J Sign Process Syst (2012) 68:415–431
DOI 10.1007/s11265-011-0631-9

Logic and Memory Design Based on Unequal Error Protection for Voltage-scalable, Robust and Adaptive DSP Systems

Georgios Karakonstantis · Debabrata Mohapatra · Kaushik Roy

Received: 23 January 2010 / Revised: 27 September 2011 / Accepted: 28 September 2011 / Published online: 27 October 2011

Abstract In this paper, we propose a system level design approach considering voltage over-scaling (VOS) that achieves error resiliency using unequal error protection of different computation elements, while incurring minor quality degradation. Depending on user specifications and severity of process variations/channel noise, the degree of VOS in each block of the system is adaptively tuned to ensure minimum system power while providing "just-the-right" amount of quality and robustness. This is achieved, by taking into consideration block level interactions and ensuring that under any change of operating conditions, only the "less- crucial" computations, that contribute less to block/system output quality, are affected. The proposed approach applies unequal error protection to various blocks of a system–logic and memory–and spans multiple layers of design hierarchy–algorithm, architecture and circuit. The design methodology when applied to a multimedia sub-system shows large power benefits (up to 69% improvement in power consumption) at reasonable image quality while tolerating errors introduced due to VOS, process variations, and channel noise.

G. Karakonstantis (✉)
Electrical Engineering, Telecommunications Circuits Lab,
Swiss Federal Institute of Technology (EPFL),
Lausanne 1015, Switzerland
e-mail: gkarakon@purdue.edu

D. Mohapatra · K. Roy
ECE School, Purdue University,
West Lafayette, IN 47907, USA

D. Mohapatra
e-mail: dmohapat@purdue.edu

K. Roy
e-mail: kaushik@purdue.edu

Keywords Low power · Variation aware design · Supply voltage scaling · Memory design

1 Introduction

Miniaturization of devices has resulted in the integration of numerous complex and power hungry processing units on a single portable device, the operation of which is mainly constrained by limited battery lifetime. Due to the quadratic dependence of power on voltage, supply voltage over-scaling (VOS) has been investigated as an effective method to reduce power [1]. However, VOS increases the delays in all computation paths and can result in incorrect or incomplete computation of certain paths. Besides power consumption, process variations also pose major design concern with technology scaling, resulting in delay errors [2]. Conventional wisdom dictates up-scaling the supply voltage or logic gate up-sizing to prevent delay errors and to achieve higher parametric yield. However, such techniques come at a cost of increased power and/or die area. Hence, meeting the contradictory requirements of high yield (in presence of unreliable components), low power and high quality is becoming exceedingly challenging in nanometer designs.

Methodologies that jointly address the issues of low power and error resiliency in digital signal processing (DSP) blocks have been proposed recently in [3, 4]. The methodology in [3] ensures tolerance to delay errors, utilizing the concept of unequal error protection. By protecting the "crucial" computations (more contributive to output quality) through algorithm/architecture co-design, low power at minor quality degradation is achieved. On the other hand, algorithmic noise tolerance (ANT) in [4] adds a reduced precision replica of the original block that acts as error control; it estimates potential errors of the main block

and tolerate any timing violations by selecting the *most correct* computed output. Such methodologies have proven to be efficient for the design of individual blocks, by being able to tolerate errors due to VOS/process variations, intelligently trading off quality-of-results.

However, when these blocks are integrated in a system, then the interactions between the individual blocks have to be considered. First, we observe that lower power or good quality for an individual block does not necessarily translate to lowest power or best quality for the overall system. Interestingly, while some blocks may operate at lower power and achieve "good" block level quality, subsequent blocks might still be consuming higher power without improving the system quality. This can be due to the fact that they are performing "redundant" computations based on VOS-affected incorrect outputs (albeit "less crucial one") of the preceding blocks. These computations can be thought of as "redundant" or "less-crucial" for subsequent blocks as they unnecessarily increase system power without improving the overall quality. Second, depending on dynamically changing system/user requirements and operating conditions (process variations, channel noise), the best quality may not always be required or achievable.

In this paper, we propose a system level design approach that achieves error resiliency under variations while considering VOS, using using unequal error protection of different computation elements, while incurring minor quality degradation. By taking into account block level interactions, the proposed approach provides the "right" amount of quality at the right amount of power consumption (by configuring the degree of VOS). Specifically, depending on power, user quality requirements and severity of process variations/channel noise, the degree of VOS for each block is adaptively tuned to ensure minimum system power while providing "just-the-right" amount of quality and robustness. This is achieved by guaranteeing that the delay errors due to VOS affect only the less-crucial (redundant) computations of the block (system) by efficient algorithmic/architectural modifications.

Apart from logic blocks, memory elements are also a ubiquitous part of today's systems and consume significant percentage of overall system power. From system perspective logic blocks interact with memory elements. It is possible, while logic blocks consume low power, the memory elements continue to consume high power, storing unnecessary computations that are potentially erroneous due to scaled voltage or process variations in logic blocks. Therefore, there is a need to expand the scope of unequal error protection to memory elements. It should be observed that standard memory cells (6 transistor SRAM cells or 6T cells) are vulnerable to failures under voltage scaling [5]. Hence, we apply a preferential storing policy in which the crucial computations are stored in robust memory banks composed of robust 8T-cells [6] and less crucial computations are stored in conventional 6T-cells. Such configuration ensures low probability of failure for the crucial banks even under VOS or process variations due to the robust nature of 8T cells, translating to lower quality loss. Furthermore, the less crucial banks can be dynamically turned off depending on quality/power user requirements, allowing system to adapt to various operating conditions and significantly reducing the overall system power when best quality is not required/achievable. The proposed approach departs from a recently published work [7] in which authors considered bit-significance rather than transform level-computations (such as in DCT/IDCT) as discussed in this paper.

The proposed approach is applied to system level joint design of logic (focusing on DCT/IDCT) and memory blocks of a sub-system which is ubiquitous in various applications (i.e. from lossy compression of audio (e.g. MP3) and images (e.g. JPEG) to spectral methods for the numerical solution of partial differential equations). In this paper we focus on a multimedia sub-system in order to show the implications and required modifications in each sub-block of such a system. Our contributions in this paper can be summarized as follows:

1) Considering VOS based DCT implementation we design a voltage scalable and robust IDCT by accounting for block level interactions in a multimedia sub-system.
2) We identify crucial/less-crucial computations for maintaining high output image quality in IDCT. The IDCT coefficients are modified such that only the less-crucial IDCT computations (based on potentially incorrect DCT outputs) are affected under VOS (unequal error protection for different computation elements). This avoids any unnecessary power increase, while ensuring minimal quality degradation of IDCT as well as multimedia sub-system.
3) Adaptively tune the degree of VOS for each logic block (DCT, IDCT) for achieving minimum system power in the presence of process variations/channel noise, meeting user specifications by using unequal error protection, unlike existing implementations [8–10].
4) A hybrid memory architecture is proposed that ensures correct read/write operations of crucial computations even under VOS.
5) The proposed architecture, inherently conceals channel noise at scaled Vdd's during transmission over noisy channel.

The rest of this paper is organized as follows. In section 2 the proposed system level approach focusing on logic is explained. In section 3 the application of our technique to a multimedia sub-system is described. Section 4 presents the unequal error protection scheme applied to memory

elements of a sub-system. Section 5 discusses the system level design for power awareness and error tolerance. Section 6 discusses the integration of the proposed multimedia sub-system within a wireless system. Finally, conclusions are drawn in section 7.

2 Design Strategy—Logic

To elucidate our design approach let us consider an example where a system consists of processing blocks A and B (Fig. 1). The outputs of block A are denoted as Z_{A1}, Z_{A2}, Z_{A3} while that of block B is $Y_B=f(Z_{A1},Z_{A2},Z_{A3})$. As mentioned earlier, VOS in block A would lead to an increase in path delays resulting in some computations not being completed within the specified target delay. Such timing violation due to VOS or process variations that results in erroneous output is shown in Fig. 2. Note that due to fluctuations in combinational output, the flip-flop latches an incorrect random value (outputting logic '1' instead of expected logic '0') due to setup time violation under scaled voltage (1.1 V). Recent approach in [3] addresses the quality degradation in conventional design under VOS by utilizing the concept of crucial/less-crucial computations. It identifies block specific computations as crucial and less-crucial and ensures that under VOS only the less-crucial computations get affected, leading to minimal quality degradation.

Let us assume that block A is designed using such a methodology in which outputs Z_{A1}, Z_{A2} are associated with crucial computations, whereas Z_{A3} is associated with less-crucial ones. The design ensures that under any delay errors due to VOS, only Z_{A3} is affected so that "reasonable" quality is retained. In this case the outputs of block A (A(x)) which are fed as inputs to block B can be written as: $A(x) = s + \eta_A$, where s is the error free output of A and η_A is the VOS error in A due to the incomplete computation of Z_{A3}. Depending on the computations in each block, such error (η_A) may degrade the output quality, while causing subsequent blocks to consume unnecessary power. Note that low power for block A does not indicate lowest achievable system power P_S, since the power consumption

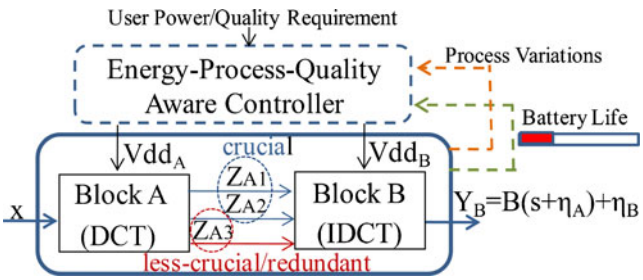

Figure 1 Power and process aware system design.

can be further reduced by applying VOS to block B. Interestingly, even if block B operates under no VOS, the overall quality of the system Q_{SYS} may not improve (potentially incorrect Z_{A3} due to VOS in A). This observation led us to think of a design strategy that would allow us to perform VOS in block B without any further degradation of system quality. It is made possible by ensuring (by proper design) that only those computations of block B that utilize the incorrect outputs of A (such as less-crucial Z_{A3}) are affected under VOS in B. These computations of block B (involving Z_{A3}) can also be thought of as less-crucial, and hence, redundant (under VOS of A) for the whole system (not contributing to Q_{SYS}). Note that if we do not follow such an approach, the system would be operating under non-optimal conditions since block B unnecessarily tries to improve Q_{SYS}, which under the assumed operating conditions (VOS in A), cannot be achieved.

