# Demonstrating the Advantages of Analog Wafer-Scale Neuromorphic Hardware

Hartmut Schmidt Andreas Grübl José Montes
Eric Müller Sebastian Schmitt Johannes Schemmel

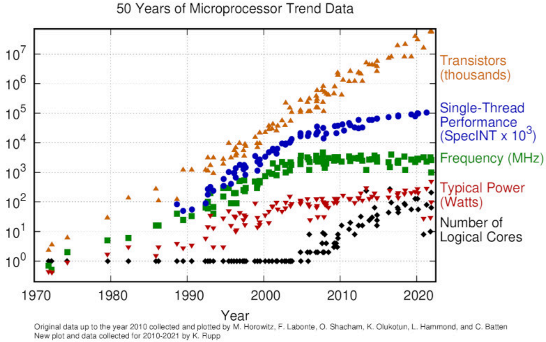
mueller@kip.uni-heidelberg.de

Kirchhoff Institute for Physics
Ruprecht-Karls-Universität Heidelberg

2025-03-25
NICE 2025

# Conventional Computing

- Significant demand from AI training and applications[1,2]

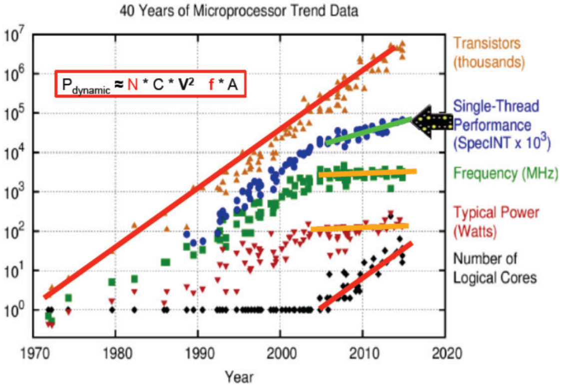



[1] N. Maslej et al., "The AI index 2024 annual report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, Apr. 2024

[2] S. Chen, "How much energy will AI really consume? the good, the bad and the unknown," Nature, vol. 639, no. 8053, pp. 22–24, 2025. DOI: 10.1038/d41586-025-00616-z

# Conventional Computing

- Significant demand from AI training and applications[1,2]
- Dennard (energy-density) scaling ended ∼2006
  - Dynamic power consumption, power wall, dark silicon, memory wall
  - → New computing stacks




[3]

[3] T. Conte, "IEEE rebooting computing initiative & international roadmap of devices and systems," in IEEE Rebooting Computer Architecture 2030 Workshop, 2015

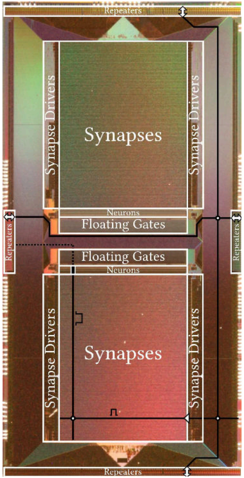
# Conventional Computing

- Significant demand from AI training and applications[1,2]
- Dennard (energy-density) scaling ended ∼2006
  - Dynamic power consumption, power wall, dark silicon, memory wall
  
  → New computing stacks
- Domain-specific hardware accelerators[4]:
  GPUs, FPGAs, and beyond

[4] W. J. Dally et al., “Domain-specific hardware accelerators,” Commun. ACM, vol. 63, no. 7, pp. 48–57, Jun. 2020. DOI: 10.1145/3361682

# Neuromorphic Hardware?

- Numerical simulation:
  - high level of parallelism is possible but latency to result is limited[1,2]
- SNNs follow an event-driven computing paradigm: sparse in space and time
- Neuromorphic hardware can complement simulation → SNN accelerators
- Functional modeling (ML-inspired?), but also in Computational Neuroscience:
  - Complex neuron dynamics, plasticity, long/repetitive experiments or guided reconfiguration!

