwise} \end{cases} \quad (3)$$

where τ is an application dependent parameter chosen to maximize the performance of ANT.

Thus, ANT detects and corrects errors approximately, but does so in a manner that satisfies the

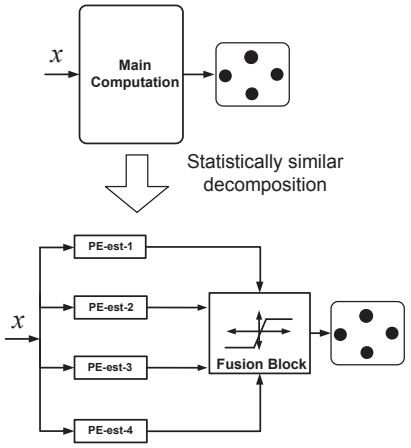


Fig. 3. The stochastic sensor network on a chip (SSNOC) block diagram.

performance specifications (SNR or BER) of the application at hand.

For ANT to enhance robustness, it is necessary that when $\eta \neq 0$, that η be large compared to e. In addition, the probability of the event $\eta \neq 0$, i.e., the component probability of error for the main PE, be small. For ANT to also provide energy-efficiency, it is necessary that the errors in the main PE are primarily due to enhancement of its energy-efficiency. In practice, these properties are easily satisfied when errors in the main PE occur due to timing violations due to voltage overscaling (VOS) [4] or a nominal case design being subjected to a worse case process corner. In VOS, the supply voltage is scaled below the critical voltage $V_{dd-crit}$ needed for error-free operation. When the supply voltage is lower than $V_{dd-crit}$, the circuit will operate slower than the designed margins, and thus timing violations will occur. The errors due to these timing violations are referred to as VOS type errors. As most computations are LSB first, VOS type errors are generally large magnitude MSB errors.

ANT has been shown to achieve up to $3\times$ energy savings in theory and in practice via prototype IC design [4] for finite impulse response (FIR) filters.

2.2 Stochastic sensor network on a chip (SSNOC)

Recently, the ANT approach has been extended into the networked domain via the concept of a SSNOC [5]. SSNOC relies only on estimator PEs or *sensors* to compute, and it permits hardware errors to occur in them (see Fig. 3). Thus, the output of the i^{th} estimator PE (PE-est-i) is given as

$$y_i = y_o + \eta_i + e_i \tag{4}$$

where η_i and e_i are the hardware and estimation errors in the i^{th} estimator, respectively.

If hardware errors are due to timing violations, one can approximate the error term in (4) as $(1-p_e)e_i + p_e\eta_i$, where p_e is the probability of $\eta_i \neq 0$, i.e., the component probability of error. Such an ϵ-contaminated model lends itself readily to the application of robust statistics [18] for error compensation. A key drawback of SSNOC is the feasibility of decomposing computation into several sensors whose outputs are *statistically similar*, i.e., its generality. SSNOC has been applied to a CDMA PN-code acquisition system, where the sensors were obtained through polyphase decomposition. Simulations indicate orders-of-magnitude improvement in detection probability while achieving up to 40% power savings.

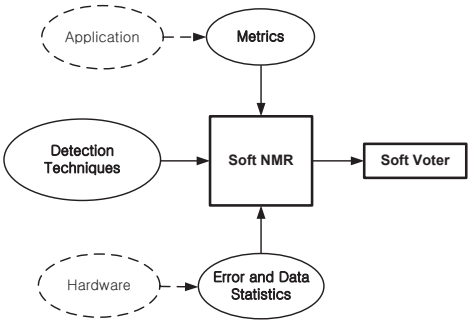


Fig. 4. The soft NMR framework.

3 Soft *N*-Modular Redundancy

Techniques such as ANT and SSNOC described in Section 2, though effective, are application dependent. NMR, though general, incurs a heavy power penalty. In this section, we propose soft NMR, which embodies the generality of NMR and the effectiveness of ANT. First, we present soft NMR by introducing a mathematical framework in which soft NMR and related techniques can be analyzed and understood. Various components of this framework are described, followed by the derivation of the soft voter algorithm and architecture.

3.1 Soft NMR Framework

The soft NMR framework (see Fig. 4) includes three components: 1) data and error statistics, 2) an application-dependent performance-metric, and 3) detection techniques. This framework enables us to systematically develop soft voting algorithms (see Fig. 1(b)) and architectures. Referring to Fig. 1(b), the parameters used to describe soft NMR are defined in Table 1.

3.1.1 Performance metrics

The soft voter is a realization of a Bayesian detector. A Bayesian detector tries to minimize the associated cost in making a decision. If we denote $C(\hat{y}, y_o)$ as the cost incurred in choosing $\hat{y}$ as the correct value, when in fact y_o is the correct value, the conditional cost (given that y_o is the correct value) is $C(\hat{y}, y_o)$.

TABLE 1
Parameters of soft NMR framework.

Notation	Description
N	total number of PEs
y_o	correct output value
Y_o	random variable corresponding to y_o
$\mathcal{V}$	*output space*: set of all possible outputs Its cardinality is m and its elements are denoted as $v_1, v_2, ..., v_m$. Note that $y_1, y_2, ..., y_N, \hat{y}, y_o \in \mathcal{V}$
$\mathcal{R}$	*observation space*: the set of all PE observations $\{y_1, y_2, ..., y_N\}$
$\mathcal{H}$	*hypothesis space*: set of hypotheses to be employed in detection
F	the event that a *fault* has occured
i	PE index
j	$\mathcal{V}$ index
r_j	*prior* defined as $P(y_o = v_j)$ Note that $\sum_{v_j \in \mathcal{V}} r_j = 1$
q_j	the probability that PE observation y_i is v_j given a fault has occurred, i.e., $q_j = P(y_i = v_j \mid F)$
$p_{e,sys}$	overall system error probability, i.e., the probability that $\hat{y} \neq y_o$
P_{e_i}	distribution of the error e_i of PE-i
p_{e_i}	error probability of PE-i
$c_j(\mathcal{R})$	occurrences of v_j in $\mathcal{R}$
$C(\hat{y}, y_o)$	cost of deciding $\hat{y}$ as the correct value, given y_o was the correct value

Denoting $\delta(y_1, ..., y_N) = \hat{y}$ as the decision rule based on the N observations, the average cost, or Bayes risk is given as

$$E(C(\delta(y_1, ..., y_N), Y_o)) \tag{5}$$

where the expectation is over the random variable Y_o representing the correct output y_o. It can be shown [19] that the decision rule δ which minimizes the Bayes risk, minimizes the posterior cost $E(C(\hat{y}, Y_o)|y_1, ...y_N)$. The soft voter is an implementation of the Bayes decision rule $\delta_{Bayes} = min_\delta E(C(\delta(y_1, ..., y_N), Y_o)|y_1, ..., y_N)$.

Thus, the cost function is the sole element that determines the functionality of the soft voter and should be closely tied to the application performance metric to maximize performance. For example, in a CPU, a single bit-flip in the instruction code regardless of its location will result in an incorrect instruction to be executed. Here, it is important to prevent any errors from occurring. Thus, an appropriate metric would be $p_{e,sys}$, the system error probability. The cost associated with this metric would be to assign a constant cost to all incorrect decisions $\hat{y} \neq y_o$ and 0 to a correct decision. In other applications, where small differences in the numeric value of an output are tolerable, we may wish to penalize large errors more than small errors. Here, a suitable metric may be the minimum mean square error (MMSE) which carries a cost function $C = (\hat{y} - y_o)^2$. For example, in image processing, the quality of an image may be assessed objectively with its peak signal-to-noise ratio (PSNR).

3.1.2 Error and data statistics

Soft NMR makes explicit use of two types of statistical information: (1) *data statistics*, and (2) *error statistics*. Data statistics are the distribution of the error-free PE output. This is referred to as the *prior* distribution, or prior. Error statistics are the distribution of the errors at the PE output. Note: the prior depends only upon input data statistics and the input-output mapping of the computation. The error distribution depends upon input data statistics, the functionality, the PE architecture, circuit style, and other implementation parameters. We, therefore, assume that both data and error statistics are obtained via a one-time Monte Carlo simulation using characterization/typical data. Data statistics is obtained via behavioral simulations of the computation. Error statistics are obtained via a register-transfer level (RTL) simulations of the PE with characterization data. The performance of soft NMR is then quantified with a separate/test data sequence, while employing the prior and error statistics obtained from the characterization data. Characterization and test data sequences are said to be statistically similar in that they are obtained from the same random process. For example, for image processing applications, one image is employed as characterization data, while the test data can be any other image(s). For architectures where *a priori* information on input statistics are not available, such as a general purpose system, or where statistics change over time, a training module that adaptively collects data and error statistics can be employed.

For error statistics to be meaningful, an error model that extends the existing model used in NMR analysis [20] is proposed. The error model separates the output into an error-free component and the error. Further elaboration on the error model and error characterization is provided in 3.2 and 3.3.