The proposed design approach is shown in Fig. 1. A power, quality, process aware controller takes feedback from individual blocks and adaptively assigns the parameters Vdd_A and Vdd_B to A, B respectively, such that minimum system power P_S and reasonable Q_{SYS} is ensured (that meets user specified quality requirements (Q_{DES})). The design problem can be expressed as:

$$\begin{aligned} &\text{minimize } P_S(Vdd_A, Vdd_B) = P_A(Vdd_A) + P_B(Vdd_B) \\ &\text{subject to } Q_{SYS} \geq Q_{DES} \end{aligned} \tag{1}$$

where Q_{SYS} can be expressed in terms of system Signal-to-noise ratio (SNR):

$$Q_{SYS} = 10\log_{10}\frac{{\sigma_{SIG}}^2}{{\sigma_{\eta SYS}}^2} \tag{2}$$

In Eq. 2, ${\sigma_{SIG}}^2$ and ${\sigma_{\eta SYS}}^2$ are the power of error free system output and system VOS noise power, respectively. Using Eq. 2, the constraint in Eq. 1 can be written as:

$${\sigma_{\eta SYS}}^2\left({\sigma_{\eta A}}^2(Vdd_A),{\sigma_{\eta B}}^2(Vdd_B)\right) \leq \frac{{\sigma_{SIG}}^2}{10^{Q_{DES}/10}} \tag{3}$$

where ${\sigma_{\eta SYS}}^2$ can be expressed in terms of VOS error power of blocks A and B. The noise power due to VOS error of A (η_A) and VOS error of B (η_B) are denoted by ${\sigma_{\eta A}}^2$ and ${\sigma_{\eta B}}^2$ respectively. Since output of block B depends on output of block A (Fig. 1), they are dependent and hence the inter-dependencies between the errors of the blocks have to be considered in evaluating ${\sigma_{\eta SYS}}^2$. Hence, a joint design of blocks A, B is necessary for meeting the constraint given in Eq. 1.

In Eq. 3, ${\sigma_{\eta SYS}}^2$ is shown to be a function of ${\sigma_{\eta A}}^2$ and ${\sigma_{\eta B}}^2$ as system noise power depends on individual block noise power. Since noise power due to VOS of a block is directly related to its Vdd, ${\sigma_{\eta A}}^2$ and ${\sigma_{\eta B}}^2$ (hence ${\sigma_{\eta SYS}}^2$) can

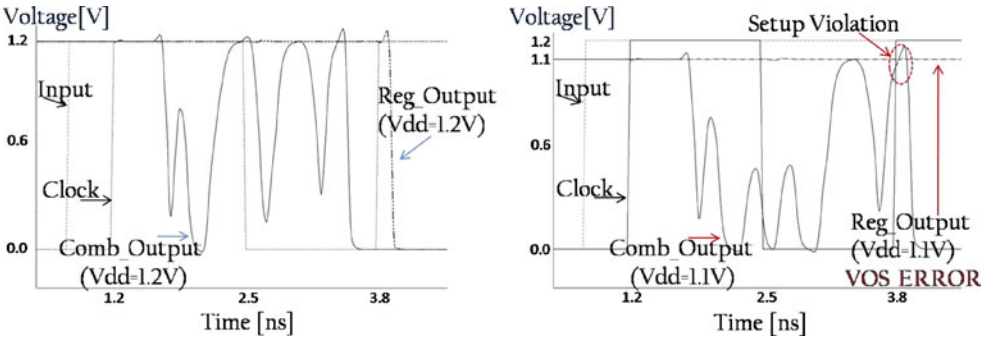

Figure 2 Typical timing violation due to VOS (Conventional DCT architecture).

be modulated by tuning Vdd_A and Vdd_B respectively. By configuring Vdd_A and Vdd_B, the system quality degradation due to VOS (reflected in $\sigma_{\eta SYS}^2$) can be controlled such that Eq. 3 is satisfied providing the "right" amount quality at minimal power consumption. The above design strategy is applied to a multimedia sub-system, analyzed next.

3 Application to A Multimedia Sub-system

We apply the proposed approach to the design of voltage over-scalable sub-system which is the core of various standards (here we focus on its utilization for image compression in JPEG) widely used in popular portable systems (i.e. digital camera) [8–10, 12]. An example of such sub-system found in a portable camera system is shown in Fig. 3. Such system consists mainly of logic and memory blocks, the operation of which can be divided into different stages. Initially, a scene is captured through an image sensor, a procedure called color interpolation [11] pre-process the captured image and the resultant image enters a discrete cosine transform (DCT) block [3] that transforms each 8x8 block of the image to the frequency domain. The output is then quantized and the resultant compressed image is stored in memory. Then an IDCT block is responsible for de-compressing and preparing the image for display. In this section we focus on the computationally intensive logic components (DCT/IDCT)

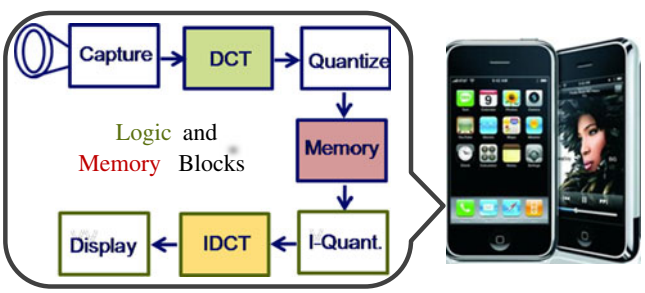

Figure 3 Multimedia sub-system within a complex portable embedded system. Logic (DCT/IDCT) and Memory highlighted.

of such sub-system. Specifically, based on interactions with a DCT architecture, we design a voltage scalable and robust IDCT that performs system level adaptive quality modulation for minimizing system power, utilizing the concept of unequal error protection. Note that although we focus on a type-II DCT/IDCT algorithm; similar modifications can be applied to other algorithms. Next paragraphs discuss the algorithmic and architectural changes required for realizing such a design. At this point we would like to note that any modifications should not lead to major changes and increase the overhead in other blocks of the target application. Rather attention has to be paid in order to design other blocks such that they can benefit from the unequal error protection. Section 4 discusses the design of scalable/robust memory blocks exploiting the proposed unequal error protection approach in logic, while Section 5 describes how other blocks of the system (i.e. quantization) can benefit from the proposed approach.

3.1 Conventional IDCT

Conventionally, IDCT transforms an NxN block (output of DCT) from frequency domain to spatial domain and in matrix notation it can be expressed as: $X = C^T Z C$, where C is the coefficient matrix and Z is the input NxN block that contains the DCT outputs. Using row-column decomposition, IDCT can be decomposed into two 1D-IDCT units, which can be expressed as:

$$X = C^T \left(C^T Z^T\right)^T \tag{4}$$

In such an implementation, the 1D-IDCT is applied to each row of input data and the result is transposed. A second 1D-IDCT is applied to the rows of the transposed data to obtain IDCT (Eq. 4). The 1D-IDCT output of a row in 8×8 block is expressed as:

$$w_i = \sum_{k=0}^{7} \frac{c(k)}{2} z(k) \cos\left(\frac{(2i+1)k\pi}{16}\right),\ i = 0, 1, 2, ..., 7$$
$$c(0) = 1/\sqrt{2}, c(k) = 1 \ if \ k \neq 0 \tag{5}$$

Setting $c_k = cos(k\pi/16)$, $a = c_1$, $b = c_2$, $c = c_3$, $d = c_4$, $e = c_5$, $f = c_6$, $g = c_7$ and exploiting symmetry of C, the 1D-IDCT can be written as [12]:

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} d & b & d & f \\ d & f & -d & -b \\ d & -f & -d & b \\ d & -b & d & -f \end{bmatrix} \begin{bmatrix} z_0 \\ z_2 \\ z_4 \\ z_6 \end{bmatrix} + \begin{bmatrix} a & c & e & g \\ c & -g & -a & -e \\ e & -a & g & c \\ g & -e & c & -a \end{bmatrix} \begin{bmatrix} z_1 \\ z_3 \\ z_5 \\ z_7 \end{bmatrix} \quad (6)$$

$$\begin{bmatrix} w_7 \\ w_6 \\ w_5 \\ w_4 \end{bmatrix} = \begin{bmatrix} d & b & d & f \\ d & f & -d & -b \\ d & -f & -d & b \\ d & -b & d & -f \end{bmatrix} \begin{bmatrix} z_0 \\ z_2 \\ z_4 \\ z_6 \end{bmatrix} - \begin{bmatrix} a & c & e & g \\ c & -g & -a & -e \\ e & -a & g & c \\ g & -e & c & -a \end{bmatrix} \begin{bmatrix} z_1 \\ z_3 \\ z_5 \\ z_7 \end{bmatrix} \quad (7)$$

3.2 Voltage Scalable and Error Resilient DCT

We can observe from Eqs. 6 and 7 that 1D-IDCT takes as input each output row of DCT (Eq. 1: $C^T Z^T$) denoted by z_0-z_7. Input z_0 is the DC coefficient that contains the low frequencies of an 8×8 image block, whereas z_1-z_7 are the AC coefficients that contain high frequencies. Taking advantage of the fact that human eye is more sensitive to lower spatial frequencies, a voltage scalable and process tolerant DCT architecture was proposed in [3]. After the identification of crucial computations (z_0), the architecture in [3] ensured that only less-crucial computations (that contribute less to output quality) are affected by delay errors due to VOS. This is achieved by algorithmic and architectural modifications considering unequal error protection of different computation elements–the crucial computations are designed to be faster at the cost of less crucial ones. This allows the architecture to provide graceful degradation of output quality under VOS. Specifically, under scaled Vdd_1, computations z_0-z_4 are protected by VOS-induced error by making them faster (possibly at the cost of making z_5-z_7 slower, so as to ensure lower area and power dissipation). Under further VOS (Vdd_2), the architecture ensures that the most crucial computation z_0 is at least computed correctly.

