# Neuromorphics

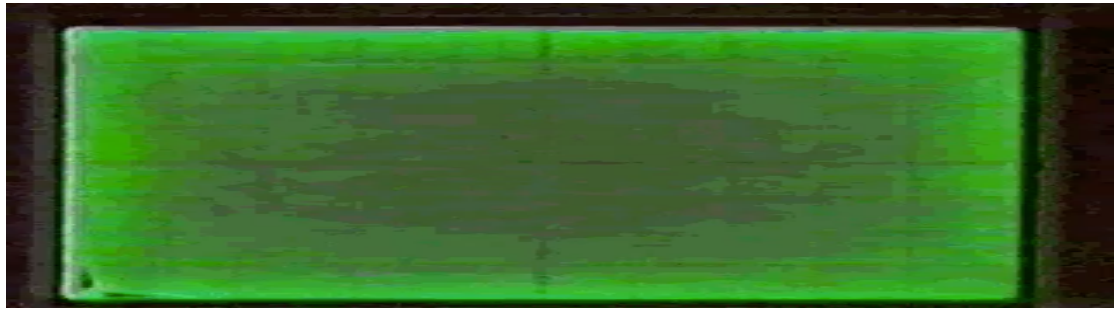
## SYDE 556/750

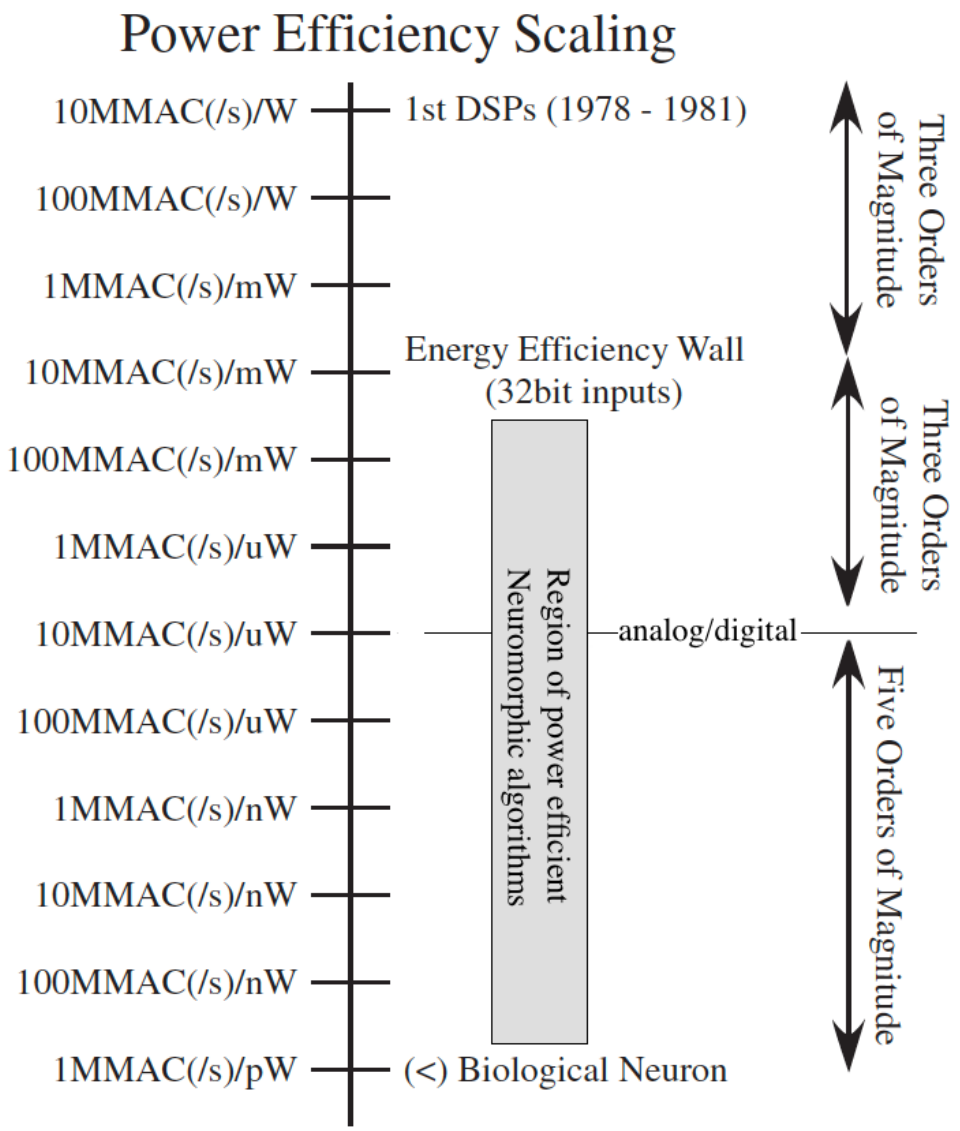
Chris Eliasmith

# Motivation

## Biological brains are event driven and energy efficient

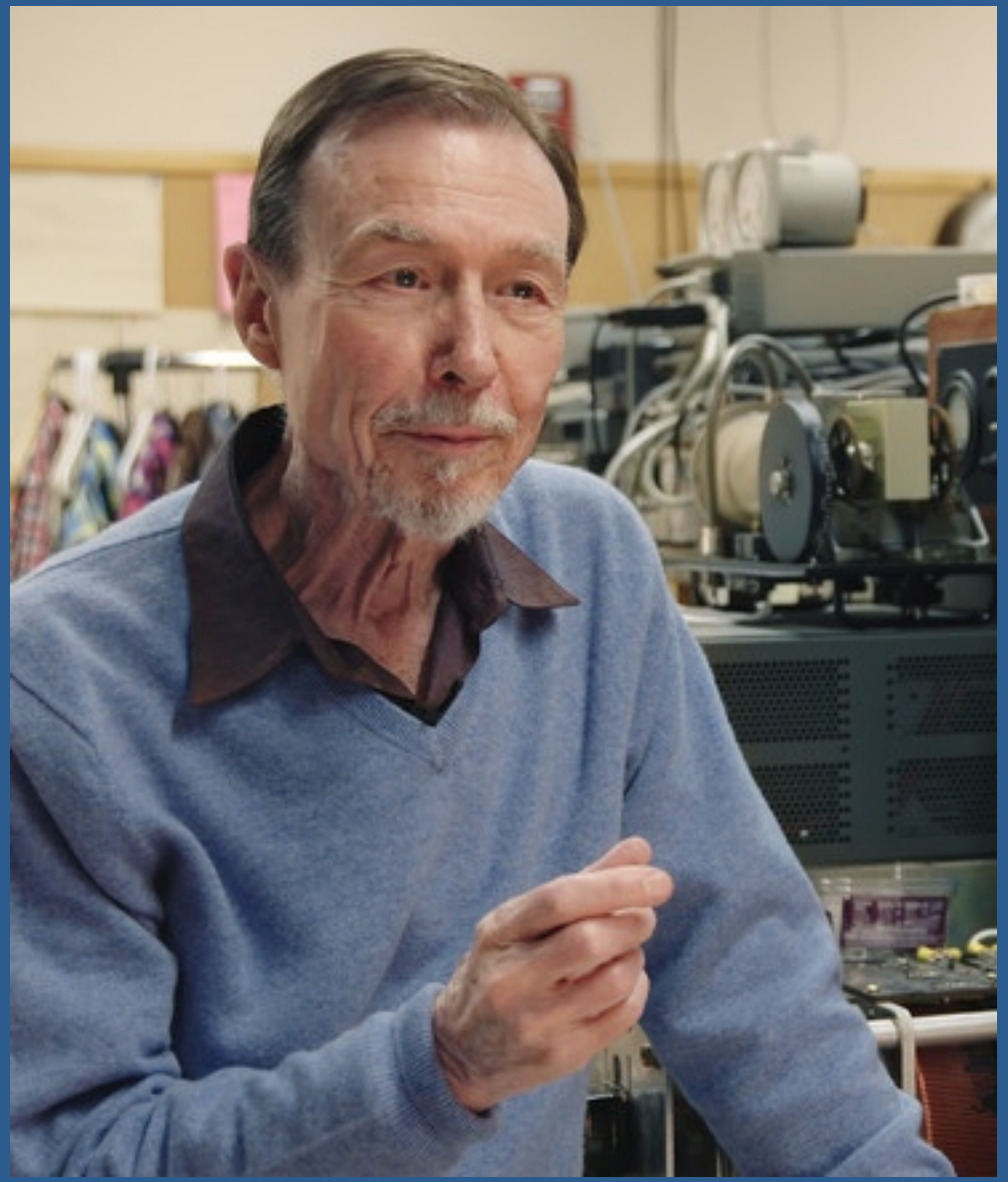
- $10^8$ less power than 32 bit digital
- Lots of room for HW and algorithm innovations now






