Process Variation Tolerant Low Power DCT Architecture

Nilanjan Banerjee, Georgios Karakonstantis and Kaushik Roy
Purdue University, West Lafayette, IN-47907, USA.
Email:{nbanerje, gkarakon, kaushik}@purdue.edu

Abstract: ***2-D Discrete Cosine Transform (DCT) is widely used as the core of digital image and video compression. In this paper, we present a novel DCT architecture that allows aggressive voltage scaling by exploiting the fact that not all intermediate computations are equally important in a DCT system to obtain "good" image quality with Peak Signal to Noise Ratio(PSNR) > 30 dB. This observation has led us to propose a DCT architecture where the signal paths that are less contributive to PSNR improvement are designed to be longer than the paths that are more contributive to PSNR improvement. It should also be noted that robustness with respect to parameter variations and low power operation typically impose contradictory requirements in terms of architecture design. However, the proposed architecture lends itself to aggressive voltage scaling for low-power dissipation even under process parameter variations. Under a scaled supply voltage and/or variations in process parameters, any possible delay errors would only appear from the long paths that are less contributive towards PSNR improvement, providing large improvement in power dissipation with small PSNR degradation. Results show that even under large process variation and supply voltage scaling (0.8V), there is a gradual degradation of image quality with considerable power savings (62.8%) for the proposed architecture when compared to existing implementations in 70 nm process technology.***

1. INTRODUCTION

Energy-aware designs are necessary to prolong the battery lifetime of portable devices, to prevent excessive heat generation which might result in device reliability problems, and to reduce the cost associated with expensive cooling techniques. With increasing demand of video messaging in multimedia/wireless communications, development of low-energy image/video transmission schemes are necessary [1, 2]. Conventional image compression schemes are designed to minimize distortion of the reconstructed image for a given bit-rate. However, applications like portable multimedia may not always require the best image quality [3]. This aspect can be effectively exploited to obtain architectures that provide the "right" trade-off between image quality and energy consumption.

Discrete Cosine Transform (DCT) is important in the field of video/image compression due to its inherent capability of achieving high compression rates at low design complexity. A lot of research has been devoted to reduce number and complexity of computations in DCT architectures [4, 5]. Low-power requirements for image compression have resulted in several other DCT architectures that used partial computation [6] or dual-threshold voltages [7]. However, low power is not the only requirement in today's designs. With technology scaling, process parameter variations pose a major design concern. Studies have shown [8] that parameter variations create a delay spread of almost 30% for 70 nm process technology, leading to delay failures in some chips. Conventional wisdom dictates a conservative design approach (e.g., scaling up the Vdd or upsizing logic gates) to prevent delay failures and to achieve high parametric yield. However, such techniques come at a cost of increased power and/or die area. Therefore, process tolerance and low power represent contradictory design requirements. In this paper, *we simultaneously target low power and process tolerance by proposing an architecture amenable to voltage scaling even under parameter variations*. Our contributions are as follows:

- Identify computational paths that are vital in maintaining high image quality
- Develop an algorithm/architecture that makes more important computations (in terms of image quality) to have shorter paths than less important ones
- Utilization of this architecture to make any path-delay errors predictable under a single scaled supply voltage and process parameter variations, and to tolerate delay failures in such paths with minimal PSNR degradation of image
- Reconfiguration of the architecture to provide trade-offs between image quality and power consumption

The paper is organized as follows. The proposed DCT architecture operated under a single scaled Vdd is presented in Section 2. Implementation details and the results of scaled-Vdd scheme are elaborated in Section 3. The process tolerance capabilities of this architecture are presented in Section 4. Section 5 concludes the paper.

2. DCT ARCHITECTURE: PRINCIPLES AND DESIGN

In this section, we briefly mention the underlying principle of conventional DCT systems and then propose a new DCT architecture for low power. Though our DCT implementation considers 8-bit coefficients, the technique can be easily extended to 12-bit or 16-bit DCT coefficients.

2.1 Conventional DCT

Conventional 2D-DCT [13] can be shown with the following block diagram (Fig. 1) which shows that the 2D-DCT can be separated in two 1-D DCTs.

Fig.1. 2-D DCT architecture expressed as two 1-D DCT transforms

The intermediate computation is a 1D-DCT unit that transforms an 8 X 8 image block from spatial domain to frequency domain. The 1-D DCT transform is expressed as:

$$w_k = \frac{c(k)}{2} \sum_{i=0}^{7} x_i \cos \frac{(2i+1)k\pi}{16}, \; k = 0,1,2,\ldots,7 \qquad (1)$$

$$c(k) = \begin{cases} 1/2 & k = 0 \\ 1 & otherwise \end{cases}$$

In vector-matrix form, the same equation can be written as:

$$w = T \bullet x^t , \qquad (2)$$

where, T is an 8 X 8 matrix with cosine functions as its elements, and x and w are row and column vectors, respectively [9].