Since the outputs z_5-z_7 or z_1-z_7 are potentially incorrect under VOS in DCT, they should not be taken into consideration in the computation of subsequent blocks (IDCT) of the system. These outputs and consequently the computations that use them can be considered as redundant, since their correctness cannot be guaranteed. In addition, they might degrade the system output quality, while causing unnecessary increase in system power dissipation. To avoid this, Vdd in IDCT block can be scaled down such that only redundant computations involving potentially incorrect outputs of DCT are affected by such scaling. This unequal error protection strategy helps in achieving minimal system quality degradation while achieving large improvement in power consumption.

In the next subsection we will present the design of VOS-IDCT (block B (Fig. 1)) that ensures correct operation except for less-crucial/redundant outputs of block A (DCT). Here, we need to highlight that even though IDCT is the reverse operation of DCT, the manner of their computation differs due to the transposition of matrix C (Eq. 4). Note also that DCT and IDCT process inputs in different domain, one in space and the other in frequency, respectively. Hence, the design of desired scalable IDCT architecture does not follow from the transformations in [3]. Therefore, algorithmic transformations, specific to IDCT, are needed in order to ensure minimal quality degradation under VOS.

3.3 Algorithmic Transformations of IDCT

In order to facilitate voltage scaling and quality modulation at the architecture layer, algorithmic transformations are required that will ensure minimum quality degradation under any delay error. Initially, the computations in IDCT need to be classified in crucial and less-crucial based on the interaction with the voltage scalable DCT discussed earlier. We denote computations in IDCT involving correct DCT outputs as crucial and those involving potentially incorrect ones (due to VOS) as less-crucial. To identify the nature of IDCT computations (crucial/less-crucial), we rearrange Eqs. 6 and 7 as:

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} = \overbrace{\underbrace{\begin{bmatrix} d \\ d \\ d \\ d \end{bmatrix}}_{C_1} z_0 + \underbrace{\begin{bmatrix} d \\ -d \\ -d \\ d \end{bmatrix}}_{C_2} z_4 + \underbrace{\begin{bmatrix} a & b & c \\ c & f & -g \\ e & -f & -a \\ g & -b & -e \end{bmatrix}}_{C_3} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}}^{\Gamma_2}}^{\Gamma_1} + \overbrace{\underbrace{\begin{bmatrix} e & f & g \\ -a & -b & -e \\ g & b & c \\ c & -f & -a \end{bmatrix}}_{C_4} \begin{bmatrix} z_5 \\ z_6 \\ z_7 \end{bmatrix}}^{\Gamma_3} \quad (8)$$

$$\begin{bmatrix} w_7 \\ w_6 \\ w_5 \\ w_4 \end{bmatrix} = \begin{bmatrix} d \\ d \\ d \\ d \end{bmatrix} z_0 + \begin{bmatrix} d \\ -d \\ -d \end{bmatrix} z_4 + \begin{bmatrix} -a & b & c \\ -c & f & g \\ -e & -f & a \\ -g & -b & e \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} - \begin{bmatrix} e & -f & g \\ -a & b & -e \\ g & -b & c \\ c & f & -a \end{bmatrix} \begin{bmatrix} z_5 \\ z_6 \\ z_7 \end{bmatrix} \quad (9)$$

Such rearrangement allows representation of each w_i in terms of crucial and less-crucial terms. The crucial term is

denoted by $\Gamma_1=C_1z_0$ whereas less-crucial terms are denoted by $\Gamma_2=C_2z_4+C_3[z_1\ z_2\ z_3]^T$ and $\Gamma_3=C_4[z_5\ z_6\ z_7]^T$. Note that Γ_2 and Γ_3 are identified as less-crucial since their computation involves potentially incorrect DCT outputs under VOS (z_1-z_7).

The differentiation of IDCT computations allows us to apply VOS in IDCT with graceful degradation in system output quality. This is achieved by applying unequal error protection, ensuring that any VOS induced error will affect only the less-crucial computations from contributing to the output, while guaranteeing correctness of crucial computations. To elucidate further, under VOS, less-crucial computations of w_i are successively omitted, preventing them from affecting the output. As shown in Eqs. 8 and 9, each output w_i is sum of one crucial (Γ_1) and two less-crucial (Γ_2, Γ_3) terms. At nominal Vdd (Vdd_{nom}), all three terms contribute to output, which equals $\Gamma_1+\Gamma_2+\Gamma_3$. At scaled Vdd_1, we exclude one of the less-crucial terms (Γ_3) from the output (which now equals $\Gamma_1+\Gamma_2$) since computation of Γ_3 involves potentially incorrect inputs z_5-z_7. Furthermore, at scaled Vdd_2, both the less-crucial terms (Γ_2, Γ_3) need to be omitted since their computation involves potentially incorrect inputs z_1-z_7. In order to facilitate gradual exclusion of less-crucial terms from the output under VOS, we require Γ_1, Γ_2 and Γ_3 to be computed independently, without any sharing of computation between them. However, in the absence of computation sharing among Γ_1, Γ_2 and Γ_3, prohibitively large area overhead may be incurred. To circumvent this issue, we synthesize IDCT coefficients such that area overhead is minimized.

In addition, efficient synthesis of coefficients plays an extremely important role in providing unequal error protection. By intelligently modifying coefficients, we constrain the crucial term (Γ_1) to be computed within k_1 adder levels, while $\Gamma_1+\Gamma_2$ and $\Gamma_1+\Gamma_2+\Gamma_3$ are constrained to be computed within k_2 and k_3 adder levels respectively ($k_1<k_2<k_3$). Under VOS, this unequal error protection technique can provide reasonable output quality since the crucial term (Γ_1 computed within k_1 adder levels) remains unaffected. In order to avoid any delay penalty associated with such unequal error protection scheme, the synthesized coefficients are required to satisfy the condition $k_3=L$, where L is the number of adder levels in a conventional multiplierless IDCT. However, this is achieved at the expense of minor quality degradation (due to modified coefficients). The synthesis of coefficients for minimizing area overhead and quality degradation without incurring any delay penalty can be viewed as an optimization problem involving vector scaling operation which is discussed below.

Vector scaling operations [3, 8] can be represented with only few shifts and adds. Each 'ONE' in the coefficient vector represents a value that needs to be shifted and added. The total number of 'ONE's' in a coefficient vector of size S can be optimally added within L adder levels given by:

$$L = \log_2 N, N = \sum_{i=0}^{S-1}\sum_{j=0}^{B-1} b_{ij} \tag{10}$$

where B is number of bits used to represent each coefficient. For instance, multiplication of 8-bit coefficient d with scalar input z_0 can be represented as $2^5z_0+2^3z_0+2^2z_0+z_0$. Since d consists of N=4 'ONE's' (Table 1), dz_0 and hence Γ_1 can be optimally computed within L=2 adder levels. Similarly, using the original 8-bit coefficients with N=29 (Table 1), the optimal number of adder levels required for computation of each path w_i (Eqs. 8 and 9) is $L=5=k_3$. However, with original coefficients, Γ_1, Γ_2 and Γ_3 cannot be computed independently (without sharing). Hence, modification of original coefficients is required in our VOS based IDCT design.

First, we note that coefficient d should not be modified since it is multiplied with the crucial DCT output z_0 (Eqs. 8 and 9) and any change would affect the output image quality significantly. Thus computations involving C_1 and C_2 are separated from the rest as shown in Eqs. 8 and 9. Second, we only modify the coefficients contained in C_3 and C_4. However, since each coefficient appears in both C_3 and C_4, any change of coefficient in one of the matrices would affect the other and may result in an increase in path delay of w_i (more than 5 adder levels) and area overhead. This leads us to the first constraint in which, the modified coefficients must guarantee that computational path delay of any w_i does not exceed $k_3=L=5$ adder levels. This ensures that the resulting architecture operates at the same frequency as conventional multiplierless architecture (no delay penalty). Second, the total number of ones in each row of modified C_3

Table 1 Original and modified 8-bit IDCT coefficients.