Based on Hasler & Marr, 2013

# Carver Mead



## People

G. Cauwenberghs

T. Delbruck

J. Hasler

K. Boahen

S. Liu

M. Mahowald

R. Sarpeshkar

L. Watts

## Companies

Foveon
(Sigma)

Synaptics
(6 bn devices)

Impinj
($600M)

Actel
($430M)

## Tech

iPhone 4
(noise chip)

Spiking
cameras/
retina

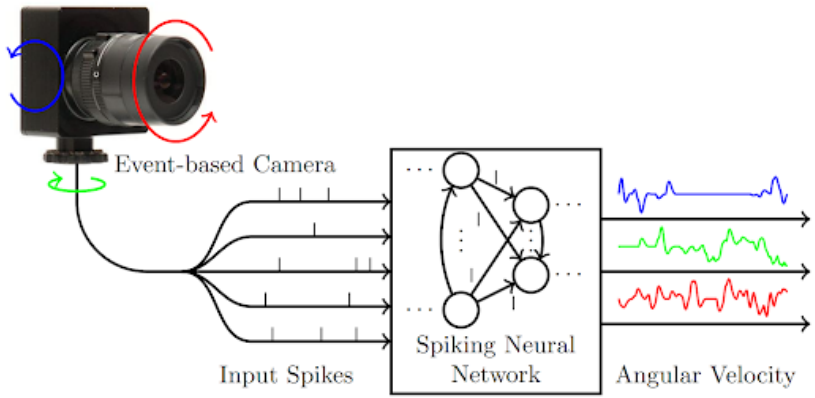
Spiking
cochlea

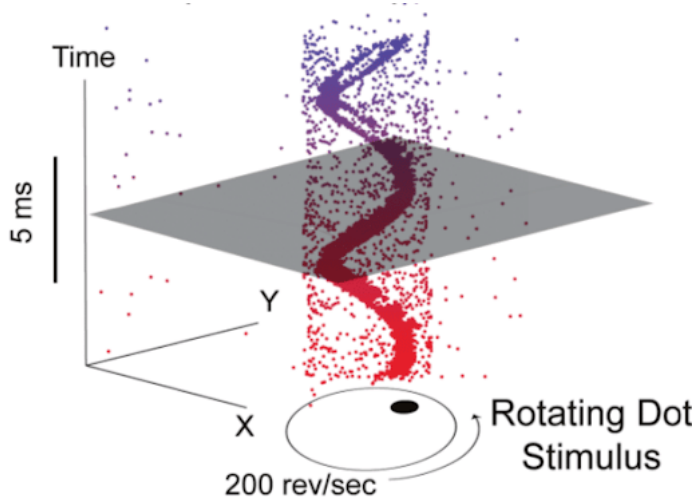
Neuromorphic
Haptic Sensors

Trackball

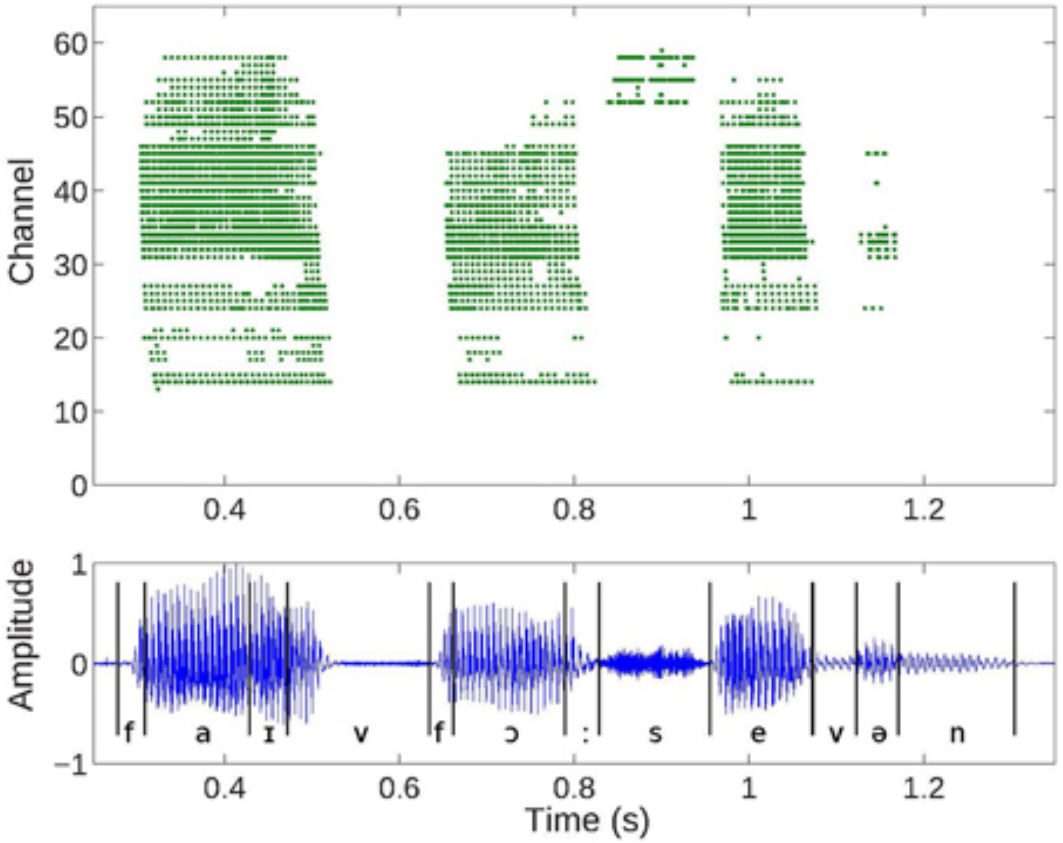
# Early days (1980-2000): Sensors

## Spiking retina (camera)



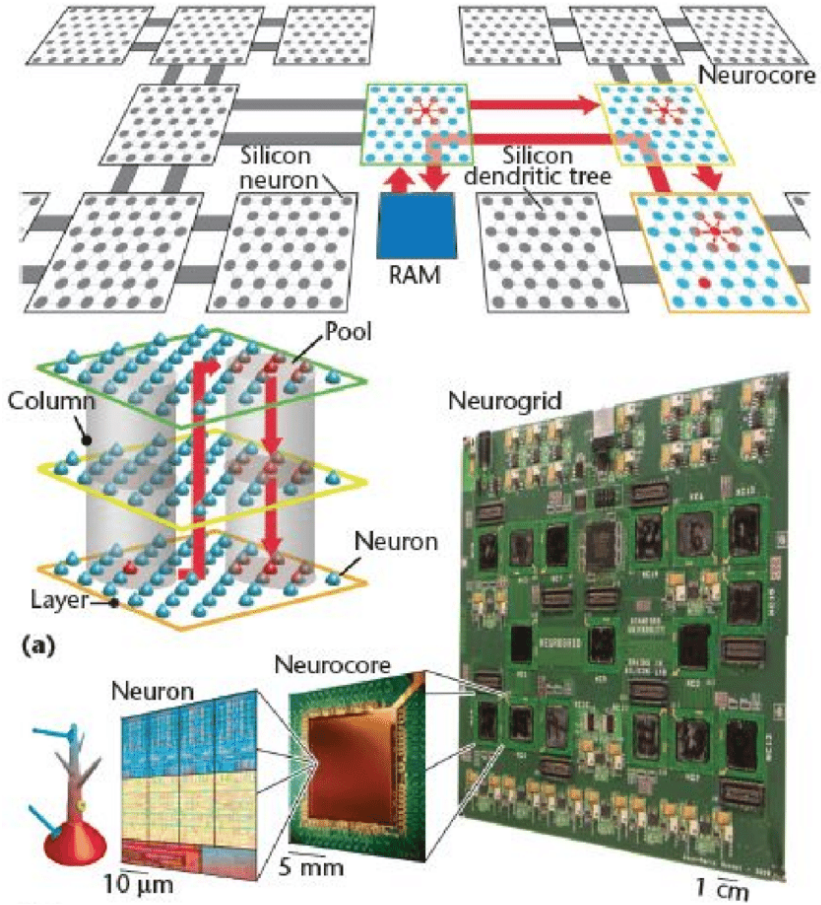

## Spiking cochlea



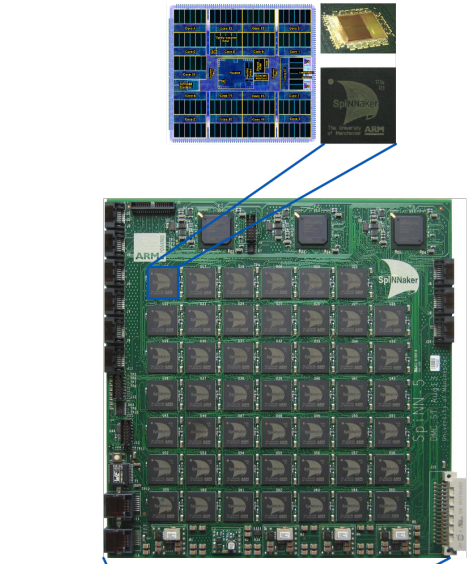
# Maturation (2000-2015): Cortical Circuits

## NeuroGrid (2009)

- 16 chips
- 1M neurons
- 6B synapses
- 2W




## SpiNNaker (2005)

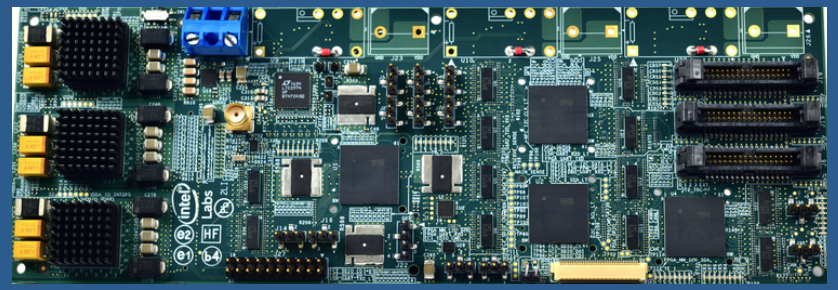