The 8 X 8 coefficient matrix T is symmetric and this property is used for even/odd 1-D DCT calculation in the following manner:

$$\begin{bmatrix} w_0 \\ w_2 \\ w_4 \\ w_6 \end{bmatrix} = \begin{bmatrix} d & d & d & d \\ b & f & -f & -b \\ d & -d & -d & d \\ f & -b & b & -f \end{bmatrix} \begin{bmatrix} x_0 + x_7 \\ x_1 + x_6 \\ x_2 + x_5 \\ x_3 + x_4 \end{bmatrix} \qquad (3)$$

978-3-9810801-2-4/DATE07 © 2007 EDAA

$$\begin{bmatrix} w_1 \\ w_3 \\ w_5 \\ w_7 \end{bmatrix} = \begin{bmatrix} a & c & e & g \\ c & -g & -a & -e \\ e & -a & g & c \\ g & -e & c & -a \end{bmatrix} \begin{bmatrix} x_0 - x_7 \\ x_1 - x_6 \\ x_2 - x_5 \\ x_3 - x_4 \end{bmatrix} \quad (4)$$

where, $c_k = cos(n\pi/16)$, $a=c_1$, $b=c_2$, $c=c_3$, $d=c_4$, $e=c_5$, $f=c_6$, *and* $g=c_7$. We can rearrange eqn. (3) in the following manner:

$$\begin{bmatrix} w_0 \\ w_2 \\ w_4 \\ w_6 \end{bmatrix} = (x_0+x_7)\begin{bmatrix} d \\ d \\ d \\ d \end{bmatrix} + (x_1+x_6)\begin{bmatrix} b \\ f \\ -f \\ -b \end{bmatrix} + (x_2+x_5)\begin{bmatrix} d \\ -d \\ -d \\ d \end{bmatrix} + (x_3+x_4)\begin{bmatrix} f \\ -b \\ b \\ -f \end{bmatrix} \quad (5)$$

Eqn. (4) can be expressed in a similar format. As shown in eqn. (5), each of the outputs in a 1D-DCT computation is simply the additions of vector scaling operations. It should be noted that several optimization techniques have been proposed [1] to reduce the number of operations in DCT computation.

1	2	6	7	15	16	28	29
3	5	8	14	17	27	30	43
4	9	13	18	26	31	42	44
10	12	19	25	32	41	45	54
11	20	24	33	40	46	53	55
21	23	34	39	47	52	56	61
22	35	38	48	51	57	60	62
36	37	49	50	58	59	63	64

Fig.2. Energy distribution for 2-D DCT matrix

Note that all coefficients of the 2-D DCT matrix do not affect the image quality in a similar manner. Analysis conducted on various images like Lena, Peppers etc. show that most of the input image energy (around 85% or more) is contained in the first 20 coefficients (Fig. 2) of the DCT matrix after the 2D-DCT operation. The coefficients beyond that (21-64) contribute significantly less in the improvement of image quality and hence, the PSNR. Fig. 2 shows the energy distribution, known as zig-zag scan, (from 1 to 64) for the final DCT matrix.

2.2 Proposed DCT architecture

In this subsection, we introduce the underlying concept utilized in designing our DCT architecture. From Fig. 2, we can infer that the energy content is distributed non-uniformly across the final DCT matrix. With this information in mind, we propose an architecture that computes the high-energy components of the final DCT matrix faster than the low-energy components. The design methodology developed for this architecture is shown in Fig. 3. While computing the 1-D DCT, we calculate the first 5x8 sub-matrix (marked as "Faster Computation" in Fig.3 (b)) earlier than the remaining 3X8 values (marked as "Slower Computation" in Fig.3 (b)). To explain this further, let us first consider how a conventional pipelined DCT system works. In a pipelined DCT system, usually eight pixel values (x_0-x_7) corresponding to one column are input at a time and 1-D DCT is performed on them to obtain the values w_0-w_7 (Fig. 3(b)) in a manner shown in eqn. 5. In the next clock cycle, the next set of intermediate values (w_8-w_{15}) is computed corresponding to inputs x_8-x_{15} and, so on. Our architecture is designed in a way that in each clock cycle the first five values (e.g. values w_0-w_4 for inputs x_0-x_7 since each w computation takes the entire column values x_0-x_7) are evaluated faster than the remaining values (w_5-w_7), which take longer time to be computed. Therefore, the 5x8 sub-matrix is "faster" than the remaining 3x8 sub-matrix.