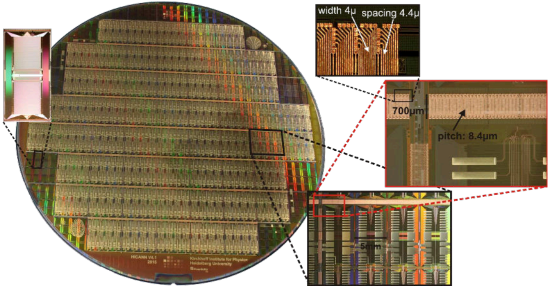



[1] A. C. Kurth et al., "Sub-realtime simulation of a neuronal network of natural density," Neuromorphic comput. eng., vol. 2, no. 2, p. 021 001, 2022. DOI: 10.1088/2634-4386/ac55fc

[2] J. Jordan et al., "Extremely scalable spiking neuronal network simulation code: From laptops to exascale computers," Frontiers in Neuroinformatics, vol. 12, p. 2, 2018. DOI: 10.3389/fninf.2018.00002

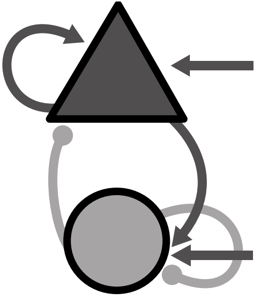
# BrainScaleS-1

- ($\leq$) 20× modules
- Wafer-scale integration (180 nm CMOS)
- 384 ASICs per 20 cm wafer
- 48 FPGAs, 40 GbE uplink to control cluster
- Typical speedup factor of 10’000

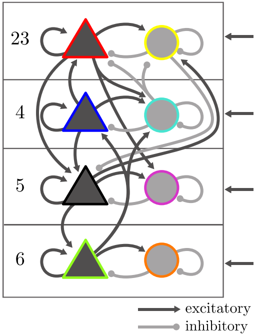



# Two Network Models from Computational Neuroscience

## Balanced Random Network[1]



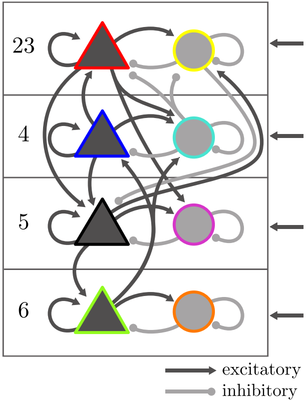
## Cortical Microcircuit Network Model[2]

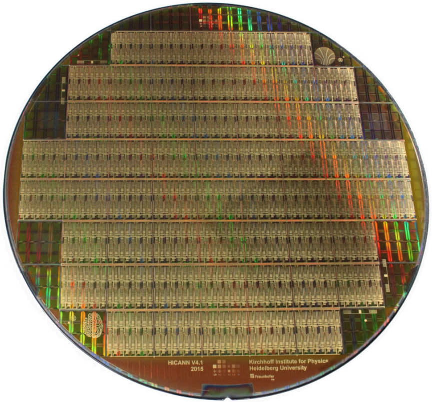



[1] N. Brunel, "Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons," Journal of Computational Neuroscience, vol. 8, no. 3, pp. 183–208, 2000. DOI: 10.1023/A:1008925309027

[2] T. C. Potjans and M. Diesmann, "The cell-type specific cortical microcircuit: Relating structure and activity in a full-scale spiking network modela," Cereb. Cortex, vol. 24, pp. 785–806, 3 2012. DOI: 10.1093/cercor/bhs358

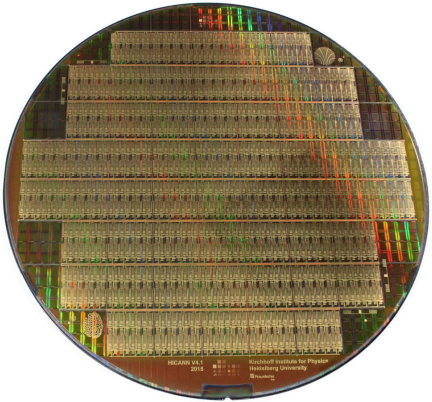
# Mapping the “Microcircuit” to BrainScaleS-1



# Mapping the “Microcircuit” to BrainScaleS-1

- 200k analog neuron circuits & 43M synapses
- Neurons follow configurable AdEx dynamics
- Configurable maximum fan-in implemented by linking multiple neuron circuits (up to 64 neurons resulting in 14k synapses)
- On-wafer sparse configurable circuit-switched network for asynchronous spike communication[1]
- Modeling API: PyNN (on top of the BSS-1 “Operating System”)

