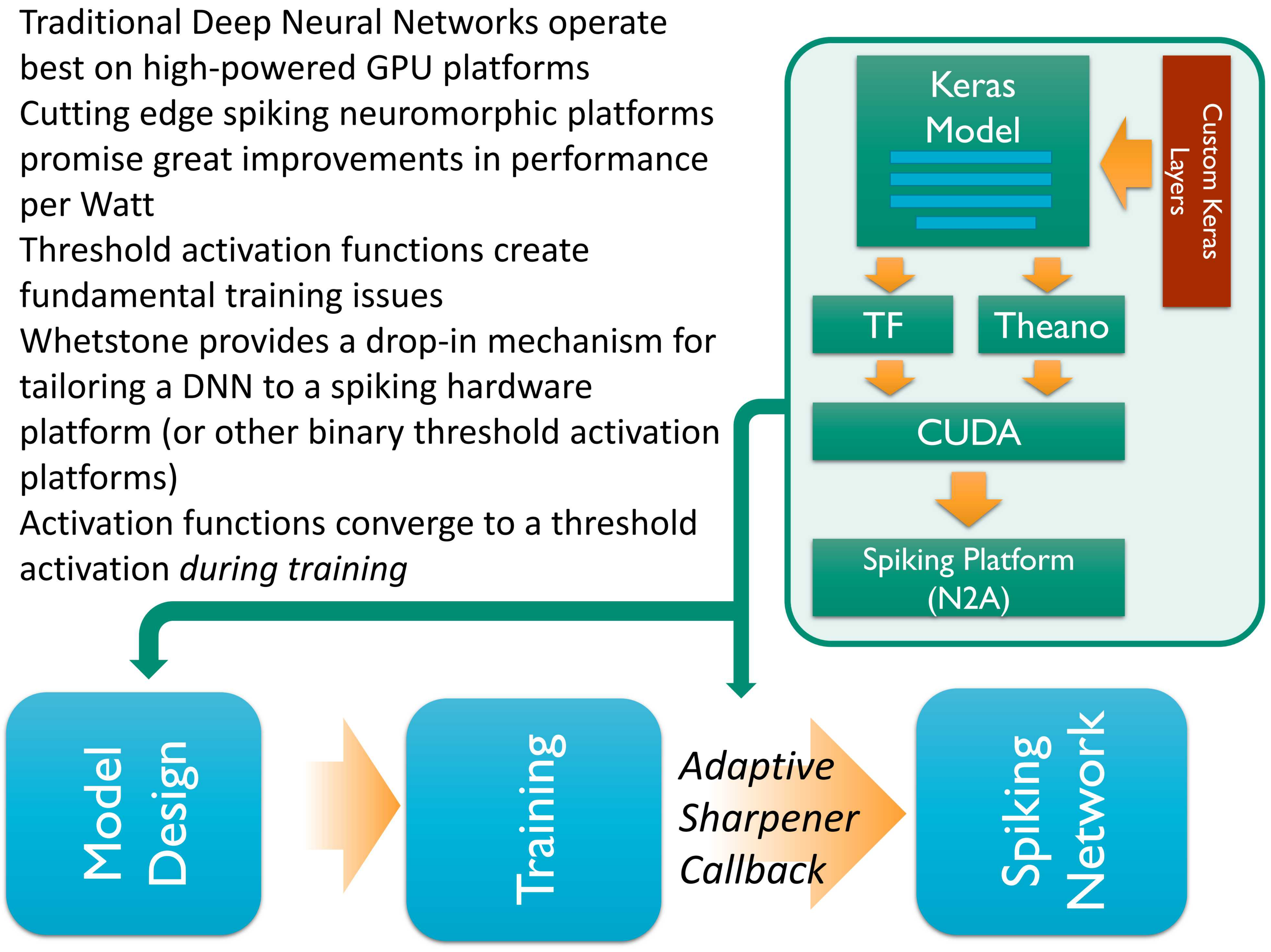
SAND2018-1383C

Whetstone: An accessible, platform-independent method for training spiking deep neural networks for neuromorphic processors

William Severa*, Ryan Dellana, Craig M. Vineyard and James B. Aimone
Sandia National Laboratories, Albuquerque, NM

