An IEEE Compliant Floating Point MAF

R.V.K. Pillai, D. Al-Khalili[†] and A.J. Al-Khalili
Concordia University, Montreal, CANADA
[†]*Royal Military College, Kingston, CANADA*

Key words: computer-arithmetic, digital - CMOS, floating-point, low-power-design, power-consumption-model, switching-activity, vlsi

Abstract:

In this paper[1], we present a new architecture for low power floating point multiply - accumulate (MAC) fusion. The proposed architecture supports IEEE and non IEEE rounding modes. The functional partitioning of the adder segment of the MAC into three distinct, clock gated data paths allows activity reduction. The switching activity function of the adder is represented as a three state FSM. During any given operation cycle, only one of the data paths is active, during which occasion, the logic assertion status of the circuit nodes of the other data paths are maintained at their previous states. Critical path delay and latency are reduced by incorporating speculative rounding and data path simplifications. The proposed scheme offers a worst case power reduction of around 25%, in contrast to a comparable scheme reported in literature.

1. INTRODUCTION

The computation of multiply - accumulate is fundamental in many scientific and engineering applications. Since the number of computational operations envisaged by a dot product process are more than one - evaluation of a product and summation of this product with another

1.This work had been supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35498-9_57

L. M. Silveira et al. (eds.), *VLSI: Systems on a Chip*

operand - the time complexity of dot product operations are relatively high. The execution time of such operations may, however, be reduced by incorporating concurrency. With floating point operands, the fusion of MAC operations is fairly intricate, owing to the requirements for significand alignments during addition. Though the complexity of floating point hardware units that envisage fusion of multiply - accumulate operations (MAF) is relatively high in comparison with traditional approaches, MAF architectures [1] are still the preferred choice for time critical applications. The IBM RISC/6000 reported in [1] had been the first FPU with multiply - accumulate fused architecture. While the IBM MAF demonstrates the feasibility of floating point multiply - accumulate fusion, this MAF is, however, not widely accepted owing to certain limitations as far as compliance with IEEE [2] [3] floating point standards is concerned. The IBM MAF doesn't produce results that conform to IEEE standards, though the numerical accuracy of results is probably better than that of IEEE conformal schemes. W. Kahan [4] terms this "a mixed blessing". In order that the results of multiply - accumulate operation be conformal with IEEE standards, the results of multiplication and addition need separate rounding. With the IBM scheme, rounding is performed only once, which is rather a compound rounding operation encompassing both multiplication and addition. This paper addresses the development of a low power floating point multiply accumulate fused architecture which produces results that comply with the IEEE norms. The proposed architecture also supports non IEEE rounding.

2. THE PROPOSED MAF

Fig. 1 illustrates the significand data path organization of the proposed MAF. With the proposed scheme, formation of IEEE product is rather straight forward. The partial product array compresses the partial products into two sum and carry vectors. The CP Add/round block performs the carry propagate addition/rounding operation. Pre - computation for rounding is envisaged. Once the final result taking into account the rounding/normalization decisions is arrived at, the rounding information of the product can be used for the rounding of the dot product (or sum).

With the IBM MAF scheme, since the position of the significand of the product is taken as the reference, the significand of C gets aligned all the time irrespective of the value of its exponent. For those values of exponents of C that are greater than that of the product AB, the significand of C is left shifted through an appropriate number of bit positions and vice

