Soft NMR: Exploiting Statistics for Energy-Efficiency

Eric P. Kim, Rami A. Abdallah, and Naresh R. Shanbhag
Coordinated Science Laboratory / Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
1308 W Main St., Urbana, Illinois, USA, 61801
{epkim2, rabdall3, shanbhag}@illinois.edu

***Abstract*—Achieving energy-efficiency in nanoscale CMOS process technologies is made challenging due to the presence of process, temperature and voltage variations. In this paper, we present *soft* N-modular redundancy (*soft NMR*) that consciously exploits statistics of errors due to these nanoscale artifacts in order to design robust and energy-efficient systems. In contrast to conventional NMR, soft NMR employs *estimation* and *detection* techniques in the voter. We compare NMR and soft NMR in the design of an energy-efficient and robust discrete cosine transform (DCT) image coder. Simulations in a commercial $45nm$, $1.2V$, CMOS process show that soft triple-MR (TMR) provides 10× improvement in robustness and 13% power savings over TMR at a peak signal-to-noise ratio (PSNR) of $20dB$. In addition, soft dual-MR (DMR) provides 2× improvement in robustness and 35% power savings over TMR at a PSNR of $20dB$.**

I. Introduction

Modern nanoscale CMOS exhibit a number of artifacts such as process, temperature and voltage variations, leakage, and soft errors due to particle hits, just to name a few. As a result, simultaneously achieving robustness and energy-efficiency is a challenge. Worst-case designs address the robustness issue but with a severe power penalty. Nominal-case design, though energy-efficient, suffer from reliability problems. Error-resiliency is an attractive approach towards achieving robust and energy-efficient operation in nanoscale CMOS. Techniques such as algorithmic noise-tolerance [1], [2], and [3] exploit the algorithmic structure of the computation in order to reduce power, and thus are application-specific.

N-modular redundancy (NMR) [4] (see Fig. 1(a)) is a well-known and general fault-tolerant technique that provides robustness to critical applications such as military, medical and server applications, but comes with a 2× complexity and power overhead. In NMR, a computation is replicated N times and the outputs are majority voted upon to select the correct one. NMR ignores error-statistics exhibited by the N processing elements (PEs).

In this paper, we propose *soft NMR* (see Fig. 1(b)) to improve robustness and energy-efficiency of NMR, while preserving its generality. Structurally, soft NMR differs from NMR in that it incorporates a *soft voter*, which is composed of an *estimator* and a *detector*. Thus, soft NMR views computation in the PEs as a noisy communication channel, and employs the estimator as an equalizer, and 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×-to-10× improvement in robustness accompanied 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.

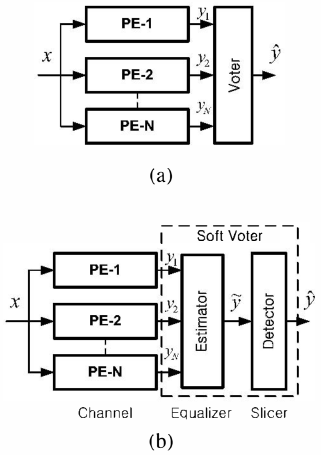


Fig. 1. Block diagram of: (a) NMR, and (b) soft NMR.

In this paper, we describe the soft NMR architecture in section II, followed by a characterization of error-statistics in section III. In section IV, an application of soft NMR to a discrete cosine transform (DCT)-based image coder is demonstrated with simulation results shown in section V. Finally section VI concludes the paper with future research directions.

II. Soft NMR Architecture

The key difference between conventional NMR and soft NMR is in the voter. Conventional NMR employs a majority voter though there are other voters such as the plurality voter. The soft voter fundamentally differs from conventional voters as it explicitly exploits statistical information. This section presents, the conventional majority voter and the soft voter in more detail.

978-1-4244-4467-0/09/$25.00 ©2009 IEEE

A. The Majority Voter

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,

$$\hat{y} = maj(y_1, y_2, ..., y_N) \tag{1}$$

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 could be chosen (plurality) or an error can be flagged. An efficient majority word voter for a TMR system [5] is shown in Fig. 2(a).

