



Exceptional service in the national interest

MEMORY TRADE-OFFS IN NEUROMORPHIC COMMUNICATION STRATEGIES OF THE FLYWIRE CONNECTOME ON LOIHI 2

Felix Wang (presenter), Brad Theilman, Fred Rothganger, William Severa (UTSA), Craig Vineyard, Brad Aimone

Neuro-Inspired Computational Elements Conference

Mar. 2026, Atlanta, GA



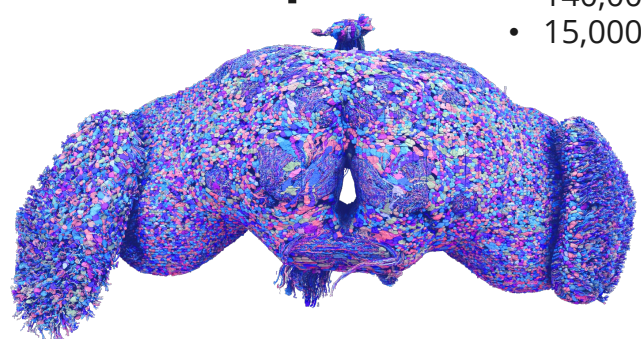
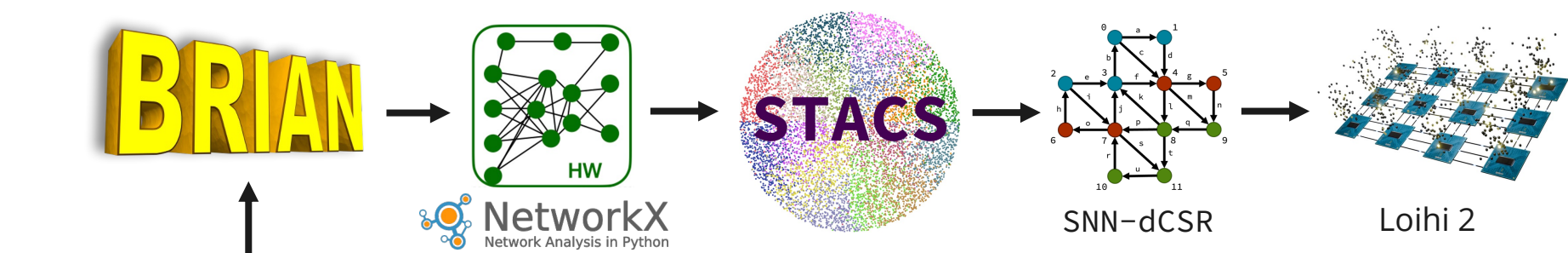
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2026-18935C

FRUIT FLY BRAIN ON A CHIP



- We developed an advanced neuromorphic compilation toolchain to support mapping a complex biological connectome in Brian 2 for execution on Loihi 2



<https://codex.flywire.ai/>

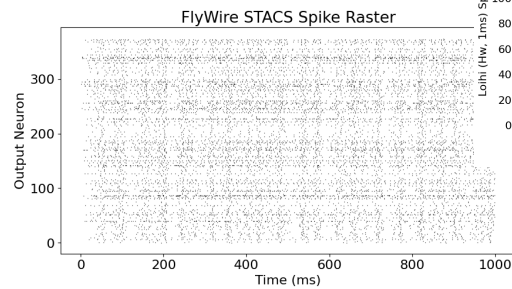
P. K. Shiu et al., "A drosophila computational brain model reveals sensorimotor processing," Nature, vol. 634, no. 8032, pp. 210–219, 2024

Network model with roughly:

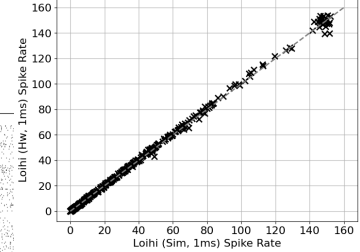
- 140,000 Neurons
- 15,000,000 Synapses

Mapped onto 12 chips, 1440 neurocores, directly through NxCore API

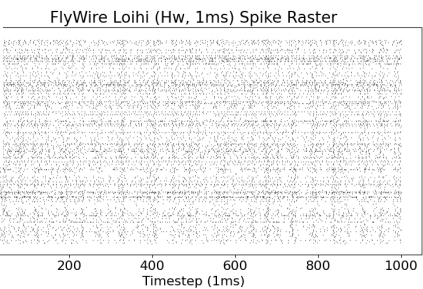
Reference simulation output (Brian 2, STACS)



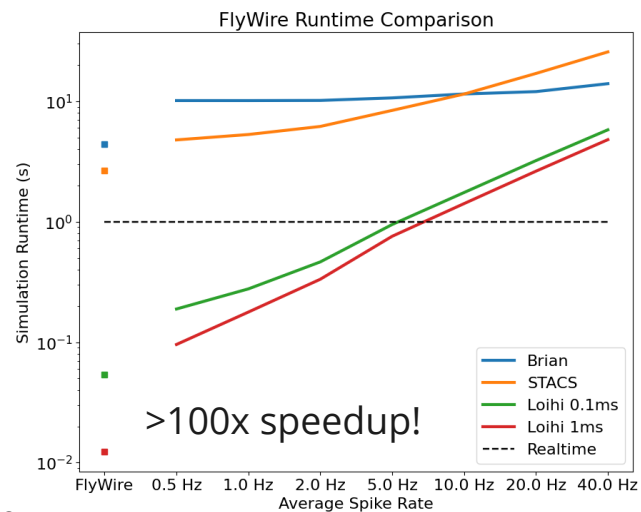
FlyWire Comparison: Loihi Hw vs Sim (1ms)