3.1.3 Detection techniques

The role of the soft voter in Fig. 1(b) is to determine the output $\hat{y}$ that would, on average, optimize a prespecified performance metric. The detector makes a decision based on the PE observations $y_1, y_2, \ldots, y_N$ that minimizes the Bayes risk. The detector, however, is constrained to a predefined *hypothesis* set $\mathcal{H}$ such that the decision rule $\hat{y} = \delta(y_1, ..., y_N)$ produces $\hat{y} \in \mathcal{H}$. Thus, the detection problem requires the definition of $\mathcal{H}$, from which the corrected output $\hat{y}$ is selected.The soft voter will perform a search over all elements of $\mathcal{H}$, and for practical implementations, the hypothesis

space $\mathcal{H}$ needs to be limited. There are several ways to limit $\mathcal{H}$.

1) Hypothesis space equals the observation space:

$$\mathcal{H} = \{\mathcal{H}_i = y_i, i = 1, \ldots, N\}$$

Here one of the observations is chosen as $\hat{y}$.

2) Hypothesis space equals a value neighborhood of the observation space:

$$\mathcal{H} = \{|\mathcal{H}_i - y_i| < r_t\},$$
where r_t is a given magnitude radius

This strategy includes values that are similar to the observations and thus has more choices for $\hat{y}$

3) Hypothesis space equals a probabilistic neighborhood of the observation space:

$$\mathcal{H} = \{P(y_i - \mathcal{H}_i) > r_p\},$$
where r_p is a given probability radius

This strategy is similar to the value based expanded hypothesis set, but the expansion is based on probabilistic measures.

More details on hypothesis set limiting is given in Section 3.6.

3.2 Error models

To enable the development of the detection framework of the soft voter, we have developed a general error model that encompasses most practical situations. We first present the conventional error model used for NMR [20], then generalize this error model. Refer to Table 1 for the definition of various parameters and notation.

3.2.1 NMR Error Model

The NMR error model [20] assumes that each PE has a probability p of exhibiting a fault F. Even though a fault F has occurred, the correct value may still be observed. With probability $1-p$, the PE will be fault-free and produce the correct value y_o. When a fault occurs, the observation will have a distribution $q_j (j = 1, ..., m)$, the probability that v_j is observed. Since for an error to occur, the PE needs to exhibit a fault and output the wrong value, the relation between PE error probability p_{e_i}, p and q_j is as follows:

$$p_{e_i} = p(1 - q_c) \tag{6}$$

where q_c is the probability the output is y_o in presence of a fault.

This model assumes the failure is independent of the input (or output), which is reasonable in cases where the failure is due to random particle hits or defects. However, timing errors due to voltage over-scaling have a direct dependence on the input and thus the errors depend on the input and the error-free output.

3.2.2 Soft NMR Error Model

Soft NMR error model is a generalization of the NMR error model given by

$$y_i = y_o + e_i(x, y_o) \tag{7}$$

The error e_i may differ for each processing element PE-i. Also the error may depend on the input, output or both as in (7). The distribution of e_i is denoted as $P_{e_i}(e_i)$. The relationship between p_{e_i} and $P_{e_i}(e_i)$ is as follows:

$$p_{e_i} = \sum_{e_i \in \mathcal{V}, e_i \neq 0} P_{e_i}(e_i) \tag{8}$$

This error model is more general than [20] in the sense that each PE is allowed to have a separate error distribution, and the dependence between the input/output of the PE and the error is captured.

The NMR error model is a special case of the soft NMR error model. For instance, q_j and p in the NMR error model can be written as:

$$q_j = P(y_i = v_j | F) = \begin{cases} P_{e_i}(e_i = v_j - y_o), & \text{when } e_i \neq 0 \\ 0, & \text{when } e_i = 0 \end{cases} \tag{9}$$

When $e_i = 0$, $q_j = P_{e_i}(0) = 0$, which implies that in case of a fault F, it will always lead to an error. Thus the probability of fault F becomes

$$p = 1 - P_{e_i}(e_i = 0) \tag{10}$$

In (9), the dependence of $P_{e_i}(e_i)$ on output y_o is clear. P_{e_i} shifts for each value of y_o for it to maintain a constant q_j.

3.3 Error statistics

Soft NMR requires the knowledge of the error statistics $P_{e_i}(e_i)$. In addition, both NMR and soft NMR work best when the individual block errors e_i are independent. In this section, methods to obtain $P_{e_i}(e_i)$, and techniques to ensure the independence of errors are discussed.

Error statistics are obtained by comparing structural RTL and behavioral simulations. The RTL simulations are conducted with delay values obtained via circuit/transister level characterization of basic macros such as a 1-bit full adder (FA). These simulations result in timing violations. The behavioral simulations provide the error-free output. Both simulations employ characteristic data.

The timing error distribution at the output of a 8×8, 8-bit input, 14-bit output, 2-D DCT block using Chen's algorithm [21], with mirror adders and array multipliers [22] as fundamental building blocks, implemented in a commercial 45 nm, 1.2 V CMOS process, is shown in Fig. 5 for two different voltages. In Fig. 5, one observes that the error PMFs become more spiky as the supply voltage decreases, and that a few large amplitude errors have a high probability of occurrence.

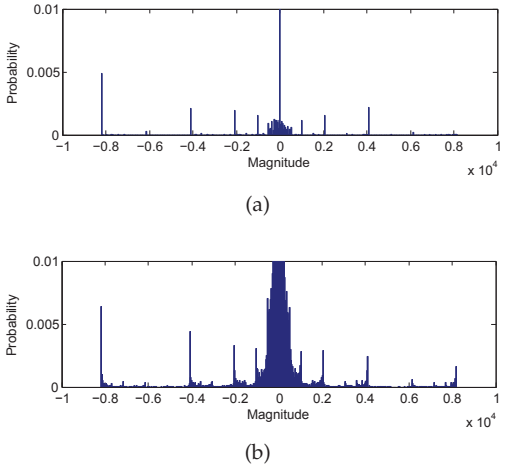


Fig. 5. Error statistics of a voltage overscaled DCT block in a 45 nm, 1.2 V CMOS process with $V_{dd,crit} = 1.2$ V: (a) $V_{dd} = 1$ V (probability of error is 0.0374), and (b) $V_{dd} = 0.8$ V (probability of error is 0.7142).

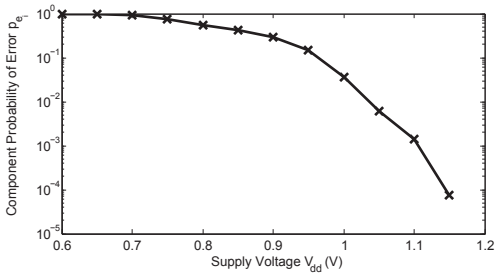


Fig. 6. Component error probability p_{e_i} vs. supply voltage V_{dd} for the DCT architecture.

This is to be expected as the DCT architecture is LSB-first and hence timing errors will appear in the MSBs, i.e., large amplitude error will occur. Such statistical characterization of macro blocks is essential if one is to design error-resilient architectures in nanoscale process technologies. We will employ the error PMFs in Fig. 5 in Section 5 to study the performance of soft NMR and its power vs. robustness trade-offs.

The component error probability p_{e_i} due to VOS of the DCT block is shown in Fig. 6. This plot was obtained from structural Verilog simulations of the DCT architecture at various supply voltages (hence delays) but with a fixed clock frequency. For the architecture being considered we find that p_{e_i} increases rapidly as the supply voltage is reduced beyond $V_{dd-crit}$.

3.3.1 Independence of Errors

Independent PE errors are essential for NMR to be effective. However, soft NMR does not need PE errors to be independent, though independent PE errors reduce the soft voter complexity, and thus are desirable. PE errors arising from the presence of faults satisfy the independence requirement. However, in this paper, we consider a broader class of errors such as those that arise from timing violations, i.e., insufficient execution time. Thus, if all the PEs in Fig. 1 have identical architectures and inputs, then the errors will be highly correlated or even identical. In this case, independent errors can be achieved by employing one or more of the following techniques:

- *architectural diversity*: employing different PE architectures.
- *scheduling diversity*: scheduling different sequences of operations on the PEs.
- *data diversity*: permitting each PE process a different sequence of inputs.
- *process diversity*: exploiting random within-die process variations.

In Section 5, we show the use of data diversity to obtain independent errors.

3.4 Soft Voter

Soft NMR employs a soft voter that uses Bayesian estimation techniques to minimize the cost. The soft voter algorithm depends on the performance metric and the error model. The soft voter will also depend on the error statistics if the latter can be described using a parametric form such as Gaussian, and if this form is employed in deriving the soft voter algorithm in order to reduce its complexity. In all cases, the output of the soft voter will be a function of the error statistics. Thus, the soft voter will be derived for both error models described in Section 3.2, and with two different metrics, $p_{e,sys}$ and MSE.

