M.Tech. Credit Seminar Report,
Electronics Systems Group, EE Dept., IIT Bombay
Submitted November 2004

High resolution floating point ADC

S.Nandakumar (04307416)
Supervisor: Prof. T.S.Rathore

Abstract

Floating-point analog-to-digital converters (FP-ADC) have been shown useful in applications where, high resolution for small signals and wide dynamic range is desired. FP-ADCs employ a non-uniform quantization and typically dissipate less power and require less hardware than comparable uniform quantizers for a given dynamic range and resolution. A FP-ADC is discussed for the purpose of achieving a wide dynamic range without demanding high resolution, when the high resolution is merely for covering the signal dynamic range rather than the quantization accuracy. Comparing floating-point quantization with uniform quantization, it is shown that the FP-ADC is superior from a statistical point of view to the ADC, but at the prize of doubling the conversion time. This report discusses sequential FP-ADC, and its key circuitry, achievable performance, conversion time and sensitivities to mismatches. Also, this report discusses, parallel FP-ADC, to realize a reduction in conversion time, while preserving the highest possible resolution.

1 Introduction

The input range of an ADC must be designed in such a way that the peak input signal is reliably covered. It means that the ADC must have a very high resolution, when the signal amplitude fluctuates in a very large dynamic range and the quantization accuracy is needed even for the smallest signal. In many cases, high resolution is designed for merely covering the signal dynamic range rather than the quantization accuracy [1]. Actually, in such a converter, the resolution for large signals is unnecessarily high like in radars and communication applications. It would be more rational, if the ADC offers more or less the same resolution for large and small signals within the input range.

A large dynamic range can be achieved by increasing the resolution of an A/D converter. This is however not so easy in CMOS. Very accurate S/H circuits, amplifiers and passive components are needed which hardly can be accomplished in a CMOS process, especially when speed, power and area are considered. In such a solution, a small input signal uses only a small portion of the A/D converter, which is sensitive to noise. A better solution would be to amplify the small and the large signals differently so that they can be converted to digital when they are more or less at the same level, analogous to an automatic gain control

(AGC) amplifier. In such a way, the resolution does not have to be increased. Instead, it can be kept constant for different signals [1]. The effective bits float in an additional range, similar to a floating-point number representation, which explains why it is called a floating-point A/D converter.

Traditionally, the problem of extending the dynamic range is done with nonlinear compression or by using automatic gain control (AGC) amplifiers. But AGC system can not handle fast fluctuating signals. Instead, a logarithmic amplifier has to be used for compressing the input signal dynamic range. The accuracy, however, will be seriously degraded for a large compression ratio. In order to produce a linear digital output, a look-up table must be used which has to match with the amplifier characteristics precisely. Unlike the solution of logarithmic amplifier, an FP-ADC directly gives a linear digital output. For large and small signals, the effective resolutions are kept constant, similar to floating-point number representation. The virtual input range of such an ADC is much larger than the actual one so it can handle a small input range imposed by a low voltage supply.

The high resolution is achieved in flash converters by increasing number comparators, which leads to increase in size and cost. By choosing two-step ADC [5], where one for coarse quantization and another one for fine quantization, resolution is increased with less number of comparators with penalty of double conversion time. Floating point quantization implies finding both the exponent and mantissa of a signal sample. Other than logarithmic amplifier approach, an approach based on a uniform quantization ADC connected to the acquired signal through a programmable gain amplifier (PGA) is presented in sec.2.1. Firstly, the ADC performs a conversion cycle to find out the exponent and then, setting the PGA gain correspondingly, it executes a second conversion to determine the mantissa. This converter type is called sequential FP-ADC or two-cycle FP-ADC [2]. This solution preserves high precision of uniform ADCs and avoids separate coding stages at the expense of doubling the conversion time. To speed up the conversion process, architecture with two ADCs (one for coarse quantization and the other for fine one) is presented in the section 2.2. Still having the two quantizers connected in cascade, the two quantization cycles are disjunctive and speed improvement is mainly obtained from using a variant of the flash ADC for the exponent acquisition [1]. In this solution, the key is to realize a preamplifier which produces multiple output signals with weighted gains, instead of first finding the signal domain (exponent) and

amplifying like in single ADC sequential FP-ADC.

To overcome the conversion time and amplifier mismatches problem, a parallel ADC [2] has been discussed in sec.3. It is based on uniform quantizers that preserve the precision of the sequential FP-ADC, while minimizing the conversion time close to the one of the non-uniform one-cycle quantizer. The parallel FP-ADC consists of two ADCs that work simultaneously: one determines the exponent, while the other one, connected to the quantizer input over a programmable gain amplifier, finds the mantissa. If two time adjacent samples have the same exponent, the conversion result is delivered immediately; if not, the mantissa is acquired again with PGA gain reset to the most recently acquired exponent.

2 Sequential Floating Point ADC

2.1 Single ADC Sequential FP-ADC

The sequential FP-ADC [2] is schematically presented in Fig. 1. It consists of an ADC with uniform quantization, a programmable gain amplifier (PGA) and a gain control unit (GCU). This sequential converter uses same ADC for mantissa and exponent extraction. The ADC is an ideal quantizer with rounding quantization and a quantization step; its input range is $(-V_{FS}, V_{FS})$. The mantissa (y_m) is expressed using m bits, and the exponent (e) is expressed as an E-bit number. The sequential FP-ADC [2] performs a quantization in two cycles as described in the *Statechart* diagram of Fig. 2. To begin with, in the *AcqExponent* state, GCU sets the PGA gain to the minimum (PGA gain: = 1), letting the input voltage (x) to be presented as it is ($x'=x$) to the ADC input.

