TENNLab Research on Neuromorphic Computing Systems From Devices to Algorithms

James S. Plank
EECS Department
University of Tennessee

Intel Neuromorphic Research Community
June 9, 2021

Note for the PDF version of these slides.

- The LIDAR demonstration videos are not in this PDF, but you may view them at:

  http://web.eecs.utk.edu/~jplank/2019-03-27-NICE-Video.mp4

- The video of the spiking convolutional layer may be viewed as part of my 2020 ICRC talk:

  https://www.youtube.com/watch?v=Q-7FJOS7dhI

- All of our publications may be accessed via:

  http://neuromorphic.eecs.utk.edu/

Neuromorphic Computing With TENNLab

TENNLab Leadership and Funding:

Jim Plank
University of Tennessee

Katie Schuman
Oak Ridge Natl. Lab

Garrett Rose
University of Tennessee

Nate Cady
SUNY Polytech

TENNLab Researchers (since 2014)

TENN LAB
NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

http://neuromorphic.eecs.utk.edu

TENNLab Mission

To explore all facets of brain-based, Neuromorphic computing via recurrent, spiking neural networks.

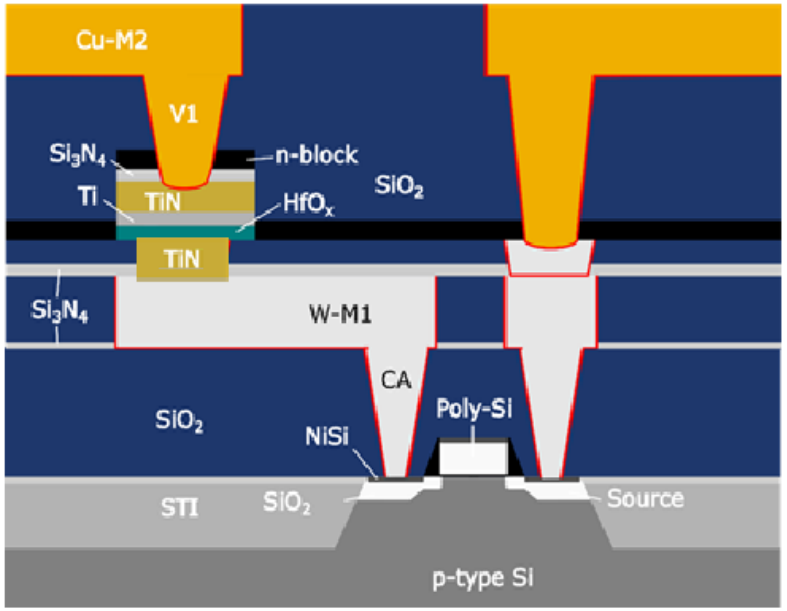
Nanoelectronics,
Circuits, ASICs, FPGAs,

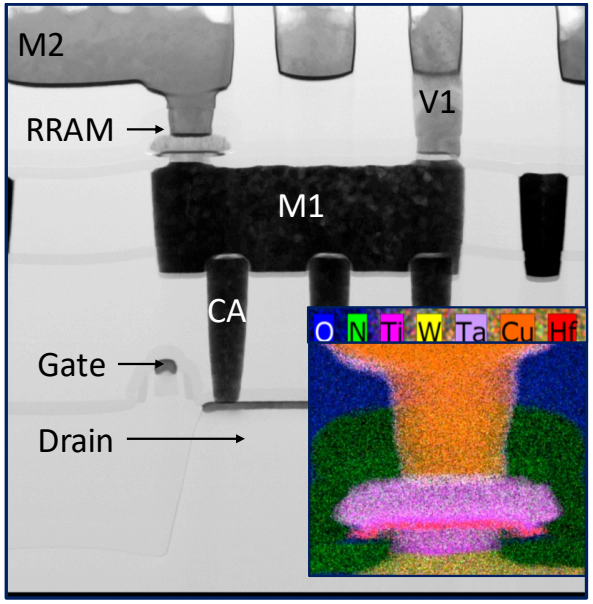
Applications, Systems,
Algorithms, Training.

In this talk

- I'll highlight the research of my colleagues
- And give you a deeper dive into some of my projects
  - Software Framework
  - Control applications
  - Input Encoding
  - Compressing convolutional layers

CMOS/RRAM Hardware (RAVENS)
Fabrication / Integration Approach

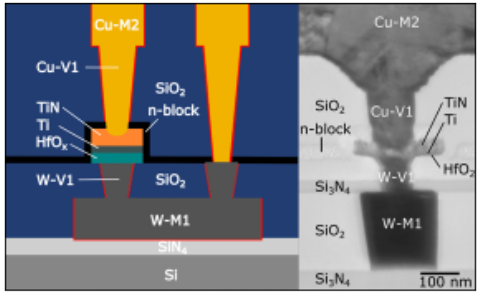




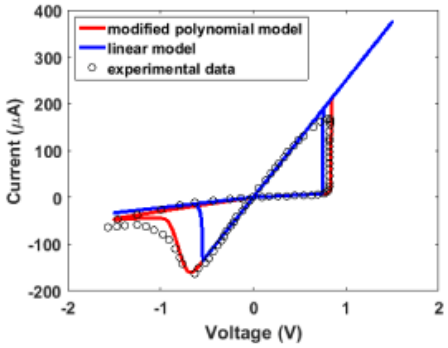
- **Hafnium oxide based RRAM**
- **FEOL/BEOL compatible process**
- **1 Transistor / 1 RRAM configuration for memory arrays and individual memory cells**

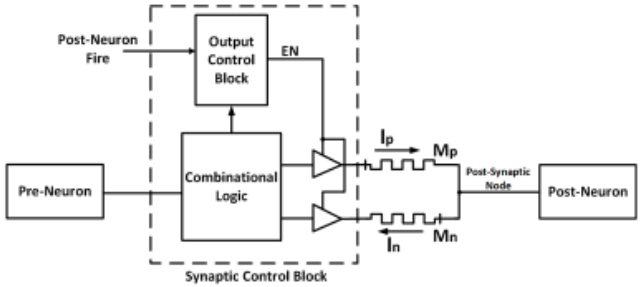
- **Integrated CMOS/RRAM neuronal learning circuits**
- **SUNY Poly's 65nm CMOS + RRAM process design kit (PDK) successfully transferred to UT-Knoxville, UT-San Antonio, UT-Austin, and Arizona State.**

SENECA Lab Research Overview

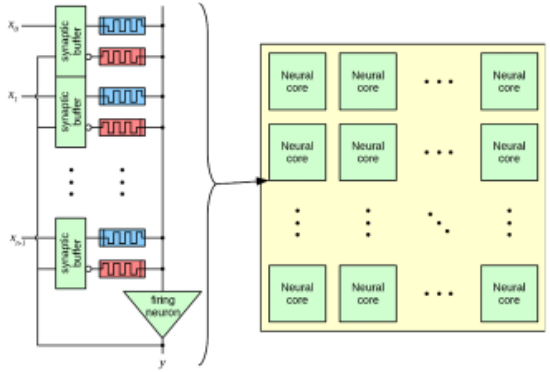


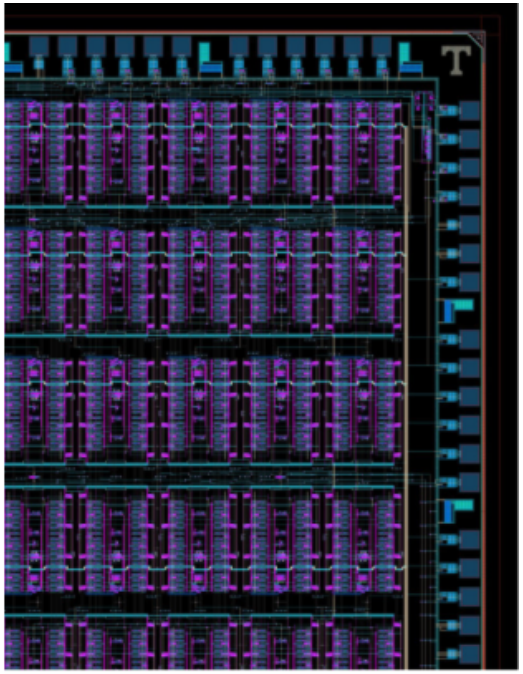
Memristor Modeling





CMOS/Memristor Circuits





Neuromorphic System Integration

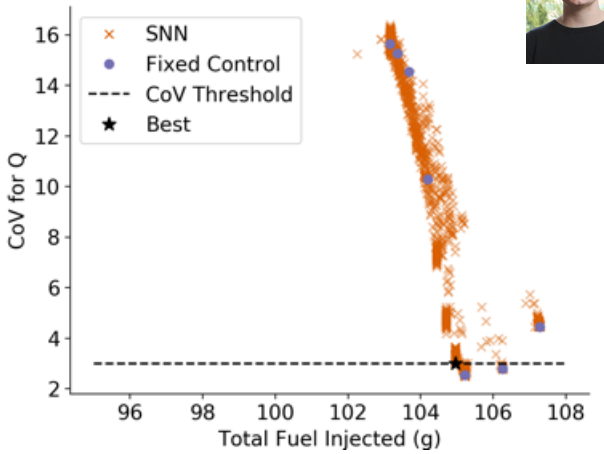
S. Sayyaparaju, M. M. Adnan, S. Amer and G. S. Rose "Device-aware Circuit Design for Robust Memristive Neuromorphic Systems with STDP-based Learning", *ACM Journal on Emerging Technologies in Computing Systems*, 16(3), May, 2020.

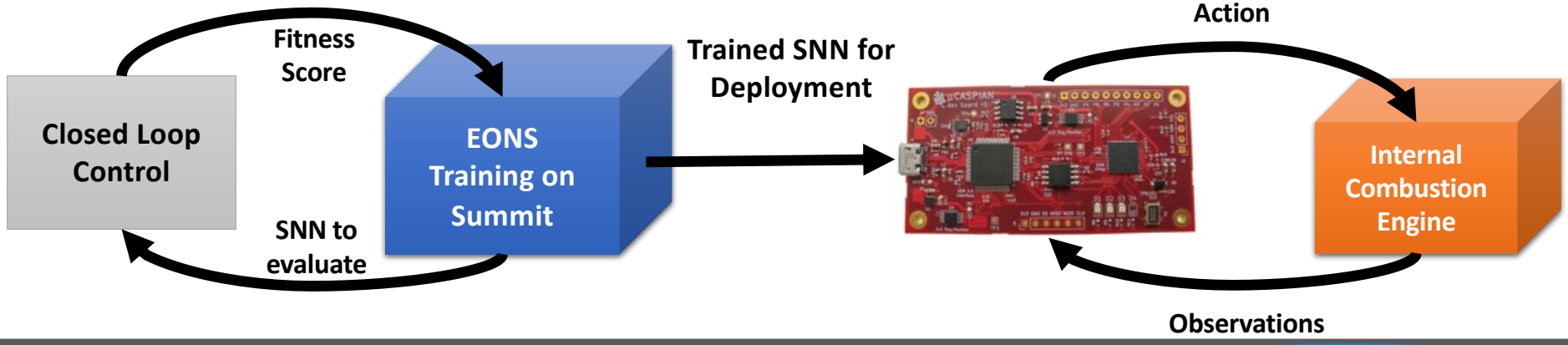
TENN LAB NEUROMORPHIC ARCHITECTURES. LEARNING. APPLICATIONS.

http://neuromorphic.eecs.utk.edu

Neuromorphic Engine Control for Fuel Efficiency

- Developed a complete workflow to train a spiking neural network (SNN) to deploy to an FPGA-based neuromorphic hardware system for internal combustion engine control.
- SNN-based approach outperforms fixed control strategies in terms of fuel efficiency in simulation while still meeting acceptable performance metrics.
- Currently deploying SNN trained on Summit to neuromorphic hardware in-the-loop with engine at National Transportation Research Center.





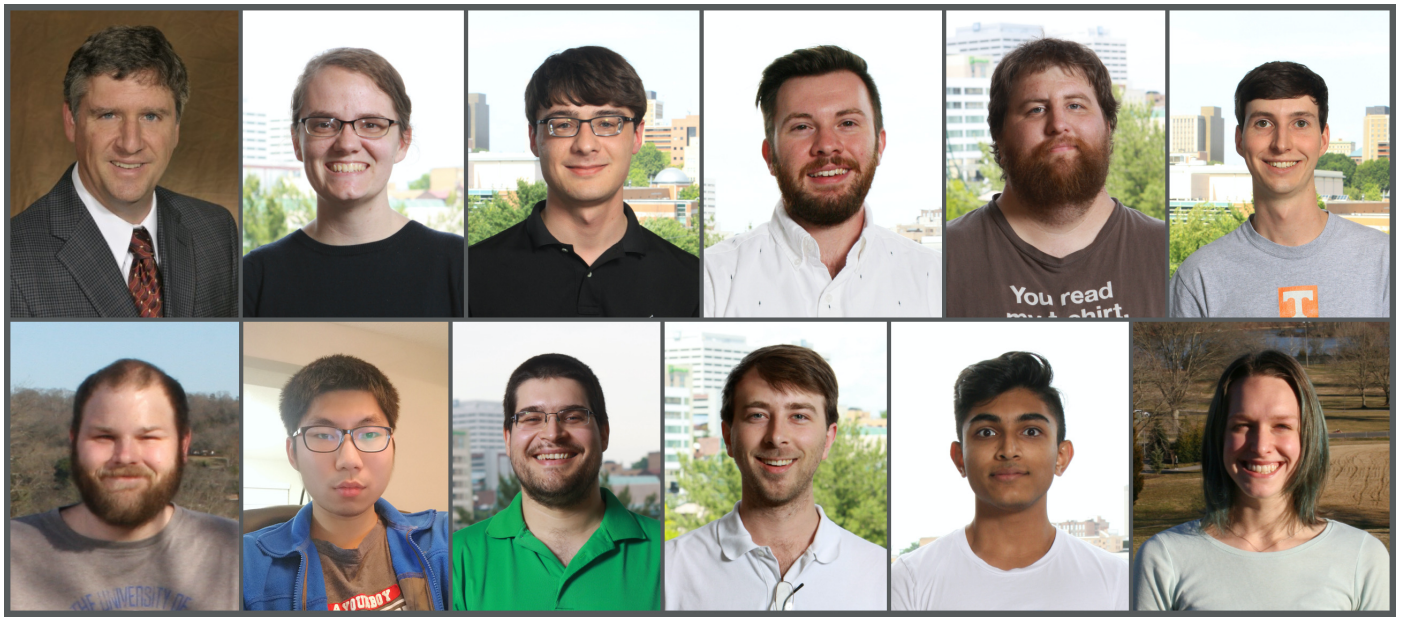
Catherine D. Schuman, Steven R. Young, J. Parker Mitchell, J. Travis Johnston, Derek Rose, Bryan P. Maldonado, Brian C. Kaul. "Low Size, Weight, and Power Neuromorphic Computing to Improve Combustion Engine Efficiency." *International Conference on Green and Sustainable Computing* 2020.