Coef.	Original		Modified	
	Value	Binary	Value	Binary
a	0.49	0011 1111	0.5	0100 0000
b	0.46	0011 1011	0.47	0011 1100
c	0.42	0011 0101	0.41	0011 0100
d	0.35	0010 1101	0.35	0010 1101
e	0.28	0010 0100	0.28	0010 0100
f	0.19	0001 1000	0.2	0001 1001
g	0.10	0000 1100	0.10	0000 1100

and C_4 must not exceed that of original C_3 and C_4, which can be expressed as

$$\forall \text{ row of } C_3,C_4: \left(\sum_{i=0}^{S_{C3}-1}\sum_{j=0}^{B-1} b_{ij}\right) + \left(\sum_{i=0}^{S_{C4}-1}\sum_{j=0}^{B-1} b_{ij}\right) \leq 21 \qquad (11)$$

In addition, we desire the crucial (Γ_1) and less-crucial (Γ_2, Γ_3) terms to be computed separately so that under VOS, potentially incorrect less-crucial terms are prevented from contributing to the output. For instance, at scaled Vdd_1, due to increase of path delay, the computation of $k_3=L=5$ adder levels in IDCT may not be completed. However, reasonable output quality can be obtained if crucial Γ_1 and one of the less-crucial terms Γ_2 were evaluated correctly by the $k_2=4$ adder level. This is due to the fact that even if we omit Γ_3 (output at Vdd_{nom} equals $\Gamma_1+\Gamma_2+\Gamma_3$), minor quality degradation is incurred due to its less-crucial nature. Of course, under Vdd_{nom}, output quality can be improved by adding Γ_3 to $\Gamma_1+\Gamma_2$ at the 5th adder level. Hence, the modified coefficients need to be synthesized such that, in the resulting architecture, Γ_1, Γ_2 are computed within $k_1=3$ adder levels, $\Gamma_1+\Gamma_2$ and Γ_3 within $k_2=4$ adder levels and $\Gamma_1+\Gamma_2+\Gamma_3$ within $k_3=L=5$ adder levels. Note that since the number of adder levels depends on the number of ones in the coefficient vector, the constraint that requires Γ_3 to be computed within $k_2=4$ adder levels can be expressed as:

$$\forall \text{ row of } C_4: \log_2\left(\sum_{i=0}^{S_{C4}-1}\sum_{j=0}^{B-1} b_{ij}\right) \leq 4 \Rightarrow \sum_{i=0}^{S_{C4}-1}\sum_{j=0}^{B-1} b_{ij} \leq 16 \qquad (12)$$

Equivalently, the total number of 'ONE's' in each row of coefficient matrix C_4 should be less than 16. Similarly, the constraint that requires Γ_2 to be computed within $k_1=3$ adder levels can be expressed as (the total number of 'ONE's' in each row of coefficient matrix C_3 should be less than 8):

$$\forall \text{ row of } C_3: \log_2\left(\sum_{i=0}^{S_{C3}-1}\sum_{j=0}^{B-1} b_{ij}\right) \leq 3 => \sum_{i=0}^{S_{C3}-1}\sum_{j=0}^{B-1} b_{ij} \leq 8 \qquad (13)$$

Finally, in order to ensure that there is no sharing of any computation between less-crucial terms $C_3[z_1\ z_2\ z_3]^T$ and Γ_3, the numbers of 'ONE's' in each row of at least one of the matrices C_3 or C_4 must be even.

Based on the above constraints, the absolute difference between the peak-signal-to-noise ratios (PSNR) of the image obtained using original coefficients and image PSNR using modified coefficients is minimized (objective function). The total number of 'ONE's' in the modified set of coefficients (Table 1) is reduced from $N=29$ to $N=23$ which overcomes the area overhead (due to no sharing) at minor quality degradation (~0.5 dB) yielding a voltage scalable IDCT architecture. Furthermore, the modified coefficients reduce the variance of VOS induced errors in IDCT, $\sigma_{\eta B}^2$, taking into account system level interactions with a VOS based DCT, while meeting the system constraint stated in Eq. 3.

3.4 Proposed Architecture

The proposed voltage-scalable, architecture, based on the above algorithmic modifications is shown in Fig. 4. As shown in Fig. 4, crucial computation Γ_1 (dz_0) is unaffected by VOS since it is computed within $L=2$ adder levels for each w_i. Interestingly, even if resource sharing across groups is absent, minimal hardware overhead (compared to conventional multiplierless architecture) is incurred. This is due to the reduction in the number of 'ONE's' in the modified coefficients (N decreases from 29 to 23). However, modification of coefficients comes at the expense of minor quality degradation (0.2 dB on average on set of 25 images [13]). In order to further optimize overall hardware, we maximized resource sharing within a group (Γ_1, Γ_2, Γ_3) of computations. Next, we analyze the architecture under i) VOS and ii) process variations.

i) **VOS:** From Fig. 4, the critical path for this design is {five adders + one 3-input multiplexer}. The multiplexer allows us to select the correct output under VOS. At Vdd_{nom}, both VOS signals (V_1, V_2) are set to zero and the multiplexer chooses the output of $k_3=L=5$th adder level ($O_3=\Gamma_1+\Gamma_2+\Gamma_3$) for each w_i. Under VOS, since computation of 5th adder level is no longer possible, the output of $k_2=4$th adder level ($O_2=\Gamma_1+\Gamma_2$) is selected by setting the multiplexer control signal V_1='1'. In such a scenario, Vdd can be scaled to an extent (Vdd_1), at which the computation of {four adders + one multiplexer} is valid. Further VOS can be obtained by setting the value of the second voltage control signal V_2='1'. At this voltage, the multiplexer chooses the 2nd adder level output ($O_1=\Gamma_1$) (Fig. 4). The path delay of {two adders + one multiplexer} determines minimum Vdd (Vdd_2) for correct operation.

ii) **Variations tolerance:** Power is not the only constraint in scaled technologies. Process variations may cause delay errors, affecting computations in some chips, which are at the slow process corner (under Vdd_{nom}). Under such a scenario, the output of the 5th adder (Fig. 4) in our architecture would be incomplete and hence invalid. However, with the help of leakage/delay sensor [14], one can easily detect the process corner and correspondingly set the first Vdd control signal V_1='1'. This will result in minor output quality degradation compared to quality at Vdd_{nom}, but will provide a valid output. Furthermore, at scaled Vdd (say Vdd_1) due to process variations, the computation of the 4th adder might be invalid. In that case, we still

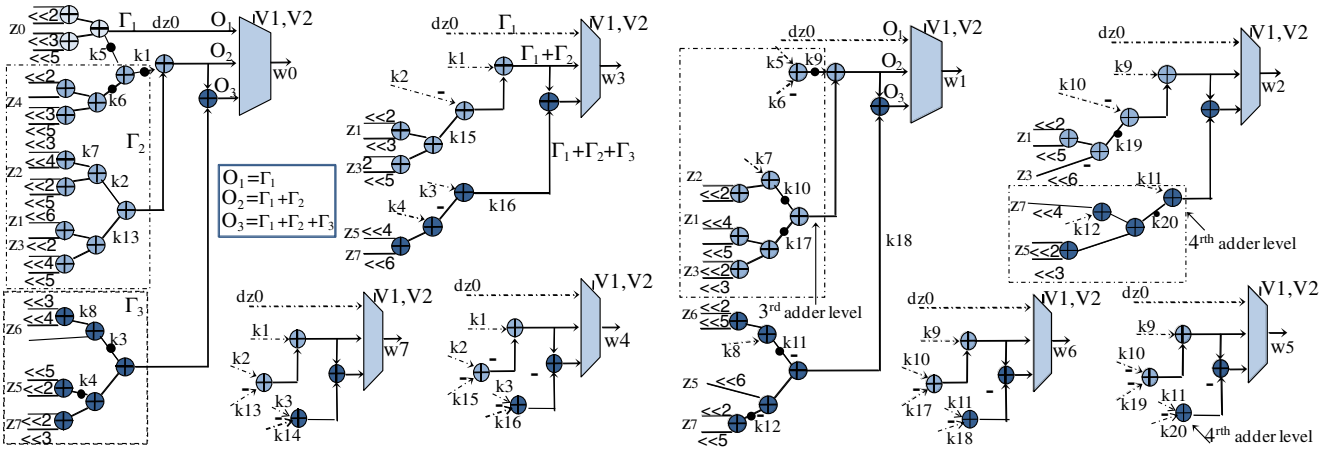

Figure 4 Proposed voltage scalable and process tolerant 1D-IDCT architecture.

have an option of setting V_2='1' and obtain a valid output at further expense of output quality. Hence, our architecture is capable of achieving VOS while providing tolerance to possible process variations induced errors.

3.5 Results

We implemented the proposed IDCT and a conventional multiplierless IDCT (using original 8-bit coefficients) architecture in Verilog. We obtained synthesized Verilog netlists from Design Compiler [15], using ARM standard cell libraries [16] which were converted to HSPICE netlists using Calibre [17]. The average power consumption at 300 MHz in IBM 90 nm technology was obtained using NanoSim. Moreover, using a set of 24 images [13], average PSNR for the proposed and conventional IDCT were obtained. The comparison results are shown in Table 2. We would like to mention that the variance of the PSNR at different Vdd levels, 1.2 V, 1.1 V, 0.71 V were 2.84, 4.77 and 5,71, respectively. Note that the proposed architecture allows large power savings with graceful quality degradation while conventional design fails under VOS. In addition note that the multiplexers impose a 4% delay overhead in the proposed design. Table 3 presents the PSNR values of some representative images at various voltages. Note that quality degradation at scaled voltages depends on the image and its information (i.e. number of high frequencies). However, in all cases our architecture provides acceptable output quality (PSNR above 20 dB) since it protects intelligently the significant parts that carry-out most of the information.

As we mentioned apart from unequal error protection, other techniques for the design of voltage scalable and error resilient logic blocks exist such as ANT [4]. In ANT, an estimator block (which is actually a reduced precision replica) is utilized to identify any error in computations in presence of delay variations due to Vdd scaling; additionally, a decision block is used to select the correct output at scaled Vdds. It is important to note that such techniques have also to be modified in order to account for the system level interactions between the various blocks. In addition, we need to note that unequal error protection provide more power savings compared to ANT since it doesn't require the addition of any hardware to protect the significant computations [3]. This leads to minor area and power overhead even at nominal voltage of operation. All in all, algorithm along with architectural innovations, as suggested in the proposed approach, can yield smaller efficient designs/systems that allow extreme power savings at both the

Table 2 Comparison of architectures at various Vdds.