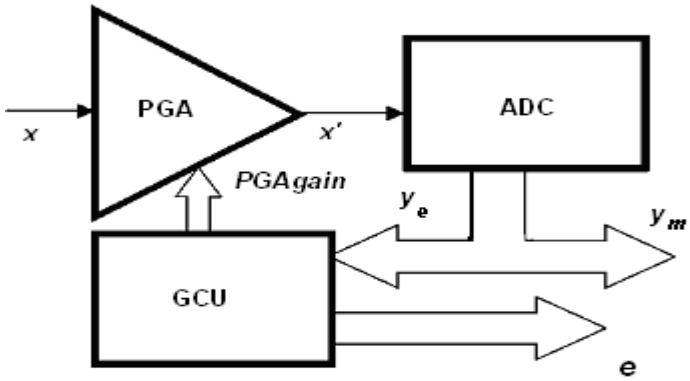

Fig.1. Block diagram of a sequential FP-ADC. [2]

The converter performs the first conversion and its result (y_e) is employed by the GCU to calculate the exponent (e) using the following equation

$$e = f(y_e) = \begin{cases} 0, & x \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right] \\ \log_2 \left|\frac{y_e}{\Delta}\right| + 1, & x \notin \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]. \end{cases} \quad (1)$$

To express the exponent as an *E*-bit number, the ADC has to have a resolution $m \geq 2^E$, with *m* and *E* being natural numbers. To take full advantage of the *m*-bit resolution in expressing the exponent values, powers of 2 are chosen for *m* i.e., $m=2^E$. Based on the exponent *e* calculated with equation (1), a new gain value is determined by the GCU, finalizing the second sub-state of *AcqExponent*

$$gain := 2^{\max E - e}, \text{ with } \max E = 2^E - 1. \quad (2)$$

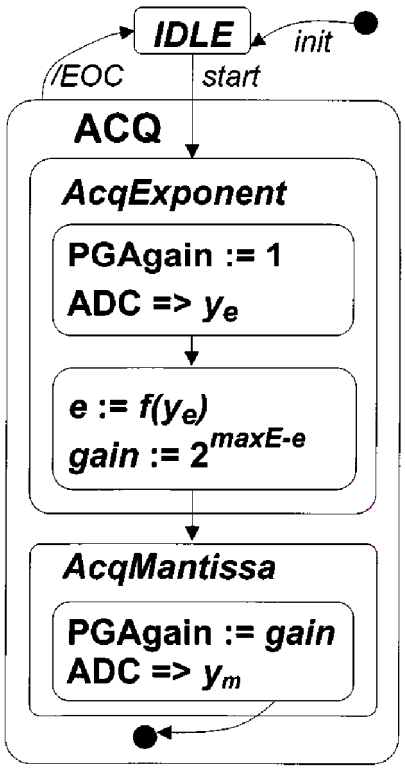

Fig.2. Statechart diagram of sequential floating-point ADC. [2]

In the next state (*AcqMantissa*, Fig. 2), the PGA is set accordingly to the *gain* value that was calculated previously in the *AcqExponent* state. The input signal (x) is thus amplified such as to maintain the PGA output signal (x' := gain.x) in the upper half of the ADC conversion range, ($V_{FS} > |x'| > V_{FS}/2$), i.e., to get the best possible use of the ADCs resolution. The ADC now performs a second conversion cycle to measure the mantissa (y_m). Based on the exponent (e) and the mantissa (y_m), the quantization function of the FP-ADC, $Q(x)$, is given by equation (3)

$$Q(x) = y_m.\left(\frac{1}{gain}\right) = y_m.2^e.\left(\frac{1}{2^{2^E-1}}\right). \tag{3}$$

The graphic representation of this quantization function (for m=4 and E=2) is given in Fig. 3, along with the correspondent exponent function e = f (x) and the quantization function of the employed uniform quantizer (b=4). To minimize the overload noise of a uniform quantizer, its input range is usually set to an interval that reduces the probability of a quantized signal falling outside it. The input range of a uniform quantizer that acquires a normally distributed signal with mean 0 and variance σ is usually set to [-4σ, 4σ] (the four sigma loading rule [4]). To use the same FP-ADC [2] model for quantizing such Gaussian signals, the input range of its quantization function from Fig. 3 is normalized to [4, 4]. The FP-ADC quantization function presents a visible improvement of precision for quantizing small signals, a progress that is achieved by reducing the length of the quantization interval for that range.

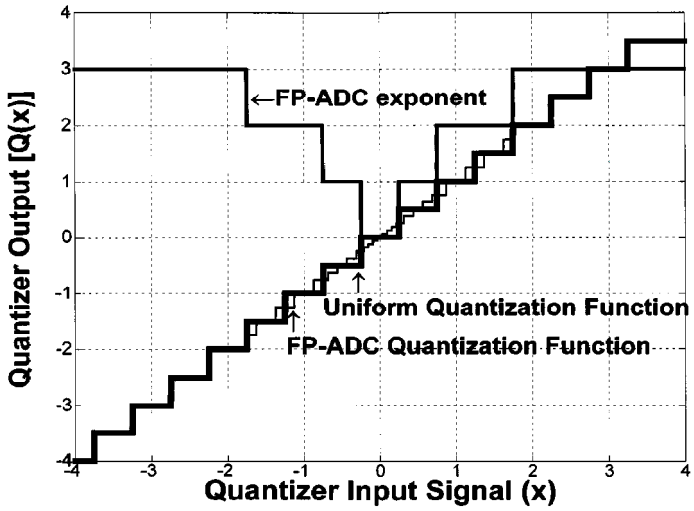

Fig.3. Quantization functions: FP-ADC (thin line) and uniform ADC (thick line). [2]

This “classic” procedure provides the floating-point result directly in its normalized form, but, even by using the fastest analog-to-digital parallel conversion technique, it doubles the conversion time.