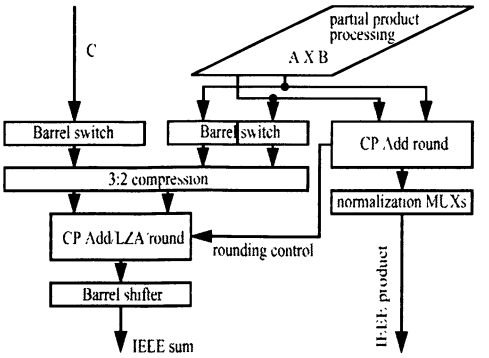

Fig. 1 - Significand data path organization of the proposed MAF

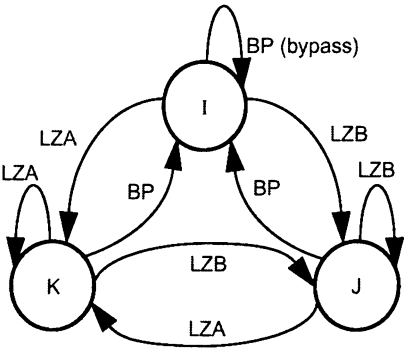

Fig. 2 - FSM representation of FADD operation

versa. With the proposed scheme, in contrast to the IBM MAF, the significands of both AB and C can be aligned in accordance with the relative magnitudes of the product and sum. During situations when the exponent of C is larger than that of AB, the output sum and carry vectors from the partial product compression circuits of the multiplier segment of the MAF are simultaneously right shifted, by using a double barrel shifter. When the exponent of AB is greater than that of C, the significand of C is right shifted by a single barrel shifter. The pre - alignment barrel shifter that shifts C can be merged with the double barrel shifter by incorporating suitable significand selection schemes. The output of the pre-alignment shifters (3 operands) is further compressed into two vectors. The significand adder accepts rounding control signals from the multiplier segment of the MAF. With the proposed scheme, support for IEEE/non IEEE rounding modes can be easily accomplished through the selection of appropriate rounding control signals from the multiplier. Pre-computation of different copies of results taking into account the various rounding/normalization requirements is envisaged. Conditional sum/carry select adders are ideal for such applications.

3. TRANSITION ACTIVITY SCALING FOR LOW POWER

In CMOS logic structures, the switching activities of functional units exhibits sensitivities towards architectural/algorithmic design decisions [5]. At the architecture level, transition activity scaling of functional units offers a viable approach for power minimization [6]. In [7], we reported the architectural design of a transition activity scaled triple data path floating point adder (TDPFADD). The approach outlined in [7] is appli-

cable for the design of floating point MAFs as well. Significant among the observations made in [7] are: (1) The leading zero estimation circuits of FADDs that handle a variable number of leading zeros need be operational only during a limited set of additions. (2) FADDs may be bypassed during certain situations. Functional partitioning of the FADD segment of the MAF into three distinct, mutually exclusive, clock gated data paths allows activity reduction. During any computing cycle, only one of the data paths is active, during which state, the logic assertion status of the circuit nodes of the other data paths are maintained at their previous states. Fig. 2 illustrates the finite state machine representation of the transition activity scaled TDPFADD [7]. State *I* represents bypass conditions. State *J* represents the operation of the FADD during those situations when the signed magnitude addition of significands can produce at the most one leading zero while state *K* represents FADD operations that can produce a pre-normalized significand with a variable number of leading zeros. The time averaged power consumption of the FADD is represented by

$$P = P(I)P_I + P(J)P_J + P(K)P_K \quad \textbf{(1)}$$

where $P(I)$, $P(J)$ and $P(K)$ represent the probability that the FADD is operating in states I, J and K respectively. P_I, P_J and P_K represent the time averaged power consumption of the FADD when the FADD is operating in the respective state. With non activity scaled FADDs, the power consumption can be as high as $P_I + P_J + P_K$.

The proposed partitioning of floating point additions also leads to data path simplifications. Since the significand pre - alignment shifts are ≤ 2 during situations when the MAC operation produces a pre - normalized significand with a variable number of leading zeros, significand pre - alignment operations of this data path (LZA data path) can be effected by using a single level of 3X1 MUXs. With p bit significands, this data path requires a normalization barrel shifter that can handle a maximum right shift of 1 and a maximum left shift of p bits. With the leading zero bounded data path (LZB data path), significand pre - alignment shifts can be anywhere between 0 and $p + 1$. The normalization shifts for this data path are bounded, a maximum right shift of 2 bit positions and a maximum left shift of 1 bit position. For both the computing data paths, only one large barrel shifter is present, by virtue of which the power consumption, logic depths and circuit delays of the data paths are minimized.

