# Fixed-Point Error Analysis and Word Length Optimization of $8 \times 8$ IDCT Architectures

Seehyun Kim and Wonyong Sung

***Abstract*—Complete fixed-point error models that include the coefficient quantization are derived for two popular 8 × 8 two-dimensional (2-D) IDCT architectures; one is based on distributed arithmetic, and the other is the multiplier–adder chain. The error models are evaluated in the integer domain to accurately measure the effects of rounding. The analysis results show that the overall mean-square error performance (OMSE) is the most critical condition for meeting the IEEE specification (IEEE Std. 1180-1990) when the rounding scheme is employed. On the other hand, the mean error effects (OME and PME) are dominant for truncation. Finally, the analysis results are compared with those of bit-accurate simulation.**

***Index Terms*—Distributed arithmetic, fixed-point error analysis, IDCT, IEEE Standard 1180-1990, word length optimization.**

## I. Introduction

THE two-dimensional (2-D) discrete cosine transform has been widely used for various image and video processing standards, such as JPEG, H.261 for videotelephony, MPEG, and HDTV. Efficient implementation of the transform requires fixed-point arithmetic, which may result in a noticeable mismatch between the encoder and decoder. In particular, this problem can be magnified when the IDCT (inverse discrete cosine transform) is used in a reconstruction loop for motion compensation purposes because the quantization error is accumulated. To solve this problem, IEEE specifies the fixed-point performance of the $8 \times 8$ IDCT for use in visual telephony and similar applications using the IEEE Std. 1180–1990 [1]. They require that the peak error (PPE), the peak mean-square error (PMSE), the overall mean-square error (OMSE), the peak mean error (PME), and the overall mean error (OME) should not exceed certain values, and the all-zero input has to produce the all-zero output. The test bed for measuring the accuracy of a proposed IDCT is shown in Fig. 1. The "reference" IDCT output is generated by the double-precision floating-point arithmetic, while the "test" output is the result of the fixed-point arithmetic. Random integers of nine bits are used for the input. Details of the test procedure are described in [1].

There have been a few studies on the fixed-point error modeling of several fast DCT/IDCT algorithms [2], [3]. However, those models are not directly applicable to the word length optimization of actual hardware because of the following reasons. First, the previous works were conducted on the algorithm level. But the quantization effects are very much dependent on the implementation architecture. Second, the fixed-point error models are not complete. For example, those studies did not consider the quantization effects of coefficients. Finally, the IEEE Standard specifications are described in terms of rounded values instead of original unquantized error signals. This means that not only are the mean and the variance of the error important, but the distribution as well. In this paper, complete fixed-point error models are derived for two of the most popular architectures of 2-D IDCT. Then, we evaluate the integer domain fixed-point error, and determine the cost optimum word lengths conforming to the IEEE Standard specification. The analytical results are also proved by experiment with the aid of the *fixed-point optimization utility* that was developed by the authors [4].

Although a few fast 2-D IDCT algorithms have been proposed, the row–column decomposition technique is preferred for VLSI implementations due to its numerical characteristics and structural regularity. In order to reduce the number of arithmetic operations without sacrificing the regularity, the one-step decomposed Chen's algorithm [5] has been widely employed. For the matrix–vector product operator, the distributed arithmetic (DA)- and the multiplier–adder-based architectures are usually considered. Although some implementations using the systolic array have been reported recently [6], [7], most actual VLSI implementations of the 8 × 8 IDCT have been based on the DA or multiplier–adder-based architecture, as shown in the survey by Pirsch *et al.* [8].

This paper is organized as follows. A technique for analyzing the fixed-point error in the integer domain is explained in Section II. In Section III, the fixed-point error model and the optimum word lengths of a DA-based 8 × 8 2-D-IDCT architecture is discussed. Section IV presents the error model of a multiplier–adder-chain-based architecture and the optimized internal word lengths. Concluding remarks are given in Section V.

## II. Integer Domain Fixed-Point Error Analysis

The IEEE specifications are based on integer domain quantization errors that are measured after rounding the output of the fixed-point implementation as illustrated in Fig. 1. In order to analyze the fixed-point error in the integer domain, it is necessary to redefine the specifications in a stochastic manner.

Consider a general additive noise model

$$\tilde{X} = X + N \tag{1}$$

where $X, \tilde{X}$, and $N$ are all random variables representing a floating-point result, a fixed-point result, and the fixed-

Manuscript received June 27, 1996; revised October 15, 1997 and July 20, 1998. This paper was recommended by Associate Editor W. Li.