2.2 Double ADC Sequential FP-ADC

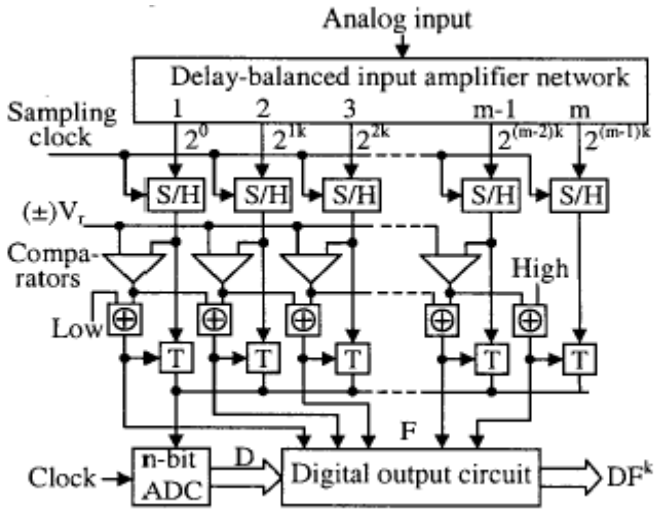

Fig.4. Double ADC Sequential FP-ADC. [1]

The double ADC sequential FP-ADC [1] shown in Fig.4., has got first ADC, which consists delay balanced input amplifier network, S/H circuits, comparators and output selector using Ex-OR gates. Second ADC is a uniform ADC which converts the analog output from balanced amplifier network according to selection of logic circuit. The delay-balanced input amplifier network amplifies the input analog signal and produces *m* amplified signals with amplifications 2(i-1).*k***,** where *k* =constant and i = l, 2**,**..., *m*, respectively. When *k* = l or 2**,** the signal amplitudes are weighted in the form of binary or quaternary. Part of the signals may be nonlinear or saturated. The largest linear signal will appear at one of the outputs. This particular output may change from one to another, depending on the input signal amplitude. The signals are sampled with identical sampling-and-hold (S/H) circuits using the same sampling clock to produce *m* voltage samples fed to corresponding switches. In order to select the largest linear signal, these samples, except for sample *m*, are compared through comparators with a reference voltage $V_r = (V_{max}/2)$ where V_{max}, is the maximum linear voltage output of the amplifiers. For a bi-directional signal, e.g. radio IF signal, $\pm V_r$ are used instead of V_r. Since no principle

difference, mainly focus is on one-directional signals below. The thermometer code produced by the comparators is further transferred to an *m*-bit flag code *F* by the XOR gates. The XOR gate located at the interface of 0- to-1 will generate logic high to turn on the switch at the same position. Note that one input of XOR gate 1 is always connected to logic low, and one input of XOR gate *m* is always connected to logic high. In such a way, when the signal is out of the designed dynamic range, switch 1 will be kept on, and when the signal is too small, switch *m* will be kept on. The *n*-bit ADC, therefore, will not miss any signal even in the two extreme cases. In the same time, the *m*-bit flag code *F* is fed to the digital output circuit. Assuming *F*=00001000 (*m*=8), the position of flag "1" indicates the location and the amplification of the selected signal. The digital output circuit combines the *n*-bit data code *D*, the *m*-bit flag code *F* and the constant *k* to generate a final output DF^k with $n+(m\text{-}1)k$ bits, i.e. the dynamic range has been expanded by $(m\text{-}1)k$ bits. If *k* is an integer, only simple shifting operations are involved. In contrast to the converting curve of a logarithmic ADC shown in Fig. 5(a), the converting curve of an FP-ADC [1] is formed by the combination of multiple linear lines, the bold part in Fig, 5(b) for *m*=4 and *k*=l.

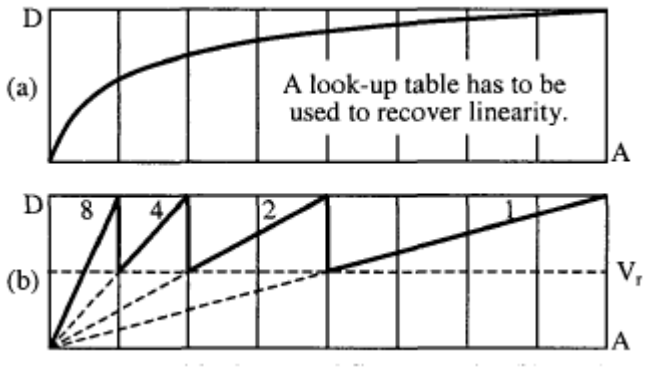

Fig.5. Logarithmic (a) and Floating point-ADC (b). [1]

Note that unlike a Single ADC sequential FP-ADC (sec-2.1), the signal in the Double ADC sequential FP-ADC is not sampled and converted until it becomes large enough so the impact of errors is reduced. It can be seen as that the input range has been expanded by a factor of *m*. This is a very useful feature for low supply voltage applications where the actual input range is increasingly smaller. The *n*-bit ADC can be any type in principle but a small input load is preferred so a pipelined ADC or a parallel successive approximation ADC [5] is more suitable here. The part that is different from a traditional ADC, particularly the input amplifier

network has been discussed. In the following discussion, with assumption of k= l, i.e. the signals are weighted in a binary form. The most important yet difficult issues are delay skew, amplification mismatch, DC offset, and amplifier over-voltage; these issues are discussed in the following sections.

