Mismatch-Induced Tradeoffs and Scalability of Mixed-Signal Vision Chips

A. Rodríguez-Vázquez, G. Liñán, S. Espejo and R. Domínguez-Castro
Instituto de Microelectrónica de Sevilla. IMSE-CNM-CSIC
Avda. Reina Mercedes s/n 41012 Sevilla (SPAIN)
Tel.:+34955056666, Fax: +34955056686
E-mail: angel@imse.cnm.es

Abstract[1]

This paper explores different trade-offs associated to the design of analog VLSI chips. These trade-offs are related to the necessity of keeping the analog accuracy while taking advantage of the possibility of reducing the power consumption, increasing the operation speed, and reducing the area occupation (i.e., increasing the density of processors), as fabrication technologies scale down into deep sub-micron.

1. Introduction

During the last few years several sensory-processing analog VLSI chips have been reported that are capable to "sense" images and "process" them concurrently [1]. Some of these chips also incorporate the basic functional features of general-purpose processors (stored-programmability, data memories, etc.) [2] and some basic decision-making and actuation circuitry; they hence define a first step towards the implementation of Vision Systems on Chip [3] [4].

The design of these chips is challenging due to the necessity to combine analog accuracy with small power consumption, large speed and large pixel density. Some of the reported chips feature quite good performance figures: 1.35×10^{11} OPS/W and 4.4×10^{9} OPS/mm^2 are reported for the chip in [4], which clearly outperforms conventional digital-based vision machines.

It can be expected that these figures can be further enhanced through the use of scaled-down technologies (the chip in [4] is realized in 0.5µm CMOS). However, one problem arises due to the necessity of keeping the analog accuracy, and hence the quality of the analog design, as transistor sizes decrease. In this paper we first identify mismatch as the main limit for the analog accuracy and then explore different trade-offs associated to the analog design of vision chips in the presence of mismatch.

2. Mismatch vs. Noise as Limiting Factor

The precision of any analog circuit is always constrained by different sources of unpredictable errors. Among them mismatch make two nominally identical devices to behave differently when they are used in a real integrated circuit. Based on the formulation of mismatch as a function of device geometries in [5], the variance of the large signal transconductance parameter β , the threshold voltage V_{T0}, and the slope factor n_p *[2] as function of the device area and aspect ratio can be represented as,

$$\sigma^2(V_{T0}) = \frac{A^2_{V_{T0}}}{A} + \frac{B^2_{V_{T0}}}{\sqrt{A^3 \cdot S}} + \frac{C^2_{V_{T0}}}{\sqrt{A^3 \cdot S^{-1}}}$$

$$\frac{\sigma^2(\beta)}{\beta^2} = \frac{A^2_{\beta}}{A} + \frac{B^2_{\beta}}{\sqrt{A^3 \cdot S}} + \frac{C^2_{\beta}}{\sqrt{A^3 \cdot S^{-1}}} \qquad (1)$$

$$\sigma^2(n_p) = \frac{A^2_{n_p}}{A} + \frac{B^2_{n_p}}{\sqrt{A^3 \cdot S}} + \frac{C^2_{n_p}}{\sqrt{A^3 \cdot S^{-1}}}$$

where A is the transistor channel area and S is the transistor aspect ratio.

Another accuracy limiting factor is noise. Noise is an unpredictable contribution to the instantaneous current of a MOS transistor whose effect in analog array implementations is often neglected due to the relative low accuracy needed by the process when compared to the limits that noise imposses.This limit have been some orders of magnitude above the required processing accuracy for many years [2]. However, the continuous trend towards low-voltage power supplies is reducing the signal range whereas noise limits are not reduced. This effect reduces the theoretic possible *SNR* and is becoming an actual limitation for even moderate accuracy analog processing.

[1] This work has been partially funded by DICTAM IST-1999-19007, and CICYT TIC1999-0826

[2] In the original model, the variance was formulated for the body effect factor γ. $\sigma^2(n_p)$ can be obtained as a function of $\sigma^2(V_{TO})$ and $\sigma^2(\beta)$.

0-7803-7448-7/02/$17.00 ©2002 IEEE

The equivalent noise current for a MOS transistor can be expressed as [6]:

$$\frac{\overline{i_n(t)^2}}{\Delta f} = 4 \cdot K \cdot T \cdot G_{ch} + \frac{K_F \cdot g_m^2}{A \cdot C_{ox}^h} \cdot \frac{1}{f^{A_f}} \qquad (2)$$

where h and A_f vary between 1 and 2; $G_{ch} =$ $= \beta \cdot (V_G - V_{T0} - n_p V_S)$ within ohmic region and 2/3 of this quantity in saturation; and $g_m = \partial I_{DS} / \partial V_{GS}$ is the small signal transconductance parameter .