Output through spike counters on Loihi 2



We validated our hardware implementation against expected spike rates

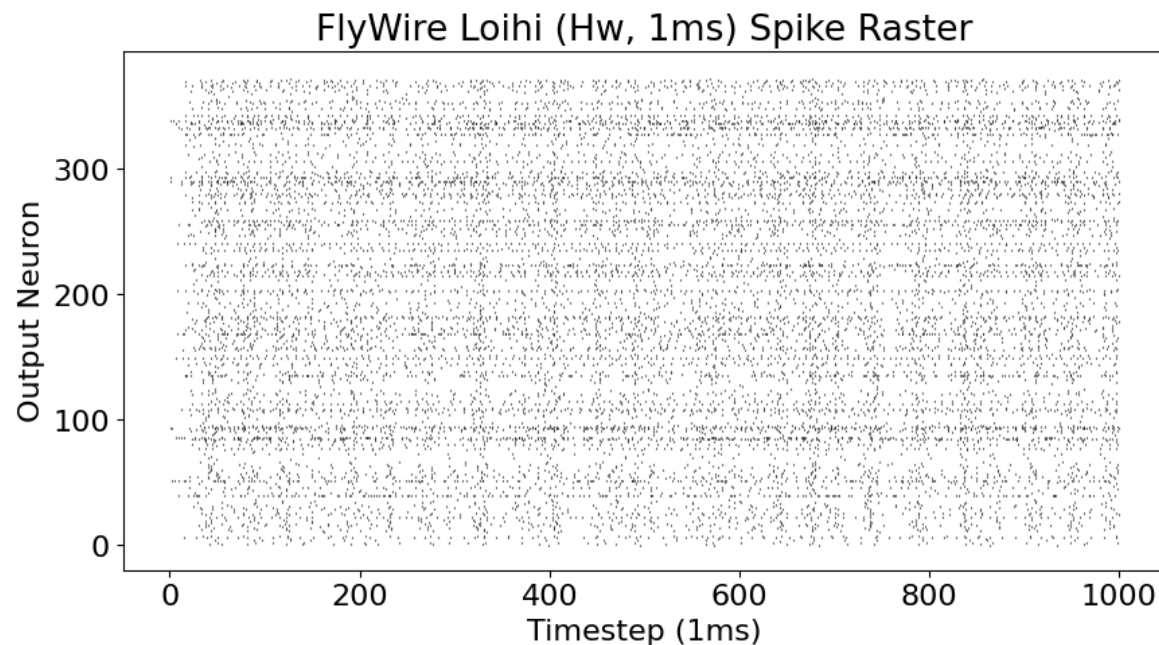


We achieved significant spike activity-dependent acceleration of biological network simulations

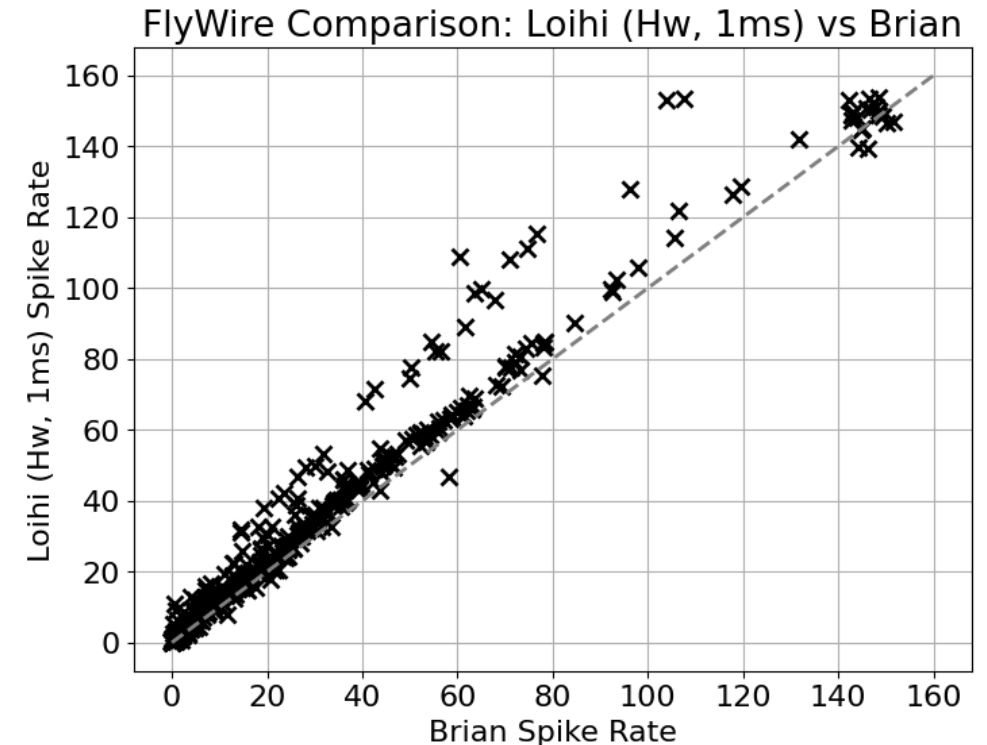
Preprint results in Wang et al., "Neuromorphic Simulation of Drosophila Melanogaster Brain Connectome on Loihi 2", arXiv, <https://arxiv.org/abs/2508.16792>

FLYWIRE NETWORK MODEL

- We replicated and validated our model against a Brian 2 point-neuron network implementation and computational neuroscience sugar neuron experiment (Shiu et al. 2024), demonstrating connectome-constrained behavior