2.2.1 Input Amplifier Network

Two basic issues for the input network are, to balance the delays and to weight the amplifications as accurate as possible. For simplicity, it is possible to use a binary amplifier tree [1]. To further improve the matching, a parallel structure has been presented here. The idea is to only use passive components for weighting the amplifications while keeping the amplifiers as identical as possible, see Fig. 6 for an example of m=5.

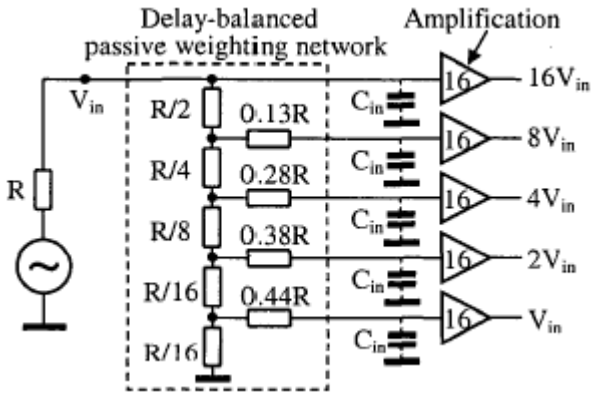

Fig.6. Input Amplifier network with m=5. [1]

A passive network is more accurate and stable than amplifiers. It is possible for a resistive divider to achieve up to a 10-bit accuracy by a careful design. By using the delay compensating resistors (horizontal ones), the delay differences are negligible [1]. The amplifier outputs are used only when they fall into the window between V_r and $2V_r$, or the amplifier input signals fall into the window between $V_r/16$ and $V_r/8$ because of the resistor divider network. It means that the noise impact is the same for all channels. The input network can be made in two steps, as shown in Fig. 7, to facilitate a larger dynamic range with limited division ratio of resistive divider to maintain accuracy.

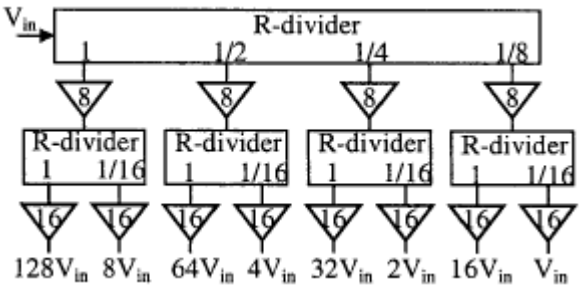

Fig.7. Input amplifier network for *m*=8. [1]

2.2.2 DC-offset and over-voltage

The amplifier DC-offsets seriously limits the achievable accuracy. For a radio IF signal, AC coupling can effectively remove the DC-offsets as presented in [3]. It is, however, not applicable for a universal ADC. Over-voltage causes another kind of problem. The amplifier saturated by an over-voltage input will seriously distort and delay the signal during its recovery period, leading to large sampling errors. A limiter can be used at the input to prevent the amplifier from over-voltage [3]. It is however difficult to eliminate the problem completely. The method introduced [1] below is different from above approaches to solve the two problems simultaneously, see Fig. 8.a.

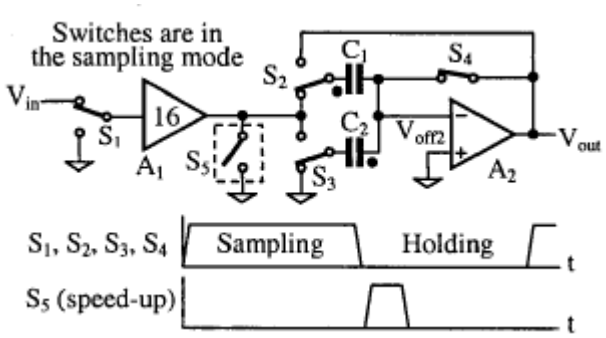

Fig.8.a. Amplifier and S/H channel in sampling mode. [1]

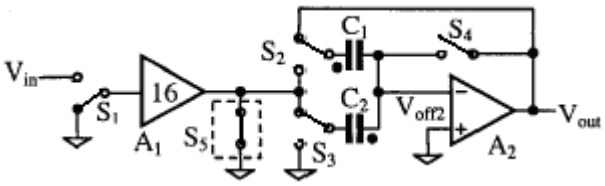

Fig.8.b. Amplifier and S/H channel in S_5 discharge. [1]

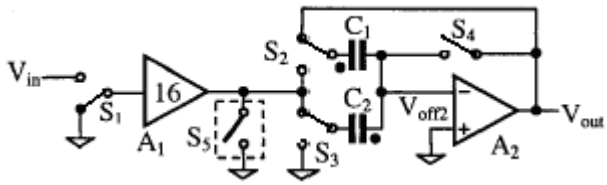

Fig.8.c. Amplifier and S/H channel in holding mode. [1]

Capacitors C_1 and C_2 are equal to implement differential charge transfer technique by switching actions.

In sampling mode:

See Fig.8.

$V_{c1}=16V_{in} + 16V_{off1} - V_{off2}$

$V_{c2}=V_{off2}$

$V_{out}=V_{off2}$

Where, V_{c1}: Voltage across capacitor C_1

V_{c2} : Voltage across capacitor C_2

V_{off1}, V_{off2} are offset voltages of amplifier A_1 and A_2 respectively

Charge in C_1 ($q1$) = C_1 ($16V_{in} + 16V_{off1} - V_{off2}$)

Charge in C_2 ($q2$) = C_2 (V_{off2})

In holding mode:

See Fig.8.c.