The transposition of the intermediate 1-D matrix (Fig. 3(b)) results in the matrix shown in Fig. 3(c). Since the 5x8 sub-matrix is

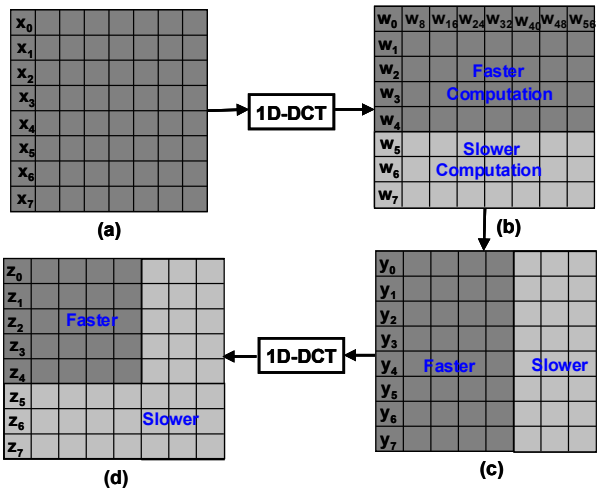


Fig. 3.(a) Input Pixel Matrix (b) Matrix after first 1-D DCT (c) Matrix after being transposed (d) Final 2-D DCT matrix

computed faster in Fig. 3(b), the corresponding transpose results in earlier computation of the 8x5 sub-matrix as shown in Fig. 3(c). The second 1-D DCT operation results in faster evaluation of the first 5 (e.g. z_0-z_4) values for each of the input columns (x_0-x_7 etc.). This enables fast computation of the first 5x5 sub-matrix of the final DCT matrix. This 5x5 sub-matrix includes all the high energy components (1-20) of the DCT matrix. The rest of the matrix is computed "slowly". Therefore, the computational paths for high energy components are shorter than their low energy counterparts. Designing the architecture in such a manner provides three distinct advantages:

a) it helps isolate the computational paths based on high energy and low energy contributing components,

b) it allows supply voltage scaling (single supply) to trade-off power dissipation and image quality even under process parameter variations.

2.3 Modification of path-lengths

To scale the supply voltage and to obtain power savings, it is necessary to skew the different path-lengths in the DCT computation. In this sub-section, we describe the step-by-step procedure that guarantees that the path-lengths for computing the first five elements (w_0-w_4) of the DCT computation are shorter than that of the remaining three elements (w_5-w_7). We achieve this by creating a relationship between the DCT coefficients. Later we will observe that such path-skewing leads to supply voltage scaling for low-power operations even under process parameter variations.

Let us consider the original 8-bit DCT coefficients shown in Table 1. We slightly alter the coefficient values as shown in Table 2. The reason for such modification will be clear shortly. This modification should be performed carefully so that it has minimal effect on the image quality. We keep the value of the coefficient "d" unchanged in

Table 1. Original 8-bit DCT coefficients

DCT Coef	Value	Binary Number
a	0.49	0011 1111
b	0.46	0011 1011
c	0.42	0011 0101
d	0.35	0010 1101
e	0.28	0010 0100
f	0.19	0001 1000
g	0.10	0000 1100

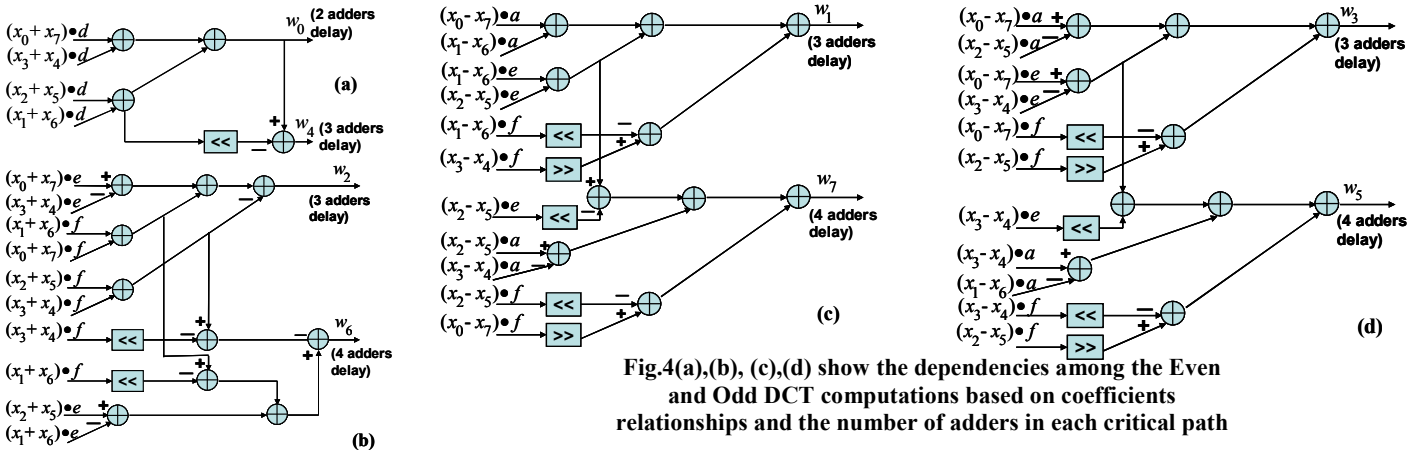


Fig.4(a),(b), (c),(d) show the dependencies among the Even and Odd DCT computations based on coefficients relationships and the number of adders in each critical path