S. Kim is with the Information Technology Laboratory, LG Corporate Institute of Technology, Seoul, Korea.

W. Sung is with the School of Electrical Engineering, Seoul National University, Seoul 151-742, Korea.

Publisher Item Identifier S 1051-8215(98)09286-6.

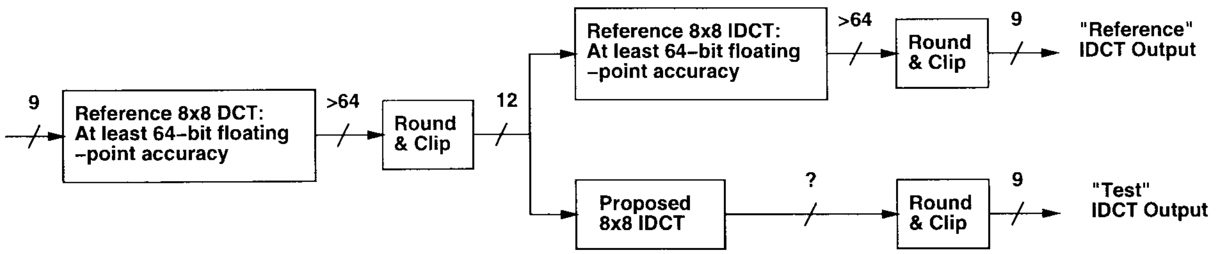

9
Reference 8x8 DCT: At least 64-bit floating -point accuracy
>64
Round & Clip
12
Reference 8x8 IDCT: At least 64-bit floating -point accuracy
>64
Round & Clip
9
"Reference" IDCT Output
Proposed 8x8 IDCT
?
Round & Clip
9
"Test" IDCT Output

Fig. 1. Testbed for measuring the accuracy of a proposed IDCT.

point error, respectively. We also assume that $X$ and $N$ are independent of each other. Then the integer domain fixed-point error $E$ can be defined as follows:

$$E = \left[\tilde{X}\right]_R - [X]_R. \tag{2}$$

Note that $[X]_R$ refers to the rounded value of $X$, i.e., the largest $n \cdot \Delta$ which is smaller than or equal to $X + (\Delta/2)$, where $n$ is an integer and $\Delta$ is the quantization step size. The probability that the integer domain fixed-point error is an integer $i$, $P_i$, can be shown to be [9]

$$P_i = \int_{-0.5}^{0.5} f_X(x) \cdot F_{N,i}(x)\, dx \tag{3}$$

where $f_X(x)$ is the probability density function (pdf) of $X$ and

$$F_{N,i}(x) \triangleq \int_{i-0.5}^{i+0.5} f_N(x - \tilde{x})\, d\tilde{x}. \tag{4}$$

Now, we can reformulate the IEEE criteria in the integer domain. For example, the OMSE and the PME criteria can be computed as follows:

$$\text{OMSE} \triangleq \frac{1}{64} \sum_{i=0}^{7} \sum_{j=0}^{7} \sum_{e=e_l}^{e_u} e^2 \cdot f_{E_{ij}}(e) \tag{5}$$

$$\text{PME} \triangleq \underset{i,j}{\text{amax}} \left\{ \sum_{e=e_l}^{e_u} e \cdot f_{E_{ij}}(e), \quad i, j = 0, \cdots, N-1 \right\} \tag{6}$$

where $E_{ij}$ represents the integer fixed-point error at pixel location $(i, j)$, and is defined as

$$E_{ij} = \left[\tilde{X}_{ij}\right]_R - [X_{ij}]_R. \tag{7}$$

Note that $e_l$ and $e_u$ are the lower and the upper bounds of the fixed-point error, respectively, and $f_{E_{ij}}(e)$ is the corresponding probability density function of $E_{ij}$. The "amax" operator selects the element whose absolute value is the maximum. All other criteria, such as PPE, PMSE, and OME, can be defined in the same fashion [10].

## III. Optimization of a DA-Based Architecture

Distributed arithmetic is one of the most popular VLSI implementation methods for computing a matrix–vector product because multiplications are not needed, and as a result, the hardware cost can be greatly reduced. An architecture for computing the transformation by employing the distributed arithmetic is shown in Fig. 2. As shown in the figure, there are three quantization error sources: coefficients for the first and the second transform $(Q_{c1}, Q_{c2})$, and the output of the limiter for the first transform $(Q_l)$, which can be assumed to be independent of each other. Note that the limiter at the output of the second transform is modeled separately by considering the probability of the integer domain output error after rounding. The word lengths of input and output signals are specified in the IEEE Standard as 12 and 9 bits, respectively.