During holding mode all switches change their position. The voltage across capacitors C_1 and C_2 changes accordingly.

$V_{c1}= V_{out} - V_{off2,}$

$V_{c2}= V_{off2} - 16V_{off1,}$

Charge in C_2 ($q2$) = ($V_{off2} - 16V_{off1}$) C_2,

Charge discharged from C_2 = ($V_{off2} - V_{off2} + 16V_{off1}$) C_2,

Due to differential charge transfer, C_1 is discharged (due to polarity) by equal amount of charge discharge of C_2.

C_1 ($V_{out} - V_{off2}$) = ($16V_{in} + 16V_{off1} - V_{off2}$) C_1 – ($V_{off2} - V_{off2} + 16V_{off1}$) C_2,

Since, $C_1 = C_2$

$V_{out} = 16Vin.$

S_5 is an optional switch (see Fig.8.b), temporarily connecting the output of A_1 to ground in the beginning of holding phase to speed up the recovery of A_1 from saturation. During the holding phase, the input of A_2 should still be V_{off2}. As a result, the offset voltages of both A_1 and A_2 are cancelled. Note that, S_4 is the only switch sensitive to clock and charge feed-through. All other switches just connect well-defined voltages and do not affect the accuracy. Therefore, this is quite feasible [1]. Moreover, since the input of A_1 is connected to the signal ground after each sampling phase, the amplifier input voltage always starts from minimum, which effectively eliminates the over-voltage delay problem. The speed, however, will be affected by the settling time of A_1.

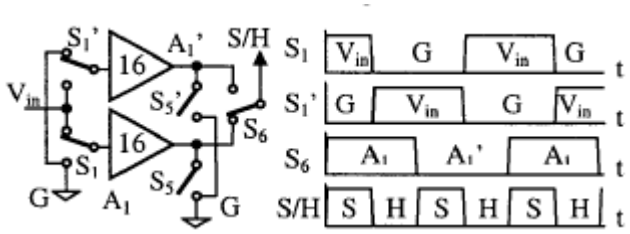

Fig.9. Two amplifiers in parallel to speed up. [1]

In order to increase speed [1], two amplifiers (A_1 and A_1') can be used for each channel, as shown in Fig. 9. A_1 and A_1' will gain more time for signal settling. When they are switched to the holding phase, S_5 or S_5' will help them to quickly come back to the signal ground (see Fig.9.for timing). The power of A_1 and A_1' should be kept low to limit the total power consumption. A low power amplifier has to be used.

2.2.3 Impacts of mismatches

The double ADC sequential FP-ADC [1], which is discussed, is insensitive to the variation of *Vr*, as long as *Vr* < (*Vmax*/2), see Fig.10. *Vmax* is the maximum input voltage of uniform ADC. The only impact is that the utilized linear region moves up or down. The switching point between two amplifiers is still well defined, see the bold lines. If the *n*-bit ADC covers the variation region, there will be very little impact. The system is also insensitive to the variation of absolute amplifications as long as they are matched. The reason is that the amplification ratio is defined through the passive components rather than the active amplifiers. The impact will be the same as the variation of *Vr*, but, instead of *Vr*, the scales of vertical lines (i.e. the absolute amplifications) change equally. As long as the linear region is still larger than *2Vr*, the impact is very little.

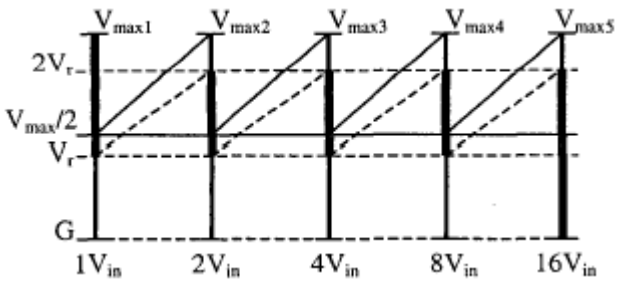

Fig.10. Impact of reference voltage variation. [1]

The impact of amplification discrepancy on the linearity is shown in Fig. 11, where the amplification of channel 2 is lower.

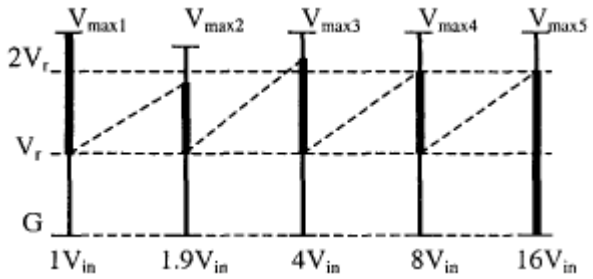

Fig.11. Impact of amplification mismatch. [1]

The gain mismatch in the amplifier network is due to two factors: the accuracy of the passive divider and the amplifiers. With careful layout the mismatch in the passive divider can be minimized. The gain in the amplifiers is harder to match. The negative feedback (NFB) with high loop gain, $A\beta$, makes total gain inversely proportional to feedback gain β. It is clear that a high loop gain will transfer the gain matching problem from the active element to the passive feedback network.

Delay Time Error:

When the delays of the amplifiers are not completely matched, the signal will have different delays through the amplifier tree. The sample point will then be changing relatively to the input signal when switching the active output from one node to another. This will introduce a time jitter and add distortion to the signal. As seen in Fig.12.the error is maximum, where the signal slope is at its maximum.

