Some Design Trade-Offs for Large CNN Chips using Small-Size Transistors

A. Rodríguez-Vázquez, G. Liñán, R. Domínguez-Castro, J.L. Huertas and S. Espejo

Instituto de Microelectrónica de Sevilla-Universidad de Sevilla
Edificio CICA, C/Tarfia s/n, 41012-Sevilla, SPAIN
FAX:: 34 5 4231832; Phone:: 34 5 4239923
email: angel@cnm.us.es

ABSTRACT

Small-size MOS transistors (MOST) exhibit a bunch of second-order effects which limit their application to design Cellular Neural Network (CNN) chips. The inverse dependency of mismatch with transistor sizes may result in severe accuracy degradation. Also, because of the down scaling of supply voltages with the technology feature size, noise and distortion produce large additional errors in submicron technologies. To reduce the influence of all these errors requires to properly choose the interconnection synapse circuitry, to perform intensive parametric optimization, and to use large enough transistor sizes. Consequently, the cell density and the operation speed cannot be scaled up to their limits because they have to be traded-off for accuracy. This trade-off is illustrated by the evaluation of the composed Power/ (Precision × Speed) figure, which results independent on the sizes. In addition to the parametric errors, catastrophic faults impose a limit on the maximum chip size for given yield, and open the issues of fast go/no-go testing, fault-driven reconfiguration and/or multi-chip architectures.

1. Introduction

The potentials of CNNs and other vision-oriented systems are only fully realized chips [1][2]. Basic system-level targets in the design of these chips are to increase the cell density (number of processing cells per unit area), and the cell operation speed. As the technology scales down to submicron all the lateral dimensions decrease by the scaling factor λ, and the vertical dimensions scale as λ^{-a}, where a is typically around 1/2 [1] [3]. Consequently, the MOS gate capacitance per unit area increases as λ^a, and the small-signal transconductance per unit channel-ratio increases also as λ^a for fixed bias. As a result one could ideally expect,

$$\text{cell density} \propto \lambda^2 \qquad \text{time constant} \propto \lambda^{-2} \tag{1}$$

with constant current and, hence, with no penalty on the power consumption, provided that the voltage ranges remain constants [2]. However, the actual scaling scenario is more pessimistic because of the increased influence of second-order phenomena on small-size MOSTs. Some recently reported submicron CNN CMOS chips [5][6] feature smaller cell density and operation speed, and larger power consumption, than expected from these formulae. Particularly, [6] obtains 27.5cells/mm^2, a time constant of 0.4μs, and 7bit analog accuracy in a 0.8μm 5V technology. Obviously, there is still room to improve these chips through structural and parametric optimization. However, if precision is a design goal, the cell density and operation speed will be inevitably constrained by mismatch and noise. And these constraints are expected to become harder as the signal dynamic ranges decrease because of the down scaling of supply voltages in deep submicron technologies.

Out of the circuitry of a CNN cell, the interconnection synapse consume much of the area [2] and is the critical part in terms of precision and speed. This paper addresses the influence of mismatch and noise on the operation of the strong-inversion MOST synapse, and outlines some practical trade-offs and issues induced by these errors.

2. MOST Synapse

We will describe the non-ideal synapsis behavior by,

$$\begin{aligned} y = {} & Y_{os} + k\left(1+\frac{\Delta k}{k}\right)(w-W_{os})\,(x-X_{os}) + D_{x_2}x^2 + D_{w_2}w^2 \\ & + D_{x_2w}x^2w + D_{xw_2}xw^2 + D_{x_3}x^3 + D_{w_3}w^3 + \ldots \approx kwx \end{aligned} \tag{2}$$

where w and x are the weight and input signals. In the ideal case, k is an error-free coefficient; all the remaining coefficients above are null in this ideal scenario.

