Published jointly in *Journal of VLSI Signal Processing*, vol.8, pp.27-44, and in *Analog Integrated Circuits and Signal Processing*, pp.27-44, July 1994.

Analog VLSI Signal Processing: Why, Where and How?

Eric A.Vittoz
CSEM, Centre Suisse d'Electronique et de Microtechnique SA
Maladière 71, 2007 Neuchâtel, Switzerland

Abstract

Analog VLSI signal processing is most effective when precision is not required, and is therefore an ideal solution for the implementation of perception systems. The possibility to choose the physical variable that represents each signal allows all the features of the transistor to be exploited opportunistically to implement very dense time- and amplitude-continuous processing cells. This paper describes a simple model that captures all the essential features of the transistor. This symmetrical model also supports the concept of pseudo-conductance which facilitates the implementation of linear networks of transistors. Basic combinations of transistors in the current mirror, the differential pair and the translinear loop are revisited as support material for the description of a variety of building blocks. These examples illustrate the rich catalogue of linear and nonlinear operators that are available for local and collective analog processing. The difficult problem of analog storage is addressed briefly, as well as various means for implementing the necessary intrachip and interchip communication.

Introduction

It is a commonly accepted view that the role of analog in future VLSI circuits and systems will be confined to that of an interface, the very thin "analog shell" between the fully analog outer world and the fully digital substance of the growing signal processing "egg"[1]. The main advantage of all-digital computation is its cheap and potentially unlimited precision and dynamic range, due to the systematic regeneration of the binary states at every processing step. Other advantages are the cheap and easy design and test of the circuits. Digital circuits are designed by computer scientists working with advanced synthesis tools to implement specific or programmable algorithms. According to this view, the task of analog designers will be essentially concentrated on developing converters to translate, as early as possible, all information into numbers, with an increasing demand on precision, and to bring the results of the all-digital computation back to the real world.

This view is certainly correct for the implementation of all systems aiming at the precise *restitution* of information, like audio or video storage, communication and reproduction. Together with very fast but plain computation on numbers, these are the tasks at which the whole electronic industry has been most successful.

However, the very precise computation on sequences of numbers is certainly not what is needed to build systems intended for the quite different category of tasks corresponding to the *perception* of a continuously changing environment. As is witnessed by even the most simple animals, what is needed for perception is a *massively parallel collective processing* of a large number of signals that are continuous in time and in amplitude. Precision is not required (why would 16-bit data be needed to recognize a human face?), and nonlinear processing is the rule rather than the exception.

Low-precision analog VLSI circuits therefore seem to be best suited for the implementation of perception machines, specially if cost, size and/or power consumption are of some concern [2]. The following chapters are intended to show how very efficient small-area processing cells can be implemented by exploiting in an opportunistic manner all the features of the available components, in particular those of the transistor. This will allow massively parallel architectures to be implemented on single chips to carry out the required collective processing on real-time signals. Whether or not these architectures should be inspired from biological systems depends on the particular task to be solved. For example, handwriting is made for the human eye and should therefore be handled by taking inspiration from the human vision system, whereas the bar code has been invented for machine-reading and can thus be handled without reference to the eye.

The global perception of all relevant data is fundamental for advanced control systems. Therefore, analog VLSI also seems to be an ideal medium for the implementation of fuzzy controllers [3], [4].

Signal representation in analog processing circuits

Signals in an analog circuit are not represented by numbers, but by physical variables, namely voltage V, current I, charge, frequency or time duration. The various modes of signal representation have respective advantages and drawbacks and can therefore be used in different parts of the same system.

The voltage representation makes it easy to distribute a signal in various parts of a circuit, but implies a large stored energy $CV^2/2$ into the node parasitic capacitance C.

The current representation facilitates the summing of signals, but complicates their distribution. Replicas must be created which are never exactly equal to the original signal.

The charge representation requires time sampling but can be nicely processed by means of charge coupled devices or switched capacitor techniques.

Pulse frequency or time between pulses is used as the dominant mode of signal representation for communication in biological nervous systems. Signals represented in this pseudo-binary manner (pulses of fixed amplitude and duration) are easy to regenerate and this representation might therefore be preferred for long distance transfers of information. It is discontinuous in time, but the phase information is kept in asynchronous systems.

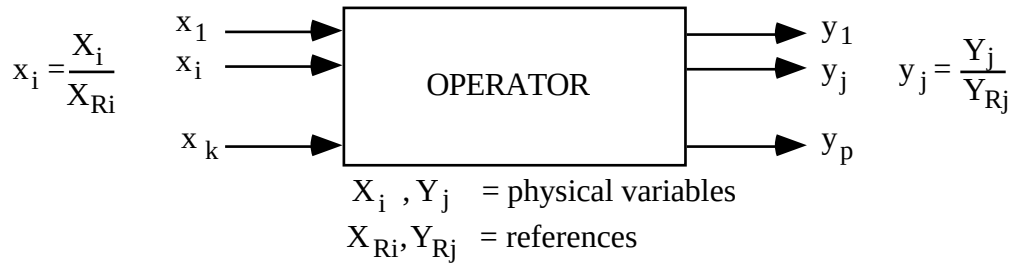


Fig. 1 : Physical variables and references.

In any case, each signal x is represented by the value of a physical variable X related to the corresponding reference value X_R. Thus, an operator with k input x_i and p output y_j (Fig.1) may require as many as k+p references X_{Ri},Y_{Rj} . These references must either be produced internally, with their origin clearly identified to keep them under control with respect to process and ambiance variations, or provided from outside.

For some particular operations, the number of physical references required by the operator itself may be reduced to just one or even none at all , as shown in table 1. If the operation is time dependent, at least one time reference T_R is needed.

Product of k variables: $y = a\, x_1 x_2 x_3 \ldots x_k$

Thus: $Y = a \dfrac{Y_R}{X_{R1} X_{R2} X_{R3} \ldots X_{Rk}} X_1 X_2 X_3 \ldots X_k$

single reference

input var. X	output var. Y	Unit of reference	reference for k=1	
I	I	$[A^{1-k}]$	current gain	A_i
I	V	$[VA^{-k}]$	transresistance	R_m
V	I	$[AV^{-k}]$	transconductance	G_m
V	V	$[V^{1-k}]$	voltage gain	A_v

Table 1 : Product of k variables and necessary physical reference.

All circuit architectures should be based on *ratios* of matched component values [5], to eliminate the dependency on any other absolute value than the required normalization references X_R. As a simple example, the gain of a linear current or voltage amplifier (same representation at input and output) can and must be implemented by means of matched components to achieve complete independence of any absolute value. In a squarer or in a two-input multiplier with same representation at input and output, matched components must be used to limit the dependency on absolute values to that on the single voltage or current required by the operation, which must be provided by a clearly identified and well-controlled reference. These basic rules for sound analog design can only be respected if the circuits are analyzed with respect to their physical behavior and not simply numerically simulated on a computer.

The MOS transistor: a very resourceful device.

Fig.2 shows the schematic cross-section and the symbol of an n-channel MOS transistor. In order to maintain the intrinsic symmetry of the device, the source voltage V_S, the gate voltage V_G and the drain voltage V_D are all referred to the local p-type substrate (p-well or general substrate). This practice is very convenient for analog circuit design. The symbol and the definitions for the p-channel device are also shown in the figure. In schematics, the connection B to substrate is only represented when it is not connected to the negative supply rail V^- (for n-channel) or to the positive rail V^+ (for p-channel).

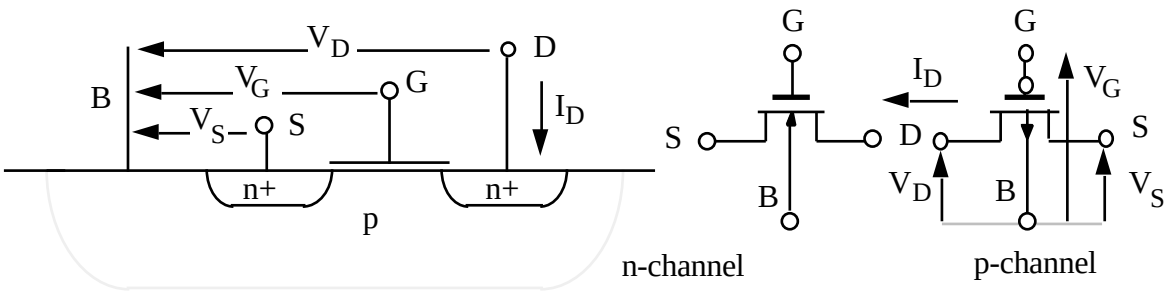


Fig. 2 : Schematic cross-section of an n-channel MOS transistor and symbols for n- and p-channel transistors. All voltages are referred to the local substrate. Positive current and voltages for the p-channel transistor are reversed to maintain the validity of the model derived for the n-channel.

Only three basic parameters are needed for a first order characterization and modelling of the transistor. These are:

V_{T0} : the gate threshold voltage for channel at equilibrium,

n : a slope factor usually smaller than 2 which tends to 1 for very large values of gate voltage V_G,

$\beta = \frac{W}{L}\mu C_{ox}$: the transfer parameter, measured in A/V^2, which can be adapted by the designer by changing the channel width-to-length ratio W/L. C_{ox} is the gate oxide capacitance per unit area and μ is the carrier mobility in the channel.

This last parameter may be conveniently replaced by

$I_S = 2n\beta U_T^2$: specific current of the transistor (typically 20 to 200 nA for W=L), where $U_T = kT/q$ is the thermal voltage (26mV at 300°K).

The drain current I_D of the transistor may be expressed as [6]:

$$I_D = I_F - I_R \tag{1}$$

where $I_F(V_G, V_S)$ is the forward component of current, independent of V_D
and $I_R(V_G, V_D)$ is the reverse component of current, independent of V_S.