IDCT	Proposed			Conventional
Vdd(V)	1.2	1.1	0.71	1.2
Power (mW)	15.8	12.3	4.7	15.9
PSNR (dB)	34.9	30.04	24.8	35.1
Area (um^2)	70192			71146

Table 3 PSNR (dB) values of 5 most representative images [13] at various Vdds.

IDCT	Proposed			Conventional
Vdd(V)	1.2 V	1.1 V	0.71 V	1.2 V
kodim2	36.5	32.6	28.6	36.8
kodim9	33.5	31.21	26.1	35.59
kodim11	36	30.1	25.2	36.2
kodim15	33.4	30.2	26.4	33.6
kodim18	38.38	28.6	24.2	34.5
kodim24	34.4	28.01	23.1	34.7

nominal (due to reduced effective switched capacitance) and scaled Vdds.

4 Design Strategy—Memory

In addition to logic blocks, today's systems consist of memory units into which computations are written to and read back from as depicted in Fig. 3. From that perspective, apart from the interactions between logic blocks, there are interactions with memory blocks which must not be neglected. Additional power savings can be obtained considering such interactions jointly with user quality requirements, by scaling Vdd or even turning off sections of the memory when they are not required, as explained in the following paragraphs.

Let us consider the example shown in Fig. 1 and assume that there exists a memory block between blocks A and B. The outputs of block A are stored in the memory block, while block B reads the stored values. Let us now assume that we apply VOS in block A. Then, as mentioned earlier, the less crucial computation Z_{A3} may not be computed correctly (potentially incorrect). In this case, according to the proposed strategy, we could apply VOS in block B, leading to further reduction of system power without degrading the output system quality Q_{SYS} (already reduced due to errors in block A). The question that arises now is: "what is the lowest system power that can be achieved under the given conditions (VOS, quality loss)?" Note that, if we take into consideration the memory block, then it is possible to further reduce power by avoiding storage of the potentially incorrect (due to VOS) less crucial computation Z_{A3} since it is not going to be used in subsequent blocks. Furthermore, we can design memory such that the sections that store less crucial computations are completely turned off, thereby yielding additional power savings. The proposed memory will be able to adapt to user requirements; provide best possible quality consuming high power or trade off quality for reduced power consumption by scaling the number of computations accessed. Before we delve into the details of the proposed memory, it is worth mentioning that under VOS and parameter variations, standard 6T memory cells suffer largely from read/write, and access failures [6]. Hence, in case of memory, quality is associated with cell and array failure probability; smaller the memory failure probability, better the output quality.

4.1 Memory Failures

Random variations and reduction of cell supply voltage result in memory failures [5, 7] in standard 6T cells. Such failures are generally categorized as read, write and access failures. By definition higher access delay implies higher probability of access failures and vice-versa. For medium frequency of operation the access failures are important only at low voltages for 65 nm technology. Therefore, for image processing applications that require low operating frequency (~10–30 MHz) the delay constraint can be easily satisfied even at low Vdd (~0.6 V) [7]. Thus, for these applications, access failures (due to increase in access delay under VOS) can be neglected for all practical purposes.

The flipping of the data during reading of the cell [5] results in read failures, while lower strength of the cell access transistor results in higher write failure at lower Vdd. Since read and write margin have conflicting design requirements [5, 7] we can ignore the probability of their simultaneous negative effect on the cell. Hence, the overall cell failure probability P_F can be approximated as:

$$P_F \approx P_{RF} + P_{WF} \tag{14}$$

where P_{RF} and P_{WF} are the read and write failure probability, respectively.

In order to overcome the lack of read static noise margin (SNM) which is the main obstacle to applying VOS in 6T SRAM array, 8T SRAM cells [6] can be utilized. Specifically, due to decoupling of the memory nodes during read/write, the 8T-bit cell provides resiliency to VOS induced failures. This is evident in Fig. 5 which presents the P_{RF} obtained from extensive Monte Carlo simulations.

Figure 6 demonstrates the write failure probability for both cells. It can be observed that P_{WF} is similar for both cells since the main improvement of 8T cell over 6T cell is during the read operation. Figures 5 and 6 also show the failure probabilities for the upsized 6T cell under iso-area condition. We can observe that 8T cell provides better immunity against VOS induced failures. Due to the improved robustness of the 8T cell at scaled supply voltages, we can lower Vdd of memories more aggressively using the 8T bit-cell. However, the 8T bit-cell incurs large area penalty (larger than 30% compared to 6T bit-cell) limiting its application. In order to address the large area overhead incurred by 8T bit-cells, we propose a scalable

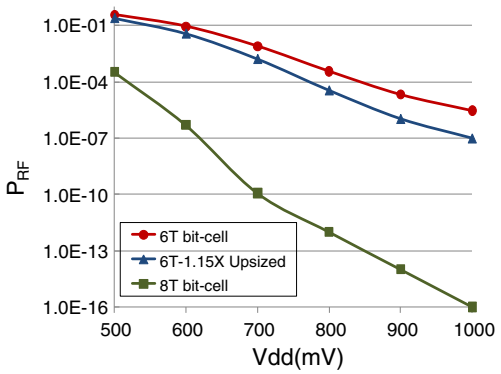

Figure 5 Read failure probability.

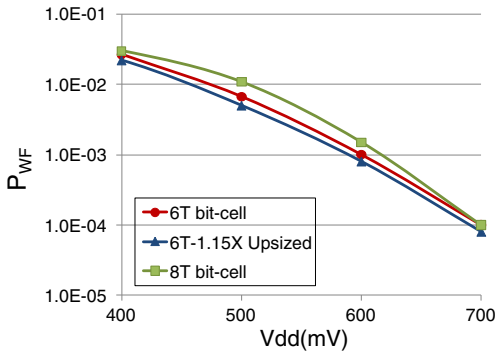

Figure 6 Write failure probability.

memory architecture that not only takes advantage of the robustness associated with 8T bit-cells but also ensures low area overhead by adopting a significance driven storage (or unequal error protection) policy: storing computations according to their significance. We discuss the architecture and circuit level details for implementing such a scalable memory in the subsequent section.

4.2 Architecture Level

To implement our unequal error protection policy, we partition the memory array into K smaller banks according to the type of computations they store; crucial or less-crucial. Note that the value of K is determined by the number of crucial and less crucial parts within a target application. Interestingly, such technique can improve the power efficiency of the SRAM because the word line capacitance being switched and the number of bit cells activated are reduced [1].

For instance, in case of DCT we have distinguished computations into 3 parts. The least crucial group (z_5-z_7), the less crucial group (z_2-z_4) and crucial part (z_1) for each 8×8 image block. Therefore, we divide memory into 3 banks as shown in Fig. 7. Each of the banks is equivalent to an SRAM array consisting of row and column decoders and sense amplifier. An extra address word called the block address is also used in order to select one of the K banks to be read from or written to. The main advantage of such an

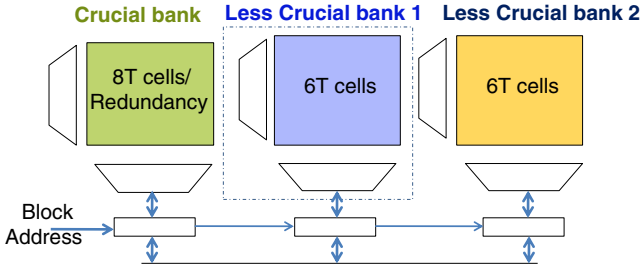

Figure 7 Memory organization based on unequal error protection approach.

organization is that only the addressed block is activated. Depending on user power/quality requirements, less crucial blocks are put in sleep mode by disabling the sense amplifiers and row and column decoders. This allows us to achieve significant power reduction since less crucial banks are not required and Vdd of unnecessary banks are set to zero. Note that muxes are used to route the inputs and clocks to the appropriate memory blocks, and to gate the inputs and clocks to the memory banks that are not accessed.

4.3 Circuit Level

Apart from significance driven banking organization (shutting down less crucial banks), utilization of 8T SRAM cells provides additional protection to data from crucial computations under VOS/parametric variations. In order to reduce the area penalty we consider the usage of 8T cells for only the crucial bank (mentioned above). This configuration results in less than 10% area overhead in the case of DCT/IDCT system as discussed in detail in subsequent sections.

The proposed memory achieves power savings not only due to preferential turning off of less crucial banks but also due to VOS. This can be achieved since the 8T cells provide enough immunity to crucial computations even at 600 mV, achieving large power savings at good output quality. Note that we assume that the sense-amplifier driving the output buses always operates at the nominal voltage. This provides an inherent level conversion for the data bit being read from the low-voltage memory domain under VOS.

It should be noted that the hold stabilities of 8T bit-cells are significantly degraded during writing of other bit-cells that shares its write line (WWL), namely half selection problem [6, 7]. This issue can be avoided by assigning an independent word line per each computation within the crucial bank. In general depending on the size of basic computation stored in each bank, the bit-cell number per one row can also vary. For instance, in case of multimedia system, memory bank 2 can store four computations while bank 3 could store three computations per row.

4.4 Array Failure and Redundancy

Note that other techniques such as redundant rows and columns [5] could also be applied in order to increase the stability of the banks under voltage scaling. Specifically, redundant columns could be also used within the crucial banks instead of 8T-bit cells. The failure probability of a column (P_{col}) or row (P_{row}) is defined as the probability that any of the cells in that column or row fails. Assuming column redundancy, the failure probability of a memory

array (P_{bank}) designed with N_{col} number of columns and N_{rc} number of redundant columns is defined as the probability that more than N_{rc} columns fail [5]. Hence, P_{col} and P_{bank} can be defined as:

$$P_{col} = 1 - (1 - P_F)^{N_{row}} \tag{15a}$$

$$P_{bank} = \sum_{i=N_{rc}+1}^{N_{col}+N_{rc}} \binom{N_{col}+N_{rc}}{i} \times P_{col}^{i}(1-P_{col})^{N_{col}+N_{rc}-i} \tag{15b}$$

Figure 8 presents probability of success (1-P_{bank}) comparison between crucial bank using redundant columns (with 6T cells) and 8T-bit cells. It is evident that 8T bit-cells provide better immunity even under aggressive VOS compared to the redundancy based scheme, thereby justifying our selection of 8T bit-cells. Apart from the failure probability, we also analyze the area penalty associated with each of the two design options mentioned above. Of course redundancy can improve fault tolerance but comes at a cost of significant area overhead as discussed next.