TENN LAB NEUROMORPHIC ARCHITECTURES. LEARNING. APPLICATIONS.

http://neuromorphic.eecs.utk.edu

Project #1

TENNLab Neuromorphic Computing Framework, V6

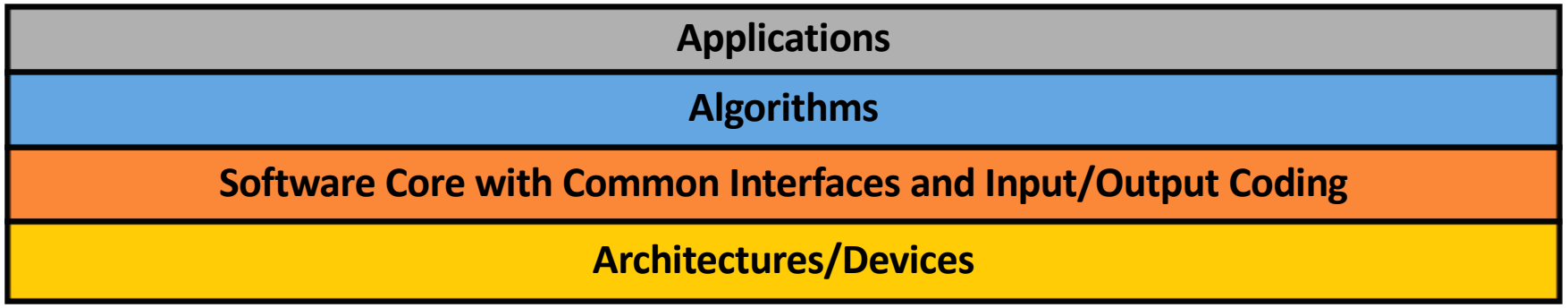


TENN LAB NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

J. S. Plank, C. D. Schuman, G. Bruer, M. E. Dean and G. S. Rose, The TENNLab Exploratory Neuromorphic Computing Framework, *IEEE Letters of the Computer Society*, 1 (2), 2019.

http://neuromorphic.eecs.utk.edu

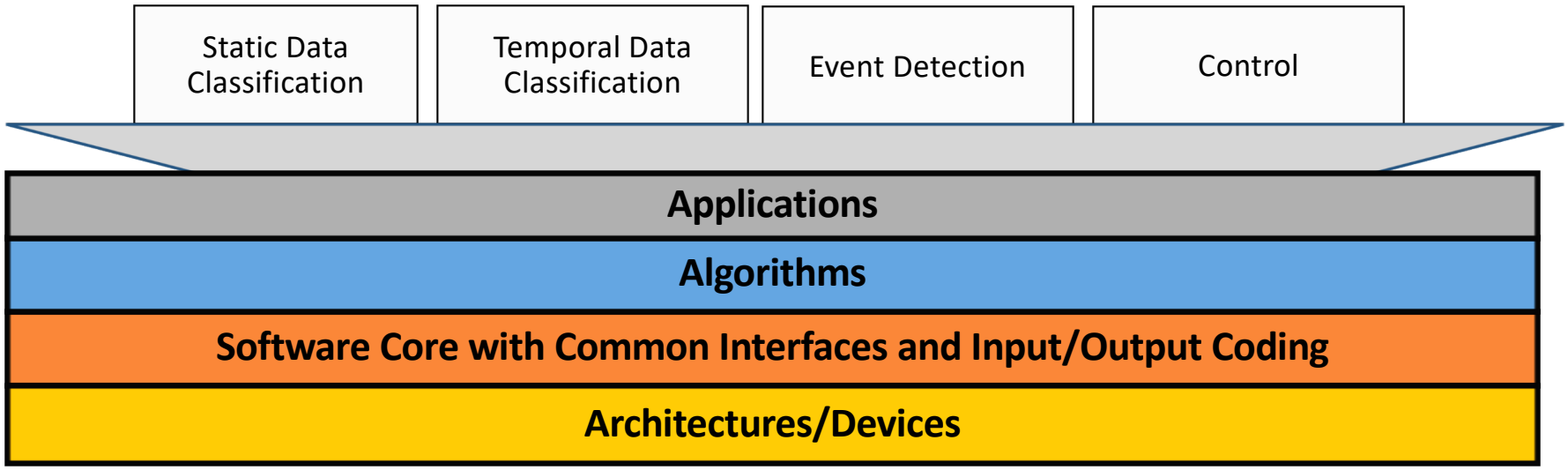
TENNLab Neuromorphic Computing Framework, V6



This picture has looked pretty much the same for years.
The difference is interoperability.

J. S. Plank, C. D. Schuman, G. Bruer, M. E. Dean and G. S. Rose, The TENNLab Exploratory Neuromorphic Computing Framework, *IEEE Letters of the Computer Society*, 1 (2), 2018.

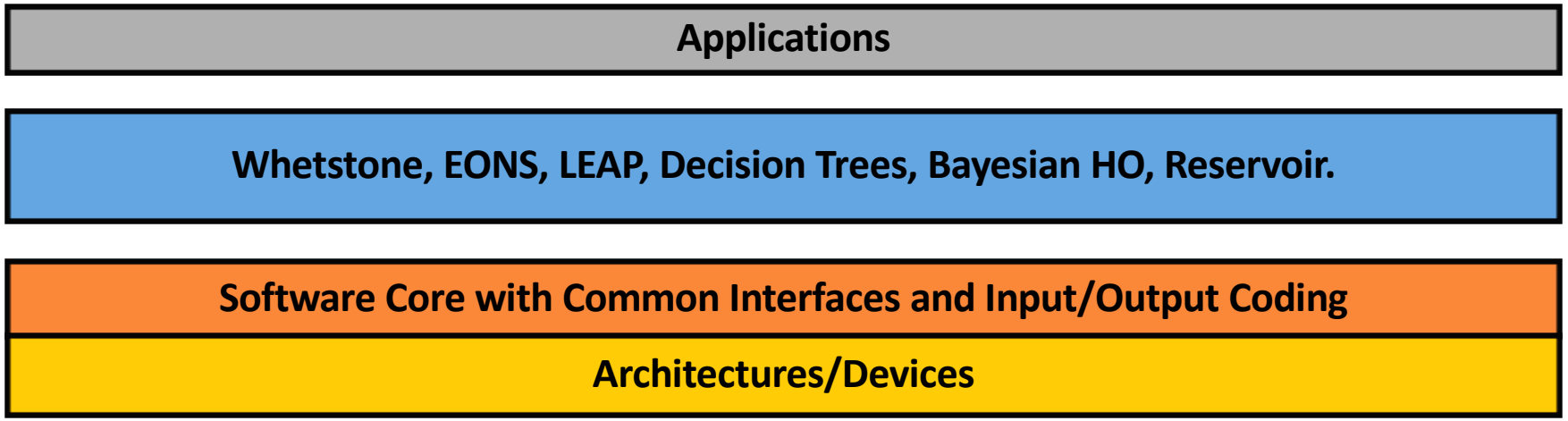
TENNLab Neuromorphic Computing Framework, V6



- Through 2019, our team wrote all of the applications in C++.
- With V6, we interoperate with Scikit Learn, Scikit, Optimize, OpenAI Gym (ALE)
- Workflow in C++ and in Python

J. S. Plank, C. D. Schuman, G. Bruer, M. E. Dean and G. S. Rose, The TENNLab Exploratory Neuromorphic Computing Framework, *IEEE Letters of the Computer Society*, 1 (2), 2018.

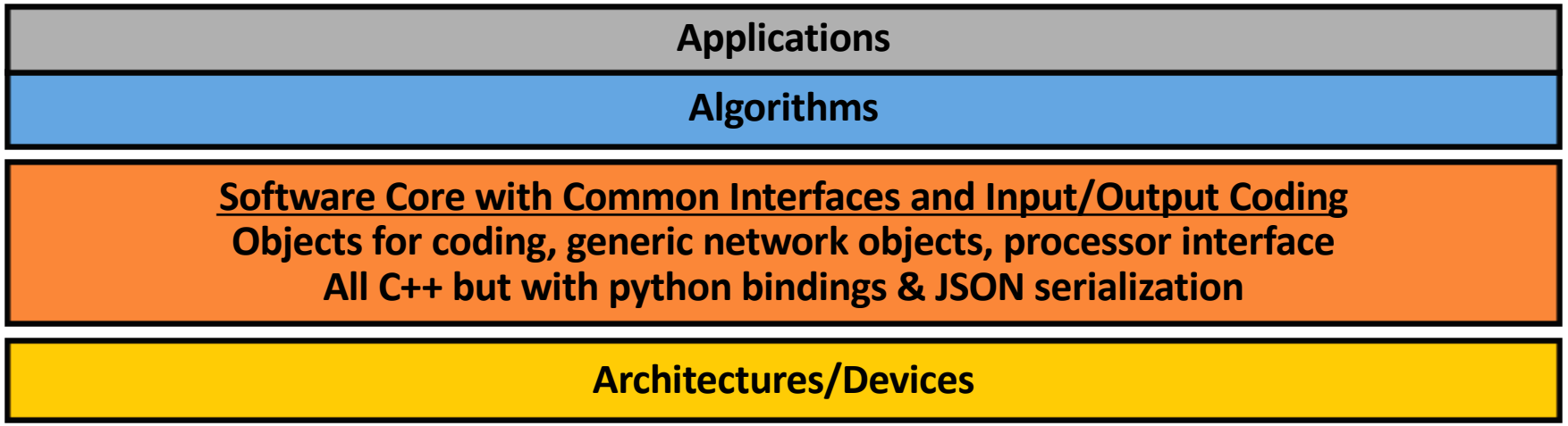
TENNLab Neuromorphic Computing Framework, V6



- Through 2019, EONS was it.
- Interoperation with python enables other ML techniques.

J. S. Plank, C. D. Schuman, G. Bruer, M. E. Dean and G. S. Rose, The TENNLab Exploratory Neuromorphic Computing Framework, *IEEE Letters of the Computer Society*, 1 (2), 2018.

TENNLab Neuromorphic Computing Framework, V6

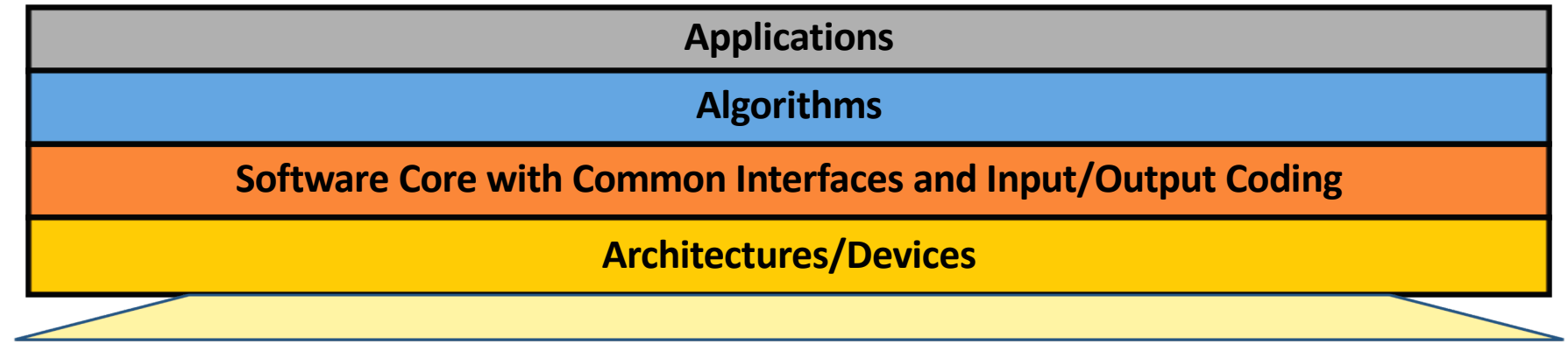


TENN LAB NEUROMORPHIC ARCHITECTURES. LEARNING. APPLICATIONS.

J. S. Plank, C. D. Schuman, G. Bruer, M. E. Dean and G. S. Rose, The TENNLab Exploratory Neuromorphic Computing Framework, *IEEE Letters of the Computer Society*, 1 (2), 2018.

http://neuromorphic.eecs.utk.edu

TENNLab Neuromorphic Computing Framework, V6



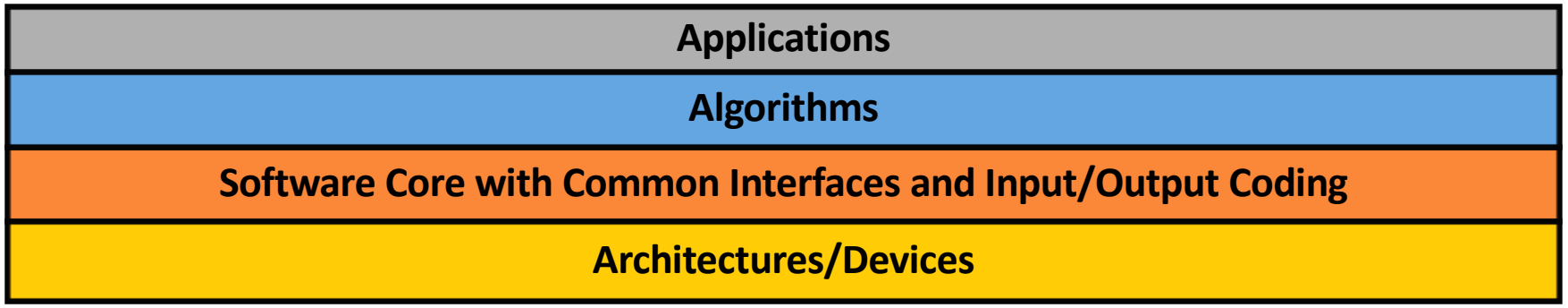
mrDANNA DANNA2 Biom SOENS Caspian

← Written by us or with our help. → ← Written by others. →

J. S. Plank, C. D. Schuman, G. Bruer, M. E. Dean and G. S. Rose, The TENNLab Exploratory Neuromorphic Computing Framework, *IEEE Letters of the Computer Society*, 1 (2), 2018.

http://neuromorphic.eecs.utk.edu

TENNLab Neuromorphic Computing Framework, V6



If you're interested in trying it out, please let me know.

J. S. Plank, C. D. Schuman, G. Bruer, M. E. Dean and G. S. Rose, The TENNLab Exploratory Neuromorphic Computing Framework, *IEEE Letters of the Computer Society*, 1 (2), 2018.

Project #2

LIDAR-based control applications

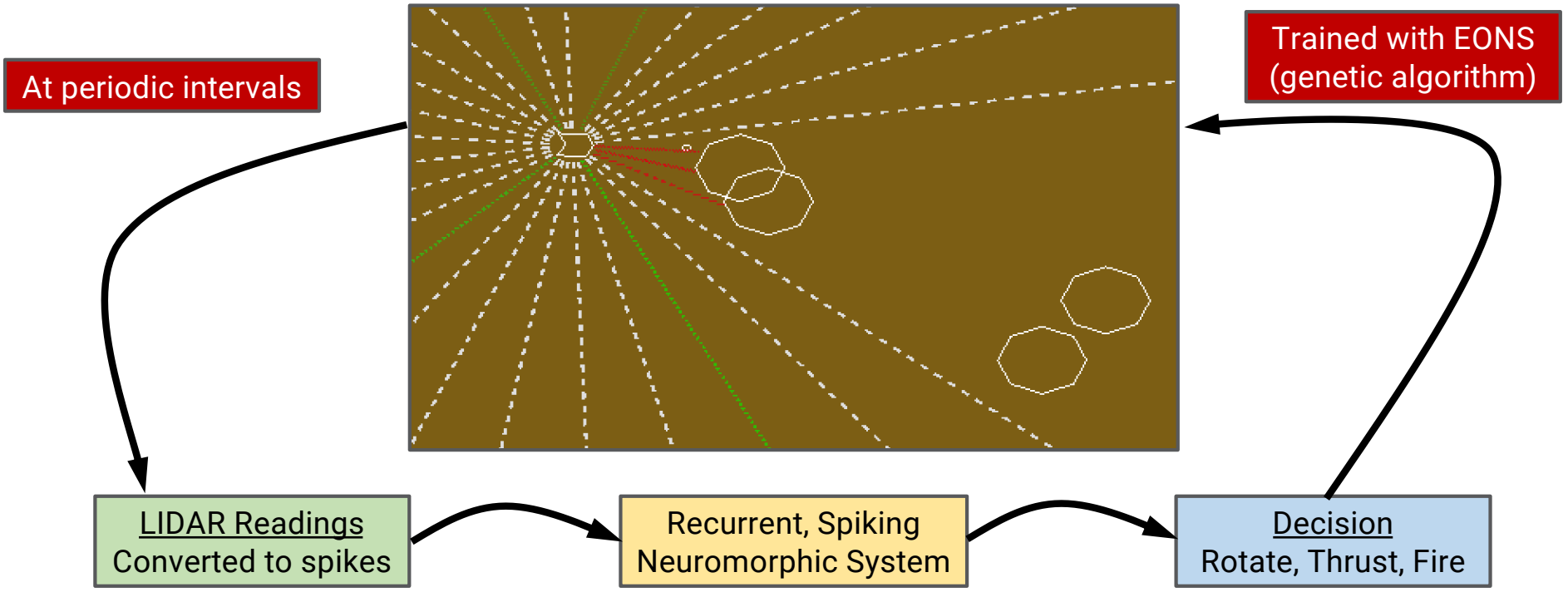
TENN LAB NEUROMORPHIC ARCHITECTURES. LEARNING. APPLICATIONS.