The forward (reverse) component can be modelled with an acceptable precision in a very wide range of currents as [6]

$$I_{F(R)} = I_S \ln^2\left[1 + \exp\left(\frac{V_G - V_{T0} - nV_{S(D)}}{2nU_T}\right)\right] \tag{2}$$

If $I_{F(R)} \ll I_S$ ($V_G < V_{T0} + nV_{S(D)}$), this component is in *weak inversion* (also called *subthreshold*) and (2) can be approximated by the *exponential* expression

$$I_{F(R)} = I_S \exp\left(\frac{V_G - V_{T0} - nV_{S(D)}}{nU_T}\right) \tag{3}$$

If $I_{F(R)} \gg I_S$ ($V_G > V_{T0} + nV_{S(D)}$), this component is in *strong inversion* (also called *above threshold*) and (2) can be approximated by the *quadratic* expression

$$I_{F(R)} = \frac{I_S}{4}\left(\frac{V_G - V_{T0} - nV_{S(D)}}{nU_T}\right)^2 = \frac{\beta}{2n}(V_G - V_{TO} - nV_{S(D)})^2 \tag{4}$$

If $I_{F(R)}$ is neither much smaller nor much larger than I_S, this component of current is in *moderate inversion* [7] and must be expressed with the full expression (2).

The value of $V_{S(D)}$ for which the argument in (3) and (4) becomes zero is called the *pinch-off voltage* V_P, given by

$$V_P = \frac{V_G - V_{T0}}{n} \tag{5}$$

The component $I_{F(R)}$ is in weak inversion for $V_{S(D)} > V_P$ and in strong inversion for $V_{S(D)} < V_P$, with some margin needed to be outside the transition range of moderate inversion.

The global mode of operation of an MOS transistor depends on the combined levels of I_F and I_R, as shown in Fig. 3.

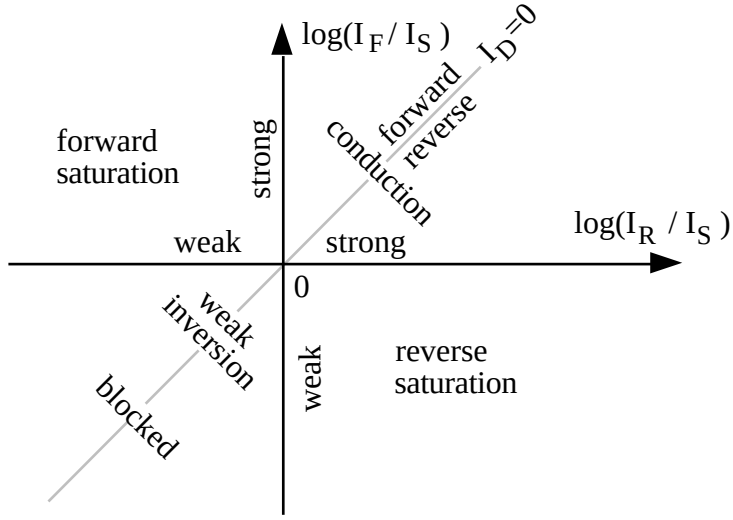


Fig.3: Modes of operation of an MOS transistor.

The transistor is said to be in *conduction* when both I_F and I_R are in strong inversion. The drain current given by (4) can then be expressed as

$$I_D = I_F - I_R = \beta(V_D - V_S)\left[V_G - V_{T0} - \frac{n}{2}(V_D + V_S)\right] \qquad \text{for } V_S \text{ and } V_D < V_P \quad (6)$$

If only one of the components is in strong inversion, then the other is negligible and the transistor is in *saturation*, forward and independent of V_D with

$$I_D = I_F = \frac{\beta}{2n}(V_G - V_{T0} - nV_S)^2 \qquad \text{for } V_S < V_P < V_D \quad (7)$$

or reverse and independent of V_S with

$$I_D = -I_R = -\frac{\beta}{2n}(V_G - V_{T0} - nV_D)^2 \qquad \text{for } V_D < V_P < V_S \quad (8)$$

Strong inversion is the natural mode of operation at high saturation currents, typically above 1μA. Its use at very low currents requires a large value of channel length L and therefore a large area per transistor. The ultimate limit is given by leakage currents generated in the depletion layer underneath the channel.

When both I_F and I_R are in weak inversion, the whole transistor is said to operate in the *weak inversion* mode, for which (3) and (1) yield [8]

$$I_D = I_F - I_R = I_S\, e^{\frac{V_G - V_{T0}}{nU_T}}\left(e^{\frac{-V_S}{U_T}} - e^{\frac{-V_D}{U_T}}\right) \qquad \text{for } V_S \text{ and } V_D > V_P \quad (9)$$

which can also become saturated (independent of V_D or V_S) as soon as $|V_D - V_S| \gg U_T$. Weak inversion is the natural mode of operation at low saturation currents, typically below 10nA. Its use at high currents requires a large value of channel width W and therefore a large area per transistor.

When both components are smaller than a minimum arbitrary level, usually determined by the leakage currents, the transistor is considered to be *blocked*.

It is worth remembering that an MOS transistor may also be operated as a bipolar transistor if the source or the drain junction is forward-biased in order to inject minority carriers into the local substrate. If the gate voltage is negative enough (for an n-channel device), then no current can flow at the surface and the operation is purely bipolar [9].

Fig. 4 shows the major flows of current carriers in this mode of operation, with the source, drain and well terminals renamed emitter E, collector C and base B.

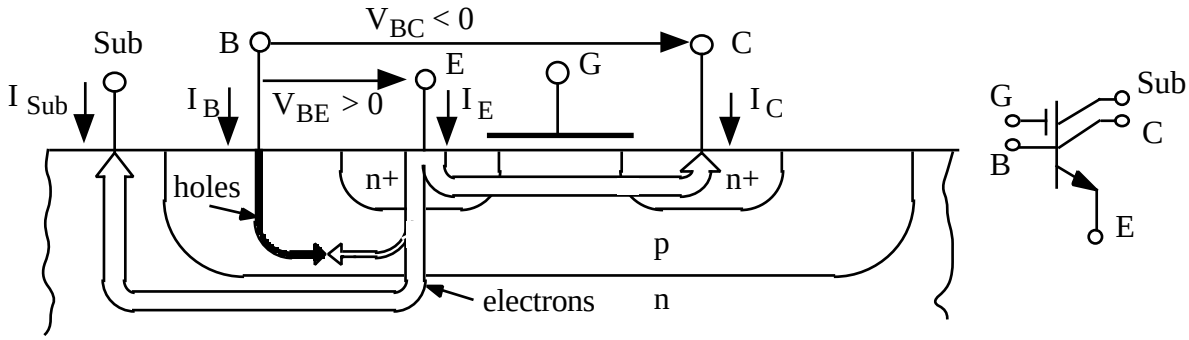


Fig. 4 : Bipolar operation of the MOS transistor : carrier flows and symbol.

Since there is no p+ buried layer to prevent injection to the substrate, this lateral npn bipolar (lateral pnp for an n-well process) is combined with a vertical npn. The emitter current I_E is thus split into a base current I_B, a lateral collector current I_C and a substrate collector current I_{Sub}. Therefore, the common-base current gain $\alpha = -I_C / I_E$ cannot be close to 1. However, due to the very small rate of recombination inside the well and to the high emitter efficiency, the common-emitter current gain $\beta = I_C / I_B$ can be large. Maximum values of α and β are obtained in concentric structures using a minimum size emitter surrounded by the collector.

For $V_{CE} = V_{BE}-V_{BC}$ larger than a few hundred millivolts, this transistor is in active mode and the collector current is given, as for a normal bipolar transistor, by

$$I_C = I_{Sb}\, e^{\frac{V_{BE}}{U_T}} \qquad (10)$$

where I_{Sb} is the specific current in bipolar mode, proportional to the cross-section of the emitter to collector flow of carriers.

The collector diffusion and the gate may be omitted, which leaves the vertical bipolar to substrate, limited in its applications by its grounded collector (substrate). This vertical bipolar can be used as a sensitive light sensor. Incident photons create electron-hole pairs, which are collected by the well-to-substrate junction to produce a component of base current. If the base terminal (well) is kept open, this current is multiplied by the high current gain β to become the emitter current.

Better bipolar transistors are of course available in BiCMOS processes.

Coming back to the field-effect modes of operation, the generic $I_D(V_D)$ characteristics for V_S and V_G constant are shown in Fig.5, where two qualitatively different behaviors can be identified: voltage-controlled current source and voltage-controlled conductance.

The transistor behaves as a *voltage-controlled current source* $I=I_F$ in forward saturation, i.e. for $V_D>V_{Dsat}$ given by

$$V_{Dsat} = V_S + 3 \text{ to } 4\, U_T \qquad \text{(from (9) for } \textit{weak inversion}\text{)} \quad (11)$$

$$V_{Dsat} = V_P = \frac{V_G - V_{T0}}{n} \qquad \text{(from (7) and (5) for } \textit{strong inversion}\text{)} \quad (12)$$

The transfer function $I(V_G)$ is exponential in weak inversion and quadratic in strong inversion. Connecting the gate to the drain and imposing a current I provides the inverse functions $V_G(I)$ which are respectively *logarithmic* and *square-root.*

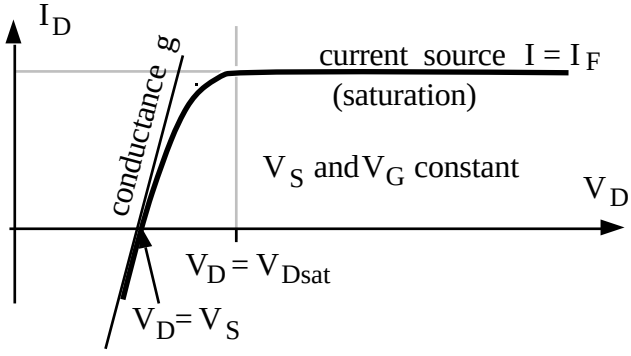


Fig. 5 : Generic output characteristics for V_G and V_S constant.

The small-signal dependence of this current source on the gate voltage can be characterized by the transconductance $g_m = \partial I/\partial V_G$. Equations (4) and (3) yield

$$g_m = \frac{I}{nU_T} = 2\beta U_T \exp\left(\frac{V_G - V_{T0} - nV_S}{nU_T}\right) \qquad \textit{(weak inversion)} \quad (13)$$

$$g_m = \frac{2I}{V_G - V_{T0} - nV_S} = \frac{\beta}{n}(V_G - V_{T0} - nV_S) = \sqrt{\frac{2\beta I}{n}} = \frac{\sqrt{I.I_S}}{nU_T} \quad \textit{(strong inversion)} \quad (14)$$

Weak inversion provides the maximum possible value of g_m for a given value of current I and the minimum value of saturation voltage V_{Dsat}.

