Neuromorphic Silicon Neural Interfaces

Gert Cauwenberghs

Department of Bioengineering Institute for Neural Computation UC San Diego

http://isn.ucsd.edu

Gert Cauwenberghs

Neuromorphic Silicon Neural Interfaces

gert@ucsd.edu

Neuromorphic Engineering *"in silico" neural systems design*



Gert Cauwenberghs

Computational Systems Neuroscience

Brain 1 m

Systems







Neurons 100 μm





Neuromorphic Systems Engineering



V_{Pre}

Φ

Ō

ð

0

C_{Membrane}

V_{Post}

Drain

Multi-scale levels of investigation in analysis of the central nervous system (adapted from Churchland and Sejnowski 1992) and corresponding neuromorphic synthesis of highly efficient silicon cognitive microsystems. Boltzmann statistics of ionic and electronic channel transport provide isomorphic physical foundations.

G. Cauwenberghs, "Reverse Engineering the Cognitive Brain," PNAS, 2013

Gert Cauwenberghs

Analysis

Neuromorphic Silicon Neural Interfaces

gert@ucsd.edu

Physics of Neural Computation

Silicon and Lipid Membranes Mead, 1989



Voltage-dependent *p*-channel

- Hole transport between source and drain
- Gate controls energy barrier for holes across the channel
- Boltzmann distribution of *hole energy* produces exponential *decrease* in channel conductance with gate voltage



Voltage-dependent conductance

- K^+/Na^+ transport across lipid bilayer
- Membrane voltage controls energy barrier for opening of ion-selective channels
- Boltzmann distribution of *channel energy* produces exponential increase in K^+/Na^+ conductance with membrane voltage

Event-Coding Silicon Retina

Zaghloul and Boahen, 2006



- Models coding and communication of visual events in the mammalian retina and optic nerve
 - Integrated photosensors (rods)
 - On and off transient and sustained ganglia cell outputs
 - Spatiotemporal compressed coding and communication in optic nerve
 - Address-event coding of spikes

Change Threshold Detection APS CMOS Imager

Chi, Mallik, Clapp, Choi, Cauwenberghs and Etienne-Cummings (2007)



- Event-driven video compression
 - Change detection and threshold encoding on the focal plane
- 6T pixel combines APS and change event coding
- 4.3mW power at 3V and 30fps





Video Out



Slow Rotation



Change Events Out

Reconfigurable Synaptic Connectivity and Plasticity *From Microchips to Large-Scale Neural Systems*



Gert Cauwenberghs

Scaling of Task and Machine Complexity



Achieving (or surpassing) human-level machine intelligence requires a convergence between:

- Advances in computing resources approaching connectivity and energy efficiency levels of computing and communication in the brain;
- Advances in deep learning methods, and supporting data, to adaptively reduce algorithmic complexity.

Scaling and Complexity Challenges

- Scaling the event-based neural systems to performance and efficiency approaching that of the human brain will require:
 - Scalable advances in silicon integration and architecture
 - Scalable, locally dense and globally sparse interconnectivity
 - *Hierarchical address-event routing*
 - High density (10¹² neurons, 10¹⁵ synapses within 5L volume)
 - Silicon nanotechnology and 3-D integration
 - High energy efficiency (10¹⁵ synOPS/s at 15W power)
 - Adiabatic switching in event routing and synaptic drivers
 - Scalable models of neural computation and synaptic plasticity
- **Neuro** Convergence between cognitive and neuroscience modeling
 - Modular, neuromorphic design methodology
- **CogSci** Data-rich, environment driven evolution of machine complexity

NanoE Phys

CS

EE

Large-Scale Reconfigurable Neuromorphic Computing Technology and Performance Metrics

