Analog VLSI Neural Networks

Microelectronics group DIBE

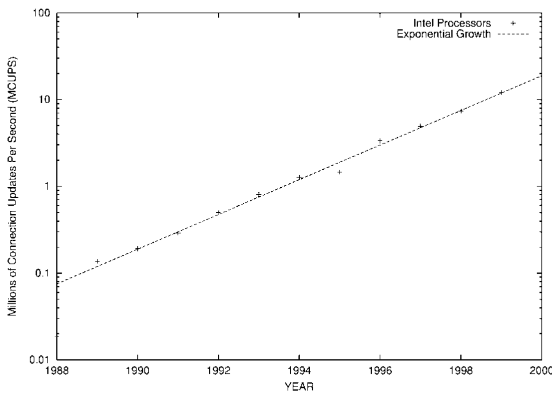
Maurizio Valle

Digital vs. analog VLSI implementations

	digital technology	*analog technology*
signal representation	numbers (symbol)	physical signals (e.g. voltages, currents, charge, etc.)
time	sampling	continuous/sampling
signal amplitude	quantized	continuous
signal regeneration	along path	degradation
resolution (S/N)	cheap and easy	area and power expensive
transistor mode of operation	switch mode	all modes
energy efficiency	low	high
area per processing element (i.e. computational density)	large	small
architecture	low degree of parallelism	high degree of parallelism
design and test	easy	difficult/expensive

Exponential growth of computing power for Neurocomputing

General Purpose Microprocessors



Signal representation in analog processing circuits

signals in an analog circuit are represented by physical variables, e.g. voltage V, current I, charge Q, frequency or time duration

- V: easy distribution of a signal but large stored energy (e.g. $CV^2/2$) into the node parasitic capacitance
- I: easy implementation of sum of signals but complicate distribution
- Q: requires time sampling, nice processing e.g. switched capacitor techniques
- Pulse frequency or time between pulses: dominant mode of signal representation for communication in biological nervous systems. Easy signal regeneration

Signal processing in analog processing circuits

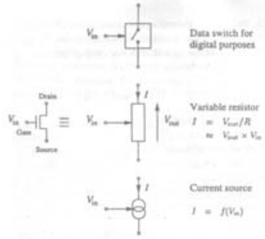
Primitives of computation arise from the physics of the computing devices.

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

a MOS transistor can provide many functions:

- switch;
- generation of square, square root, exponential and logarithmic functions;
- voltage controlled current source;
- voltage controlled conductance;
- analog multiplication of voltages;
- short term and long term storage;
- light sensor;
- etc.

The MOS transistor: modes of operation



- **switch mode**
- **variable resistor** $I = K' \frac{W}{L}\left[(V_{GS} - V_T)V_{DS}\right]$
- **controlled current source (1)** $I = K' \frac{W}{L}(V_{GS} - V_T)^2$
- **controlled current source (2)** $I = I_M' \frac{W}{L} \exp\left(\frac{V_{GS}}{n\phi_t}\right)$

Signal processing in analog processing circuits

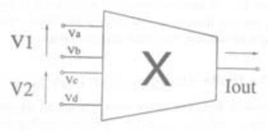


Fig. 3. A schematic view of the Gilbert multiplier.

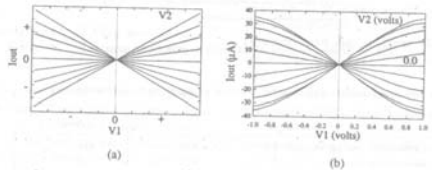
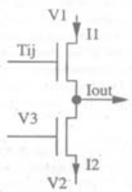
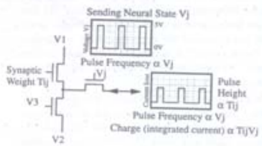