J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

http://neuromorphic.eecs.utk.edu

LIDAR-based control applications

Real-Time Application Equipped with LIDAR Sensors



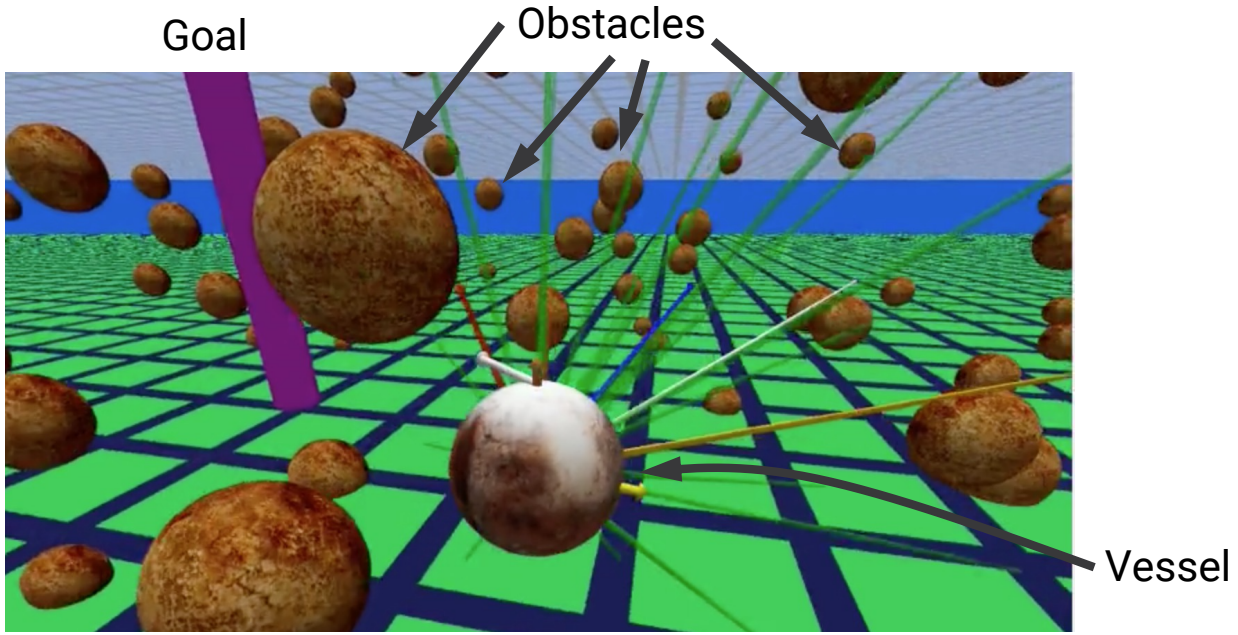
J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#1: Front-Facing Sense-and-Avoid (FFSA) - Navigation

- Vessel with a 5X5 array of LIDAR Sensors
- Boost in 6 directions

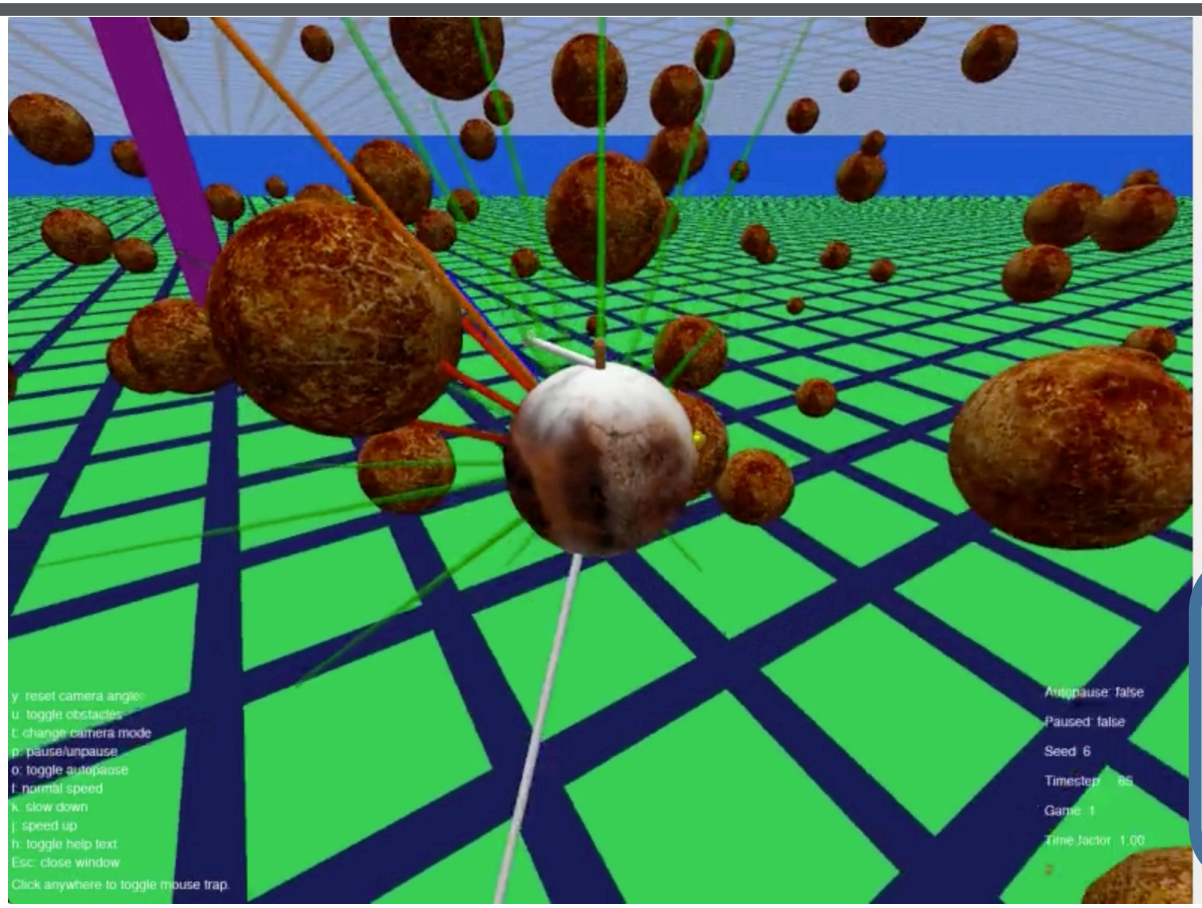
Objective:

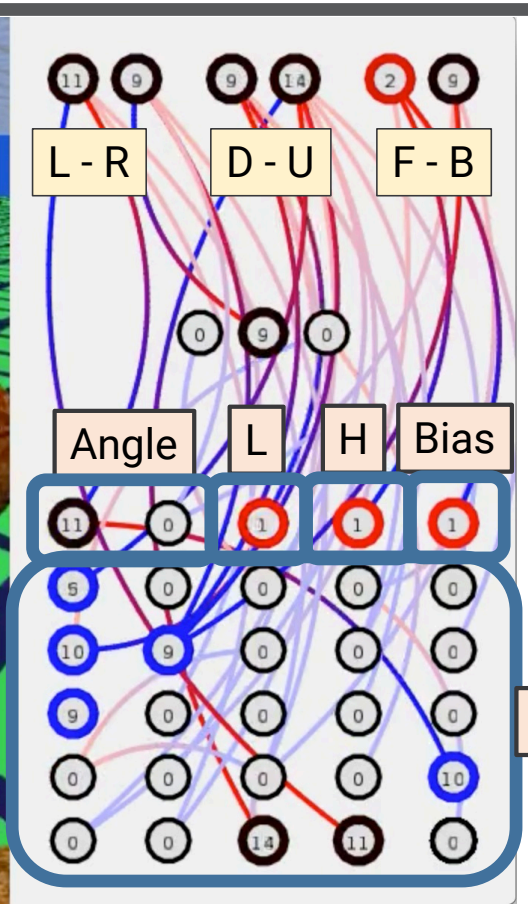
$$\left(\frac{\text{Distance attained to goal}}{\text{Distance to goal}}\right) * \left(\begin{array}{c}\text{0.6 penalty}\\ \text{for crash}\end{array}\right)$$



J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#1: Front-Facing Sense-and-Avoid on mrDANNA (memristive)



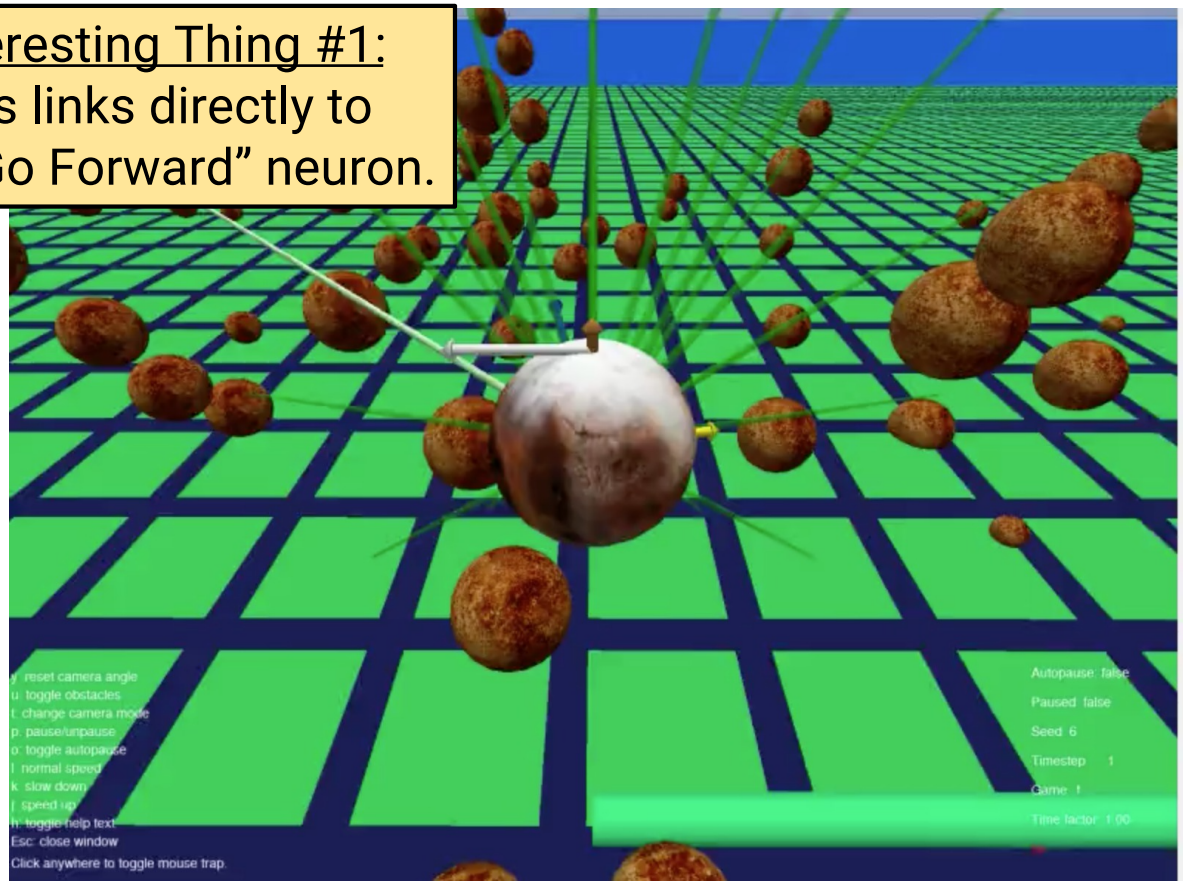


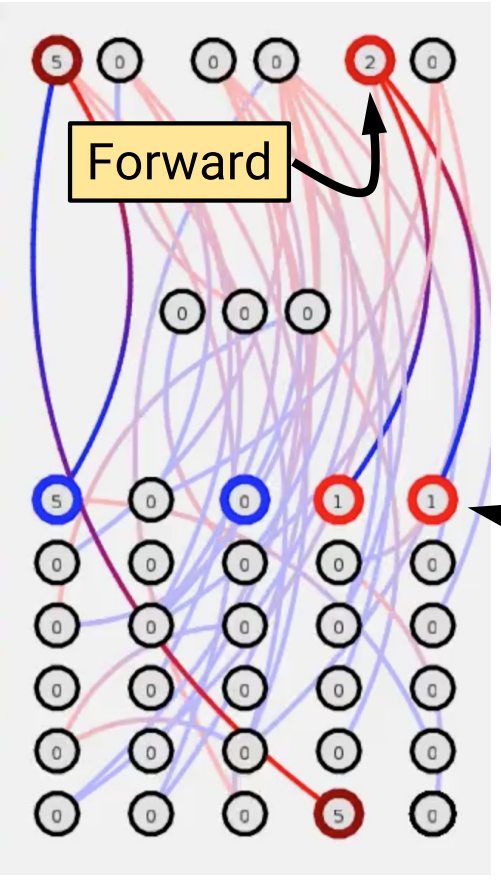
J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

to the vessel.

#1: Front-Facing Sense-and-Avoid on mrDANNA (memristive)

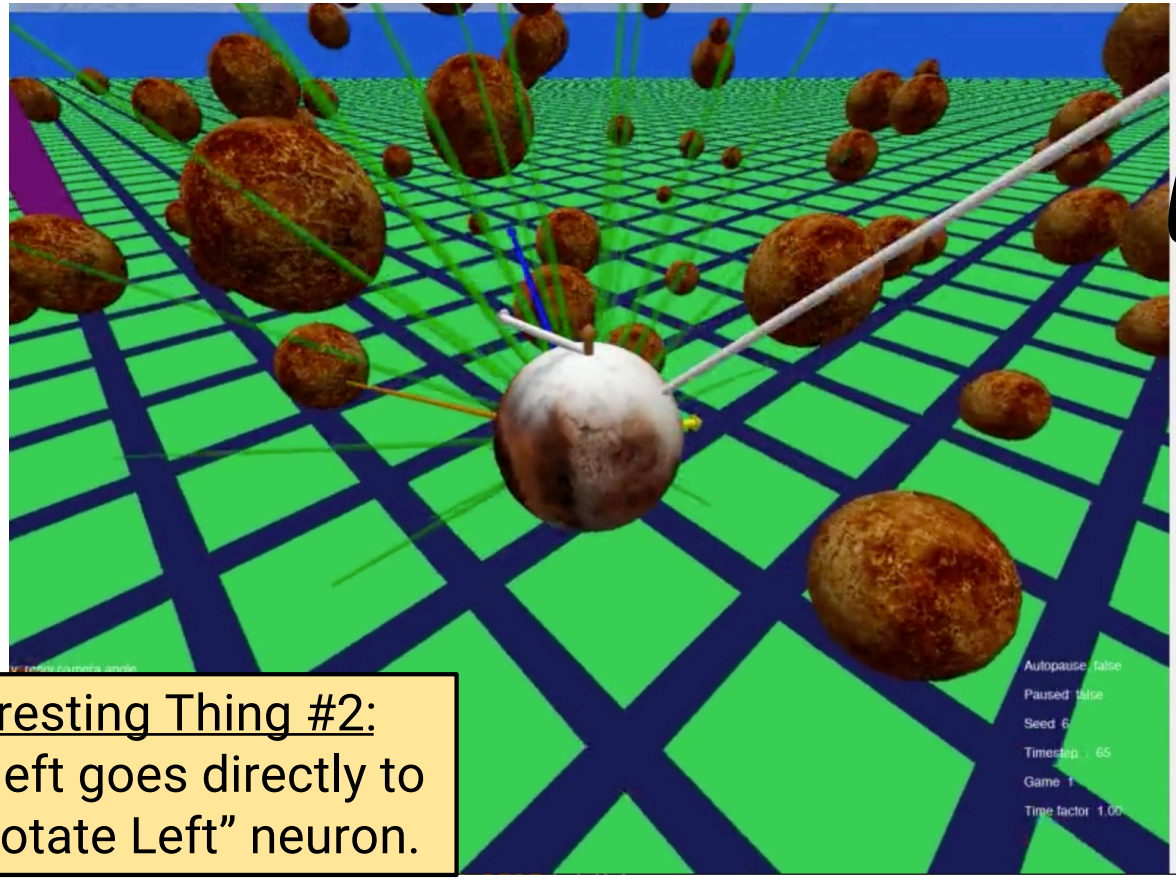
Interesting Thing #1:
Bias links directly to the "Go Forward" neuron.