4.5 Memory Area

The memory area of each bank (without any redundancy) can be written as:

$$A_{bank-cl} = N_{row}N_{col}A_{cell} \tag{16}$$

where A_{cell} is the size of the cell used in each bank and N_{row} and N_{col} is the number of rows and columns within each bank. Similarly, the array size of the redundant columns can be written as:

$$A_{bank-rc} = N_{row}N_{rc}A_{cell} \tag{17}$$

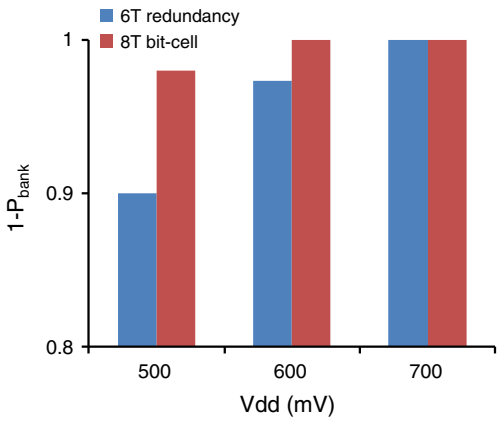

Figure 8 Success probability (1- P_{bank}) of Crucial Bank 1 using column redundancy (6T cells) and 8T-bit cell.

Including the area of the redundant columns then the overall bank area can be written as:

$$A_{bank} = A_{bank-cl} + A_{bank-rc} = N_{row}(N_{col} + N_{rc})A_{cell} \tag{18}$$

Therefore, the overall memory area is given by:

$$A_{array} = A_{bank1} + A_{bank2} + A_{bank3} \tag{19}$$

The decoder size is not included in the above estimation of area since it depends on the number and not in the type of cells used or redundancy, thereby not affecting the overall area comparison.

As mentioned earlier, greater stability of 8T cell comes at the expense of larger area overhead compared to 6T cell. Specifically, the size of 8T cell is 0.83 μm^2, whereas the size of 6T cell is 0.64 μm^2 in the 65 nm technology node. Furthermore, the size of each bank depends on the size of processed image. Let us assume that image size is MxM. Since DCT processes an 8×8 image block the number of the DCT blocks to be processed are (M/8)×(M/8). In each block the most crucial computation is the DC output which according to our significance driven approach, is stored in crucial bank 1. Therefore the size of bank 1 is ((M/8)×(M/8))×N_{bits}, where N_{bits} is the number of bits of each DCT output. On the other hand, the size of the less crucial bank 2 (group 1) is 24×((M/8)×(M/8))×N_{bits} and least crucial bank 3 is 39×((M/8)×(M/8))×N_{bits}. Therefore, for a 256×256 image the usage of 8T bit-cells leads to approximately 28% area overhead for crucial bank which translates to less than 6% area overhead for the overall memory array.

On the other hand, redundancy may improve the failure probability of 6T cells but comes at a cost of significantly larger area overhead. Depending on the degree of redundancy it can lead to 100% area overhead.

4.6 Array Power Estimation

Intuitively, the proposed approach results in read, write and leakage power savings either due to VOS or turning off less crucial banks. For instance, during read operation, power is dissipated due to switching of word line and bit lines. Under VOS, word line power can be reduced significantly as it is evident from the following equation:

$$P_{WL} = N_{col}C_{bitWL}V_{dd}^{2} \tag{20}$$

where, C_{bitWL} is the word line capacitance per cell. 8T bit-cell may have a larger C_{bitWL} than 6T cell, however, it achieves much lower failure probability even under aggressive VOS as shown in previous sections. As evident from Eq. 20, large VOS results in significant (quadratic) power savings. Figure 9(a),(b) show the read and write power of crucial banks using 6T and 8T bit-cells under

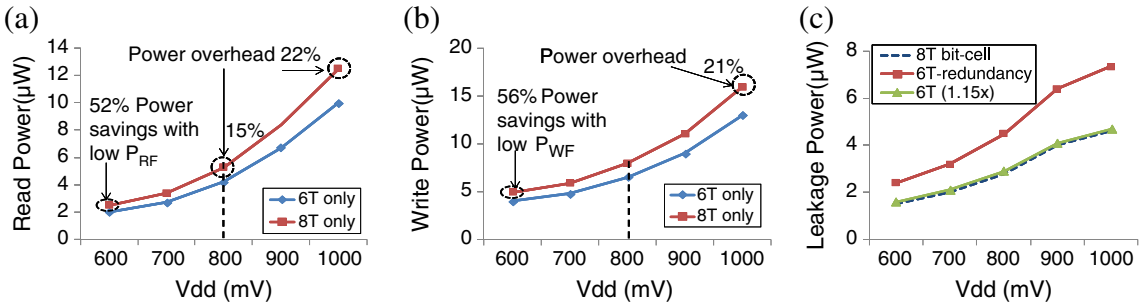

Figure 9 **a** Read, **b** Write and **c** Leakage power comparison for a 32 KBit array using 8T and 6T cells.

VOS. At nominal Vdd the proposed scheme results in 21%–22% read and write power overhead compared to conventional 6T bit-cells. However, the usage of 8T cells allows voltage over-scaling in the crucial bank with negligible impact on overall failure probability (Figs. 5 and 6). Results show that it achieves ~55% power savings for both read and write at 30 MHz frequency.

In addition, the total array leakage, dissipated in the unselected cells in the accessed bank can be approximated by [5]:

$$I_{leak-bank} = \sum_{i=1}^{N_{cells}} I_{leak-cell} \tag{21}$$

where N_{cells} is the total number of cells and $I_{leak\text{-}cell}$ is the leakage of a cell (dominated by the subthreshold current component, since an active array will be operating at higher voltage). It is evident that the usage of redundancy results in larger number of inactive cells (within an accessed bank), leading to higher leakage as shown in Fig. 9(c). On the other hand, the proposed hybrid memory array (crucial 8T and less crucial 6T banks) provides good tradeoff between memory failures, area overhead and power savings. Not only it facilitates aggressive VOS (Fig. 9), resulting in reduction of active power, but also can effectively lower the leakage power dissipated in the accessed bank due to less number of unselected cells (compared to redundancy) and VOS (exponential dependence of leakage on Vdd).

Moreover, active and leakage is further reduced by turning off the less crucial banks in case they are not required to be accessed. By doing so more than 65% of overall memory power can be reduced. The combination of VOS (crucial bank) and turning off of less crucial banks can result in more than 80% memory power savings. All in all, the proposed scheme provides just in time power and quality with small area overhead, which are necessary characteristics of today's complex systems.

5 System Level Trade-offs

In section 3.5 we presented results for the proposed IDCT assuming DCT operation in the absence of VOS. In this section, we show the power/quality trade-offs of proposed IDCT, operating in conjunction with a Vdd scalable DCT under different operating conditions (VOS, process variations). We also discuss power reduction techniques in other parts of the sub-system, memory and quantization blocks and provide the details of a low overhead adaptive compensation circuit. Note that we focus on the DCT/IDCT, quantization and memory blocks that can benefit from the proposed approach, however other parts of such a multimedia system are also present and taken into consideration. Specifically, in order to verify the proposed approach a JPEG system was considered and the proposed modified blocks were integrated in it. Note that there is no any implication in other blocks, such as the zig-zag scan or run-length and entropy encoder other that the ones that are evident (i.e. under low power mode there is no need to encode and access the less-crucial computations). In addition the proposed approach does not impose any modification in the representation of the data and the bit widths conform to conventional blocks used in JPEG

5.1 Impact of VOS on DCT/IDCT System

In Fig. 10 and Table 4 the output image quality of the multimedia sub-system at different scaled Vdds of DCT (Vdd_A) and IDCT (Vdd_B) is presented. Let us assume that user specification does not require highest possible quality, but rather demands minimization of energy to prolong battery life. In such a scenario, system quality (Q_{SYS}) obtained by operating both the blocks at Vdd_{nom} (best quality for this system) may not be desirable. Let us consider a situation in which VOS is applied only to DCT (Vdd_{A2}) while IDCT is allowed to operate at Vdd_{nom}. However, note that these parameters (Vdd_{A2} for block A and Vdd_{Bnom} for block B) may not be optimum for

Figure 10 Proposed system with DCT at Vdd_1 and IDCT at **a** Vdd_{nom}, **b** Vdd_1, **c** Vdd_2, **d** Conventional System at scaled Vdd_1.

achieving minimum system power (P_S) under the given conditions. Interestingly, the system quality Q_{SYS} obtained when DCT operates at Vdd_{A2} and IDCT at Vdd_{Bnom} is comparable to Q_{SYS} that is obtained when IDCT is also operating at Vdd_{B2} as shown in Table 4. Hence, we note that, even if IDCT operates at Vdd_{nom} (consuming higher power), it cannot improve Q_{SYS} which has been already degraded due to VOS in DCT. In such a scenario, the proposed approach dictates VOS in IDCT (Vdd_2) to reduce P_S, with minor degradation in quality as shown in Fig. 10(a, b,c). This is achieved by ensuring that, only the less-crucial computations in IDCT (that use DCT outputs affected by VOS (z_1-z_7)) can get affected under VOS. However, if IDCT continues to operate at Vdd_{nom} (using z_1-z_7), Q_{SYS} may be degraded further at additional cost of processing power. Conversely, when VOS is applied to IDCT, Vdd in DCT can also be scaled down for lowering system power consumption. Such VOS in DCT will incur minor degradation in Q_{SYS} since it is already degraded due to VOS in IDCT (Table 4 and Fig. 10). Figure 10(d) shows the image obtained from a conventional DCT/IDCT system affected by VOS errors. Note that low compression ratio was used in order to be able to capture the impact of modified coefficients on quality. Furthermore, note that in case of a memory consisting of 6T cells, then under VOS the output image has artifacts similar to Fig. 10(d) due to large failure probability that was shown earlier in section 4.