With the proposed architecture, the FADD endures bypass conditions whenever the results are known apriori. During situations when $|C| > |AB|$

and the exponent difference is greater than $p + 1$, the result is C. When $|C| < |AB|$ and the exponent difference is greater than p, the result is AB. Apart from these situations, the FADD can also be bypassed during operations of the type $0 \pm$ operand, $\pm \infty \pm$ operand and operations that leads to NaNs.

With MAFs, the multiplier may also be activity scaled for reduced power operation. Whenever the product $|AB| \ll |C|$, the multiplier segment of MAF can be activity scaled [8]. However, for applications that need IEEE sums and products, irrespective of their relative magnitudes, the question of transition activity scaling of the multiplier segment is not very relevant. During situations when the product is known apriori, e.g., 0 X number, NaN X number, $\pm \infty$ X number, the multiplier can be activity scaled. The activity scaling of the multiplier is best addressed from a control path perspective. In instruction driven processors, pre-computation of multiplier/MAF bypass conditions during an early stage of instruction scheduling is possible. With such schemes, transition activity scaling of multiplier doesn't slow down the speed performance of the MAF.

4. POWER MODELS

During FP additions, the transition activities of barrel shifter control lines exhibit sensitivities towards the significand alignment behavior of FADDs. In particular, the magnitudes and rate of change of alignment shifts reflect the power implications of significand alignment operations. The switching activities within the significand data paths also exhibit sensitivities towards the above parameters. The following paragraphs highlight the development of analytical models, that capture the impact of significand alignments on the power consumption of FADDs. Before we go into the specifics of this aspect of power consumption, the following definitions shall be introduced.

Definition 1 - Expected shift: The expected shift of any data alignment operation is defined by

$$E_{SH} = E[x] = \sum_{k=0}^{k_{max}} kP(x = k) = \sum_{k=0}^{k_{max}} kP_k \qquad \textbf{(2)}$$

In equation (2), x represents the shift distance, which is not necessarily the exponent difference; the precise relation between these parameters is a function of FADD data path/barrel shifter organization [8]. With equation (2), it is assumed that x is wide sense stationary (WSS). In floating

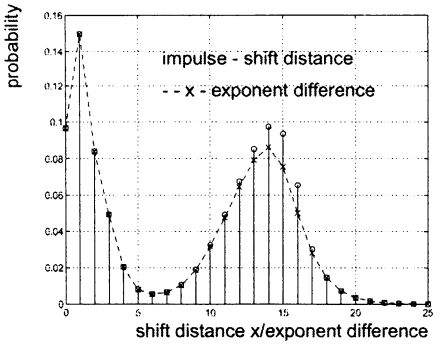

Fig. 3 - pdf of shift distances

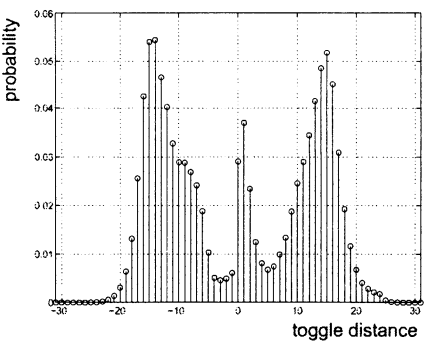

Fig. 4 - pdf of z[n] = x[n] - x[n - 1]