Practical synapse circuits may have offset, error gain and distortion even with ideal MOSTs. The simplest synapse consist of one transistor (Fig. 1),

$$\begin{aligned} & y \approx \beta\left[(x-V_T)\,w - \frac{w^2}{2}\right] && \text{ohmic} \\ & y \approx \frac{\beta}{2}(w-V_T)^2 + \beta(w-V_T)\,x\left\{1+\frac{1}{2(w-V_T)}x\right\} && \text{sat.} \end{aligned} \tag{3}$$

where β may include the effect of mobility degradation [7]. Their large non-linearities result into small values of the maximum input and weight signals, x_{max} and w_{max}, thus decreasing the maximum output signal y_{max} and increasing the influence of mismatch and noise. Linearization is needed to enhance these ranges in practical circuits.

1. This corresponds to the constant voltage (CV) scaling law[3] —largely used because it keeps the digital noise margin [4].

2. In [6], 45% of the area is occupied by the synapse, 8% is for an integrator and a non-linear block, 15% is devoted to analog memories, 25% to the digital and control circuitry, and 7% to the optical interface.

0-7803-3583-X/97 $10.00 ©1997 IEEE

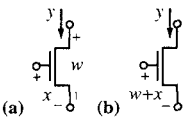

Fig. 2 *Single MOST Synapse: (a) ohmic region; (b) saturation region.*

Fig. 2 shows two common linearized synapse: Fig. 2(a) [8] operates in ohmic region, while Fig. 2(b) (a MOST version of the Gilbert multiplier [9]) operates in saturation. Assuming matched transistors and neglecting mobility degradation [3] both circuits obtain null second-order nonlinear errors and zero offset, and, hence, qualify a priori for practical CNN chips.

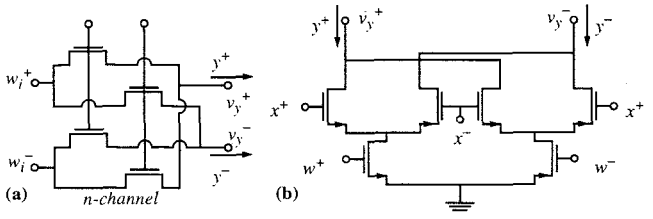

Fig. 1 *Linearized ohmic-MOST and saturation-MOST synapse.*

3. Errors due to Matching and Noise

Mismatch

The precision of CNN analog synapses, and any other analog circuit, is constrained by mismatch [10]. Its formulation as a function of the sizes is cornerstone for parametric yield optimization and design centering [11]. These models show that the variance of the fluctuations in the threshold voltage (V_T) and the transconductance factor (β) show an inverse dependence with transistor sizes, which for small-size transistors can be approximated as [10][12][13] [4],

$$\sigma_{V_T}^2 \approx \frac{A_{V_T}^2}{WL} + \frac{B_{V_T}^2}{W^2L} + \frac{C_{V_T}^2}{WL^2} \qquad \frac{\sigma_\beta^2}{\beta^2} \approx \frac{A_\beta^2}{WL} + \frac{B_\beta^2}{W^2L} + \frac{C_\beta^2}{WL^2} \tag{4}$$

The coefficients are technology-dependent and are reasonably expected to become smaller as the technology scales down. However, despite this reduction, mismatch will always limit the scaling of silicon area for given accuracy.

Mismatch errors are smaller for ohmic than for saturation region; it will be illustrated through the comparison of the two synapse of Fig. 2, in Section 5. The current error for the single MOST in ohmic region is,

$$\frac{\sigma_y^2}{y_{max}^2} \approx \frac{\sigma_\beta^2}{\beta^2} + \sigma_{V_T}^2 \frac{1}{(x - V_T)_{max}^2} \tag{5}$$

where signal ranges are determined from linearity issues.

Fig. 3(a) illustrates the dependency of this error with the channel area WL (log. scale) and the aspect ratio W/L (log. scale), for $(x-V_T)_{max}$=1V in a 0.8μm technology. Each curve corresponds to a different percentage of error (%). With these data, to obtain around 7 equivalent-bits requires a channel area of about 50μm^2 for square transistors and larger for rectangular transistors. On the other hand, if the signal range is decreased by α^{-1} ($\alpha > 1$), the area might has to be increased by up to α^2 to keep the precision. Similar qualitative relations are observed for Fig. 1(b) and the linearized synapse of Fig. 2.