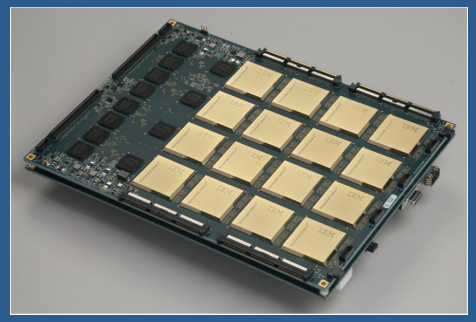
- 18 cores/chp
- 48 chp/bd
- 100 bd/rack
- 1M cores (2018)
- 1B neurons
- 100kW

# Recent Advances (2015-2020)

## Hardware

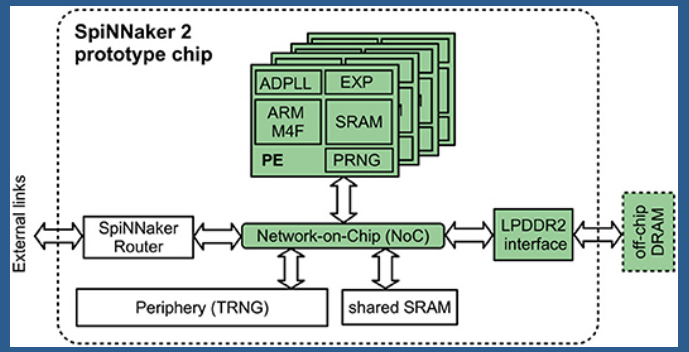
Intel Loihi





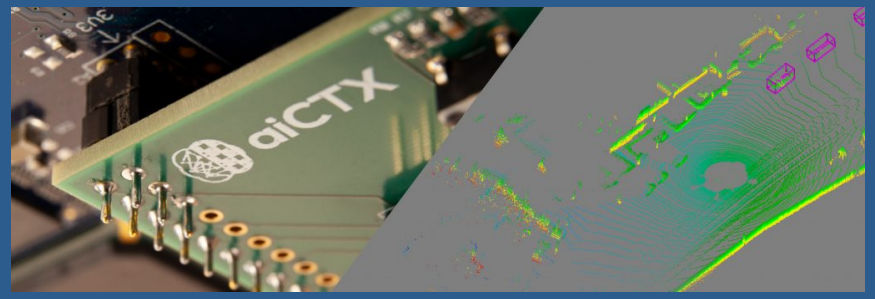
IBM TrueNorth

Spinnaker 2

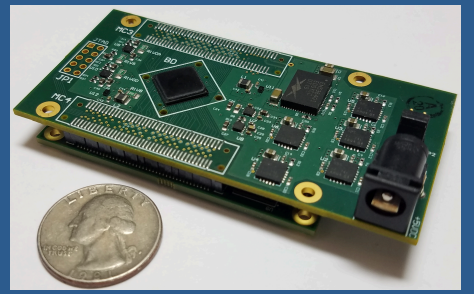



Akida


BrainDrop




BrainDrop



## Algorithms

LMU

LSNN

Feedback
Alignment

SLAYER

SNN
Mapping

SNN
Backprop

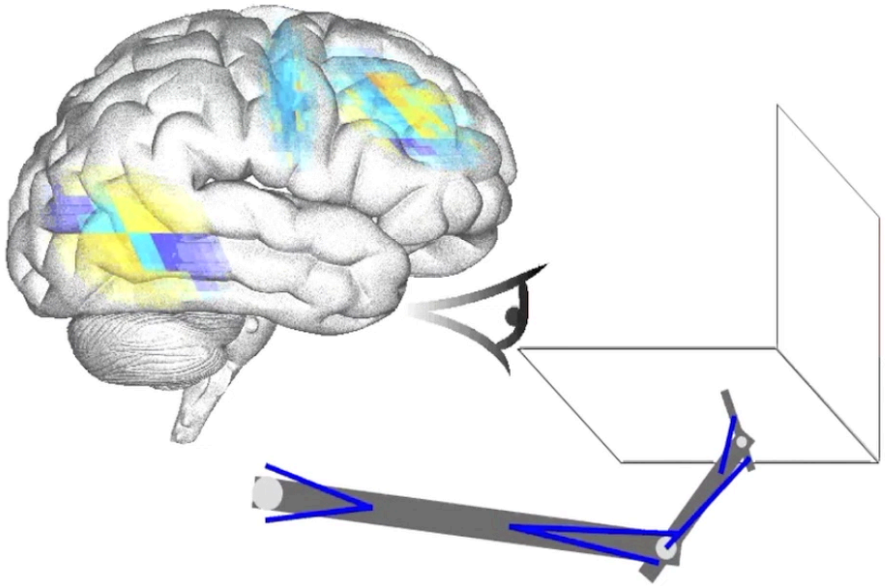
eprop

Temporal
Dithering

# Recent Advances (2015-2020)

## Scaling Spiking Algorithms

- VGG16 (25K neurons, 138M weights), ResNet34 on ImageNet (Purdue, 2020)
- 1M element k-nearest neighbor (Intel, 2020)
- Spaun (6M neurons; 20B weights; Waterloo, 2018)



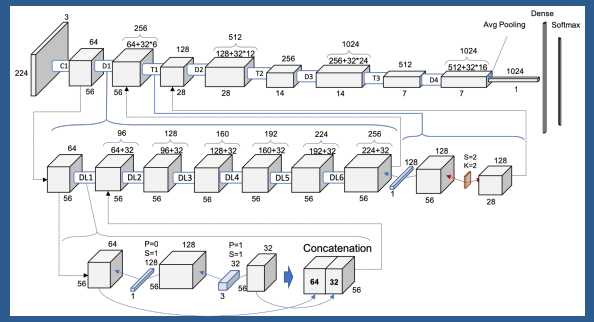
# Recent Advances (2015-2020)

