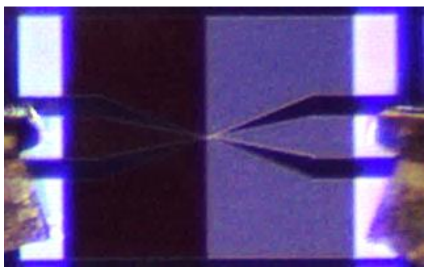
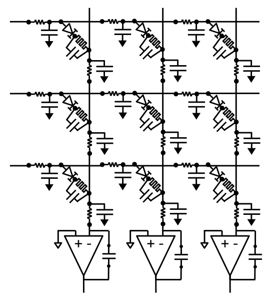
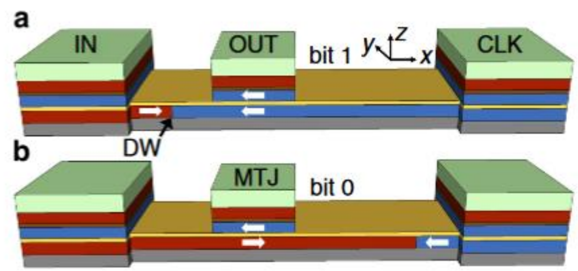
*Exceptional service in the national interest*

Sandia National Laboratories








# Evaluating complexity and resilience trade-offs in emerging memory inference machines

Christopher H. Bennett*, Ryan Dellana, T. Patrick Xiao , Ben Feinberg, Sapan Agarwal, Suma Cardwell, Matthew J. Marinella, William Severa, Brad Aimone