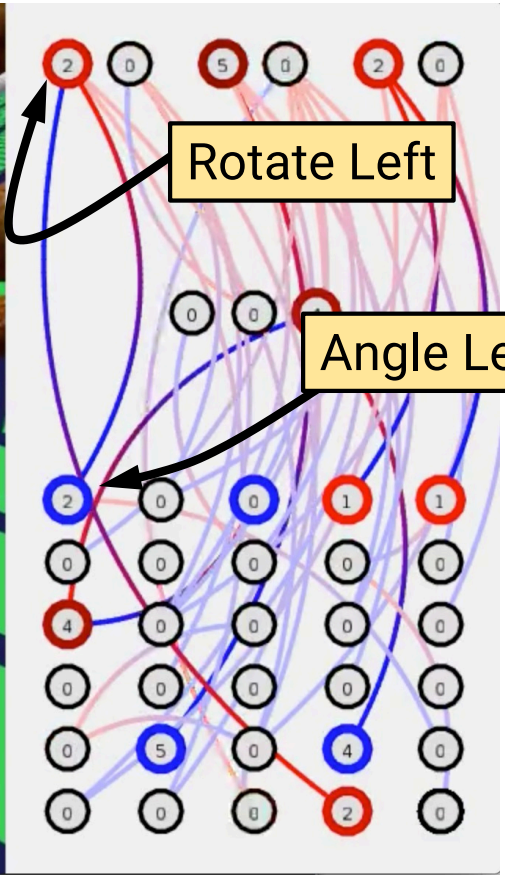




J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#1: Front-Facing Sense-and-Avoid on mrDANNA (memristive)

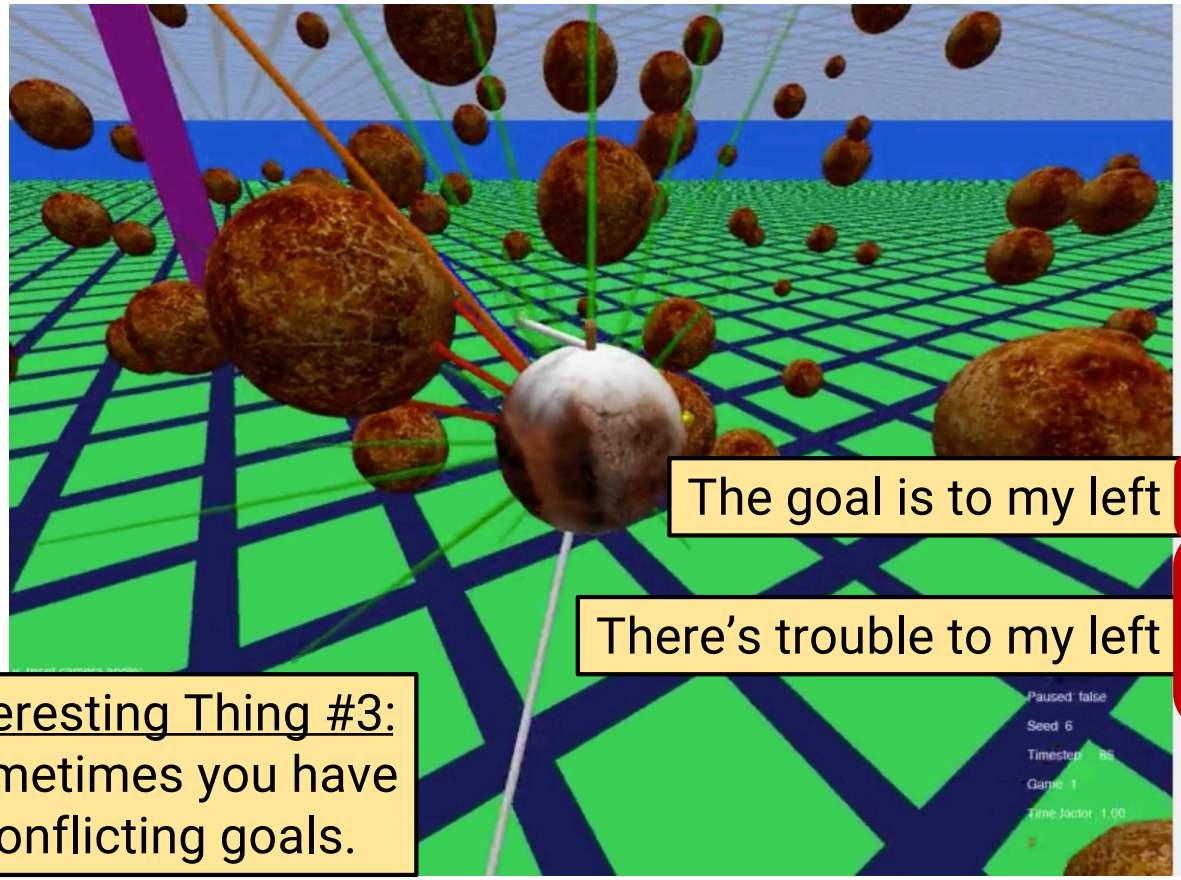


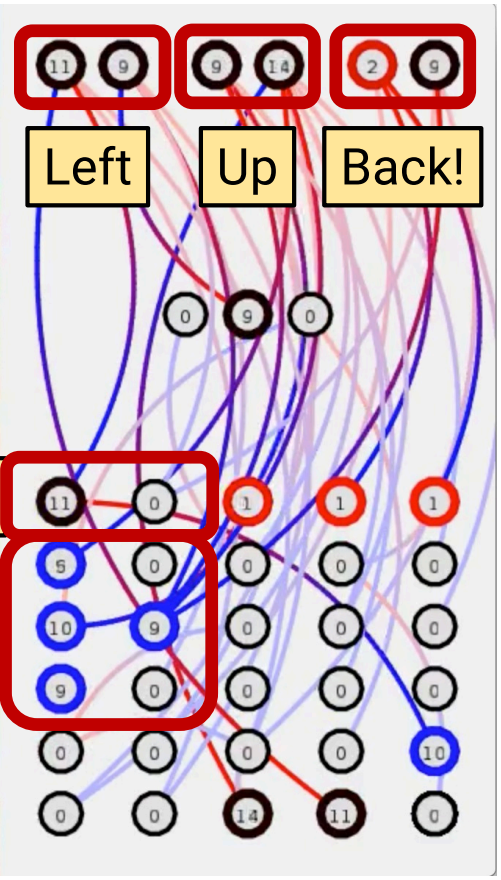


Interesting Thing #2:
Angle left goes directly to the "Rotate Left" neuron.

J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#1: Front-Facing Sense-and-Avoid on mrDANNA (memristive)





Interesting Thing #3:
Sometimes you have conflicting goals.

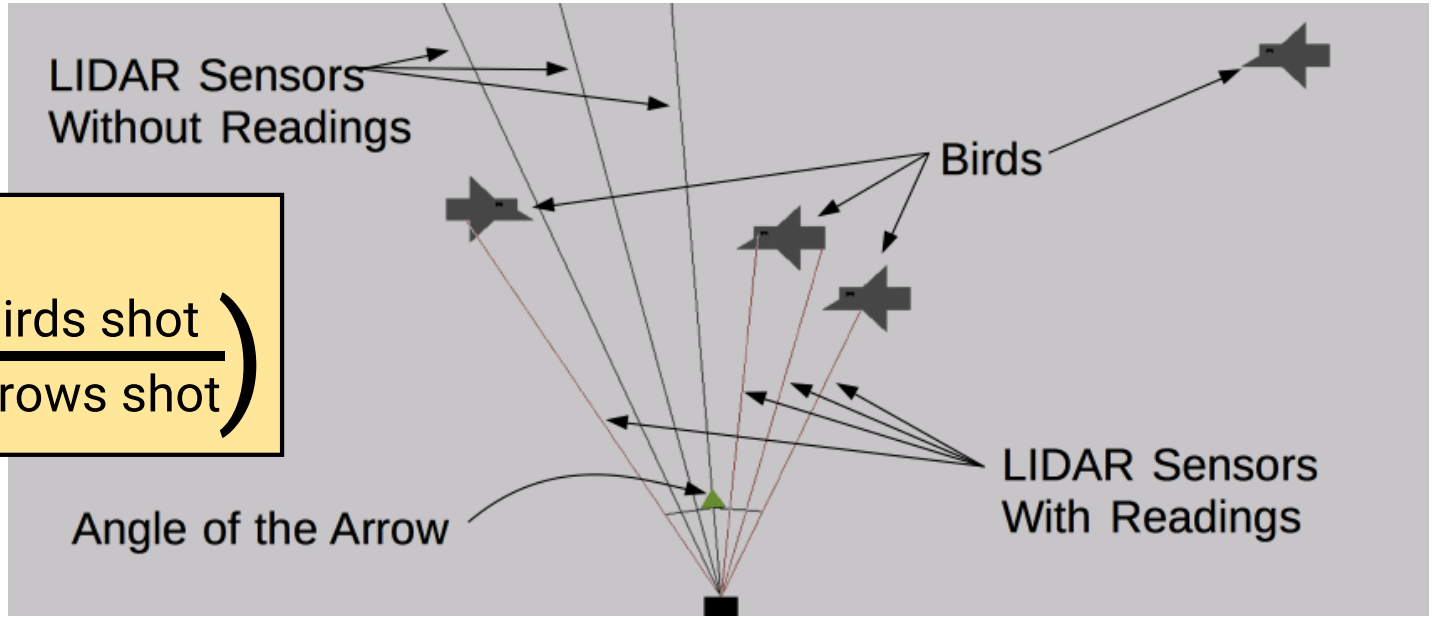
J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#2: Bowman - Targeting

- Bow & arrow with 7 LIDAR sensors, equally spaced.
- Rotate left, rotate right, shoot (with cool down)

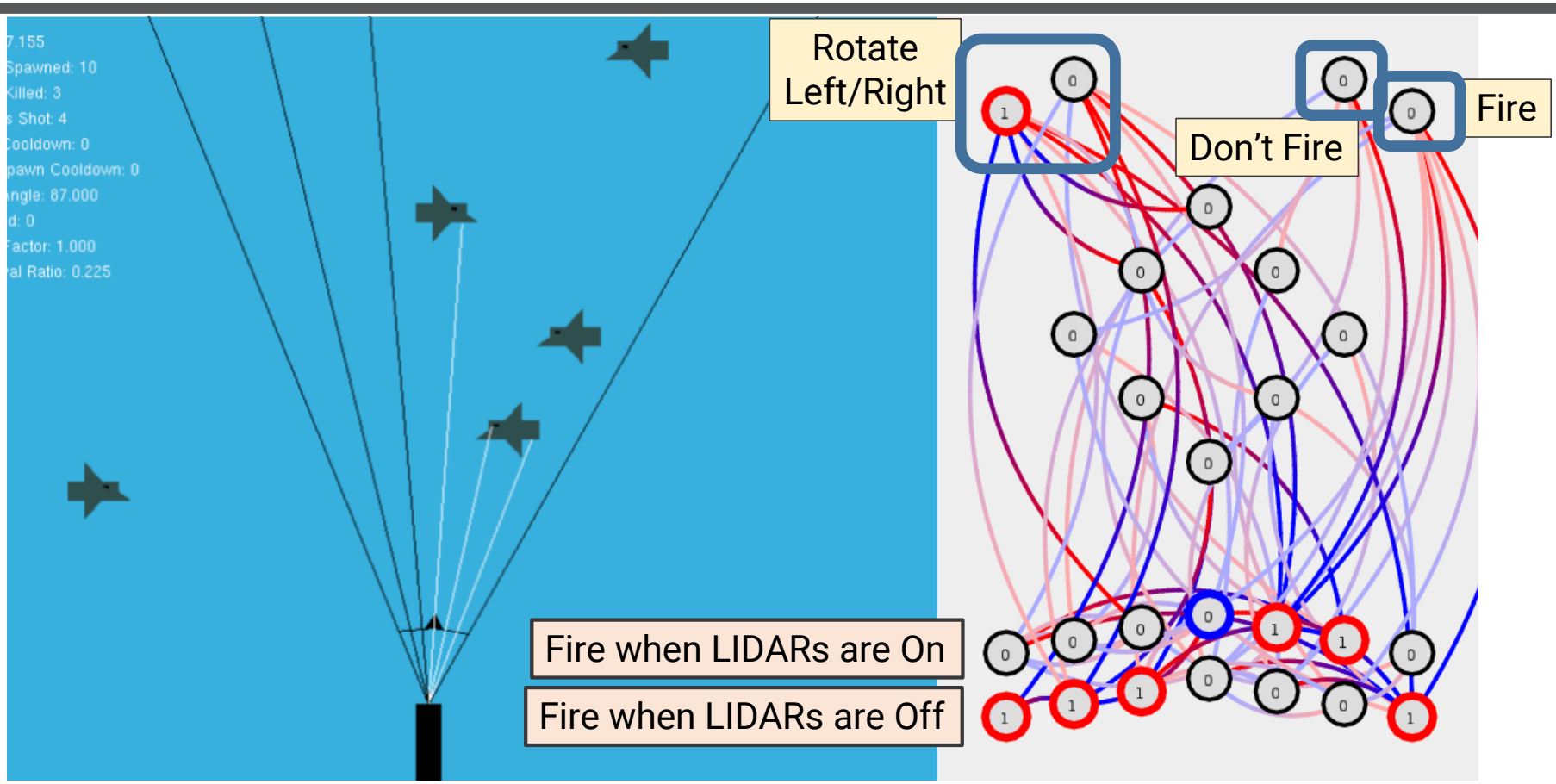
Objective:

$$\left(\frac{\text{birds shot}}{\text{birds spawned}}\right)\left(\frac{\text{birds shot}}{\text{arrows shot}}\right)$$



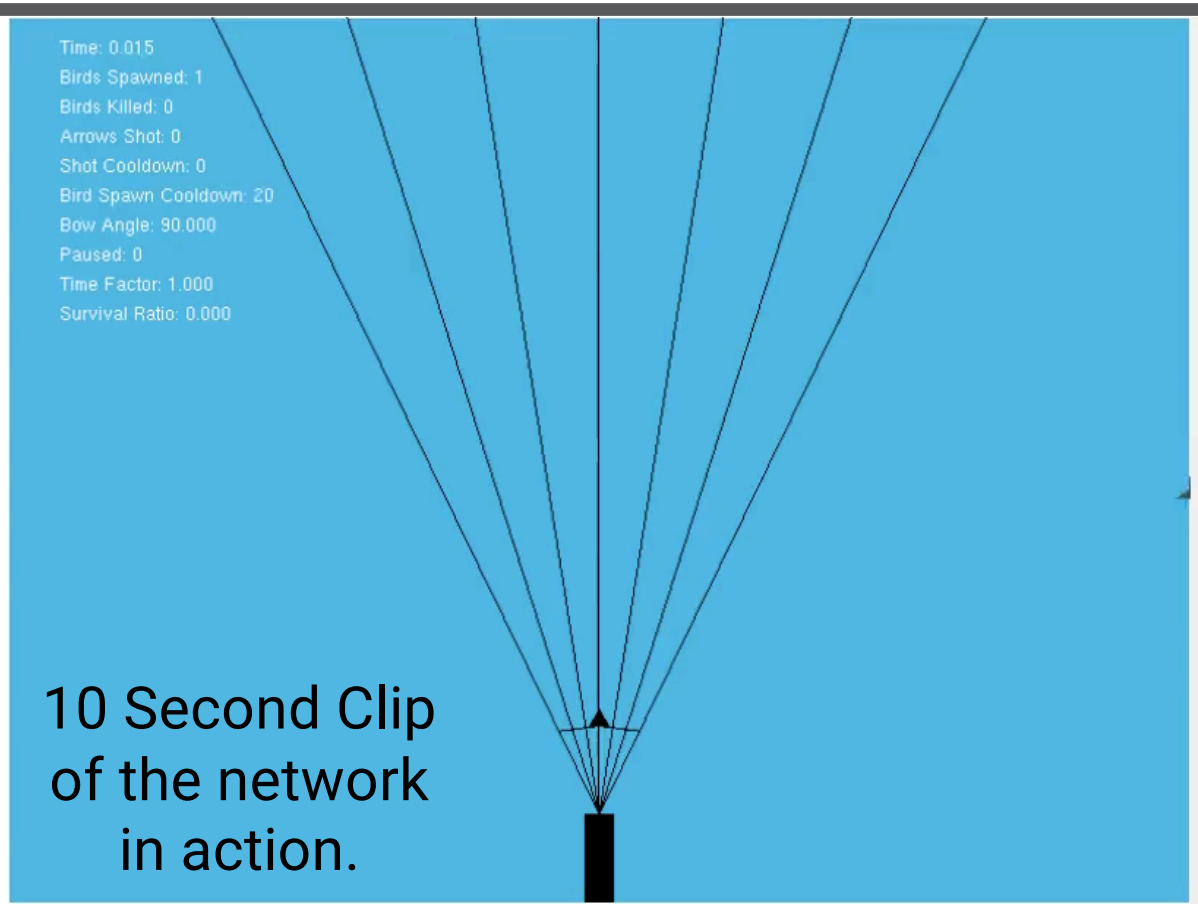
J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#2: Bowman on DANNA 2 (FPGA/ASIC)

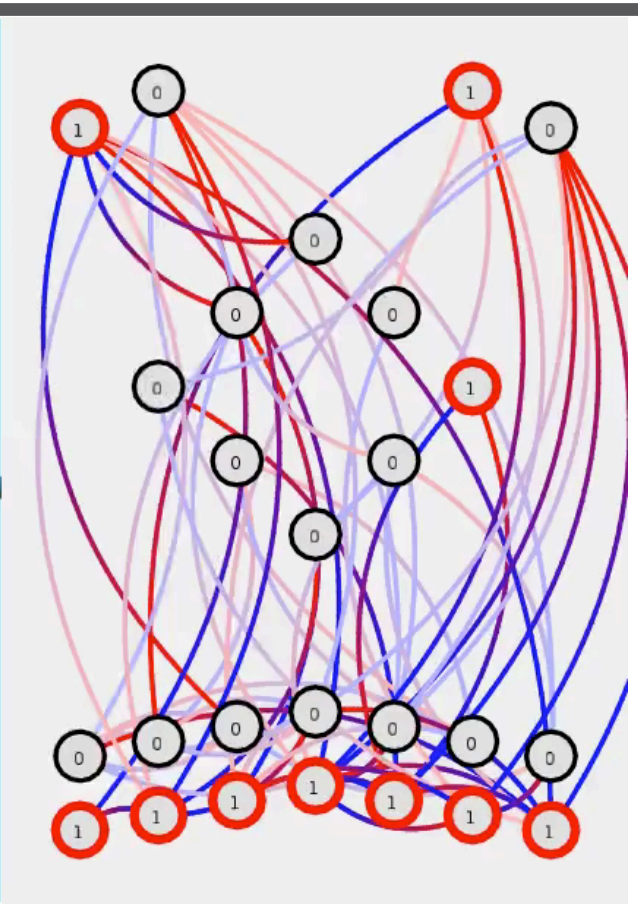


J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#2: Bowman on DANNA 2 (FPGA/ASIC)