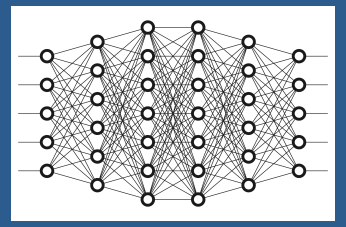
## Scaling Hardware

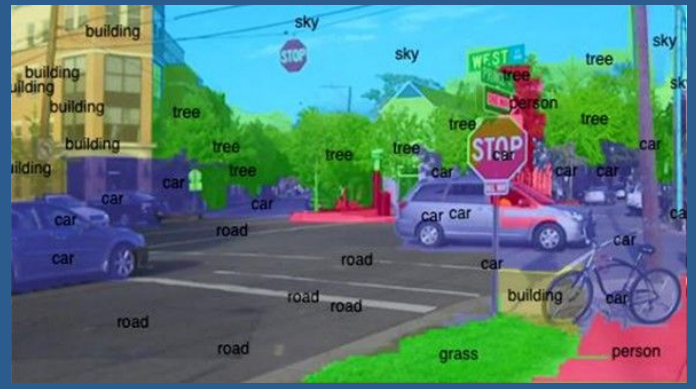
- Note: Spaun scaled up on current GPUs = 1.21 GW
- Braindrop (2017; analog/digital;
  4k neurons; Spaun = 38kW)
- Loihi (2019; digital;
  100M neurons; Spaun = 2.4MW)
- SpiNNaker 2 (2021; digital;
  10B neurons; Spaun = 2.4MW)

# Relation to Current AI

## Similarities

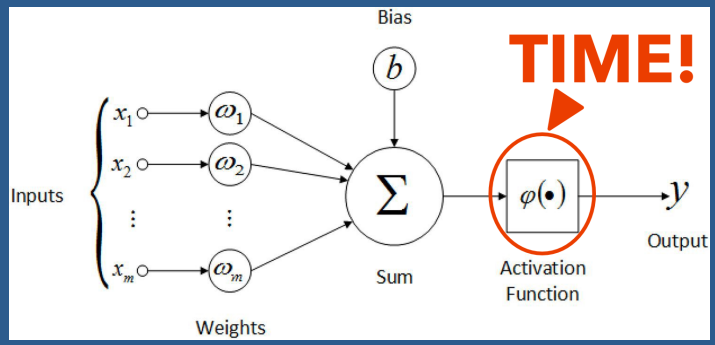


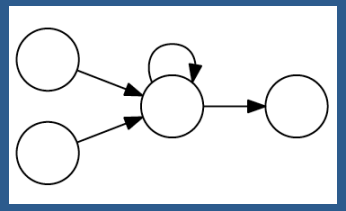


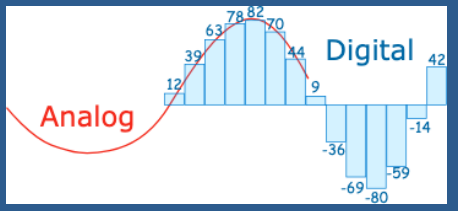

## Differences

### Fundamental

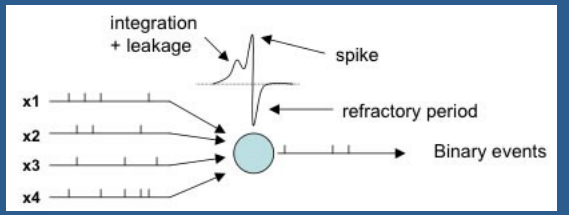



### Emphasis





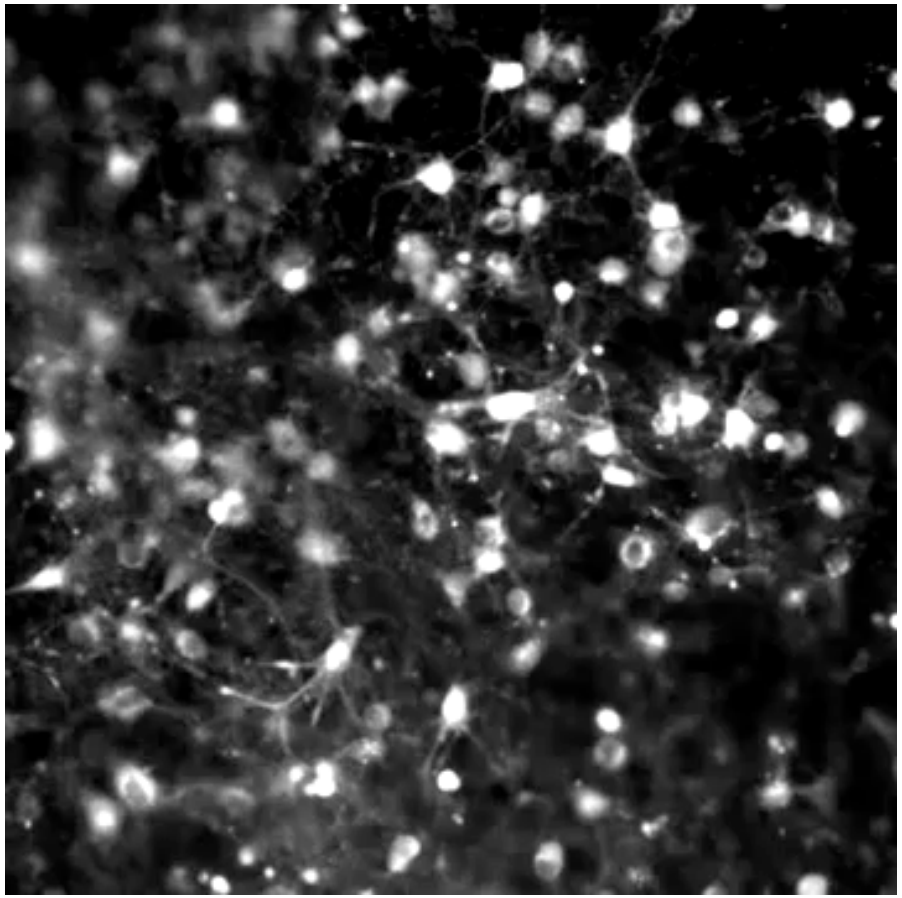

Online Learning



# Biological Computation

Behavior is *dynamic*

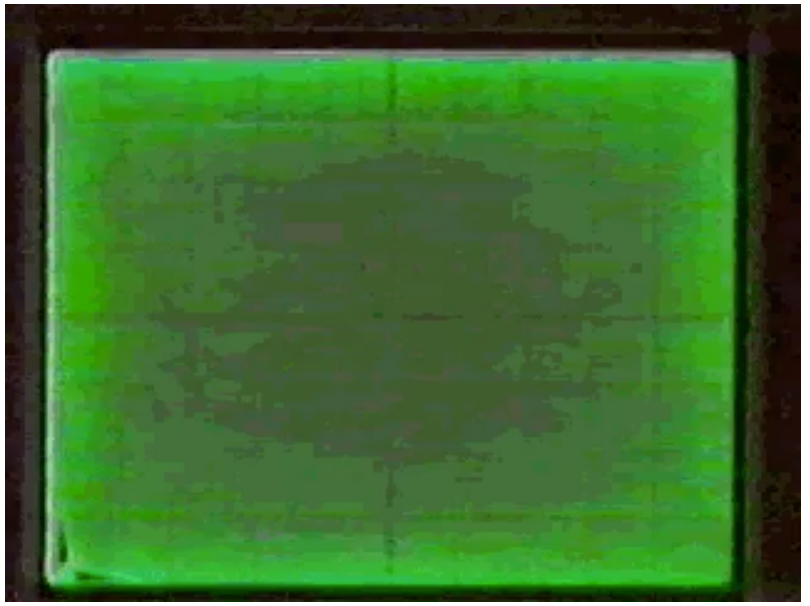
# Biological Computation



GeorgiaTech NeuroLab

Groups of neurons are *dynamic*

# Biological Computation



Single neurons are *dynamic*

## Why it matters?

***Efficiency***

Human brain: 20W

GPUs: 1,210,000,000W (2020)

(Stewart, 2021)

For things the brain is good at
(i.e., not rendering video)

