Power aware accuracy-guaranteed fractional bit-widths optimization

Linsheng Zhang, Yan Zhang, Wenbiao Zhou

Department of Electronic and Information Engineering
Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
email: zhanglinsheng@gmail.com, ianzh@hit.edu.cn, zhouwenbiao@hit.edu.cn

Abstract

This paper presents a novel power aware fractional bit-widths optimization scheme during floating-point to fixed-point transformation of digital signal processing (DSP) algorithms. The scheme guarantees accuracy at output and saves power in multipliers. Quantization-Operation-Error (QOE) model is used to construct the worst case quantization error propagation. Based on QOE, a power reduction technique is proposed to dynamically reduce switching activity in multipliers when not the worst case is confronted. Results of four case studies demonstrate that 1.65% to 2.14% system power is saved with the power reduction technique, which is nearly free.

Keywords: Fixed-point, accuracy-guaranteed, low-power, digital signal processing

1. Introduction

Power consumption has become a primary design criterion for modern DSP systems. The majority of low-power design techniques and analyses in electronic designs are targeting low levels, such as transistor level, gate level and Register Transfer Level (RTL). However, the most effective power reductions often stem from system level [1].

Most algorithms with high precision in computation is wasteful and significant hardware reductions are possible. Bit-widths can be optimized to achieve desired performance and efficient implementation cost: higher speed, smaller area and lower power. The process, called floating-point to fixed-point transformation, can directly reduce power at system level.

There are mainly two kinds of methods for bit-widths optimization. One is simulation-based [2, 3, 4] and the other is analytical [5, 6]. The former methods use large simulations to search the bit-widths. The analytical methods deploy interval analysis and error models to analyze signals' ranges and precisions. Many computer arithmetic and scientific applications restrict the maximum absolute error bound at output. The simulation-based methods do not guarantee to find results within the error constraint for every input. So, analytical methods are used for this kind of accuracy-guaranteed problem.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 presents our power aware accuracy-guaranteed fractional bit-widths optimization scheme. Section 4 gives results and comparisons of four case studies. Conclusions are summarized in section 5.

Proceedings of the 11th Joint Conference on Information Sciences (2008)
Published by Atlantis Press

2. Related work

Fang *et al.* [5] use Affine Arithmetic (AA), which considers the correlations among signals, to model range and precision analyses. It serves much better than Interval Arithmetic, but signal's range and precision are solved in one single affine expression, which may limit the optimization.

Lee *et al.* [6] develop an approach called MiniBit, which also uses AA, but separates the range and precision problem apart. MiniBit guarantees output accuracy while minimizing area cost. Power is not considered in MiniBit.

Mallik *et al.* [4] propose algorithms for trading off error constraint with power and area. They employ SystemC to accelerate their simulations and use a safety factor to tighten the error constraint for more convincing results. However, accuracy at output is not guaranteed in their work.

3. Proposed scheme

3.1. Background

3.1.1. Wordlength

A fixed-point signal's wordlength (WL) is composed of integer part (IWL, including the sign bit for signed arithmetic) for preventing overflow and fraction part (FWL, or fractional bitwidth) for sustaining output accuracy.

$$WL = IWL + FWL$$

Our scheme focuses on output accuracy, which only concerns about signals' FWLs optimization. Signals' IWLs can be derived by adopting method like *Range Analysis* in [6].

After signal's FWL is determined, the signal is quantized. There are mainly two types of quantization: truncation and round to nearest, which respectively cause maximum error of 2^{-FWL} and 2^{-FWL-1} to the signal. Truncation is more implementation efficient than round to nearest. From now on, (Wx, Ix, Fx) is used to represent signal x's (WL, IWL, FWL) and truncation is considered as the default quantization type.

3.1.2. Affine Arithmetic

Affine arithmetic (AA) [7] is developed as a refinement in range analysis. It not only keeps track of signals' intervals, but also preserves correlations among them. Using AA, the quantized version x_q of signal x is:

$$x_q = x - Q_x, \quad Q_x = 2^{-Fx}\varepsilon_x \qquad (1)$$

where $\varepsilon_x \in [0, 1]$ represents the independent uncertainty of x that propagates through dataflow and contributes to the uncertainties of intermediate signals and output.

3.1.3. Power

Power consumption in a CMOS digital system consists of dynamic, short-circuit and leakage power. Short-circuit power is due to short circuit current conducting directly from the supply to ground, and leakage power is primarily determined by fabrication technology. They take very small portion of the total power consumed in a system and are rather low level issues. We are interested in higher levels of abstraction, so only dynamic power is considered.

$$Power_{dynamic} = \alpha C_L V_{DD}^2 f_{clk}$$

where α is the switching activity parameter, C_L is the load capacitance, V_{DD} is the operating voltage and f_{clk} is the clock frequency. From

high-level view, power can be saved through switching activity reduction, which minimizes the number of operations in computation.