Table 2. Modified DCT coefficients

DCT Coef	Value	Binary Number
a	0.50	0100 0000
b	0.47	0011 1100
c	0.41	0011 0100
d	0.35	0010 1101
e	0.28	0010 0100
f	0.19	0001 1000
g	0.10	0000 1100

Table 3.Relationships Among coefficients

Coef	Dep Expr
a	a
b	e + f
c	a + e – 2*f
d	d
e	e
f	f
g	f/2

this process since it computes the DC component (lowest frequency) in a 2D-DCT matrix and is most important in determining image quality. Any modification to this coefficient degrades the image quality by a considerable amount [10]. The PSNRs with the original and the modified coefficients are shown in Table 6. The next step involves the establishment of a dependency among the various coefficients. As shown in Table 3, we retain coefficients a, d, e and f to be the same as that in Table 2. The coefficients b, c and g are expressed in terms of a, e, and f (possible because of the slight modification of the coefficients).

While incorporating these dependencies among the coefficients, we make sure that the delay impact on the even and odd DCT computations is minimal. It should be noted that the clock-cycle of the DCT pipeline is determined by the delay of the longest path (either w_5 or w_6 or w_7).

Based on these modifications, path-lengths for Even and Odd DCT components are shown in Fig. 4 (with the dependencies shown in Table 3). We observe from Fig. 4(a), (b), (c) and (d) that the computations of w_4, w_5, w_6, w_7 are dependent on w_0, w_1, w_2 *and* w_3. In terms of critical path-lengths, w_0 has 2 adders, w_1-w_4 has 3 adders and w_5-w_7 has 4 adders. This ensures that the delay in paths w_0-w_4 to be always less than paths w_5-w_7. As mentioned in Section 2, we exploit the delay difference in computational path-lengths to effectively scale down the voltage and to make trade-offs between power dissipation and image quality under process parameter variations.

3. IMPLEMENTATION OF SCALED-Vdd DESIGN

To implement the scaled-Vdd approach for low power, we employ the CSHM scheme proposed in [11]. CSHM scheme is based on the principle of vector scaling, where a set of small bit sequences (alphabets) are selected that covers the set of coefficients. Multiplication of the alphabets and the inputs are computed ahead and the final multiplication results are obtained by shift and add operations on the precomputed values. For instance, a simple vector scaling operation $[c_0,c_1]\cdot x$, c_0 = *01100111*, c_1=*10001011*, can be easily decomposed as $c_0.x=2^5\cdot(0011)x+(0111)\cdot x$, $c_1.x=2^7\cdot(0001)x+(1011)\cdot x$. If x, $(0011)x$, $(0111)x$ and $(1011)x$ are available (precomputed), the entire multiplication process is reduced to a few add and shift operations only. An alphabet set is a set of alphabets spanning all the coefficients in vector ***C*** as mentioned earlier. In this example, the alphabet set is {*0001, 0011, 0111, 1011* }. The advantage of using this scheme for the proposed architecture is explained in the following paragraphs. Fig. 5(a) shows the generic implementation of the CSHM architecture. In this architecture, the precomputer banks are shared across the select/shift and adder (SSA) units. The number of SSA units required is determined by the number of multiplications needed to pre-compute the inputs (e.g. x_0+x_7 with d, e, f etc.) shown in Fig. 5. The outputs of the SSA units (e.g. (x_0+x_7)*d etc.) are shared across all computational paths (Fig. 5).

3.1. Hardware Optimization by Reduced Alphabet Set

Let us first consider the original 8-bit DCT coefficients and their corresponding alphabets shown in Table 4.

The CSHM scheme uses the precomputation of 6 alphabets denoted by {1X, 3X, 5X, 11X, 13X, 15X} for the original DCT coefficients. This implies that each of the pre-computers used for the vector scaling operation require 6 precomputer banks and the subsequent multiplexers are of size 6:1 [11]. On the other hand, for our scheme, first, we slightly alter the coefficient values and then express the other coefficients [b, c, g] in terms of [a, e, f, d]. The new set of alphabets {1, 13} required for representing the remaining coefficients are shown in Table 5.

This design modification considerably reduces the sizes of the precomputers and SSA units used for the vector scaling operation of the CSHM implementation [11] since we require only 2 pre-computer banks followed by 2:1 muxes. Moreover, since the number of coefficients are reduced to four (a, d, e, f), the number of SSA units also reduces. Therefore, both the dynamic and leakage power consumption for these pre-computing blocks reduces significantly without any significant degradation to the image quality. The simplified pre-computer unit of the 2-alphabet CSHM is shown in the Fig. 5(b).