**Lesson 1: Hardware should match algorithms**

**Dynamics (High-level)**

## ‘Matching’ algorithms

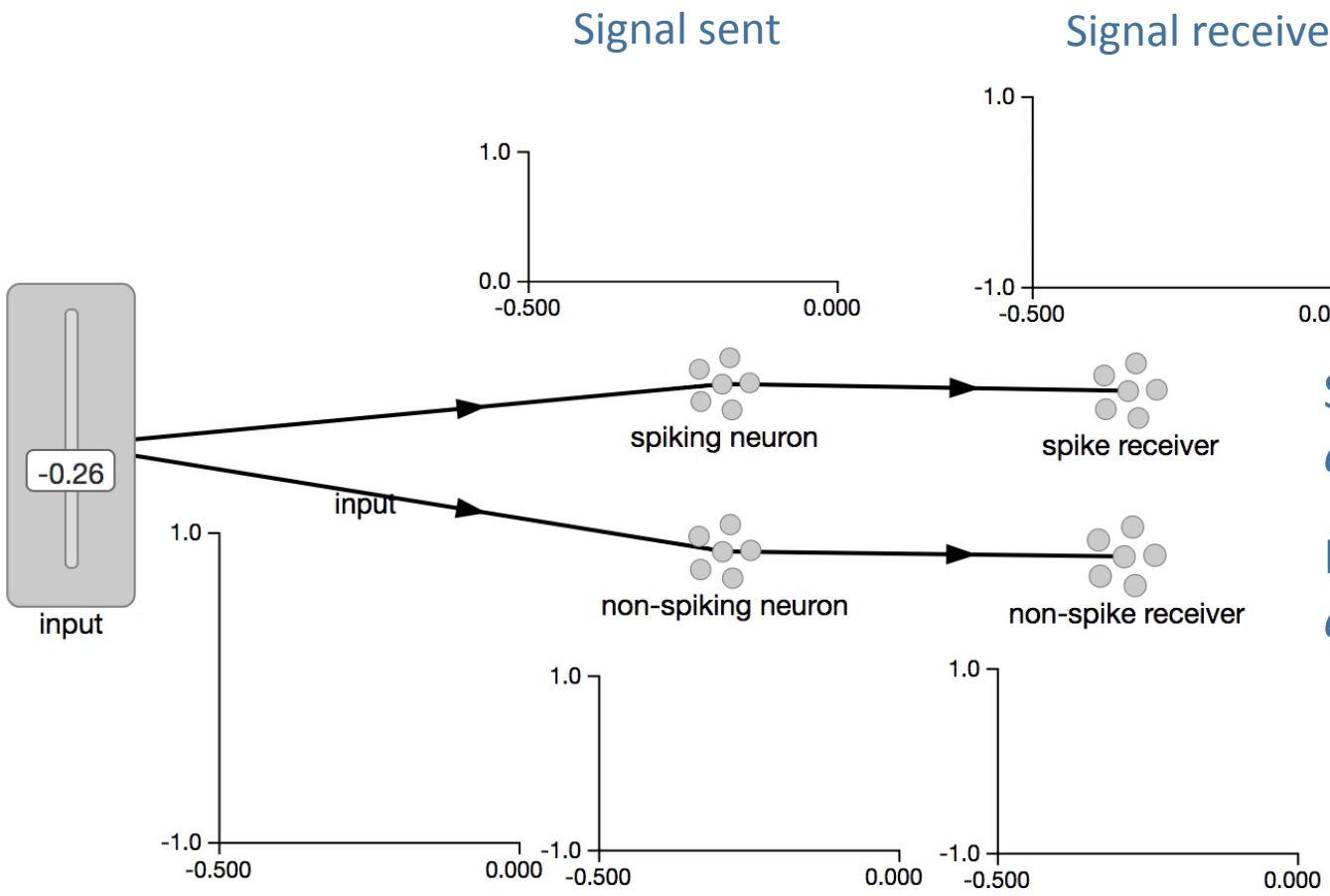
- Not new: Generality/efficiency tradeoff (Kolmogorov)

- Algorithms for dynamics
  - LSTMs, Transformers – are not continuous *dynamic* sys
  - LMU – new (optimal) dynamic sys

We’ll get there later

# Dynamics (Low-level)

# ‘Matching’ algorithms

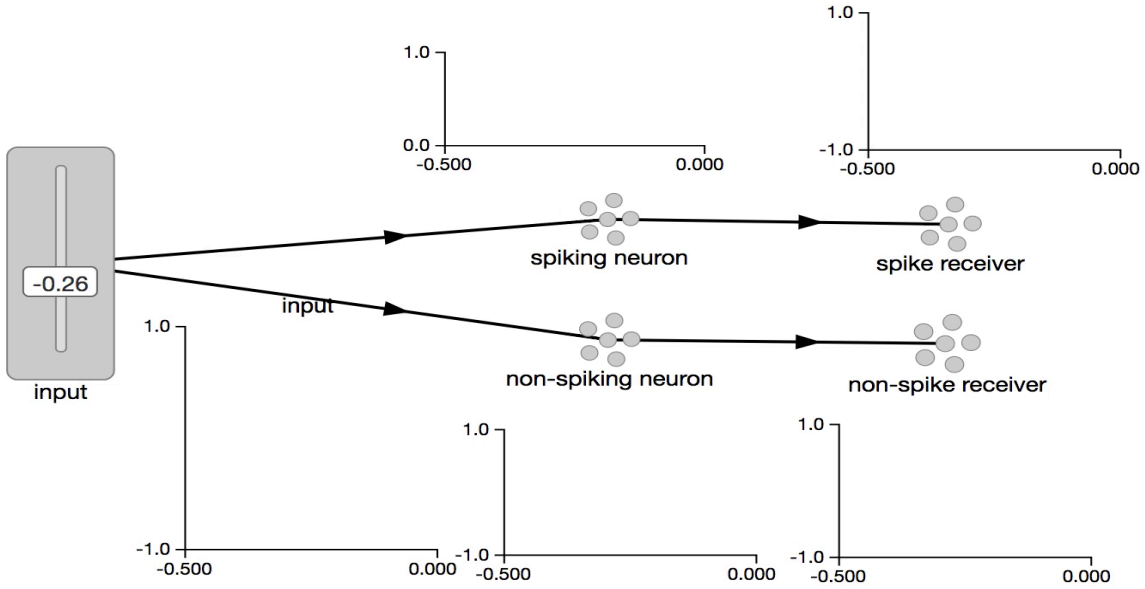



Spiking Neurons
*a=f(x,t)*

Rate Neurons
*a=f(x)*

Nengo

# Dynamics (Low-level)




- Sparsification over time
  - Less communication
- Less computation
  - Fewer memory lookups
- Cheaper computation
  - Sum instead of multiply

## Biological Computation

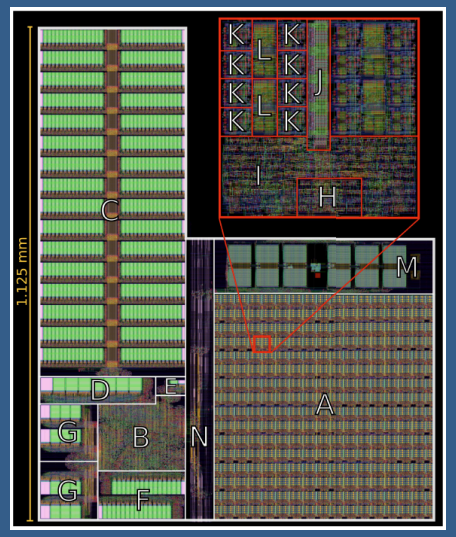
- Dynamic (real-time)
- Unclocked
- Analog
- Slow (<KHz)
- Efficient (<1fJ)
- (Parallel)

## Engineered Computation

- Step based
- Clocked
- Digital
- Fast (>MHz)
- Less efficient (>1pJ)
- (Less parallel)

## What matters?

# BrainDrop



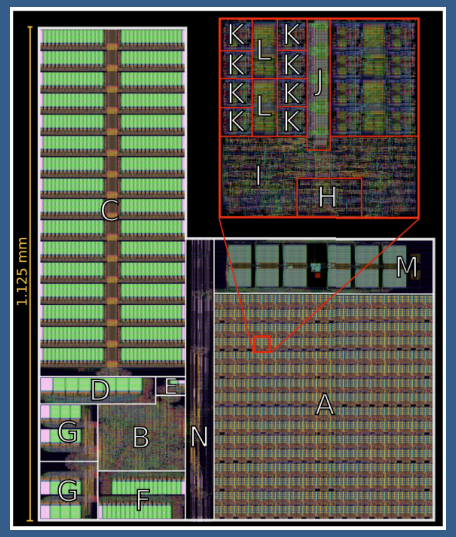

Boahen (Stanford), Manohar (Yale), Eliasmith (Waterloo)

## All of the above (prototype HW)

- ✓ Dynamic (real-time)
- ✓ Unclocked
- ✓ Analog (dendrites and neurons)
- ✓ Slow
- ✓ Efficient (300 fJ)
- ✓ (Parallel)

(38kW Spaun)

# BrainDrop




Boahen (Stanford), Manohar (Yale), Eliasmith (Waterloo)

## But…

- Every chip is different
- Unique optimizations for each algorithm for each chip
- I.e. you can't copy 'programs' from one chip to another

Therefore not commercially deployable

## Lesson 2: Not all constraints are computational

## What matters?

## Need to consider each in detail

- Dynamics (real-time) – what level?
- Unclocked – just some parts?
- Analog – just some parts?
- Slow – different clock regimes?
- Efficient – result of other choices
- (Parallel)

Huge design space (yay!?)