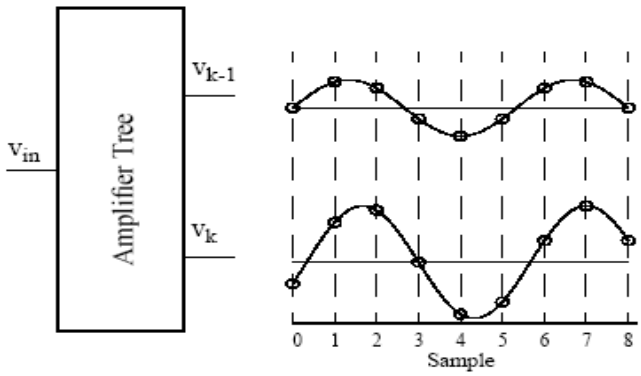

Fig.12. Jitter when switching mode. [1]

3 Parallel FP-ADC

A parallel FP-ADC structure that exhibits a faster conversion rate is presented. The quantization function of the parallel FP-ADC [2] is identical to the one of the sequential FP-ADC, but the conversion speed is increased by a factor that depends on the statistical properties of the acquired signal. The presented circuit consists of two ADCs (ADC-M and ADC-E) and PGA, as shown in Fig. 13. The ADC-M is a uniform quantization ADC that measures the mantissa, and has an *m*-bit resolution. To find out the exponent of the quantized signal, the uniform quantizer ADC-E and the GCU blocks implement a logarithmic conversion characteristic. The two steps of the classic sequential FP-ADC are performed in parallel, as they are presented in the *Statechart* diagram of Fig. 14. The conversion process (*ACQ*) is completed in two concurrent states, *AcqExponent* and *AcqMantissa* (Fig. 14) that are carried out by ADC-E and ADC-M (Fig. 13), respectively; both converters are fired up synchronously by the *start* signal. The ADC-E performs a coarse conversion and based on its result (y_e), the GCU determines the exponent *e* and the corresponding PGA gain^t. At the same time, the ADC-M performs a conversion for getting the mantissa y_m. The PGA gain is set to a value that was previously predicted based on the history of the acquired signal, aiming to maintain the PGA output (*x*') in the upper half of the ADC-M conversion range, thus getting the best accuracy out of the ADC *m*-bit resolution. The simplest approach of a zero degree polynomial extrapolation is considered to illustrate the gain prediction of the parallel FP-ADC (PGA gain:= $gain^{t-1}$).

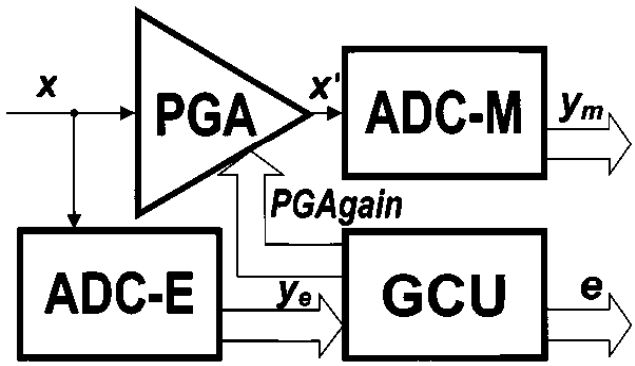

Fig.13. Block diagram of parallel FP-ADC. [2]

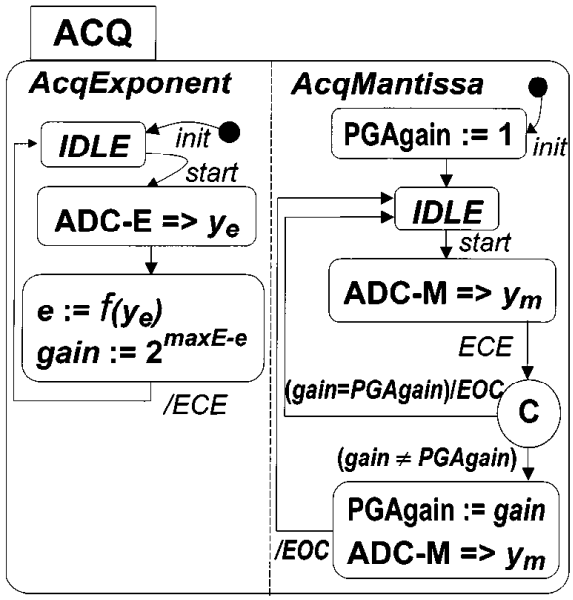

Fig.14. Statechart diagram of parallel FP-ADC. [2]

The result of the floating-point quantization is correct if the absolute value of the current quantized signal ($x' = x \cdot gain^{t-1}$) falls in the upper half of the ADC-M conversion range ($V_{FS} > x' > V_{FS}/2$), i.e., the ADC-E has currently acquired the same exponent as the preceding sample ($gain^t = gain^{t-1}$).This gain comparison is carried out as soon as the GCU determines the current gain and sends the termination logic signal *ECE* over to the ADC-M to synchronize its operation. If the predicted gain coincides with the current acquired one, the GCU generates the *EOC* signal to indicate the end of conversion and the ADC-M comes back to its initial *IDLE* state, as well. If the previously predicted gain of the PGA ($gain^{t-1}$) has driven the signal out of the conversion range of the ADC-M ($|x'| > V_{FS}$) or it falls inside the lower half of

it ($|x'| < V_{FS}/2$), the GCU sets the PGA gain to the current acquired exponent (PGA $gain^{t-1}$) and the ADC-M repeats the measurement to get the correct mantissa. The *EOC* signal is generated after this second conversion cycle is accomplished, and the ADC-M is driven back into its initial *IDLE* state. In conclusion, a correct acquisition is accomplished in one conversion cycle if the voltage at the input of the ADC-M is in the upper half of its conversion range otherwise in a series of two conversion cycles if a readjustment of the gain of the PGA is required.