Table 4. Original DCT coefficients and their alphabets

Coef	Value	Binary Number	Alphabet x
a	0.49	0011 1111	3x, 15x
b	0.46	0011 1011	3x, 11x
c	0.42	0011 0101	3x, 5x
d	0.35	0010 1101	1x,13x
e	0.28	0010 0100	1x
f	0.19	0001 1000	1x
g	0.10	0000 1100	3x

Table 5. Reduced coefficient set and their alphabets

Coef	Value	Binary Number	Alphabet x
a	0.50	0100 0000	1x
d	0.35	0010 1101	1x,13x
e	0.28	0010 0100	1x
f	0.19	0001 1000	1x

Further optimization on the precomputers can be performed if we consider the odd-DCT and the even-DCT components, separately. We observe from Fig. 4 that the even-DCT components require computations with the coefficients (d, e, f) which require two alphabets {1X, 13X}. On the other hand, the odd-DCT components are calculated with coefficients (a, e, f) requiring only one alphabet {1X}. Moreover, no muxes are required for this implementation, further reducing the power consumption in the pre-computers. The optimized pre-computer for the Odd–DCT implementation is shown in Fig. 5(c). It should be noted that the optimization of the pre-computers also reduces the delay involved in the DCT computations.

3.2. Results

To verify the effectiveness of the scaled-Vdd DCT design option, we compared our architecture to both a conventional architecture [13] implemented with Wallace Tree Multipliers (WTM) and the 2-alphabet CSHM architecture [11] in terms of power consumption, delay, area overhead and image quality (PSNR) under supply voltage scaling. It should be noted that the range of the supply voltage scaling is determined by the difference in delays between the shortest (w_0, path1) and the longest computational paths (w_5-w_7, path 6-8) for the proposed architecture. The design is implemented in VHDL and synthesized using Synopsys Design Compiler [14] to obtain a Verilog netlist. The Verilog netlist is then converted into a Hspice file. The Hspice files were subsequently simulated using BPTM 70 nm technology [12] using 1000 patterns and the average power consumption is determined. The delays of the critical paths are also

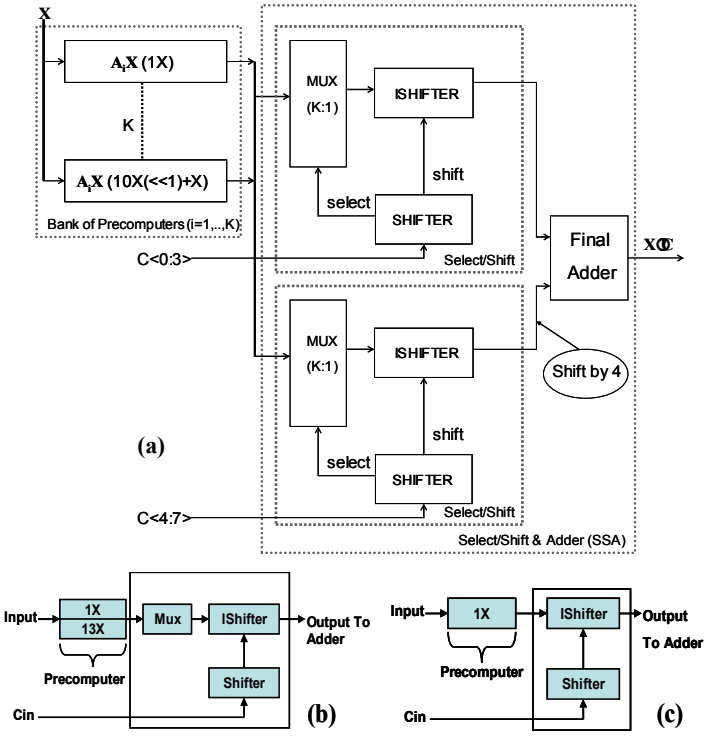


Fig.5. (a) Generic CSHM architecture
(b) Optimized CSHM for Even-DCT (c) for Odd-DCT

Table 6(a) Comparison of different architectures

Vdd=1.0V	CSHM DCT (2 alphabets)	DCT with WTM	Proposed DCT
Power (mW)	25.1	29.8	26
Delay (ns)	3.2	3.64	3.57
Area (um^2)	80490	108738	90337
PSNR (dB)	21.97	33.23	33.22

Table 6.(b) Scaled Vdd results for proposed design (Other architectures fail at scaled voltages)

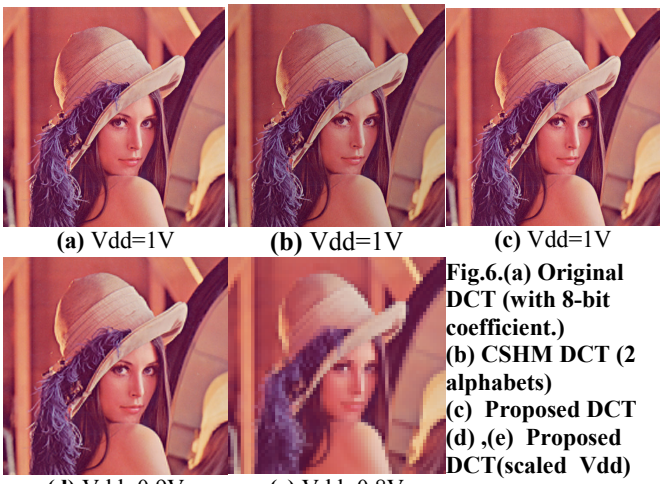
	Proposed DCT Vdd=0.9V	Proposed DCT Vdd=0.8V
Power (mW)	17.53	11.09
PSNR (dB)	29	23.41

noted. The area estimates consists of active transistor areas. The results have been summarized in Table 6.