3.4.1 Soft voter that minimizes $p_{e,sys}$

As mentioned in Section 3.1.1, the cost function that minimizes $p_{e,sys}$ is given by:

$$C(\hat{y}, y_o) = \begin{cases} 1, & \text{if } \hat{y} \neq y_o \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Under this cost function, the posterior cost becomes

$$E(C(\hat{y}, Y_o)|y_1, ..., y_N) = \sum_{\forall v_j \in \mathcal{V}} P(v_j = y_o|y_1, ..., y_N) \mathbb{1}_{\hat{y} \neq v_j} \tag{12}$$

where $\mathbb{1}_{statement}$ denotes the indicator function that is 1 when *statement* is true and 0 otherwise. Thus, to minimize (12), $\hat{y}$ should be chosen to be v_j that maximizes $P(v_j = y_o|y_1, ..., y_N)$. This result is equivalent to minimizing $p_{e,sys}$, where we would like to find the value that has the largest probability of being correct. Thus, the soft voter maximizes:

$$P(v_j = y_o|y_1, y_2, ..., y_N) = \frac{P(y_1, y_2, ..., y_N|v_j = y_o)P(v_j = y_o)}{P(y_1, y_2, ..., y_N)} \tag{13}$$

Since the denominator of (13) is independent of v_j, it can be ignored. Thus, the soft voter finds the solution to the following:

$$\arg\max_{\forall v_j \in \mathcal{V}} P(y_1, y_2, ..., y_N|v_j = y_o)P(v_j = y_o) \tag{14}$$

Equation (14) will be referred to as the optimal rule. In general, the optimal rule is computationally intensive to implement if the hypothesis space $\mathcal{H} = \mathcal{V}$. This is because an exhaustive search needs to be executed over all possible m values of the output space, and the output space grows exponentially with increasing bit width. Soft NMR is an approximation of this rule and its performance will be compared to the optimal rule.

Assuming NMR error model, and defining the event where k PEs are faulty as F_k, the optimal rule (14) can be further simplified by application of Bayes' theorem $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$:

$$\arg\max_{\forall v_j \in \mathcal{V}} P(y_1, y_2, ..., y_N | v_j = y_o) P(v_j = y_o) = r_j \sum_{k=0}^{|c_j(\mathcal{R})|} P(y_1, y_2, ..., y_N | \{v_j = y_o\} \cap F_{N-k}) \times P(F_{N-k} | v_j = y_o) \quad (15)$$

where

$$P(y_1, y_2, ..., y_N | \{v_j = y_o\} \cap F_{N-k}) = q_j^{|c_j(\mathcal{R})|-k} \prod_{l=1,...,m, l \neq j} q_l^{|c_l(\mathcal{R})|} \quad (16)$$

$$P(F_{N-k} | v_j = y_o) = \binom{|c_j(\mathcal{R})|}{k} (1-p)^k p^{N-k} \quad (17)$$

Substituting (16) and (17) into (15), we obtain

$$\arg\max_{\forall v_j \in \mathcal{V}} P(y_1, ..., y_N | v_j = y_o) P(v_j = y_o) =$$

$$= \arg\max_{\forall v_j \in \mathcal{V}} r_j \left(\prod_{l=1,...,m, l \neq j} q_l^{|c_l(\mathcal{R})|} \right) \times \sum_{k=0}^{|c_j(\mathcal{R})|} \binom{|c_j(\mathcal{R})|}{k} (1-p)^k p^{N-k} q_j^{|c_j(\mathcal{R})|-k} \quad (18)$$

$$= \arg\max_{\forall v_j \in \mathcal{V}} r_j \left(\prod_{l=1,...,m, l \neq j} q_l^{|c_l(\mathcal{R})|} \right) p^N \left(\frac{1-p}{p} + q_j \right)^{|c_j(\mathcal{R})|} \quad (19)$$

We first simplify the optimal rule under the NMR error model, then apply the results to the soft NMR error model with slight modifications.

- Soft Voter using NMR Error Model
  It can be shown [20] that under the NMR error model with the assumption that the outputs of each block are subject to independent errors, the optimal rule can be simplified from (19) as follows:

$$\arg\max_{\forall v_j \in \mathcal{V}} \frac{r_j}{q_j^{|c_j(\mathcal{R})|}} \left(\frac{1-p}{p} + q_j \right)^{|c_j(\mathcal{R})|} \quad (20)$$

  using the fact that q_j is assumed to be independent of y_j.

From (20), we see that for cases where $|c_j(\mathcal{R})| = 0$ the expression simplifies to $\arg\max_{\forall v_j \in \mathcal{V}} r_j$. This implies that instead of searching for all m values of the output space, only the N outputs of each PE and the maximum *a priori* value need to be evaluated. Thus, in this case the hypothesis space becomes the observation space with one more element, the maximum a priori value. Furthermore if $q_j = 0$ then (20) goes to infinity which implies that v_j is the correct output. The final soft voter equation is:

$$\arg\max_{\forall v_j \in \mathcal{H}} \frac{r_j}{q_j^{|c_j(\mathcal{R})|}} \left(\frac{1-p}{p} + q_j \right)^{|c_j(\mathcal{R})|}, \quad (21)$$

$$\text{where } \mathcal{H} = \mathcal{R} \cup \{\arg\max_{\forall v_j \in \mathcal{V}} r_j\} \quad (22)$$

Note that this simplification is exact, i.e., without any approximations. At an extreme, all the values needed to compute (20) can be precomputed and the resulting complexity of the soft voter is $O(N)$.

- Soft Voter using Soft NMR Error Model
  Under this scenario, the equality that leads from (19) to (20) is no longer valid.
  From (9), we note that $q_j = 0$ in (19). Also p^n is independent of j and can be omitted. The resulting equation is:

$$\arg\max_{\forall v_j \in \mathcal{V}} r_j \left(\prod_{l=1,...,m, l \neq j} q_l^{|c_l(\mathcal{R})|} \right) \left(\frac{1-p}{p} \right)^{|c_j(\mathcal{R})|} \quad (23)$$

  Here the product term will only involve a maximum of N multiplications. The prior r_j will be known to the voter, and $(\frac{1-p}{p})^{|c_j(\mathcal{R})|}$ can be precomputed and stored given that N is small. Thus, at a minimum, $N+2$ multiplications are needed in evaluating (23).
  However, it is not possible to reduce the hypothesis space in evaluating (23) as was done in the case of using the NMR error model, because q_l changes with every value of y_j. Assuming that at least one output is correct, limiting the hypothesis space to the N outputs (just the observation space) will be a reasonable approximation. The resulting final equation for the soft voter is:

$$\arg\max_{\forall v_j \in \mathcal{R}} r_j \left(\prod_{l=1,...,m, l \neq j} q_l^{|c_l(\mathcal{R})|} \right) \left(\frac{1-p}{p} \right)^{|c_j(\mathcal{R})|} \quad (24)$$

  Various simulation results prove that such an approximation has little impact on performance. The resulting complexity for the soft voter also becomes $O(N^2)$. As replication of blocks has a high overhead, N tends to be small, usually 3, which is feasible to implement.

3.4.2 Soft voter using MSE metric

As another example in deriving the soft voter, we consider the MSE metric. In this case, the cost function is $C(\hat{y}, y_o) = (\hat{y} - y_o)^2$. The posterior cost becomes

$$E(C(\hat{y}, Y_o)|y_1, ..., y_N) = \sum_{\forall v_j \in \mathcal{V}} P(v_j = y_o|y_1, ..., y_N)(\hat{y} - v_j)^2 \tag{25}$$

Assuming Gaussian error statistics, differentiating (25) and equating to zero, it is easily shown that (25) is minimized when $\hat{y}$ is the mean of the observations. Thus, the soft voter chooses $\hat{y}$ to equal the hypothesis closest to the mean. Assuming $\mathcal{H} = \mathcal{R}$, the soft voter functionality is described by:

$$\hat{y} = \arg\min_{\forall v_j \in \mathcal{R}} |v_j - \text{mean}(y_1, y_2, ..., y_N)| \tag{26}$$

For $N = 3$, the (26) reduces to:

$$\hat{y} = \text{median}(y_1, y_2, y_3) \tag{27}$$

These two examples demonstrate that the soft voter algorithm depends upon the system level performance metric and the error model. The soft voter algorithm can be simplified greatly at times if the error statistics can be described via a parametric form and if this form is exploited in the derivation of the soft voter.

3.5 Voter Architectures and Complexity

The architecture for the soft voter that minimizes $p_{e,sys}$ with $N = 3$ is shown in Fig. 7. The soft voter selects one of the PE outputs as the final output $\hat{y}$ only when all three inputs are equal. If not, the maximum *a posteriori* (MAP) block is activated. The MAP block is where actual computation of the soft voter occurs.

The MAP block architecture depends on the error model. Under the NMR error model, the MAP block implements (21), which is shown in Fig. 8(a), where the prior r_j and the value $\frac{1-p}{p}\frac{1}{q_j} + 1$ are stored in memory. Each value corresponding to the PE outputs is read from memory and cross-multiplied to compute the MAP equation (21). Finally, the value y_i that gives the maximum is chosen as the output.