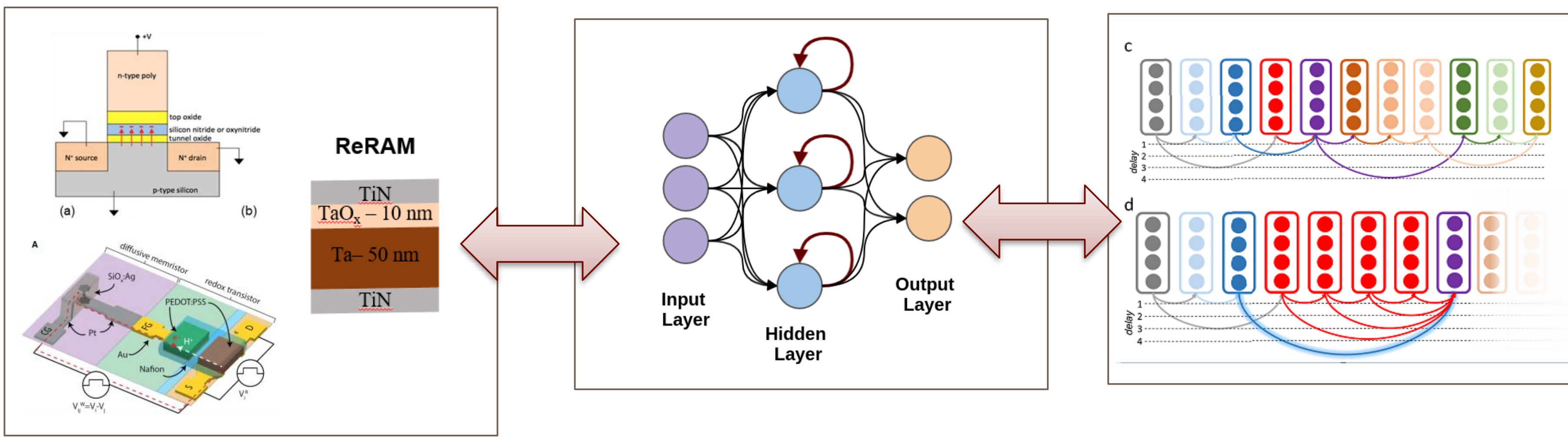
*cbennet@sandia.gov

Center for Computing Research, Sandia National Laboratories

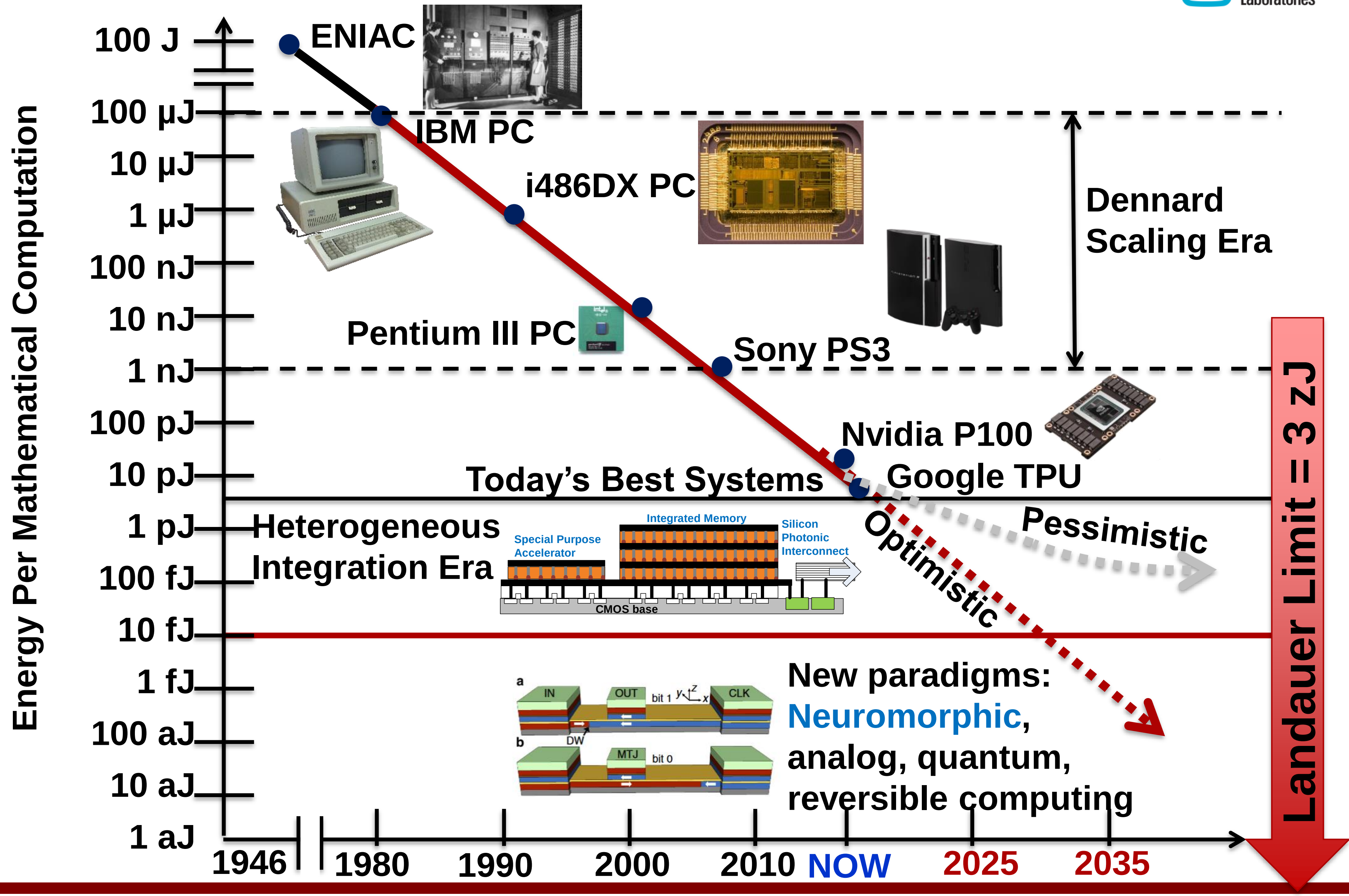
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

# Outline

- **Problem Statement :**
  - **Interest in emerging memory for efficient inference engines**
- **Early results on recurrent neural networks**
- **Future steps and summary**

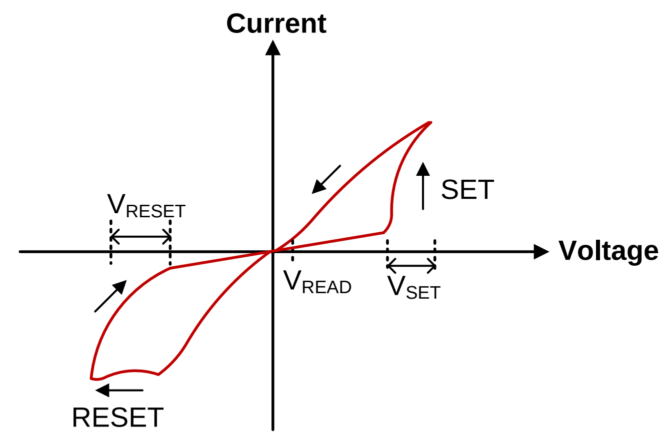


# Evolution of Computing Machinery



# Realizing physical matrix kernels

- Ideal Vector-Matrix Mulitply :
  - Electrically realisable using Kirchoff's + Ohm's laws
- Programmable resistors - e.g. ReRAM/MRAM devices- key component
  - <u>Small voltages to read (inference)</u>
  - Large voltages to program



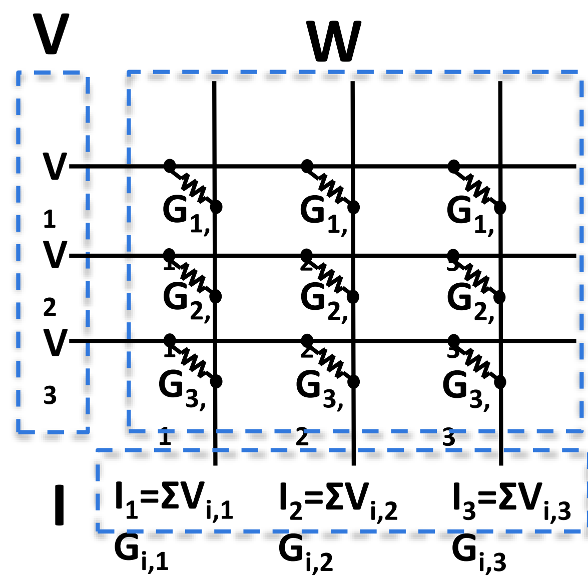

**Mathematical**

$$V^T W = I$$

$$\begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix} =$$

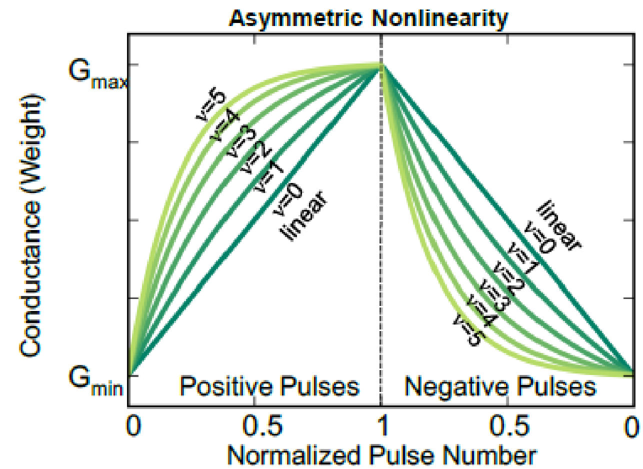
$$\begin{bmatrix} I_1=\Sigma V_{i,1} W_{i,1} & I_2=\Sigma V_{i,2} W_{i,2} & I_3=\Sigma V_{i,3} W_{i,3} \end{bmatrix}$$