The transistor behaves as a *voltage-controlled conductance* g for $V_D \approx V_S$. It can be easily shown that:

$$g = ng_m \qquad (15)$$

Inspection of (6) shows that in *strong inversion*, the $I_D = f(V_D - V_S)$ function can be kept linear if V_S and V_D are varied *symmetrically* with respect to a fixed voltage V. The conductance g is then constant and is given by (15) and (14) with $V_S = V$. This is true as long as the transistor remains in conduction, that is in the range $|V_D - V_S| \leq 2(V_{Dsat} - V)$.

According to (1) and (2), the current I_D flowing through the transistor can be expressed by

$$I_D = I_S \left[f(V_S, V_G) - f(V_D, V_G) \right] \qquad (16)$$

where $f(V, V_G)$ is a decreasing and always positive function of V. By defining a *pseudo-voltage*

$$V^* = -V_0 f(V, V_G) \qquad (17)$$

where V_0 is a scaling voltage of arbitrary value, (16) may be rewritten

$$I_D = g^* (V_D^* - V_S^*) \qquad (18)$$

This is the linear Ohm's law with a constant *pseudo-conductance* of the transistor

$$g^* = I_S/V_0 \qquad (19)$$

Thus, networks of transistors interconnected by their sources and drains with a common gate voltage behave linearly with respect to currents [42]. The pseudo-voltage V* cannot change sign, but its 0-reference level (pseudo-ground) is reached as soon as the corresponding real voltage V is large enough ($f(V, V_G)$ tends to 0 for V large), which facilitates the extraction of any current flowing to the pseudo-ground.

The pseudo-conductance given by (19) is in general not voltage-controllable and can only be changed by changing the channel width-to-length ratio W/L in I_S. However, according to (3), in weak inversion

$$f(V,V_G) = e^{\frac{V_G - V_{T0}}{nU_T}} \; e^{\frac{-V}{U_T}} \tag{20}$$

is the product of separate functions of V_G and V, and the linear Ohm's law (18) is still valid with the following new definitions of V* and g*:

$$V^* = -V_0 \, e^{\frac{-V}{U_T}} \tag{21}$$

$$g^* = \frac{I_S}{V_0} \; e^{\frac{V_G - V_{T0}}{nU_T}} \tag{22}$$

The network of transistors remains linear with respect to currents but the pseudo-conductance g* of each transistor is now controllable independently by the value of its gate voltage V_G. This linearity is available in the whole range of weak inversion, which may correspond to 3 to 6 orders of magnitude. The corresponding range of real voltages is only 7 to 14 U_T. It should be pointed out that, although they do tranfer linearly currents from input to output, these pseudo-linear networks should not be confused with translinear circuits [13].

A transistor behaves as a *switch* if its conductance g is changed from a very low to a very high value by a large swing of the gate voltage V_G.

For V_D and V_S « V_P, relation (6) becomes

$$I_D = \beta(V_D - V_S)(V_G - V_{T0}) \tag{23}$$

The transistor in strong inversion can thus be used as a *multiplier* of the drain-to-source voltage by the gate voltage overhead to produce the drain current.

Another possible function offered by a single transistor is that of an *analog memory.* The gate control voltage, and therefore the resulting drain current, can be stored on the gate capacitance. The gate electrode has just to be isolated, either by a switch, or by avoiding any connection to the gate which is left floating.

If the gate is isolated by a switch, the leakage current of the junctions associated with this switch limits the storage time to no more than a few minutes. If it is left floating, wrapped within a silicon dioxide layer, the storage time can be many years, but the problem is to write in the information by changing the gate voltage. This can be done by one of the various mechanisms used for digital E^2PROM.

Beyond the function of switch, which is the only one exploited in digital circuits, we have identified so far the following functions provided by a single transistor :

- Generation of square, square-root, exponential and logarithmic functions.
- Voltage-controlled current source.
- Voltage controlled conductance, linear in a limited range.
- Fixed or voltage-controlled pseudo-conductance, linear in a very wide current range.
- Analog multiplication of voltages.
- Short term and long term analog storage.
- Light sensor.

Additional functions offered by some basic combinations of just a few transistors are discussed in the following section.

Basic combinations of transistors.

As stated previously, analog circuits should be based upon ratios of matched components to eliminate whenever possible any dependency on process parameters. The amount of mismatch of the electrical parameters in similar components depends on the process, on the type of component, and on their size and layout [5].

The mismatch between two identical MOS transistors must be characterized by two statistical parameters [6], the threshold mismatch δV_{T0}, with RMS value σ_{VT}, and $\delta\beta/\beta$ with RMS value σ_β. These components of mismatch are usually very weakly correlated and their mean value is zero in a good layout. The RMS mismatch of drain current $\delta I_D/I_D$ of two saturated transistors with the same gate and source voltages can then be expressed as

$$\sigma_{ID} = \sqrt{\sigma_\beta^2 + \left(\frac{g_m}{I_D}\,\sigma_{VT}\right)^2} \tag{24}$$

whereas the mismatch δV_G of the gate voltage of two saturated transistors biased at the same drain current and source voltage is given by

$$\sigma_{VG} = \sqrt{\sigma_{VT}^2 + \left(\frac{I_D}{g_m}\,\sigma_\beta\right)^2} \tag{25}$$

As an example, these results are plotted in Fig.6 for σ_T = 5mV and σ_β = 2 %, by using a continuous value of $g_m(I_D)$ computed from equation (2) with nU_T = 40mV.

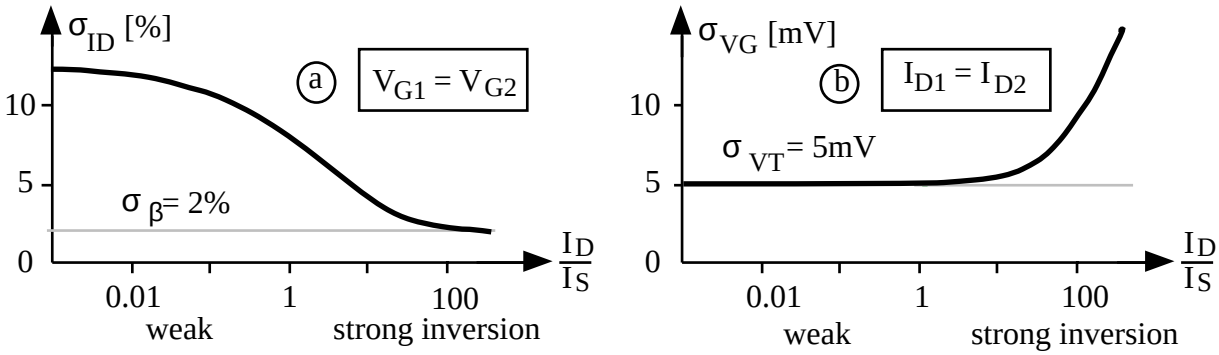


Fig.6 : Matching of drain currents (a, for same gate voltage) and gate voltages (b, for same drain current) as functions of the current levels of two saturated transistors.

It can be seen that weak inversion results in a very bad mismatch σ_{ID} of the drain currents but a minimum mismatch σ_{VG} of the gate voltages, equal to the threshold mismatch σ_{VT}. Operation deep into strong inversion is needed to reduce σ_{ID} to the minimum possible value σ_β.

The mismatch of bipolar transistors or of bipolar operated MOS transistors is that of their specific currents according to (10). The resulting mismatches of ∂V_{BE} and $\partial I_C/I_C$ do not depend on the collector current level and are much smaller than those of a MOS transistor, thanks to the much lower influence of surface effects.

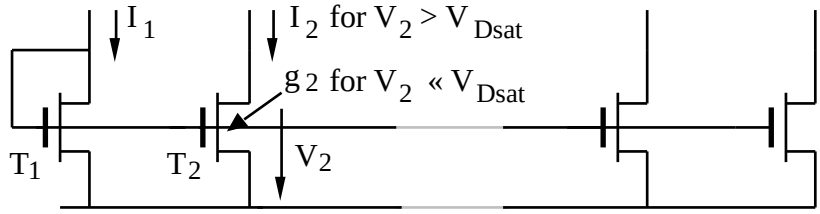


Fig. 7 : Multiple current mirror or current-controlled conductance.

One of the most useful elementary combinations of transistors is the current mirror, shown by Fig.7, which provides a *weighted copy* I_2 of its input current I_1 according to

$$I_2 = \frac{I_{S2}}{I_{S1}}\ I_1 \qquad (26)$$

provided the output voltage V_2 is above the saturation voltage V_{Dsat} given by (11) or (12). The copy I_2 is approximately equal to I_1 if the two transistors are identical (I_{S1}=I_{S2}). The imprecision due to mismatch is worst in weak inversion as illustrated in Fig.6a. Many copies can be obtained from the same mirror by repeating the output transistor T_2. A current ratio m/n is best obtained by using m+n identical transistors and by connecting in parallel n and m transistors to replace T_1 and T_2 respectively. Any value of the ratio can be obtained by using transistors of different width W and/or length L, but the precision is then reduced by a degraded matching.

For V_2 « V_{Dsat}, T_2 (and any repetition of it) behaves as a *current-controlled conductance* g_2. From (15) and (13) or (14):

$$g_2 = \frac{I_1}{I_{S1}} \cdot \frac{I_{S2}}{U_T} \qquad \textit{in weak inversion} \qquad (27)$$

$$g_2 = \sqrt{\frac{I_1}{I_{S1}}} \cdot \frac{I_{S2}}{U_T} \qquad \textit{in strong inversion} \qquad (28)$$

Current memorization (or storage) can be obtained by adding transistor T_d as shown in Fig.8a. The gate capacitance C (augmented if necessary with some additional capacitor) is charged through T_d at the value of gate voltage V corresponding to I_1. If I_1 is removed, T_d is blocked without changing the charge on C and the weighted copy I_2 is thus maintained. It can be reset to zero by means of a switch discharging C. Current *sampling*, *holding* and *weighting* is provided by the clocked scheme of Fig.8b, which is the basic building block of switched-current filters [10].