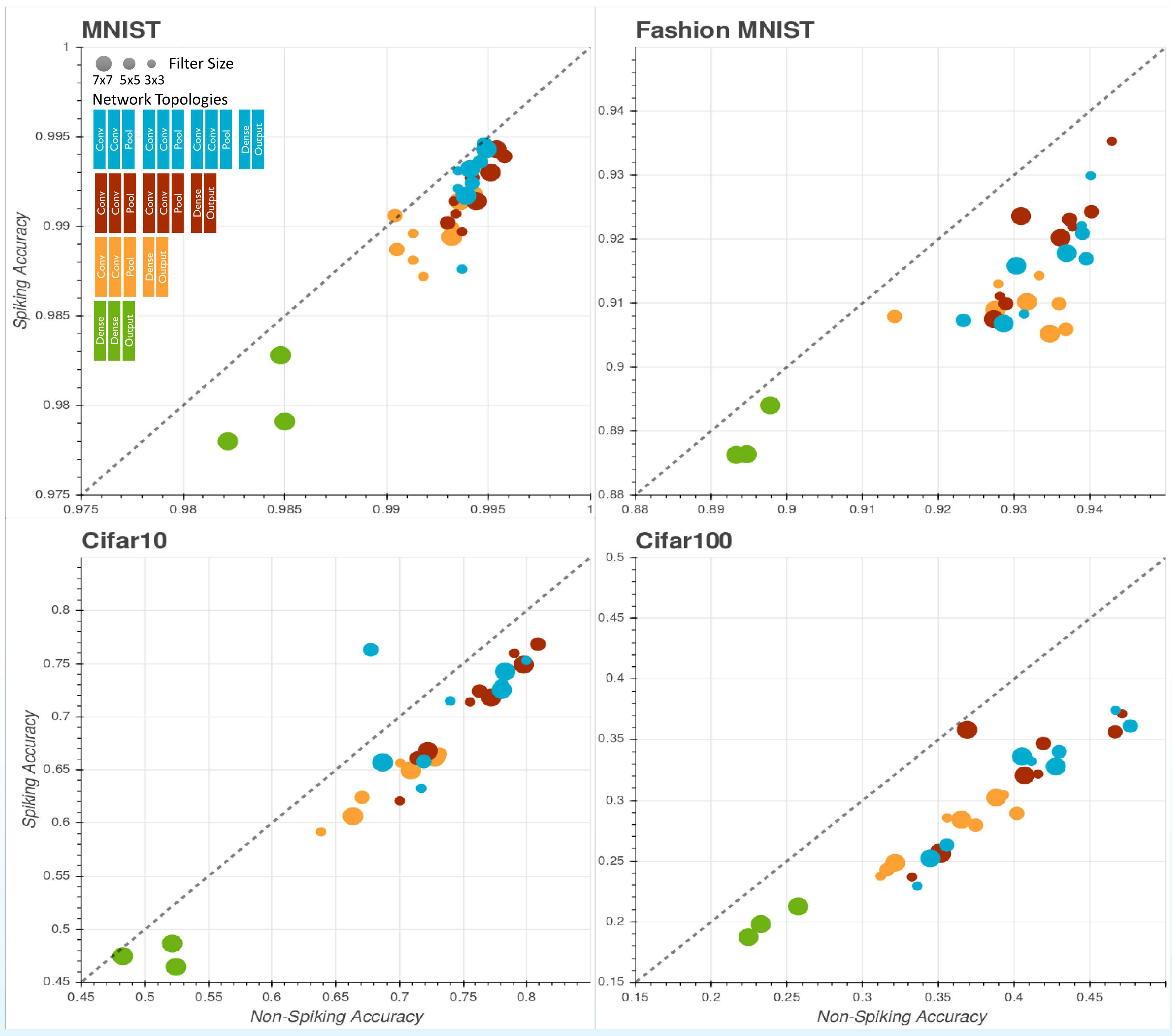
Utilizing Spiking Neuromorphic Hardware

- Traditional Deep Neural Networks operate best on high-powered GPU platforms
- Cutting edge spiking neuromorphic platforms promise great improvements in performance per Watt
- Threshold activation functions create fundamental training issues
- Whetstone provides a drop-in mechanism for tailoring a DNN to a spiking hardware platform (or other binary threshold activation platforms)
- Activation functions converge to a threshold activation *during training*



