LOW-POWER FLOATING-POINT ENCODING FOR SIGNAL PROCESSING APPLICATIONS.

Giuseppe Visalli and Francesco Pappalardo

Advanced System Technology
ST Microelectronics
Stradale Primosole 50, 95121 Catania , Italy
Giuseppe-ast.Visalli@st.com, Francesco.Pappalardo@st.com

ABSTRACT

IEEE organization defined a standard for floating-point arithmetic, used by processing systems, in its directive 754 [1]. This directive encodes floating-point numbers using a maximum of 64 bits: 23 bit of fractional as single precision format and 52 bit of fractional as double precision format. The new multimedia terminals require low-power applications; the most important floating-point units (adders and multipliers) represent a significant part of total power wasted by a modern System-On-Chip. They might dissipate less power, using a reduced format representation. To verify this possibility, real systems simulate floating - point operations using different formats. In this conference paper, multimedia systems operate in different scenarios: wireless communication and image manipulation.

1. INTRODUCTION.

Digital applications make an intensive use of floating - point (FP) operations. They use IEEE-754 compliant arithmetic units. These units rarely meet the goal of power reduction; they are significant power consumers in a modern System -on -Chip (SoC). *Tong et. al* [2] indicates different actions to reduce power from FP units; in particular he proposed the reduction of precision/range in floating-point arithmetic, introducing a precision error. This error, as Euclidean distance from the used format to 754 standard, increases during iterative multiply -and- add operations (MAC). “Limit of the Sum” (LOS) represents the maximum number of iterative sums beyond which the precision error exceeds 50% (-3dB). Average LOS calculation performs:

$$\sum_{k=0}^{LOS-1} a_k \cdot b_k \tag{1}$$

a_k and b_k are random real numbers. Table 1 shows LOS varying mantissa width from 8 to 22 bits using a reduced format similar to 754 standard. Iterative random MAC operation allows calculating average LOS. This paper shows the performances of wireless and multimedia systems, which work with a reduced precision/range arithmetic, violating the 754 standard. These systems are:

1. Soft-Output Viterbi (SOVA) for Serially Concatenated Convolutional Decoding (SCC).
2. Jakes Fading Process filtered by Auto- regressive model.
3. Gaussian Noise filtered by Auto-Regressive model.

Mantissa Width	LOS
8	24
10	27
12	38
14	69
16	330
18	5160
20	20153
22	$\rightarrow \infty$

Table 1. Average LOS.

4. Image scaling interpolation using the bi-cubic algorithm.

A particular C++ class (*CFloat*) represents floating - point numbers with variable precision - range. The key operators are overloaded. This class can be exported in SystemC language without difficulties. You may use the IEEE double variables or CFloat variables; C++ compiler conditional directives select variable type. The paper is organized as follows: section 2 briefly introduces IEEE-754 standard for floating-point representation. Section 3 presents the performance of a communication system, which uses the Soft-Output Viterbi as inner convolutional decoder. Section 4 and 5 shows the performance of systems, which filter in the order the Jakes fading model and the Gaussian noise. Section 6 presents a typical image manipulation algorithm: the bi-cubic interpolation for image scaling. The systems described in section 4,5, and 6 works with a reduced floating-point format. The remainder of the document provides conclusion and point of discussion.

2. IEEE 754 - 1985.

Floating-point numbers mostly uses the directive IEEE-754 [1] as standard representation; this directive represents fractional (mantissa) and exponent using binary numbers. IEEE-754 standard uses the formats illustrated in Table 2. The single precision format uses 32-bits; the high precision format requires 64-bits.

-	Sign	Exp	Mantissa	Bias
Single Prec	1	8	23	127
Double Prec	1	11	52	1023

Table 2. IEEE 754 standard Layout.