3.2. Accuracy guaranteed FWLs optimization

3.2.1. Module definition

Figure 1 shows a module in dataflow. a, b and c respectively represent the module's infinite-accurate two inputs and one output. Δ_x represents the absolute error introduced to signal x because of quantization error bound propagations. "op" represents "$+, -$ or $\times$", which are the most commonly used fundamental arithmetics in DSP algorithms.

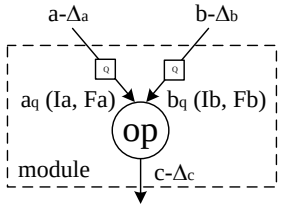

Fig. 1: A module in dataflow

In this paper, we consider the maximum absolute error Δ_{output}, which is constrained at final output, is less than or equal to 1. Through interval analysis, it is easy to have that any of inputs or intermediate signals x's absolute error Δ_x:

$$0 \leq \Delta_x \ \leq \Delta_{output} \leq 1 \qquad (2)$$

From Eq.(1), signal x's quantization error $0 \leq Q_x \leq 1$ and $Fx \geq 0$.

3.2.2. Error propagation and QOE

In Figure 1, for "$op = \pm$", $c = a \pm b$.

$$\begin{aligned} \Delta_c &= |(\Delta_a \pm \Delta_b) + (Q_a \pm Q_b)| \\ &= |(\Delta_a \pm \Delta_b) + QOE_{a\pm b}| \end{aligned} \qquad (3)$$

where $QOE_{a\pm b} = Q_a \pm Q_b$.

Quantization-Operation-Error (QOE) is proposed to formulate the error introduced to module's output because of inputs' quantization and the following operation.

For "$op = \times$", $c = ab$. From Eq.(2) and because $xy \leq (x^2 + y^2)/2 \leq (|x| + |y|)/2$, while $|x|, |y| \in [0, 1]$, we have

$$\begin{aligned} \Delta_c &= |(a - \Delta_a - Q_a)(b - \Delta_b - Q_b) - ab| \\ &= | - a\Delta_b - b\Delta_a - aQ_b - bQ_a \\ &\quad + \Delta_a\Delta_b + \Delta_aQ_b + \Delta_bQ_a + Q_aQ_b| \\ &\leq |a\Delta_b + b\Delta_a + aQ_b + bQ_a| \\ &\quad + (\Delta_a + \Delta_b)/2 + (\Delta_a + Q_b)/2 \\ &\quad + (\Delta_b + Q_a)/2 + (Q_a + Q_b)/2 \\ &\leq (|a| + 1)\Delta_b + (|b| + 1)\Delta_a \\ &\quad + (|a| + 1)Q_b + (|b| + 1)Q_a \\ &= (|a| + 1)\Delta_b + (|b| + 1)\Delta_a \\ &\quad + QOE_{a\times b} \end{aligned} \qquad (4)$$

where $QOE_{a\times b} = (|b| + 1)Q_a + (|a| + 1)Q_b$.

From Eq.(3-4), we can derive that the absolute error at algorithm's output is absolute linear summation of inputs quantization errors and all dataflow modules' QOE.

The approximations in Eq.(4) enlarge the error expression, which result a more rigorous FWLs result and larger area cost, but signals' quantization errors are separated from each other. That makes the FWLs search much easier and power reduction possible.

3.2.3. FWLs Search Algorithm

To guarantee accuracy at output, the worst case must be considered, which all the coefficients and quantization errors' maximum absolute value should be taken, like $\max|a|, \max|b|, \max Q_a$ and $\max Q_b$ are taken in Eq.(4). From

Eq.(1), $\max \Delta_{output}$ is a linear function of all signals' maximum quantization errors (2^{-FWL}).

Area model is another hot research topic. Different area models can be adopted targeting different logic synthesizers and device technologies. For example in [6], area model of $x \pm y$ is taken as $\max(Ix + Fx, Iy + Fy)$ and $x \times y$'s area is $(Ix+Fx)(Iy+Fy)$. The total area of an algorithm is all modules' area summation. After signals' IWLs are determined (Section 3.1.1), total area is a function of all signals' FWLs combination.

When error constraint *err_spec* at output is provided, our object is to find the FWLs combination which makes $\max \Delta_{output} \leq err\_spec$ while minimizing the total area.