Spike raster from the sugar neuron experiment on Loihi 2



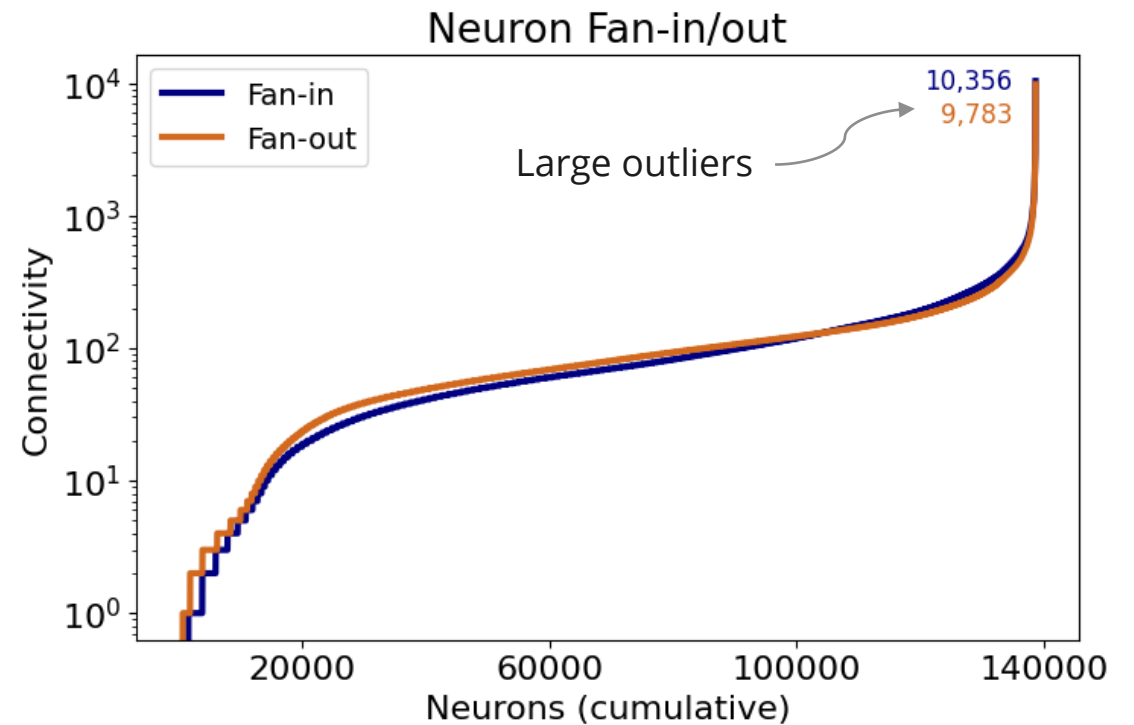
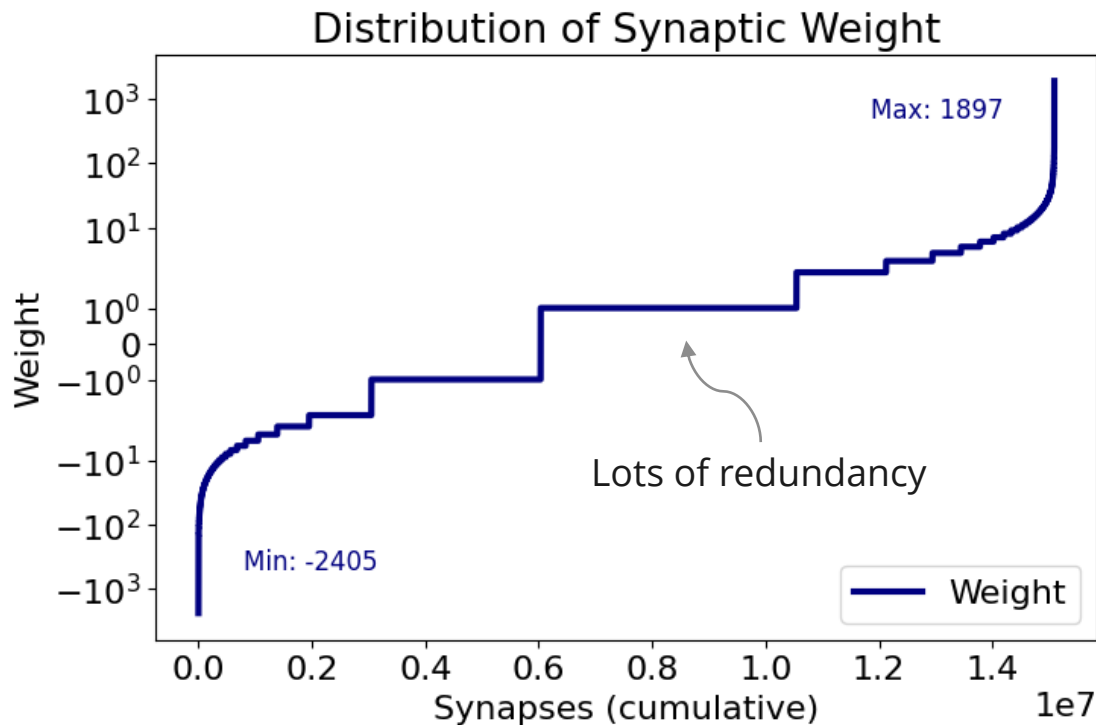
Parity plot comparing spike rates between Loihi 2 hardware and Brian 2 simulation
Index-matched pairs of neurons are marked with 'x', parity is shown as dashed line