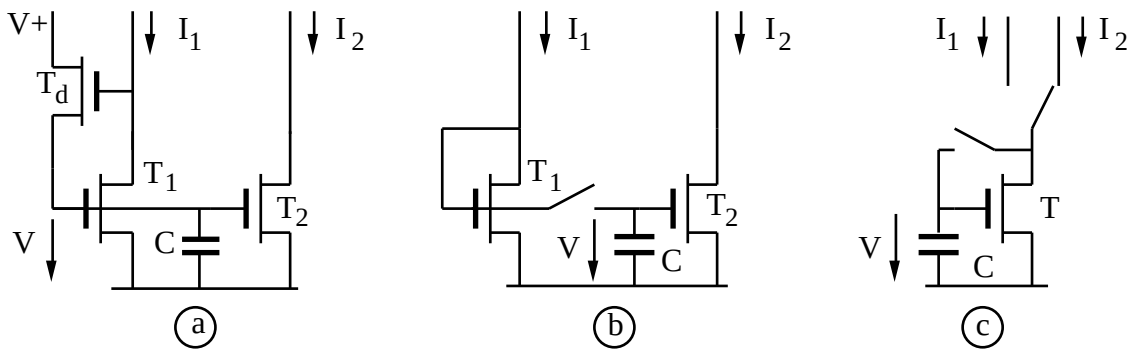


Fig.8 : Current memorization : peak value (a), sample-and-hold (b), precise replica with a single transistor (c).

The precision of the stored current can be drastically increased by using the same transistor T sequentially as the input and output device of the mirror, as shown in Fig.8c [11]. Several such *dynamic mirrors* (or *current copiers*) can be combined and clocked sequentially to provide continuous currents that are precisely weighted by integer numbers [12].

Combinations of n-channel and p-channel transistors can be used to carry out the addition or subtraction of replicas of two currents. Any attempt to impose a negative input current blocks the two transistors and results in a zero output current. This can be exploited to implement special functions such as min(A,B) and max(A,B) shown in Fig.9 [3].

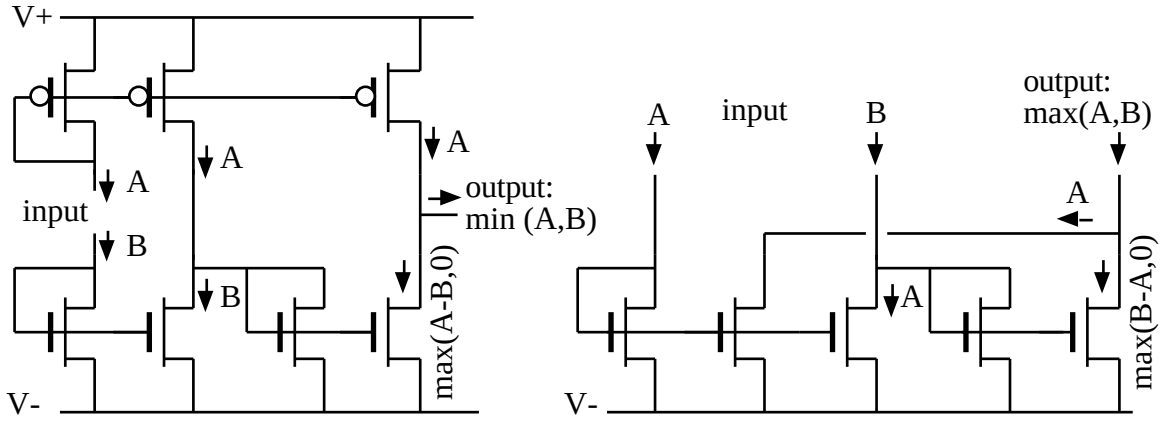


Fig.9 : Combination of current mirrors for computation of min(A,B) and max(A,B).

Another most important elementary combination of transistors is the well-known differential pair represented in Fig. 10 with its transfer characteristics for saturated transistors.

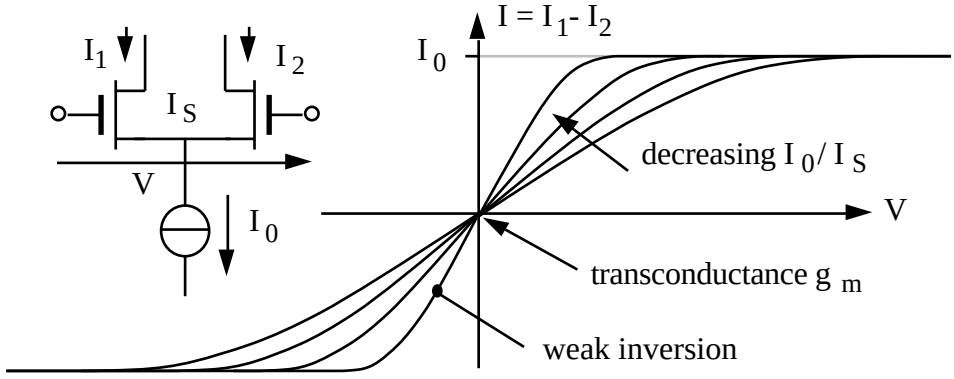


Fig.10 : Differential pair and transfer characteristics in saturation.

The constant current source I_0 is progressively steered from one to the other transistor by the differential input voltage V. For $I_0 \ll 2I_S$, the transistors are in weak inversion. By using the normalized voltage

$$v = \frac{V}{2nU_T} \tag{29}$$

the transfer characteristics (for saturated transistors) can be expressed from (3) as

$$I = I_0 \tanh(v) \quad I_0.v \ (\text{for } |v| \ll 1) \tag{30}$$

For small values of V, the differential pair in weak inversion thus provides *multiplication* of this input voltage by the bias current I_0. The output current saturates at the value $\pm I_0$ for $|v| \gg 1$. Weak inversion offers thus a good approximation of the *signum function* for $|v| \gg 1$.

For $I_0 \gg 2I_S$, the transistors are in strong inversion and the transfer characteristics derived from (4) are

$$I = v\sqrt{I_0 I_S}\ \sqrt{2 - \frac{I_S}{I_0} v^2} \quad (\text{for } v^2 < \frac{I_0}{I_S}) \qquad v\sqrt{2I_0 I_S} \quad (\text{for } v^2 \ll \frac{I_0}{I_S}) \tag{31}$$

The differential pair in strong inversion allows the saturating value of V to be increased by increasing the bias current. It provides multiplication of small values of V by the square root of the bias current I_0 .

The slope of the transfer characteristics for $I_1 \quad I_2$ is the transconductance g_m of the differential pair. It is given by (13) or (14) with $I = I_0/2$. The transconductance of the differential pair is thus controllable by the bias current I_0.

Another basic combination of transistors is the translinear loop [13], shown in Fig.11 for bipolar transistors.

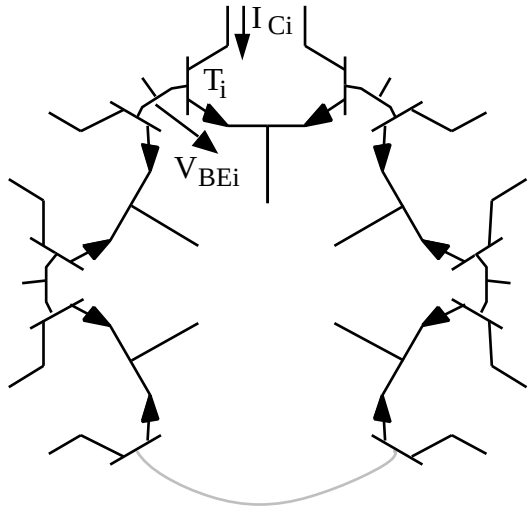


Fig. 11 : Translinear loop.

The base-emitter junctions of an even number of transistors are connected in series, half of them in each direction (clockwise cw and counter-clockwise ccw) so that

$$\sum_{cw} V_{BEi} = \sum_{ccw} V_{BEi} \tag{32}$$

Expressing V_{BEi} as a function of the collector current I_{Ci} according to (10) yields:

$$\frac{\prod_{cw} I_{Ci}}{\prod_{ccw} I_{Ci}} = \frac{\prod_{cw} I_{Sbi}}{\prod_{ccw} I_{Sbi}} = \lambda \tag{33}$$

where the loop factor $\lambda = 1$ for perfectly matched identical transistors. This very general result does not depend on temperature nor on the current gain of the transistors. The transistors can be in any sequence inside a loop, and several loops may share some common transistors. Translinear circuits are based on the availability of an exponential voltage-to-current law. They are best implemented by bipolar transistors or bipolar operated MOS, in order to obtain a reasonably good precision. They can also be implemented by means of saturated MOS

transistors operated in weak inversion with V_S=0 (each source connected to a local well) according to (9). The loop factor λ will then be very inaccurate and variable with temperature, due to the large variation of V_{T0} from device to device.

Building blocks for local operations

A large variety of linear and nonlinear building blocks can be obtained by exploiting in an opportunistic manner the features offered by transistors and their elementary combinations.

The difference operation $I = I_1 - I_2$ on the currents through the two transistors of a differential pair can be easily carried out by means of one or several current mirrors. An operational transconductance amplifier (OTA) is then obtained [5], which is the most fundamental building block of traditional linear analog processing. It may be combined with switches and capacitors to realize a wide range of time-sampled switched-capacitor circuits, including very precise linear filters [14]. The controlled transconductance g_m of the OTA may be combined with capacitors to implement all kinds of continuous-time linear filters [15].

The simple combination of current mirrors shown in Fig.12 realizes a current conveyor. By symmetry with the real ground, a virtual ground node N is obtained, where incoming currents can be summed without changing the node potential. The sum I is available as an output current source. This circuit can be used to aggregate several voltage signals V_i weighted by the $I_i(V_i)$ characteristics of dipoles, as shown in the same figure. Using a transistor in strong inversion for each dipole, the circuit implements the sum of n products V_i (V_{Gi}-V_{T0}) with just 5+n transistors.