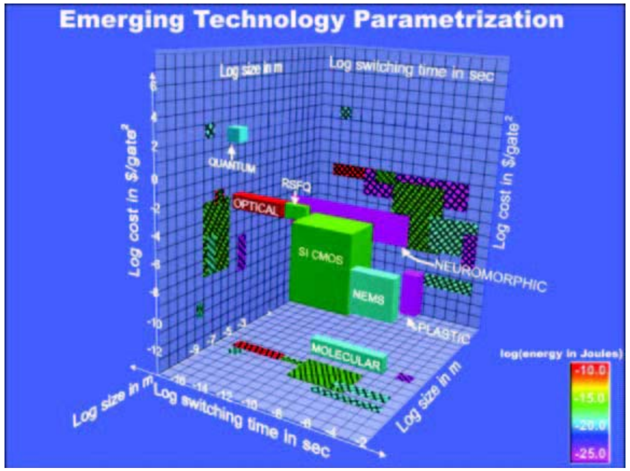
Fig. 4. Output characteristics of the Gilbert multiplier.



$I_{out} \approx K(V_1 - V_{out})(T_{ij} - V_{out})$



Technological trends *(Hutchby et al 2002)*



Technological trends *(Hutchby et al 2002)*

	Tmin [s]	Tmax [s]	CD min [m]	CD max [m]	Energy [J/op]
Si CMOS (22 nm node, 2001 ITRS)	3E-11	1E-6	3E-7	5E-6	4E-18
Neuromorphic	1E-13	1E-4	6E-6	6E-6	3E-25

Technological trends

Year	2001	2003	2005	2008
DRAM ½ pitch (nm)	150	120	100	70
MPU gate length (nm)	100	80	65	45
Memory size at introduction (bits)	2G	4G	8G	?
ASIC usable transistors / cm^2 (million)	40	73	133	328
Power supply voltage (V) (minimum logic V_{dd} for lowest power)	1.2	1.2	0.9	0.6
Chip frequency (MHz)	1,400	1,700	2,000	2,500
Maximum number of wiring levels	7	8	9	9
Number of total package pins / balls (ASIC)	2,000	2,500	3,100	4,400

Table: *Trends based on the International Technology Roadmap for Semiconductors, December 1999*

Rationale

Analog VLSI NNs intend to create biologically inspired structured neural systems that perform (specific) computations with high efficiency:

- the computational power of biological NNs derives not only from massive parallelism but also from analog processing [Mead 1989];
- full potential of silicon technology can be better exploited by using the physics of the devices to do the computation (i.e. considering the analog operation of integrated circuits [Mead 1990]);
- the possibility of mimicking the functions of biological neurons and networks (e.g. [Andreou 1991], [Meador 1989]).

Rationale

Analog VLSI technology looks attractive for the efficient implementation of artificial neural networks

- Massively parallel neural systems are efficiently implemented in analog VLSI technology, thus allowing high processing speed.
- Fault tolerance: to ensure fault tolerance to the hardware level it is necessary to introduce redundant hardware and, in analog VLSI technology, the cost of additional nodes is relatively low.
- Low power: the use of weak inversion operated MOS transistors reduces the synaptic and neuron power consumption, thus offering the possibility of low power neural systems.
- Real-world interface: analog neural networks eliminate the need for A/D and D/A converters and can be directly interfaced to sensors and actuators.

Basic research milestones

- Hopfield and Tank proposed the first electronic implementation of a NN in 1986. Their implementation is not suited for the direct VLSI implementation because: i) it is not area efficient; ii) it is difficult to integrate on silicon; iii) the circuit is not programmable.
- Tsividis and Satyanarayana in 1987 proposed a set of analog circuit primitives for adaptive NNs.
- In 1989, Mead designed circuits for early sensory functions and emphasized the role of analog processing, learning, self-organization, low power processing and area-efficient circuits.
- Vittoz, in 1990 outlined that analog neural processing is a low precision analog signal processing task.