The MAP block under the soft NMR error model which implements the computation in (24) is shown in Fig. 8(b) for a general value of N. In Fig. 8(b), the error PMF q_j and the prior r_j are stored in memory. Signals *hyp_sel* and *input_sel* select a hypothesis and an observation from Y to calculate e_i and hence the expression in (24). Thus, the power overhead of the soft voter will be small for small values of component error probability even though its gate complexity is large.

Finally, the soft voter architecture for minimizing the mean square error (MMSE) under Gaussian noise is shown in Fig. 9. An efficient majority word voter for a triple-MR (TMR) system can be found in [23].

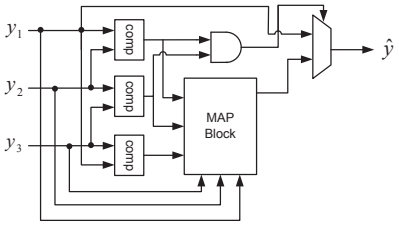
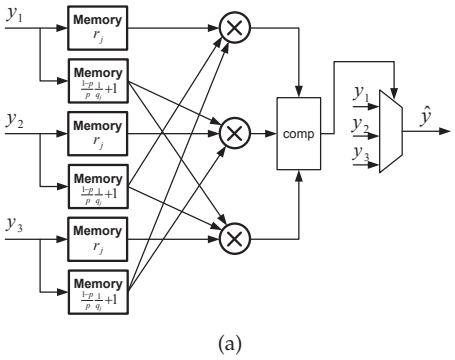
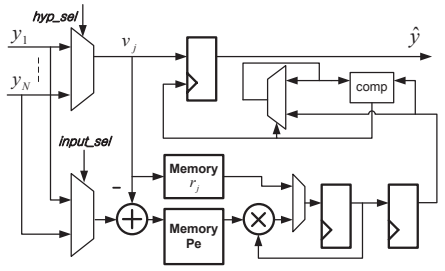


Fig. 7. Block diagram of top-level architecture of the soft voter.



(a)



(b)

Fig. 8. Architecture of MAP block: (a) using NMR error model, and (b) using soft NMR error model.

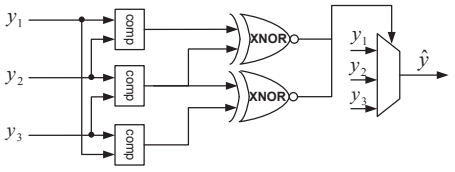
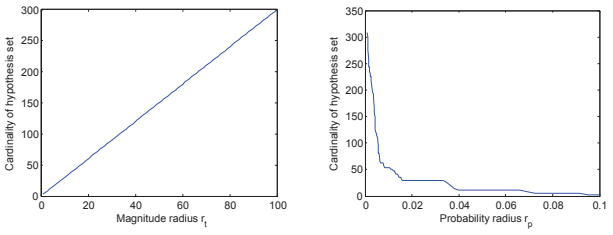


Fig. 9. Block diagram of median soft voter.

Given these voter architectures, the complexity of an n-bit majority voter and soft voter are compared in Table 2. For NMR, we consider triple modular redundancy (TMR), while for soft NMR, we consider two versions: *soft double modular redundancy (DMR)* and *soft TMR*. Unlike DMR, which can detect but cannot correct errors, soft DMR can detect and correct errors. Table 2 shows that the complexity of a soft voter increases exponentially with bit width n, because of the memory requirements for storing error statistics. We later show that soft NMR provides power savings in spite of its large complexity.

TABLE 2
Complexity comparison for n-bit voters.

	Complexity (transistor count)
TMR	$54n$
Soft DMR	$3n2^{n+2} + 52n^2 + 1300n$
Soft TMR	$3n2^{n+2} + 468n^2 + 1400n$



(a) (b)

Fig. 10. Plot of $|\mathcal{H}|$ vs. radius for $N = 3$ using: (a) magnitude-based hypothesis expansion, and (b) probability-based hypothesis expansion.

3.6 Hypothesis expansion

Both NMR and soft NMR will fail to correct errors if all N PE outputs y_i $(i = 1, \ldots, N)$ are in error. This is because both approaches choose one of the N PE outputs y_i as the corrected output $\hat{y}$, i.e., the hypothesis space $\mathcal{H}$ equals the observation space $\mathcal{Y}$. Unlike NMR, soft NMR can overcome this problem by expanding the hypothesis space $\mathcal{H}$ as discussed in Section 3.1. $\mathcal{H}$ is expanded by including values that are close to $\mathcal{Y}$ within a certain distance measure, i.e., within a radius (r_t or r_p). Two metrics for distance are used: (1) magnitude, and (2) probability. The relationship between the radius and size of $\mathcal{H}$ is linear for magnitude based expansion, while it depends on the probability distribution (Fig. 5) for probability based expansion. The size of $\mathcal{H}$ for a DCT increases exponentially with r_p as shown in Fig. 10.

The voter complexity also increases exponentially as it depends on the size of $\mathcal{H}$. Simulation results show that the performance does increase, and soft DMR approaches soft TMR (Fig. 21). The increase in performance is due to the fact that in some cases where all observations are in error, soft NMR is still able to correct errors. This is shown explicitly in Fig. 11, where one can see that TMR, soft TMR and soft TMR with hypothesis expansion (soft TMR-HE) all perform perfect error correction when only one error is present. However, when two errors are present, TMR fails catastrophically while both versions of soft TMR are able to correct about 90% of the errors. Furthermore, in the presence of three errors, soft TMR-HE is able to correct 39% errors even though TMR and soft TMR both fail completely.

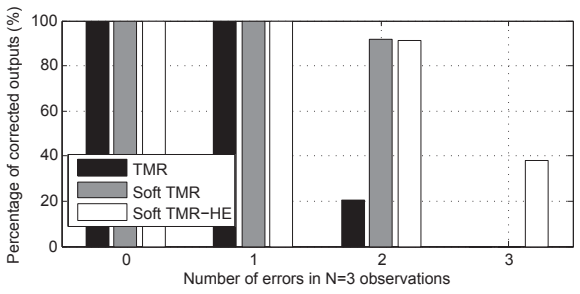


Fig. 11. Plot of percentage of corrected errors for $N = 3$ for a 2D-DCT image coder operated at $V_{dd} = 0.95V$. A probability radius r_p of 0.3 was used for soft TMR-HE.

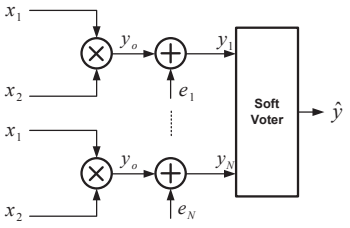


Fig. 12. The multiplier simulation setup.

3.7 Simulation Results

An 8-bit multiplier is employed as an example to demonstrate the benefits of soft NMR. The simulation setup is depicted in Fig. 12. The characterization and test input data are obtained independently from identical (uniform) distributions. Given the input distribution, The output distribution, i.e., the prior r_j is calculated from the input distribution.

Figure 13 shows the results of soft NMR vs. NMR majority and plurality voters with $N = 3$. The optimal soft voter for soft NMR is also included for comparison. Figure 13(a) shows the results with binomial noise. Figure 13(b) shows the results for a practical situation where the error statistics are those of a 16-bit RCA with timing errors [24]. Overall, reduction in $p_{e,sys}$ of $4\times$-to-$10\times$ can be achieved by soft NMR over NMR.

4 Statistical Analysis of Soft NMR, NMR and ANT

The benefits of soft NMR in Section 3 motivates us to analyze its performance systematically and compare it with ANT and NMR. Indeed, all three techniques achieve their robustness via a voter, which makes a decision based on a set of observations and other relevant information. In this section, statistical analysis of soft NMR, NMR and ANT is presented [25]. Analysis of NMR has been previously done under the NMR error model in [11], [20]. In this section, we will present an analysis of NMR and ANT under the soft NMR error model. The analysis assumes the statistical

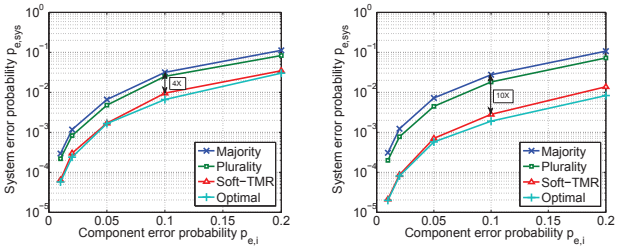


Fig. 13. Performance of soft NMR applied to a multiplier with: (a) e being binomial with parameters $P_{e_i}(k) = \binom{m}{k} p_{e_i}^k (1-p_{e_i})^{m-k}, m = 2^{16}$ and $p_{e_i} = 0.5$, and (b) e obtained from a 16-bit RCA using IBM 90 nm process with V_{dd} set to 66% of $V_{dd,crit}$.