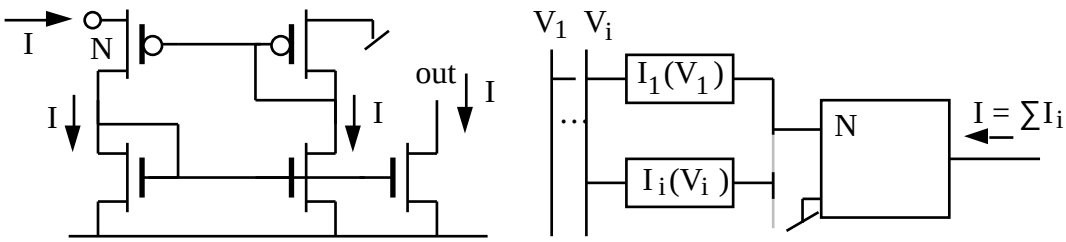


Fig.12 : Current conveyor and its application to the aggregation of signals.

The multiplication of voltage signals may also be obtained by combining 3 differential pairs operated in strong inversion [16], whereas the multiplication/division of current signals is best achieved by using translinear circuits. A simple example is illustrated in Fig.13.

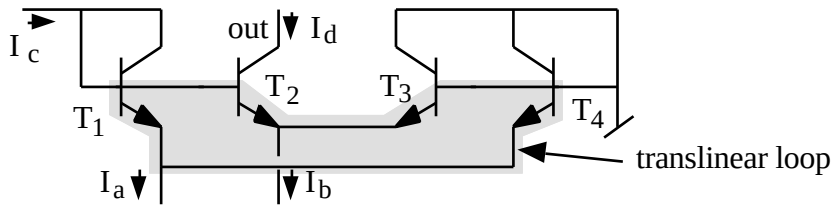


Fig.13 : Multiplication/division of currents by a translinear loop.

If all transistors are identical (loop factor λ =1), then (32) and (33) yield

$$I_{C1}I_{C3} = I_{C2}I_{C4} \qquad (34)$$

Thus, by inspection of Fig.13 and neglecting the base currents while keeping the non-unity value of the common-base current gain α:

$$I_c\,(\alpha I_b - I_d) = I_d\,(\alpha I_a - I_c) \qquad I_d = \frac{I_c\,I_b}{I_a} \tag{35}$$

This result is independent of α and valid as long as $0 < I_c < I_a$.

Nonlinear manipulations on currents can also be done by exploiting the square-law behavior of MOS transistors in strong inversion [17], as illustrated by the squarer/divider of Fig.14.

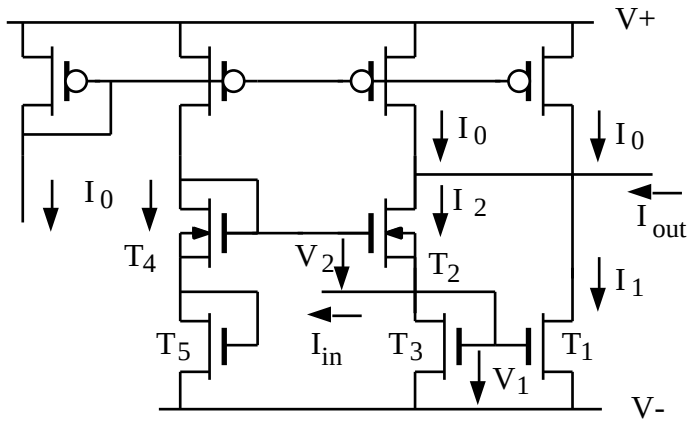


Fig.14 : Current squarer/divider using MOS transistors in strong inversion.

For saturated transistors in strong inversion with V_S=0 (sources connected to local wells), equation (7) may be rewritten

$$I_D = \frac{\beta}{2n}\,(V_G - V_{T0})^2 \tag{36}$$

All transistors of the same type have the same size and thus the same value of β for perfect matching, and n is assumed to be constant. The gate voltages of T_1 and T_2 are V_1 and V_2 respectively, whereas those of T_4 and T_5 are $(V_1+V_2)/2$. The input current I_{in} is I_2-I_1 since T_3 and T_1 have the same drain current I_1. The application of (36) to transistors T_1, T_2 and T_4 yields

$$I_1 + I_2 = 2I_0 + \frac{I_{in}^2}{8I_0} \qquad (\text{for } I_{in}^2 < 16I_0^2) \tag{37}$$

The term $2I_0$ is then removed from the output current I_{out} by the pair of identical p-channel transistors.

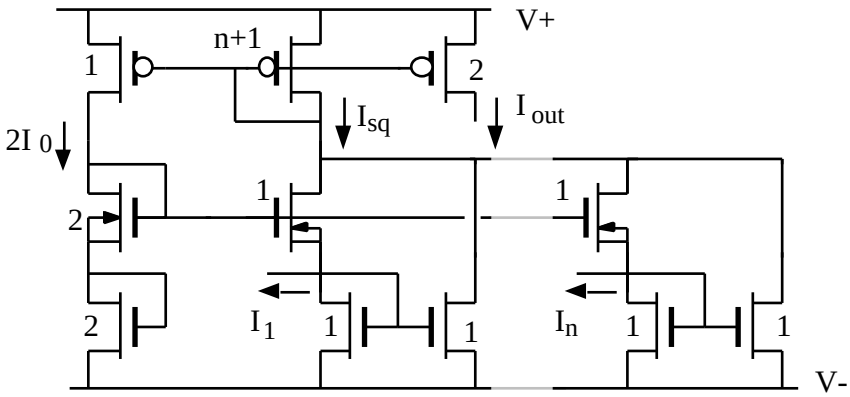


Fig.15 : Computation of the length of a n-dimensional vector of currents.

The same principle can be used to compute the length of an n-dimensional vector with current components I_i by modifying the circuit as shown in Fig.15 [18].

The configuration of transistors T_1, T_2 and T_3 of Fig.14 is repeated n times with the n input currents I_i to produce the total current I_{sq}. From (37):

$$I_{sq} = 2nI_0 + \frac{1}{8I_0}\sum_{i=1}^{n} I_i^2 \tag{38}$$

Each transistor in Fig.15 is a parallel combination of the number of unit transistors indicated, with a total of 4n+8 units.The current I_{sq} is used to autobias the structure by imposing, through the weighted p-channel current mirrors:

$$I_0 = \frac{I_{out}}{4} = \frac{I_{sq}}{2(n+1)} \tag{39}$$

which yields, by introducing (38):

$$I_{out} = \sqrt{\sum_{i=1}^{n} I_i^2} \tag{40}$$

Since each $|Ii| < I_{out} = 4I_0$, the result is valid as long as the significant input components are large enough to keep the corresponding input cell in strong inversion. For n=1, the circuit provides the absolute value of the single input current. If each $I_i=I_{ai}-I_{bi}$, the circuit computes the euclidean distance between two points at coordinates I_{ai} and I_{bi} in the n-dimensional space.

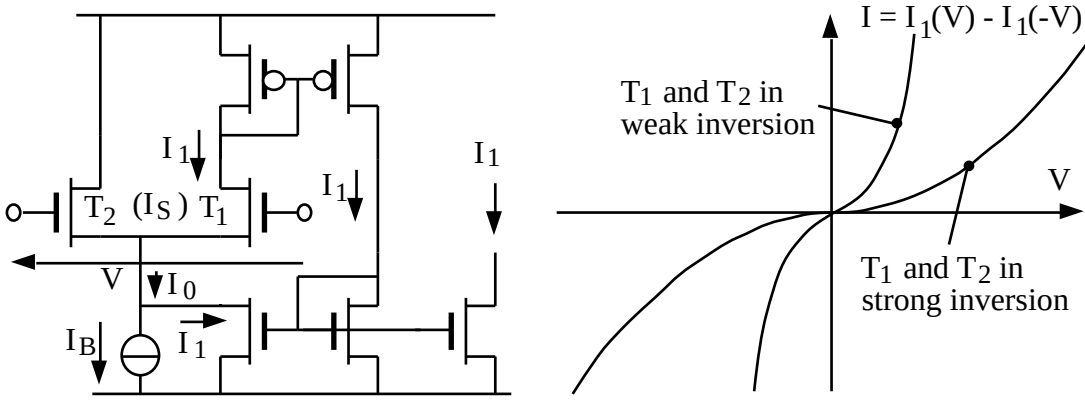


Fig.16 : Implementation of expansive nonlinear functions.

Expansive nonlinear functions may be implemented by re-injecting one of the output currents of a differential pair into its bias current as shown in Fig.16 [19]. Using two such circuits in a push-pull combination provides, by application of (9) or (7) [20]:

$$I = 2I_B \sinh(2v) \qquad \text{for } T_1 \text{ and } T_2 \text{ in } \textit{weak inversion} \tag{41}$$

$$I = I_S \, \mathrm{sgn}(v) \, v^2 \quad \text{for } |I| \gg I_B \text{ and } T_1 \text{ and } T_2 \text{ in } \textit{strong inversion} \tag{42}$$

where v is the normalized value of input voltage V defined by (29).

The elevation to any power k of a signal represented by a current can be effected by one of the circuits shown in Fig.17 which are based on a generalization of the translinear principle [21]. I_X is the input current, I_Y the output current and I_R is a common reference current.

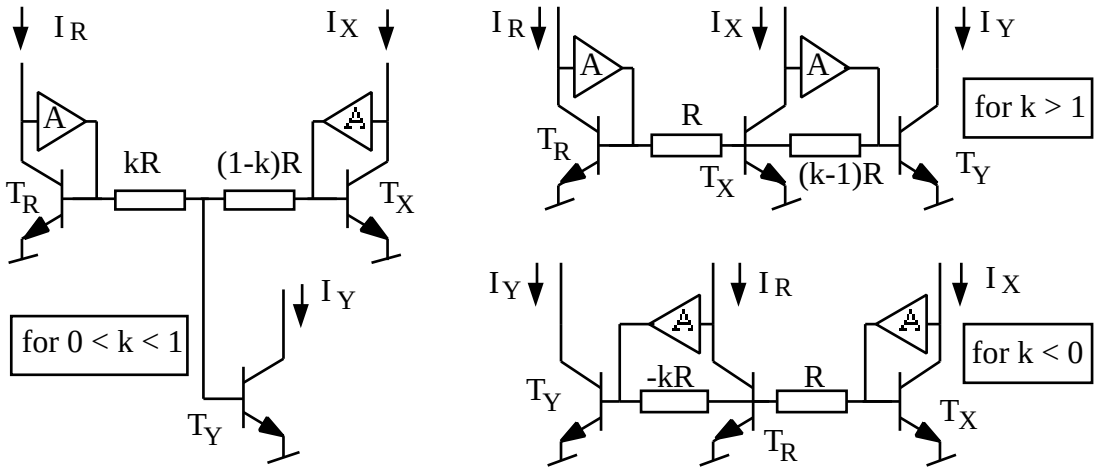


Fig.17 : Circuit implementing $y = x^k$ for signals represented by currents I_X and I_Y.