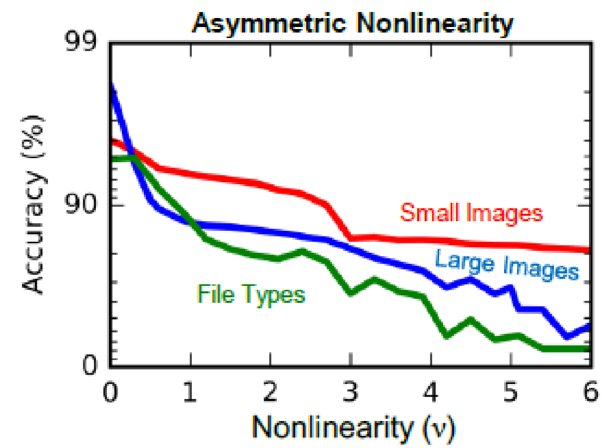
**Electrical**



# Challenges for adaptive analog accelerators

- Emerging ReRAM : far from ideal , floating-point ‘weights’
- Several key problems:
  - Limited resolution
  - Read and write noise
  - Device stochasticity
  - Device non-linearity
  - Device asymmetry
- Preliminary analysis: most severe impact from *asymmetric non-linearity*
- How can we get around this??
  - A) Increase bio-realism of learning accelerators
  - B) Focus on implementation of pre-trained networks, and use on-chip fine-tuning
    - Seeking natural computing: efficient combination of physical properties and algorithms



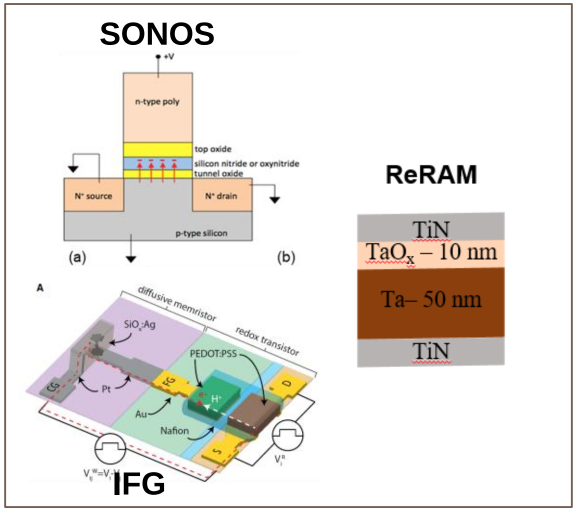




**Agarwal et al, IJCNN 2016**

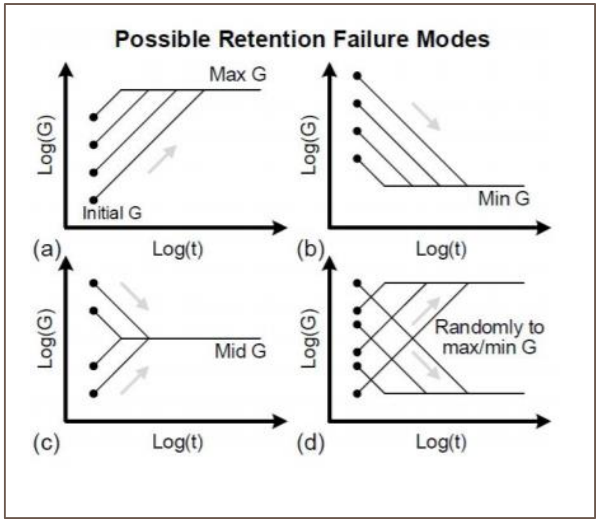
# Major opportunity:

## Emerging devices to implement neural network inference

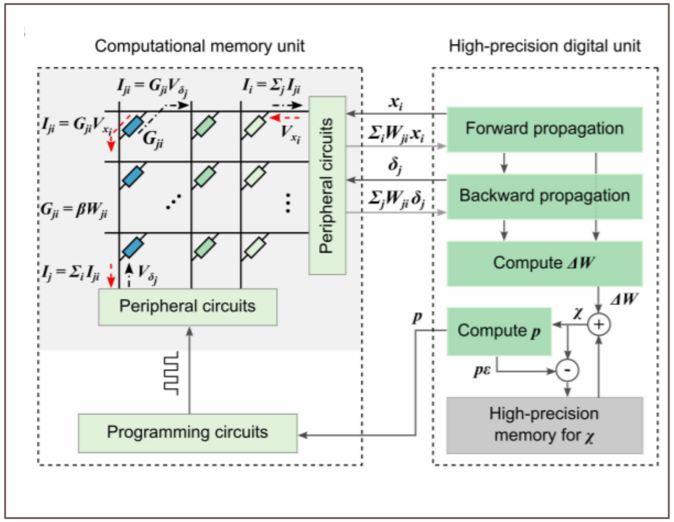
- Focus implementation around highly analog devices with linearity for updating/fine-tuning , if still needed
- Remaining serious issues:
  - Physical limits exist on minimal cycle-to-cycle noise (combination of generic [thermal/Johnson-Nyquist] and device specific [RTN])
  - Retention failure and drift in floating-gate, charge-trapping and ReRAM are real concerns
- Possibility to do mixed-computing using low-precision devices and high precision CMOS - > we explore limits of this using <u>highly analog weights</u>