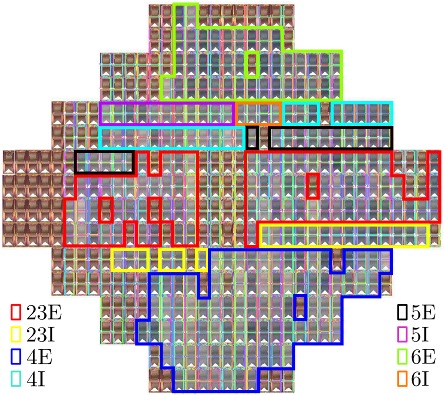



---

[1] H. Schmidt et al., “From clean room to machine room: Commissioning of the first-generation BrainScaleS wafer-scale neuromorphic system,” Neuromorphic comput. eng., vol. 3, no. 3, p. 034 013, 2023. DOI: 10.1088/2634-4386/acf7e4

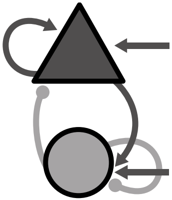
# Mapping the “Microcircuit” to BrainScaleS-1

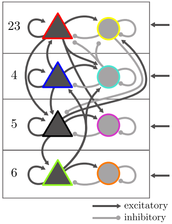
- 384 ASICs (each marked w/ white triangle at the bottom)
- Neuron placement represented by shading
- Darker shades indicate higher neuron counts
- Routed connections visualized as colored lines
- Colored borders indicate model populations



# Adapting Network Models to BrainScaleS-1 I

- Size of the network models:






| Balanced Random Network | Cortical Microcircuit |
|---|---|
| 12’400 neurons | 80’000 neurons |
| 15’625’000 synapses | 300’000’000 synapses |

- Number of model neurons $<$ neuron circuits per wafer, but average neuron fan-in requires interlinked neuron circuits.

$\rightarrow$ Reduced amount of (model) neurons available.

# Adapting Network Models to BrainScaleS-1 II

⇒ Downscaling of neuron count and in-degree
  - maintaining the original connectivity probability, and
  - compensating[2] for the reduced input by linear weight increase following the approach by Albada et al.[1]

- Due to random network structure, some additional "synapse loss" occurs across all populations.
  - We incorporate this network model "distortions" into our simulations comprising
    - 2'083 neurons and 690'157 synapses (Balanced Random Network)
    - 7'712 neurons and 2'373'933 synapses (Cortical Microcircuit)

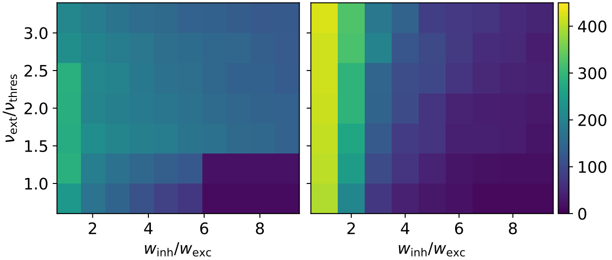
[1] S. J. van Albada et al., "Scalability of asynchronous networks is limited by one-to-one mapping between effective connectivity and correlations," PLoS Comput. Biol., vol. 11, pp. 1–37, Sep. 2015. DOI: 10.1371/journal.pcbi.1004490

[2] D. Brüderle et al., "A comprehensive workflow for general-purpose neural modeling with highly configurable neuromorphic hardware systems," Biological Cybernetics, vol. 104, pp. 263–296, 4 2011

# Result: (Downscaled) Balanced Random Network

- Varying relative inhibitory weight and external input spike rates.
- For firing rates exceeding 50 Hz, saturation effects on the hardware introduce deviations in network behavior.