5.2 Adaptive Compensation Circuit

Figure 11 shows the adaptive compensation circuit which acts as a power/quality aware controller (shown in Fig. 1). It controls the mux-enable signals of the DCT/IDCT architecture based on Vdd and process corner. The muxes in the architecture prevent incomplete or invalid computations from propagating to the final IDCT output thereby preventing possible degradation in Q_{SYS}. When Vdd of DCT is scaled down, the outputs affected by VOS (z_1-z_7) can be muxed to zero and prevented from contributing to the IDCT outputs, reducing power further (less switching activity). In such a scenario, Vdd in IDCT can also be scaled down by using the proposed architecture. Furthermore, the compensation circuit provides the block address in memory, selecting the banks that are going to be needed. Specifically, under VOS in DCT, less-crucial computations are not required by IDCT, hence the less crucial banks are disabled leading to further power reduction.

The compensation circuit consists also of a process corner detector that identifies the process corner of the chip. A lookup table, pre-calibrated at design time through extensive HSPICE simulations stores the details of how many adders are properly computed at each voltage and process corner. The lookup table receives this information and generates the corresponding mux-control logic signals. Interestingly, there is no delay overhead due to mux control signals during runtime, since the lookup table statically provides them way in advance. Note, that efficient low overhead level converters [1,

Table 4 Image PSNR with DCT (A) and IDCT (B) at various Vdds.

PSNR(dB)	Vdd_{Anom} (1.2 V)	Vdd_{A1} (1.03 V)	Vdd_{A2} (0.88 V)
Vdd_{Bnom}(1.2 V)	34.49	28.98	22.4
Vdd_{B1} (1.1 V)	28.9	28.98	22.4
Vdd_{B2}(0.71 V)	23.81	23.84	22.4

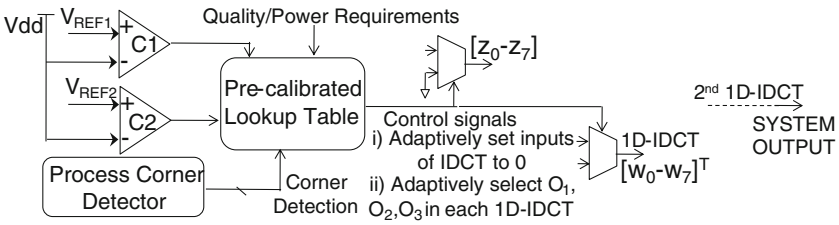

Figure 11 Power/Quality/Process aware controller.

18] can be used for blocks to operate at different Vdd levels. Alternatively, under scaling we could assign only the maximum lower voltage of the blocks in order to avoid the overhead of level converters.

5.3 Memory Blocks Within 2D-IDCT

Memory is an integral part in the implementation of 2D-IDCT involving computation of two 1D-IDCT in a sequential manner (apart from memory storing DCT outputs). Specifically, 64 outputs need to be stored after the application of 1D-IDCT and after transposition, a second 1D-IDCT is applied. Since the size of memory required is small, a register file can be used. In this case, less crucial computations do no need to be stored under user low power requirements since they are not used by the following 1D-IDCT. Therefore, in case that z_5-z_7 are not computed correctly due to VOS in DCT, then 39 memory registers are not required. In this case, as we explained above, the adaptive compensation circuit sets the intermediate 1D-IDCT outputs that are not needed to zero. Therefore, extra power savings are obtained.

5.4 Quantization/de-quantization

Additional power savings can also be obtained by optimizing the quantize/de-quantize blocks. Note that the quantization (Q) block that follows 2D-DCT [9, 10, 12] is responsible for setting the DCT outputs with small magnitude to zero. This is achieved by multiplying the DCT outputs with quantization tables that have the property of setting small DCT coefficients to zero. Note that there are various quantization tables specific to each image coding standard. In this paper, the quantization constants of JPEG standard are used. Conventionally such operation is implemented with multipliers. By using Wallace tree multipliers we found out that the quantization of an 8×8 block requires 12 mW (at Vdd_{nom}). Same is true for the de-quantization (IQ) block that precedes the 2D-IDCT block [9, 10, 12].

However, using the proposed scheme and utilizing the adaptive compensation circuit, extra power savings can be obtained. Specifically, setting DCT outputs (affected by VOS) to zero (through the muxes of sensing circuit (Fig. 11)) can be considered as a quantization operation. Therefore, under low power/low quality user requirements, the number of multiplications required for Quantization can be adaptively reduced, unlike [9, 10] where all multiplications need to be performed. Specifically, 39 (at Vdd_1) or 63 (at Vdd_2) multiplications can be eliminated for each 8×8 block resulting in further power reduction.

5.5 System Level Exploration—Process Variations

Quality/power/process variations trade-offs for the DCT/IDCT system are shown in Fig. 12. In particular, the minimum Vdd required for correct operation of DCT/IDCT at various process corners (slow-slow (SS), slow-fast (SF), typical (TT), fast-fast (FF), fast-slow (FF)) and different complexity/quality levels were determined by using HSPICE and ensuring that the activated block's critical path meets the block's delay target (delay at TT corner and Vdd_{nom}). For instance, if the chip operates at SS corner then in order to obtain a desired quality (say Q1=34 dB) the DCT/IDCT blocks have to be operated at 1.27 V, 1.31 V respectively, which results in 21% increase in system power when compared to nominal conditions (Vdd_{nom}, TT corner). However, the proposed system allows aggressive VOS in DCT/IDCT up to 0.92 V/0.79 V, achieving 33% power reduction even at the slow corner operation at the expense of quality degradation (~5 dB). Under aggressive VOS and FF corner, 69% power savings can be obtained.

Note that extra power savings are obtained by turning off less crucial banks of the memory, when they are not required. Such power savings increase further by applying VOS in

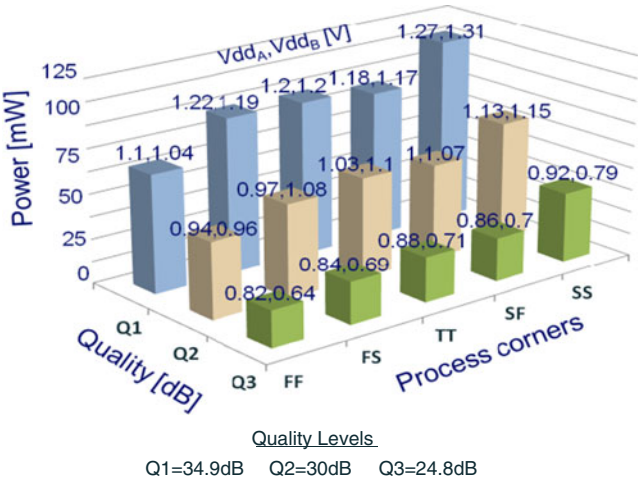

Figure 12 Design space exploration of multimedia sub-system.

memory. The utilization of robust 8T cells in crucial bank of memory ensure correct operation even at reduced Vdd of 0.6 V. Note that in image applications the performance degradation resulting from supply scaling can be considered as a minor issue since the frequency requirement is not high (~30 MHz), thus any deterioration in access delay due to VOS, would not affect the overall operation [7].

6 VOS as a Channel Noise Concealment Technique

In this section, rather than considering DCT and IDCT as part of the same portable device, we assume that DCT outputs are transmitted over a noisy wireless channel and decoded by IDCT on the receiver side. We show that by applying VOS in the receiver side, the proposed IDCT design provides intelligent tradeoff between channel noise and actual quality of the received/decoded image. In the following analysis we do not consider channel encoder/decoder (assuming an un-coded wireless system) for easier understanding and mathematical representation of the system. We model the wireless channel as additive white Gaussian noise (AWGN). The noisy channel output (input to IDCT) can be represented as:

$$Y_i = Z_i + \varepsilon_i \tag{22}$$

where Z_i is the 'input' to the channel and ε_i is an independent and identical distributed (i.i.d.) random variable with zero mean and variance σ_n^2 (noise). Interestingly, the proposed VOS based IDCT inherently achieves channel noise concealment in the receiver part. In presence of channel noise, the mean value of 1D-IDCT outputs equals E[w_i] (since E[ε_i]=0), while the variance is given by:

$$Var([w_i]) = \left(C_1^2 + C_2^2 + C_3^2 + C_4^2\right)\sigma_n^2 \tag{23}$$

which is obtained by using Eq. 22, Eq. 8 and property of random variables, Var($a\chi_1+b\chi_2$)=a^2Var(χ_1)+b^2Var(χ_2) for independent χ_1 and χ_2 (Z_i deterministic). By substituting the values of the coefficients in Eq. 23 we obtain $\mathrm{Var}(w_i) = 1.01\sigma_n^2$. When Vdd in IDCT is scaled down to Vdd_1, less-crucial term (Γ_3) involving matrix C_4 may not get computed. In such a scenario, we observe that the variance of 1D-IDCT outputs (Eq. 23) due to channel noise is reduced:

$$Var\left(\begin{bmatrix} w_{0,4} \\ w_{1,5} \end{bmatrix}\right) = \left(\begin{bmatrix} 0.88 \\ 0.46 \end{bmatrix}\right)\sigma_n^2,\ Var\left(\begin{bmatrix} w_{2,6} \\ w_{3,7} \end{bmatrix}\right) = \left(\begin{bmatrix} 0.61 \\ 0.55 \end{bmatrix}\right)\sigma_n^2$$

Furthermore, under aggressive VOS in IDCT (Vdd_2), the Var(w_i) is reduced to $0.12\sigma_n^2$. Hence, in addition to obtaining power benefits, VOS in IDCT can be used for channel error concealment as shown in Fig. 13. Under "bad" (high σ_n^2) channel conditions, the output image (obtained from IDCT) is degraded due to channel noise as shown in Fig. 13(a). In this case the system would be unnecessarily consuming higher power without achieving better quality. However, under VOS in IDCT, the image artifacts (salt and pepper) due to channel noise are reduced as shown in Fig. 13(b,c). But VOS also results in image quality degradation due to the omitted less-crucial terms. Therefore, the proposed IDCT can be used for trading-off quality degradation due to VOS in favor of channel error immunity. Note that, even if the mean of IDCT outputs is shifted (due to exclusion of less-crucial terms) the randomness (σ_n^2) in each output is reduced. This results in reduction of salt and pepper noise at the expense of an increase in image "blurriness" (Fig. 13(c)). In summary, the proposed architecture not only reduces power by VOS but also eliminates the need for power hungry concealment methods [19], with graceful image quality degradation.