Source: Fuller , Agarwal, et al, IEEE/Science



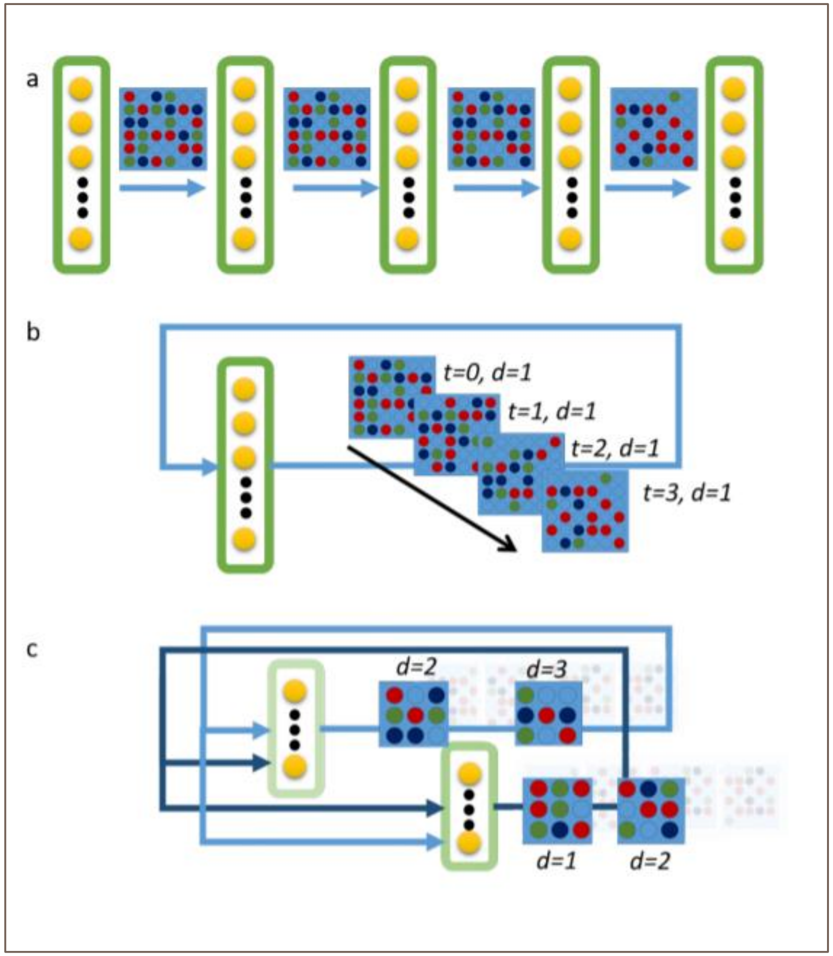

Source: Sun, et al , IEEE



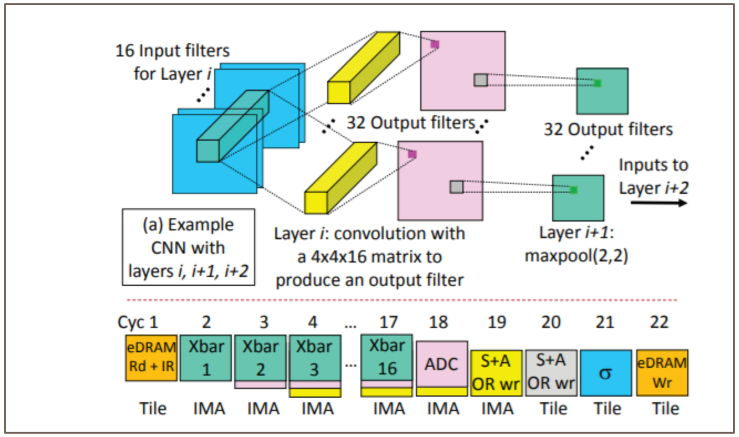

Source: Nandakumar et al, Frontiers

# NVM inference systems- overemphasis on CNNs

- Kernel-wise multiplication can result in massive crossbar requirements
  - Issues with energy and parasitics in large crossbars
  - ISAAC design: 40mW + /tile, 20W for chip.
    - 10-50x what we need for true low power computation (<1pJ per MAC)
- Massive opportunity for efficient synapse and neuron activation multiplexing - "*Mosaics*" framework
- *We focus on time-multiplexed activations*