10 Second Clip of the network in action.

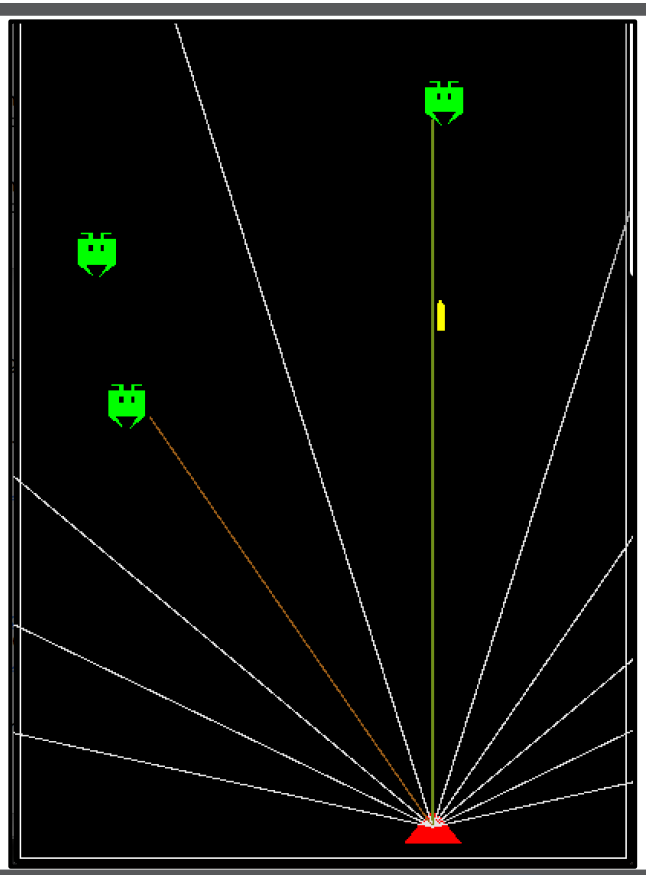


#3: Space Invaders – Movement Planning

- Robot with upward-firing missiles, 11 LIDAR sensors, plus a cooldown.
- Move left, move right, fire

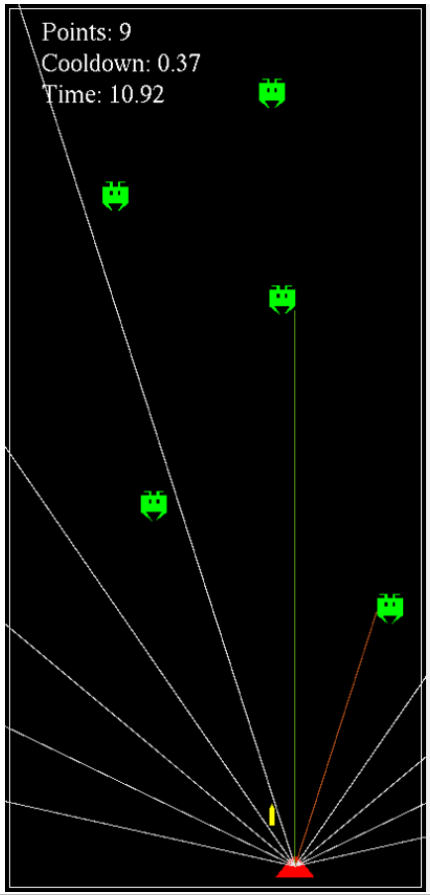
Objective:

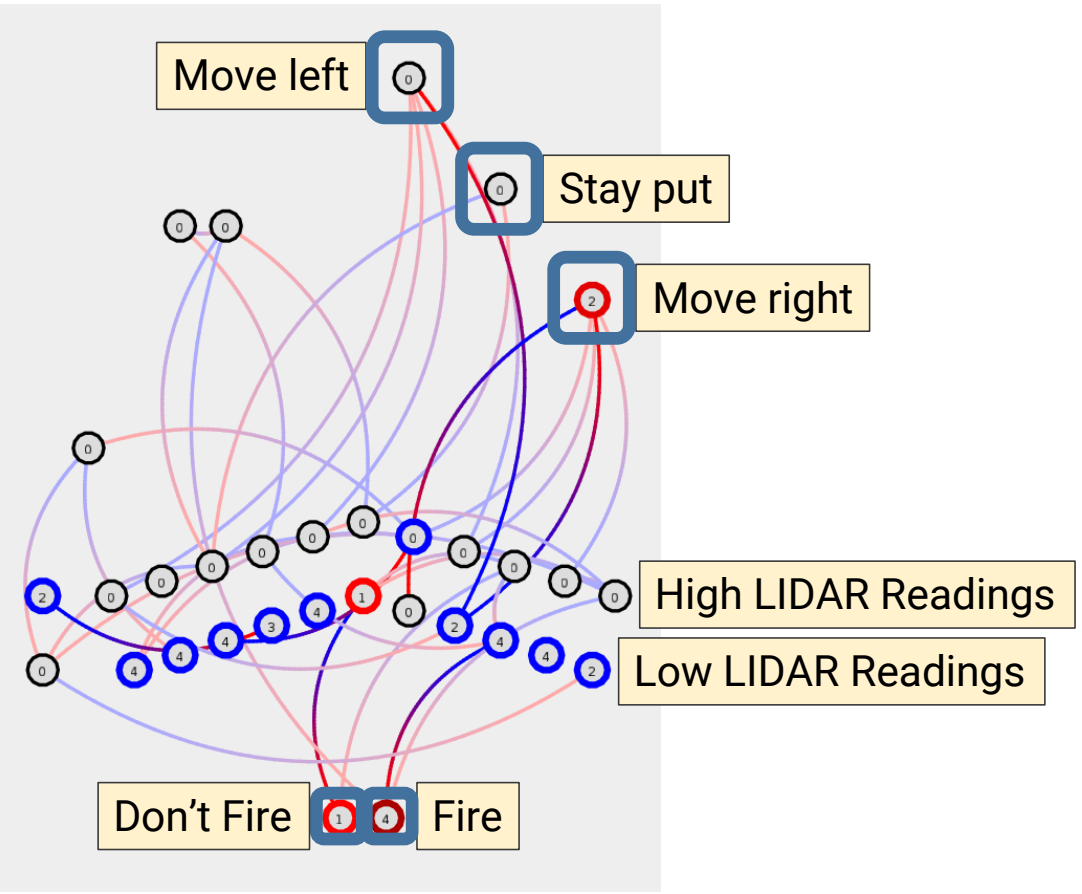
$$\left(\frac{\text{aliens shot}}{\text{missiles fired}}\right)\left(\frac{\text{survival time}}{\text{mission time}}\right)$$



J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#3: Space Invaders on NIDA (Analog, 3D)





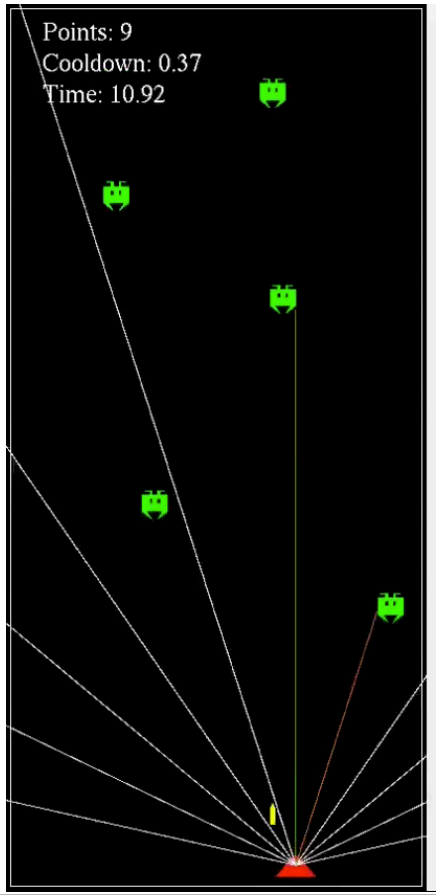
TENN LAB NEUROMORPHIC ARCHITECTURES. LEARNING. APPLICATIONS.

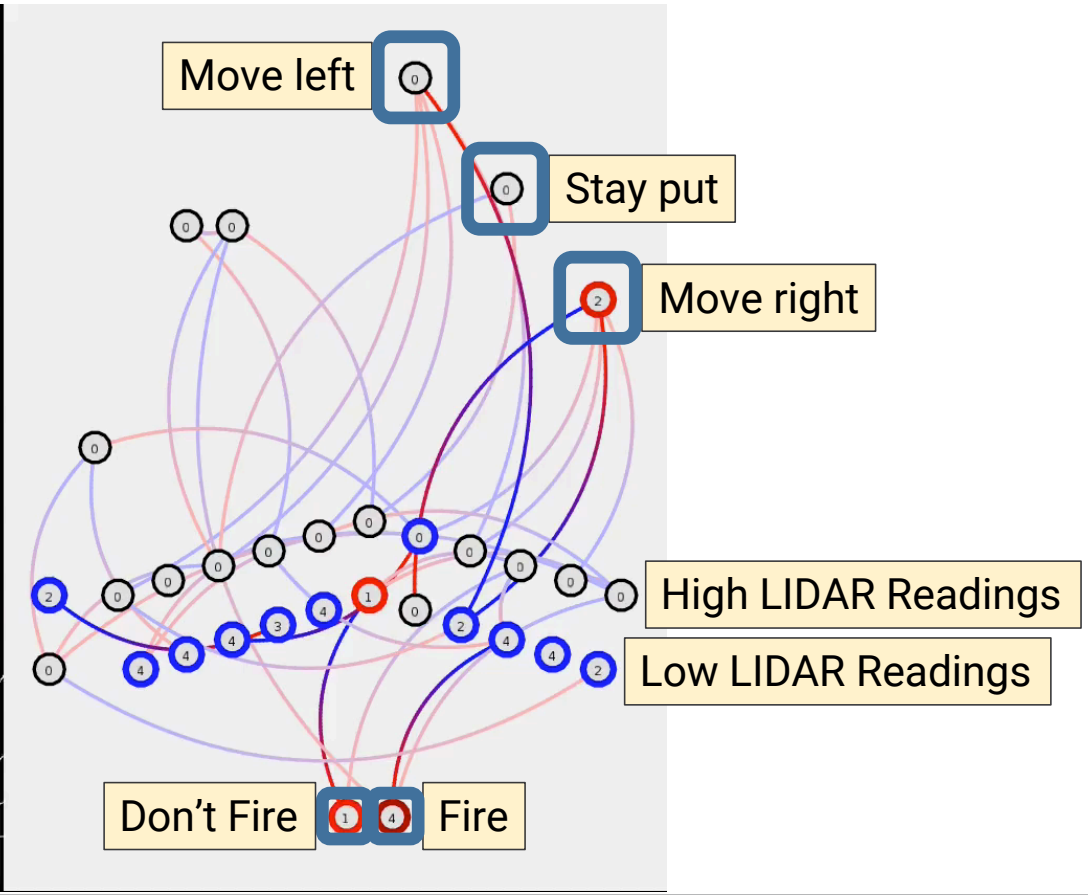
J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

http://neuromorphic.eecs.utk.edu

#3: Space Invaders on NIDA (Analog, 3D)

10 Second Clip of the network in action.





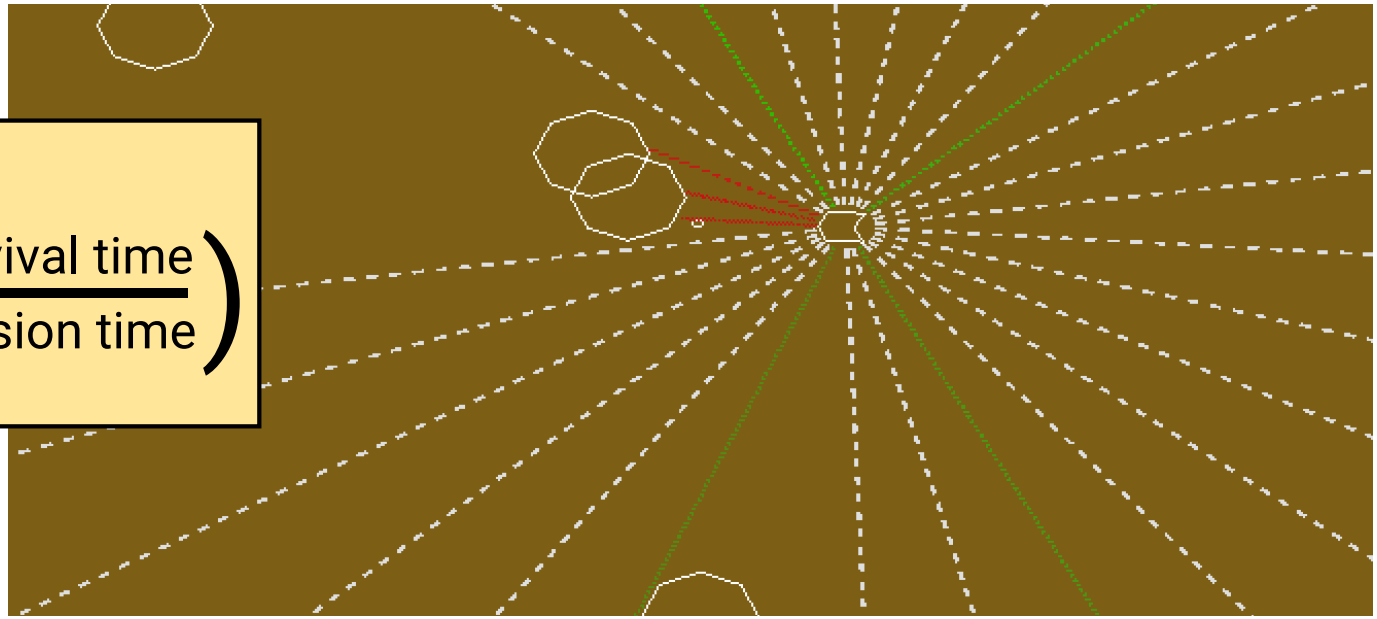
J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

#4: Asteroids – Targeting & Avoidance

- Spaceship with 30 LIDAR sensors in 360 degree field of vision.
- Rotate left, rotate right, thrust, fire.

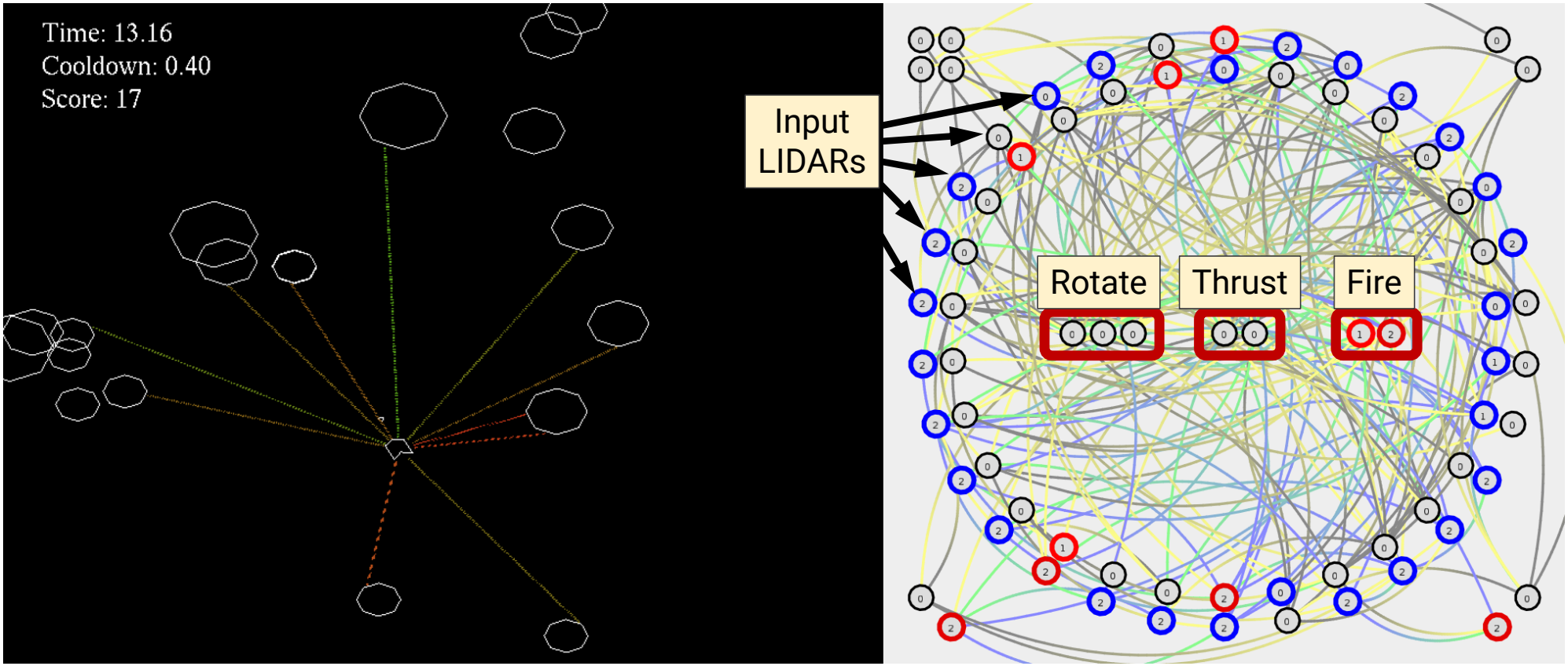
Objective:

$$\left(\frac{\text{asteroids shot}}{\text{max shots}}\right)\left(\frac{\text{survival time}}{\text{mission time}}\right)$$