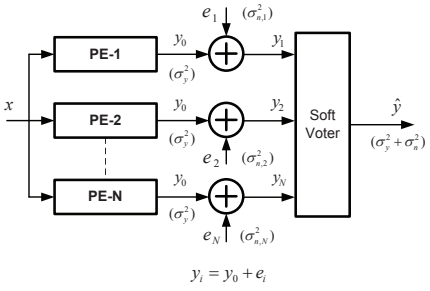


Fig. 14. Analysis framework for soft NMR. The power of each signal is written in parentheses.

properties of the characterization data and test data are identical.

4.1 Analysis Framework

Soft NMR, NMR and ANT have a common feature where several PEs perform computations, and then a voter combines the observation of all PEs to form the final outputs. Soft NMR and NMR have N identical PEs (see Fig. 1), whereas in ANT (see Fig. 2), there are 2 dissimilar PEs: the main block, and the estimator.

The analysis framework is depicted in Fig. 14 in the context of soft NMR, where N PEs compute in parallel to produce an output $y_i = y_o + e_i$, where y_o is the correct value, and e_i is the error. The notation used in this analysis is summarized in Table 1. In addition, we employ the notation σ_y^2 for the signal variance/power, and $\sigma_{n,i}^2$ for the variance/power of the error e_i.

Soft NMR assumes that the error statistics at the PE output are known. Figure 15(a) shows the error-statistics (error probability mass function (PMF)) due to timing violations at the output of a 16-b ripple carry adder. The error PMF indicates large magnitude errors to have a higher probability than small magnitude errors. In order to simplify the analysis, we assume the error PMF in Fig. 15(b). We define the probability of $e \neq 0$ as the error probability p_e, and the corresponding (large) error magnitude as d. In addition, we

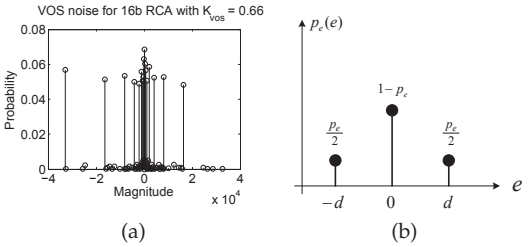


Fig. 15. Probability mass function (PMF) of errors due to timing violations: (a) error statistics from a 16-b RCA, and (b) simplified statistics used in analysis.

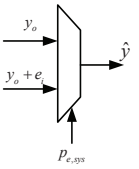


Fig. 16. The output of ANT, NMR and soft NMR modeled as a mixture distribution.

have e_i ranging from $-d$ to d, $-d \leq e_i \leq d$, we have $P_{e_i}(e_i = -d) = P_{e_i}(e_i = d) = \frac{p_e}{2}$, $P_{e_i}(e_i = 0) = 1 - p_e$ and $P_{e_i}(e_i) = 0$ for all other values of e_i.

Analysis is performed employing two metrics: (1) the system error probability $p_{e,sys}$, which is the probability that the final output $\hat{y} \neq y_o$ (see Fig. 14), i.e., $p_{e,sys} = 1 - P(\hat{y} = y_0)$, and (2) the signal-to-noise ratio (SNR), which is calculated as the ratio of the signal power to the noise power $\frac{\sigma_y^2}{E[(y_o - \hat{y})^2]}$, where the E operator denotes the expected value.

The distribution of output $\hat{y}$ can be viewed as a mixture distribution (see Fig. 16). Thus, the distribution of $\hat{y}$ is given by (see Table 1):

$$P_{\hat{y}}(v_j) = p_{e,sys}(r_j * P_{e_i}) + (1 - p_{e,sys}) r_j \tag{28}$$

where $*$ denotes the convolution operation.

It can be easily shown that the first and second moments of $\hat{y}$ are given by

$$E[\hat{y}] = E[y_o] + p_{e,sys} E[e_i] \tag{29}$$

$$E[(\hat{y} - E[\hat{y}])^2] = \sigma_{y_o}^2 + p_{e,sys}\sigma_{e_i}^2 + p_{e,sys}^2 E^2[e_i] \tag{30}$$

from which $SNR = \frac{\sigma_y^2}{E[(y_o - \hat{y})^2]}$ can be easily calculated.

4.2 Analysis of Soft NMR

The soft voter employs the *maximum a posteriori* (MAP) principle, which is optimal in the sense of minimizing the system error probability $p_{e,sys}$ by choosing the most probable value from a hypotheses set $\mathcal{H}$, given observations $\mathcal{Y}$, error statistics $P_{e_i}(e_i)$ along with the prior information r_j. The soft voter algorithm is given in (24).

An expression for the probability of error can be obtained by using the error statistics provided in Fig. 15(b). First, it should be noted that for a given

correct output, there are only three possible values the PE can produce: $y_o, y_o + d, y_o - d$. This is because the error PMF has only three error magnitudes that have nonzero probability. If all three values are observed, the soft voter can conclude the mid value is the correct output. Thus, in cases where all three possible values are observed, the soft voter will choose the correct output.

When only two distinct values are observed, there are two cases: (a) y_o never being observed: the soft voter has no way to estimate the correct output and will produce an erroneous output, and (b) y_o and one other value is observed: the soft voter chooses the correct value by taking into account the error probability and the priors. The detection rule is to choose the one with higher probability of occurrence. Assuming two values v_i and v_j are observed n times and $N-n$ times, respectively, the detection rule becomes:

$$r_i \left(\frac{p_e}{2}\right)^{N-n} (1-p_e)^n \underset{v_j}{\overset{v_i}{\gtrless}} r_j \left(\frac{p_e}{2}\right)^{n} (1-p_e)^{N-n} \tag{31}$$

Equation (31) can be further simplified to

$$n \underset{v_j}{\overset{v_i}{\gtrless}} \frac{N}{2} + \frac{1}{2} \frac{log\frac{r_j}{r_i}}{log\frac{2-2p_e}{p_e}} \tag{32}$$

where it is assumed that $p_e < \frac{2}{3}$ or else the direction of the inequality is reversed.

The trivial case is when only one value is observed. Either the correct value is observed resulting in no errors, or an erroneous value is observed leading to an error.

Combining these cases, the probability of error for soft NMR is:

$$p_{e,sys} = \sum_{i=1}^{m} r_i \left\{ p_e^N \left(1 - \frac{1}{2^{N-1}}\right) + 2 \sum_{k=0}^{\lfloor \frac{N}{2} + \frac{log\frac{r_j}{r_i}}{2log\frac{2-2p_e}{p_e}} \rfloor} \binom{N}{k} (1-p_e)^k \left(\frac{p_e}{2}\right)^{N-k} \right\} \tag{33}$$

If the priors r_j are uniform, (32) simplifies to the majority voter:

$$n \underset{v_j}{\overset{v_i}{\gtrless}} \frac{N}{2} \tag{34}$$

In this case, the system error probability simplifies to:

$$p_{e,sys} = 2 \sum_{i=0}^{\frac{N}{2}} \binom{N}{i} (1-p_e)^i \left(\frac{p_e}{2}\right)^{N-i} + p_e^N - 2\left(\frac{p_e}{2}\right)^N \tag{35}$$

which is the well-known NMR system error probability.

In general, a closed form expression for the $p_{e,sys}$ of the soft voter is difficult to obtain without specific knowledge of the error statistics. Instead, we present a numerical procedure to compute $p_{e,sys}$. For an arbitrary error PMF, $p_{e,sys}$ is given by

$$p_{e,sys} = \sum_{v_j \in \mathcal{V}} r_j \left(\sum_{e_i \in \mathcal{A}} P_{e_i}(e_i) \right) \tag{36}$$

$$\mathcal{A} = \left\{ e_i : r_j \left(\prod_{l=1,\ldots,m, l \neq i} q_l^{|c_l(R)|} \right) \left(\frac{1-p_{e_i}}{p_{e_i}} \right)^{|c_i(R)|} > r_j \left(\prod_{l=1,\ldots,m, l \neq j} q_l^{|c_l(R)|} \right) \left(\frac{1-p_{e_i}}{p_{e_i}} \right)^{|c_j(R)|} \right\} \tag{37}$$

which is easier to compute than Monte Carlo simulations. This process can be time-consuming for output spaces with large cardinality as the complexity is $O(m^N)$; however, it is faster to compute at low probability of error values than Monte Carlo simulations.

Substituting (33) in (29) and (30) provides the SNR estimate.

4.3 Analysis of NMR

In this section, we present the analysis of NMR under the soft NMR error model. We have chosen the most popular voting schemes, majority and median, as our target of analysis. Analysis of the voters using the NMR error model has been previously done in [20]. Here we present the analysis using complete statistical information via the more general soft NMR error model.

4.3.1 Majority

When presented with a set of N PE outputs $Y = \{y_1, y_2, ..., y_N\}$, a majority voter produces an output $\hat{y}$ given by

$$\hat{y} = maj(y_1, y_2, ..., y_N) \tag{38}$$

where $maj(Y)$ selects that element of Y which occurs more than $\lfloor N/2 \rfloor$ times. In the absence of a majority, the element with the most occurrences can be chosen or an error can be flagged.