If the voltage drop due to the base current of the central transistor is made negligible by choosing a sufficiently low value for the resistors, then for each of these circuits:

$$V_{BEY} - V_{BER} = k\,(V_{BEX} - V_{BER}) \tag{43}$$

If all transistors are identical, application of (10) yields:

$$\frac{I_Y}{I_R} = \left(\frac{I_X}{I_R}\right)^k \tag{44}$$

Each amplifier A must be able to source or sink the current flowing through the resistors, and can be realized by a simple combination of MOS transistors. The resistors can be implemented with the gate polysilicon layer. The exponent k can be made adjustable by providing taps along the resistors to which the output transistor T_Y can be selectively connected by switches. The transistors can be bipolar-operated MOS devices. The function of the circuit is not affected by the current that flows to the substrate in these particular devices. They could also be replaced by MOS transistors in weak inversion, but the precision achieved would be poor.

The storage of information in analog VLSI circuits is a very difficult problem for which many solutions have been worked out [22].

Fixed values can be stored in analog ROM cells as W/L ratios of transistors or resistors, or as areas of capacitors. One could also consider external storage by an optical pattern of gray scales that is projected on the chip and locally transformed into currents by light sensors.

Short term storage can be obtained by sampling and holding a voltage on a capacitor. The storage time is then limited by the leakage current of the reverse biased junctions associated with the transistor that implements the switch. This current can be minimized by minimizing the voltage across the junction, as illustrated by the current-storage cell shown in Fig.18 [23].

To store current I, the transistor T_S is switched on. The difference of currents I-I' is integrated on capacitor C until equilibrium is reached, for the value of voltage V across C giving I'=I. The switch T_S can then be blocked without changing V (except for some parasitic charge injection) and the current I is still available from the output transistor T_O. Since the voltage gain of the OTA is very large, the voltage across the leaking junction at the critical node N is limited to a small offset V_{os}.

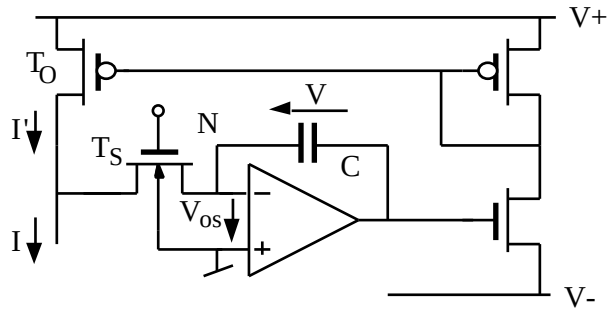


Fig.18 : Low-leakage storage of current I as a voltage V across C.

Long term storage can be achieved by refreshing the voltage of the storage capacitor. This technique requires amplitude quantization. Direct refresh to predefined analog levels can be obtained locally with a very small number of components [22], but the storage time is still limited by the small noise margin equal to one half of the quantization step. The capacitor voltage can also be refreshed from a digital RAM combined with A/D and D/A converters. However, these converters require a large chip area that is usually not compatible with the density looked for in analog implementations. They may be centralized and time-shared to save area, but the necessary high-speed multiplexing of analog busses is difficult to implement.

The preferred solution for long-term storage is to use a floating gate transistor with its gate electrode completely wrapped inside silicon dioxide. The leakage is then so small that the storage time is extended to many years, but the problem is then to find a way to change the charge stored on this gate in order to change the content of the memory.

One possibility is to give the carriers enough energy to pass the high energy barrier of the oxide. This energy can be provided by avalanching a reverse-biased junction, which works well for electrons whereas hole injection efficiency is 3 to 4 orders of magnitude lower. Hole injection can be enhanced by a field created by capacitively pulling the floating gate to a negative potential. As an alternative method, the required energy can be provided by illuminating selected areas with UV light [22]. The advantage of these techniques is their full compatibility with standard CMOS VLSI processes.

The standard technique employed in modern E^2PROMs uses the Fowler-Nordheim mechanism [24], which is a field-aided tunneling of electrons across the oxide barrier. Special technologies, with additional process steps, are needed to limit the control voltage below 20 volts and to ensure reliable operation. New circuit techniques are being developed to use these kinds of processes for analog storage [25],[26],[22].

Collective computation

Very fast analog processing of a large number of signals is made possible by circuits capable of truly parallel collective computation. A simple example is given by the normalization circuit shown in Fig.19 [27].

All cells have the same voltage difference between the emitters of their input and output transistors. Thus, according to the exponential law (10) and for all transistors identical, each cell provides the same current gain (or attenuation)

$$\frac{I_{out\ i}}{I_{in\ i}} = e^{\frac{V}{U_T}} \qquad (45)$$

Since the total output current is forced to be I_{tot} by an external current source, V is collectively adjusted by the cells to adapt the common gain to this constraint. This circuit is

very useful to adapt the levels of a set of signals to the range of best operation of the subsequent processing circuits.

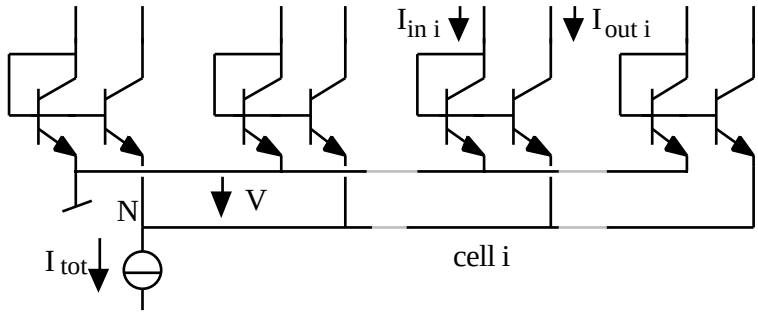


Fig.19 : Normalization of a set of signals.

All cells have the same voltage difference between the emitters of their input and output transistors. Thus, according to the exponential law (10) and for all transistors identical, each cell provides the same current gain (or attenuation)

$$\frac{I_{out\ i}}{I_{in\ i}} = e^{\frac{V}{U_T}} \qquad (45)$$

Since the total output current is forced to be I_{tot} by an external current source, V is collectively adjusted by the cells to adapt the common gain to this constraint. This circuit is very useful to adapt the levels of a set of signals to the range of best operation of the subsequent processing circuits.

The gain may alternatively be imposed either by imposing directly V (which must then be proportional to the absolute temperature T), or by imposing the ratio of the sums of input and output currents. In the latter case, the input and output nodes of one cell must be grounded to ensure proper biasing, and the circuit is an extrapolation of the multiplier/divider of Fig.13. It can be efficiently used to adapt the parameters of all the cells simultaneously in a cellular network.

Another classical example of efficient collective computation is the winner-take-all circuit shown in Fig.20 [28].

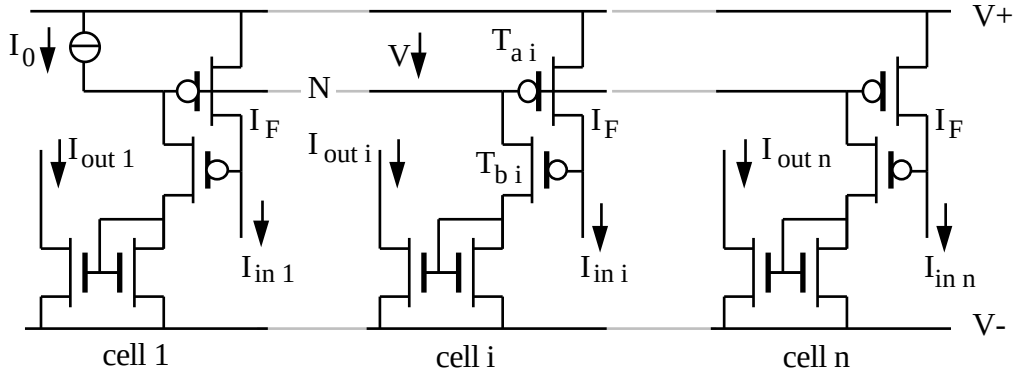


Fig.20 : Winner-take-all circuit.

Transistors T_a of all the n cells have the same gate voltage $V_G = V$, and therefore the same forward (or saturation) current I_F, according to (2). As long as the input current $I_{in\ i}$ of a cell i is larger than I_F, T_{ai} remains saturated and the current through T_{bi} keeps pulling down the common gate node N. This results in increasing the common gate voltage V and therefore the common value of I_F. When the value of I_F reaches that of $I_{in\ i}$, the transistor T_{ai} of this cell

leaves saturation. T_{bi} is thus blocked and stops pulling down the common node N. This occurs successively in cells having increasing values of input current, until $I_F = I_{inmax}$. Transistor T_b of the corresponding (winning) cell takes all the bias current, which is mirrored to the output of the cell.

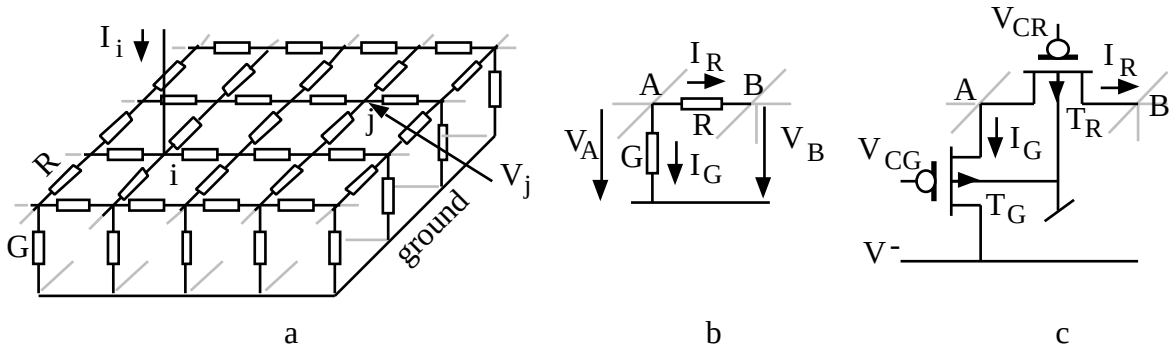
Since the comparison from cell to cell is effected via the the common gate voltage V, the precision of the decision is limited by the mismatch of drain currents according to Fig.6a. The comparison is therefore inaccurate when the maximum input current is very low.