	Stromatias 2013 SpiNNaker Manchester	Davies 2018 Loihi Intel	Merolla 2014 TrueNorth IBM	Schemmel 2010 FACETS/BrainScaleS Heidelberg	Benjamin 2014 NeuroGrid Stanford	Park 2014 IFAT UCSD
Technology (nm)	130	14	28	180	180	90
Die Size (mm ²)	102	60	430	50	168	16
Neuron Type	Digital Arbitrary	Digital Conductance Integrate & Fire	Digital Accumulate & Fire	Analog Conductance Integrate & Fire	Analog Shared-Dendrite Conductance I&F	Analog 2-Compartment Conductance I&F
# Neurons	5216 ¹	128k ²	1M ²	512	65k	65k
Neuron Area (μm²)	N/A ¹	240 (240k) ²	14 (3325) ²	1500	1800	140
Peak Throughput (Events/s)	5M	3.4G	1G	65M	91M	73M
Energy Efficiency (J/SynEvent)	8n	24p	26p	N/A	31p	22p

¹ Software-instantiated neuron model

² Time-multiplexed neuron processor

- Benjamin, B., P. Gao, E. McQuinn, S. Choudhary, A. Chandrasekaran, J. Bussat, R. Alvarez-Icaza, J. Arthur, P. Merolla, and K. Boahen, "Neurogrid: A mixed analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, 102(5):699–716, 2014.
- Davies, M. et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro, vol. 38 (1), pp. 82-99, 2018.
- Merolla, P.A., J.V. Arthur, R. Alvarez-Icaza, A S. Cassidy, J. Sawada, F. Akopyan, B.L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S.K. Esser, R. Appuswamy, B. Taba, A. Amir, M.D. Flickner, W.P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, 345(6197):668–673, 2014.
- Park, J., S. Ha, T. Yu, E. Neftci, and G. Cauwenberghs, "65k-neuron 73-Mevents/s 22-pJ/event asynchronous micro-pipelined integrate-and-fire array transceiver," Proc. 2014 IEEE Biomedical Circuits and Systems Conf. (BioCAS), 2014.
- Schemmel, J., D. Bruderle, A. Grubl, M. Hock, K. Meier, and S. Millner, "A waferscale neuromorphic hardware system for large-scale neural modeling," *Proc. 2010 IEEE Int. Symp. Circuits and Systems (ISCAS)*, 1947–1950, 2010.
- Stromatias, E., F. Galluppi, C. Patterson, and S. Furber, "Power analysis of largescale, real-time neural networks on SpiNNaker," *Proc. 2013 Int. Joint Conf. Neural Networks (IJCNN)*, 2013.

Long-Range Configurable Synaptic Connectivity



Comparison of synaptic connection topologies for several recent large-scale event-driven neuromorphic systems and the proposed hierarchical addressevent routing (HiAER), represented diagrammatically in two characteristic dimensions of connectivity: expandability (or extent of global reach), and flexibility (or degrees of freedom in configurability). Expandability, measured as distance traveled across the network for a given number of hops *N*, varies from linear and polynomial in *N* for linear and mesh grid topologies to exponential in *N* for hierarchical tree-based topologies. Flexibility, measured as the number of target destinations reachable from any source in the network, ranges from unity for point-to-point (P2P) connectivity and constant for convolutional kernel (Conv.) connectivity to the entire network for arbitrary (Arb.) connectivity. MMAER: Multicasting Mesh AER; WS: Wafer-Scale.

Park et al, "Hierarchical Address Event Routing for Reconfigurable Large-Scale Neuromorphic Systems," IEEE TNNLS, 2017

Neuromorphic Silicon Neural Interfaces

Hierarchical Address-Event Routing (HiAER)



 $L3^1$ $L2^2$ L_2 L12 L1 2(3) (1)(2)(3) (4)(1)(2)(3) (4) 2 3 4 1

Example network with 16 neurons and weighted synaptic connections.

Example partitioning into hierarchical neural network with ascending and descending projections through inserted relay neurons.



Corresponding edge-vertex-dual HiAER implementation with synaptic routing tables (SRT) at each level in the hierarchy.

