

### Custom VLSI implementations of neural networks

### Early attempts

The idea of making custom analog VLSI implementations of neural networks dates back to the late '80's - early '90s:

[Holler et al. 1989, Satyanarayana et al. 1992, Hammerstrom 1993, Vittoz 1996]

- General purpose computing
- Full-custom analog implementation
- Neural network accelerator PC-boards
- Competing with "Intel steamroller"
- Communication bandwidth limited
- Difficult to "program"

### Current research Technological progress Application-specific focus Power-dissipation/computational Embedded system integration power

ICANN09 Tutorial

### Silicon neural network characteristics

## Silicon neuron designs

### Many VLSI models of spiking neurons have been developed in the past, and many are still being actively investigated:

- [1] C. Mead, Analog VLSI and Neural Systems. Reading, MA: Addison-Wesley, 1989.
- [2] M. Mahowald and R. Douglas, "A silicon neuron," Nature, vol. 354, pp. 515-518 1991. [3] A. van Schaik, "Building blocks for electronic spiking neural networks,"
- Neural Networks, vol. 14, no. 6-7, pp. 617-628, Jul-Sep 2001. [4] K. Hynna and K. Boahen, "Space-rate coding in an adaptive silicon neuron," *Neural Networks*, vol. 14, pp. 645–656, 2001.
- [5] G. Indiveri, "A low-power adaptive integrate-and-fire neuron circuit," in Proc. IEEE International Symposium on Circuits and Systems. IEEE, May 2003, pp. IV-820-IV-823.
- [6] L. Alvado, J. Tomas, S. Saighi, S. Renaud-Le Masson, T. Bal, A. Destexhe, and G. Le Masson, "Hardware computation of conductance-based neuron models," Neurocomputing, vol. 58-60, pp. 109-115, 2004. [7] M. Simoni, G. Cymbalyuk, M. Sorensen, and S. Calabrese, R.L. De-
- Weerth, "A multiconductance silicon neuron with biologically matched dynamics," IEEE Transactions on Biomedical Engineering, vol. 51, no. 2, pp. 342-354, February 2004.
- [8] J. Schemmel, K. Meier, and E. Mueller, "A new VLSI model of neural microcircuits including spike time dependent plasticity," in Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 3. IEEE, July 2004, pp. 1711-1716.
- [9] J. Arthur and K. Boahen, "Recurrently connected silicon neurons with active dendrites for one-shot learning," in *IEEE International Joint Conference on Neural Networks*, vol. 3, July 2004, pp. 1699–1704.
- [10] E. Farquhar and P. Hasler, "A bio-physically inspired silicon neuron," *IEEE Transactions on Circuits and Systems*, vol. 52, no. 3, pp. 477–488, March 2005.
- [11] K. M. Hynna and K. Boahen, "Neuronal ion-channel dynamics in silicon," in 2006 IEEE International Symposium on Circuits and Systems, May 2006, pp. 3614–3617.

- [12] J. Arthur and K. Boahen, "Synchrony in silicon: The gamma rhythm. IEEE Transactions on Neural Networks, vol. 18, pp. 1815–1825, 2007
- [13] J. Wijekoon and P. Dudek, "Compact silicon neuron circuit with spiking and bursting behaviour," Neural Networks, vol. 21, no. 2-3, pp. 524-534, March-April 2008.
- [14] P. Livi and G. Indiveri, "A current-mode conductance-based silicon neuron for address-event neuromorphic systems," in IEEE International Symposium on Circuits and Systems, ISCAS 2009. IEEE, May 2009, pp. 2898-2901.
- [15] F. Folowosele, A. Harrison, A. Cassidy, A. G. Andreou, R. Etienne-Cummings, S. Mihalas, Niebur, and T. Hamilton, "A switched capacitor implementation of the generalized linear integrate-and-fire neuron," in IEEE International Symposium on Circuits and Systems, ISCAS 2009. IEEE, May 2009, pp. 2149–2152.
- [16] M. Rastogi, V. Garg, and J. Harris, "Low power integrate and fire circuit for data conversion," in IEEE International Symposium on Circuits and Systems, ISCAS 2009. IEEE, May 2009, pp. 2669-2672. [17] E. Chicca, A. M. Whatley, V. Dante, P. Lichtsteiner, T. Delbrück,
- P. Del Giudice, R. J. Douglas, and G. Indiveri, "A multi-chip pulsebased neuromorphic infrastructure and its application to a model of orientation selectivity," IEEE Transactions on Circuits and Systems I, Regular Papers, vol. 5, no. 54, pp. 981–993, 2007.
   J. Schemmel, J. Fieres, and K. Meier, "Wafer-scale integration of analog neural networks," in *Proceedings of the IEEE International Joint*
- Conference on Neural Networks, 2008, (In Press).
- [19] Y. Wong, P. Xu, and P. Abshire, "Ultra-low spike rate silicon neuron," Biomedical Circuits and Systems Conference, 2007 BIOCAS 2007, pp 95-98, Nov. 2007
- [20] R. Vogelstein, U. Mallik, J. Vogelstein, and G. Cauwenberghs, "Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses," IEEE Transactions on Neural Networks, vol. 18, no. 1, pp. 253-265, Jan. 2007.