Let us consider that the only significant mismatch error is that on the large signal transconductance parameter β – as it actually happens in many practical circuits used for stablishing interconnections in analog array processors [4]. In terms of the transistor area A and aspect S this error is expressed as,

$$\frac{\sigma_I^2}{I_{max}^2} = \frac{A_\beta^2}{A} + \frac{B_\beta^2}{\sqrt{A^3 \cdot S}} + \frac{C_\beta^2}{\sqrt{A^3 \cdot S^{-1}}} \qquad (3)$$

Under similar assumptions, the noise contribution can be approximated by:

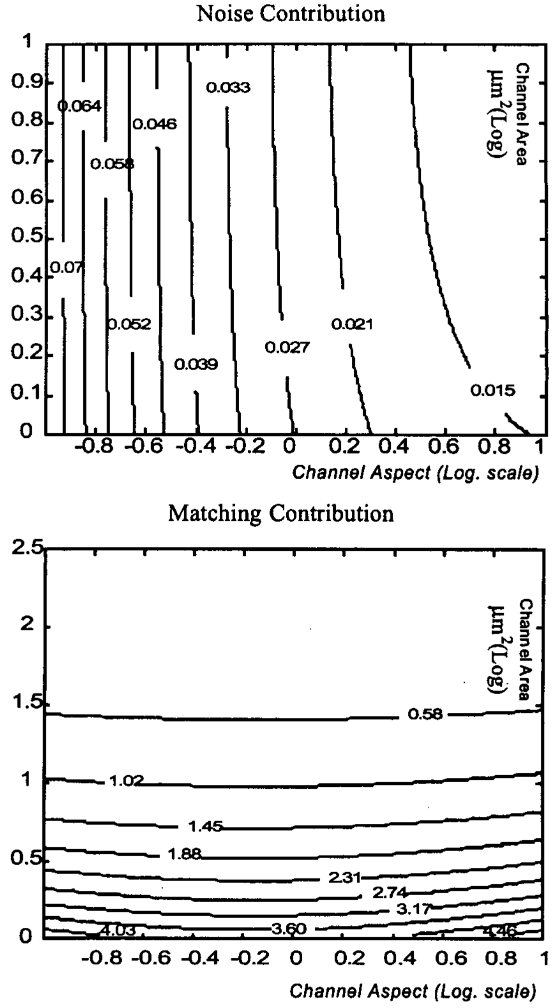
$$\frac{\sigma_I^2}{I_{max}^2} = \frac{4 \cdot k \cdot T \cdot [X_c + x_{max} - V_{TO} - n_p \cdot W_c - n_p \cdot w_{max}]}{\mu_0 \cdot C_{ox} \cdot S \cdot x_{max}^2 \cdot w_{max}^2} \cdot \Delta f + \\ + \frac{K_F}{C_{ox} \cdot x_{max}^2} \cdot \frac{\ln\left(\frac{f_{max}}{f_{min}}\right)}{A} \qquad (4)$$

We can now assign some typical numerical values for a $0.5\,\mu m$ CMOS technology and graphically represent noise and mismatch contributions as functions of the channel area and aspect ratio in order to compare them and, thus, to obtain which is the limiting one. Using typical parameters for CMOS 0.5μm: $X_c = 2.75\text{V}$, $x_{max} = w_{max} = 0.4\text{V}$, $V_{T0} = 0.65\text{V}$, $\mu_0 = 588\text{cm}^2\text{V}^{-l}\text{s}^{-l}$, $C_{ox} = 3.4 fF \mu\text{m}^{-2}$, $K_F = 3.6 \times 10^{-25} \text{V}^2\text{F}$ and considering a bandwidth $1 mHz$ to $5 MHz$, the graphs in Fig.1 results. There we see that for devices with channel areas of about $50 \mu m^2$ the matching level sets an accuracy slightly above 8 bits while for this same area and a channel aspect ratio of 0.1 the noise poses a limit in the resolution of 10.48 bits , far beyond from that posed by mismatching phenomena.