At the nominal voltage (1V), the longest path delay for the proposed design is similar to conventional DCT approach. Similar performance is attributed to the optimization of the proposed architecture (simpler pre-computer/selection mux), which reduces the number of arithmetic operations (additions/shifts) required to compute the longest path in the proposed design. This improves both the computation delay (1.92%) as well as the power consumption (12.75%) for the proposed technique (Table 6(a)). The area for the proposed architecture is also less than the conventional DCT since the latter consists of area-intensive WTMs. PSNR has been used as a metric to evaluate the image quality. It is observed that our architecture produces a high quality image at nominal Vdd. It should be noted that the 2-alphabet CSHM DCT consumes less power than our architecture at Vdd=1V. However, the PSNR for the various images (>30 dB) is much higher for our design compared to the low power CSHM approach [10] (PSNR =21.9 dB). Fig. 6(a), (b) and (c) show the transformed Lena image for the different designs at a nominal voltage of Vdd=1V for 70nm technology.

We also observe the DCT outputs of different designs as we scale down the voltage from the nominal value. At a supply voltage of 0.9V, the conventional architecture fails! This is because all computational paths in this design are approximately of similar length as shown in Table 7. Hence, reducing the supply voltage prevents any complete DCT computation and drastically affects the PSNR.

The 2-alphabet CSHM also suffers from same drawbacks as previous case (similar pathlengths) and fails to compute DCT outputs



(a) Vdd=1V **(b)** Vdd=1V **(c)** Vdd=1V

(d) Vdd=0.9V **(e)** Vdd=0.8V

Fig.6.(a) Original DCT (with 8-bit coefficient.) (b) CSHM DCT (2 alphabets) (c) Proposed DCT (d) ,(e) Proposed DCT(scaled Vdd)

Table 7. Computational Path delays of DCT outputs at Nominal Vdd(1V) for a 70 nm technology

Computation Path	Delay (ns)	
	Conv. WTM Arch.	Proposed Arch.
Path1(w_0)	3.55	2.12
Path2(w_1)	3.6	2.7
Path3(w_2)	3.63	2.68
Path4(w_3)	3.64	2.82
Path5(w_4)	3.63	2.81
Path6(w_5)	3.59	3.55
Path7(w_6)	3.6	3.56
Path8(w_7)	3.64	3.57

at scaled Vdd's. However, at a nominal Vdd this design had a lower computing delay than the proposed architecture. For a fair comparison, we scale the Vdd to a value that allows this design to operate with same frequency as our design at nominal Vdd. Even under this condition, we find that it is impossible to obtain reasonable image quality below 0.95V.

For the proposed design, we observe a gradual degradation in the image quality as the voltage is scaled down. At 0.9V, only 5 of the 8 paths (w_0-w_4) are computed (design operating as same frequency as nominal Vdd) and a PSNR of 29dB is obtained with power savings of 41.2% compared to the WTM implementation. At 0.8V complete computation of only the first path (w_0) is possible resulting in a PSNR of 23dB and power savings of 62.8%. However, it should be noted that below 0.77V, none of the DCT outputs are computed at the nominal frequency of operation.

4. PROCESS VARIATION TOLERANCE

As mentioned in Section 1, parameter variation is becoming an increasingly serious issue with technology scaling, and it is a very challenging task to concurrently address process variation (represented in terms of delay failures) and low power dissipation. In this section, we discuss how our architecture (Section 2 and 3) is suitable for both low-power consumption and tolerance to parameter variations. The delay failures along these computational paths (due to voltage scaling/process variations) are considered as errors in computations in this discussion. To evaluate the advantages of our approach under effects of process variation, we compare our design to a conventional one.

4.1. Conventional Design under Process Variation

In a conventional DCT design all path-lengths for evaluating the 1-D DCT computations are of almost similar lengths (Table 7). Under process variation the delays in computing these paths may vary depending on the process corner that the chip is in (assuming only inter-die variations; however, our design style is also tolerant to intra-die variations). The worst case computation path-length delay determines the operating frequency of the system. Also, since all the paths have similar delays, it is possible that the paths (w_0-w_4) contributing to the computation of the high-energy components get affected. Fig. 7 shows the effects of process variation for the conventional DCT architecture. The original design was operating at a delay (D) with nominal supply voltage Vdd. Let us consider a case, where under parameter variation, one of the important paths w_0 (denoted by Path 1 in Fig. 7(a)) has increased delay (D_{new} > D) so its computation is not completed at the operational frequency 1/D. In this case, the output image (Fig. 8(a)) is hardly recognizable. To avoid such delay errors and to operate the system at same frequency, we have to either increase Vdd (Vddnew) or upsize transistors along this critical path (Fig. 7(b)). Both solutions result in additional power consumption and/or area overhead. In the worst case, all the important paths might get affected and the overhead might be quite significant to obtain a reasonable output image.