Joshi et al, 2010; Park et al, 2011, 2017

Gert Cauwenberghs

Large-Scale Reconfigurable Neuromorphic Computing



Hierarchical Address-Event Routing (HiAER) Integrate-and-Fire Array Transceiver (IFAT) for scalable and reconfigurable neuromorphic neocortical processing. (a) Biophysical model of neural and synaptic dynamics. (b) Dynamically reconfigurable synaptic connectivity is implemented across IFAT arrays of addressable neurons by routing neural spike events locally through DRAM synaptic routing tables. (c) Each neural cell models conductance based membrane dynamics in proximal and distal compartments for synaptic input with programmable axonal delay, conductance, and reversal potential. (d) Multiscale global connectivity through a hierarchical network of HiAER routing nodes. (e) HiAER-IFAT board with 4 IFAT custom silicon microchips, serving 256k neurons and 256M synapses, and spanning 3 HiAER levels (L0-L2) in connectivity hierarchy. (f) The IFAT neural array multiplexes and integrates (top traces) incoming spike synaptic events to produce outgoing spike neural events (bottom traces). The latest IFAT microchip measured energy consumption is 22 pJ per spike event, several orders of magnitude more efficient than emulation on CPU/GPU platforms.

Yu et al, BioCAS 2012; Park et al, BioCAS 2014; Park et al, TNNLS 2017; Broccard et al, JNE 2017

Memristive Synapse Arrays for Neuromorphic Processing-in-Memory



Intel/STmicroelectronics (Numonyx) 256Mb multi-level phase-change memory (PCM) [Bedeschi et al, 2008]. Die size is 36mm2 in 90nm CMOS/Ge2Sb2Te5, and cell size is 0.097μm2. (a) Basic storage element schematic, (b) active region of cell showing crystalline and amorphous GST, (c) SEM photograph of array along the wordline direction after GST etch, (d) I-V characteristic of storage element, in set and reset states, (e) programming characteristic, (f) I-V characteristic of pnp bipolar selector.

- Scalable to high density and energy efficiency

- < 100nm cell size in 12nm CMOS
- < pJ energy per synapse operation
- Vertically stacked integration in Intel-Micron 3D Xpoint/Optene SSD persistent memory

Memristive Synapse Arrays for Neuromorphic Processing-in-Memory



Hybridization and nanoscale integration of CMOS neural arrays with phase change memory (PCM) synapse crossbar arrays. (a) Nanoelectronic PCM synapse with spike-timing dependent plasticity (STDP) [Kuzum *et al*, 2011]. Each PCM element implements a synapse with conductance modulated through phase transition as controlled by timing of voltage pulses. (b) CMOS IFAT array vertically interfacing with nanoscale PCM synapse crossbar array by interleaving via contacts to crossbar rows. The integration of IFAT neural and PCM synapse arrays externally interfacing with HiAER neural event communication combines the advantages of highly flexible and reconfigurable HiAER-IFAT neural computation and long-range connectivity with highly efficient (fJ/synOP range energy cost) local synaptic transmission.

CMOS-RRAM Reconfigurable Neurosynaptic Array

Wan et al, ISSCC 2020

- Integration of CMOS neurons and resistive random-access memory (RRAM) memristive synapses with *in-situ* revertible dataflow at record efficiency



W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, H-S. P. Wong, "33.1: A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models," *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco CA, Febr. 15-19, 2020.

CMOS-RRAM Reconfigurable Neurosynaptic Array

Wan et al, ISSCC 2020

- Gibbs stochastic sampling for Bayesian generative inference
 - Alternating between INFerence and GENeration in transpose datapaths
 - Restricted Boltzmann Machine (RBM) / Variational Autoencoder (VAE)
- Real-time image reconstruction from corrupted/noisy MNIST input



W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, H-S. P. Wong, "33.1: A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models," *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco CA, Febr. 15-19, 2020.