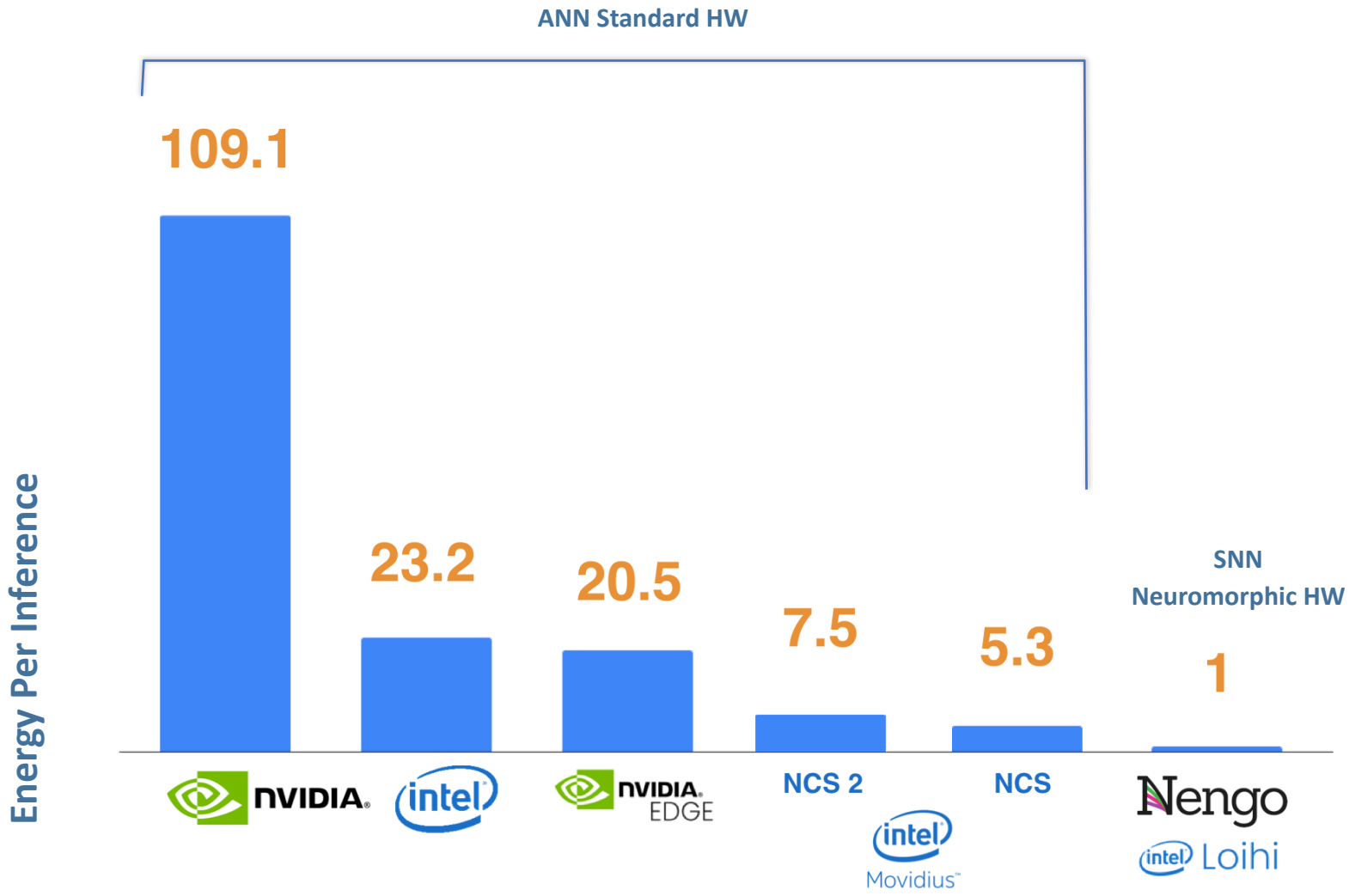
# Intel Loihi


## Another set of choices

- Fully digital
- Asynchronous
- Supports on chip learning
- 128 cores 1024 neurons/core
- Hard to program (Nengo helps)
- Easier to use/larger than Braindrop
- Loihi 2 just released (1M neurons)