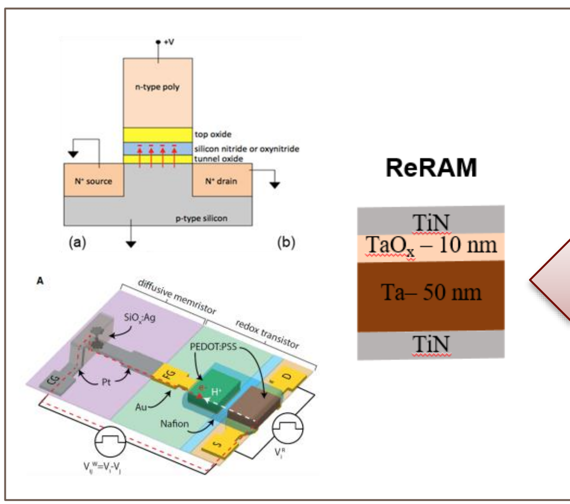

*Source: Bennett .. Aimone, et al*

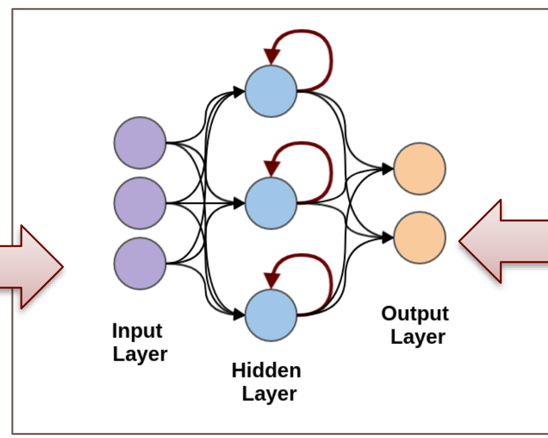


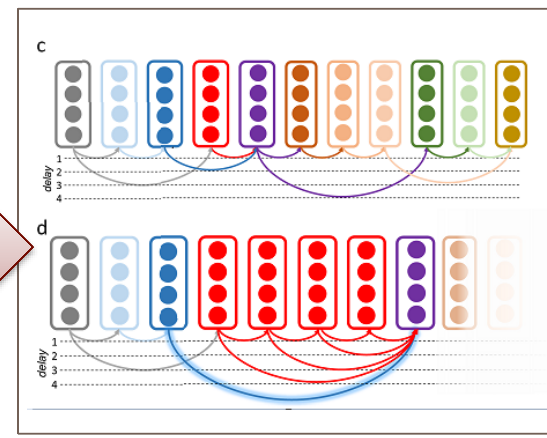

*Source: Shafiee et al, ISCA 2016*

# Outline

- **Problem Statement :**
  - **Interest in emerging memory for efficient inference engines**
- **Early results on recurrent neural networks**
- **Future steps and summary**

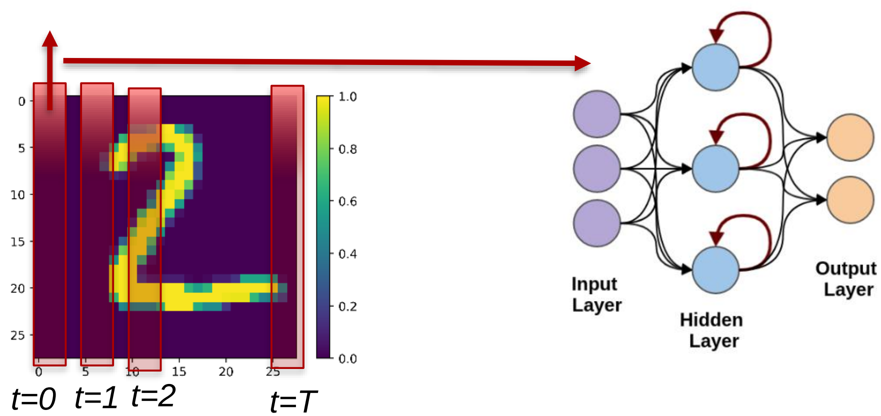


# Machine learning tasks

- In increasing order of difficulty:
  - MNIST: small images
    - 60k training, 10k test
    - MLP typical result: 96%+
    - CNN typical result: 98%+
  - F-MNIST: small images
    - 60k training, 10k test
    - MLP typical result: 83%+
    - CNN typical result: 91%+
- Presentation style for recurrent networks
  - Standard image presentation is subdivided into pixel-wise partitions that correspond to number of time steps, *T*
  - *T* must therefore be a natural divisor of *Num_pixels*