3. Effect of Scaling Process

Let us assume that a factor of λ is applied to obtain the next minimum feature size, that is, lateral dimensions scale as,

$$l_{min}^{new} = \frac{1}{\lambda} \cdot l_{min}^{old} \qquad (5)$$


Noise Contribution

Matching Contribution

Fig. 1. Comparing matching vs noise errors

Therefore, the gate oxide thickness, that approximately evolves in current technologies as $t_{ox} \propto \sqrt{l_{min}}$, scales as:

$$t_{ox}^{new} \cong \frac{1}{\sqrt{\lambda}} \cdot t_{ox}^{old} \qquad (6)$$

Assume that the synapse size define the achievable cell density,

$$Density \propto \frac{1}{L_X \cdot L_Y} \qquad (7)$$

where L_X and L_Y are the synapse width and length; as technologies scale down, $Density$ will evolve according to:

$$Density_{new} = \lambda^2 \cdot Density_{new} \qquad (8)$$

Another important parameter whose evolution needs to be considered is the time constant which for transconductance type synapses can be expressed as:

$$\tau = C/g_m \tag{9}$$

In the case of the one transistor synapse [4] employed to implement large analog processing arrays, the transconductance parameter is approximately given by:

$$g_m = \beta \cdot (V_D - V_S) = \mu_o \cdot C_{ox} \cdot S \cdot w \tag{10}$$

where w is the weight control signal [4].

On the other hand, assuming that the capacitor is implemented by using the gate capacitance of a MOS transistor (either N or P type) whose drain and source terminals are connected to the appropriated DC level (either supply, ground or any convenient level), the capacitance value neglecting border effects is approximately given by

$$C \cong C_{ox} \cdot A \tag{11}$$

This strategy for the implementation of the capacitor results into the smallest area occupation; also, the capacitor nonlinearity is not crucial for analog array processors [2].

From (10) and (11) the time constant becomes;

$$\tau = \frac{1}{\mu_o \cdot w} \cdot \frac{A}{S} \tag{12}$$

and scales as:

$$\tau_{new} = \tau_{old} \cdot \lambda^{-2} \tag{13}$$

We should also examine the way in which the reduction of the technology sizes affects accuracy. The question is, what happens with the technological parameters related to the accuracy issue when the technology scales down? Do they scale down also?

The answer is that not all of them scale as technology does. The historical trend shows [7] that scaling down produces a reduction of the main important parameter related with mismatching effects on the V_{T0} parameter, $A_{V_{T0}}$, as technology shrinking evolves – see Fig.2.

However, accuracy in the behavior of the one transistor synapse mainly refers to the existence on a random spatial distributed error in the effective weight value ($\beta \cdot w$) due to the random fluctuations on the β technological parameter.

The historical trend of A_β parameter differs from that of $A_{V_{T0}}$. It is observed that, contrary to $A_{V_{T0}}$, A_β parameter has remained practically unchanged for many generations of technologies having smaller and smaller minimum feature sizes.

Therefore, the accuracy related errors of the transistor current are approximately given by:

$$\frac{\sigma_I^2}{I_{max}^2} \approx \frac{A_\beta^2}{A} \tag{14}$$

and will evolve as:

$$\frac{\sigma_I^2\big|_{new}}{I_{max}^2\big|_{new}} \approx \frac{A_{old}}{A_{new}} \frac{\sigma_I^2\big|_{old}}{I_{max}^2\big|_{old}} \tag{15}$$

Consequently the relative error,

$$\varepsilon = \frac{\sigma_I}{I_{max}} \tag{16}$$

will grow according to:

$$\varepsilon_{new} = \varepsilon_{old} \cdot \lambda \tag{17}$$

We can conclude that accuracy is expected to be degraded as technologies scale down, if transistors are designed keeping the same area relative to the technology feature size. Indeed, accuracy can only be kept by approximately maintaining the same absolute area. Of course this statement is based on the historical trend of shrinking process and we do not know what will happen tomorrow but no indication about a possible change in this trend is observed in available technologies.

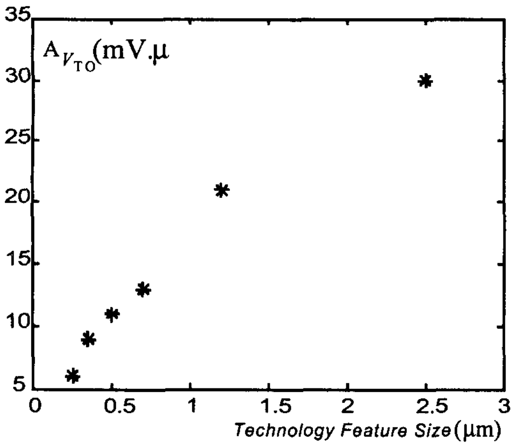


Fig. 2. Historical trend of $A_{V_{TO}}$ parameter

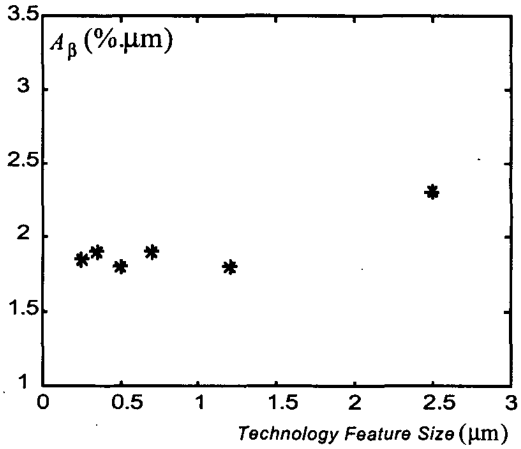


Fig. 3. Historical trend of A_β parameter.