2.4MW Spaun

# Deep Spiking Network

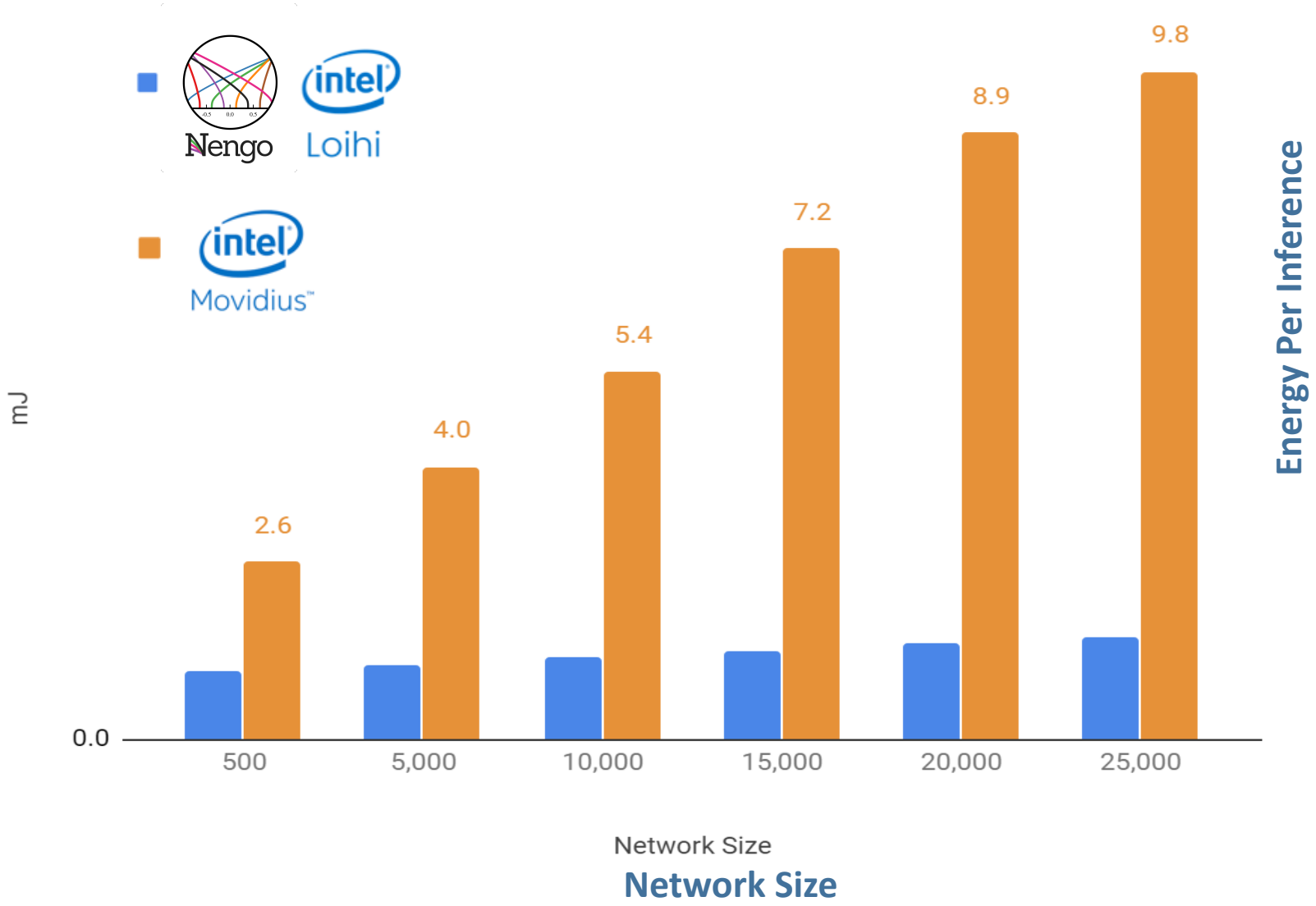



ANN Accuracy – 92.7%

SNN Accuracy – 93.8%

APPLIED BRAIN RESEARCH

# DSN Scaling



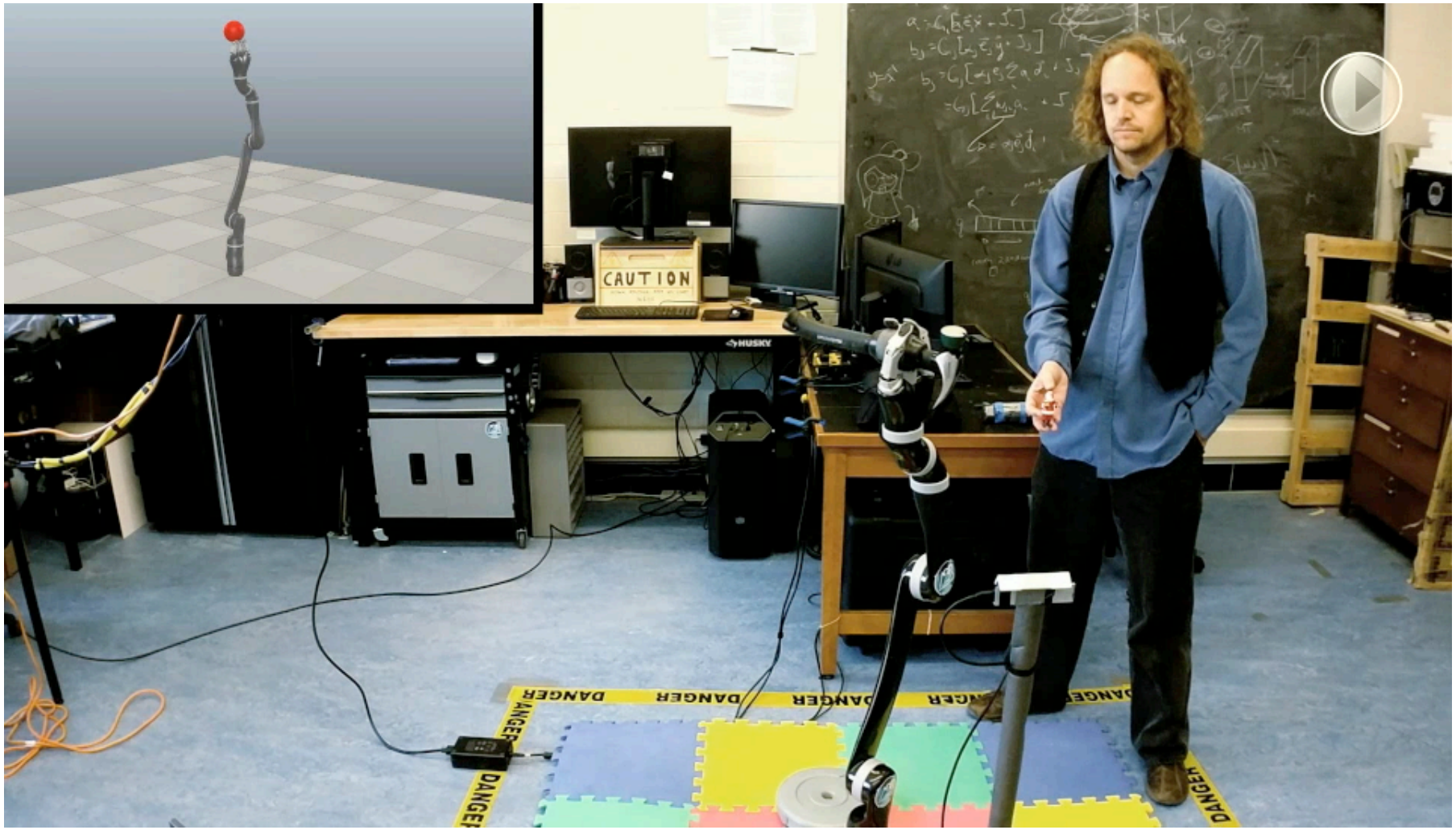

APPLIED BRAIN RESEARCH

# Adaptive Control

## Challenges

- Build the adaptive controller in spiking hardware (Loihi)
- Demonstrate it is better than other solutions (e.g. PID)
- Demonstrate an industrially relevant result
- Demonstrate the benefits of neuromorphic adaptive control