In order to increase the precision, the 754 standard uses double encoding: real numbers very close to zero use de-normalized encoding, if not the standard uses the normalized encoding. The standard adds the exponent value with the *Bias*, allowing the double encoding. A normalized number has a leading one as integer part. Normalized mantissa own to the range [1,2). The littlest number in normalized encoding is $1.0 \cdot 2^{-126}$. A de-normalized number (F) uses a 23-bit field (X), operating in single precision format, as follows:

$$F = 0.X \cdot 2^{-126} \tag{2}$$

De-normalized number has zero in exponent field. Real number zero uses this encoding; 754 standard provides positive and negative zero. The directive 754 also define the rules for floating point operations, the technique for conversion to other formats (e.g. integers), the techniques for rounding in multiplication and exception rules. In this conference paper we simulate different systems operating with a reduced mantissa and exponent field. The mantissa width regulates the format's precision; the exponent field has role in the range of representable numbers.

3. THE SOFT-OUTPUT VITERBI ALGORITHM.

In 1989, Hagenauer [3] [4] introduced the Soft-Output Viterbi Algorithm (SOVA) as alternate choice to symbol-by-symbol MAP decoding. SOVA performs Viterbi decoding with reliability information computation. This system makes an intensive use of floating-point operations: add-compare-select units, update metrics, reliability estimation and related updating in the trellis. The reliability information depends by the probability that the source symbol has been incorrectly detected:

$$P_{c,k} = Pr\{\hat{u_k} \neq u_k | y_k\} \tag{3}$$

Where u_k represents the input symbol and y_k the received signal sample. Each survivor in the trellis has its reliability information. A communication system, which uses SOVA as inner convolutional decoder, works with a reduced format floating-point arithmetic. Figure 1 shows the transmitter side; an interleaving block separates the two stages of encoding in order to achieve statistical independence. Figure 2 shows the receiver side; the de-interleaver follows the SOVA inner decoder. The inner and outer convolutional codes have rate 1/2 with 64 states. Moreover, the outer decoder perform maximum likelihood detection using the reliability information as follows:

$$L_k = \log\{\frac{1 - P_{c,k}}{P_{c,k}}\} \tag{4}$$

The SOVA needs a simpler formula in order to update the reliability coefficients in the trellis; it uses the following rule:

$$L_k \Leftarrow min\{L_k, \Delta\frac{E_s}{N_0}\} \tag{5}$$

The difference between the metrics of concurrent and the survivor (Δ) allows estimating the reliability information for the considered destination state. E_s/N_0 is the signal to noise ratio (SNR). The optimal outer convolutional decoder uses the following metric:

$$\sum_k x_k \cdot \hat{u_k} \cdot L_k \tag{6}$$

Where x_k is the kth source symbol, $\hat{u_k}$ represents the kth hard decision from SOVA and L_k stand for the kth reliability information derived from (5). The channel symbols own to $Q - PSK$ constellation. Figure 3 shows the BER (Bit Error Rate) regression. At $SNR = 6dB$ the system decreases its performance by -12dB using 14-bit mantissa. The performance degradation cannot be accepted using 12-bit mantissa, where the Uncoded Q-PSK is more convenient. This suggests that a mantissa greater than fourteen could be a good trade-off between performances and low-power architecture.

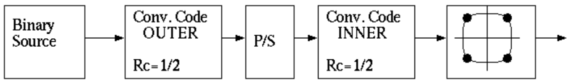

Fig. 1. Soft-Output Viterbi: TX-Side

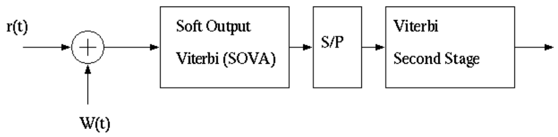

Fig. 2. Soft-Output Viterbi: RX-Side

4. TERRESTRIAL CHANNEL MODEL: THE JAKES FADING.

Terrestrial communications often use the Jakes model as fading representation [5] [6]. This model represents the multi-path interference by a sum of sinusoids, with initial phase as uniform random variable. The sinusoids have frequency related to the Doppler effect; the frequency deviation depends on the mobile speed and the carrier frequency:

$$f_D = f_c \cdot \frac{v}{c} \tag{7}$$

Where f_c is the carrier frequency, v represents the speed of mobile and c stand for the light speed. A modern communication system RF front-end filters the received signal. In this paper an auto-regressive (AR) model filters the Jakes fading process working with different precision floating - point arithmetic. AR model has four stables poles.