- Above threshold (strong inversion)
- Mixed analog/digital

G.Indiveri (NCS @ INI

- Rate-based
- Real-time
- Conductance-based
- Large-scale, event-based networks

- Below threshold (weak inversion)
- Spiking
- Accelerated-time
- Integrate-and-Fire

#### G.Indiveri (NCS @ INI)

Most designs can be traced back to one of two types of silicon neurons.

ini |uzh|eth|zürich 3/78

- Fully analog
  - ۲

  - Small-scale, hard-wired

## Why subthreshold neuromorphic VLSI



Exploit the physics of silicon to reproduce the *bio*-physics of neural systems.



#### G.Indiveri (NCS @ INI)

ICANN09 Tutorial

### Diffusion and saturation



$$I_{ds} = I_0 e^{\kappa_n V_g / U_T} \left( e^{-V_s / U_T} - e^{-V_d / U_T} \right)$$

is equivalent to:

$$I_{ds} = I_0 e^{\kappa \frac{V_g}{U_T} - \frac{V_s}{U_T}} - I_0 e^{\kappa \frac{V_g}{U_T} - \frac{V_d}{U_T}}$$
$$I_{ds} = I_f - I_r$$

If  $V_{ds} > 4U_T$  the  $I_r$  term becomes negligible, and the transistor is said to operate in the saturation regime:

$$I_{ds} = I_0 e^{\kappa_n V_g / U_T - V_s / U_T}$$

### MOSFETs in subthreshold



where

- *I*<sub>0</sub> denotes the nFET current-scaling parameter
- κ<sub>n</sub> denotes the nFET subthreshold slope factor
- $U_T$  the thermal voltage
- $V_g$  the gate voltage,  $V_s$  the source voltage, and  $V_d$  the drain voltage. The current is defined to be positive if it flows from the drain to the source

G.Indiveri (NCS @ INI)	ICANN09 Tutorial	ini  uzh eth zürich	9 / 78

### Exponential voltage dependence



Е

ini |uzh|eth|zürich

6/78

## n-FETs and p-FETs

In Complementary Metal-Oxide Semiconductor (CMOS) technology, there are two types of MOSFETs: n-FETs and p-FETs



In traditional CMOS circuits, all n-FETs have the common bulk potential  $(V_b)$ connected to Ground (Gnd), and all p-FETs have a common bulk potential (typically) connected to the power supply rail ( $V_{dd}$ ).

The corresponding (complementary) equation for the p-FET is

$$I_{ds} = I_0 e^{\kappa_p (V_{dd} - V_g)/U_T} \left( e^{-(V_{dd} - V_s)/U_T} - e^{-(V_{dd} - V_d)/U_T} \right)$$

#### G.Indiveri (NCS @ INI)

ICANN09 Tutorial

## The differential-pair



## One, two, and three transistor circuits



### The transconductance amplifier







In the linear region  $(|V_1 - V_2| < 200 mV)$ :

$$I_{out} \approx g_m (V_1 - V_2)$$



is a tunable conductance.

0.2

0.3

ini |uzh| eth | zürich 12 / 78

### What is a synapse?



In 1897 Charles Sherrington introduced the term synapse to describe the specialized structure at the zone of contact between neurons as the point in which one neuron communicates with another.

2005 winner of the Science and Engineering Visualization Challenge

by G. Johnson, Medical Media, Boulder, CO.





• Excitatory | Inhibitory

• Depressing | Facilitating

G.Indiveri (NCS @ INI)

ICANN09 Tutoria

ini |uzh|eth|zürich 19/78

## Synaptic transmission



In chemical synapses the presynaptic and postsynaptic membranes are spearated by extracellular space.

The arrival of a presynaptic action potential triggers the release of neurotransmitter in the extracellular space.

The neurotransmitters react with the postsynaptic receptors and depolarize the cell.