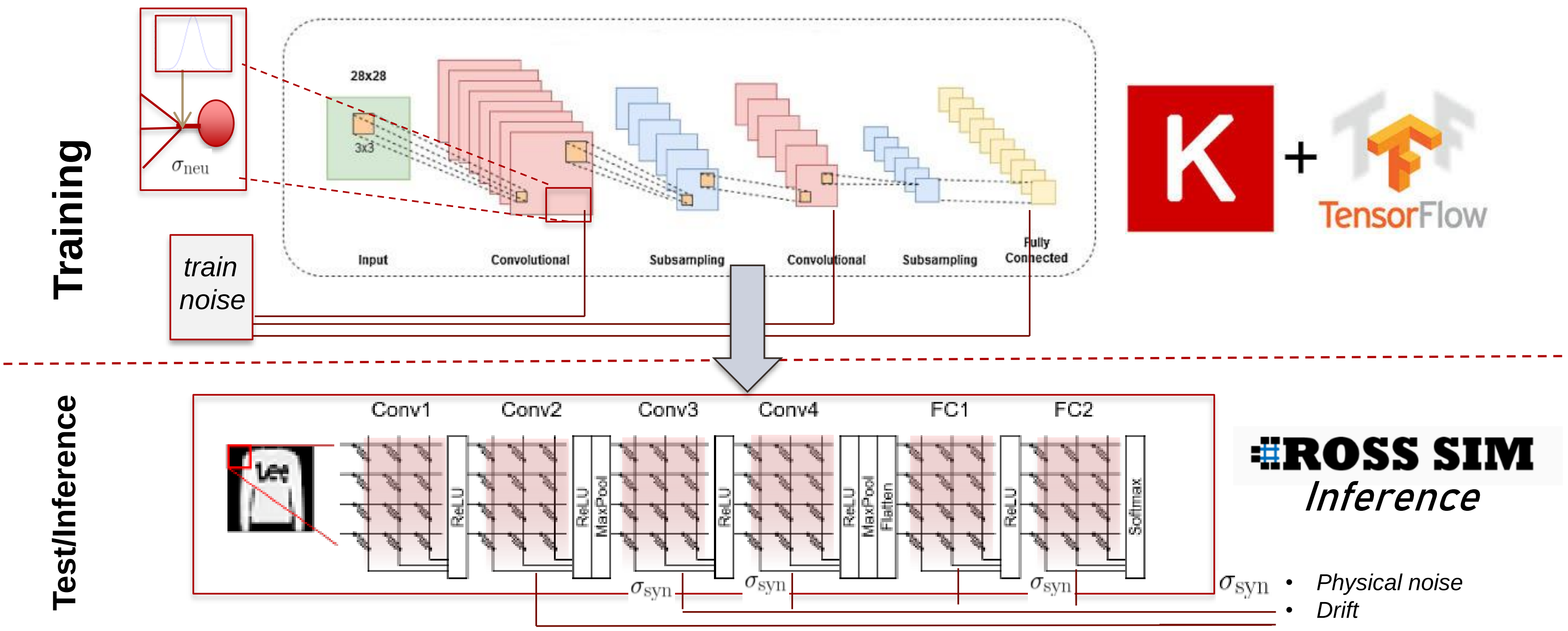



**MNIST Task**

**Fashion-MNIST Task**

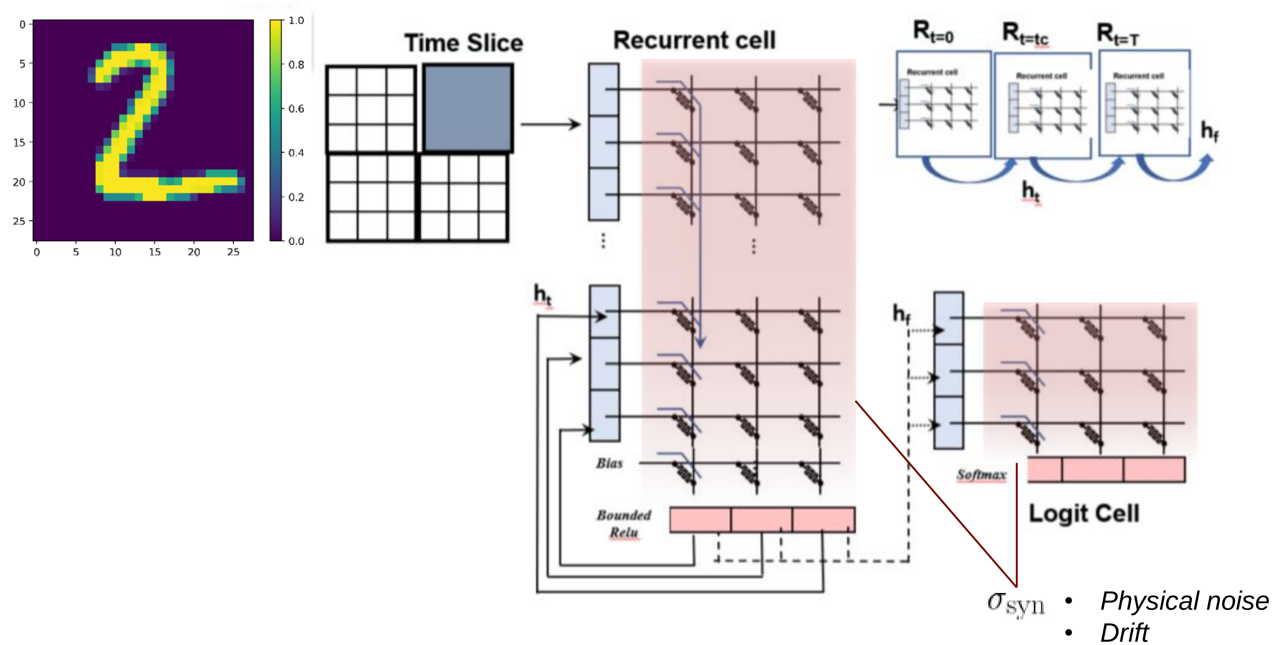
# Methodology I

- Networks were trained upfront using Keras/Tensorflow 2.2
  - MNIST, Fashion-MNIST
- Neural networks were pre-trained with, and without, a gaussian injected regularization term applied at pre-activation of neurons
  - Applied to activations in both convolutional filter crossbars and dense-layer crossbars
- During test-time, synaptic noise applied to all synapses (devices) in crossbars
- Equivalence between these effects given by: $\sigma_{\mathrm{neu}} = \sigma_{\mathrm{syn}} \left(W_{\mathrm{max}} - W_{\mathrm{min}}\right) \sqrt{n}\, \gamma_{\mathrm{act}}$



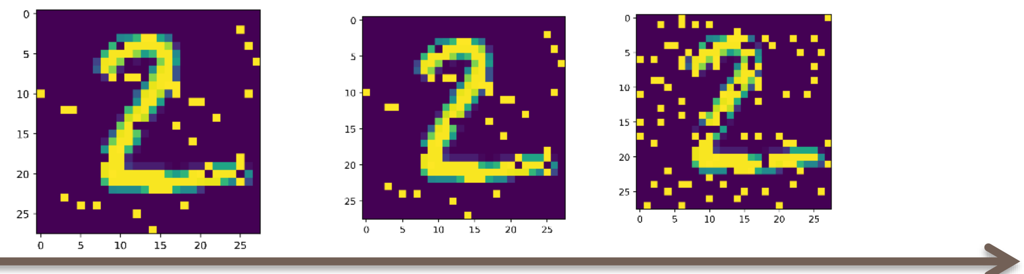
# Methodology II

- We propose a novel design for recurrent neural networks exploiting natural time re-use in a dense NVM crossbar
  - Less peripheral overhead than complex software RNN schemes such as Gated Recurrent Unit (GRU) , Long Short Term Memory (LSTM)
    - Only Rectified Linear Units (ReLU)-- > less complex circuit than tanh() etc
- We consider both normal and noise injected cases
  - As in CNN case, Gaussian noise injected before the ReLU activation