SYNAPTIC MAPPING CHALLENGES



- To fit the FlyWire network with its large and irregular graph structure onto the highly distributed, memory-limited neurocores of Loihi 2, we focused on the synapses
 - Orders of magnitude difference: 15,000,000 synapses \gg 140,000 neurons
 - Each neurocore only has 128kB of locally available memory (analogous to L1 cache)





PARTITIONED NETWORK REPRESENTATION

- We leveraged the SNN extension to dCSR from the STACS simulator to transform the network model

- Brian 2: Global indexing
Shared memory



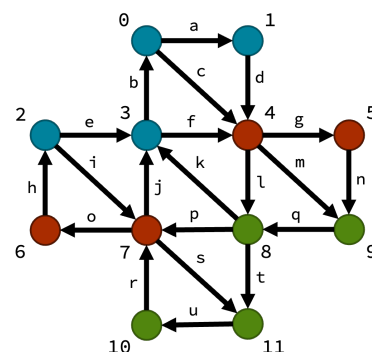
- STACS: Global indexing
Distributed memory



- Loihi 2: Local indexing
Distributed memory

- Each partition provides “self-contained” state information (except for external events)

The SNN-dCSR format supports network (re)partitioning, making it suitable for computational parallelism, whether its target platform is between nodes on an HPC system or between chips/cores on neuromorphic hardware



dist	adjcy.0	adjcy.1	adjcy.2
0 0	1 3 4	0 1 3 5 8 9	3 4 7 9 11
4 13	0 4	4 9	4 5 8
8 29	3 6 7	2 7	7 11
12 42	0 2 4 7 8	2 3 6 8 10 11	7 8 10

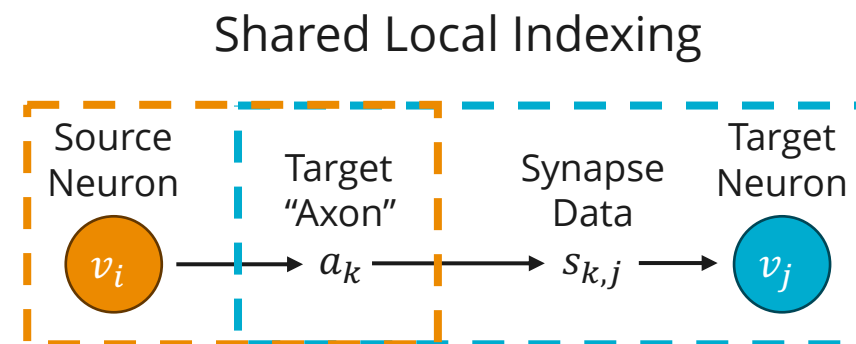
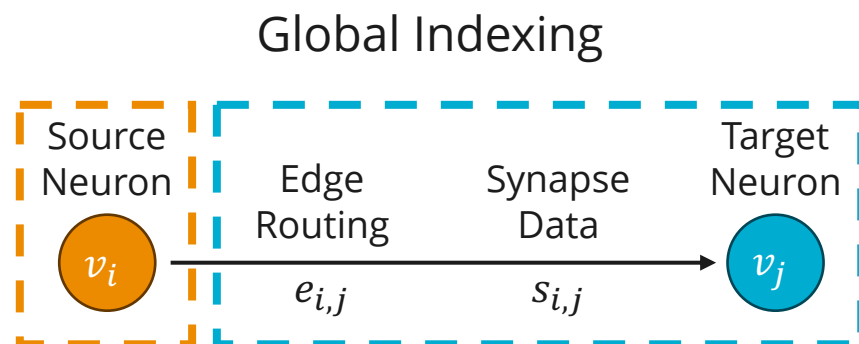
state.0	state.1	state.2
0 \emptyset b \emptyset	4 c d f $\emptyset \emptyset \emptyset$	8 \emptyset l \emptyset q \emptyset
1 a \emptyset	5 g \emptyset	9 m n \emptyset
2 \emptyset h \emptyset	6 \emptyset o	10 \emptyset u
3 \emptyset e \emptyset j k	7 i $\emptyset \emptyset$ p r \emptyset	11 s t \emptyset

Example 3-way partitioning of a simple network with 12 vertices and 21 directed edges using the SNN-dCSR format, directed edges with associated state are bolded

MEMORY ADDRESSING AND COMMUNICATION ON LOIH1 2



- The distributed memory and locally indexed addressing of the neurocores is also used in spike communication, which are mediated through local "axon" indexes
 - This intermediate addressing system provides an additional level of indirection and allows for better memory efficiency (no need for global indexes)



- Spikes may be routed by their source neuron (e.g., address-event representation)
- Connectivity and synapse data are more coupled
- Memory burden is light on the source core, but heavier on the target core (e.g., routing tables)

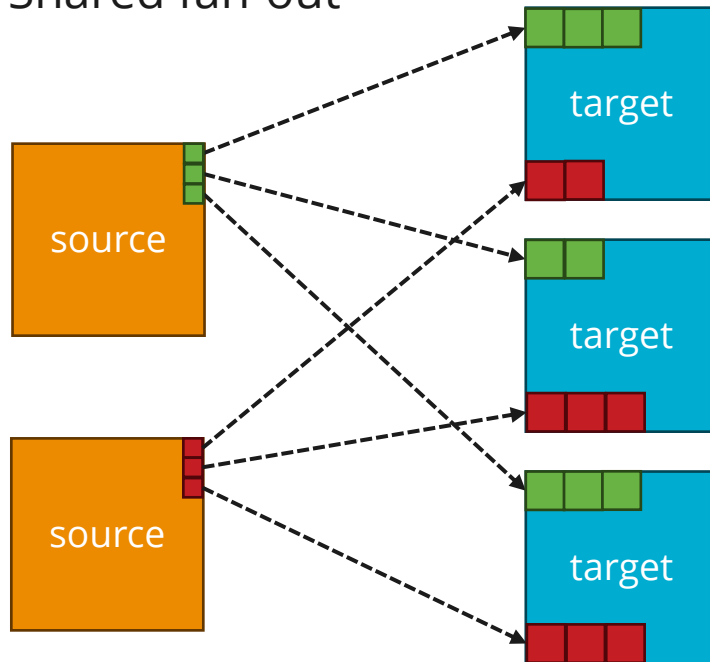
- Spikes are routed by their target information (i.e., destination-based)
- Connectivity and synapse data are decoupled
- More memory overhead on the source core, but increased memory flexibility on the target core