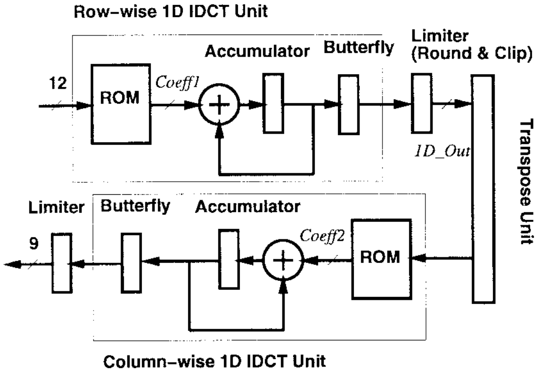


Fig. 2. Block diagram of a distributed arithmetic based 2-D IDCT.

### A. Fixed-Point Error Model

The 1-D IDCT matrix $\mathcal{D}$ can be decomposed as follows [9]:

$$\mathcal{D} = S_1 B_o C_o S_2 \tag{8}$$

where $S_1$ and $S_2$ simply shuffle the data, and $B_o$ performs a butterfly operation. $C_o$ is a block diagonal matrix, and its two 4 × 4 matrices can be obtained by decomposing the 8 × 8 IDCT coefficient matrix according to Chen's method.

In order to construct a DA hardware, the partial sum of coefficients $C_o$ should be computed in advance. It can be easily shown that the maximum value of partial sums is 2.7208. This means that at least two integer bits are needed for the representation of the coefficient ROM [4]. Since this format can represent all numbers less than 4, the upper 0.56 bit $(= \log_2(4/2.7208))$ is a waste. Thus, by scaling up coefficients as much as $\sqrt{2}$, we can reduce the waste of the integer bits in the coefficient ROM. The scaling effect can easily be compensated in the last stage by a 1 bit right shift because the overall effect of the 2-D transform is a magnification by 2. Now, let us introduce a scaled IDCT matrix $\mathcal{T}_d$ which is defined as

$$\mathcal{T}_d = S_1 B C_d S_2 \tag{9}$$

where $B = 2B_o$, $C_d = \sqrt{2}C_o$. Then, by the row–column decomposition, the 2-D IDCT matrix $Z$ becomes

$$Z = \tfrac{1}{8}\mathcal{T}_d X \mathcal{T}_d^T. \tag{10}$$

Note that the scale factor of 1/8 corresponds to just a 3-bit right shift.

By elaborating the equation for the 2-D IDCT, we can obtain the following[1, 2]:

$$\tilde{Z} = Z + \Gamma_d \tag{11}$$

where $\Gamma_d \triangleq \Gamma_1+\Gamma_2+\Gamma_3$ and $\Gamma_1 \triangleq (1/2\sqrt{2})\{\mathcal{D}(S_1B\Lambda_r)^T\}^T$, $\Gamma_2 \triangleq (1/2\sqrt{2})(\mathcal{D}\Psi^T)^T$, and $\Gamma_3 \triangleq \frac{1}{8}(S_1B\Lambda_c)^T$. Note that $\Gamma_1, \Gamma_2$, and $\Gamma_3$ describe the fixed-point errors caused by $Q_{c1}, Q_l$, and $Q_{c2}$, respectively. $\Lambda_r$ and $\Lambda_c$ denote the fixed-point error occurring in the DA hardware of the rowwise and the columnwise transforms, respectively. The $ij$th element of $\Lambda_r$, $\lambda_{ij}^r$, is defined as[3]

$$\lambda_{ij}^r \triangleq \sum_{n=0}^{N_1-1} \alpha_m^i (-1)^{\delta_n} 2^{-n+I_1} \tag{12}$$

where $\alpha_m^i$ is the quantization error of the DA coefficient, and $\delta_n$ denotes the discrete delta function, i.e., $\delta_0 = 1$ and $\delta_i = 0$ if $i \neq 0$. $N_1$ and $I_1$ represent the word length and the integer word length of the input data to the rowwise transform, respectively, and $m$ is a discrete random variable that depends on the index $j$ and the input data. Similarly, the $ij$th element of $\Lambda_c$, $\lambda_{ij}^c$, is defined as

$$\lambda_{ij}^c \triangleq \sum_{n=0}^{N_3-1} \beta_m^i (-1)^{\delta_n} 2^{-n+I_3} \tag{13}$$

where $\beta_m^i$ is the rounding noise of the DA coefficient in the columnwise transform, and $N_3$ and $I_3$ denote the word length and the integer word length of the input data, respectively. $\Psi$ represents the rounding error generated at the limiter in front of the transpose unit.