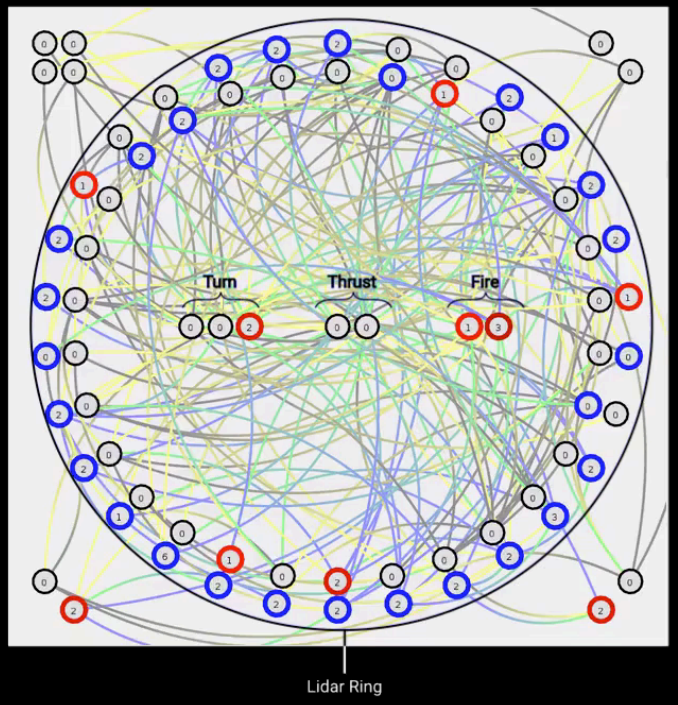
J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

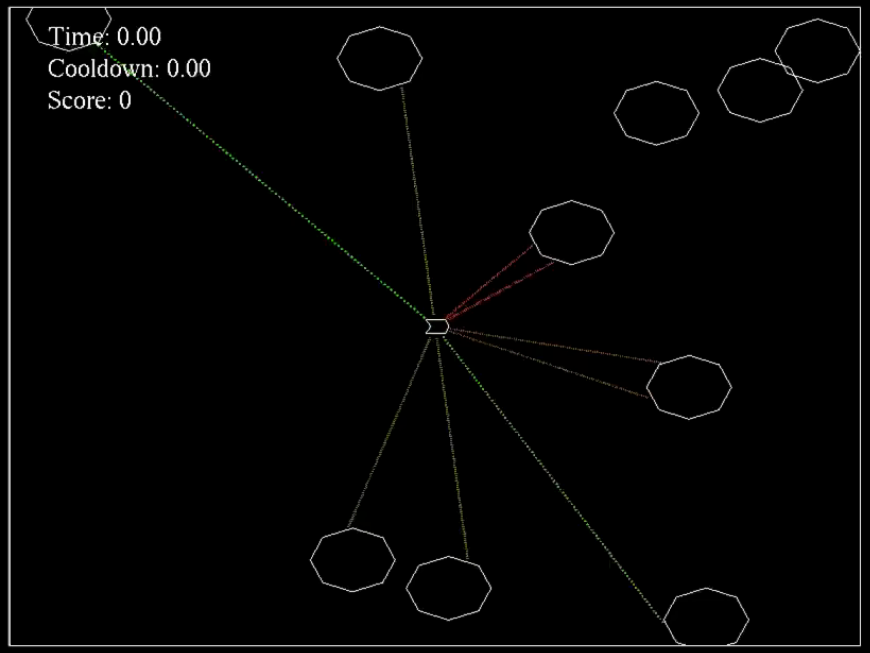
#4: Asteroids on DANNA 2 (FPGA/ASIC)



J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

Danna2 Sparse Neuromorphic Device Plays Asteroids





The right outputs are "don't fire" and "fire". Ties are broken to not fire.

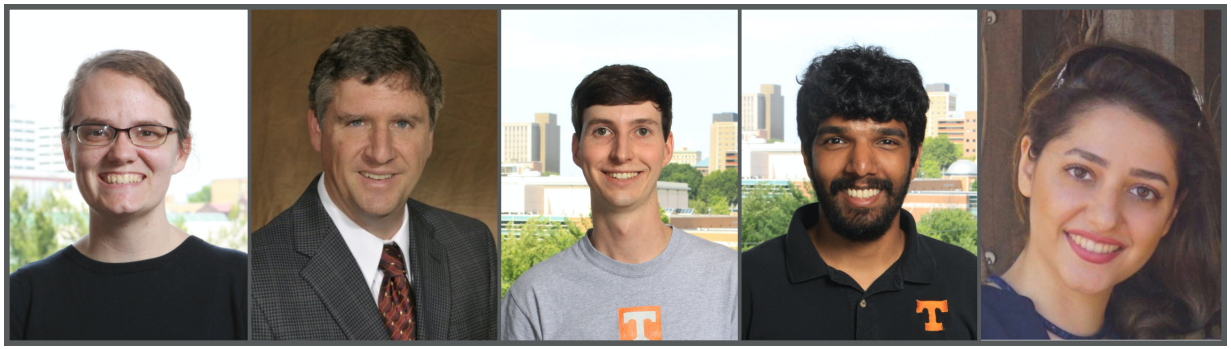
Control Applications - Takeaways

- LIDAR input simple, effective.
- Evolved networks are small.
- EONS paradigm is a natural fit for training.
- Can give insights into network design.
- Would like to compare better with RL algorithms.

J. S. Plank, C. Rizzo, K. Shahat, G. Bruer, T. Dixon, M. Goin, G. Zhao, J. Anantharaj, C. Schuman et al, "The TENNLab Suite of LIDAR-Based Control Applications for Recurrent, Spiking, Neuromorphic Systems" *GOMACTech*, 2019.

Project #3

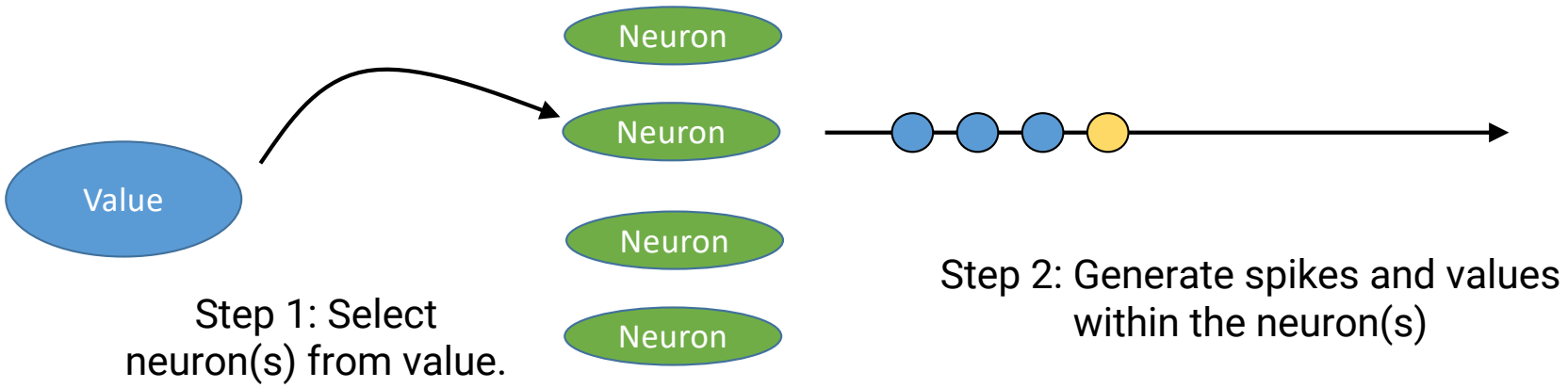
Evaluating Input Encodings



C. D. Schuman, J. S. Plank, G. Bruer and J. Anantharaj, Non-Traditional Input Encoding Schemes for Spiking Neuromorphic Systems, *IJCNN: The International Joint Conference on Neural Networks*, 2019

Evaluating Input Encodings

Explore the ramifications of variants of population coding on the training success of classification and control applications.



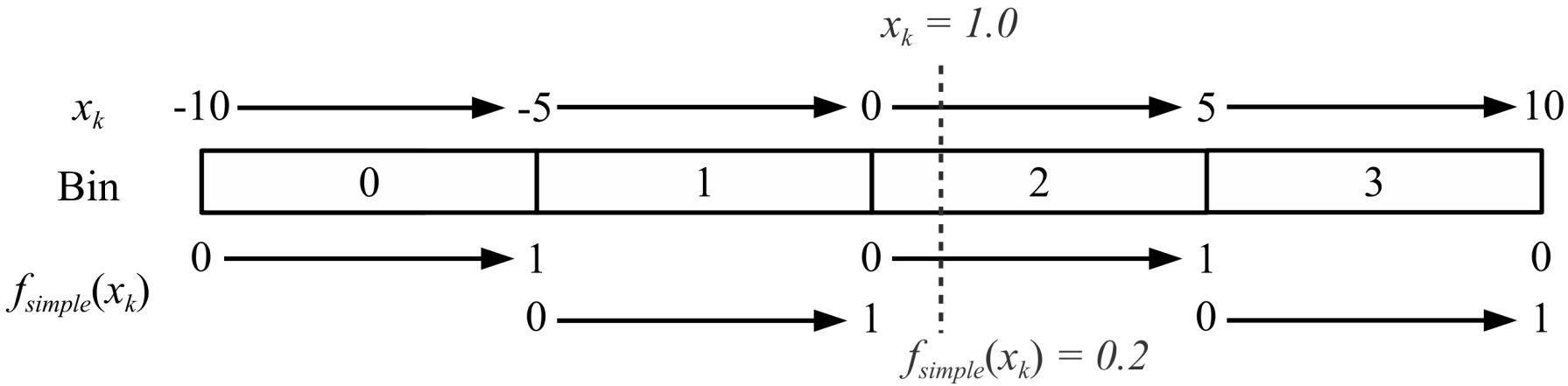
Step 1: Select neuron(s) from value.

Step 2: Generate spikes and values within the neuron(s)

C. D. Schuman, J. S. Plank, G. Bruer and J. Anantharaj, Non-Traditional Input Encoding Schemes for Spiking Neuromorphic Systems, *IJCNN: The International Joint Conference on Neural Networks*, 2019

Evaluating Input Encodings

To convert input values to spikes, partition the domain into bins, and use one input neuron per bin.

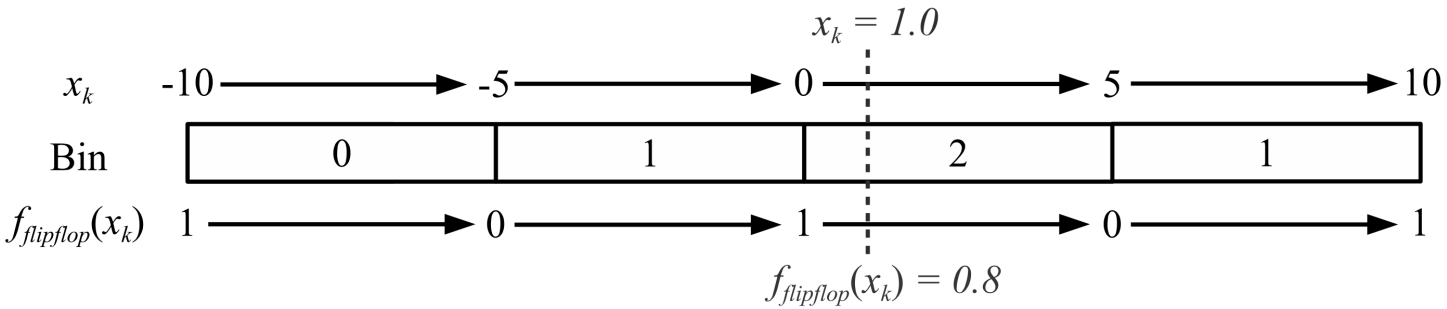


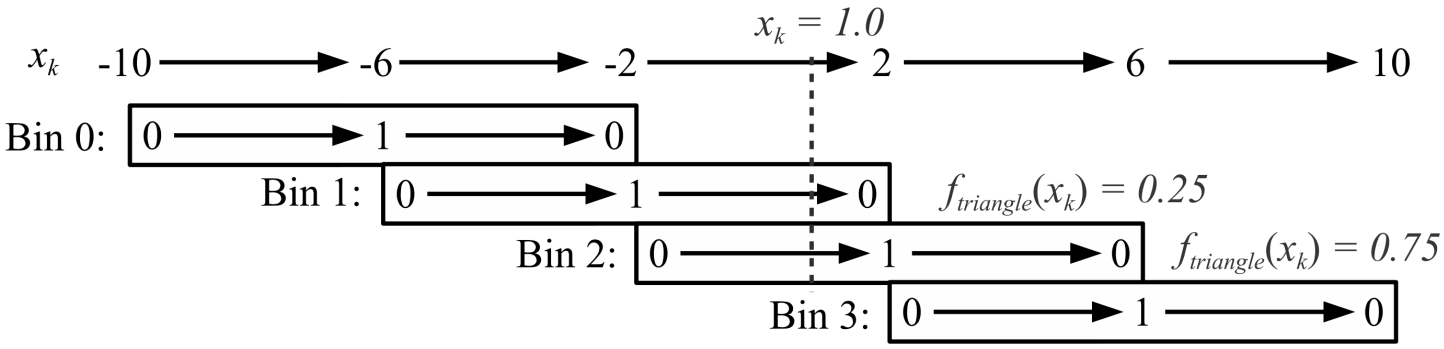
Once the bin is determined, the value is converted to a second value between 0 and 1, which is the input into the bin.

Evaluating Input Encodings

The second value's determination can vary: Simple, Flip-flop, Triangle.

Flip-Flop





Triangle

Evaluating Input Encodings

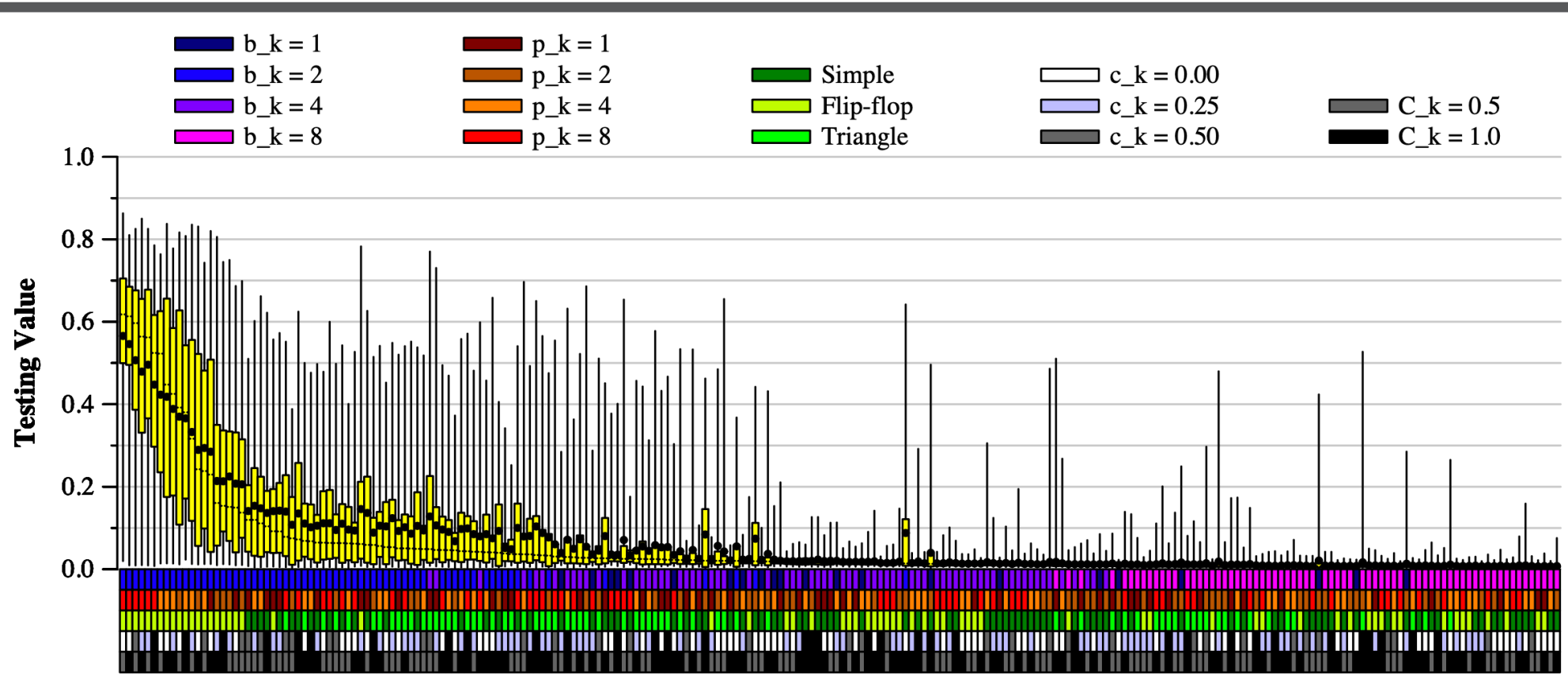
And within the bin, there are multiple ways to enter the second value:

- Vary minimum spike value
- Vary maximum spike value
- Number of spikes

Experiment

- Two classification applications (Radio, 3-MNIST)
- Two control applications (Pole balance, Robonav)
- Four neuroprocessors
- Training with genetic algorithm (EONS)
- TENNLab Software Framework
- 230 Hyperparameter combinations per test

Sample Results (Pole Balancing)



C. D. Schuman, J. S. Plank, G. Bruer and J. Anantharaj, Non-Traditional Input Encoding Schemes for Spiking Neuromorphic Systems, *IJCNN: The International Joint Conference on Neural Networks*, 2019

Input Encoding: Conclusions

- Encoding technique has a profound effect on training success.
- Of course, there's no "one-size-fits all".
- Can use Bayesian optimization to reach same conclusions with far fewer tests (2nd paper).
- Gives us a starting point for new applications.