Chemical synaptic transmission is characterized by specific temporal dynamics.

## EPSC and EPSP

G.Indiveri (NCS @ INI)



Superimposed excitatory post-synaptic currents (EPSCs) recorded in a neuron at different membrane potentials (from Sacchi et al., 1998).



ini |uzh| eth | zürich

20 / 78

Excitatory post-synaptic potential (EPSP) in response to multiple pre-synaptic spikes (from Nicholls et al. 1992).

G.Indiveri (NCS @ INI)

ICANN09 Tutorial

### Neural network models



In classical neural network theory

- signals are (tipically) continuous values that represent the neuron's mean firing rate,
- neurons implement a saturating non-linearity transfer function (*S*) on the input's *weighted* sum,
- the synapse implements a multiplication between the neuron's input signal (*X<sub>i</sub>*) and its corresponding synaptic weight (*w<sub>i</sub>*).

G.Indiveri	(NCS @ INI)	

ICANN09 Tutorial

ini |uzh|eth|zürich 24/78

## VLSI synapses in pulse-based neural networks



In pulse-based neural networks the *weighted* contribution of a synapse can be implemented using a single transistor.

In this case p-FETs implement excitatory synapse, and n-FETs implement inhibitory synapses.

The synaptic weight can be set by changing the  $W_i$  bias voltage or the  $\Delta t$  duration.

## VLSI synapses in classical neural networks

The role of the VLSI synapse in implementations of "classical" neural network models is that of a *multiplier*.

Multiplying synaptic circuits have been implemented using a wide range of analog circuits, ranging from the single MOS-FETs to the Gilbert multiplier.



Figure: Schematic of half of a Gilbert multiplier. This circuit multiplies  $I_{in1}$  and  $I_{in2}$  by  $I_b$  if  $I_{in1} + I_{in2} = I_b$ .



### Linear pulse integrators

A *linear* integrator is a linear low-pass filter. Its impulse response should be a *decaying exponential*.

With VLSI and subthreshold MOSFETS its fairly easy to implement exponential voltage to current conversion, and linear voltage increase or decrease over time.



### Linear charge-and-discharge integrator



G.Indiveri (NCS @ INI)	ICANN09 Tutorial	ini  uzh eth züric
The diff-pair integrate	or (DPI)	
$V_{thr}$	$\tau \frac{d}{dt} I_{syn} + I_{syn} \approx \frac{d}{dt} I_{syn} + I_{syn} + I_{syn} \approx \frac{d}{dt} I_{syn} + I_{syn} + I_{syn} \approx \frac{d}{dt} I_{syn} + I_{syn} \approx \frac{d}{dt} I_{syn} + I_{syn} \approx \frac{d}{dt} I_{syn} + I_{syn} + I_{syn} \approx \frac{d}{dt} I_{syn} + I_{syn} + I_{syn} \approx \frac{d}{dt} I_{syn} + I_{syn} \approx \frac$	$I_{w} = I_{0}e^{\frac{\kappa V_{w}}{U_{T}}}$ $I_{\tau} = I_{0}e^{\frac{\kappa (V_{dd} - V_{\tau})}{U_{T}}}$ $I_{c} = C\frac{d}{dt}(V_{dd} - V_{syn})$ $I_{in} = I_{w}\frac{e^{\frac{\kappa V_{syn}}{U_{T}}}}{e^{\frac{\kappa V_{syn}}{U_{T}}} + e^{\frac{\kappa V_{thr}}{U_{T}}}}$ $I_{syn} = I_{0}e^{\frac{\kappa (V_{dd} - V_{syn})}{U_{T}}}$ $\frac{d}{dt}I_{syn} = -\frac{\kappa}{U_{T}}I_{syn}\frac{d}{dt}V_{syn}$ $\frac{I_{w}I_{gain}}{I_{\tau}}$

Log-domain pulse integrator



### **DPI** equations

**DPI transfer function:** 

$$\tau \frac{d}{dt} I_{out} + I_{out} = \frac{1}{I_{\tau}} \frac{I_{out}}{1 + \left(\frac{I_{out}}{I_g}\right)} I_{in}$$
  
$$\tau \frac{d}{dt} I_{out} + I_{out} \approx \frac{I_g}{I_{\tau}} I_{in}, \quad \text{if } I_{out} \gg I_g, \ I_w \gg I_{\tau}$$