4. Design Trade-Offs

Analog design art mainly consists, apart from many other things, of the combination of many design equations involving power consumption, speed, and accuracy. Typically, the objective is to meet the design requirements by minimizing (or maximizing) a certain F.O.M., Figure Of Merit, using the physical parameters of the transistors – area and aspect – as design variables.

Unfortunately, it is not possible to obtain a design by optimizing independently each of the existing design equations since they show cross-relationships among them. The problem, in these cases, becomes a problem of optimizing, at the same time, several design equations.

4. 1. Accuracy vs. Density

The dependence of the mismatch level with the channel aspect ratio is low for moderately large values of channel areas. Due to this, the channel area (A) is constrained by the required accuracy and therefore it can be written that the precision P satisfies:

$$P^{-2} = \frac{A_\beta^2}{A} \qquad (18)$$

where P is defined as:

$$P = \frac{I_{max}}{\sigma_I} \qquad (19)$$

On the other hand, the density of synapses, that is, the number of synapses per area unit, can be basically expressed as:

$$Density = \frac{K_{area}}{A} \qquad (20)$$

where A is the channel area ($W \cdot L$) and K_{area} is a constant that includes the influence of the routing lines, diffusion regions etc. on the achievable density.

Hence, a first trade-off can be formulated

$$\frac{P^{-2}}{Density} = A_\beta^2 \cdot K_{area} \qquad (21)$$

Accordingly to this, maximum achievable accuracy and cell density can not be separately optimized since the maximum the accuracy, the minimum the density and vice versa.

4. 2. Speed vs. Power.

The maximum power consumption of a synapse is expressed as:

$$\begin{aligned} Pow = V \cdot I = w_{max} \cdot \mu_o C_{ox} S \cdot x_{max} \cdot w_{max} = \\ = \mu_o \cdot C_{ox} \cdot S \cdot w_{max}^2 x_{max} \end{aligned} \qquad (22)$$

While the minimum time constant - maximum weight value - is given by

$$\tau = \frac{1}{\mu_o \cdot w_{max}} \cdot \frac{A}{S} = Speed^{-1} \qquad (23)$$

Therefore,

$$\frac{Pow}{Speed} = A \cdot C_{ox} \cdot x_{max} \cdot w_{max} \qquad (24)$$

Consequently, it seems that the only way to minimize this figure – reduce the power consumption and increase the speed – is by reducing the synapse's area. Nevertheless it automatically leads to a reduction on the achievable accuracy. On the other hand, reducing the signal ranges – x_{max} or w_{max} – will directly degrade the signal to noise ratio and then the accuracy.

A global figure of merit involving speed accuracy and trade-off can be formulated in the following way,

$$\frac{Pow \cdot Speed^{-1}}{P^2} = A_\beta^2 \cdot C_{ox} \cdot x_{max} \cdot w_{max} \qquad (25)$$

Since A_β does not show any evolution as technology is scaled down, this F.O.M. only depends on technology scaling process as C_{ox} does. Therefore, since $C_{ox} \propto \sqrt{\lambda}$ it is expected that the F.O.M. will increase – will worsen – in the future.

5. References

[1] A. Moini, *Vision Chips*. Kluwer Academic Publishers, 2000.

[2] T. Roska and A. Rodríguez-Vázquez (eds.), Towards the Visual Microprocessor. John Wiley & Sons, 2001.

[3] A. Rodríguez-Vázquez et al, "CMOS Design of Focal Plane Programmable Array Processors". *Proc. of the 9th ESAN*, pp. 57-62, Bruges, April 2001.

[4] G. Liñán et al., "A 0.5µm CMOS 10^6 Transistors Analog Programmable Array Processor for Real-Time Image Processing". *Proc. of the 1999 ESSCIRC*, pp. 358-361, September 1999.

[5] M.J.M. Pelgrom et al., "Matching Properties of MOS Transistors". *IEEE J. Solid-State Circuits*, Vol. 24, pp. 1433-1440, October 1989.

[6] E.A. Vittoz, "Future of Analog VLSI in the VLSI Environment". *Proc. of ISCAS 1990*, pp. 1372-1390, 1990.

[7] M. Steyaert et al., "Speed-Power-Accuracy Trade Off in High-Speed Analog-to-Digital Converters: Now and in the Future". *Proc. of the 9th Workshop in Analog Circuit Design*, April 2000.