We call it Fred: `{ "spikes": { "flip_flop": 2, "ov_interval": 8 }`

- Could use some better theory.

C. D. Schuman, J. S. Plank, G. Bruer and J. Anantharaj,
"Non-Traditional Input Encoding Schemes in Spiking Neuromorphic Systems," *IJCNN*, 2019.

M. Parsa, P. Mitchell, C. D. Schuman, R. M. Patton, T. E. Potok and K. Roy
"Bayesian-Based Hyperparameter Optimization for Spiking Neuromorphic Systems," IEEE Conference on Big Data, 2019.

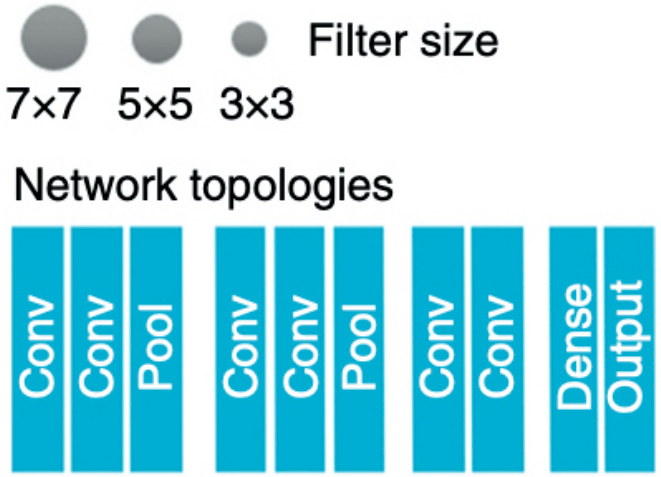
Project #4:

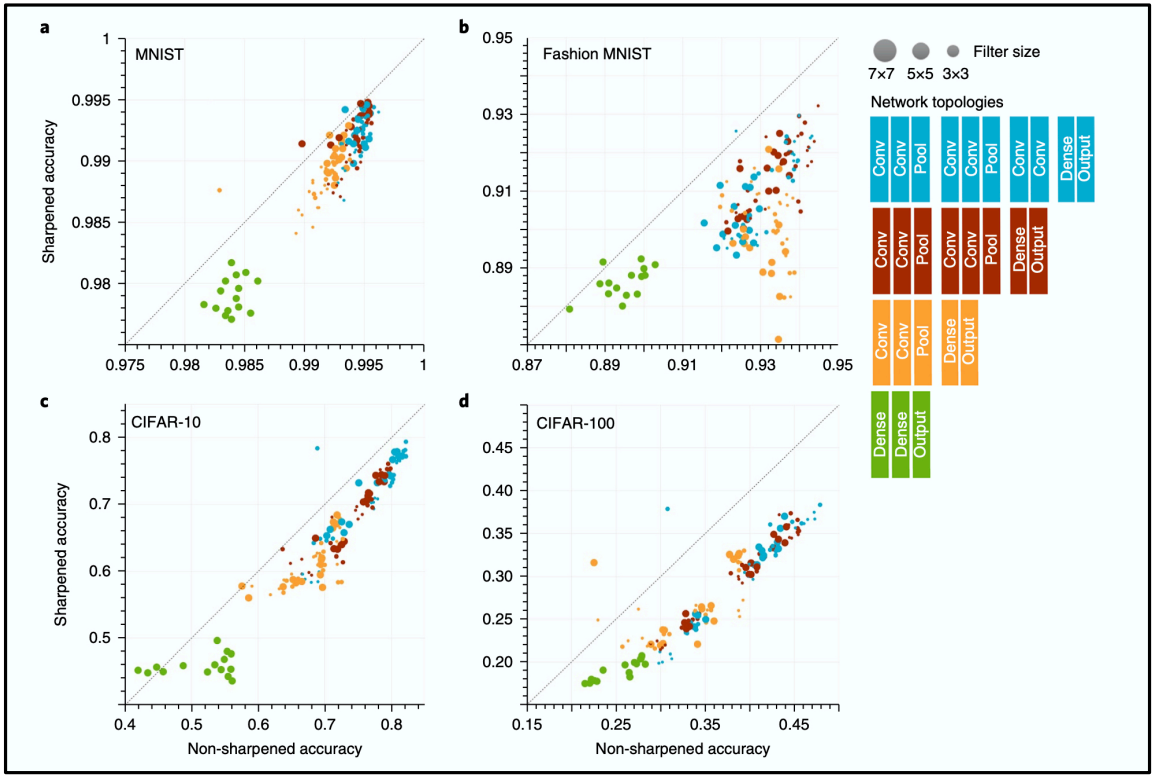
Compressing convolutional layers in spiking Deep networks.

J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

The Whetstone Approach

- Train in Keras
- Port to your favorite SNN.

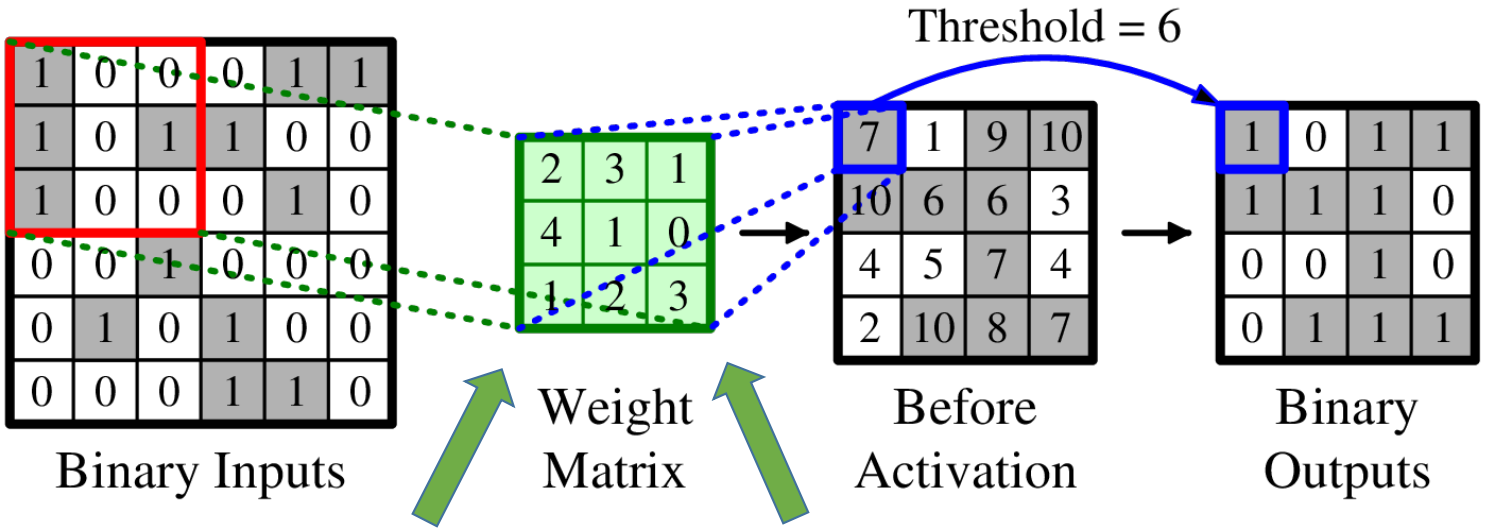




W. Severa, C. M. Vineyard, R. Dellana, S. J. Verzi and J. B. Aimone, "Training Deep Neural Networks for Binary Communication with the Whetstone Method", *Nature Machine Intelligence*, January, 2019, pp. 86-95.

The Whetstone Approach

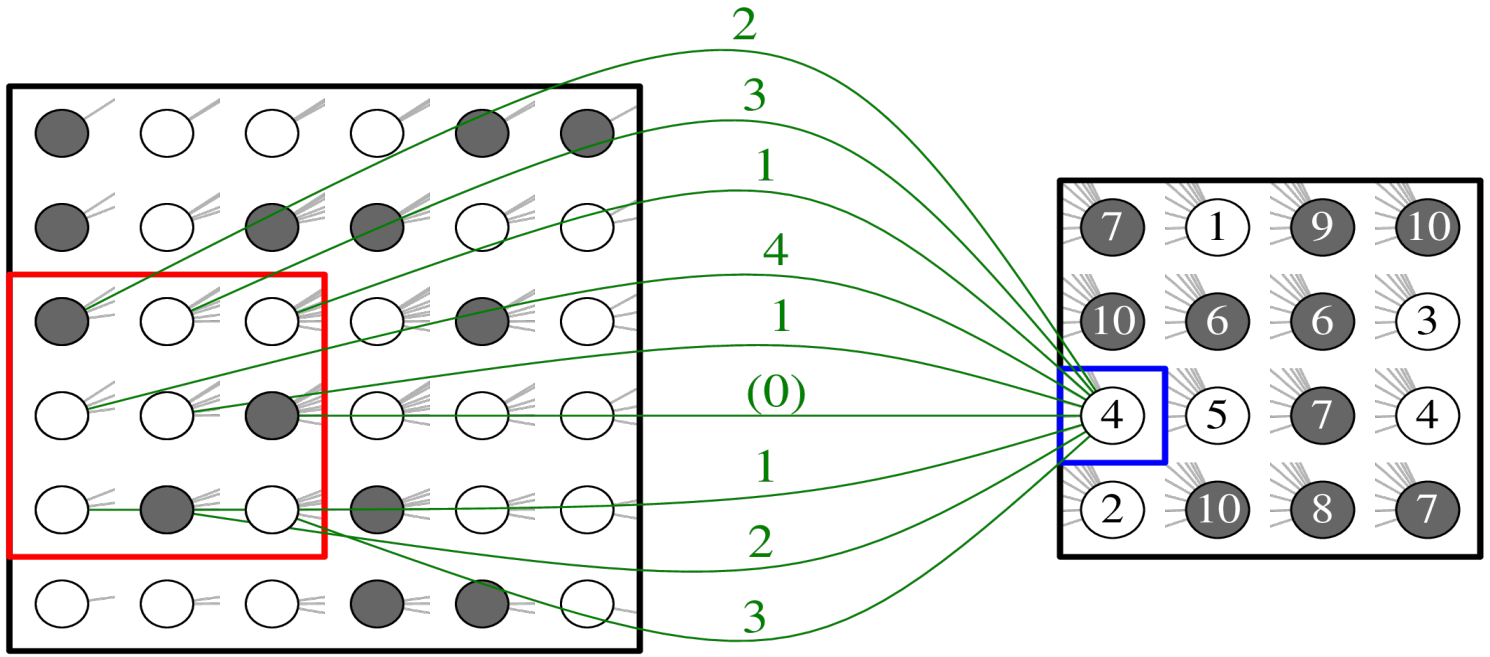
Let's take a look at the convolution operation.



In a standard ANN, there is one of these used for all inputs/outputs in a layer.

The Whetstone Approach

When you convert this to a SNN, each multiplication becomes a synapse.



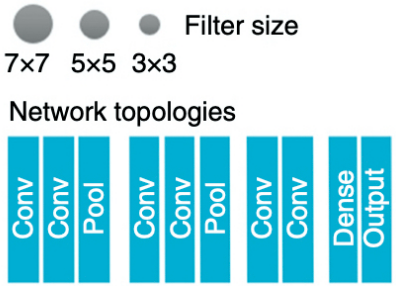
2	3	1
4	1	0
1	2	3

Weight Matrix

J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

The Whetstone Approach

And these synapses add up!

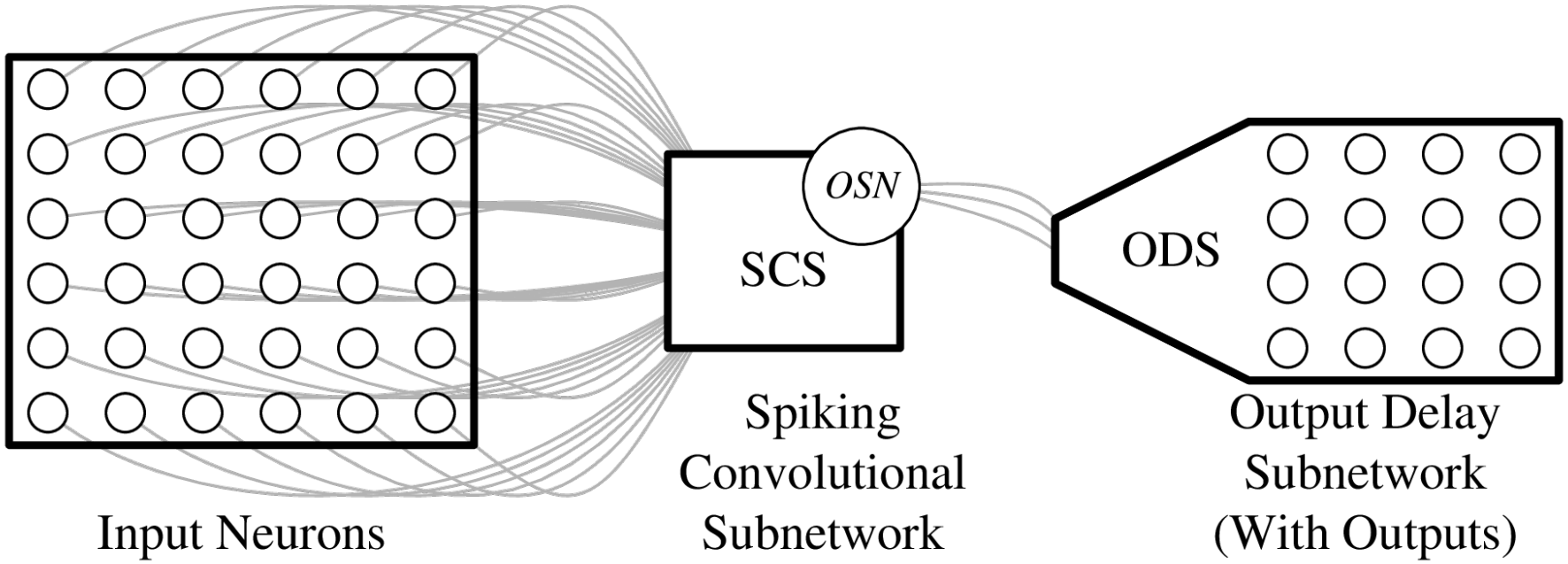


Largest MNIST Network

Layer	Outputs	Synapses	Weights
1. Input	784	0	0
2. 7x7 convolution, 32x28x28	25,088	1,229,312	1,568
3. 7x7 convolution, 64x28x28	50,176	78,675,968	100,352
4. Maxpool 2x2, 64x14x14	12,544	12,544	0
5. 5x5 convolution, 64x14x14	12,544	20,070,400	102,400
6. 5x5 convolution, 64x14x14	12,544	20,070,400	102,400
7. Maxpool 2x2, 64x7x7	3,136	3,136	0
8. 3x3 convolution, 128x7x7	6,272	3,612,672	73,728
9. 3x3 convolution, 128x7x7	6,272	7,225,344	147,456
10. Flatten to 254, fully conn.	254	1,593,088	1,593,088
11. Flatten to 40, fully conn.	40	10,160	10,160
Total	129,654	132,503,024	2,131,152