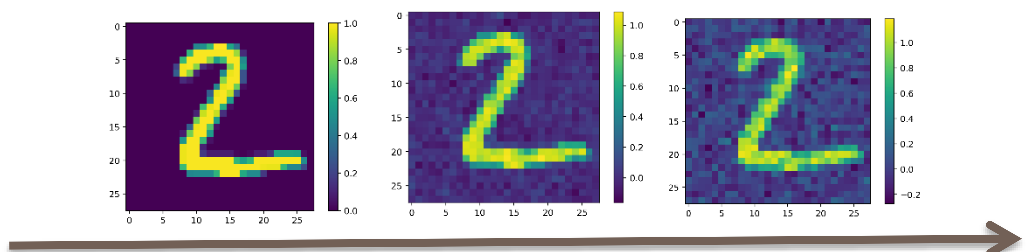


# Methodology III

- Test-set noise added on top of internal (synaptic ) noise
  - Gaussian:  test_set + test_set_noise(mean=0,std = sigma).
    - *Additive = more info loss*
  - Speckle: test_set*test_set_noise(mean=0,std=sigma)
    - *Scaled = less  info loss*
  - Salt and pepper noise (random_noise from sklearn); proportion of total pixels pushed to max/min vals .
    - Direct info loss, but *localized*



Increasing s&p noise

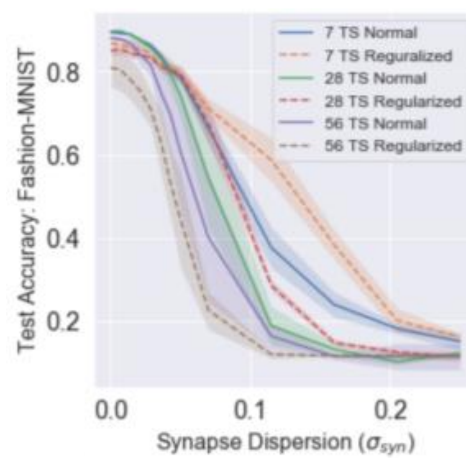


Increasing gaussian noise

# Result I

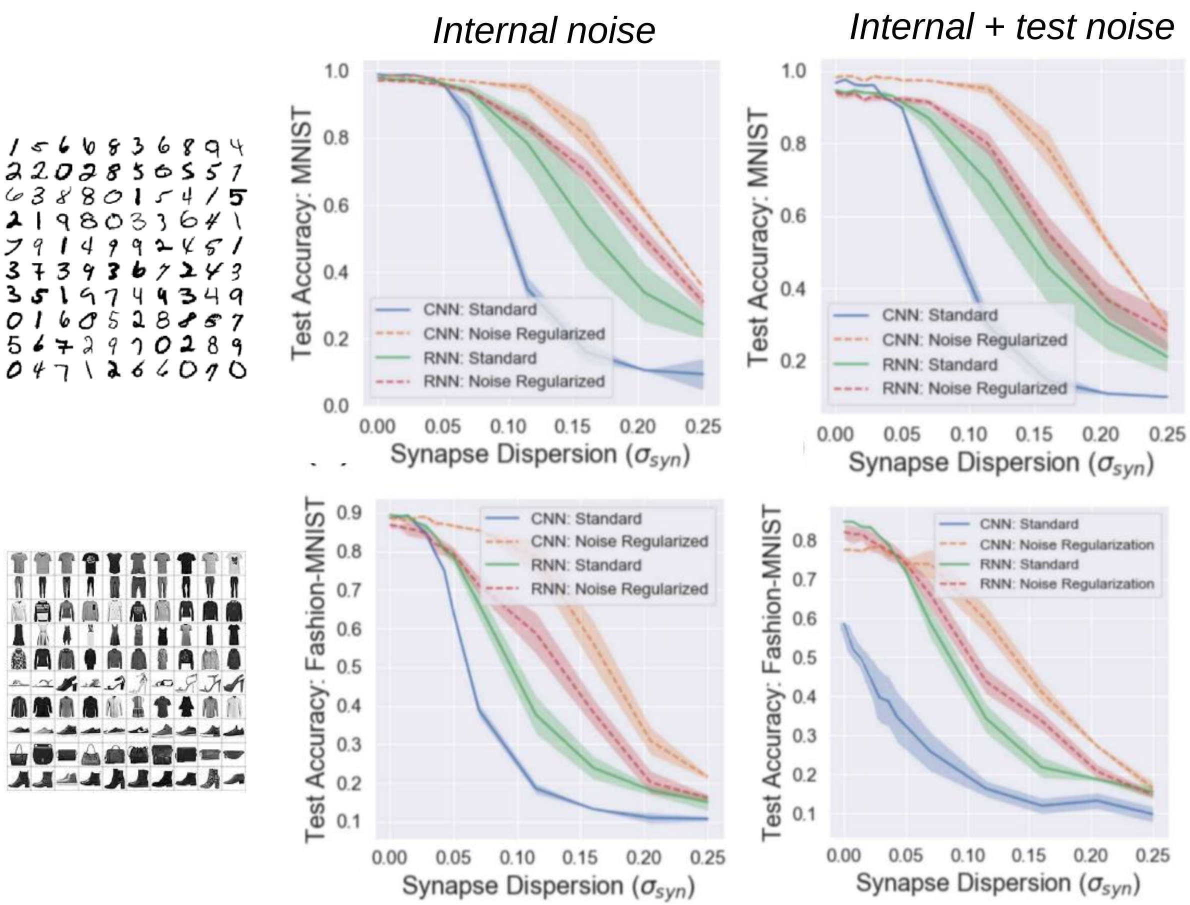
- Given training optimization (best optimizer and learning rates chosen for each system),
  - CNN systems , deployed with realistic (2.5%) internal noise, outperform RNN and MLP on both tasks
  - RNN systems, achieve near parity when test-set noise is applied, and <u>beat CNN system on harder task if effects combined</u>
- RNN systems perform best at a lower number of time-steps
  - Internal system noise is sub-linearly additive over temporal cycles (some cancellation exists)