B. The Soft Voter

The soft voter employs the *maximum a posteriori* (MAP) principle [6], 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 H, given observations Y, and error statistics $P_e()$. The soft voter algorithm is obtained by choosing $H = Y$ for complexity reasons, and assuming the knowledge of $P_e()$, is given as follows:

$$\hat{y} = \underset{v_j \in \{y_1, ..., y_N\}}{\arg\max} \; r_j P_e((e_1 = y_1 - v_j), ..., (e_N = y_N - v_j)) \tag{2}$$

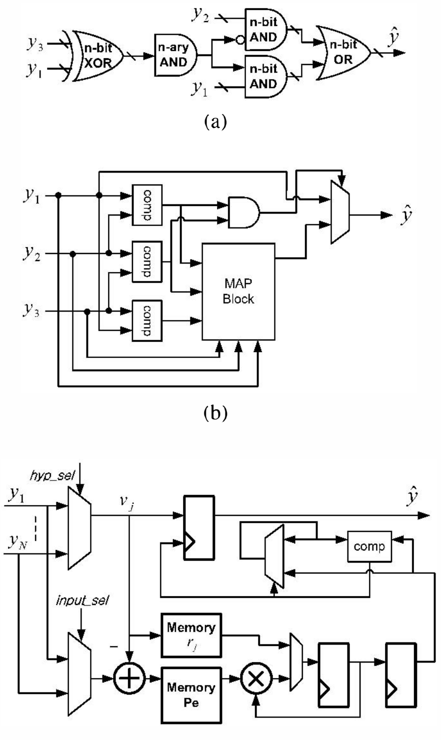
where $H = Y = \{y_1, y_2, ..., y_N\}$ are the outputs of the N PEs, r_j is the *a priori* probability that v_j is the correct output, and $P_e(e_1, ..., e_N)$ is the joint error probability mass function (PMF). Figure 2(b) shows the soft voter architecture for $N = 3$. Here, the soft voter selects one PE outputs as the final output $\hat{y}$ only when all three inputs are equal. If not, the MAP block is activated. The MAP block, which implements the computation in (2), is shown in Fig. 2(c) for a general value of N. In Fig. 2(c), the error PMF $P_e()$ 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 (2). 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.

III. Error Characterization

Soft NMR requires the knowledge of the error statistics $P_e()$. In addition, both NMR and soft NMR work best when the individual PE errors e_i are independent. In this section, we describe a methodology to obtain $P_e()$, and techniques to sure the independence of errors.

A. Simulation Methodology

We assume that block errors are due to timing violations. For example, timing errors in the PEs can be generated via voltage overscaling (VOS) [1], where the supply voltage is set to be lower than the critical voltage ($V_{dd,crit}$), i.e., the voltage at which the PE is error-free. We first characterize basic building blocks such as a 1-b full-adder (FA) in a commercial $1.2V$, $45nm$ process technology using ***HSPICE*** to obtain delay vs. supply voltage curves. Next, a register transfer-level (RTL) model of the PE architecture, a 2-D DCT in this case, is developed in *Verilog* which incorporates the delay values at a specific supply voltage. The error PMF $P_e()$ is obtained by comparing the correct output and the erroneous output obtained through RTL simulations at various supply voltages using typical input sequences.



(a)

(b)

(c)

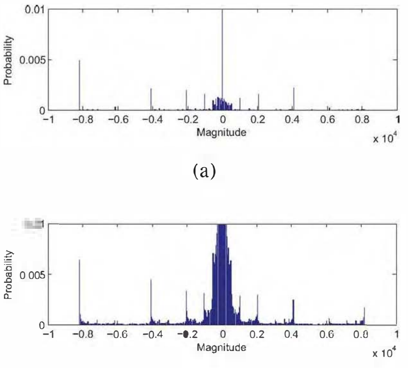
Fig. 2. Voter architectures: (a) majority voter, (b) soft voter, and (c) block diagram of the MAP block.

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 [7], employing mirror adders and array multipliers [8], implemented in a commercial $45nm$, $1.2V$ CMOS process, is shown in Fig. 3 for two different supply voltages. Figure 3 shows that the error PMFs become spiky as the supply voltage is reduced, and that a few large amplitude errors have a high probability of occurrence. This is to be expected as the DCT architecture computes LSB-first, and hence timing errors appear first in the MSBs, i.e., large amplitude error will occur. We will employ the error PMFs in Fig. 3 in Section V to study the power vs. robustness trade-offs.