We can build a 2-D $\max QOE$ Look Up Table (LUT) and an $AREA$ LUT with respect to (Fa, Fb) combination for every module, which respectively store module's maximum QOE (multiplied by coefficients like in Eq.(4)) and area. Starting from the minimum FWLs, we choose the most efficient module's FWLs change as every greedy search step, which decreases most error with unit area increase. It makes the results move very quickly towards *err_spec*. After *err_spec* is fulfilled, we can furthermore reduce area with least error increase, which makes use of the gap between $\max \Delta_{output}$ and *err_spec* to find a lower area solution. This greedy search algorithm will quickly find an area-efficient FWLs combination result.

3.3. Power reduction technique

Multipliers are the major sources of power consumption in typical DSP applications. Based on QOE in Section 3.2.2, we can reduce multipliers' power by decreasing their switching activities without sacrificing the required error constraint.

From Eq.(4) $QOE_{a\times b} = (|b|+1)Q_a + (|a| + 1)Q_b$, $\max|b|$ and $\max|a|$ are taken for the worst case. So, a multiplicand's quantization error is inversely proportional to the other one's value plus 1. We can dynamically relax multiplicand's quantization error when the other one's value does not achieve its peak. Take a's Q_a in module $a \times b$ as the example:

$$\begin{aligned} Q'_a &= 2^{-Fa'} \leq \frac{\max|b|+1}{|b|+1} 2^{-Fa} \\ Fa' &\geq Fa - \log_2 \frac{\max|b|}{|b|} \\ Fa' &= Fa - Fa\_t \end{aligned} \tag{5}$$

where $Fa\_t = \lfloor \log_2 \frac{\max|b|}{|b|} \rfloor$ means the truncated fractional bits from Fa, because of the variance of b's value.

One thing to be noted is that Fa' in Eq.(5) maybe negative, which means the resulted effective least significant bit can be higher than decimal point. Just like Eq.(5), the dynamically truncated fractional bits $Fb\_t = \lfloor \log_2 \frac{\max|a|}{|a|} \rfloor$. Figure 2 gives a circuit design example of Eq.(5) in signed multiplier. Unsigned multiplier is simpler.

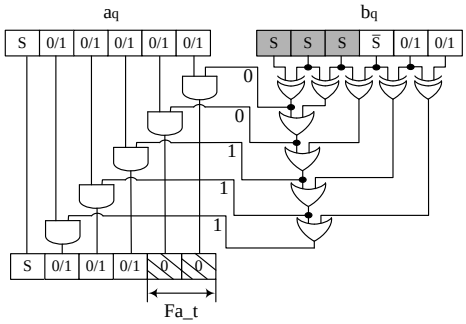

Fig. 2: A circuit design example of Fa_t in signed $a \times b$

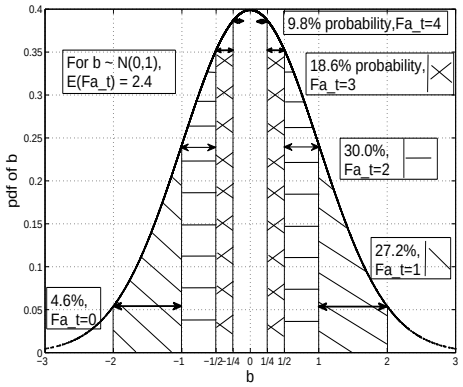

Fig. 3: Expectation of Fa_t when $b \sim N(0,1)$

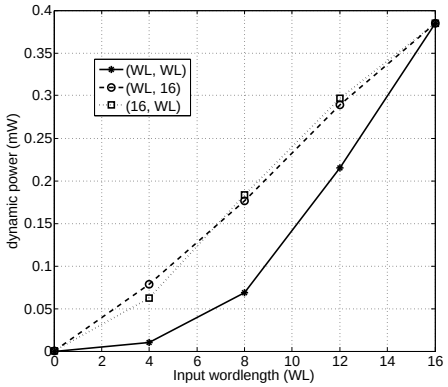

Fig. 4: Dynamic power in 16×16 Wallace multiplier (1MHz)

For uniform distributed b, all bits of b_q has $\frac{1}{2}$ possibility to be 0 (or 1). The expectation of Fa_t: $E(Fa\_t) = \frac{1}{4}\times 1 + \frac{1}{8}\times 2 + \cdots + \frac{1}{2^{Wb-1}}\times(Wb-2) \to 1$. However, it is more accurate to model signal's distribution as gaussian process [8]. Figure 3 demonstrates the calculation of $E(Fa\_t) = 2.4$ when $b \sim N(0,1)$. Generally speaking, for $b \sim N(\mu,\sigma)$, $E(Fa\_t)$ increases when $|\mu|$ increases or σ decreases.