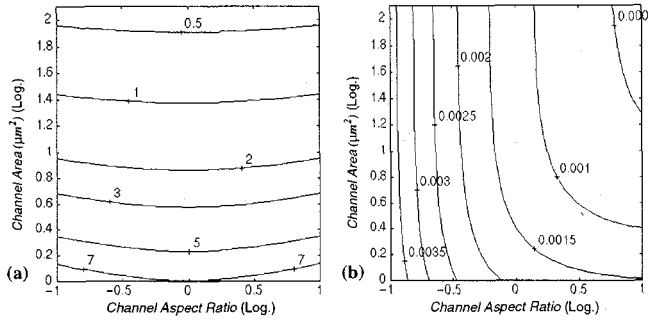

Fig. 3 *Full-Scale Percentage Errors: (a)Mismatch; (b) Noise.*

Noise

The noise-induced errors are due to the thermal and the flicker fluctuations of the channel current [7]. Since the latter depends inversely on the channel area, it dominates across a large frequency band for small-size transistors. However, the flicker component becomes less signicant for non-minimum area transistors, as those required to keep the accuracy, and the thermal contribution cannot be discarded.

The equivalent noise current is [7][13],

$$\frac{\overline{y_n^2}}{\Delta f} \approx 4kTG_{ch} + \frac{K_f g_m^2}{WLC_{ox}^h f^{A_f}} \tag{6}$$

where h and A_f vary between 1 and 2, and $G_{ch} \approx \beta (V_{GS} - V_T)$ in ohmic and 2/3 of this quantity in saturation. The errors in the former region are given by,

$$\frac{\sigma_y^2}{y_{max}^2 \Delta f} = \frac{4kT}{\mu C_{ox}} \frac{L}{W} \frac{1}{(x - V_T)_{max}^2 w_{max}^2} + \frac{K_f}{C_{ox}^m} \frac{1}{WL} \frac{1}{w_{max}^2} \frac{1}{f^{A_f}} \tag{7}$$

Fig. 3(b) shows the integrated noise error as a function of WL and W/L; the regions dominated by the thermal and the flicker contributions are clearly discernible. As for Fig. 3(a), we have assumed $(x-V_T)_{max}=w_{max}$=1V in a 0.8μm technology. The noise has been integrated from 1mHz to 1MHz. For these data, the noise error is much smaller than the mismatch error — a positive consequence of the relatively large time constant needed to guarantee correct dynamic operation of the system [5]. Assuming square transistors of 50μm^2 and keeping the inte-

3. It does not preclude to obtain INL values smaller than 0.4% and THD values below 0.2% for up to 2V input range in a 5V technology [6].

4. This model is somewhat controversial. Much more modelling effort is required to accurately capture the mismatch of small-size transistors.

5. The processing advantages of CNNs are not tied to the speed of the cells, but to the parallel operation of the system.

gration band, mismatch and noise become comparable only for $(x-V_T)_{max}=w_{max}$ <400mV. Because these tiny dynamic ranges will only appear for about 2.5V supply (for the linearized synapsis of Fig. 2(a) in the 0.8μm technology) noise can be expected to become a serious drawback for deep submicron technologies. Otherwise, mismatch can be expected to constitute the major problem for accuracy.

4. Mismatch-Induced Trade-offs

Let us assume for simplicity that the border effects are negligible in (4); i.e. that $B_{V_T} = C_{V_T} = B_\beta = C_\beta = 0$. The following expressions are found for the performance aspects related to the ohmic region synapse,

$$\text{Prec.}^{-2} \approx \frac{A_\beta^2}{WL} + \frac{1}{s_{max}^2}\frac{A_{V_T}^2}{WL} \qquad \text{Density}^{-1} \propto WL$$
$$\text{Power} \approx \mu C_{ox} s_{max}^3 \frac{W}{L} \qquad \text{Speed}^{-1} \approx \frac{1}{\mu s_{max}} L^2 \tag{8}$$

where Prec. stands for precision and we assume equal range s_{max} for x and w. The top expressions highlight a constraint on the cell density,

$$\text{Density}^{-1} \propto \frac{1}{\text{Prec.}^{-2}}\left(A_\beta^2 + \frac{A_{V_T}^2}{s_{max}^2}\right) \tag{9}$$