Response to  $\Delta t$  pulse at time  $t_i$ :

$$I_{out}(t_i + \Delta t) = \frac{I_g I_{in}}{I_\tau} \left( 1 - e^{-\frac{\Delta t}{\tau}} \right) + I_{out}(t_i) e^{-\frac{\Delta t}{\tau}}, \quad I_{out}(t_i) = I_{out}(t_{i-1} + \Delta t) e^{-\frac{t_i - t_{i-1}}{\tau}}$$

Response to an arbitrary spike train  $\rho(t) = \sum_i \delta(t - t_i)$ , with  $\Delta t \ll \tau$ :

$$I_{out}(t) = \left(\frac{I_g I_{in}}{I_{\tau}}\right) e^{-\frac{t}{\tau}} \int_0^t e^{\frac{\xi}{\tau}} \rho(\xi) d\xi$$

Mean response to spike train of mean frequency  $\bar{v}$ :

$$< l_{out} > = \left(\frac{l_g l_{in}}{l_\tau}\right) \Delta t \bar{v}, \ \ \bar{v} = \frac{1}{\Delta t + l \bar{S} l}$$

ICANN09 Tutorial

G.Indiveri (NCS @ INI)

(Bartolozzi, Indiveri, 2007)

29 / 78

G.Indiveri (NCS @ INI)

## **DPI** measured response







### DPI response to spike-trains



## Short-term depression





(M. Boegerhausen, P. Suter, and S.-C. Liu, 2003)

G.Indiveri (NCS @ INI)

### STDP and beyond



#### Alternative spike-driven learning algorithm

Spike-driven weight change depends on the value of the post-synaptic neuron's membrane potential, and on its recent spiking activity.

Fusi et al. 2000; Brader et al. 2007

### Recipe for efficient VLSI implementation

- bistability: use two synaptic states;
- redundancy: implement many synapses that see the same pre- and post-synaptic activity'
- stochasticity & inhomogeneity: induce LTP/LTD only in a subset of stimulated synapses.
- Slow learning: only a fraction of the synapses memorize the pattern.
- + The theory is matched to the technology: use binary states, exploit mismatch and introduce fault tolerance by design.

ICANN09 Tutorial

### Neurons ... in a nutshell A quick tutorial

Real Neurons

Complexity

- Conductance based models
- Integrate and fire models
- Rate based models
  - Sigmoidal units
  - Linear threshold units

## Spike-driven learning in VLSI I

	J. Arthur and K. Boahen.			
	Learning in silicon: Timing is everythin In Y. Weiss, B. Schölkopf, and J. Platt, 2006.	g. editors, Advances in Neural Information Processin	g Systems 18. MIT Press, Cambridge, MA,	
	A. Bofill-i Petit and A. F. Murray.			
	Synchrony detection and amplification IEEE Transactions on Neural Network	by silicon neurons with STDP synapses. s, 15(5):1296–1304, September 2004.		
	E. Chicca, D. Badoni, V. Dante, M. D'A	ndreagiovanni, G. Salina, S. Fusi, and P. Del Giudi	ce.	
_	A VLSI recurrent network of integrate- IEEE Transactions on Neural Network	and-fire neurons connected by plastic synapses w s, 14(5):1297–1307, September 2003.	ith long term memory.	
	P. Häfliger, M. Mahowald, and L. Watts	6		
	A spike based learning neuron in anal In M. C. Mozer, M. I. Jordan, and T. Pe	og VLSI. tsche, editors, Advances in neuralinformation proc	essing systems, volume 9, pages 692–698.	
	G. Indiveri, E. Chicca, and R. Douglas			
	A VLSI array of low-power spiking neu IEEE Transactions on Neural Network	rons and bistable synapses with spike-timing depe s, 17(1):211-221, Jan 2006.	ndent plasticity.	
	S. Mitra, G. Indiveri, and S. Fusi.			
	Learning to classify complex patterns of In J.C. Platt, D. Koller, Y. Singer, and S. Cambridge (MA). 2020. MIT Press.	using a VLSI network of spiking neurons. . Roweis, editors, <i>Advances in Neural Information</i>	Processing Systems 20, pages 1009–1016,	
	Gambhuge (IVIA), 2008. MIT Press.			
	G.Indiveri (NCS @ INI)	ICANN09 Tutorial	ini  uzh  eth   zürich 3	8 / 78

### Neurons of the world



#### (adapted from B. Mel, 1994)

ICANN09 Tutoria

ini |uzh| eth | zürich 37 / 78

### **Equivalent Circuit**



If excitatory input currents are relatively small, the neuron behaves exactly like a first order low-pass filter.



### Spike generating mechanism



If the membrane voltage increases above a certain threshold, a spike-generating mechanism is activated and an action potential is initiated.