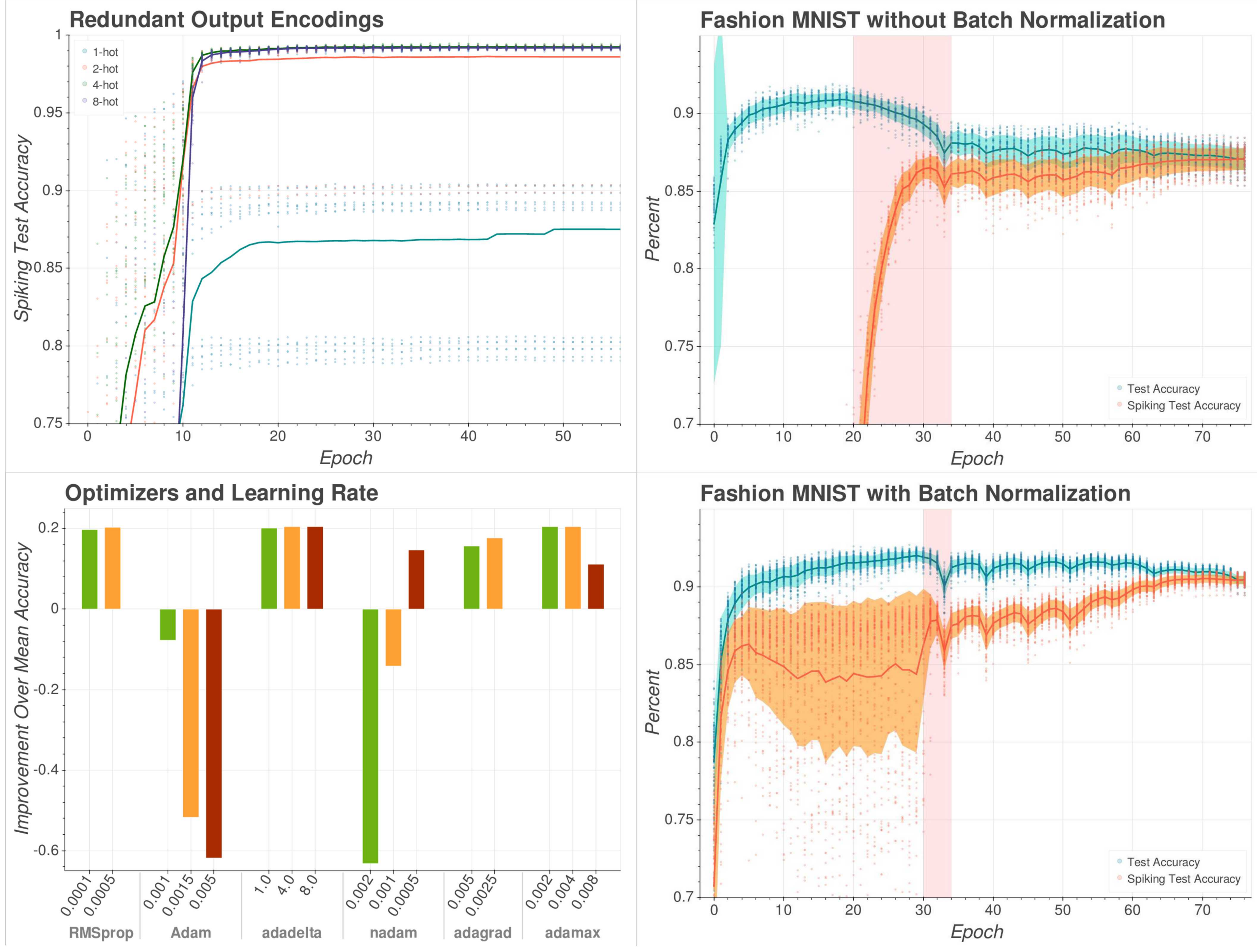
Model Modifications Ensure Minimum Accuracy Loss

- Modest performance losses across several datasets (MNIST, Fashion MNIST, Cifar10, Cifar100) on *unmodified network models*
- Spiking Neural Networks are typically more brittle than traditional deep neural networks
- This suggests design considerations for best performance
  - Redundancy in output encodings offers a voting-scheme-type decision process
  - Large convolutional filters improve training stability and network performance for some topologies
  - Choice of optimizer is critical for consistent convergence
- With these basic modifications, spiking accuracy can be competitive
- We require no post-hoc analysis or additional time cost (1 DNN layer = 1 SNN layer)



Compartmented Design Leverages Existing Proven Techniques

- Modifications for the network topology are limited to the activation function and output layer
- Many standard, effective techniques translate immediately to the spiking neural network
  - Dropout
  - Max Pooling
  - Batch Normalization
- Batch normalization greatly improves convergence to spiking activations
  - Majority of accuracy degradation occurs during the sharpening of the first layer
  - Batch normalization helps mitigate this loss
  - Useful for even smaller networks
- Activation sharpening is optimizer agnostic → However, certain optimizers are better suited. Moving average modulation improves repeatability.
- Adaptive sharpener allows easy convergence to spiking thresholds
  - Automated, controls-based mechanism
  - Implemented as a callback
  - More consistent than hand-tuning



Enabling Wide and Easy-to-Implement Adoption

- Neuromorphic hardware platforms are appealing for a wide variety of low-power, embedded applications
- Sophistication and expertise required to make use of these platforms creates a high barrier of entry
- Whetstone enables deep learning experts to easily incorporate spiking hardware architectures
- Networks are portable and hardware-agnostic
- Some benefits of the convergent activation method:
  - Low barrier of entry, built on standard libraries (Keras, Tensorflow, CUDA, etc.)
  - No post-hoc analysis, no added time complexity
  - Only simple integrate-and-fire neurons are required
  - Compatible with standard techniques like dropout and batch normalization

*POC: wmsever@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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.