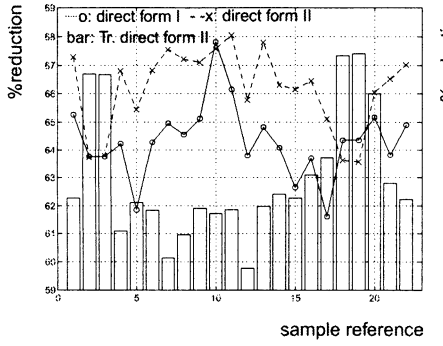

Fig. 5 - Reduction during IIR filtering

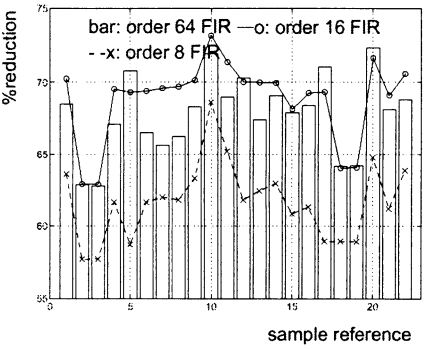

Fig. 6 - Reduction during FIR filtering

point DSP operations, strictly speaking, $x[n]$ represents a non stationary random process [9]. Though $x[n]$ is non stationary, the time averaged power consumption of FP units may still be characterized on the basis of an 'average' behavior of $x[n]$ [8]. In our simulation based study, the probability density function (pdf) of x is evaluated on the basis of the frequency distribution of $x[n]$ over the whole set of FP addition operations during the underlying experiment. Fig. 3 illustrates sample pdfs of x (significand pre alignment shift of a non activity scaled FADD) as well as the underlying exponent differences, the relevant frequency distributions of which had been observed during filtering of white noise ($N(0,1)$). A sequence of white noise samples (128K) had been low pass filtered (single precision FP operations) using an 8th order elliptical filter (transposed direct form II IIR filter - pass band ripple 0.03 dB, stop band ripple 140 dB, normalized cut-off frequency 0.2). In Fig. 3, only a truncated region of the pdf is shown, that is, the probabilities are illustrated for shift distances/exponent differences upto 25 only. The expected shift in this case is 8.5886 bits while the expected exponent difference is 14.2360 bits.

Definition 2 - Toggle distance: Toggle distance is defined as the difference

between present shift and last shift, i.e., $z[n] = x[n] - x[n-1]$.

Definition 3 - Expected toggle distance: The expected toggle distance is defined as the average number of bit positions through which shift operations oscillate about the mean or expected shift. The expected toggle distance is computed as the mean value of absolute toggle distances, as given by,

$$E[|z[n]|] = \sum_{l=0}^{l_{max}} l P(|x[n] \quad x[n \quad 1]| = l) \tag{3}$$

With conventional IEEE single precision FADDs, the index variable l can assume values between 0 and 31 (0 and 63 for double precision). With activity scaled FADDs, l is a function of the organization of barrel shifters.

With transition activity scaled FADDs, most of the parameters discussed above are scaled versions of the relevant parameters of conventional FADDs. Fig. 4 illustrates a pdf of pre-alignment toggle distances of a conventional FADD, that had been experimentally observed during filtering of random noise samples, described earlier. The expected toggle distance for this case is 10.4004 bits.

4.1 CONTROL PATH SWITCHING

In general, the control path power consumption of FADDs is dominated by the power consumption of barrel shifter control lines (both pre-alignment and normalization) as well as various data selection signals that facilitate the presentation of exponents, default results etc. With this, the control path power measure of FADDs can be modeled by

$$P_C \propto \sum_{i \in S} t_i Y_i \tag{4}$$

where S represents the set of all control signals, t_i and Y_i represent the transition activity and fanout of the ith control signal. With different architectural schemes, the number of control signals, their transition activities as well as fanouts differ.

