Neuromorphic VLSI Event-Based devices and systems

Giacomo Indiveri

Institute of Neuroinformatics
University of Zurich and ETH Zurich

LTU, Lulea
May 28, 2012
Outline

1 “Neuromorphic Engineering”

2 Spike-based sensory systems

3 Spiking Neural Networks
   - Silicon Neurons
   - Silicon synapses
   - Winner-Take-All networks
   - Multi-chip networks

4 Learning

5 Neuromorphic Cognitive Systems
The brain of the worker honeybee occupies a volume of around 1 mm$^3$ and weighs about 1 mg. The total number of neurons in the brain is estimated to be 950,000.

- Flies acrobatically
- Recognizes patterns
- Navigates
- Forages
- Communicates
Energy consumption: $10^{-15}$ J/op, at least $10^6$ more efficient than digital silicon (20 watts vs. 1 Mil. watts)

The brain of the worker honeybee occupies a volume of around $1 \text{mm}^3$ and weighs about 1 mg. The total number of neurons in the brain is estimated to be 950,000.

- Flies acrobatically
- Recognizes patterns
- Navigates
- Forages
- Communicates

G. Indiveri (http://ncs.ethz.ch/)
Neocortex → Neural computation → Silicon → Behavior

G. Indiveri (http://ncs.ethz.ch/)

Neuromorphic spiking chips
Goals:

- To exploit the physics of silicon to reproduce the bio-physics of neural systems, using subthreshold analog VLSI circuits.
Goals:

- To exploit the physics of silicon to reproduce the bio-physics of neural systems, using subthreshold analog VLSI circuits.
- To develop multi-chip spike-based computing systems, using the Address-Event Representation (AER) and asynchronous digital VLSI technology.
Neuromorphic VLSI systems

Goals:

- To exploit the physics of silicon to reproduce the bio-physics of neural systems, using subthreshold analog VLSI circuits.
- To develop multi-chip spike-based computing systems, using the Address-Event Representation (AER) and asynchronous digital VLSI technology.
- To automatically configure and “program” neuromorphic processing systems distributed across multiple chips, to carry out real–time behavioral tasks.
Outline

1 “Neuromorphic Engineering”

2 Spike-based sensory systems

3 Spiking Neural Networks
   - Silicon Neurons
   - Silicon synapses
   - Winner-Take-All networks
   - Multi-chip networks

4 Learning

5 Neuromorphic Cognitive Systems
AER silicon retinas
Tobi Delbruck
Silicon retina properties
http://siliconretina.ini.uzh.ch

(movie)
An AER silicon cochlea
Shih-Chii Liu

AER EAR: A Matched Silicon Cochlea Pair With Address Event Representation Interface
Vincent Chan, Student Member, IEEE, Shih-Chii Liu, Member, IEEE, and André van Schaik, Senior Member, IEEE
Silicon cochlea properties

High f → Basilar membrane → Inner hair cells → Ganglion cells → Low f

Gain vs. Freq (Hz)

- channel 5
- channel 15
- channel 25
Silicon cochlea properties

High f → Basilar membrane → Low f

− Inner hair cells
− Ganglion cells
Outline

1. “Neuromorphic Engineering”
2. Spike-based sensory systems
3. Spiking Neural Networks
   - Silicon Neurons
   - Silicon synapses
   - Winner-Take-All networks
   - Multi-chip networks
4. Learning
5. Neuromorphic Cognitive Systems
Implement neural computation in silicon

Classical neural networks

Neuromorphic multi-neuron networks

G. Indiveri (http://ncs.ethz.ch/)
Spiking multi-neuron architectures

- Networks of silicon neurons with adaptation, refractory period, etc.
- Silicon synapses with realistic temporal dynamics
- Winner-Take-All architectures
- Spike-based plasticity mechanisms
Outline

1. “Neuromorphic Engineering”

2. Spike-based sensory systems

3. Spiking Neural Networks
   - Silicon Neurons
   - Silicon synapses
   - Winner-Take-All networks
   - Multi-chip networks