CMOS-RRAM Reconfigurable Neurosynaptic Array

Wan et al, ISSCC 2020



W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, H-S. P. Wong, "33.1: A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models," *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco CA, Febr. 15-19, 2020.

Gert Cauwenberghs

Hierarchical Address-Event Routing Neural and Synaptic Array Transceiver HiAER-NSAT



Kerneltron: Adiabatic Support Vector "Machine"

Karakiewicz, Genov and Cauwenberghs



$$y = \operatorname{sign}(\sum_{i \in S} \lambda_i y_i K(\mathbf{x}_i, \mathbf{x}) + b)$$



Classification results on MIT CBCL face detection data



Karakiewicz, Genov, and Cauwenberghs, VLSI'2006; CICC'2007; JSSC 2007; SJ 2013

m

M1

m.n

w

Write

M2

 $RS_{m}^{(i)} x_{n}^{(j)} Vout_{m}^{(i)}$

1.2 TMACS / mW

- adiabatic resonant clocking conserves charge energy
- energy efficiency on par with human brain (10¹⁵ SynOP/S at 15W)



 $Vout^{(i)}_{m}$

0

Vdd/2

Vdd

MЗ

CID

gert@ucsd.edu

Closing the Loop: Interactive Neural/Artificial Intelligence



gert@ucsd.edu



Neuromorphic Systems Engineering

F. Broccard, S. Joshi, J. Wang and G Cauwenberghs, "Neuromorphic neural interfaces: from neurophysiological inspiration to biohybrid coupling with nervous systems," *JNE*, 2017

Gert Cauwenberghs

Minimally Invasive Neurotechnologies



S. Ha, C. Kim, A. Akinin, J. Park, H. Wang, C. Maier, P. Mercier and G. Cauwenberghs, "Silicon Integrated High-Density Electrocortical Interfaces," *Proceedings of the IEEE*, vol.105 (1), pp. 11-33, 2017.

ENIAC:

Encapsulated Neural Interfacing and Acquisition Chip



S. Ha, C. Kim, A. Akinin, J. Park, H. Wang, C. Maier, P. Mercier and G. Cauwenberghs, "Silicon Integrated High-Density Electrocortical Interfaces," *Proceedings of the IEEE*, vol.105 (1), pp. 11-33, 2017.

Gert Cauwenberghs

Neuromorphic Silicon Neural Interfaces

Highly Sensitive, Low-Noise Low-Power Integrated Biopotential Sensing and Acquisition

Kim, Joshi, Courellis, Wang, Miller, and Cauwenberghs, 2018

First biopotential integrated ADC to deliver greater than 90dB dynamic range, lower than 1µVrms input-referred noise, and faster than 1ms settling to 200mVpp input transients, at less than 1µW power per channel, with 16 recording channels integrated within 1 sq. mm in 65nm CMOS:



Kim, Joshi, Courellis, Wang, Miller, and Cauwenberghs, "A 92dB Dynamic Range sub-µV rms-noise 0.8 µW/ch Neural-Recording ADC Array with Predictive Digital Autoranging," IEEE ISSCC 2018.

Electrophysiology Lab-on-a-Chip



J. Wang, A. Paul, D. Zhang, J. Wu, Y. Xu, Y. Zou, C. Kim, and G. Cauwenberghs, "1024-Electrode Hybrid Voltage/Current-Clamp Neural Interface System-on-Chip with Dynamic Incremental-SAR Acquisition," 2020 IEEE Symposium on VLSI Circuits (VLSI-Circuits), Honolulu HI, June 14-19, 2020.