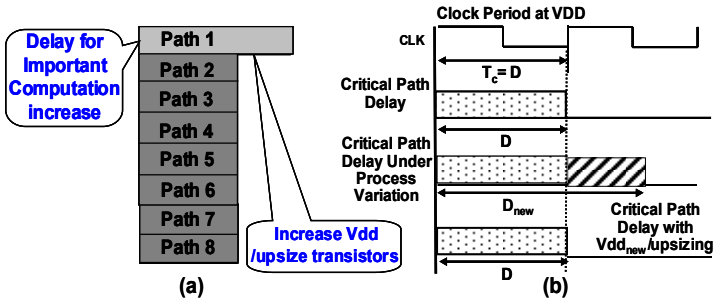
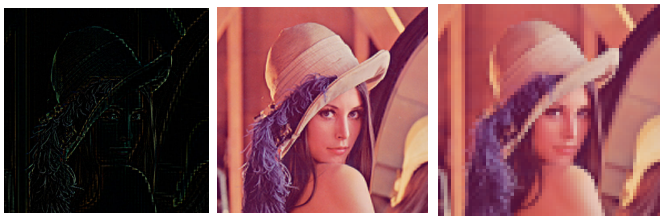


Fig. 7.(a) Short paths of Conv. Design affected by process variation (b) Higher Vdd or upsizing for compensation



(a) (b) (c)

Fig.8. Worst case scenarios for various architectures at nominal and scaled Vdds under parameter variations (a) Conventional (b) Proposed design at Nominal Vdd (c) At scaled Vdd

4.2. Proposed Architecture under Process Variation

We consider two different scenarios to show the process tolerance benefits of our architecture (Fig. 9).

Case 1: Nominal Vdd

At nominal operating voltage, the frequency of operation (1/D1) is dictated primarily by delays of the longer paths (D1 for either w_5 or w_6 or w_7). Under process parameter variations, if the delays of one or more of the longer paths increase, the outputs of those paths may lead to latching wrong values. Since, those paths contribute less towards image quality, there is only a slight degradation in PSNR. Fig. 9(b) shows the paths containing w_5 and w_6 (paths 6 and 7) which are affected by process variation and have a longer delay ($D_2 > D_1$). The rest of the paths maintain the delay target D_1 and their outputs are computed. *Therefore, it is possible to maintain same delay as original design (D_1) under process variation, without any power/area overhead*. In the worst case, none of the longer paths are computed/used. The resulting image for this case is shown in Fig.8(b). It is important to note that the critical path lengths of each of the longer paths (Fig. 4) contain at least one additional adder delay compared to the shorter paths -- we determined through simulations that parameter variations will not increase the delays of those paths (to more than D_1) contributing more towards PSNR even under large process variation (30% delay spread). Therefore, at nominal Vdd our architecture ensures that there are no delay failures in calculating the short paths.

Case 2: Scaled Vdd

With Vdd scaling (Section 3), the delays of the shorter paths increase. Fig.9(c) shows the scenario when under scaled Vdd the delays of the shorter paths are affected by process variation. Let us consider that under no process variation all the short paths (w_0-w_4) are computed at a scaled supply voltage, Vdd2. It should be noted that the operating frequency of the system is still maintained at $1/D_1$,

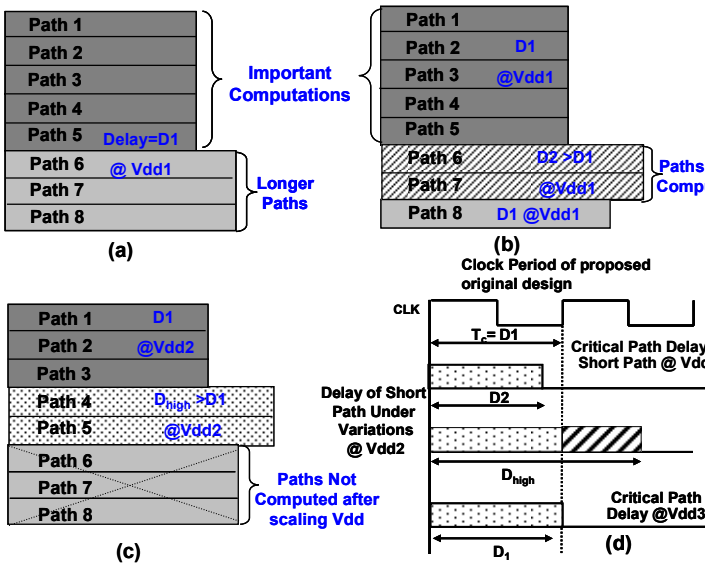


Fig. 9.(a) Proposed Design with high/low delay paths
(b) Nominal Vdd: Longer paths under process variation
(c) Scaled Vdd: Shorter paths affected by variations
(d) Delay variation and tolerance for shorter paths

the nominal Vdd frequency. Under these conditions, the delays of some paths (say for instance, w_3 and w_4) experience delay failures ($D_{high} > D_1$) due to process variation. We have two options to consider in this case:

Option 1: We continue operating the system at Vdd2 and ignore the outputs of the paths which have delay errors (delay higher than D1 at nominal supply, Vdd1) under process variation. Since the short paths are vital for determining the final image quality, the transformed image quality is affected significantly in this process. In the worst case only w_0 is computed (shortest). The resulting image is shown in Fig.8(c).