4. Learning

5. Neuromorphic Cognitive Systems
Silicon neurons

The low-power adaptive exponential I&F neuron

\[ \tau \frac{d}{dt} I_{\text{mem}} + I_{\text{mem}} \approx \frac{I_g I_{\text{in}}}{I_{\tau}} + f(I_{\text{mem}}) \]

[Indiveri et al., ISCAS 2010]
The low-power I&F neuron

Positive Feedback

![Graph showing the relationship between time (ms) and $I_{\text{mem}}/I_0$. The graph includes data points and a fit line.](http://ncs.ethz.ch/)
The low-power I&F neuron
Spike frequency adaptation

![Graph showing spike count vs. instantaneous firing rate](image_url)
The low-power I&F neuron

Basic response properties

Leaky I&F model

F-F curve (note mismatch)
Outline

1 “Neuromorphic Engineering”

2 Spike-based sensory systems

3 Spiking Neural Networks
   - Silicon Neurons
   - Silicon synapses
   - Winner-Take-All networks
   - Multi-chip networks

4 Learning

5 Neuromorphic Cognitive Systems
Synapses are often modeled as instantaneous multipliers.

Science and Engineering Visualization Challenge

The diff-pair integrator (DPI) circuit

\[ I_{\text{syn}}(t) = I_0 e^{-\frac{\kappa}{U_T} (V_{\text{syn}}(t) - V_{dd})} \]

\[ I_{\text{thr}} = I_0 e^{-\frac{\kappa(V_{\text{thr}} - V_{dd})}{U_T}} \]

\[ C_{\text{syn}} \frac{d}{dt} V_{\text{syn}} = -(I_{\text{in}} - I_{\tau}) \]

\[ \tau \frac{d}{dt} I_{\text{syn}} + I_{\text{syn}} = \frac{I_{\text{thr}} I_{\text{w}}}{I_{\tau}} \]

[Bartolozzi and Indiveri, Neural Computation, 2007]
The DPI synapse

Temporal dynamics

![Diagram showing temporal dynamics of EPSC (nA) over time (s) for different values of Vw (300mV, 320mV, 340mV)].

- **Vw = 300mV**
- **Vw = 320mV**
- **Vw = 340mV**

G. Indiveri (http://ncs.ethz.ch/)
Outline

1. “Neuromorphic Engineering”

2. Spike-based sensory systems

3. Spiking Neural Networks
   - Silicon Neurons
   - Silicon synapses
   - Winner-Take-All networks
   - Multi-chip networks

4. Learning

5. Neuromorphic Cognitive Systems
Recurrent cooperative-competitive architectures

- Hardwired local synapses
- Local excitatory connections
- Global inhibitory connections

[Chicca et al., *Nips*, 2006]
Local recurrent connectivity

Winner-take-All architectures

- Input signals are encoded with mean firing rates
- Computation and information transfer is data driven
Without local connectivity activated output spike rates represent linearly input spike rates (modulo mismatch effects)
Local recurrent connectivity

Winner-take-All architectures

With local WTA connectivity the network exhibits:

- Selective amplification
- Signal normalization
- Signal restoration

[Chicca et al., *Nips*, 2006]
1 “Neuromorphic Engineering”

2 Spike-based sensory systems

3 Spiking Neural Networks
   - Silicon Neurons
   - Silicon synapses
   - Winner-Take-All networks
   - Multi-chip networks

4 Learning

5 Neuromorphic Cognitive Systems
Spikes and Address-Event Systems

**Inputs**

- Encode
- Decode
- Address Event Bus

**Source Chip**

**Action Potential**

**Address Event representation of action potential**

**Outputs**

**Destination Chip**

- 1
- 2
- 3
- 12

**Graph**

- V_{mem} (V)
- Time (s)