Each PE produces the correct output y_o with a probability $1 - p_{e_i}$ independently of other processes. As the probability of choosing the correct output can be given as

$$\sum_{v_j \in V} P\left(c_j(\mathcal{R}) > \frac{N}{2} \middle| v_j = y_o \right) P(v_j = y_o) \tag{39}$$

the error probability of the majority voter becomes:

$$p_{e,sys} = \sum_{v_j \in V} r_j \sum_{k=0}^{\lfloor \frac{N}{2} \rfloor} \binom{N}{k} (p_{e_i})^{N-k} (1-p_{e_i})^k \tag{40}$$

4.3.2 Median

When presented with a set of N PE outputs $Y = \{y_1, y_2, ..., y_N\}$, a median voter produces an output $\hat{y}$ given by

$$\hat{y} = med(y_1, y_2, ..., y_N) \tag{41}$$

where $med(Y)$ selects that element of Y which is the median of all values. Assuming $\{y_1, y_2, ..., y_N\}$ are ordered in increasing order and N is odd, then the median will be $y_{\lfloor \frac{N}{2} \rfloor}$.

For v_j to be the median, $\frac{N-1}{2}$ outputs need to be less than v_j and the other $\frac{N-1}{2}$ need to be greater than v_j. Therefore, the probability of median giving the correct value can be derived as follows:

$$\begin{aligned}
&\sum_{v_j \in V} r_j P \left(\sum_{\forall v_k < v_j} c_k(\mathcal{R}) = \frac{N-1}{2} \bigcap \right. \\
&\left. \sum_{\forall v_k > v_j} c_k(\mathcal{R}) = \frac{N-1}{2} \middle| v_j = y_o \right) \\
&= \sum_{v_j \in V} r_j \left[1 - P \left(\sum_{\forall v_k < v_j} c_k(\mathcal{R}) > \frac{N}{2} \middle| v_j = y_o \right) \right. \\
&\qquad \left. - P \left(\sum_{\forall v_k > v_j} c_k(\mathcal{R}) > \frac{N}{2} \middle| v_j = y_o \right) \right]
\end{aligned}$$

The error probability is $1 - P(correct)$ and thus is given by:

$$\begin{aligned}
&\sum_{v_j \in V} r_j \left[P \left(\sum_{\forall v_k < v_j} c_k(\mathcal{R}) > \frac{N}{2} \middle| v_j = y_o \right) + \right. \\
&\qquad \left. P \left(\sum_{\forall v_k > v_j} c_k(\mathcal{R}) > \frac{N}{2} \middle| v_j = y_o \right) \right]
\end{aligned} \tag{42}$$

And each term in the square brackets will be the probability that $\mathcal{R}$ has more than $\frac{N}{2}$ being greater or less than the correct output y_o. Therefore, the error probability of a median voter is given as:

$$\begin{aligned}
p_{e,sys} = \sum_{v_j \in V} r_j \Bigg\{ & \sum_{k=\lceil \frac{N}{2} \rceil}^{N} \binom{N}{k} \left[p_{e_i} \sum_{l=1}^{j-1} q_l \right]^k \\
& \left[1 - p_{e_i} \sum_{l=1}^{j-1} q_l \right]^{N-k} && (43) \\
& + \sum_{k=\lceil \frac{N}{2} \rceil}^{N} \binom{N}{k} \left[p_{e_i} \sum_{l=j+1}^{m} q_l \right]^k && (44) \\
& \left[1 - p_{e_i} \sum_{l=j+1}^{m} q_l \right]^{N-k} \Bigg\} && (45)
\end{aligned}$$

As is the case with soft NMR, substituting (40) or (45) in (29) and (30) will give SNR estimates.

4.4 Analysis of ANT

ANT can be viewed as a special case of NMR with two PEs, a main block (PE-1) and an estimator (PE-est), each with different error statistics (see Fig. 2), and the detector chooses among the two PE outputs by comparing $|y_1 - y_{est}|$ to a threshold τ. In this analysis, we will assume the estimation error is $-d/2 \leq e_{est} \leq d/2$, i.e., the estimation error is smaller than hardware errors.

Considering the error PMF in Fig. 15(b), if $\tau \geq \frac{3d}{2}$, then PE-1 will be chosen regardless of PE-est. If $\tau < \frac{d}{2}$, then PE-1 is chosen only when PE-1 is error-free (which is the correct output), else PE-est is chosen. When $\frac{d}{2} \leq \tau < \frac{3d}{2}$, $p_{e,sys}$ will depend on the values of e_1 and e_{est}. Thus, $p_{e,sys}$ for ANT can be calculated as:

$$p_{e,sys} = \begin{cases} p_e(1 - P_{e_{est}}(0)), & \text{when } \tau < \frac{d}{2} \\ \frac{p_e}{2} \sum_{d-\tau \leq |e_{est}| \leq \frac{d}{2}} P_{e_{est}}(e_{est}), & \text{when } \frac{d}{2} \leq \tau < d \\ \frac{p_e}{2} \{1 - P_{e_{est}}(0) + \\ \sum_{0 < |e_{est}| \leq \tau - d} P_{e_{est}}(e_{est}) \}, & \text{when } d \leq \tau < \frac{3d}{2} \\ p_e, & \text{when } \tau \geq \frac{3d}{2} \end{cases} \tag{46}$$

The probability of error of ANT for an arbitrary error statistics is:

$$1 - P(E_1) - P(E_2) \tag{47}$$

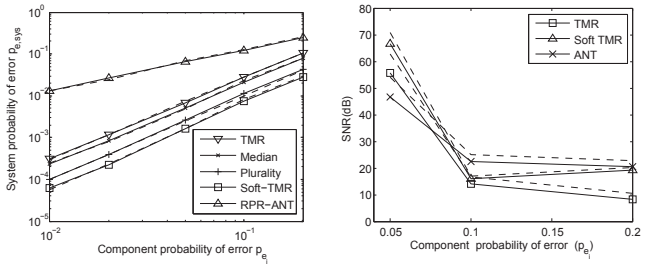
where E_1 is the event when $\hat{y} = y_1$ and $e_1 = 0$, and E_2 is the event when $\hat{y} = y_{est}$ and $e_{est} = 0$. Thus (47) becomes:

$$p_{e,sys} = 1 - P_{e_1}(0) \sum_{|e_{est}| < \tau} P_{e_{est}}(e_{est}) - P_{e_{est}}(0) \sum_{|e_1| > \tau} P_{e_1}(e_1) \tag{48}$$

4.4.1 Reduced precision redundancy

Reduced precision redundancy (RPR) is a technique where the estimator uses smaller bit precision to be able to produce the correct result in a shorter time, which enables the estimator to be free from timing errors. However as the precision is reduced, there is a small estimation error. Assuming the lower bits have a uniform distribution, the estimation error will also have a uniform distribution. Thus, in RPR, $P_{e_{est}}$ can be assumed to be uniform, and its magnitude will depend on the number of bits by which the precision was reduced. If the precision was reduced by b-bits, the estimation error statistics will become:

$$P_{e_{est}} = \frac{1}{2^b} \tag{49}$$



(a) (b)

Fig. 17. Comparison of analytical (dashed) and simulation (bold) results for an 8-bit multiplier: (a) $p_{e,sys}$ metric, and (b) SNR metric.

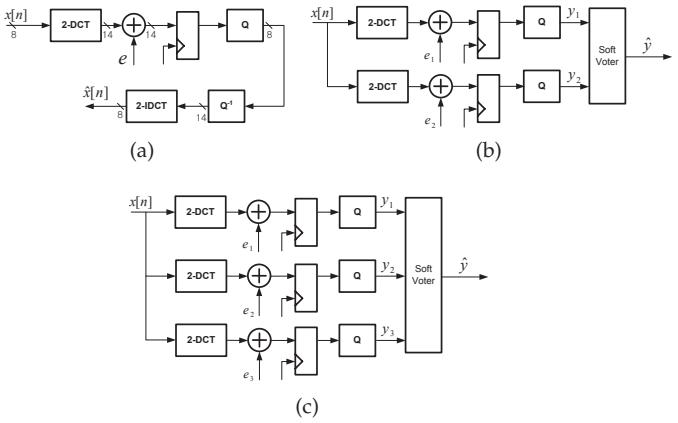
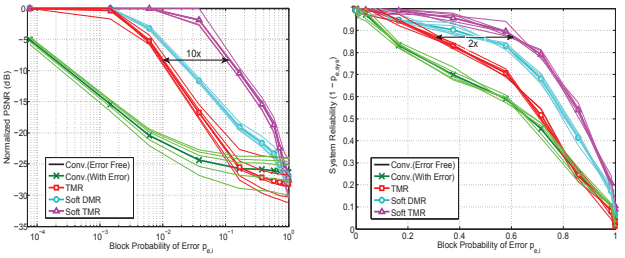


Fig. 18. DCT-based image compression architecture: (a) conventional architecture, (b) soft DMR, and (c) soft TMR.