J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

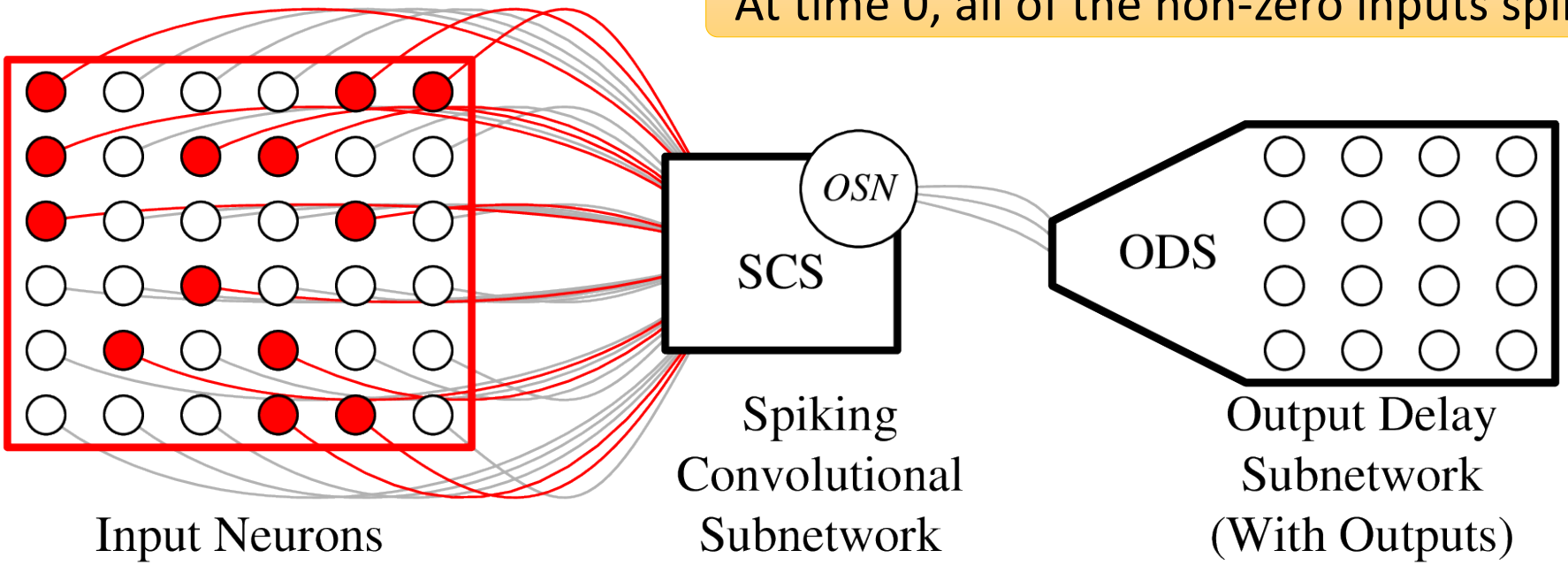
The High Level Picture



J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

The High Level Picture

At time 0, all of the non-zero inputs spike.

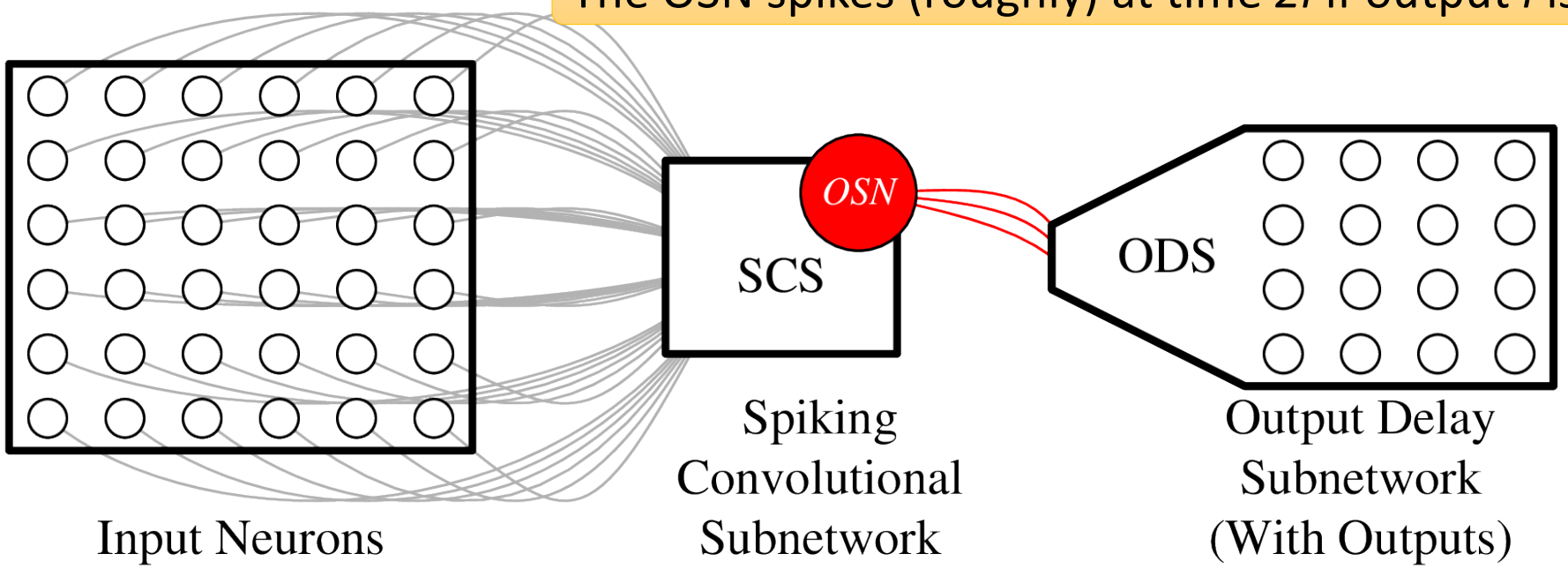


And then the SCS gets to work.

J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

The High Level Picture

The OSN spikes (roughly) at time *2i* if output *i* is 1.

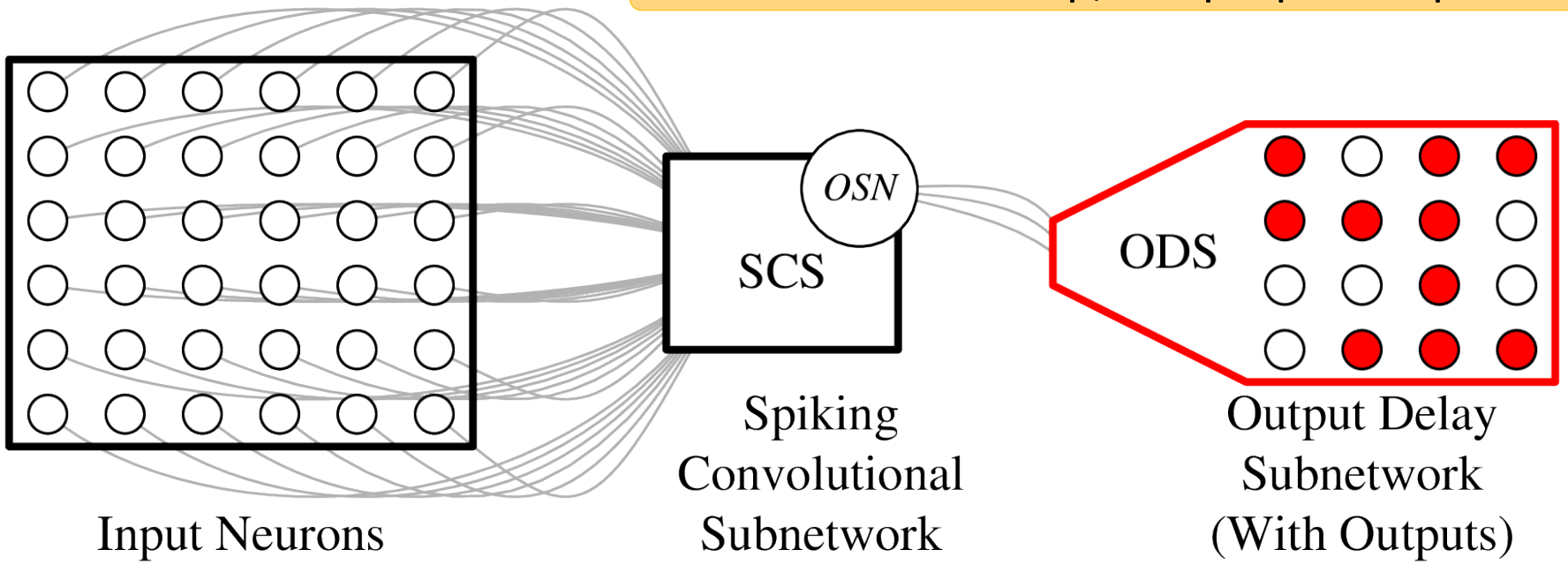


The ODS routes those spikes to the correct outputs.

J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

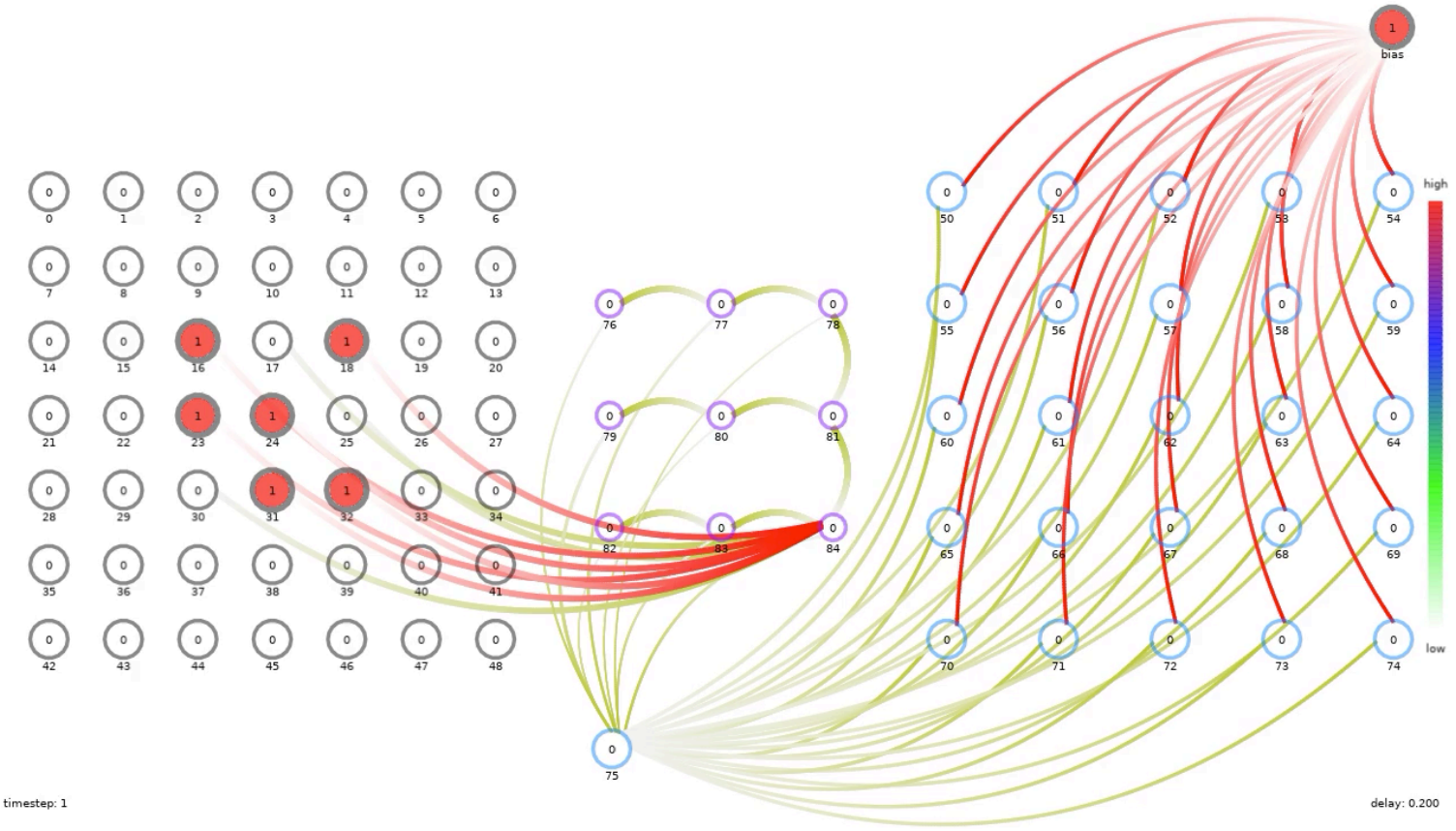
The High Level Picture

At the final timestep, the proper outputs fire.



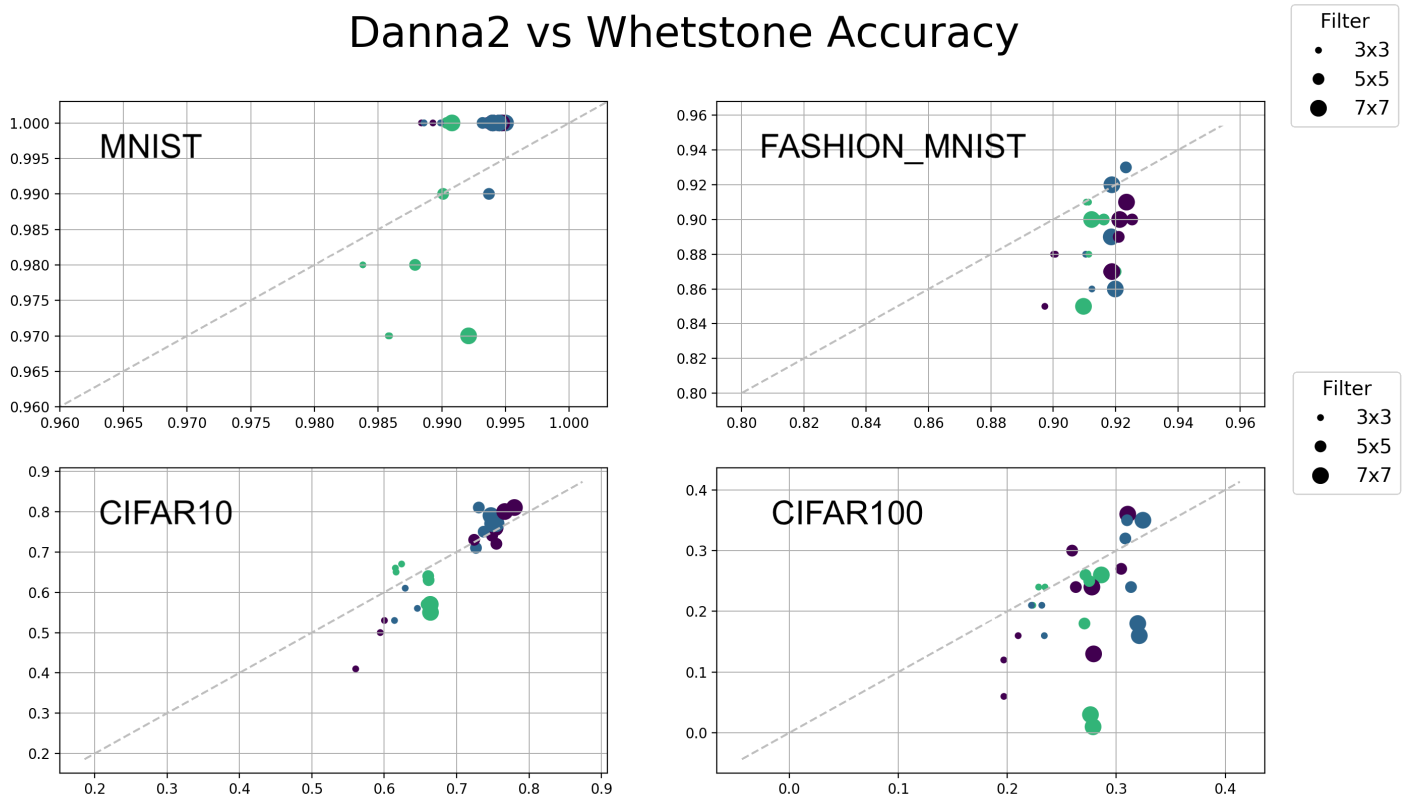
J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

The details are beyond this talk, but look at this example:



Experiment: Keras -> Whetstone -> DANNA2

Nearly 20x fewer synapses.



J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

Status

- Training module in TENNLab Software Framework
- Verified with DANNA 2 and “GNP” neuroprocessors
- Nice target application for SNN neuroprocessors.

J. S. Plank, J. Zhao and B. Hurst, Reducing the Size of Spiking Convolutional Neural Networks by Trading Time for Space, *IEEE International Conference on Rebooting Computing (ICRC)*, December, 2020.

Recap

- Nate Cady: Fabricating CMOS/RRAM Hardware
- Garrett Rose: Memristor modeling, circuits and systems
- Katie Schuman: Neuromorphic fuel efficiency
- Sampling of my projects
  - Software Framework
  - Control applications
  - Input Encoding
  - Compressing convolutional layers