G. Indiveri (http://ncs.ethz.ch/)

Neuromorphic spiking chips
Hierarchical or multi-layer networks

The basic problem with these models is, of course, generalization: a look-up table cannot deal with new events, such as viewing a face from the side rather than the front, and it cannot learn in the prediction-free sense described earlier. One of the simplest and most powerful types of algorithm developed within learning theory corresponds to the basic ability to generalize depends on the combination of cells tuned by visual experience. Notice that in the model of Fig. 2, the tuning of the units depends on learning, probably unsupervised (for instance, through random fluctuations of the weights, as discussed in reference 18). However, the weights of the combination (see Fig. 3) depend on learning the task and require at least some feedback (see Box 2).

Circuits in infratemporal cortex and prefrontal cortex, which come to a variety of stimuli, although at sub-maximal firing rates — might support the existence of a similar architecture underlying generalization in domains other than vision. The question now is whether the available evidence supports the existence of a similar architecture underlying generalization in domains other than vision. The older computation-through-memory models of vision and motor control. The question now is whether the available evidence supports the existence of a similar architecture underlying generalization in domains other than vision.
Hierarchical or multi-layer networks

In visual cortex, neurons with a bell-shaped tuning are common. For example, Logothetis et al. trained monkeys to perform an object recognition task with isolated views of novel three-dimensional objects. About one tenth of the tuned neurons were view-invariant, showing specificity for a certain object view or lighting condition. About 90% of the tuned neurons were view-tuned (see Fig. 1) to one of the training objects. When recording from the animals' inferotemporal cortex, they found that the great majority of neurons selectively tuned to complex stimuli such as faces and other objects. About one tenth of the tuned neurons were view-invariant, showing specificity for a certain object view or lighting condition. About 90% of the tuned neurons were view-tuned (see Fig. 1) to one of the training objects.

Figure 1 sketches one such quantitative model, and summarizes a combination of such view-tuned neurons (Fig. 2) can provide view-invariant, object recognition in cortex. The question now is whether the available evidence supports the existence of a similar architecture underlying generalization in domains other than vision. The older computation-through-memory models of vision and motor control. The question now is whether the available evidence supports the existence of a similar architecture underlying generalization in domains other than vision. The older computation-through-memory models of vision and motor control. The question now is whether the available evidence supports the existence of a similar architecture underlying generalization in domains other than vision.

In summary, the accumulated evidence points to a visual recognition scheme for visual recognition and its quantitative consistency with categorization. Computer models have shown the plausibility of this generalization well by interpolating among the examples. That a combination of broadly tuned neurons—those that respond to the side rather than the front, and it cannot learn in the predictive sense described earlier. One of the simplest and most powerful types of algorithm developed within learning theory corresponds to a look-up table cannot deal with new events, such as viewing a face. The basic problem with these models is, of course, generalization: a look-up table cannot deal with new events, such as viewing a face.
Outline

1 “Neuromorphic Engineering”

2 Spike-based sensory systems

3 Spiking Neural Networks
   • Silicon Neurons
   • Silicon synapses
   • Winner-Take-All networks
   • Multi-chip networks

4 Learning

5 Neuromorphic Cognitive Systems
Spike-timing dependent plasticity (STDP)

1. If an input (pre-synaptic) spike arrives shortly before an output (post-synaptic) spike is emitted, the synaptic efficacy is increased.
2. If it arrives soon after the output spike is emitted, the synaptic efficacy is decreased.

Abbot, Nelson, 2000
STDP is not enough for learning complex spatio-temporal patterns

Senn, Biological Cybernetics, 2002

[...] additional non linearities are required if STDP should be relevant for both encoding information represented in a spike correlation code and a mean rate code without spike correlations.
STDP is not enough for learning complex spatio-temporal patterns

Spike-based learning mechanisms ideal for VLSI implementations

- depend on the neuron’s membrane potential;
- synaptic weights have two stable states (bi-stability);
- many synapses see the same pre- and post-synaptic mean activity (redundancy);
- LTP/LTD is induced only in a random subset of stimulated synapses (stochasticity).