(a) (b)

Fig. 19. Normalized PSNR and reliability vs. component probability of error p_{e_i} over five test images (bold lines represent the mean over the five test images): (a) normalized PSNR vs. p_{e_i}, and (b) system reliability $(1 - p_{e,sys})$ vs. p_{e_i}.

4.5 NMR and Soft NMR Comparison

Subtracting (35) from (40), we obtain:

$$p_{e,sys-NMR} - p_{e,sys-softNMR} = \sum_{k=0}^{\lfloor \frac{N}{2} \rfloor} \binom{N}{k} \left[p_e^{N-k}(1-p_e)^k \left\{ 1 - \left(\frac{1}{2}\right)^{N-k-1} \right\} - 2\left(\frac{p_e}{2}\right)^N + 4\left(\frac{p_e}{4}\right)^N \right] \quad (50)$$

As the summand in the summation is always a non-negative quantity, we see that soft NMR will always outperform NMR under this error statistics.

4.6 Simulation Results

Monte Carlo simulation is used to compare to the results obtained from analysis. We show the results for an 8-bit multiplier. The simulation setup for the multiplier is the same as in Section 3 (see Fig. 12). We employed a 6-bit reduced precision version of the multiplier as an estimator for ANT. Figure 17 shows that the analysis predicts the simulation results to within 0.1% for $p_{e,sys}$ and 2 dB for SNR, on average.

5 Application: DCT-based Image Compression

In this section, we compare soft NMR with NMR in terms of robustness, and energy-efficiency in the context of a 2D-DCT system. Figure 18 shows the various 2D-DCT architectures being considered. We replicate the DCT block and perform voting after the quantizer. Chen's algorithm [21] is used for deriving the DCT architecture and the quantizer (Q) employs the JPEG quantization table [26]. Only the DCT blocks are subject to VOS, and hence these are the only blocks that exhibit errors. All voters are operated at their critical supply voltage of $V_{dd-crit} = 0.7V$ to ensure correct operation. The quantizer, the inverse quantizer and the inverse DCT (IDCT) are all assumed to be error-free in order to isolate the effects of DCT errors. Errors are captured just before the latch at the DCT outputs in Fig. 18. These errors are independent via the use of techniques discussed in 3.3. The soft voter employs (24), which minimizes $p_{e,sys}$, and the majority voter employs (38). One characterization image was employed to determine the prior and error statistics, while five test images ($I1, I2, I3, I4$ and $I5$) were employed to evaluate performance. The error-free PSNRs for these images were $PSNR_{I1} = 28.57dB$, $PSNR_{I2} = 33.01dB$, $PSNR_{I3} = 31.89dB$, $PSNR_{I4} = 32.10dB$, and $PSNR_{I5} = 32.16dB$.

5.1 Robustness

In order to compare the system performance across the five images, we plot the normalized PSNR in Fig. 19(a), where the PSNRs achieved by any technique for a specific image is normalized with respect to the error-free PSNR for that image, i.e., the error-free PSNR is subtracted from the actual PSNR. With $\mathcal{H} = \mathcal{Y}$, Fig. 19(a) shows that, for a wide range of PSNRs ($15dB - 30dB$), soft TMR can tolerate approximately $10\times$ higher component probability of error p_{e_i}

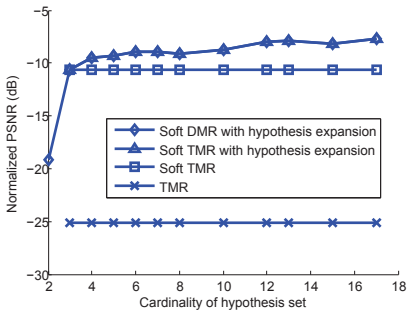


Fig. 21. Increase in PSNR vs. $\mathcal{H}$ for probability based expansion for image $I1$ and $p_{e,i} = 0.146$. Soft TMR and TMR have $|\mathcal{H}| = 3$.

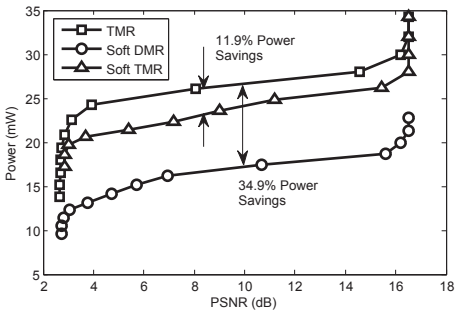


Fig. 22. Power consumption for TMR, soft DMR and soft TMR vs. PSNR.

than TMR at the same PSNR for multiple test images. Figure 19(b) shows that soft TMR is $2\times$ more robust than TMR at the same system level reliability. More interestingly, soft DMR outperforms TMR at all values of p_{e_i} (and hence voltages) even though soft DMR has $0.65\times$ the complexity (including voter complexity) of TMR. This is remarkable as it suggests that soft DMR is a viable low-complexity(power) alternative to TMR with no loss in robustness. Thus, for an application whose main block is extremely complex, instead of triplicating the system, we are able to duplicate it and employ a more sophisticated voter to achieve better performance than TMR.

Figure 20 shows the reconstructed images at the IDCT output for various techniques. It is clear that TMR is hardly able to recover from errors while, soft DMR and soft TMR perform significantly better.

5.2 Hypothesis expansion

Figure 21 plots the normalized PSNR as the cardinality of $\mathcal{H}$ ($|\mathcal{H}|$) is increased. Probability-based hypothesis expansion improves the PSNR of both soft TMR and soft DMR. The PSNR improvement for soft DMR is remarkable (approx $10dB$) bringing its PSNR on par with that of soft TMR. This is because probability based hypothesis expansion estimates and includes the most probable observations into $\mathcal{H}$, thereby reducing the difference between the hypothesis sets, and hence the performance, of soft DMR and soft TMR. This result also shows that soft NMR is capable of correcting errors even when all the observed values ($\mathcal{Y}$) are incorrect.

5.3 Power Savings

Power numbers were obtained via circuit simulations (HSPICE) of logic blocks, and via CACTI [27] for memory including leakage power. Figure 22 shows the total power consumed for a given PSNR. Power consumption of the DCTs, quantizers and the voters, i.e., the entire transmitter, were included in these comparisons. It can be seen that soft TMR achieves 10% to 15% power savings and soft DMR achieves 30% to 40% power savings compared to TMR over a wide range of PSNRs.

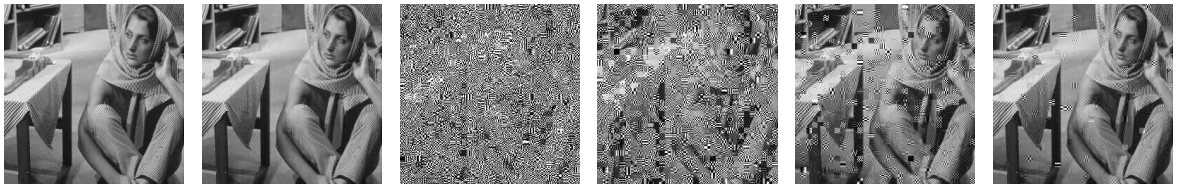
TABLE 3
Voter complexity and power overhead for 8-bit voters compared to a TMR DCT transmit chain.

	Complexity (transistor count)	Complexity Overhead	Power Overhead
TMR	432	0.1%	0.07%
Soft DMR	35848	8.2%	0.25%
Soft TMR	64868	14.9%	1.49%

Table 3 shows that the soft voter, though much more complex than the majority voter, has a very low power overhead with respect to the DCT transmit chain. This is primarily because the MAP block in Fig. 7(b) is activated infrequently and a majority of the complexity comes from memory which consumes little power. This indicates that expending computational resources in exploiting statistics is an effective way of reducing power.

6 Conclusion and Future Work

Soft NMR has been shown to improve the robustness and power efficiency over conventional NMR by explicit use of error statistics. This has been done by employing a soft voter that is based on detection techniques. Two detection criteria were explored, minimizing probability of error, and also minimizing mean squared error. Analysis of soft NMR was performed to show its benefits over NMR. The accurate results obtained through analysis shows that it can also be employed in designing robust systems. Soft NMR was then applied to a DCT image coding application. Simulations in a commercial $45nm$, $1.2V$, CMOS process show that soft triple-MR (TMR) provides $10\times$ improvement in robustness, and 12% power savings over TMR at a peak signal-to-noise ratio (PSNR) of $20dB$. In addition, soft dual-MR (DMR) provides $2\times$ improvement in robustness, and 35% power savings over TMR at a PSNR of $20dB$. This work opens up a number of interesting problems to explore including:



(a) Original Image (b) Error Free (c) With Errors (d) TMR (e) Soft DMR (f) Soft TMR

Fig. 20. Reconstructed image at IDCT output with $V_{dd} = 0.95V$ and $p_{e_i} = 14.63\%$: (a) original image, (b) error-free conventional with no errors (PSNR = $33dB$), (c) conventional with VOS errors (PSNR = $6.15dB$), (d) TMR (PSNR = $7.88dB$), (e) Soft DMR (PSNR = $14.09dB$), and (f) Soft TMR (PSNR = $22.25dB$).

a) algorithms of approximating the optimal bound, b) using time and space correlation statistics, c) methods of efficiently storing the statistical information and the impact of finite precision, and d) methods of obtaining the statistical information.