SPIKING COMMUNICATION STRATEGIES



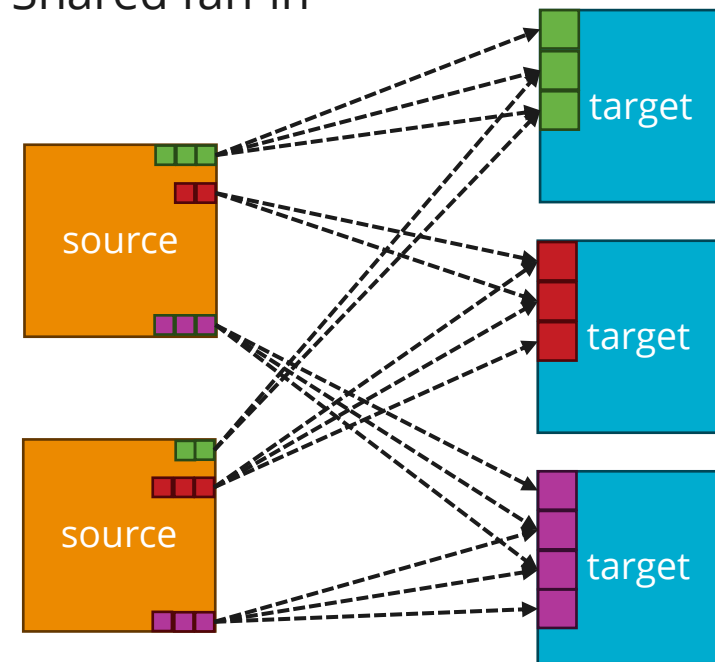
- Neuromorphic hardware is optimized for sending small messages, so we explore a counter-intuitive communication strategy to trade-off messages for memory

Shared fan-out



- One message per source-target neurocore
- Multiple target weights per message
- Weights are not shared between neurons

Shared fan-in

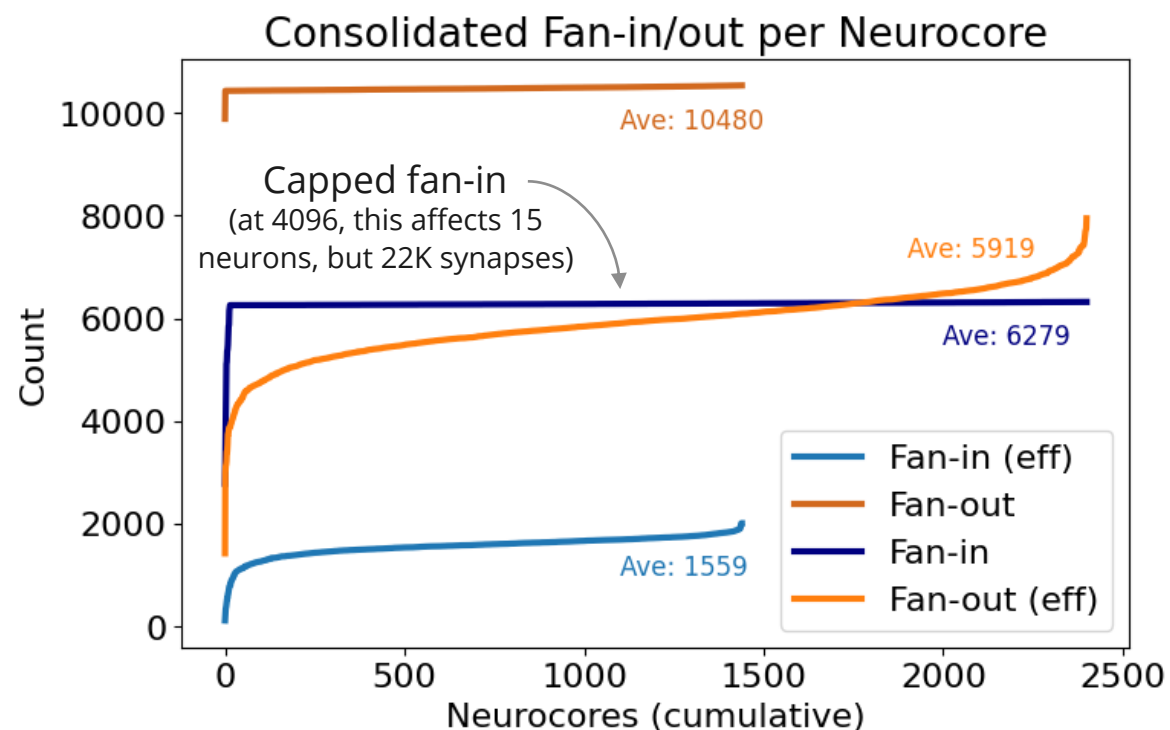
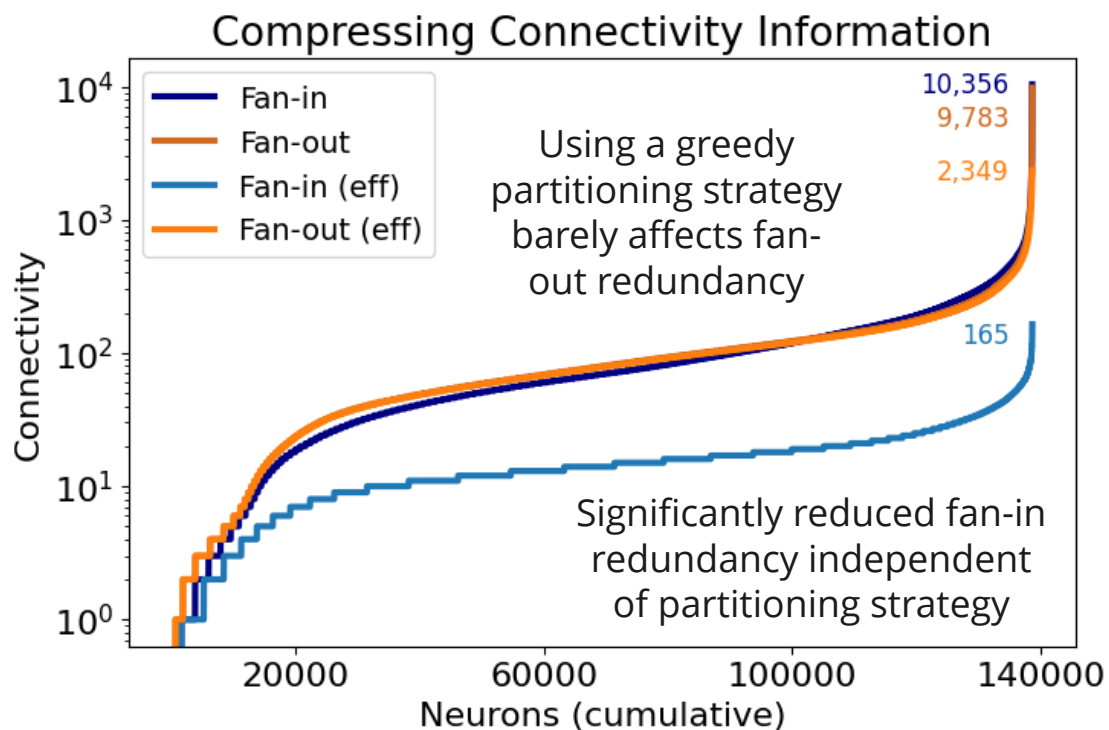