From the definitions of $\Gamma_i$'s, it can be shown that $\gamma_{ij}^1$ and $\gamma_{ij}^3$ are linear combinations of $\alpha_m^i$ and $\beta_m^i$, respectively, which are independent of each other. Also, $\gamma_{ij}^2$ is a weighted sum of $\psi_{ij}$, which are independent. Therefore, according to the well-known central limit theorem, we can approximate the probability density functions of $\gamma_{ij}^1$, $\gamma_{ij}^2$, and $\gamma_{ij}^3$ to Gaussian distributions. The means and variances of $\gamma_{ij}^k$'s will be presented in the following section.

### B. Word Length Determination Conforming to the IEEE Specifications

In order to develop the mean and the variance matrices of $\Gamma_i$'s, consider the following theorems. Let us assume that for a matrix $\Omega$ whose components are random variables $\omega_{ij}$, $\sigma^2(\Omega)$ denotes a matrix whose components are the variances of $\omega_{ij}$'s.

*Theorem 1:* Let $A$ be a constant matrix, and $\Omega$ a matrix whose components are independent random variables. Then the variance matrix of $A\Omega$ is

$$\sigma^2(A\Omega) = A^{(2)}\sigma^2(\Omega) \tag{14}$$

where $A^{(2)} = A \circ A$. $A \circ B$ is the Schur product of $A$ and $B$, where the $(i, j)$th component is defined as $(A \circ B)_{ij} \triangleq a_{ij}b_{ij}$.

*Theorem 2:* Let $A$ and $B$ be constant matrices, and $\Omega$ a matrix whose components are independent random variables. Then the variance matrix of $A\Omega B$ is

$$\sigma^2(A\Omega B) = A^{(2)}\sigma^2(\Omega)B^{(2)}. \tag{15}$$

The proofs of theorems are given in the Appendix.

By applying Theorems 1 and 2 for the definitions of $\Gamma_i$'s, the mean and the variance matrices can be obtained as follows:

$$\begin{aligned}
\mu(\Gamma_1) &= \frac{1}{2\sqrt{2}} S_1B\,\mu(\Lambda_r)\mathcal{D}^T \\
\sigma^2(\Gamma_1) &= \frac{1}{8}(S_1B)^{(2)}\sigma^2(\Lambda_r)\mathcal{D}^{(2T)} \\
\mu(\Gamma_2) &= \mathbf{0}_{8\times 8}, \quad \sigma^2(\Gamma_2) = \frac{1}{8}\sigma^2(\Psi) \\
\mu(\Gamma_3) &= \frac{1}{8}\,\mu(\Lambda_c^T)(S_1B)^T \\
\sigma^2(\Gamma_3) &= \frac{1}{64}\sigma^2(\Lambda_c^T)(S_1B)^{(2T)}.
\end{aligned} \tag{16}$$

Now, consider the distribution of $Z$, which is the restored image by floating-point arithmetic. Since the IEEE Standard specifies that the transformed image $X$ is rounded to 12 bit integers before the inverse transform, the rounded image $\hat{X}$ can be expressed as follows:

$$\hat{X} = X + \Theta \tag{17}$$

where $\Theta$ denotes the rounding error. We can assume that $\theta_{ij}$ is independent and identically distributed (i.i.d.) with zero mean and the variance of $\frac{1}{12}$. The image restored by $\hat{X}$ can be written

$$Z = Z_o + \Xi \tag{18}$$

where $Z_o$ and $\Xi$ are defined as $\mathcal{D}X\mathcal{D}^T$ and $\mathcal{D}\Theta\mathcal{D}^T$, respectively. Each element of $\Xi$ is a linear combination of $\theta_{ij}$'s. According to the central limit theorem, we can also assume that the distribution of $\xi_{ij}$ is Gaussian. Since the mean of $\theta_{ij}$ is zero, that of $\xi_{ij}$ is zero too. The variance is equal to that of $\theta_{ij}$, $\frac{1}{12}$ because the IDCT is a similarity transform, i.e., $\mathcal{D}^T\mathcal{D} = I$. Since $z_{ij}^o$ is assumed to be uniformly distributed from $-L$ to $H$, the probability density function of the floating-point result $z_{ij}$ for a given $[z_{ij}]_R$ can be modeled as a sum of shifted probability density functions of $\xi_{ij}$. For notational convenience, let $P$ and $Q$ be the random variables whose values are $z_{ij}$ and $\xi_{ij}$, respectively. Then,