If precision must be kept constant, the density might not be scaled up with the technology resolution. Quite on the contrary, because of the inverse dependency with s_{max}^2, and the reduction of s_{max} in deep submicron (due to supply voltage lowering) the density might even decrease as the feature size decreases. The only way to increase the cell density without degrading the precision is by decreasing the technology-dependent coefficients of the fluctuations in V_T and β.

A similar constraint can be found on the speed,

$$\text{Speed}^{-1} \approx \frac{1}{\mu \text{Prec.}^{-2} s_{max}}\left(A_\beta^2 + \frac{A_{V_T}^2}{s_{max}^2}\right) \propto \frac{\text{Density}^{-1}}{s_{max}} \tag{10}$$

where we have assumed $W=L$. This equation highlights also a trade-off between the speed and the cell density. Similar trade-offs are found among the other performance aspects. In general, they cannot be all improved simultaneously. As the sizes change, each index changes in a different manner so that the following composed figure is found,

$$\text{Power} \times \text{Prec.}^{-2} \times \text{Speed}^{-1} \approx \mu C_{ox}\left(s_{max}^2 A_\beta^2 + A_{V_T}^2\right) \tag{11}$$

Because the vertical dimensions scale as λ^{-a} $(1<a<2)$, this figure can be expected to scale up as λ^a; further improvement can be expected due to the reduction of A_{V_T} with t_{ox}.

5. Mismatch Errors in the Linearized Synapse

Because of the trade-offs above, proper design of the CNN synapse requires parametric optimization. Structural optimization is also required to improve the overall circuit performance. Structural optimization comprise two different levels: model optimization [14] and topology selection. The importance of the latter is illustrated below by comparing the two linearized synapse of Fig. 2.

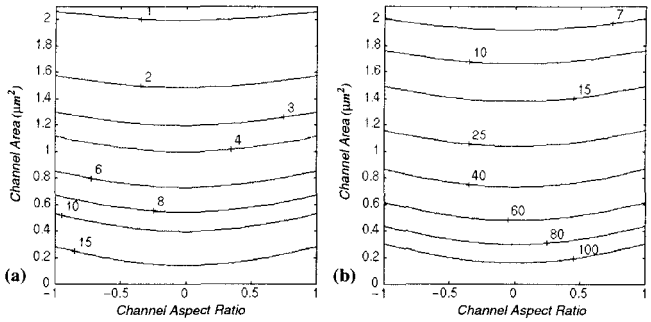

Fig. 4 *Mismatch Errors in the Linearized Synapse: (a) Fig.2(a); (b) Fig.2(b).*

Both synapse have zero offset and second-order distortion under nominal conditions. However, mismatch render these errors different from zero. The first six rows in Table 1 shows simplified expressions for these errors under the design conditions in the last three rows. From this table, we conclude that Fig. 2(b) requires larger area occupation and/or power consumption to obtain the same levels of accuracy than Fig. 2(a). Fig. 4 illustrates this by showing the mismatch-induced errors for the two synapse under the assumption of equal power consumption and area occupation. Differences in precision are dramatic, supporting our statement about the necessity of structural optimization for CNN design.

For submicron design the capability of the different synapse to operate with low-voltage supply has also to be explored. To this regard, Fig. 2(a) exhibit also excellent performance. In the 0.8μm technology it obtains about 500mV signal range for 0.5% precision (including distortion) with 3.3V supply, and about 360mV for the same precision with 2.2V supply. Above this voltage level, the influence of noise is less significant than that of mismatch. Below this level, noise becomes also important and forces using larger transistors to keep the precision.