- One message per source-target neuron
- Singular target weight per message
- Weights can be shared between neurons

NEUROCORE MAPPING COMPARISON



- The choice of communication strategy affects the memory compression available
 - Shared fan-in allows the same network to fit onto significantly fewer neurocores
 - Shared fan-out still runs into hard limits in representing the outlier neurons, axon indexes are also more dependent on the partitioning (extra bookkeeping)



SPIKE ACTIVITY-DEPENDENT SCALING EXPERIMENT



- We performed a simple activity-dependent scaling experiment increasing the baseline background spiking rate (per neuron)
 - Neuromorphic platforms such as Loihi 2 generally scale with respect to spike volume
 - Runtimes were measured without spike probes to reduce synchronization overhead (note: new asynchronous ethernet-based streaming I/O improves runtime)

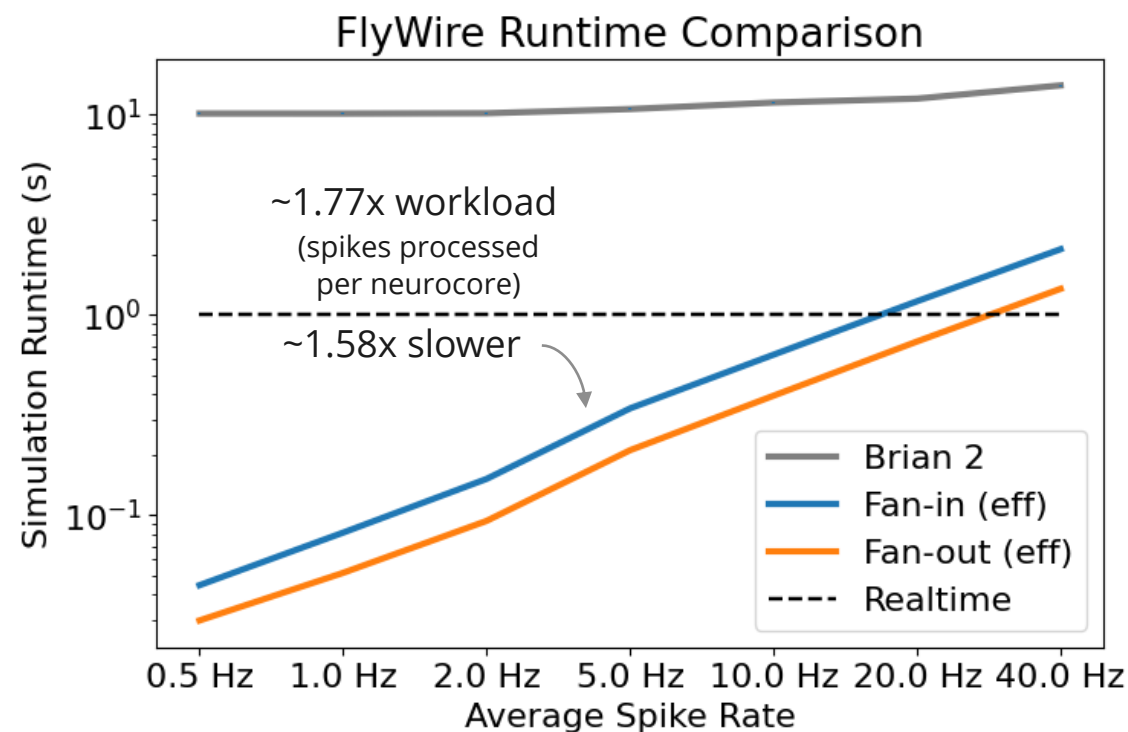
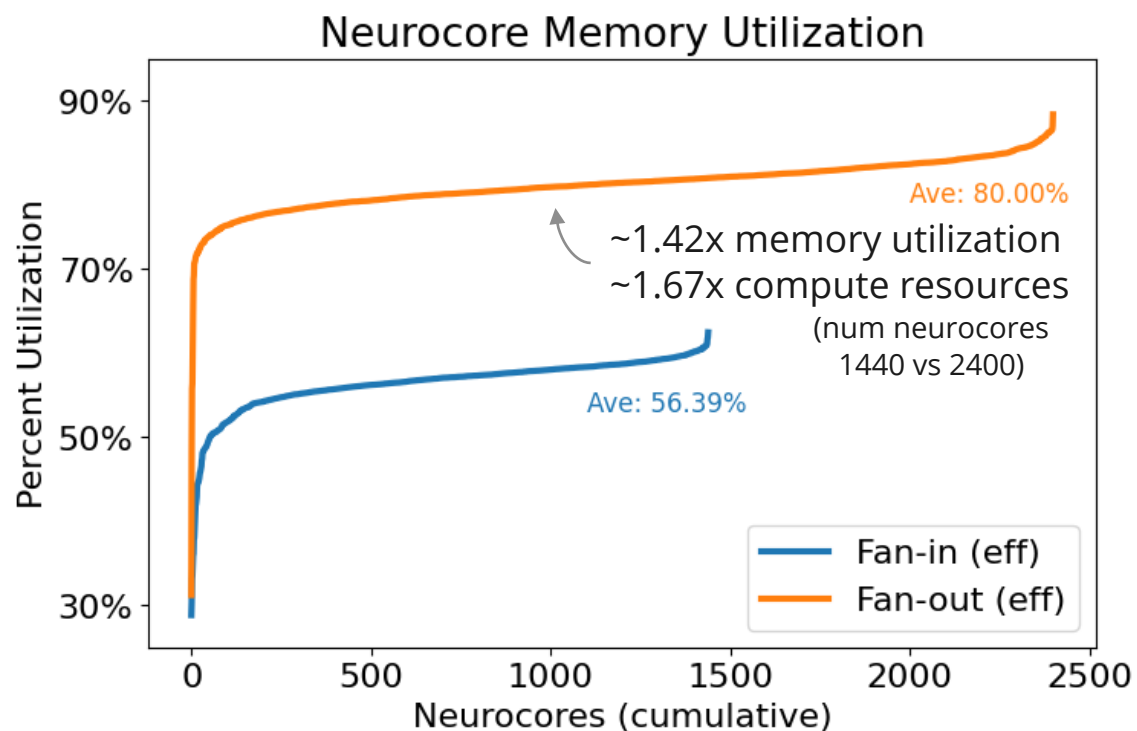
Platform / Spike Rate	0.5Hz	1.0Hz	2.0Hz	5.0Hz	10.0Hz	20.0Hz	40.0Hz
Brian 2	10.128	10.127	10.158	10.662	11.49	12.02	13.992
Loihi 2 (fan-in)	0.0447	0.0819	0.1518	0.3414	0.6329	1.1702	2.1283
Loihi 2 (fan-out)	0.0298	0.0515	0.0936	0.211	0.395	0.7369	1.3502

Runtime performance comparison between different communication strategies for different baseline background spiking rates (per 1 second simulated time, units are in seconds)



MEMORY UTILIZATION AND RUNTIME TRADE-OFFS

- We found that both the runtime performance and relative difference between the two communication strategies scaled roughly linearly with respect to spike volume
 - For the shared fan-in strategy, fewer computational resources performed roughly the same total workload at proportionally the same latency (e.g., preserving parallel cost)



SUMMARY AND TAKEAWAYS



- Neuromorphic computing refers to a class of brain-inspired emerging technologies characterized by distributed, sparse, event-based computation for enabling low-latency neural algorithms on energy-efficient computing architectures
 - For the FlyWire connectome model, we were able to achieve faster than realtime execution, and two orders of magnitude faster than the conventional simulator
 - Useful tool for advancing neuro-inspired AI and computational neuroscience
- Many important software and tooling challenges for taking full advantage of neuromorphic hardware
 - Understanding the space of trade-offs, which may be different than on conventional hardware
 - Use of scalable intermediate representations
 - Robust software infrastructure for compilation

NERL at Sandia
<https://neuroscience.sandia.gov/>



BACKUP: FLYWIRE NETWORK MODEL DETAILS



$$\frac{dv}{dt} = (v_0 - v + g)/\tau_m \quad \text{if } v > v_{th} : \quad v = v_r$$