The weighted average of n voltage sources can easily be obtained by connecting each source V_i to a common output node through a conductance G_i. The output voltage V_{out} at the common node is then given by

$$V_{out} = \sum_i G_i V_i \Big/ \sum_i G_i \qquad (46)$$

Each conductance may be implemented by an OTA connected in unity gain configuration [29]. The voltage sources do not have to provide any current and the value of G_i is then that of the transconductance g_m of the OTA. It can be controlled by the bias current I_{0i} of the OTA, as explained by Fig.10. If the input pair is operated in weak inversion, then $G_i \sim I_{0i}$; if V_i represents the spatial position of the current sources I_{0i} along an axis, then V_{out} represents the center of gravity of the current sources (or their median if most of the OTAs are saturated) . This property can be exploited to compute the center of gravity of a 2-dimensional light spot [30].

Local averaging, in which the contributions of spatially distant sources are reduced, can be obtained by the resistive diffusion network represented in Fig.21a for the 2-dimensional case.



a b c

Fig.21 : 2-dimensional resistive diffusion network.

The contribution to voltage V_j of an input current I_i is maximum locally (i=j) and vanishes with the distance with a rate that can be approximated by

$$\frac{V_j}{I_i} \sim r^{-1/2} \, e^{\frac{-r}{L}} \qquad (47)$$

where $$L = 1 \Big/ \sqrt{RG} \qquad (48)$$

and r is the distance between nodes i and j measured in grid units [29]. The network could be implemented by means of real resistors, but it would not allow the diffusion distance L to be controlled electrically. A better solution is obtained by using transistors operated as controlled conductances and/or transconductance amplifiers [29]. The network is then generally nonlinear for large signals and this behavior may be opportunistically exploited to limit the effect of excessive input contributions.

A linear network can be obtained by using the concept of pseudo-conductance of transistors in weak inversion, defined by relations (20) to (22), as illustrated by the comparison of subnetworks b and c in Fig.21. Conductances G and resistances R are replaced by transistors T_G and T_R that implement pseudo-conductances G^* and pseudo-resistances R^*, which are controlled respectively by voltages V_{CG} and V_{CR}. According to (22) and (48), the diffusion length is then given by

$$L = 1/\sqrt{R^*G^*} = e^{\frac{V_{CG} - V_{CR}}{2nU_T}} \qquad (49)$$

The pseudo-ground is obtained by choosing the voltage V- sufficiently negative to make the reverse component of current through T_G negligible. The network is linear with respect to currents as long as all transistors remain in weak inversion. As output, the voltage V_A at node A is replaced by the current $I_G = GV_A$ flowing through transistor T_G. This output current can be extracted by inserting the input transistor of an n-channel current mirror between T_G and V^-.

The resistive network of Fig.21a has been used to realize a VLSI silicon retina formed of a large hexagonal array of identical cells [31]. This retina provides edge enhancement in an image projected on the chip by 2-dimensional high-pass spatial filtering. A new very dense version of a retina cell based on the pseudo-conductance concept is shown in Fig.22. This circuit resembles a previously published retina based on "diffusor" transistors [43], but it does not require matching between p and n transistors, it avoids the important residual nonlinearity due to the combination of gate- and source-modulation, and it exploits the concept of pseudo-ground to extract the currents.

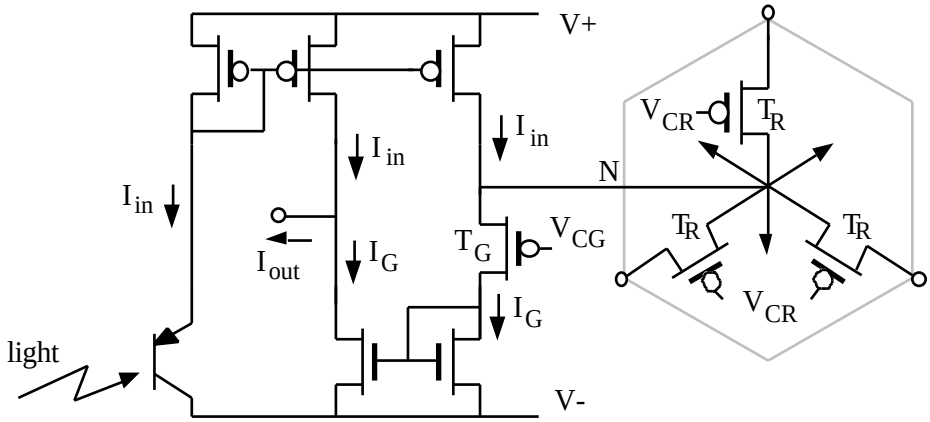


Fig.22 : Cell of an artificial retina based on the pseudo-conductance network T_R-T_G.

The local light sensor implemented by means of an open-base vertical pnp bipolar transistor available in the n-well CMOS process produces an input current I_{in}. This current is mirrored into the local node N of a hexagonal network based on pseudo-conductances implemented by T_R and T_G according to Fig. 21c. The spatially low-pass filtered signal I_G is extracted by the n-channel mirror and subtracted from another copy of I_{in} to produce the high-pass filtered output signal I_{out}. Other applications of linear resistive networks are the computation of first and second order moments to extract the axis of minimum inertia of an image [32], and the guidance of a mobile robot in presence of obstacles [33].

The linear network of figure 21 can also be very efficiently exploited to implement the weighting function of the excitatory-inhibitory lateral interconnections that provide the necessary collective behavior in a Kohonen self-organizing feature map [34]. By using the pseudo-conductance concept, the size of the "bubble" of activity can then adjusted by the diffusion distance L according to (49). In another form of implementation of the Kohonen map, the bubble is created by selecting the most active cell by a winner-take-all and by

activating all the cells within a certain radius. The nonlinear diffusion network shown in Fig.23 is then preferred [35][23].

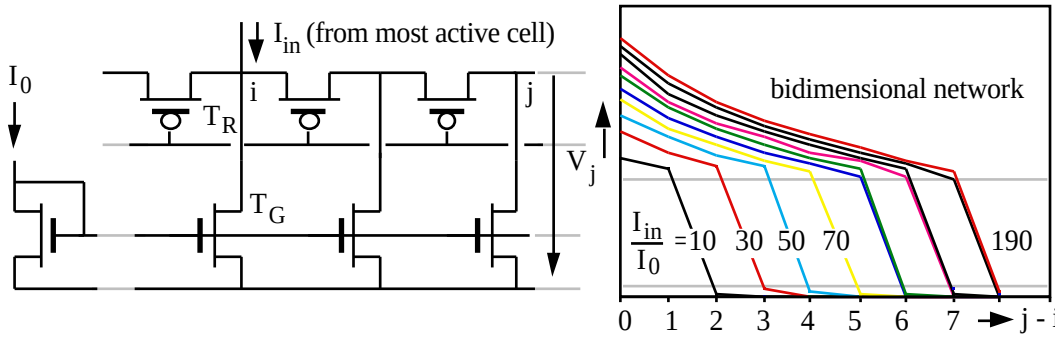


Fig.23 : Nonlinear diffusion network.

The conductances G are replaced by n-channel current sources T_G and resistances R by p-channel transistors T_R with a common gate voltage. The conductance of T_R increases with the node voltage V_j, therefore this voltage decreases more linearly with the distance j-i than in the case of the linear network. This feature may be exploited to adapt the learning gain to the distance. Furthermore, for the 2-dimensional case corresponding to the calculated curves shown in the figure, V_j decreases abruptly when the distance exceeds the radius of a circle of area I_{in}/I_0. The size of the bubble of activity can thus be easily controlled by I_{in}/I_0.

The collective processing of a large number of data in massively parallel circuits requires a dense network of communication between cells, that is not naturally available in intrinsically 2-dimensional VLSI implementations. Furthermore, very large perception systems will have to be split into several VLSI chips, on which successive layers of processing may be implemented, with the result of a massive flow of signals from chip to chip. These problems of intrachip and interchip communication can be addressed by several complementary approaches.

As shown by the examples of Fig. 19 and 20, some collective computation is already possible by exchanging information on a single node corresponding to a single wire across the whole chip. Another interesting example is that of the collective evaluation of the two components of translation of a picture by an array of cells communicating via only two wires [36] .

The need for long distance communication can be avoided by using cellular networks, which are arrays of identical analog cells communicating directly with other cells within a limited distance only, possibly just with their nearest neighbors. A collective computation is still carried out by communication from neighbor to neighbor. The simple resistive network of Fig.21 and its application to the artificial retina of Fig.22 belong to this category. The cellular neural network (CNN) is another example. In this case, each cell basically amounts to a leaky integrator and a limiting nonlinearity [37]. It has been shown that such networks are capable of a large variety of 2-dimensional processing. Their behavior is controlled by an analog template that characterizes the action of each cell on its different neighbors [38]. This template is common to all the cells and can thus be communicated to each of them by a limited number of programming wires.

In more general situations, the interconnection between cells can be organized in a hierarchical manner. The interconnections are very dense for short distances, but their number is drastically reduced between distant cells. This appears to be the "strategy" used in the brain. Collective decisions are then obtained progressively from very local groups of cells to the whole system.

The communication over long distance or from chip to chip can exploit the very large bandwidth offered by a single wire (with respect to that possible with its biological

counterpart, the axon), to multiplex a number of analog signals. A standard row-column scanning scheme can be used [39]. Its main drawbacks are that it requires time sampling, which destroys the time continuity possible with analog processing, and that the scanning clock eliminates the phase information by creating synchronized pieces of information.