Short and long term storage

The storage of information in analog VLSI circuits is not straightforward

- short term storage can be obtained by sampling and holding a voltage on a capacitor
- long term storage can be achieved:
  - by refreshing the voltage of the storage capacitor (amplitude quantization)
  - multi-level dynamic storage
  - non-volatile analogue weight storage

Short and long term storage

LTM implementation	Adaptation (learning)	Reference	Resolution [bits]
Non-volatile analog memory	Easy adaptation (on-chip learning)	[Kim 1998]	8
		[Holler 1989]	6
Local On-Chip Digital memory	Off chip learning (e.g. chip-in-the-loop learning)	[Shima 1992]	8
		[Spiegel 1992]	6
Analog self-refreshing memory cell	Easy adaptation (on-chip learning)	[Hochet 1991]	7 + 1/2
		[Castello 1991]	5
		[Cauwenberghs 1994]	8
		[Ehlert 1998]	12
Mixed digital/analog memory cell	Off chip learning (e.g. chip-in-the-loop learning)	[Castello 1991]	10

Analog signal processing issues

analog uncertainty

- process variations, non linearities, variable gains in multipliers (i.e. inaccuracies) don't appear to be a serious impediment
- component mismatch can give raise to destructive offset errors
- does noise enhance or not learning and generalization capabilities?
- accuracy of weight changes during learning is very important

Analog signal processing issues

- Analog circuits should be based upon ratios of matched components to eliminate whenever possible any dependency on process parameters
- **Mismatch:** it is the process that causes time-indipendent random variations in physical quantities of identically designed devices.
- Non-ideal behavior of circuits
- Circuit offsets
- etc.

◆ Many design trade-offs: speed/accuracy, area/accuracy, speed/area, power/accuracy, etc.

◆ High design, test and development costs

Analog signal processing issues

Following Draghici 2001, Lehmann 1999, and the usual meaning of the terms, *(absolute)* ***accuracy*** *is defined as the extent to which the results of a calculation or the readings of an instrument approach the true values of the calculated or measured quantities, and are free from errors*. What's more, ***precision*** is the measure of the range of values of a set of measurements, *and indicates reproducibility of the observations*.

Digital systems can be considered **precise**, since they always reproduce the same results in the same circumstances. However, digital systems can be considered **accurate only** to the extent to which they have **enough digits** to represent exactly the appropriate value (i.e. enough resolution).

Analog signal processing issues

Analog circuits are ***potentially accurate*** *because they are able to produce any specific value within their range*. Nevertheless analog circuits are affected by **noise** and, in analog circuits, absolute accuracy is very **expensive** (and not so meaningful) in terms of power consumption, silicon area and circuit complexity. *However analog circuits can be considered imprecise since they are unlikely to produce the same results in different occurrences of an experiment or in the same experiment with different silicon dies.*

From the previous considerations, a straightforward conclusion is that **analog circuits are not suitable for computations that need "exact"** (i.e. precise and accurate in the digital and absolute meaning) **responses**: i.e. analog circuits are poor at determining exact values.

Analog signal processing issues

In NNs, even if single processing elements exhibit low resolution, the ***collective computation*** of the whole network and the ***feedback scheme*** (i.e. **on-line, on-chip learning**) can be used to achieve the desired response.

Some authors compared analog and digital systems using digital-equivalent computing accuracy (i.e. absolute accuracy), i.e. **resolution** (i.e. S/N and equivalent number of bits), as comparison metrics.

In A/D and D/A conversion systems, the resolution (i.e. the Effective Number Of Bits, ENOB) is related in the analog domain to the Signal to Noise Ratio (i.e. SNR):

e.g. $(SNR)dB = 6.02 \times ENOB + 1.76$.

Analog signal processing issues

Shannon 1949: the capacity in bits (C) of a continuous (linear) channel in presence of additive white noise with power N is

$$C = B\log_2\left(\frac{S+N}{N}\right)$$

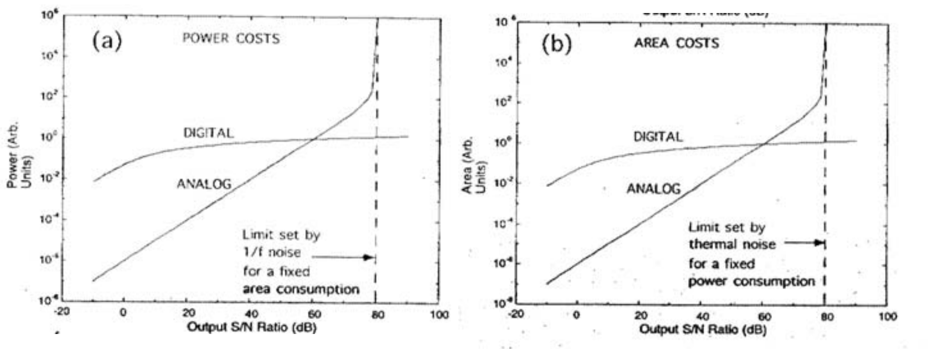
B is the bandwidth of the channel in bits per second and S is the signal power.