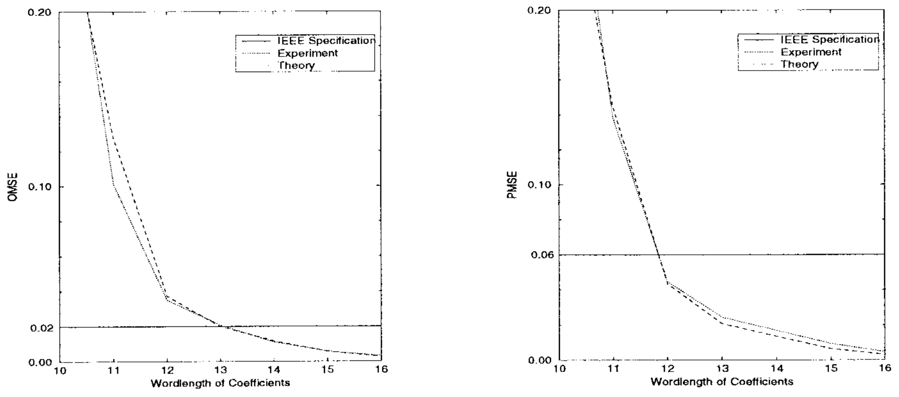
$$f_P(p) = \begin{cases} \displaystyle\sum_{i=-L}^{H} f_Q(q-i), & [P]_R - 0.5 \le p \le [P]_R + 0.5 \\ 0, & \text{elsewhere} \end{cases} \tag{19}$$

where $f_Q(q)$ is a Gaussian distribution function with $\mu = 0$ and $\sigma^2 = \frac{1}{12}$ as derived above. Now, we can evaluate the

[1] In this paper, $\tilde{a}$ represents a result by fixed-point arithmetic while $a$ indicates that of floating-point arithmetic.

[2] Greek letters denote quantization error signals.

[3] Note that capital letters denote matrices, and small characters represent their elements. For example, $\gamma_{ij}^k$ is the $ij$th element of $\Gamma_k$.



(a) (b)

Fig. 3. (a) OMSE and (b) PMSE criteria for the first IDCT block, $Coeff1$.

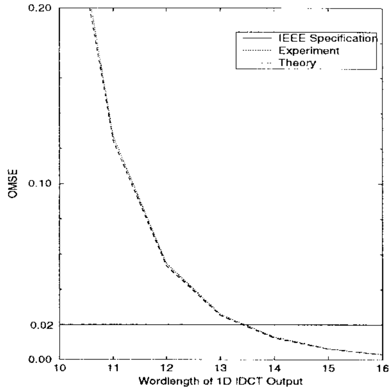


Fig. 4. OMSE criterion for the output of the limiter, $1D\_Out$.

fixed-point error performance in terms of IEEE criteria by using the results developed in Section II.

In order to carry out bit-accurate fixed-point simulation of given IDCT hardware, we take advantage of the *fixed-point optimization utility* [4]. The set of cost-optimum word lengths that requires the minimum hardware cost while satisfying the system performance can be determined by using the procedure proposed in [11].

From both analytic and simulation results, it was found that the overall mean-square error effects are dominant. The OMSE and PMSE criteria for the first coefficient are compared in Fig. 3, which shows that the OMSE condition requires at least 14 bits for the coefficients while the PMSE performance is met with 12 bits. Fig. 4 shows the OMSE criterion for the output of the limiter. The cost optimum word lengths appear in Table I. As shown in the table, analytic results are consistent with the experimental results. The numbers inside the parentheses show the word lengths of the previous implementation [12]. As for modeling the hardware cost, the cell libraries of VLSI Technologies, Inc. are used [13].

## IV. Optimization of a Multiplier- and Adder-Based Architecture

The matrix–vector product in the IDCT can be implemented in a straightforward way by using multiplier and adder chains as shown in Fig. 5. There are five quantization error sources: quantization of coefficients for the first and the second transform $(Q_{c1}, Q_{c2})$, word length reduction for the outputs of the first and the second multipliers $(Q_{a1}, Q_{a2})$, and the output of the limiter for the first transform $(Q_l)$, which are independent of each other. Note that the limiter at the output of the second transform is modeled separately by considering the probability of the output error after rounding.