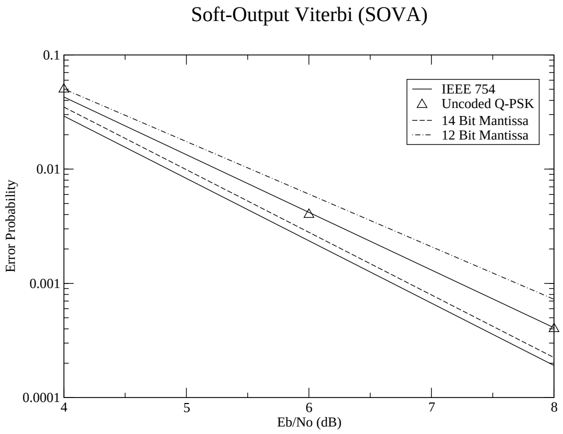

Fig. 3. Soft-Output Viterbi: BER varying mantissa width.

-	12-Bit	14-Bit	Gain
Abs Error	1.48	0.74	100%
LOS	38	69	81%

Table 3. Relation between LOS and simulation results.

We assume for simulation 90Hz as acceptable value for f_D (close to 100Km/h at 900MHz of carrier frequency) and a multi-path of eight rays. Figure 4 shows the output waveform working with 12-bit mantissa; Figure 5 shows the output waveform working with 14-bits mantissa. In both examples the relative error does not exceed the 50%. The gain in absolute error strongly agrees with the gain in precision (see Table 3).

5. IDEAL CHANNEL MODEL: THE GAUSSIAN NOISE.

The communication theory often uses, in its formal demonstrations, the Gaussian noise as ideal channel model. Similarly to Jakes fading, an AR model filters a white Gaussian noise. Because the linearity of AR filters, the random output process is still Gaussian. The statistical power of the random output process comes from an MMSE standard deviation estimator (see Fig. 6). This system works with different mantissa width, varying the precision of floating-point arithmetic. Figure 7 shows the estimated statistical power (MSE = 0.01) related to the filtered Gaussian noise. The graph shows a cut-off area placed at mantissa width less than seven bits; the output statistical power is very close to zero. Mantissa greater than twelve bits allows estimating a statistical power very close to IEEE-754 value. The transition interval, from six to ten bits, gives statistical power with a not negligible precision error.

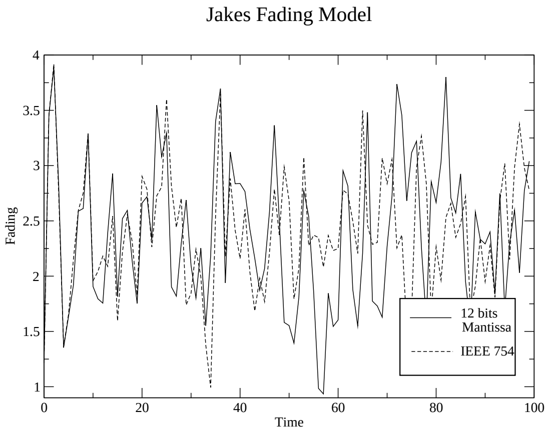

Fig. 4. Fading noise filtered (mantissa 12-bits).