3.1 Characteristics of parallel floating-point ADC

The quantizer's dynamics, the signal-to-quantization noise Ratio (SNR) and the conversion time are the most important characteristics of the FP-ADC that are considered. The performance of the proposed FP-ADC [2] that uses *E* bits to represent the exponent (i.e., ADC-E resolution) and *m* bits for the mantissa (i.e., ADC-M resolution) is compared against the characteristics of a fixed-point uniform quantizer having *b*-bits resolution.

A. Quantizers' Dynamics

The quantizer's dynamic is defined as the ratio between the largest admissible quantizer input signal and the smallest one that gives a nonzero quantized acquisition, given the condition that the magnitude of the input signal is bounded to the quantization range. The 2's *complement* code is usually used in uniform mid tread quantizers to represent bipolar signals. The largest positive number that can be expressed with *b* bits in 2's *complement* is ($y^+ = 2^{b-1}$-1), while the smallest negative number is ($y^- = 2^{b-1}$). For $y^+ \approx |y^-|$, to be considered an acceptable approximation, the dynamic D_U of a *b*-bit resolution ADC is given by equation (4)

$$D_U = 2^{b-1}. \tag{4}$$

To evaluate the efficiency of using a word of a given length, the resolutions of the uniform quantizer (*b*) and of the FP-ADC (*E*+*m*), respectively, are considered equal:

$$b = E+m. \tag{5}$$

Observing the same constraint from equation (4),i.e.$y^+ \approx |y^-|$ with max$E = 2^E$ -1, the FP-ADC dynamic can be expressed in terms of the dynamic of the fixed-point uniform quantizer (D_U)

$$D_F = 2^{\max E} \cdot 2^{m-1} = 2^{\max E - E} \cdot D_U. \tag{6}$$

Comparing a 12-bit uniform quantizer with an FP-ADC characterized by $E+m=12$, the FP-ADC can acquire signals with higher dynamics; if the exponent is expressed by a 3-bit number ($E=3$), then $D_F/D_U=16$ and, if $E=4$, then $D_F/D_U=2048$.

B. Signal-to-Quantization Noise

Then, the signal-to-quantization noise ratio [4] of the uniform quantizer can be expressed in terms of its resolution b, as given by equation (7)

$$SNR_U=3.\,4^{b-2}. \tag{7}$$

If a normally distributed signal with mean 0 and variance is quantized with this uniform quantizer having an input range of $[-4\sigma, +4\sigma]$, then the ADC quantization step Δ can be expressed in terms of its resolution and of the signal variance

$$\Delta = \frac{8\sigma}{2^b} = \sigma.2^{3-b}. \tag{8}$$

The quantizer absolute error is defined by

$$\varepsilon = Q(x) - x.$$

while the relative error is defined by

$$\varepsilon_r = \varepsilon/x = (\,Q(x) - x\,)\,/\,x. \tag{9}$$

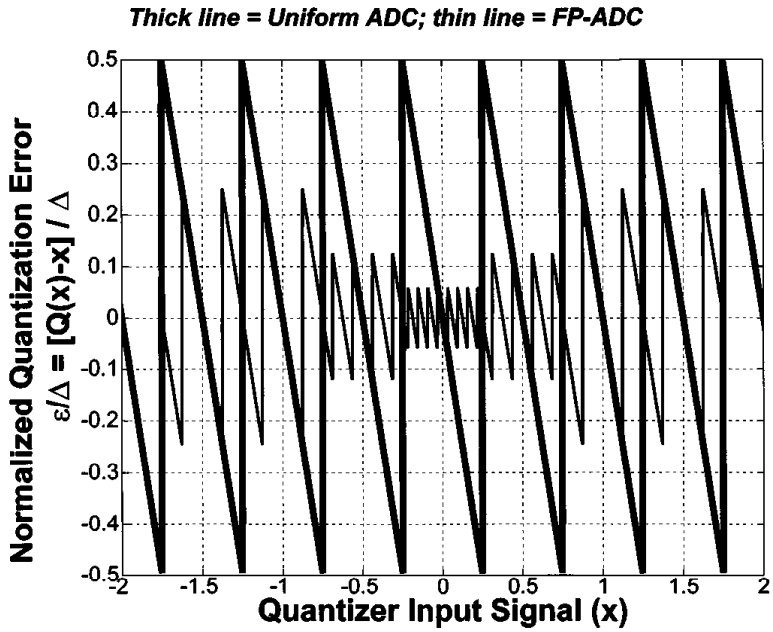

Fig.15. Normalized absolute quantization error. [2]

The uniform quantizer introduces an absolute error no larger than $\Delta/2$ along the quantizer input range. Since outside the interval [-2, 2] both quantizers present the same error, that region was not represented in Fig. 15. Inside [-2, 2], representing half of the maximum input

range, there are 2^E -1 measurement domains, where, compared to a uniform ADC, the FP-ADC exhibits a smaller quantization error, since the floating-point quantization step gets finer as the quantizer input signal becomes smaller as given by [2].

The *relative error* (ε_r) shows the superiority of the FP-ADC over the uniform ADC [2]. The *relative error* of the two compared quantizers is represented in Fig.16. The absolute error of the uniform quantizer (presented by the thick line graph in Fig.16.) equals the input signal ($\varepsilon = x$) all over the quantizer input sub-domain ($-\Delta/2$, $\Delta/2$), determining the highest relative error (100%). Since the smallest quantization step of the FP-ADC is $\Delta_E = \Delta/2^{\max E} << \Delta$. The FP-ADC exhibits a relative error remarkably smaller than that of the uniform quantizer all through the interval [-2, 2].