TABLE I
Optimized Word Lenghts for the Distributed Arithmetic-Based Architecture

| | IWL | Min WL | | Opt WL | |
|---|---|---|---|---|---|
| | | Sim. | Anal. | Sim. | Anal. |
| Coeff1 | 2 | 14 | 14 | 15 | 15(16) |
| 1D_Out | 11 | 14 | 14 | 14 | 14(16) |
| Coeff2 | 2 | 14 | 14 | 14 | 14(16) |
| HW Cost | 11287(12378) gates | | | | |

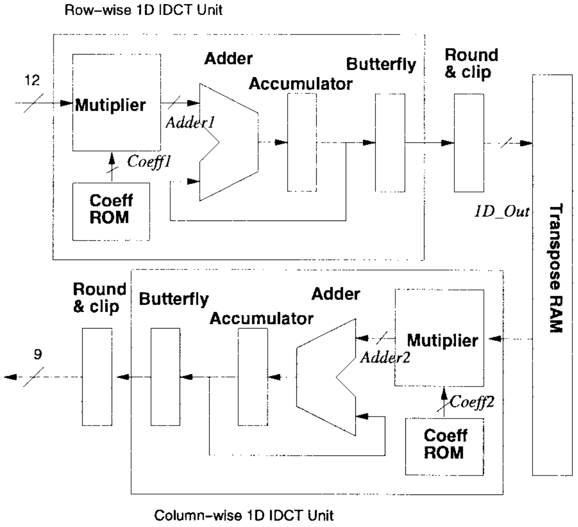
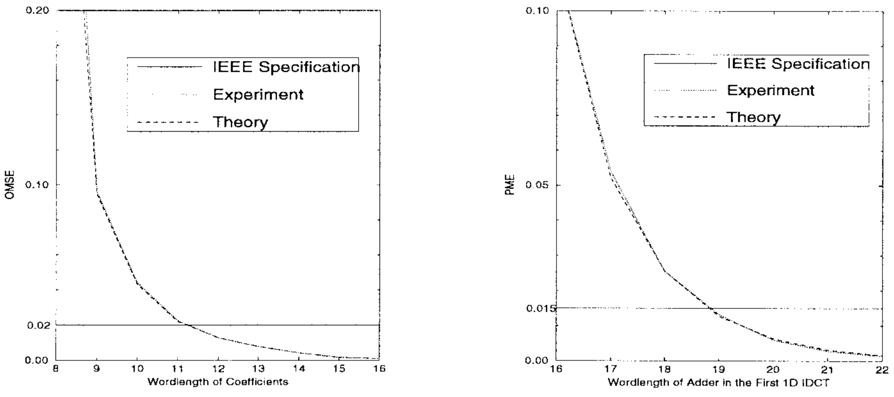


Fig. 5. Block diagram of a multiplier-added based 2-D IDCT.

### A. Fixed-Point Error Model

Let us introduce a scaled transform matrix $\mathcal{T}_m$ which is defined as

$$\mathcal{T}_m = S_1 B C_m S_2 \tag{20}$$



(a) (b)

Fig. 6. (a) OMSE criterion for the coefficients of the first 1-D IDCT unit, $Coeff1$, and (b) PME criterion for the word length of the adder in the first 1-D IDCT unit, $Adder1$.

where $B = 2B_o$, $C_m = C_o$. Then, the 2-D IDCT output $Z$ becomes

$$Z = \tfrac{1}{4}\mathcal{T}_m X \mathcal{T}_m^T. \tag{21}$$

Similarly to Section III, we can obtain the fixed-point error model by elaborating (21), the equation for the 2-D IDCT. The 2-D transformed data using fixed-point arithmetic $\tilde{Z}$ can be represented as follows:

$$\tilde{Z} = Z + \Gamma_m \tag{22}$$

where $\Gamma_m$ indicates the fixed-point error, and it can be written as $\Gamma_m \triangleq \Gamma_1 + \Gamma_2 + \Gamma_3 + \Gamma_4 + \Gamma_5$ and $\Gamma_1 \triangleq \frac{1}{2}\{\mathcal{D}(S_1 B \Theta_r S_2 X)^T\}^T$, $\Gamma_2 \triangleq \frac{1}{2}\{\mathcal{D}(S_1 B \Lambda_r)^T\}^T$, $\Gamma_3 \triangleq \frac{1}{2}(\mathcal{D}\Psi^T)^T$, $\Gamma_4 \triangleq \frac{1}{4}\{S_1 B \Theta_c S_2(\mathcal{D}X)^T\}^T$, $\Gamma_5 \triangleq \frac{1}{4}(S_1 B \Lambda_c)^T$. $\Gamma_i$, $i = 1, \cdots, 5$ represent the overall fixed-point error caused by $Q_{c1}, Q_{a1}, Q_l, Q_{c2}$, and $Q_{a2}$, respectively. $\Lambda_r$ and $\Lambda_c$ denote the truncation error matrices occurring after the multipliers of the rowwise and the columnwise transforms, respectively. The quantization errors of transform coefficients are expressed as $\Theta_r$ and $\Theta_c$, respectively. Finally, $\Phi$ represents the rounding error at the limiter in front of the transpose unit.