A better approach is possible by using priority schemes, that can be based on the activity of the cells. Such a scheme is applicable whenever the result of each layer of processing is the activation of a limited number of cells. The equivalent bandwidth of communication to the next layer is thus maximized by giving each cell access to the communication channel with a degree of priority proportional to its level of activity. A very natural way of implementing such a scheme consists in representing the level of activity by the frequency of very short asynchonous pulses, and in giving each cell access to a single bus of n parallel wires to transmit its own n-bit location code each time it produces a pulse [40].The simultaneous access to the bus by more than one cell can be arbitrated or not. In the latter case, it results either in false codes, the number of which can be made negligible by introducing some redundancy, or in non-recognized codes which just reduce the overall gain of the transmission [41].

Conclusion.

Analog circuits have a limited precision, but they allow all the features of the transistors to be exploited opportunistically in order to build very dense processing cells. Neither time-sampling nor amplitude quantization is needed, and the physical variable representing each signal can be selected locally to minimize the circuit complexity. Analog CMOS VLSI therefore constitutes an ideal medium for the implementation of perception systems, which do not require high precision, but which are based on the collective processing of a continuous flow of data by massively parallel systems. Care must be taken to maintain an acceptable level of precision by resorting to sound design techniques, based on locally matched devices, to minimize or even eliminate the dependency on process parameters.

The overall low-frequency behavior of an MOS transistor, in all possible bias situations and from very low to high current, can be captured in a simple model based on just three independent parameters. Inspection of this model shows that a single transistor can readily be used to implement a large variety of elementary processing functions. It can even be used as a bipolar transistor to provide additional features. Basic combinations of a few transistors such as the current mirror, the differential pair and the translinear loop drastically enrich the catalogue of possibilities. As a result, virtually any desired local operation can be implemented by a very limited number of transistors, as has been illustrated by a variety of examples. The storage of information is a difficult problem in analog circuits, but several complementary solutions are available already.

Collective computation is made possible by the feasibility of massively parallel analog systems, but it requires means for communicating from cell to cell. Simple circuits, with just a few transistors per cell, are available to implement some important collective operations by communicating via one or a very limited number of wires. Some more general collective computation can be achieved by cellular networks. Long distance communication can use a limited number of interconnections to multiplex several analog signals, either by row-column scanning, or by using priority schemes which better exploit the nature of the information processed in massively parallel systems.

Analog circuit designers have been continuously creative in spite of the increasingly insistent predictions that their future may be very severely limited by the conquering progression of digital systems. There is no doubt that the prospect of fully analog VLSI for very advanced perception systems will further stimulate this creativity and generate a multitude of innovative ideas.

References

1 Workshop on Future Directions in Circuits, Systems, and Signal Processing, May 4-5, 1990, New Orleans LA.

2 E. Vittoz, "Future trends of analog in the VLSI environment", invited session WEAM-9, ISCAS'90, New Orleans , May 2, 1990.

3 T. Yamakawa and T. Miki, "The current mode fuzzy logic integrated circuit fabrication by the standard CMOS process", IEEE Trans. on Computers, vol.C-35, pp.161-167, February 1986.

4 O. Landolt, "Efficient analog CMOS implementation of fuzzy rules by direct synthesis of multidimensional fuzzy subspaces", submitted to Fuzz'IEEE '93, San Francisco.

5 E. Vittoz, "The design of high-performance analog circuits on digital CMOS chips", IEEE J. Solid-State Circuits, vol. SC-20, pp. 657-665, June 1985.

6 E. Vittoz, "Micropower Techniques", in *Design of VLSI Circuits for Telecommunication and Signal Processing*, Editors J.Franca and Y.Tsividis, Prentice Hall, to be published.

7 Y.P. Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987.

8 E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operation", IEEE J. Solid-State Circuits, vol. SC-12, pp. 224-231, June 1977.

9 E. Vittoz, "MOS transistors operated in the lateral bipolar mode and their applications in CMOS technology", IEEE J. Solid-State Circuits, vol. SC-18, pp. 273-279, June 1983.

10 J.B. Hughes, “Switched current filters”, in *Analog IC Design: the current-mode approach*, C. Toumazou, F.J. Lidgey and D.G. Haigh, Editors, Peter Peregrinus, London 1990.

11 E. Vittoz, "Dynamic Analog Techniques", in *Design of VLSI Circuits for Telecommunication and Signal Processing*, Editors J.Franca and Y.Tsividis, Prentice Hall, to be published.

12 E. Vittoz and G. Wegmann, "Dynamic current mirrors", in *Analog IC Design: the current-mode approach*, Edit. C. Toumazou, F. Lidgey and D. Haigh, Peter Peregrinus, London, 1990.

13 B. Gilbert, "Translinear circuits: a proposed classification", Electron. Letters, vol.11, p.14, 1975.

14 G.C. Temes, *Integrated Analog Filters*, IEEE Press, 1987.

15 R. Schaumann and M.A. Tan, “Continuous-time filters”, in *Analog IC Design: the current-mode approach*, Edit. C. Toumazou, F. Lidgey and D. Haigh, Peter Peregrinus, London, 1990.

16 E. Vittoz, "Analog VLSI implementation of neural networks", proc. Journées d'Electronique EPFL on Artificial Neural Nets, EPF-Lausanne, Oct-10-12, 1989, pp.224-250.

17 K. Bult and H. Wallinga, "A class of analog CMOS circuits based on the square-law characteristic of an MOS transistor in saturation", IEEE J. Solid-State Circuits, vol. SC-22, No 3, pp.357-365, June 1987.

18 O. Landolt, E. Vittoz and P. Heim, "CMOS selfbiased euclidean distance computing circuit with high dynamic range", Electronics Letters, 13th Febr.1992, vol.28, No 4.

19 M. Degrauwe, J. Rijmenants, E. Vittoz and H. De Man, "Adaptive biasing CMOS amplifiers", IEEE J. Solid-State Circuits, vol. SC-17, pp. 522-528, June 1982.

20 E. Vittoz and X. Arreguit, "CMOS integration of Hérault-Jutten cells for separation of sources", *Analog VLSI Implementation of Neural Networks*, Ed. C.Mead and M. Ismail, Kluwer Academic Publ., Norwell, 1989.

21 X. Arreguit, E. Vittoz and M. Merz, "Precision compressor gain controller in CMOS technology", IEEE J. Solid-State Circuits, vol. SC-22, pp. 442-445, June 1987.

22 E.Vittoz, H.Oguey, M.A.Maher, O.Nys, E.Dijkstra and M.Chevroulet, "Analog storage of adjustable synaptic weights" in *Introduction to VLSI-Design of Neural Networks*, Editor. U.Ramacher, Kluwer Academic Publ., 1991.

23 O. Landolt, "An analog CMOS implementation of a Kohonen network with learning capability", 3rd Int. Workshop on VLSI for Neural Networks and Artificial Intelligence, Oxford, 2-4 September 1992.

24 M. Lenzlinger and E. Snow, "Fowler-Nordheim tunneling into thermally grown SiO2", J. Appl. Phys., vol.40, p.278, 1969.

25 E. Säckinger and W. Guggenbuhl, "An analog trimming circuit based on a floating-gate device", IEEE J. Solid-State Circuits, vol. SC-23, No 6, PP.1437-1440, Dec. 1988.

26 M. Holler *et al*, "An electrically trainable artificial neural network (ETANN) with 10240 floating gate synapses", Proc. Int. Joint Conf. on Neural Networks, pp.II-191-196, Washington, June 1989.

27 B. Gilbert, "A monolithic 16-channel analog array normalizer", IEEE Journal of Solid-State Circuits, vol.SC-19, p.956, 1984.

28 J. Lazzaro *et al*, "Winner-take-all networks of order N complexity", in Proc. 1988 IEEE Conf. on Neural Information Processing - Natural and Synthetic, Denver, 1988.

29 C.A. Mead, *Analog VLSI and neural systems*, Addison-Wesley, Reading, 1989.

30 S. DeWeerth and C.A. Mead, "A two-dimensional visual tracking array", *Proc. 1988 M.I.T. Conf. on VLSI, M.I.T. Press*, Cambridge, MA, pp.259-275.

31 C.A. Mead and M.A. Mahowald, "A silicon model of early visual processing", Neural Networks, vol.1, pp.91-97, 1988.

32 D. Standley and B. Horn, "Analog CMOS IC for object position and orientation", SPIE vol.1473 *Visual Information Processing: From Neurons to Chips* (1991), pp.194-201.

33 L. Tarassenko *et al*, "Real-time autonomous robot navigation using VLSI neural networks", in *Advances in Neural Information Processing Systems*, vol.3, R.P.Lippmann, J.E.Moody and D.S.Touretzky (editors), pp.422-428, Morgan Kaufmann, 1991.

34 E. Vittoz *et al.* "Analog VLSI implementation of a Kohonen Map", Proc. Journées d'Electronique on Artificial Neural Nets, EPFL, Lausanne, Oct. 1989, pp.292-301.

35 P. Heim, B. Hochet and E. Vittoz, "Generation of Learning Neighborhood in Kohonen Feature Maps by means if Simple Nonlinear network", Electronics Letters, vol.27, No 3,1991, pp.275-277.

36 J. Tanner and C.A. Mead, "A correlating optical motion detector", Proc. Conf. on Advanced Research in VLSI, January 23-25 1984, MIT, Cambridge MA.

37 L. Chua and L. Yang, "Cellular neural networks: theory", IEEE Trans. Circuits and Systems, vol.35, pp.1257-1272, Oct.1988.

38 L. Chua and L. Yang, "Cellular neural networks: applications", IEEE Trans. Circuits and Systems, vol.35, pp.1273-1290, Oct.1988.

39 C.A. Mead and T. Delbrück, "Scanners for visualizing activity of analog VLSI circuitry", CNS Memo 11, California Institute of Technology, June 27, 1991.

40 M.Mahowald, *VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function*, Ph.D.Thesis, California Institute of Technology, Pasadena, 1992.

41 A. Mortara and E.Vittoz, "A communication architecture tailored for analog VLSI artificial neural networks: intrinsic performance and limitations", submitted to IEEE Trans. on Neural Networks.

[42] E.Vittoz and X.Arreguit, "Linear networks based on transistors", Electronics Letters, vol.29, pp.297-299, 4th Febr. 1993

[43] K.W.Boahen and A.G.Andreou, "A contrast sensitive silicon retina with reciprocal synapses", Advances in Neural Information Processing Systems, Vol.4, pp.764-772, Morgan Kaufmann Publishers, San Mateo, 1992.