6. Testing Issues

Although next generation technologies will have lower defect densities, the envisaged complexity levels of CNN chips will make test and yield-oriented design capital issues. Because the careful electrical and physical design of the cell, defects translated into non-catastrophic failures are expected to have an effect mainly on the interconnects. Then, techniques for fast verification of wiring may deserve attention rather than a system-level test methodology — reserved as the final test step for those chips passing the previous one.

Concerning spot defects causing catastrophic failures, they can affect to both the active and the interconnection parts. Probably these defects will dominate, provided that the design has been optimized on the basis of a thorough formulation of parametric errors, and a second test step has to be devoted to detect them. It would be interesting to

develop procedures that can be simultaneously applied to all the cells and collectively interpreted, in order to speed-up the whole production test.

In any case, for an industrial application of CNN chips, it might be pertinent to accompany the test procedure with reconfiguration schemes. To that purpose, the cells have to be modified to accommodate some kind of rerouting allowing to disconnect those that are not working correctly. As an alternative, or a complement, multi-chip configurations have to be devised as a way to conciliate large cell counts with large production yield.

ACKNOWLEDGEMENT: This work has been funded by spanish CICYT under contract TIC96-1392-C02-02 (SIVA).

7. REFERENCES

[1] C.Koch and H.Li (editors): *Vision Chips*. New York, IEEE Press 1994.

[2] T. Roska and L.O. Chua: "The CNN Universal Machine: An Analogic Array Computer". *IEEE Trans. on Circ. and Syst.-II*, Vol., 40, No.-3, March 1993.

[3] S. Wong and C.A.T. Salama: "Impact of Scaling on MOS Analog Performance". *IEEE J. Solid-State Circ.*, Vol. 18, pp. 106-114, Feb. 1983.

[4] J.B. Kuo: *CMOS Digital IC*. McGraw-Hill, Taipei 1996.

[5] P. Kinget and M. Steyaert: "An Analog Parallel Array Processor for Real-Time Sensor Signal Processing". *1996 Int. Solid-State Circ. Conf.*, paper 6.1.

[6] S. Espejo et al.: "A 0.8μm CMOS Programmable Analog-Array-Processing Vision Chip with Local Logic and Image Memory". Proc. of 1996 *Eur. Solid-State Circ. Conf.*, pp. 280-283, 1996.

[7] H.C. de Graaf and F.M. Klaassen: *Compact Transistor Modeling for Circuit Design*. Springer-Verlag, New-York 1990.

[8] B. Song: "CMOS RF Circuits for Data Communication Applications". *IEEE J. Solid-State Circ.*, Vol.21, pp. 310-317, April 1986.

[9] B.Gilbert: "A Precise Four-Quadrant Multiplier with Subnanosecond Response". *IEEE J. Solid-State Circ.*, Vol. 3, pp. 365-373, Dec. 1968.

[10] K.R. Lakshmikumar et al.: "Characterization and Modelling of Mismatch in MOS Transistors for Precision Analog Design". *IEEE J. Solid-State Circ.*, Vol. 21, pp. 1057-1066, Dec. 1986.

[11] R. Spence and R.S. Soin: *Tolerance Design of Electronic Circuits*. Addison-Wesley, Wokingham 1988.

[12] M.J.M Pelgrom, A.C.J. Duinmaijer and A.P.G. Welbers: "Matching Properties of MOS Transistors". *IEEE J. Solid-State Circ.*, Vol.24, pp. 1433-1440, October 1989.

[13] E.A. Vittoz: "Future of Analog VLSI in the VLSI Enviroment". *Proc. of the IEEE Int. Symp. Circ. and Syst.*, pp. 1372-1375, 1990.

[14] S. Espejo et al.: "A VLSI-oriented Continuous-Time CNN Model". *Int. J. Circuit Theory and ApplicationsI*, Vol. 24, pp. 341-356, May 1996.