### *B. Word Length Determination Conforming to the IEEE Specifications*

Similarly to Section III-B, we can evaluate the integer domain error criteria. For example, the mean and the variance of $\Gamma_1$, which is the error component caused by the quantization of coefficients, can be represented as follows:

$$\sigma^2(\Gamma_1) = \tfrac{1}{4}(S_1 B \Theta_c S_2)^{(2)} \sigma^2(X) \mathcal{D}^{(2T)},$$
$$\mu(\Gamma_1) = \mathbf{0}_{8\times 8}. \tag{23}$$

It can be also shown that $\gamma_{ij}^1$'s are linear combinations of $x_{ij}$, which are independent random numbers. Thus, we can approximate the probability density functions of $\gamma_{ij}^1$, where $i, j = 0, \cdots, 7$, to Gaussian distributions with the corresponding variances and means according to the central limit theorem. Now, the integer domain error criteria can be evaluated using the probability density function of $z_{ij}$, which has been developed in Section III-B.

TABLE II
Optimized Word Lenghts for the Multiplier–Adder-Based Architecture

| | IWL | Min WL | | Opt WL | |
|---|---|---|---|---|---|
| | | Sim. | Anal. | Sim. | Anal. |
| Coeff1 | 0(1) | 12 | 12 | 13 | 13(14) |
| Adder1 | 10(15) | 19 | 19 | 19 | 19(28) |
| 1D_Out | 11(11) | 14 | 14 | 15 | 15(16) |
| Coeff2 | 0(1) | 12 | 12 | 12 | 12(14) |
| Adder2 | 11(15) | 20 | 20 | 20 | 20(33) |
| HW Cost | 19866(21374) gates | | | | |

From both analytic and experimental results, it was found that the most crucial condition for $Q_{c1}, Q_{c2}$, and $Q_l$ is the overall mean-square error OMSE. However, since the multiplier outputs are usually truncated to reduce the word length of the following adders, the means of $Q_{a1}$ and $Q_{a2}$ are not zero. And as a result, the peak mean error PME and the overall mean error OME play the key role for determining the minimum word length for *Adder1* and *Adder2*, respectively. Although we can reduce the size of the adders by inserting rounding circuits after the multipliers, it may not be a more efficient solution. The OMSE criterion for the first coefficient and the PME criterion for the adder in the 1-D IDCT unit are compared in Fig. 6. The cost optimum word lengths appear in Table II. As shown in the table, analytic results are quite consistent with the experimental results. The numbers inside the parentheses show the word lengths of the previous implementation [14].

## V. Concluding Remarks

The finite word length effects of 8 × 8 2-D IDCT algorithms were analyzed on the architectural level, and the optimum internal word lengths for the distributed arithmetic and the multiplier–adder-based architectures have been determined to satisfy the IEEE specifications while requiring the minimum hardware cost.

First, in order to analytically evaluate the IEEE specifications, which are defined in the integer domain by the ensemble sense, a simple method for analyzing the integer domain error has been presented. Also, the IEEE criteria have been reformulated in a stochastic sense. Second, the complete fixed-point error models for both the distributed arithmetic and the multiplier–adder-chain-based 8 × 8 2-D IDCT architectures were derived. Finally, the optimum set of word lengths conforming to all of the IEEE specified criteria including PPE, PMSE, OMSE, PME, and OME was determined using the analytical results. The analytical results were compared with that of the bit-accurate simulation. The hardware costs using these optimized word lengths are about 9.7 and 7.6% lower than those of the previous implementations. This study can be used for the VLSI implementation of the video rate DCT and IDCT because the distributed arithmetic and the multiplier–adder-chain-based architectures are quite regular and adequate for high throughput processing.