4.2 ALIGNMENT DRIVEN DATA PATH SWITCHING

Whenever the position of the aligned significand endures oscillations about the expected shift, the significand data path bits that fall within the

toggle range endure higher transitions. The power consumption due to this phenomenon is proportional to the expected toggle distance. During FP subtractions, the 1's (or 2's) complement of the aligned significand is added with the significand of the larger number. During such a scenario, the zeros appearing at the higher order bit positions (due to shift operation) of the aligned significand gets complemented into 1's. If the toggling between addition and subtraction operations is relatively significant, then the power consumption due to this activity is significant. The power consumption of significand adders of FADDs (which, by and large, reflects the alignment driven data path switching), taking into account the above effects can be represented by

$$P_{SI} = 2\hat{P}_{ADD}\left[1 + \frac{E[|z[n]|] + E[x]t_S}{p}\right] \quad \textbf{(5)}$$

In (5), $E[x]$ and $E[|z[n]|]$ represent the expected shift and expected toggle distance respectively while p represents the width of significand. t_s represents the probability for sign toggling. $\hat{P}_{ADD}$ represents the time averaged power consumptions of significand adders, during situations when both $E[|z[n]|]$ and $E[x]t_s$ are zeros. With FADDs that incorporate leading zero anticipatory logic, the power consumption of these units are comparable to that of significand adders. This aspect is taken into account by the scaling factor 2 in equation (5). With the proposed MAF, the power consumption of the FADD segment can be represented by

$$P_{SII} = 2\hat{P}_{ADD1}[1 + \mu_A]P(K) + \hat{P}_{ADD1}[1 + \mu_B + \gamma_B]P(J) \quad \textbf{(6)}$$

In the above equation, μ_A and μ_B represent the values of the parameter $E[|z[n]|]/p$ for the LZA and LZB data paths respectively of TDPFADD. γ_B represents the value of $E[x]t_s/p$ for the LZB data path. Since the LZA data path handles only subtractions, the question of alignment driven sign toggling doesn't arise in this case. With the IBM MAF scheme, the extra switching activity due to pre-alignment toggling largely affects the p MSB bit positions of the adder. With the lower order bit positions, the switching activity is, by and large, a function of the signal probabilities of the compressed partial products. With the proposed MAF scheme, the effect of pre-alignment toggle distances of the compressed partial products (sum and carry vectors) outweigh that of the significand of C, due to similar reasons.

TABLE I : DATA PATH UTILIZATION PROBABILITIES DURING IIR FILTERING

Sl. No	IIR1			IIR2			IIR3		
	P(I)	P(J)	P(K)	P(I)	P(J)	P(K)	P(I)	P(J)	P(K)
1	0.0589	0.5380	0.4031	0.0595	0.6769	0.2636	0.0588	0.5610	0.3802
2	0.0589	0.5944	0.3467	0.0588	0.5910	0.3502	0.0600	0.6671	0.2729
3	0.0588	0.5667	0.3745	0.0588	0.5666	0.3746	0.0598	0.6749	0.2653

TABLE II : DATA PATH UTILIZATION PROBABILITIES DURING FIR FILTERING

Sl. No	FIR1			FIR2			FIR3		
	P(I)	P(J)	P(K)	P(I)	P(J)	P(K)	P(I)	P(J)	P(K)
1	0.2002	0.6433	0.1565	0.1765	0.6390	0.1845	0.1112	0.6963	0.1925
2	0.2002	0.6687	0.1311	0.1765	0.7396	0.0839	0.1111	0.8364	0.0525
3	0.2003	0.6756	0.1241	0.1765	0.6937	0.1298	0.1112	0.7651	0.1237

5. RESULTS

Instrumented digital filter programs that envisage single precision FP operations, emulating the two MAF schemes had been developed. The experiments involved the filtering of an assorted collection of data samples - both synthetic and real data. The first among the synthetic signals is a sequence of white noise samples (N(0,1) IID RVs) of sample size 128K, while the second and third are auto regressive signals of the same sample size. Specifically, the AR model of the second signal is $y[n] = x[n] + 0.9*y[n - 1]$ while that of the third signal is $y[n] = x[n] + 0.5*y[n - 1]$. The first three filters are 8th order elliptical filters (low pass), having pass band ripple of 0.03 dB, stop band ripple of -100dB and normalized cut - off frequency 0.2. Filters I, II and III are direct form I, direct form II and transposed direct form II realizations of the same filter. The last three are low pass (normalized cut-off frequency of 0.2) FIR filters of order 64, 16 and 8 respectively. With real data, an assorted collection of bipolar audio signal samples ranging in size between 8594 and 6318040 samples had been low pass filtered using both FIR and IIR filters. During the course of filtering, frequency distributions of pre-alignment and normalization shifts, their rate of change and the relevant bit level activities had been collected.