The component error probability p_{e_i} of the DCT block due to VOS is shown in Fig. 4, where we find that p_{e_i} increases rapidly as the supply voltage is reduced beyond $V_{dd-crit}$. This plot was obtained from a structural *Verilog* simulations of the DCT architecture at various supply voltages (hence delays) but with a fixed clock frequency.

B. Independence of Errors

As mentioned earlier, independent errors are essential for NMR and soft NMR to work well. If all the PEs in Fig. 1 have identical architectures and inputs, then the errors will be highly correlated or even identical. However, timing errors can



(a)

(b)

Fig. 3. Error statistics of a voltage overscaled DCT block in a $45nm$, $1.2V$ CMOS process with $V_{dd,crit} = 1.2V$: (a) $V_{dd} = 1V$ (probability of error is 0.0374), and (b) $V_{dd} = 0.8V$ (probability of error is 0.7142).

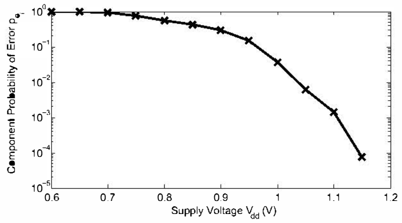


Fig. 4. Component error probability p_{e_i} vs. supply voltage V_{dd} for the DCT architecture.

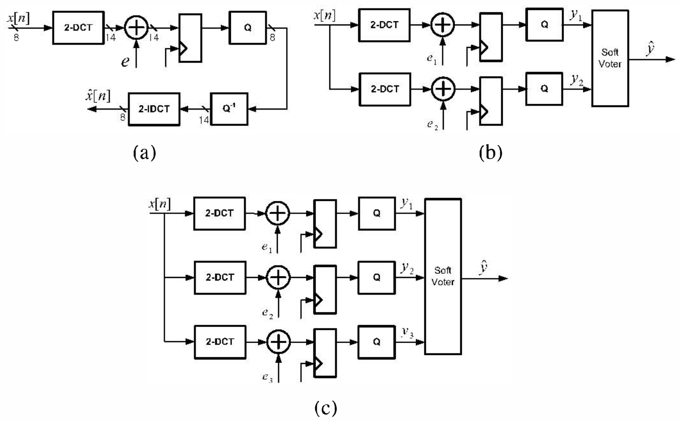
easily be made independent by employing one or more of the following techniques:

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

In this paper, we employ data diversity to obtain independent errors. First, we swap the operands in the DCT multipliers. Second, we choose between row-first or column-first processing of an 8×8 block of input pixels. Third we process a different sequence of the 1024 8×8 blocks in computing the 2-D DCT of the image. For $N = 3$, we found that the Kullback-Liebler (KL) distance between the joint PMF $P_e(e_1, e_2, e_3)$ and the product PMF $P_e(e_1)P_e(e_2)P_e(e_3)$ was close to zero (0.032) implying these two PMFs are practically equal. Furthermore, simulation in Section V show that the independence assumption is indeed valid.

IV. Soft NMR-based DCT Architecture

Figure 5 shows the various 2D-DCT architectures being considered. We replicate the DCT block and perform voting after the quantizer. Chen's algorithm [7] is used for deriving the DCT architecture and the quantizer (Q) employs the JPEG quantization table. 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 while consuming minimum power. The quantizers, 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. 5. These errors are independent by use of techniques discussed in Section III. The soft voter employs (2), which minimizes $P_{e,sys}$, and the majority voter employs (1).



(a)

(b)

(c)

Fig. 5. DCT-based image compression architecture: (a) conventional architecture, (b) soft DMR, and (c) soft TMR.

V. Simulation Results

In this section, we compare soft NMR with NMR in terms of robustness and energy-efficiency in the context of the 2D-DCT system described in section IV.

