Truncation noise in fixed-point SFGs

G.A. Constantinides, P.Y.K. Cheung and W. Luk

A new model for predicting truncation error variance in fixed-point filter implementations is introduced. The proposed model is shown to be more accurate than existing models, particularly for some direct hardware implementations. In addition, some comments are made on the applicability of existing error models.

Introduction: Traditionally, much of the research into estimating the effects of truncation and roundoff noise in fixed-point systems has been focused on the implementation within, or design of, a digital signal processing (DSP) processor. This leads to certain constraints on and assumptions about quantisation errors: for example that the wordlength at all signals in the signal flow graph (SFG) is constant, and that the wordlength before a quantisation is much greater than after (for example after a multiply). When designing custom hardware to implement DSP functions, we may often be freed from these constraints. In this context, we report in this Letter some alternative truncation-noise models that have been developed as part of a high-level synthesis system for the mapping of an SFG to hardware implementation in a field programmable gate array (FPGA). To reduce as much as possible the wastage of FPGA resources, we are specifically interested in the truncation rather than rounding of results.

The effects of finite register length in fixed-point systems have been studied for some time. Oppenheim and Weinstein [1] and Liu [2] presented standard models for quantisation errors and error propagation through linear time-invariant systems, based on linearising the truncation of signals. Error signals, assumed to be uniformly distributed, white and uncorrelated, are added whenever a truncation occurs. This approximate model has served very well, since the quantisation error power is dramatically affected by the signal width in a uniform-width structure, meaning that it is only necessary to have models accurate to within 30–40% in order to predict the required signal width [3]. However, in a system realisation where different signals may have different wordlengths, it is possible to improve on these models.

Models: In our high-level synthesis environment, it is common to perform optimisations which may often lead to truncations of just one or two bits. Under these circumstances, the assumption of a continuous error distribution, and hence the variance derived from it, breaks down and we find a mismatch between theoretical and empirical error powers. For this reason, while retaining the linearisation model and the uncorrelated assumptions, we introduce a different model of error variance based on a discrete distribution. In the proposed model, each signal in the signal-flow graph is characterised by a tuple of parameters, (n, p). These indicate the wordlength (excluding the sign bit) and the position of the binary point rightwards from the sign-bit ($p = 0$ is the standard normalisation). We let a signal (n_1, p) be truncated to (n_2, p). For two's complement arithmetic the truncation error is bounded by

$$-2^p(2^{-n_2} - 2^{-n_1}) \leq e \leq 0 \qquad (1)$$

We assume that each possible value of e has equal probability. This is a fair assumption for a signal with large enough dynamic range, since it is the low-order bits that are truncated [3]. For the two's complement, there is a nonzero mean

$$m_q = -\frac{1}{2^{n_1-n_2}} \sum_{i=0}^{2^{n_1-n_2}-1} i \cdot 2^{p-n_1} = -2^{p-1}(2^{-n_2} - 2^{-n_1}) \qquad (2)$$

and the variance is given by

$$\sigma_q^2 = \frac{1}{2^{n_1-n_2}} \sum_{i=0}^{2^{n_1-n_2}-1} (i \cdot 2^{p-n_1})^2 - m_q^2 = \frac{1}{12}2^{2p}(2^{-2n_2} - 2^{-2n_1}) \qquad (3)$$

Note that for $n_1 >> n_2$ and $p = 0$, eqn. 3 simplifies to $\sigma_q^2 = \frac{1}{12}2^{-2n_2}$ which is the well-known predicted error variance in [3] for a model with continuous PDF and $n_1 >> n_2$. To determine the effect of internal truncations on error power at the outputs of an SFG, our synthesis tool applies an L_2-type scaling [4] to these variances in order to propagate them to the system outputs.

Results: In [3], the uniformly distributed model with bounds shown in eqn. 1 is presented, which yields a variance of $\sigma_q^2 = \frac{1}{12}(2^{-n_2} - 2^{-n_1})^2$ for two's complement truncation with $p = 0$ (referred to as the unsimplified model in this Letter), and which is simplified to $\sigma_q^2 = \frac{1}{12}2^{-2n_2}$ for $n_1 >> n_2$ (referred to as the simplified model). The clear distinction between the unsimplified model and our proposed model for $p = 0$, $\sigma_q^2 = \frac{1}{12}(2^{-2n_2} - 2^{-2n_1})$ should be noted. The relative (proportional) error between the two's complement model presented in this Letter and these two previous models is illustrated in Figs. 1 and 2, respectively. For n_1 close to n_2 there is significant deviation between the models.