Tables I and II present the data path utilization statistics of the FADD segment of the MAF, that had been observed during filtering of synthetic data. With the above results, the most important observation is that the probability that a variable number of leading zeros occur during the

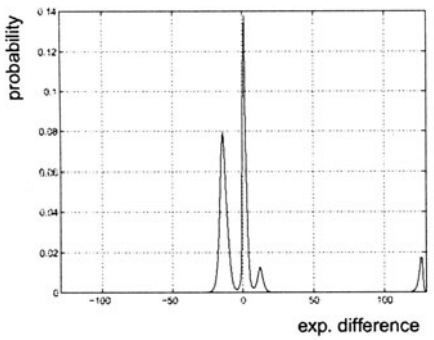

Fig. 7 - pdf of exp. diff. in IBM MAF

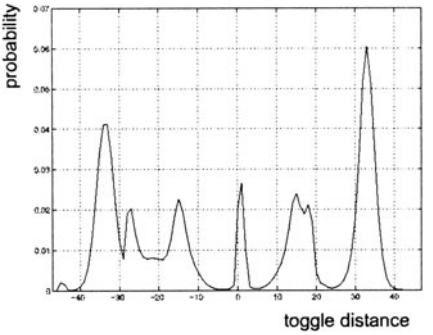

Fig. 8 - pdf of temporal shift behavior

signed magnitude addition of aligned significands is marginally low, which substantiates the efficacy of the leading zero estimation based transition activity scaling approach.

Figures 5 and 6 illustrate the percentage reduction in switching activity offered by the proposed scheme during filtering of bipolar audio signal samples, as far as significand pre-alignment control is concerned. The worst case reduction is better than 56%, which is attributed to data path simplifications and transition activity scaling. The reduction in switching activity as far as significand addition is concerned is also better than 50%. During filtering, the probability that the product is shifted is around 50%. That means, the double data path barrel shifter is not operational during 50% of the time. With the IBM MAF, the pre-alignment of the significand of C is handled by a non - activity scaled, bidirectional barrel shifter, the effective data path width of which is around twice that of the proposed scheme. Because of this, the fanout weighted switching activity of the barrel shifter control lines of the IBM scheme is significantly higher than that of the proposed scheme. Figures 7 and 8 illustrate the significand pre - alignment behavior as well as the rate of change of shift (evaluated as present shift - last shift) of the IBM MAF, that had been observed during IIR filtering of white noise samples. (The pdfs illustrated in Figures 3, 4, 7 and 8 represent various instances of exponent behavior observed during performance of the same experiment). In Fig. 7, negative values of shifts indicate situations during which the significand of C is left shifted. In contrast to the significand pre-alignment behavior depicted by Figures 3 and 4, the variances as well as entropies of the pdfs shown in Figures 7 and 8 are large. In general, the higher the variance of these pdfs, the higher the power consumption.

With normalization control, the switching activity reduction observed during our experiments is consistently better than 10X. In FADDs, nor-

malization shifts through a large number of bit positions are required only during situations when the process of significand addition results in a large number of leading zeros. During all other situations, normalization shifts are limited. However, with the IBM MAF scheme, normalization shifts can be large even during other situations. With this scheme, with p bit significands, the leading 1 after significand addition can occur within a range of $2p + 2$ bits, and hence the normalization shifts are usually large. Because of this, the leading zero estimation logic also has to work with the $2p + 2$ bit results.