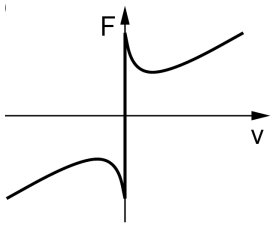

APPLIED BRAIN RESEARCH

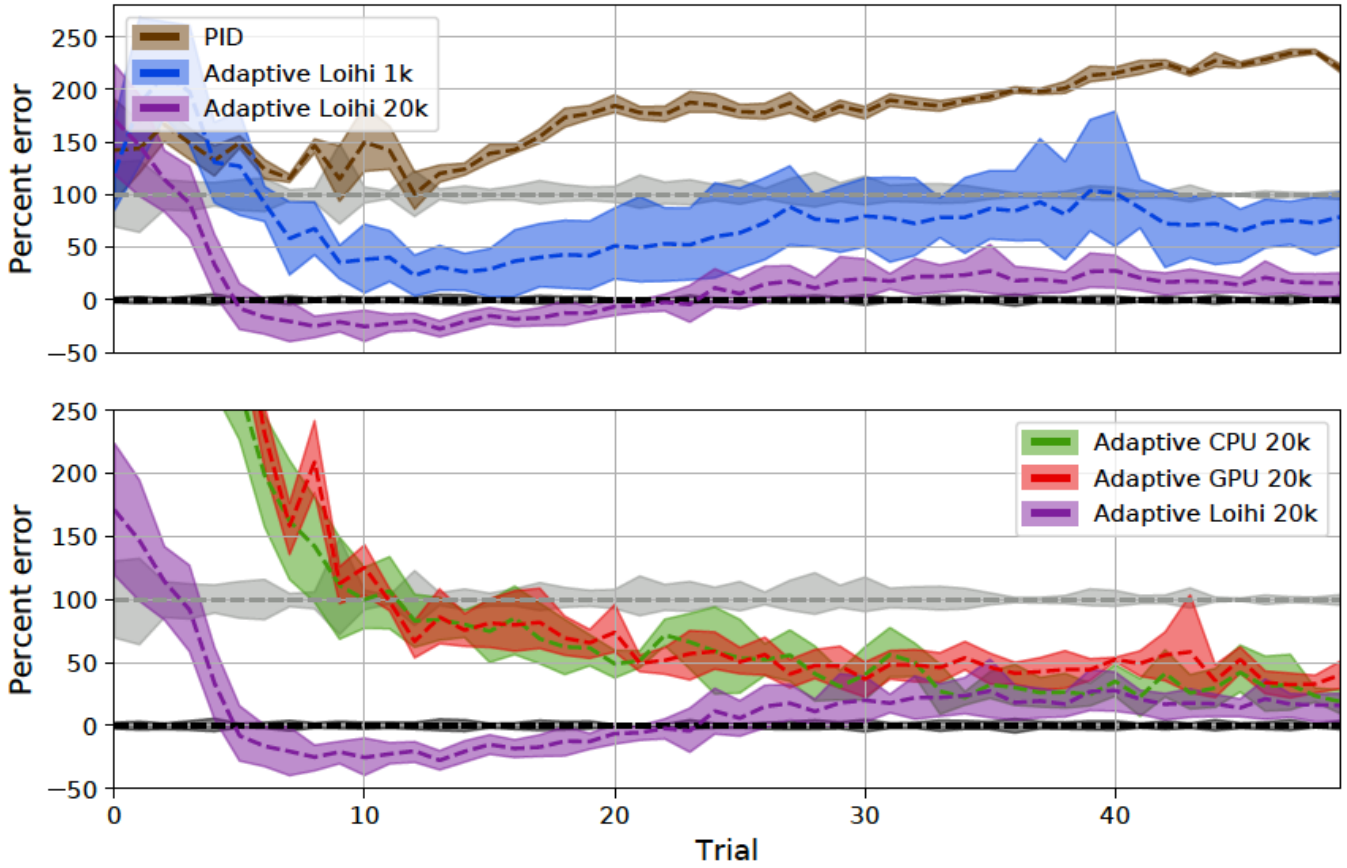



APPLIED BRAIN RESEARCH

# Accelerated Wear

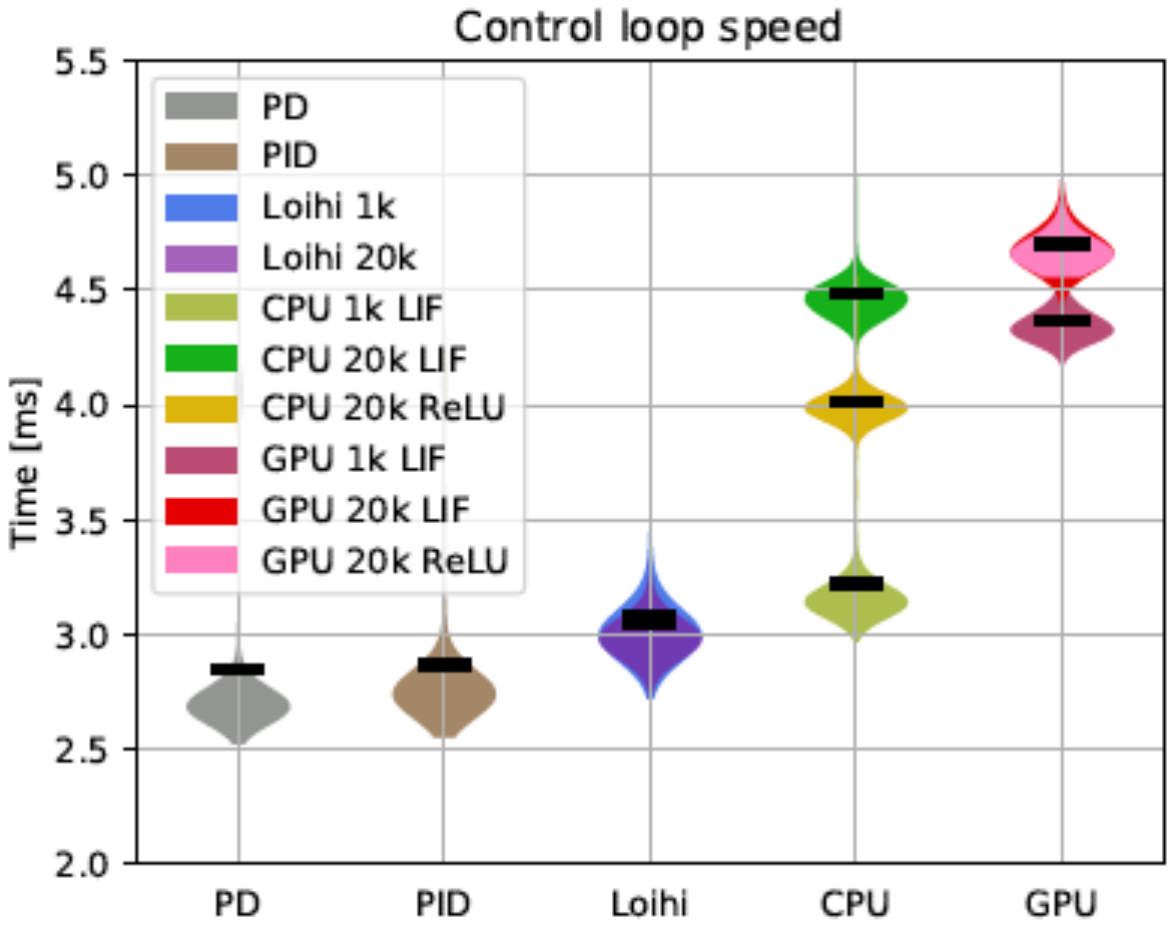
- Five reaches
- Nonlinear friction (4 yrs)
- 5 trials
- 50 runs
- Save weights




APPLIED BRAIN RESEARCH

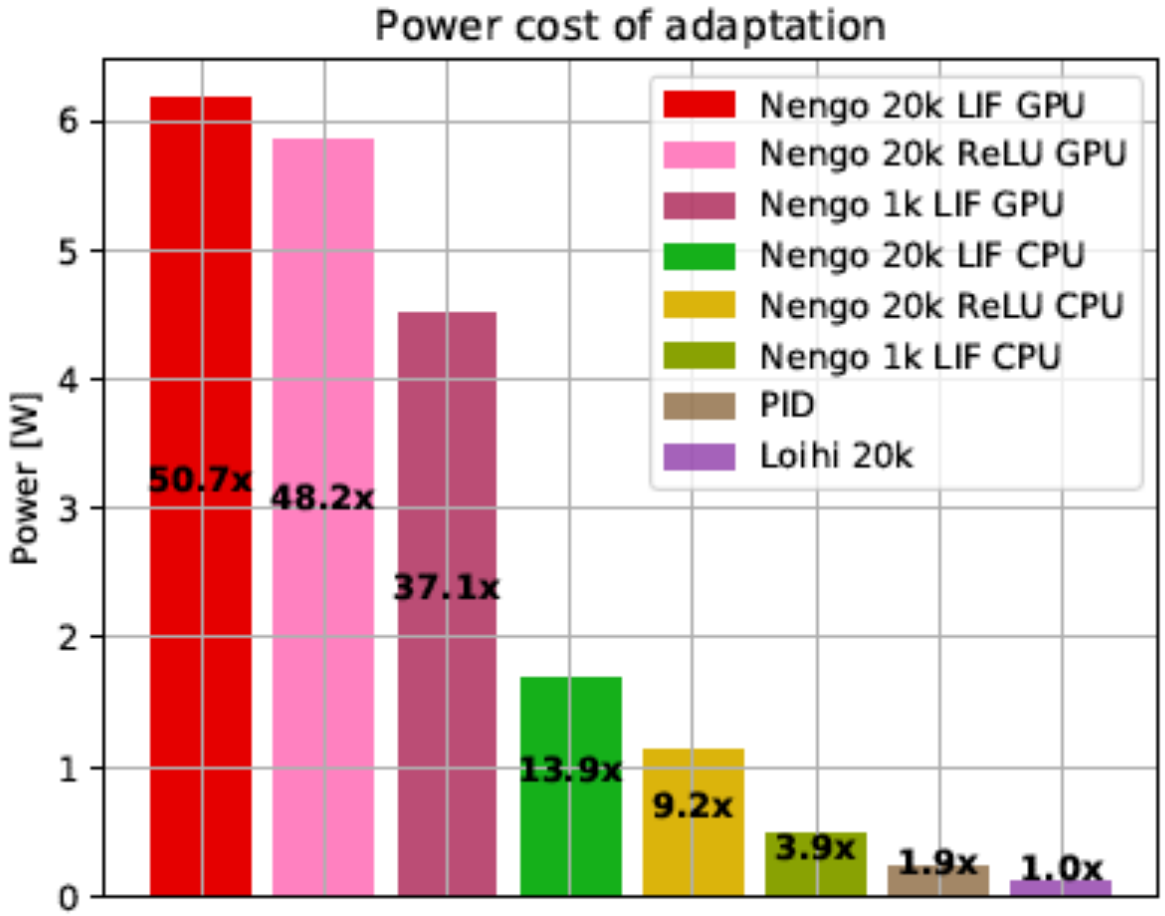
# Speed Improvements




~30% Faster

APPLIED BRAIN RESEARCH

# Energy Improvements




~10-50x Less Energy

APPLIED BRAIN RESEARCH

# Conclusion

## Lessons

- Lesson 1: Match HW to algorithms
- Lesson 2: Not all constraints are computational

- There are advantages to brain inspired HW

- These advantages are greatest for dynamic processing

# Resources

- Neuromorphics insights and challenges
- Intro to neuromorphics (popular science version)
- A neuromorph's prospectus
- Arm video
- Drone video
- http://nengo.ai