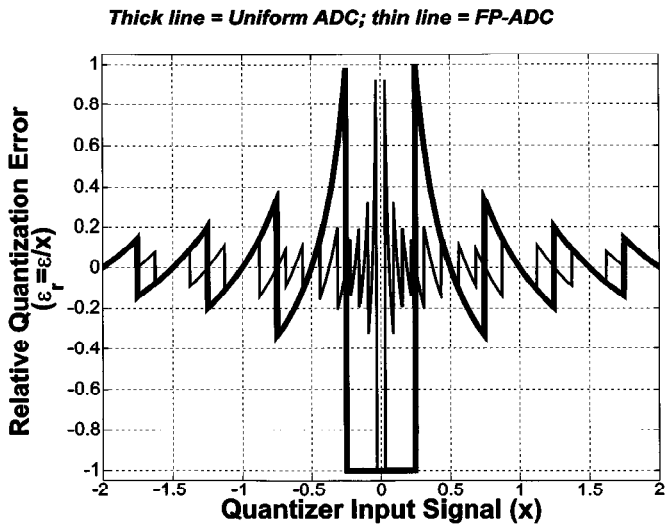

Fig.16. Relative quantization error. [2]

For a normally distributed signal with mean 0, *G(x)* is the ratio between the *SNR* of the uniform and of the floating-point quantizer, respectively (10).

$$G(x) = \frac{SNR_F}{SNR_U}. \tag{10}$$

If ADCs with the same resolutions are used to implement both the uniform and FP-ADC quantizers ($m=b$), then $G(x)>1$. The local values of this function are higher for small signals, as seen from the graph of Fig.17. Thus the smallest FP-ADC step length is set-up to quantize the smallest input signals. Increasing the resolution of the uniform ADC, increases the *SNR* of the FP-ADC highly compare to the uniform ADC for small signals, increases the dynamic range of FP-ADC [2].

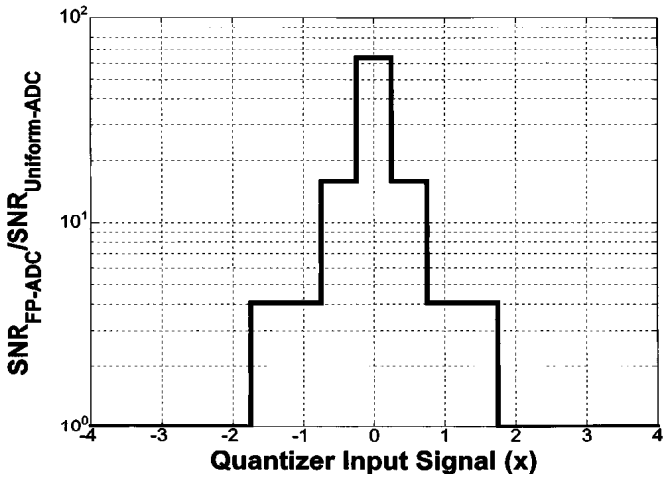

Fig.17. Relative quantization error. [2]

C. Conversion Time

The conversion time is the most important characteristics of the FP-ADC, if the FP-ADC is used, where high speed and high dynamic range are required like in radar applications. The conversion time of a classic FP-ADC with double sequential conversion is $T_F = 2T_U$. The conversion time of the parallel FP-ADC is function of statistic properties of the acquired signal [2]. If this is a second order centered Gaussian, then the average conversion time (T_F), expressed in terms of the auto correlation factor ρ is

$$c = -\frac{8}{1-\rho^2}. \tag{11}$$

$$T_F = (1+ e^c).T_U. \tag{12}$$

From the above equation (12), it is observed that the conversion time tends to be equal to that of a uniform fixed-point converter as ρ gets closer to 1. But as the signal fluctuation is more with respect to time, then conversion time cannot be closer to uniform fixed-point converter always.

4 Conclusion

The demand on very high resolution uniform ADC can be avoided by using the sequential FP-ADC discussed in sec.2.2, when the purpose is to cover the dynamic range rather than the accuracy. A deeper analysis of the pre amplifier performance regarding the delay mismatch has to be done. A parallel FP-ADC structure is discussed in sec.3 to make full use of the remarkable precision of the uniform quantizers, while minimizing the conversion time. This report presents a circuit that implements this structure along with the equations of the major static and dynamic parameters of the discussed parallel FP-ADC. For large and small signals,

the effective resolutions are kept constant, similar to a floating-point number representation by that high resolution for small signals are achieved. Further research is aiming toward the reduction of the acquisition time by a more accurate prediction of the measurement domain while minimizing the quantization noise. One aspect not looked upon in this report is the power consumption. This parameter is very important in today's portable equipment and a careful design should take this into consideration.

5 References

[1] J. Yuan and J. Piper, "Floating-point analog-to-digital converter," *6th IEEE Int. Conf. Electron. Circuits Syst., 1999. Proc. ICECS '99,* vol. 3, pp. 1385–1388, 1999.

[2] V.Groza, "High resolution floating point analog-to-digital converter," *IEEE transactions on Instrum. and Meas.*, vol. 50, no.6, pp. 1822–1829, Dec. 2001.

[3] J. Piper and J. Yuan, "A delay-balanced binary-weighted CMOS amplifier tree for a floating-point A/D converter," *Proceedings of the 16th Norchip Conference*, pp. 131-138, Nov. 1998.

[4] A. Gersho, "Principles of quantization," *IEEE Trans. Circuits Syst.*, vol.CAS-25, pp. 427–436, July 1978.

[5] Behzad Razavi, *Principles of data conversion system design*. New York: IEEE Press, 1995.