Mean firing rates of neurons

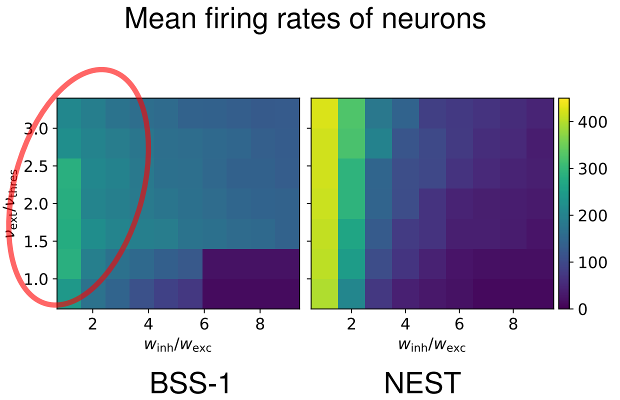



BSS-1 NEST

# Result: (Downscaled) Balanced Random Network



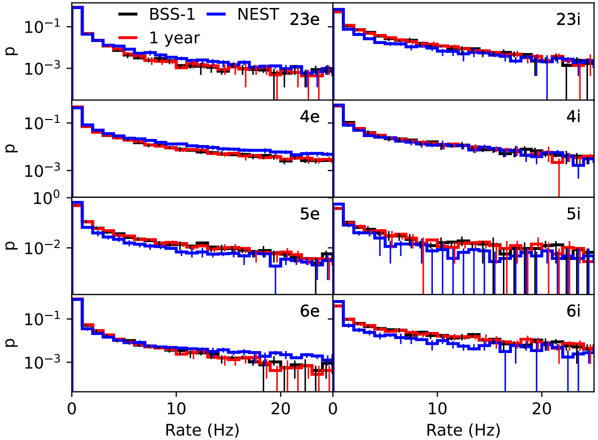
- Varying relative inhibitory weight and external input spike rates.
- For firing rates exceeding 50 Hz, saturation effects on the hardware introduce deviations in network behavior.



# Result: (Downscaled) Cortical Microcircuit

- Results are extracted from a 9 s interval of biological time, starting 1 s after the experiment onset (BSS-1 & NEST).
- Reevaluation after 53 min of wall-clock time on BSS-1.

Firing rate distribution of neurons across eight network model populations



# Result: (Downscaled) Cortical Microcircuit II

| Simulator | Performance ($10^9$ synaptic event/s) | Energy (µJ/synaptic event) |
|---|---|---|
| BrainScaleS-1 | 162 | < 0.012 |
| NeuroAIx-Framework[0,1] | 19 | 0.048 |
| CsNN[0,2] | 3.8 | 0.783 |
| NEST[0,3] | 1.8 | 0.48 |
| SpiNNaker[4] | 0.9 | 0.6 |

[0] Values are estimated from the reported speedup factor and the network behavior of the full-scale model with external Poisson inputs.

[1] K. Kauth et al., "neuroAIx-framework: Design of future neuroscience simulation systems exhibiting execution of the cortical microcircuit model 20× faster than biological real-time," Front. Comput. Neurosci., vol. 17, p. 1 144 143, 2023. DOI: 10.3389/fncom.2023.1144143

[2] A. Heittmann et al., "Simulating the cortical microcircuit significantly faster than real time on the ibm inc-3000 neural supercomputer," Front. Neurosci., vol. 15, p. 728 460, 2022. DOI: 10.3389/fnins.2021.728460

[3] A. C. Kurth et al., "Sub-realtime simulation of a neuronal network of natural density," Neuromorphic comput. eng., vol. 2, no. 2, p. 021 001, 2022. DOI: 10.1088/2634-4386/ac55fc

[4] O. Rhodes et al., "Real-time cortical simulation on neuromorphic hardware," Philos. Trans. R. Soc. A, vol. 378, no. 2164, p. 20 190 160, 2020. DOI: 10.1098/rsta.2019.0160

# Conclusion

- Speedup from physical emulation most evident for long/repetitive emulations
- Main operational overhead introduced by configuration and data transfer (e.g., read out of recorded observables)
- Comparably low energy consumption of BrainScaleS-1 can still yield advantages in comparison to numerical simulation
- Network model size limitations come from neuron, synapse, and routing resources
  - Biological connection densities difficult to efficiently scale beyond wafer-scale
- Co-execution approach:
  - validation and network topology exploration in simulation
  - neuromorphic backend handles continuous time emulation, extended-duration experiments, and iterative parameter sweeps