### Hardware implementations of spiking neurons

The first artificial neuron model was proposed in the 1943 by McCulloch and Pitts. Hardware implementations of this model date almost back to the same period.

Hardware implementations of *spiking* neurons are relatively new.



One of the most influential circuits that implements an *integrate and fire* (I&F) model of a neuron was the Axon-Hillock Circuit, proposed by Carver Mead in the late 1980s.



ini |uzh|eth|zürich 47/78

## Conductance based Si-Neurons



### Conductance-based models of spiking neurons







In 1991 Misha Mahowald and Rodney Douglas proposed a conductance-based silicon neuron and showed that it had properties remarkably similar to those of real cortical neurons.

G.Indiveri (NCS @ INI)	ICANN09 Tutorial	ini  uzh eth zürich	49 / 78

# Conductance based Si-Neurons

Silicon neuron's measurements



## The Axon-Hillock Circuit



## The Axon-Hillock Circuit



## **Capacitive Divider**







∆V2

## Axon-Hillock Circuit Dynamics



### **Power Dissipation**



The Axon-Hillock circuit is very compact and allows for implementations of dense arrays of silicon neurons BUT it has a major drawback: power consumption During the time when an inverter switches, a large amount



## Gain



### What's bad about this?

G.Indiveri (NCS @ INI)	ICANN09 Tutorial	ini  uzh  eth   zürich	58 / 78

### Conductance-based models Integrate and Fire vs Hodgkin-Huxley

Traditionally there have been two main classes of neuron models:





Integrate and fire (I-C)

## Conductance-based models

Integrate and Fire vs Hodgkin-Huxley

#### But recently proposed models bridge the gap between the two:

J Neurophysiol 92: 959–976, 2004; 10.1152/jn.00190.2004.

Generalized Integrate-and-Fire Models of Neuronal Activity Approximate Spike Trains of a Detailed Model to a High Degree of Accuracy Renaud Jolivet,<sup>1,6</sup> Timothy J. Lewis,<sup>2,4</sup> and Wulfram Gerstner<sup>1,6</sup> J. Norrophilad Docember 5, 2007. doi:10.1152/p.01107.2007. Dynamic *I-V* Curves Are Reliable Predictors of Naturalistic Pyramidal-Neuron Voltage Traces Laurent Badel,<sup>1</sup> Sandrine Lefort,<sup>2</sup> Romain Brette,<sup>3</sup> Carl C. H. Petersen,<sup>2</sup> Wulfram Gerstner,<sup>1</sup> and Magnus J. E. Richardson<sup>1,4</sup> Biol Cybern (2008) 99:361-370 DOI 10.1007/s00422-008-0259-4 Biological Cybernetics

ICANN09 Tutorial

Extracting non-linear integrate-and-fire models from experimental data using dynamic I-V curves

Laurent Badel · Sandrine Lefort · Thomas K. Berger · Carl C. H. Petersen · Wulfram Gerstner · Magnus J. E. Richardson

G.Indiveri (NCS @ INI)

ini |uzh|eth|zürich

### DPI neuron sub-threshold equations





### An ultra low-power generalized I&F circuit



(G. Indiveri, P. Livi, ISCAS 2009)

### G.Indiveri (NCS @ INI) ICANN09 Tutorial ini | uzh | eth | zürich 63 / 78

### SPICE simulations



62 / 78

## Experimental results Single spike



### Experimental results

Population activity



## Spiking multi-neuron architectures







- Networks of I&F neurons with adaptation, refractory period, *etc.*
- Synpases with realistic temporal dynamics
- Winner-Take-All architectures
- Spike-based plasticity mechanisms

Spikes and Address-Event Systems



### A spike-based learning chip



A minimum-size chip implementing a reconfigurable AER neural network. Neurons and synapses have realistic temporal dynamics. Local circuits at each synapse implement the bi-stable spike-based plasticity mechanism.

Technology:	AMS 0.35μm
Size:	3.9 <i>mm</i> × 2.5 <i>mm</i>
Neurons:	128
AER plastic synapses:	28  imes 128
AER non-plastic synapses	4 × 128
Dendritic tree multiplexer:	32 × 128     1 × 4096

Indiveri, Fusi, 2007

ini |uzh|eth|zürich 74/78

# **Distributed multi-layer networks**

Analog processing, asynchronous digital communication



### Summary

G.Indiveri (NCS @ INI)





ICANN09 Tutorial







# Thank you for your attention

Additional information available at:

### http://ncs.ethz.ch/