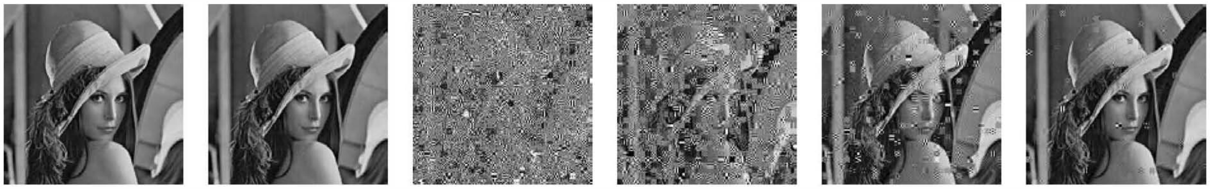
A. Robustness

Figure 6(a) shows that, for a wide range of PSNRs ($15dB - 30dB$), soft TMR can tolerate approximately 10× higher component probability of error p_{e_i} than TMR at the same PSNR. Figure 6(b) shows that soft TMR is 2× 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× 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.

Figure 7 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.

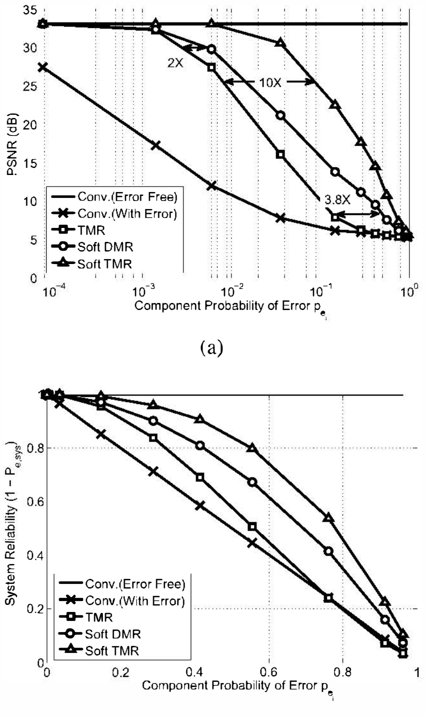
B. Power Savings

Power numbers were obtained via circuit simulations (HSPICE) of logic blocks, and via CACTI [9] for memory. Figure 8 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 20% power



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

Fig. 7. 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)

(b)

Fig. 6. System performance and reliability vs. component probability of error (p_{e_i}): (a) PSNR vs. p_{e_i}, and (b) system reliability ($1 - P_{e,sys}$) vs. p_{e_i}.

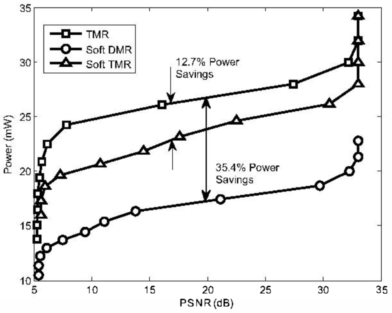


Fig. 8. Power consumption for TMR, soft DMR and soft TMR vs. PSNR.

savings and soft DMR achieves 30% to 40% power savings compared to TMR over a wide range of PSNRs.

TABLE I

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%	0.23%
Soft TMR	64868	14.9%	1.48%

Table I 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. 2(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.

VI. CONCLUSION AND FUTURE WORK

We demonstrated the value of incorporating statistics (data and error), and estimation and detection techniques to improve the robustness and energy-efficiency of NMR. This work opens up a number of interesting problems to explore including: 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] R. Hegde and N. Shanbhag, "A voltage overscaled low-power digital filter IC," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 2, pp. 388–391, Feb. 2004.

[2] M. Shafique, L. Bauer, and J. Henkel, "3-tier dynamically adaptive power-aware motion estimator for H.264/AVC video encoding," in *ISLPED '08: Proceeding of the thirteenth international symposium on Low power electronics and design*. New York, NY, USA: ACM, 2008, pp. 147–152.

[3] M. Michael and K. Hsu, "A low-power design of quantization for H. 264 video coding standard," in *2008 IEEE International SOC Conference*, 2008, pp. 201–204.

[4] A. Avizienis and J. P. J. Kelly, "Fault tolerance by design diversity: Concepts and experiments," *Computer*, vol. 17, no. 8, pp. 67–80, 1984.

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

[6] R. Blahut, *Digital transmission of information*. Addison-Wesley, 1990.

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

[8] J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital integrated circuits*, 2nd ed. Prentice Hall Upper Saddle River, NJ, 2002.

[9] S. Thoziyoor, N. Muralimanohar, and N. Jouppi, "CACTI 5.0," *HP Laboratories, Technical Report*, 2007.