Rabaey 1996: if the number of devices switching per clock cycle is N, the clock frequency f, the average load capacitance C, the power supply VDD, the power consumption of digital circuits is given by:

$$P_D = NfCV_{DD}^2$$

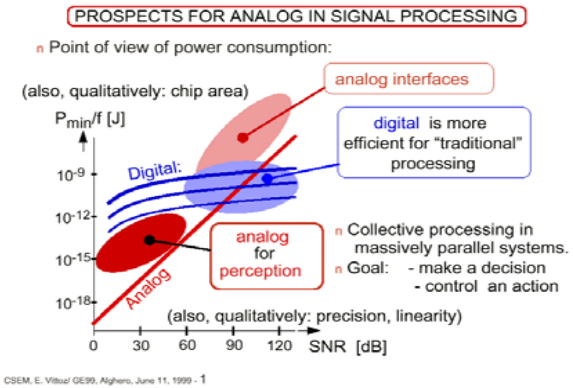
Es: $N = 10^5$, $f=10^8$ Hz, $C=10^{-12}$F, V_{DD}=2V then P_D=40 W

Analog signal processing issues



Sarpeshkar (1998) analysed a generic analog system and evidenced that analog is advantageous over digital (both in terms of power consumption and die area) up to about S/N = 60 dB.

Analog signal processing issues



Vittoz, 1990 and 1999, analysed filters (analog and digital): he evidenced that analog filters may consume much less power then their digital counterparts if a small dynamic range (i.e. SNR) is acceptable. Analog becomes extremely power inefficient when a large dynamic range is needed. Analog remains potentially advantageous over digital at low SNR ranges (less than about 60 dB) i.e. at low values of the ENOB (e.g. less than 10 bits).

Analog signal processing issues

It is worth noting that the previous analyses refer to ***linear*** systems ***without any feedback*** (and digital systems don't need feedback to increase accuracy but only to compute the system coefficients). Moreover, previous comparisons are made on digital perspective, i.e. in terms of "absolute" accuracy.

A **proper feedback schema** (i.e. **learning**, preferably implemented on-chip) can account for **relative accuracy** even if the analog circuits are inherently **not accurate and precise in absolute way**.

The **inherent feedback** structure provided by **learning** can, in principle, **compensate** for most of the **non-ideal effects** and errors. A **small ENOB** of an analog circuit doesn't prevent the overall system from achieving **correct results** as a digital system would do with the same resolution, in particular when the results consist of a non-linear complex computation (e.g. comparison, classification, recognition, etc.) on the inputs to the network.

Design methodology

Neural models	Computational primitives
Feed-forward (MLP)	neuron transfer function
Feed-forward (MLP)	synaptic multiplication
Feed-forward (MLP)	neuron input sum
Feed-forward (MLP)	weight storage
Back Propagation	neuron transfer function derivative
Back Propagation	adaptive and local control of the learning rate
Self Organizing features maps	winner-take-all networks
Boltzmann Machine	annealing method
Boltzmann Machine	co-occurrence computation

Design methodology

Computational primitives	Physical and circuit primitives
+ (sum)	Kirchoff Current Law
× (multiplication)	MOS transistor
	Operational Transconductance Amplifier
logarithm	translinear principle, [Andreou 1991b]
normalization	translinear principle, [Andreou 1991b]
"annealing"	thermal noise in the channel of a transistor, [Alspector 1991]
integration	sum of charges on a capacitor.
storage	dynamic storage of charges on a capacitor
Winner-Take-All	MOS channel length modulation [Lazzaro 1989].

Learning primitives

The learning primitives basically implement all the backward computations; for instance, in the case of the BP:

- neuron transfer function derivative;
- adaptive and local control of the learning rate;
- weight update;
- computation of error terms;
- etc.