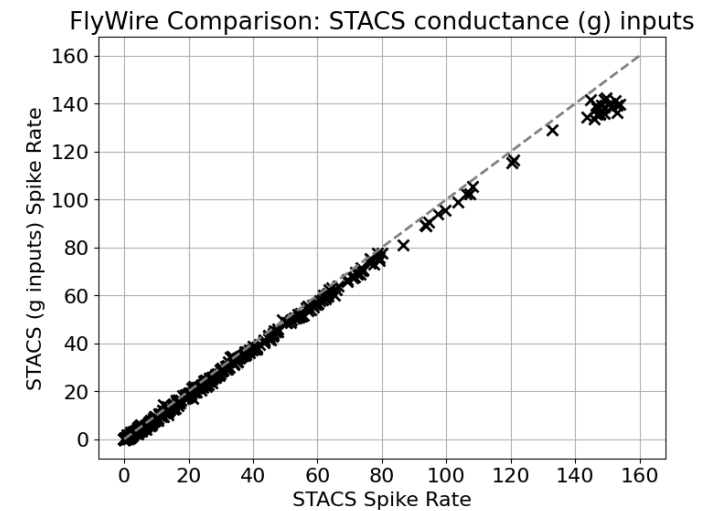
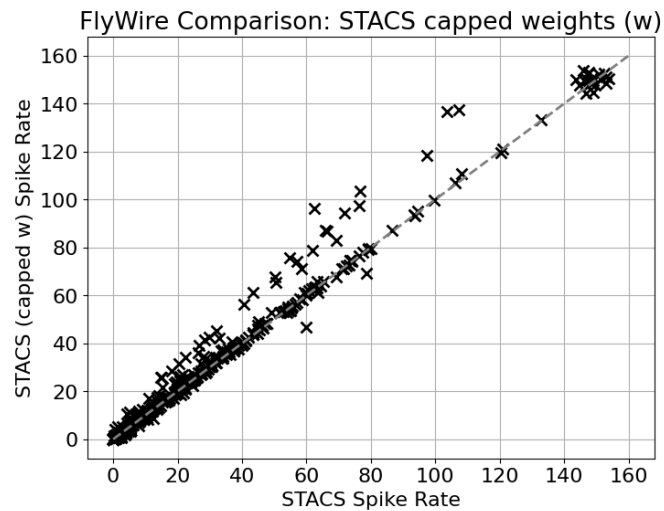
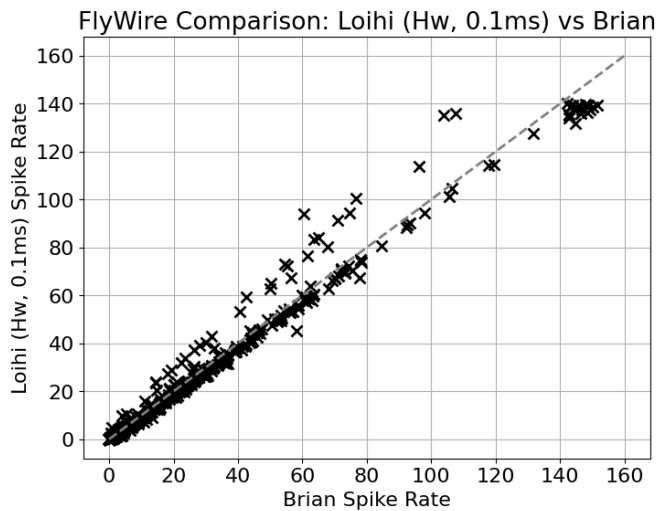
$$\frac{dg}{dt} = -g/\tau_g \quad \quad \quad g = 0$$

(unless refractory) spike and enter refractory for τ_{ref} (on incoming spike)

$$\frac{dg}{dt} = w_{ij} \sum_m \delta(t - t_i^m)$$

$\tau_g = 5ms$
 $\tau_{ref} = 2.2ms$
 $v_{th} = 7mV$
 $v_0 = v_r = 0mV$

FlyWire neuron state dynamics (left), spiking conditional (center), and synaptic input (right)



Parity plots comparing spike rates between Loihi 2 hardware and Brian 2 simulation (left), STACS simulations investigating capped weights (center) and conductance-only inputs (right)

BACKUP: NETWORK CONNECTIVITY AS A SPARSE MATRIX



		Sparse Matrix										Compressed Sparse Row	Distributed Compressed Sparse Row
		0	1	2	3	4	5	6	7	8	9		
I	0					a			b			Row: [0, 2, 4, 6, 9, 14,	Part : [0, 4, 7, 10]
	1			c			d					17, 19, 21, 23, 26]	
	2							e		f		Col: [4, 7, 2, 5, 6, 8, 2,	Row I : [0, 2, 4, 6, 9]
II	3			g					h		i	7, 9, 0, 3, 5, 7, 9,	Row II : [0, 5, 8, 10]
	4	j			k		l		m		n	1, 2, 6, 0, 4, 3, 5,	Row III: [0, 2, 4, 7]
	5		o	p				q				5, 9, 0, 1, 8]	Col I : [4, 7, 2, 5, 6, 8, 2, 7, 9]
III	6	r				s						Val: [a, b, c, d, e, f, g,	Col II : [0, 3, 5, 7, 9, 1, 2, 6, 0, 4]
	7				t		u					h, i, j, k, l, m, n,	Col III: [3, 5, 5, 9, 0, 1, 8]
	8						v				w	o, p, q, r, s, t, u,	Val I : [a, b, c, d, e, f, g, h, i]
	9	x	y								z	v, w, x, y, z]	Val II : [j, k, l, m, n, o, p, q, r, s]
													Val III: [t, u, v, w, x, y, z]

Sparse matrices are efficiently represented by the CSR format

- Numbered values correspond to row and column indexes
- Lettered values correspond to non-zero entries
- Roman numerals correspond to partitions
- Entries within a row are uniquely colored for ease of correlation within their corresponding arrays

We extend the dCSR format for SNNs

- ‘Part’ information in `dist` file
- ‘Row’ information implicit in `adjcy` files (adjacency lists)
- ‘Col’ information explicit in `adjcy`
- ‘Val’ information in `state` files