REFERENCES

[1] International technology roadmap for semiconductors 2008 update. International Technology Roadmap for Semiconductors. [Online]. Available: http://www.itrs.net/Links/2008ITRS/Home2008.htm

[2] Y. Cao and L. Clark, "Mapping statistical process variations toward circuit performance variability: An analytical modeling approach," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 26, no. 10, pp. 1866–1873, Oct. 2007.

[3] D. Ernst et al., "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture*, Dec. 2003, pp. 7–18.

[4] R. Hegde and N. R. Shanbhag, "Soft digital signal processing," *IEEE Trans. VLSI Syst.*, vol. 9, no. 6, pp. 813–823, Dec. 2001.

[5] G. V. Varatkar, S. Narayanan, N. R. Shanbhag, and D. Jones, "Sensor network-on-chip," in *2007 International Symposium on System-on-Chip*, Nov. 2007, pp. 1–4.

[6] E. Karl, D. Sylvester, and D. Blaauw, "Timing error correction techniques for voltage-scalable on-chip memories," in *2005 International Symposium on Circuits and Systems (ISCAS)*, May 2005, pp. 3563–3566 Vol. 4.

[7] G. V. Varatkar and N. R. Shanbhag, "Error-resilient motion estimation architecture," *IEEE Trans. VLSI Syst.*, vol. 16, no. 10, pp. 1399–1412, Oct. 2008.

[8] R. Abdallah and N. Shanbhag, "Error-resilient low-power viterbi decoder architectures," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4906–4917, Dec. 2009.

[9] J. Von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," *Automata Studies*, pp. 43–98, 1956.

[10] W. Brown, J. Tierney, and R. Wasserman, "Improvement of electronic-computer reliability through the use of redundancy," *IEEE Trans. Electron. Comput.*, vol. 10, pp. 407–416, 1961.

[11] N. Vaidya and D. Pradhan, "Fault-tolerant design strategies for high reliability and safety," *IEEE Trans. Comput.*, vol. 42, no. 10, pp. 1195–1206, Oct. 1993.

[12] I. Koren and S. Su, "Reliability analysis of N-modular redundancy systems with intermittent and permanent faults," *IEEE Trans. Comput.*, vol. C-28, no. 7, pp. 514–520, Jul. 1979.

[13] Y. Tamir, M. Tremblay, and D. Rennels, "The implementation and application of micro rollback in fault-tolerant VLSI systems," in *Proceedings of IEEE Fault-Tolerant Computing Symposium*, 1988, pp. 234–239.

[14] S. J. Piestrak, "Design of fast self-testing checkers for a class of berger codes," *IEEE Trans. Comput.*, vol. 36, no. 5, pp. 629–634, 1987.

[15] A. Avizienis, "Arithmetic error codes: Cost and effectiveness studies for application in digital system design," *IEEE Trans. Comput.*, vol. 20, no. 11, pp. 1322–1331, Nov. 1971.

[16] N. R. Shanbhag, "Reliable and efficient system-on-a-chip design," *IEEE Computer*, vol. 37, no. 3, pp. 42–50, Mar. 2004.

[17] E. P. Kim, R. A. Abdallah, and N. R. Shanbhag, "Soft NMR: Exploiting statistics for energy-efficiency," in *2009 International Symposium on System-on-Chip (SOC)*, Oct. 2009, pp. 52–55.

[18] P. Huber, *Robust Statistics*. New York, NY: Wiley, 1981.

[19] H. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY: Springer-Verlag, 1994.

[20] D. Blough and G. Sullivan, "A comparison of voting strategies for fault-tolerant distributed systems," in *Proceedings of the 1990 Ninth Symposium on Reliable Distributed Systems*, Oct. 1990, pp. 136–145.

[21] W.-H. Chen, C. Smith, and S. Fralick, "A fast computational algorithm for the discrete cosine transform," *IEEE Trans. Commun.*, vol. 25, no. 9, pp. 1004–1009, Sep. 1977.

[22] J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2002.

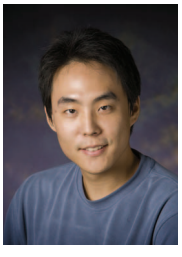
[23] S. Mitra and E. McCluskey, "Word-voter: A new voter design for triple modular redundant systems," in *Proceedings of the 18th IEEE VLSI Test Symposium*, 2000, pp. 465–470.

[24] C. T. Kong, "Study of voltage and process variations impact on the path delays of arithmetic units," M.S. thesis, University of Illinois at Urbana-Champaign, Urbana, IL, May 2008.

[25] E. P. Kim and N. R. Shanbhag, "Soft NMR: Analysis & application to DSP systems," in *2010 Proc. International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2010, pp. 1494 –1497.

[26] International Telecommunication Union, "JPEG standard," ITU-T Recommendation T.81, 1993.

[27] S. Thoziyoor, N. Muralimanohar, and N. Jouppi, "CACTI 5.0," Hewlett Packard Laboratories, Palo Alto, CA, Tech. Rep. HPL-2007-167, 2007.



Eric P. Kim Eric P. Kim is a IEEE student member. He has received his B.S. degree from Seoul National University in 2004 in electrical engineering. From 2004 to 2005 he has worked in Bluebird Soft, a PDA company as a key architect in hardware design. He has also worked in Alticast, a digital broadcasting solution company as a software engineer developing standards on digital multimedia broadcasting (DMB) and digital video broadcasting (DVB). Since 2007, he has joined the University of Illinois at Urbana-Champaign and received his M.S. degree in 2009 in electrical and computer engineering. He is now continuing his pursuit towards a Ph.D. degree. His research interests encompass a wide range of topics, including error tolerant and energy efficient VLSI architectures, digital signal processing for communication systems, and distributed estimation and detection. He is particularly interested in error-tolerant multimedia and communication systems design. Mr. Kim was awarded the Margarida Jacome best poster award at the 2009 Gigascale Systems Research Center (GSRC) annual symposium.

Naresh R. Shanbhag Naresh R. Shanbhag received his Ph.D. degree from the University of Minnesota in 1993 in Electrical Engineering. From 1993 to 1995, he worked at AT&T Bell Laboratories at Murray Hill where he was the lead chip architect for AT&T's 51.84 Mb/s transceiver chips over twisted-pair wiring for Asynchronous Transfer Mode (ATM)-LAN and very high-speed digital subscriber line (VDSL) chip-sets. Since August 1995, he is with the Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory where he is presently a Professor. He was on a sabbatical leave of absence at the National Taiwan University in Fall 2007. His research interests are in the design of integrated circuits and systems for communications including low-power/high-performance VLSI architectures for error-control coding, equalization, as well as integrated circuit design. He has more than 150 publications in this area and holds four US patents. He is also a co-author of the research monograph Pipelined Adaptive Digital Filters published by Kluwer Academic Publishers in 1994. Dr. Shanbhag is leading research themes in the DOD and Semiconductor Research Corporation (SRC) sponsored Microelectronics Advanced Research Corporation (MARCO) center under their Focus Center Research Program (FCRP) since 2006.

Dr. Shanbhag became an IEEE Fellow in 2006, received the 2006 IEEE Journal of Solid-State Circuits Best Paper Award, the 2001 IEEE Transactions on VLSI Best Paper Award, the 1999 IEEE Leon K. Kirchmayer Best Paper Award, the 1999 Xerox Faculty Award, the Distinguished Lecturership from the IEEE Circuits and Systems Society in 1997, the National Science Foundation CAREER Award in 1996, and the 1994 Darlington Best Paper Award from the IEEE Circuits and Systems Society.

Dr. Shanbhag served as an Associate Editor for the IEEE Transaction on Circuits and Systems: Part II (97-99) and the IEEE Transactions on VLSI (99-02 and 09-present), respectively. He is the technical program chair of the 2010 IEEE International Symposium on Low-Power Design (ISLPED), and is currently serving on the technical program committees of the International Solid-State Circuits Conference (ISSCC), the International Conference on Acoustics, Speech and Signal Processing (ICASSP), and others.

In 2000, Dr. Shanbhag co-founded and served as the chief technology officer of Intersymbol Communications, Inc., a venture-funded fabless semiconductor start-up that provides DSP-enhanced mixed-signal ICs for electronic dispersion compensation of OC-192 optical links. In 2007, Intersymbol Communications, Inc., was acquired by Finisar Corporation, Inc., where Dr. Shanbhag continues to provide counsel on technology strategy.