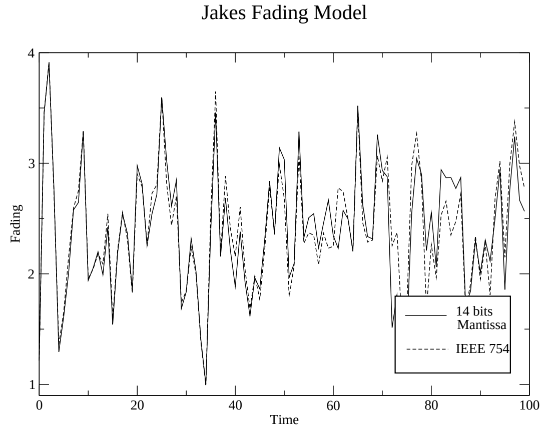

Fig. 5. Fading noise filtered (mantissa 14-bits).

6. THE BI-CUBIC ALGORITHM USED IN IMAGE SCALING.

Different from SOVA bi-cubic algorithm, as interpolation method for image scaling, uses few

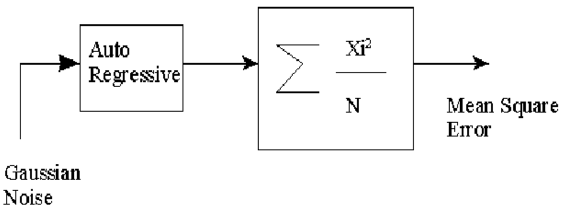

Fig. 6. Gaussian noise filtered by AR model.

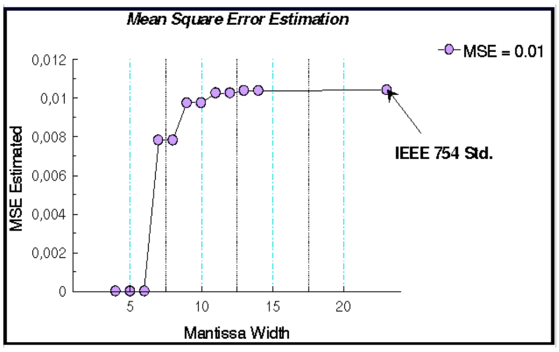

Fig. 7. Gaussian noise filtered varying precision.

floating-point resources. Modern digital cameras use bi-cubic algorithm for images format reduction; scaled image become available over a digital display. This algorithm estimates the color of destination pixel averaging sixteen colors in adjacent positions to the source pixel. Let define *srx* and *sry*, in the order, the horizontal and vertical scale ratio. Each destination point with coordinate (i,j) (see Fig.8) correspond to a non integer position in the original image (see Fig.9) given by :

$$x = i \cdot srx$$

$$y = j \cdot sry$$

The following formula gives the interpolated color (F) in the destination image:

$$F(i,j) = \textstyle\sum_{m=-1}^{2} \sum_{n=-1}^{2} F(xi+m, yi+n) \cdot$$

$$\cdot R(m-dx) \cdot R(dy-n)$$

$R(x)$ represents the *weight function*; it has a polynomial form. Bi-cubic algorithm scales images using different precision /range floating -point

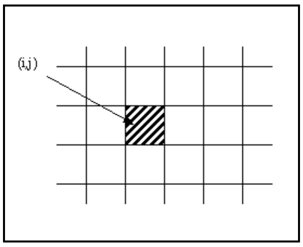

Fig. 8. Final Image : Point to estimate

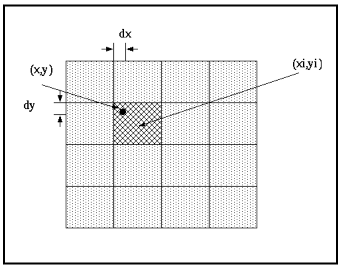

Fig. 9. Original Image : source point

arithmetic. Figure 10 shows source and destination images using 8-bit mantissa field only. Table 4 reports the percentage *mean absolute difference* (MAD), compared to the image obtained with 754 standard arithmetic, and the execution time measured using a SPARC workstation at 350 MHz.

$$MAD = \sum_{i,j} |Y_{i,j} - \bar{Y}|$$

MAD sums the absolute difference between the element luminance $Y_{i,j}$ and the average luminance $\bar{Y}$. The algorithm achieves good performance working with 8-bit precision only.