Table 1

	Fig.2(a)	**Fig.2(b)**
$\left(\frac{\sigma k}{k}\right)^2$	$\frac{1}{4}\left(\frac{\sigma\beta_x}{\beta_x}\right)^2$	$\frac{1}{16}\left(\frac{\sigma\beta_x}{\beta_x}\right)^2+\frac{1}{8}\left(\frac{\sigma\beta_w}{\beta_w}\right)^2$
$\left(\frac{\sigma X_{os}}{x_{max}}\right)^2$	$\left(\frac{\sigma V_{Tx}}{x_{max}}\right)^2+\frac{1}{4}\left(1+\frac{w_{max}}{x_{max}}\right)^2\left(\frac{\sigma\beta_x}{\beta_x}\right)^2$	$\left(\frac{\sigma V_{Tx}}{x_{max}}\right)^2+\frac{\beta_w}{2\beta_x}\left(\sqrt{\frac{\beta_x}{\beta_w}}+\frac{w_{max}}{2x_{max}}\right)\left(\frac{\sigma\beta_x}{\beta_x}\right)^2$
$\left(\frac{\sigma W_{os}}{w_{max}}\right)^2$	$\left(\frac{\sigma w}{w_{max}}\right)^2$	$2\left(\frac{\sigma V_{Tw}}{w_{max}}\right)^2+\frac{1}{4}\left(\sqrt{\frac{\beta_x}{\beta_w}}\frac{x_{max}}{w_{max}}+\frac{1}{2}\right)\left(\left(\frac{\sigma\beta_x}{\beta_x}\right)^2+2\left(\frac{\sigma\beta_w}{\beta_w}\right)^2\right)$
$\left(\frac{\sigma Y_{os}}{y_{max}}\right)^2$	$\left(1+\frac{w_{max}}{x_{max}}\right)^2\left(\frac{\sigma v_y}{w_{max}}\right)^2$	$\left(1+2\sqrt{\frac{\beta_x}{\beta_w}}\frac{x_{max}}{w_{max}}\right)^2\left(\frac{\sigma V_{Tx}}{x_{max}}\right)^2+\frac{\beta_w}{2\beta_x}\left(\frac{w_{max}}{x_{max}}\right)^2\left(\sqrt{\frac{\beta_x}{\beta_w}}\frac{x_{max}}{w_{max}}+\frac{1}{2}\right)^4\left(\frac{\sigma\beta_x}{\beta_x}\right)^2$
$\left(\frac{\sigma D_{xx}x_{max}^2}{y_{max}}\right)^2$	0	$\frac{1}{8}\left(\frac{x_{max}}{w_{max}}\right)^2\left(\frac{\beta_w}{\beta_x}\right)\left(\frac{\sigma\beta_x}{\beta_x}\right)^2+9\left(\frac{\beta_x}{\beta_w}\frac{x_{max}}{w_{max}}\right)^2\left(\frac{\sigma V_{Tx}}{\frac{w_{max}}{2}+\sqrt{\frac{\beta_x}{\beta_w}}x_{max}}\right)^2$
$\left(\frac{\sigma D_{ww}w_{max}^2}{y_{max}}\right)^2$	$\frac{1}{4}\left(\frac{w_{max}}{x_{max}}\right)^2\left(\frac{\sigma\beta_x}{\beta_x}\right)^2$	$2\frac{\beta_x}{\beta_w}\left(\frac{w_{max}}{x_{max}}\right)^2\left(\frac{\sigma\beta_x}{\beta_x}\right)^2$
Design Conditions	$w_{cm}=v_{ycm}$ $w_{max}=x_{max}$	$\beta_w=2\beta_x$ $w_{max}=x_{max}$
Maximum Supply Current	$I_{DDmax}=\frac{\beta_x}{2}w_{max}^2$	$I_{DDmax}=\frac{\beta_x}{2}x_{max}^2+\frac{\beta_w}{2}w_{max}^2+\frac{\sqrt{\beta_w\beta_x}}{2}x_{max}w_{max}$
Maximum Current	$y_{max}=2I_{DDmax}$	$y_{max}=\frac{1}{(\sqrt{2}+1)}I_{DDmax}$