# Outlook

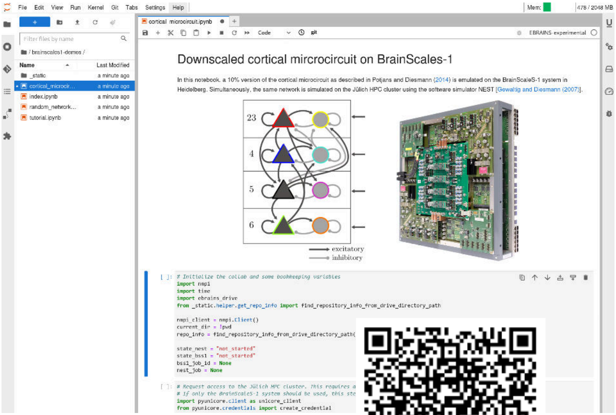
- Area efficiency limited by use of “plastic” synapses in fully static networks
  $\rightarrow$ dedicated static (higher-density) synapses in future hardware systems?
- Newer technology node! (BrainScaleS-1 uses 180 nm CMOS)
- No plasticity was involved, i.e. the model dynamics are numerically “cheap”; introducing, e.g., synaptic plasticity would amplify the benefit of physical emulation.
- No “dependent” reconfiguration was used — neuromorphic hardware can also deliver in latency-to-result use cases.

# BrainScaleS is an Open Research Platform

- Integrated into the EBRAINS Software Distribution

ESD

- Access to accelerated neuromorphic BrainScaleS via EBRAINS
- Register for EBRAINS:



# References I

N. Maslej et al., “The AI index 2024 annual report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, Apr. 2024.

S. Chen, “How much energy will AI really consume? the good, the bad and the unknown,” Nature, vol. 639, no. 8053, pp. 22–24, 2025. DOI: 10.1038/d41586-025-00616-z.

T. Conte, “IEEE rebooting computing initiative & international roadmap of devices and systems,” in IEEE Rebooting Computer Architecture 2030 Workshop, 2015.

W. J. Dally, Y. Turakhia, and S. Han, “Domain-specific hardware accelerators,” Commun. ACM, vol. 63, no. 7, pp. 48–57, Jun. 2020. DOI: 10.1145/3361682.

A. C. Kurth et al., “Sub-realtime simulation of a neuronal network of natural density,” Neuromorphic comput. eng., vol. 2, no. 2, p. 021 001, 2022. DOI: 10.1088/2634-4386/ac55fc.

J. Jordan et al., “Extremely scalable spiking neuronal network simulation code: From laptops to exascale computers,” Frontiers in Neuroinformatics, vol. 12, p. 2, 2018. DOI: 10.3389/fninf.2018.00002.

N. Brunel, “Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons,” Journal of Computational Neuroscience, vol. 8, no. 3, pp. 183–208, 2000. DOI: 10.1023/A:1008925309027.

# References II


T. C. Potjans and M. Diesmann, “The cell-type specific cortical microcircuit: Relating structure and activity in a full-scale spiking network modela,” Cereb. Cortex, vol. 24, pp. 785–806, 3 2012. DOI: 10.1093/cercor/bhs358.

H. Schmidt et al., “From clean room to machine room: Commissioning of the first-generation BrainScaleS wafer-scale neuromorphic system,” Neuromorphic comput. eng., vol. 3, no. 3, p. 034 013, 2023. DOI: 10.1088/2634-4386/acf7e4.

S. J. van Albada, M. Helias, and M. Diesmann, “Scalability of asynchronous networks is limited by one-to-one mapping between effective connectivity and correlations,” PLoS Comput. Biol., vol. 11, pp. 1–37, Sep. 2015. DOI: 10.1371/journal.pcbi.1004490.

D. Brüderle et al., “A comprehensive workflow for general-purpose neural modeling with highly configurable neuromorphic hardware systems,” Biological Cybernetics, vol. 104, pp. 263–296, 4 2011.

K. Kauth et al., “neuroAIx-framework: Design of future neuroscience simulation systems exhibiting execution of the cortical microcircuit model 20× faster than biological real-time,” Front. Comput. Neurosci., vol. 17, p. 1 144 143, 2023. DOI: 10.3389/fncom.2023.1144143.

A. Heittmann et al., “Simulating the cortical microcircuit significantly faster than real time on the ibm inc-3000 neural supercomputer,” Front. Neurosci., vol. 15, p. 728 460, 2022. DOI: 10.3389/fnins.2021.728460.

O. Rhodes et al., “Real-time cortical simulation on neuromorphic hardware,” Philos. Trans. R. Soc. A, vol. 378, no. 2164, p. 20 190 160, 2020. DOI: 10.1098/rsta.2019.0160.

# References III

H. Schmidt, “Large-scale experiments on wafer-scale neuromorphic hardware,” Ph.D. dissertation, Ruprecht-Karls-Universität Heidelberg, 2024. DOI: 10.11588/heidok.00034446.