In general, the power consumption of FADDs outweigh that of FP multipliers. As discussed previously, owing to the relatively large magnitudes of switched capacitances associated with significand alignments, the power consumption of barrel shifters dominate the power consumption of FADDs. Assuming that the power consumption of the multiplier segment of the MAF is comparable that of the adder segment, it is relatively straight forward to conclude that the worst case power advantage offered by the proposed scheme is around 25%.

6. DISCUSSION

Compared to the IBM scheme, the salient features of the proposed MAF scheme that renders it an ideal choice for DSP applications as well as general purpose computing are summarized below.

(1) IEEE compatibility: The requirement for IEEE compatible floating point results is mandatory for many computing applications. The availability of IEEE product and sum is a definite advantage.

(2) Data path simplifications: With the proposed scheme, the width of the significand data path is around half of that of the IBM scheme. The removal of one barrel shifter from the critical path of the significand adder is another notable feature. Because of these simplifications, the estimated speed performance of the proposed scheme is better than that the IBM scheme. The data path simplifications also results in area reduction. The area measures of significand adders, normalization barrel shifter and leading zero anticipatory logic of the proposed MAF are less than that of the IBM scheme. The proposed scheme also envisage the handling of certain arithmetic operations by using 1's complement arithmetic units [8], which results in power/area reductions. To put it briefly, though the proposed scheme envisage a separate data path for the handling of significand additions that are likely to result in a variable number of leading zeros, the additional area implications of this data path is offset by the area reduction measures.

(3) Transition activity scaling: The transition activity scaled data path partition renders power optimal operation.

7. CONCLUSION

A proposal for floating point multiply - accumulate fusion is presented. The proposed scheme delivers IEEE compatible (as well as non IEEE) sums and products. The estimated worst case reduction in switching activity offered by the proposed scheme is around 25%. The power/ delay advantages of the proposed scheme renders it an ideal choice for floating point dot product computations.

8. REFERENCES

[1] Erdem Hokenek, Robert K. Montoye and Peter W. Cook, "Second generation RISC floating point with multiply - add fused", *IEEE Journal of Solid State Circuits*, Vol. 15, pp. 1207 - 1213, October 1990.

[2] IEEE Standard for Binary Floating - Point Arithmetic", ANSI/IEEE Std 754 - 1985, New York, *The Institute of Electrical and Electronics Engineers Inc.*, August 12, 1985.

[3] Israel Koren, "*Computer Arithmetic Algorithms*", Prentice Hall, Englewood Cliffs, 1993.

[4] W. Kahan, "Lecture Notes on the Status of IEEE Standard 754 for Binary Floating-Point Arithmetic", (http://http.cs.berkeley.edu/~wkahan/ieee754status/ieee754.ps) Elect. Eng. & Computer Science, University of California, Berkeley.

[5] Kurt Keutzer and Peter Vanbekbergen, "The Impact of CAD on the Design of Low Power Digital Circuits," in *Proceedings of the 1994 IEEE Symposium on Low Power Electronics*, pp. 42 - 45.

[6] Anantha P. Chandrakasan, Randy Allmon, Anthony Stratakos, and Robert W. Brodersen, "Design of Portable Systems," in *Proceedings of the 1994 IEEE Custom Integrated Circuits Conference*, pp. 259 - 266.

[7] R. V. K. Pillai, D. Al - Khalili and A. J. Al - Khalili, "A Low Power Approach to Floating Point Adder Design", *in Proceedings of the 1997 International Conference on Computer Design*, pp. 178 - 186.

[8] R. V. K. Pillai, "On Low Power Floating Point DSP Data Path Architectures", Ph. D Thesis (under preparation) - Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec.

[9] R. W. Hamming, "On the Distribution of Numbers", *Bell System Technical Journal*, Vol. 49, No. 8, pp. 1609 - 1625, October 1970.