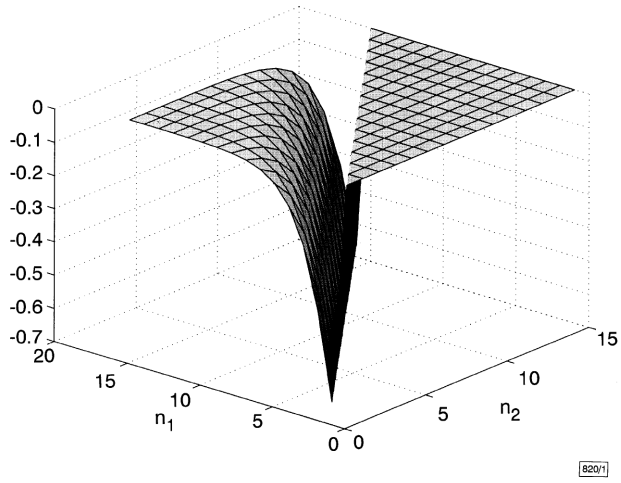

Fig. 1 *Relative error surface for unsimplified model*

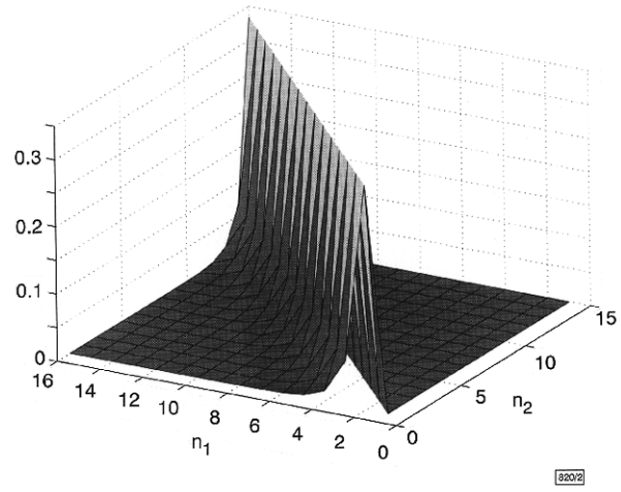

Fig. 2 *Relative error surface for simplified model*

To demonstrate the suitability of our models, we have applied them to predicting the error power in 15 real-life circuits generated by our synthesis system. Eight of these are third-order direct-form-I FIR filters, and seven are second-order direct-form-II^t IIR filters. For each circuit, 200 realisations of a 1000 sample independent, identically uniformly distributed random-process were input to a bit-true fixed-point simulator. The error power at the output was estimated by using a double precision floating point implementation as reference. Only the last 800 samples were included in the estimation, in order to avoid start-up transients. The mean error power over all realisations was evaluated. Then the absolute percentage error between the three models (the proposed model, the simplified model and the unsimplified model) and this simulated error power was obtained. The mean, minimum, and maximum percentage error of the three models over the two classes of circuit are illustrated in Table 1.

Conclusions: The conclusions of this Letter are two-fold. First, we have presented a new model for the fixed-point truncation of two's complement arithmetic based on a discrete probability distribution (we have also developed similar models for sign-and-magnitude representation). This model has been introduced to cope with implementations in hard-

Table 1: Error of two's complement models

Class	Proposed model			Simplified model			Unsimplified model		
	min	mean	max	min	mean	max	min	mean	max
	%	%	%	%	%	%	%	%	%
FIR	0.03	1.43	2.91	0.87	9.06	20.97	1.62	20.24	42.53
IIR	0.64	2.41	6.38	0.70	3.58	11.24	3.13	17.49	28.98

ware when the usual assumptions, that quantisation is only performed after multiplication, and that the wordlength before quantisation is much larger than the wordlength after quantisation, no longer hold. The model has been shown to consistently outperform more traditional truncation models in its prediction of the noise variance at the output of real circuits, allowing our synthesis tool to produce circuits that tightly meet their noise-performance specification.

Secondly, we have shown that the oft-quoted model $\sigma_q^2 = 2^{-2n_2}$ is better than just a simplification of the variance of a uniformly distributed model for the case $n_1 >> n_2$. In fact, this simplification outperforms the unsimplified model in practice when n_1 is comparable to n_2. This is an important point, the reason being that the extra variance introduced by the simplification compensates for the consistent underestimation of the original model. Indeed, it can easily be shown that the model presented in this Letter always produces a variance bounded below by the unsimplified model and above by the simplified model. This is consistent with Figs. 1 and 2, which clearly demonstrate that the simplified model is far closer to the proposed model than the unsimplified version.

16 August 1999
Electronics Letters Online No: 19991375
DOI: 10.1049/el:19991375

G.A. Constantinides and P.Y.K. Cheung (*Electrical and Electronic Engineering Department, Imperial College, London SW7 2BT, United Kingdom*)

W. Luk (*Department of Computing, Imperial College, London SW7 2BZ, United Kingdom*)

E-mail: g.constantinides@ic.ac.uk

References

1 OPPENHEIM, A.V., and WEINSTEIN, C.J.: 'Effects of finite register length in digital signal processing', *Proc. IEEE*, 1972, **60**, (8), pp. 957–976
2 LIU, B.: 'Effects of finite word length on the accuracy of digital filters - a review', *IEEE Trans.*, 1971, **CT-18**, (6), pp. 670–677
3 OPPENHEIM A.V., and SCHAFER R.W.: 'Digital signal processing' (Prentice-Hall, New Jersey, 1975)
4 JACKSON, L.B.: 'On the interaction of roundoff noise and dynamic range in digital filters', *Bell Syst. Tech. J.*, 1970, **49**, (2), pp. 159–184