## Appendix

*Proof of Theorem 1* Let $\Gamma \triangleq A\Omega$; then $\gamma_{ij} = \Sigma_{k=1}^{N} a_{ik}\omega_{kj}$. The square of the first moment of $\gamma_{ij}$ is

$$E^2(\gamma_{ij}) = \sum_{k=1}^{N} a_{ik}^2 E^2(\omega_{kj}) + 2\sum \cdot E(a_{i1}\omega_{1j})^{n_1}\cdots E(a_{iN}\omega_{Nj})^{n_N}$$

and the second moment of $\gamma_{ij}$ is

$$E(\gamma_{ij}^2) = \sum_{k=1}^{N} a_{ik}^2 E(\omega_{kj}^2) + 2\sum \cdot E(a_{i1}\omega_{1j})^{n_1}\cdots E(a_{iN}\omega_{Nj})^{n_N}$$

where $n_1, \cdots, n_N$ are either 0 or 1, and $\Sigma_{i=1}^{N} n_i = 2$. Thus, the variance of $\gamma_{ij}$ becomes

$$\sigma^2(\gamma_{ij}) = \sum_{k=1}^{N} a_{ik}^2 \sigma^2(\omega_{kj}).$$

Therefore

$$\sigma^2(A\Omega) = A^{(2)}\sigma^2(\Omega).$$

■

*Proof of Theorem 2* Let $\Gamma \triangleq A\Omega$ and $\Delta \triangleq \Gamma B = A\Omega B$. According to Theorem 1

$$\sigma^2(\Delta^T) = B^{(2T)}\sigma^2(\Gamma^T) = B^{(2T)}\sigma^2(\Omega^T)A^{(2T)}.$$

Therefore

$$\sigma^2(A\Omega B) = A^{(2)}\sigma^2(\Omega)B^{(2)}.$$

■

## References

[1] CAS Standards Committee of the IEEE Circuits and Systems Society, *IEEE Standard Specifications for the Implementations of* $8 \times 8$ *Inverse Discrete Cosine Transform,* 1991.

[2] C. Y. Hsu and J. C. Yao, "Comparative performance of fast cosine transform with fixed-point roundoff error analysis, " *IEEE Trans. Signal Processing,* vol. 42, no. 5, pp. 1256–1259, May 1994.

[3] I. D. Yun and S. U. Lee, "On the fixed-point-error analysis of several fast DCT algorithms," *IEEE Trans. Circuits Syst.,* vol. 42, pp. 685–693, Nov. 1995.

[4] S. Kim, K. I. Kum, and W. Sung, "Fixed-point optimization utility for C and C++ based digital signal processing programs," in *Proc. 1995 IEEE Workshop VLSI Signal Processing,* Oct. 1995, pp. 197–206.

[5] W. H. Chen, C. H. Smith, and S. C. Fralick, "A fast computational algorithm for the discrete cosine transform," *IEEE Trans. Commun.,* vol. COM-25, pp. 1004–1009, Sept. 1977.

[6] D. Slawecki and W. Li, "DCT/IDCT processors design for high data rate image coding," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 2, pp. 135–146, June 1992.

[7] Y. T. Chang and C. L. Wang, "New systolic array implementation of the 2D discrete cosine transform and its inverse," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 5, pp. 150–157, Apr. 1995.

[8] P. Pirsch, N. Demassieux, and W. Gehrke, "VLSI architectures for video compression—A survey," *Proc. IEEE,* vol. 83, pp. 220–246, Feb. 1995.

[9] S. Kim and W. Sung, "Fixed-point error analysis and wordlength optimization of a distributed arithmetic based $8 \times 8$ 2D-IDCT architecture," in *Proc. IEEE Workshop VLSI Signal Processing,* 1996, pp. 398–407.

[10] S. Kim, "A study on the fixed-point implementation of digital signal processing algorithms," Ph.D. dissertation, Seoul National Univ., Korea, Feb. 1996.

[11] W. Sung and K. I. Kum, "Simulation-based word-length optimization method for fixed-point digital signal processing systems," *IEEE Trans. Signal Processing,* vol. 43, pp. 3087–3090, Dec. 1995.

[12] S. I. Uramoto *et al.*, "A 100-MHz 2-D discrete cosine transform core processor," *IEEE J. Solid-State Circuits,* vol. 27, pp. 492–498, Apr. 1992.

[13] VLSI Technologies, Inc., *1-Micron Cell Compiler Library,* Nov. 1991.

[14] T. Miyazaki, T. Nishitani, M. Edahiro, and I. Ono, "DCT/IDCT processor for HDTV developed with DSP silicon compiler," *J. VLSI Signal Processing,* vol. 5, pp. 39–47, 1993.