Option 2: We increase the Vdd to a new voltage Vdd3 (Fig.9(d)) to compute all the short paths and prevent any delay failures. Interestingly, this voltage Vdd3 is less than nominal voltage Vdd1. This is due to the fact that each of the longer paths (Fig. 4) contain at least one extra adder delay, which may not be offset by the delay increase through process variation. Therefore, we can operate the shorter paths at Vdd3 (Vdd2 < Vdd3 < Vdd1) to achieve the delay target D_1. In this case, we have less power savings than DCT operation at voltage Vdd2. However, the PSNR remains unchanged in this case. Comparisons of the images from Fig. 8 show that our architecture can provide process tolerance at both nominal and scaled Vdds.

4.3. Impact of our Architecture on Manufacturing Yield

The process tolerance capabilities of the proposed architecture have a positive impact on the manufacturing yield of DCT chips. As mentioned in Section 4.1, the conventional architectures fail to compute the DCT outputs because of delay failures under parameter variations. Therefore, the yield of these architectures is considerably reduced in such cases. To restore the original yield, either Vdd has to be increased or transistors need to be upsized, both of which result in extra power overhead. For our architecture, we consider two cases:

i) At nominal voltage, when one/more longer paths fail, the PSNR for such chips (chips at the worst case process corner) have a minor degradation in image quality

ii) At scaled voltages, under worst case process corners, shorter paths (more contributive towards PSNR) may fail, resulting in significant degradation in image quality. However, they can be operated at a Vdd higher than the minimum possible Vdd in order to obtain a higher PSNR.

It is evident that the proposed architecture is able to provide reasonably high manufacturing yield under process variations by making the "right" image quality/power/yield trade-off.

5. CONCLUSION

We have presented a novel DCT architecture that simultaneously satisfies low energy requirements and tolerance to large process variations, while maintaining a reasonable PSNR for image compression. This is achieved by making the high energy computational paths shorter than their low-energy counterparts. To do this, we express the approximated values of some DCT coefficients in terms of other coefficients. As we scale the voltage from nominal values (1V) to lower values (0.8V) we observed that conventional architectures are unable to compute the output image (at same frequency as nominal voltage). On the other hand, as the voltage scales down, our architecture computes the output image with a gradual degradation in image quality. Another important aspect of the proposed architecture is that it maintains high image quality even under large process variation. This is possible since the architecture predicts and tolerates any path delay failures in longer paths under process variations. We believe that the proposed design concept of computing important computations with higher priority may be applicable to other areas of signal processing where proper trade-off between power and quality of service is required.

REFERENCES

[1] S.Appadwedula et al., "Total System Minimization for wireless image transmission", VLSI Signal Processing, vol 27, 2001, pp. 99-117

[2] T.Lan et al., "Adaptive low power multimedia wireless communications", CISS 1997, pp. 377-382

[3] L. C. Yun et al., "Digital Video in a Fading Interference Wireless Environment", ICASSP 1996, pp. 1069-72

[4] C.Loeffler et al., "Practical Fast 1-D DCT Algorithm with 11 Multiplications", ICASSP 1989, pp. 988-991.

[5] Jen-Shiun Chiang et al., "A High Throughput 2-D DCT/IDCT Architecture for Real-Time Image and Video System", ICECS 2001, pp. 867-870

[6] Tarek Darwish et al., "Energy aware Distributed Arithmetic DCT Architectures", SIPS 2003, pp. 351-356.

[7] T. Kuroda et al., "A 0.9 V, 150 MHz, 10 mW, 4 mm^2, 2-DCT core processor with variable V scheme," JSSC, vol. 31, 1996, pp. 1770—1778.

[8] S.Borkar et al., "Parameter variations and impact on circuits and microarchitecture", DAC 2003, pp. 338–342.

[9] V.Bhaskaran et al., "Image and Video Compression Standard Algorithms and Architecture", Kluwer Academic Publishers, 1996.

[10] S.Kwon et al., "DCT processor architecture based on computation sharing", OCCSC 2002, pp. 162-165.

[11] J.Park et al., "Low power reconfigurable DCT design based on sharing multiplication", ICASSP 2002, pp. III-3116-3119.

[12] Predictive Technology Model, http://www.eas.asu.edu/~ptm/

[13] R.C.Gonzalez et al., "Digital Image Processing", Prentice Hall, 2002.

[14] www.synopsys.com