In order to achieve the best implementation efficiency, bits of a_q connected to the circuit in Figure 2 can be b-statistics-dependent. For example, for the distribution of b having larger $E(Fa\_t)$, the bits of a_q connected to the circuit should be more.

Han *et al.* [9] have studied truncation of multiplicands in multipliers. For one of the most commonly used multipliers, which uses Wallace algorithm, the reduction in multiplicands' least significant bits via truncation mask reduces multiplier's power nearly linearly with the truncated wordlength.

Figure 4 gives the experimental results of 16×16 Wallace multiplier implemented on Xilinx Virtex-4 XC4VLX100-11 FPGA. "(WL, 16), (16, WL)" means that one of the multiplicand is truncation masked to WL, while the other one is 16-bit.

From Figure 4 we can see that, the proposed scheme can efficiently reduce multipliers' power, without sacrificing error constraint at output.

4. Case Studies and Comparisons

Four applications: degree four polynomial approximation (D4PA), RGB to YCbCr color space conversion (RGB2YCbCr), 2×2 matrix multiplication (2×2 MM) using Strassen's algorithm and 8×8 Discrete Cosine Transform (8×8 DCT), are carried out. Error constraint at output is set to 2^{-16}.

All inputs are considered as gaussian distributions. Multipliers use Wallace tree algorithm. Target device is Xilinx Virtex-4 XC4VLX100-11 FPGA and clock frequency is 1MHz. Xilinx tool "XPower" is used to estimate accurate dynamic power after design is synthesized, placed and routed by ISE9.1i. Area and power of applications, without (w/o) and with (w) the power reduction technique described in Section 3.3, are compared in Table 1.

Case studies			Comparisons of area and dynamic power					
Application	Out num	Mult num	Area [gates]			Dynamic power [uW]		
			w/o	w	Incr [%]	w/o	w	Decr [%]
D4PA	1	4	23028	23046	0.08	764.66	752.04	1.65
RGB2YCbCr	3	7	21795	21834	0.18	453.48	444.36	2.01
2×2 MM	4	7	26283	26511	0.87	875.26	856.53	2.14
8×8 DCT	8	28	54160	54642	0.89	2117.35	2073.05	2.09

Table 1: Case studies and comparisons

5. Conclusions

A novel power aware accuracy-guaranteed fractional bit-widths optimization scheme for floating-point to fixed-point transformation of DSP algorithms is presented in this paper. Quantization-Operation-Error (QOE) model is used to construct the worst case quantization error propagation. Based on QOE, a power reduction technique is proposed to dynamically reduce switching activities in multipliers, without sacrificing required accuracy at output. The power save is nearly free.

References

[1] A. Chandrakasan, S. Sheng, and R. W. Brodersen, Low-power CMOS design. *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4, Apr 1992, pp. 472–484.

[2] H. Keding, M. Willems, M. Coors, and H. Meyr, FRIDGE: A fixed-point design and simulation environment. *Design, Automation and Test in Europe Conf.*, 1998.

[3] K. I. Sum and W. Sung, Combined word-length optimization and high-level synthesis of digital signal processing systems. *IEEE Trans. on Computer Aided Design*, vol. 20, no. 8, Aug 2001, pp. 921–930.

[4] A. Mallik, D. Sinha, P. Banerjee and H. Zhou, Low-power optimization by smart bit-width allocation in a SystmC-based ASIC design environment. *IEEE Trans. on Computer Aided Design of Integrated Circuits and Systems*, Mar 2007, pp. 447–455.

[5] C. F. Fang, R. Rutenbar and T. Chen, Fast, accurate static analysis for fixed-point finite-precision effects in DSP designs. *International Conference on Computer Aided Design*, 2003, pp. 275–282.

[6] D.-U. Lee, A. A. Gaffar, R. C. C. Cheung, O. Mencer, W. Luk and G. A. Constantinides, Accuracy-Guaranteed bit-width optimization. *IEEE Trans. on Computer Aided Design of Integrated Circuits and Systems*, Oct 2006, pp. 1990–2000.

[7] L. H. de Figueiredo and J. Stolfi, Self-validated numerical methods and applications. *Brazilian Mathematics Colloquium monograph, IMPA, Rio de Janeiro, Brazil*, Jul 1997.

[8] P. E. Landman, J. M. Rabaey, Architectural power analysis: the Dual Bit Type method. *IEEE Trans. on VLSI System*, Jun 1995, pp. 173–187.

[9] K. Han, B. L. Evans and E. E. Swartzlander Jr., Data wordlength reduction for low-power signal processing software. *IEEE Workshop on Signal Processing Systems 2004*, pp. 343–348.