Gert Cauwenberghs

Neuromorphic Silicon Neural Interfaces

gert@ucsd.edu

Deep Brain Stimulation (DBS) for Parkinson's Disease Remediation

- Intrusive intervention
 - "Brain's pacemaker"
 - Electrode is implanted in the deep brain's thalamus
 - Periodic (130-185Hz) activation of electrical impulses delivered by the electrode to suppress Parkinsoninduced tremor
- Highly invasive procedure
 - Surgical insertion of electrode and stimulation electronics
 - Battery needs to be replaced
- Open-loop
 - Adaptation (e.g. Medtronic Activa PC+S) limited to user-mediated control of stimulation amplitude



Surgery to insert electrode deep in the brain. Parkinson's patient remains awake during surgery. http://en.wikipedia.org/wiki/Deep_brain_stimulation

Distributed Brain Dynamics of Human Motor Control

G. Cauwenberghs, K. Kreutz-Delgado, T.P. Jung, S. Makeig, H. Poizner, T. Sejnowski, M. Arnold, F. Broccard, Y.M. Chi, J. Iversen, C. Maier, E. Neftci, D. Peterson,

A. Akinin, S. Das, N. Govil, S. Hsu, T. Mullen, A. Ojeda, C. Stevenson

NSF EFRI-1137279: Mind, Machines and Motor Control (M3C)



Gert Cauwenberghs

gert@ucsd.edu

Wireless Non-Contact Biopotential Sensors

Chi et al, 2010-



Chi, Maier, Cauwenberghs, "Ultra-high input impedance, low noise integrated amplifier for noncontact biopotential sensing," IEEE JETCAS 1(4), 526-535, 2011.

Joshi, Kim and Cauwenberghs, "A 6.5-µW/MHz Charge Buffer With 7-fF Input Capacitance in 65-nm CMOS for Noncontact Electropotential Sensing," IEEE TCAS-II, 63(12), 1161-1165, 2016.

Chi, Wang, Wang, Maier, Jung, and Cauwenberghs, "Dry and Noncontact EEG Sensors for Mobile Brain-Computer Interfaces," IEEE TNSRE 20(2), 228-235, 2012.

Gert Cauwenberghs

Neuromorphic Silicon Neural Interfaces

Unobtrusive Dry-EEG Functional Brain Imaging

Mullen, Kothe, Chi, Ojeda, Kerth, Makeig, Jung, and Cauwenberghs, 2015



Mullen et al, Real-time Neuroimaging and Cognitive Monitoring Using Wearable Dry EEG. IEEE TMBE, 2015.

Opportunities for In-Ear Health Sensing

• Prevalence of wireless personal audio devices:



Rapidly aging global population:

 Over the next few decades, people 65 years and older will account for 20% of the global population, an unprecedented shift. New healthcare challenges and opportunities will arise for which reliable and continuous high-bandwidth health data will be critical.



In-Ear Health Sensing Platform

 An in-ear healthcare platform has the convenience, comfort, and discretion of a consumer audio device, while offering valuable electrophysiological and biochemical data.



In-Ear Electrophysiology

Paul et al, IEEE NER 2019; IEEE EMBC 2019

High-density dry-contact electrodes capture a wealth of physiological information from an integrated in-ear device



ASSR PSD

Impedance Imaging

- In-ear, high-density dry-contact electrode recording platform records electroencephalography (EEG) signals from the brainstem, temporal, and visual cortexes with quality comparable to commercial scalp EEG.
- Electrical impedance measurement provides electrodermal activity (EDA).
- Opportunities for closed-loop auditory neurofeedback (tinnitus, insomnia, apnea, etc).

Paul, A., Deiss, S., Tourtelotte, D., Kleffner, M., Zhang, T., and Cauwenberghs, G. Electrode-Skin Impedance Characterization of In-Ear Electrophysiology Accounting for Cerumen and Electrodermal Response. IEEE EMBS Int. Conf. Neural Engineering (NER'19), 2019.

Paul, A., Akinin, A., Cauwenberghs, G. Integrated In-Ear Device for Auditory Health Assessment. 2019 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC'19), 2019.

Integrated Systems Neuroengineering



Neuromorphic Silicon Neural Interfaces