[Fusi et al. 2000]; [Gütig, Sompolinsky 2006]; [Brader et al. 2007]
Spike-driven plasticity in silicon

Pre-synaptic Inputs

\[ w_1 \quad w_2 \quad w_3 \quad w_4 \quad \ldots \quad w_n \]

Post-synaptic Output

Pre-synaptic component

Post-synaptic component

G. Indiveri (http://ncs.ethz.ch/)

Neuromorphic spiking chips
Spike-driven plasticity in silicon

\[ \text{post} \]

\[ \text{Up} \]

\[ \text{Dn} \]

\[ \text{~weight} \]

\[ \text{pre} \]

[Mitra et al. 2009]
Stochastic weight update
LTP/LTD probabilities and stop-learning

LTD consolidation

No LTD consolidation
Outline

1. "Neuromorphic Engineering"

2. Spike-based sensory systems

3. Spiking Neural Networks
   - Silicon Neurons
   - Silicon synapses
   - Winner-Take-All networks
   - Multi-chip networks

4. Learning

5. Neuromorphic Cognitive Systems
Distributed event-driven systems

Neuromorphic “cognitive” systems can be assembled by using:

1. Full custom hybrid analog/digital neural processing VLSI devices.
2. A spike based communication protocol (e.g., the Address-Event Representation).
State dependent computation

sWTA diagram

- Inhibitory neurons
- Excitatory neurons
- Global Inhibition
- Nearest-N Excitation

sWTA networks as building blocks

- Linear behaviors
  - Analog gain
  - Locus invariance
  - Gain control by common mode input

- Non linear behaviors
  - Selective amplification
  - Signal restoration
  - Multi-stability

- Configure key parameters of the WTA network automatically.
- Implement “state-holding” elements.
- Learn network connectivity patterns.

[Douglas and Martin, 2007]
Conclusions

Toward neuromorphic cognitive behaving systems

By using event based sensors and spike-based neural processing circuits it is possible to implement real-time sensory-motor systems. By using soft WTA multi-chip networks it is possible to implement real-time state-dependent computation. The AER communication infrastructure and automated parameter tuning techniques, allow us to synthesize spike-based neural finite state machines.

experimental neuroscience
neural computation
analog VLSI design
modeling
theoretical neuroscience
system integration
theory of computing
digital VLSI development
Conclusions

Toward neuromorphic cognitive behaving systems

- By using event based sensors and spike-based neural processing circuits it is possible to implement real-time sensory-motor systems.

---

- experimental neuroscience
- neural computation
- analog VLSI design
- modeling
- system integration
- theoretical neuroscience
- theory of computing
- digital VLSI development
Conclusions
Toward neuromorphic cognitive behaving systems

- By using event based sensors and spike-based neural processing circuits it is possible to implement real-time sensory-motor systems.
- By using soft WTA multi-chip networks it is possible to implement real-time state-dependent computation.

---

Excitatory
Inhibitory

Silicon Retina

Silicon Cochlea

experimental neuroscience
neural computation
analog VLSI design
modeling
system integration
theoretical neuroscience
time of computing
digital VLSI development
Conclusions
Toward neuromorphic cognitive behaving systems

- By using event based sensors and spike-based neural processing circuits it is possible to implement real-time sensory-motor systems.
- By using soft WTA multi-chip networks it is possible to implement real-time state-dependent computation.
- The AER communication infrastructure and automated parameter tuning techniques, allow us to synthesize spike-based neural finite state machines.
Acknowledgments

The Institute of Neuroinformatics

Elisabetta Chicca
Stefano Fusi
Chiara Bartolozzi
The NCS group (http://ncs.ethz.ch/)

Rodney Douglas
Kevan Martin
Richard Hahnloser

Funding sources

neuroP (257219) ERC
SCANDLE (ICT-231168)
eMorph (ICT-231467)
nAttention (121713) SNF
SoundRec (119973) SNF