At this point we would like also to note that unequal error protection was applied in various standalone blocks such as motion estimation, color interpolation, wavelet transform and FIR filtering [20] apart from DCT. As this paper points out their integration in systems needs to consider the block level

Figure 13 Channel noise (SNR=9 dB) reduction (DCT at Vdd_{nom}) during transmission of DCT outputs using BPSK modulation.

interactions with other logic and memory sub-blocks in order to limit any overhead and benefit also other blocks.

7 Conclusion

In this paper, we propose a design approach based on system level interactions that provides "right" amount of quality and power for each block, while dissipating minimal system power under the given operating conditions. This is achieved by ensuring that, only the less-crucial computations can get affected by the operating conditions (VOS, process variations, channel noise), thereby incurring minimal quality degradation. An unequal error protection scheme was also applied to memory design that allows significant power reduction by disabling adaptively, less crucial banks of the memory. Furthermore, the utilization of robust 8T cells in crucial banks allowed voltage over-scaling resulting in further power reduction in the system memory. In addition, the VOS based design methodology can be used as a channel error concealment technique while minimizing system power consumption (by VOS). To demonstrate the efficiency of the proposed approach, it was applied to a multimedia sub-system. However, the proposed methodology can also be applied for designing other DSP systems where interaction between blocks need to be considered for providing "just-the-right" amount of power/quality and robustness.

Acknowledgement This research was funded in part by Gigascale Systems Research Center (GSRC) and by National Science Foundation (NSF).

References

1. Rabaey, J., Chandrakasan, A., & Nikolic, B. (2002). Digital integated circuits. Prentice Hall.
2. Borkar, S., Karnik, T., Narendra, S., Tschanz, J., Keshavarzi, A., & De, V. (2003). Parameter variations and impact on circuits and microarchitecture. *IEEE Design Automation Conf.*, 338–342.
3. Karakonstantis, G., Banerjee, N., & Roy, K. (2010). Process-variation resilient & voltage scalable DCT architecture for robust low-power computing. *IEEE Trans VLSI Systems, 18*(10), 1461–1470.
4. Shim, B., Sridhara, S. R., & Shanbhag, N. (2004). Reliable low-power digital signal processing via reduced precision redundancy. *IEEE Trans VLSI Systems, 12*(5), 497–510.
5. Mukhopadhyay, S., Mahmoodi, H., & Roy, K. (2005). Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS. *IEEE Trans CAD, 24*(12), 1859–1880.
6. Chang, L., Montoye, R. K., Nakamura, Y., Batson, K. A., Eickemeyer, R. J., Dennard, R. H., et al. (2008). An 8T- SRAM for variability tolerance and low-voltage operation in high-performance caches. *IEEE Journal of Solid-State Circuits, 44*(4), 956–963.
7. Ik Joon Chang, Mohapatra, D., & Roy, K. (2009). A voltage-scalable & process variation resilient hybrid SRAM architecture for MPEG-4 video processors. *IEEE Design Automation Conference*, 670–675.
8. Chiang, J.-S., Chiu, Y.-F., & Chang, T.-H. (2001). A high throughput 2- dimensional DCT/IDCT architecture for real-time image and video system. *IEEE ICECS*, 867–870.
9. Xanthopoulos, T., & Chandrakasan, A. (1999). Low-Power IDCT macrocell for MPEG-2 MP@ML exploiting data distribution properties for minimal activity. *IEEE Journal of Solid-State Circuits, 34*(5), 693–703.
10. Lengwehasatit, K., & Ortega, A. (1998). DCT computation based on variable complexity fast approximations. *IEEE ICIP*.
11. Banerjee, N., Karakonstantis, G., & Roy, K. (2009). Design methodology for low power dissipation and parametric robustness through output quality modulation: application to color interpolation filtering. *IEEE Trans CAD, 28*(8), 1127–1137.
12. Bhaskaran, V., & Konstantinides, K. (1996). *Image and video compression standard algorithms and architecture*. Kluwer Publishers.
13. http://r0k.us/graphics/kodak/
14. Kim, C. H., Roy, K., Hsu, S., Krishnamurthy, R. K., & Borkar, S. (2005). On-die CMOS leakage current sensor for measuring process variation in sub- 90 nm generations. *IEEE ICIDT*, 221–222.
15. Synopsys Design Compiler, www.synopsys.com
16. ARMLIBS 90 nm, http://www.arm.com/
17. Mentor Graphics Calibre, www.mentor.com
18. Kulkarni, S. (2003). New level converters and level converting Logic circuits for multi-VDD low power design. *IEEE System-on-Chip*.
19. Hsia, S. C., & Chou, S.-W. (2007). VLSI implementation of high-performance error concealment processor for TV broadcasting. *IEEE Trans CSVT, 17*(8), 1054–1064.
20. Karakonstantis, G., Roy, K., & Chatterjee, A. (2011). Concealing the Nanometer "Pandora Box": Cross Layer Design Techniques for Variation Aware Low Power Systems. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 19–29.

Georgios Karakonstantis received the Diploma degree in Computer and Communications Engineering from Polytechnic School of University of Thessaly, Volos, Greece, in 2005 and his PhD and Master degree in 2010 after pursuing his research work in the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN. In the summer of 2008, he was with the Advanced Technology Group, Qualcomm Incorporated, San Diego, CA as an intern. His research interests include cross-layer hardware design methodologies for low power and process variation tolerant circuits, architectures and wireless communication systems. His work has appeared in more than 20 referred journals and conferences. In January 2011 he joined the Telecommunications Circuits Lab at EPFL as a research scientist and leads the research efforts for the design of energy efficient and reliable circuits and systems

Debabrata Mohapatra received the B.Tech. degree (Hons.) in electrical engineering and a minor in electronic communication engineering from Indian Institute of Technology, Kharagpur, India, in 2005. He received his Ph.D in the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN. He is currently a research scientist with the Microprocessor Research Laboratory, Intel, Santa Clara, CA. He spent the summer of 2008 in Qualcomm"s memory design group. His research interests include low-power, process-variation aware circuit and system design for multimedia hardware in nanometer technologies.

Kaushik Roy received B.Tech. degree in electronics and electrical communications engineering from the Indian Institute of Technology, Kharagpur, India, and Ph.D. degree from the electrical and computer engineering department of the University of Illinois at Urbana-Champaign in 1990. He was with the Semiconductor Process and Design Center of Texas Instruments, Dallas, where he worked on FPGA architecture development and low-power circuit design. He joined the electrical and computer engineering faculty at Purdue University, West Lafayette, IN, in 1993, where he is currently a Professor and holds the Roscoe H. George Chair of Electrical & Computer Engineering. His research interests include Spintronics, VLSI design/CAD for nano-scale Silicon and non-Silicon technologies, low-power electronics for portable computing and wireless communications, VLSI testing and verification, and reconfigurable computing. Dr. Roy has published more than 500 papers in refereed journals and conferences, holds 15 patents, graduated 50 PhD students, and is co-author of two books on Low Power CMOS VLSI Design (John Wiley & McGraw Hill).

Dr. Roy received the National Science Foundation Career Development Award in 1995, IBM faculty partnership award, ATT/Lucent Foundation award, 2005 SRC Technical Excellence Award, SRC Inventors Award, Purdue College of Engineering Research Excellence Award, Humboldt Research Award in 2010, and best paper awards at 1997 International Test Conference, IEEE 2000 International Symposium on Quality of IC Design, 2003 IEEE Latin American Test Workshop, 2003 IEEE Nano, 2004 IEEE International Conference on Computer Design, 2006 IEEE/ACM International Symposium on Low Power Electronics & Design, and 2005 IEEE Circuits and system society Outstanding Young Author Award (Chris Kim), 2006 IEEE Transactions on VLSI Systems best paper award. Dr. Roy is Purdue University Faculty Scholar. He was a Research Visionary Board Member of Motorola Labs (2002) and held the M.K. Gandhi Distinguished Visiting faculty at Indian Institute of Technology (Bombay). He has been in the editorial board of IEEE Design and Test, IEEE Transactions on Circuits and Systems, and IEEE Transactions on VLSI Systems. He was Guest Editor for Special Issue on Low-Power VLSI in the IEEE Design and Test (1994) and IEEE Transactions on VLSI Systems (June 2000), IEE Proceedings–Computers and Digital Techniques (July 2002). Dr. Roy is a fellow of IEEE.