7. LOW-POWER FLOATING POINT UNITS: DESIGN OF CIRCUITS WITH REDUCED PRECISION.

The IEEE 754 standard uses the de-normalized encoding in order to increase the precision and preserve the property:

$$x = y \Leftrightarrow x - y = 0 \qquad (8)$$

As far as the low-power purpose is concerned, denormalized numbers, not necessary in wireless /

Precision	time (sec)	MAD %
8-bit	104	98.3%
10-bit	124	99.5%
12-bit	146	99.9%

Table 4. Mean Absolute Difference (MAD).

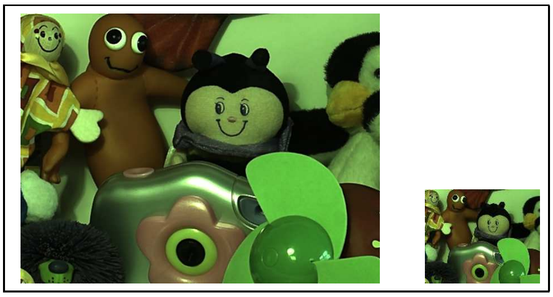

Fig. 10. Image Scaling using floating-point reduced format.

multimedia terminals, represent the weak point. The compiler can introduce instructions to detect the production of de-normal, emulating the standard IEEE. Standard 754 floating -point units perform calculation normalizing the operands. This extra-hardware represents a further contribution in total power dissipation. A new floating -point format, which uses the normalized encoding only, surely meets the goal of power reduction. Reducing the precision / range in FP arithmetic, as proposed by Tong, requires a new standard for *low-power floating -point format*. This new format has to consider the precision loss (compared to 754 standard) as mostly acceptable for low-power multimedia applications. LOS, in this case, represents an useful parameter in order to evaluate the trade-off limit.

8. CONCLUSION.

Although the floating -point arithmetic units widely use the IEEE-754 standard, they may limit the power dissipated using the reduced formats. The simulations show a threshold under that performance degradation is not negligible. This threshold changes operating in different scenarios; communication systems reach good performance working with a mantissa precision greater than 14-bits. Image manipulation algorithms often work using integer variables, so we recommend a minimum mantissa precision of 8-bits. Because area and power are relevant constraints, portable systems might as well work with a reduced precision /range arithmetic. Future studies will concern the analytical computation of average LOS, in order to find some relations between the precision and simulation results. However, the physical implementation of FP arithmetic represents the key factor in analytical average LOS calculus. The success of reduced formats in floating -point arithmetic will depend from the savings in power versus the precision loss.

9. REFERENCES

[1] “IEEE standard for binary floating-point arithmetic,” *The institute of Electrical and Electronics Engineers*, 1985.

[2] J. Tong, D.Nagle, and R.Rutenbar, “Reducing power by optimizing the necessary precision/range of floating-point arithmetic.,” *IEEE Transaction on VLSI*, vol. 8, 2000.

[3] J. Hagenauer and P. Hoer, “A Viterbi algorithm with soft-decision outputs and its applications.,” in *IEEE Global Communication (GLOBECOM)*, 1989.

[4] E. Angui S. Faudeil C. Berrou, P. Adde, “A low complexity soft-output viterbi decoder achitecture.,” in *IEEE International Conference on Communication (ICC)*, 1993.

[5] N.C. Belieu M.F. Pop, “Limitation of sum of sinusoids fading channel simulators.,” *IEEE Transaction on Communications*, vol. 49, 2001.

[6] T.Croft P. Dent, G.E. Bottomley, “Jakes fading model revisited.,” *IEEE Electronics Letter*, vol. 29, 1993.