| Architecture | Noise Scenario | | |
|---|---|---|---|
| | Internal ($\sigma_{syn}$*) | External ($\sigma_{te}$*) | Both Effects |
| MLP- MNIST | 96.8% | 94.1% | 93.1% |
| RNN - MNIST | 97.4% | 95.1% | 94.9% |
| CNN-MNIST | 98.5% | 96.7% | 96.05% |
| MLP- f-MNIST | 82.2% | 69.91% | 62.35% |
| RNN - f-MNIST | 86.3% | $\boxed{84.22\%}$ | $\boxed{81.11\%}$ |
| CNN-f-MNIST* | 85.1% | $\boxed{57.91\%}$ | $\boxed{42.35\%}$ |



# Result II

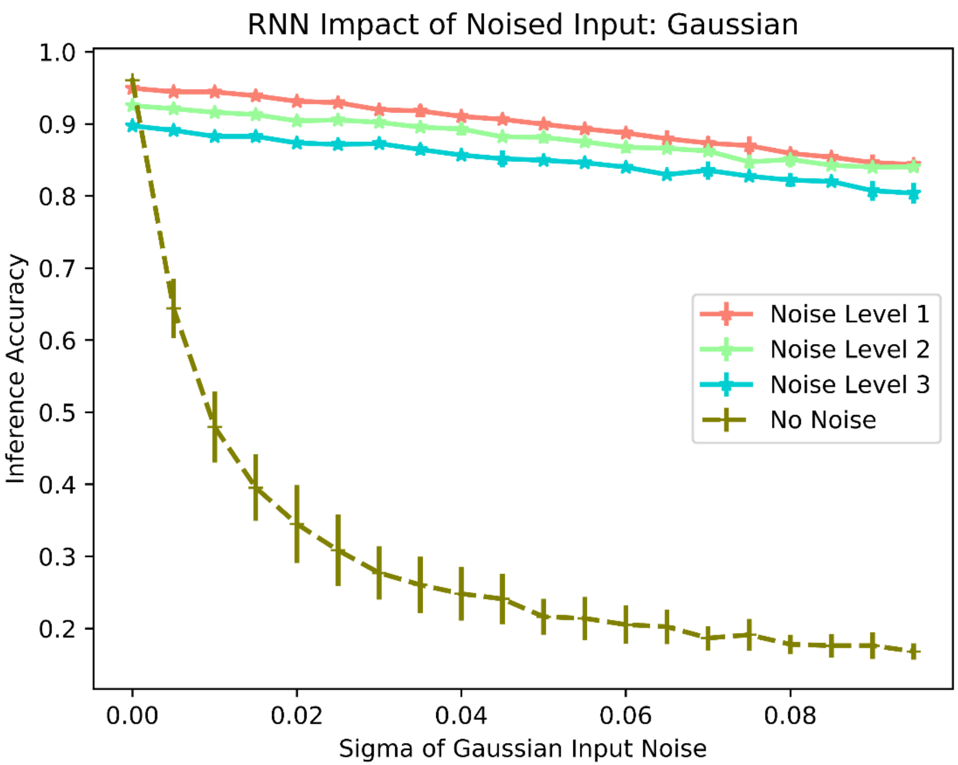
- Broader sweeps conducted on both effects
  - Regularization is useful in both deployed CNN, RNN systems
  - On easier task (MNIST), RNN does not show much benefit , but shines on fMNIST - > *CNN like results with less complexity*

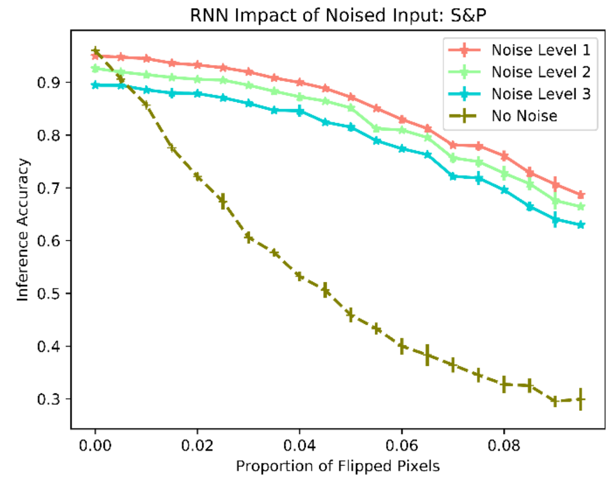


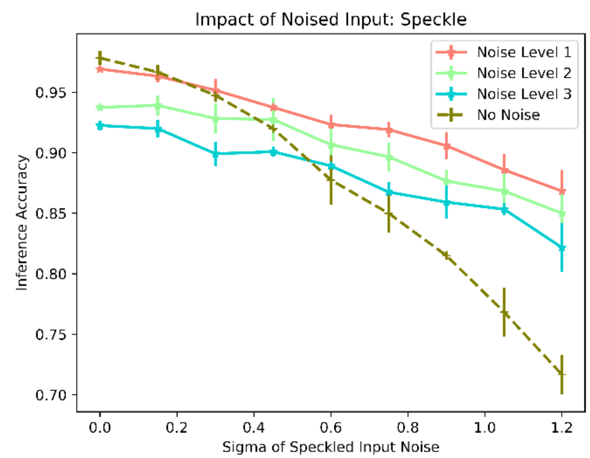
# Results III

Sandia National Laboratories

- Broader sweeps of RNN networks to test set noise were also conducted
  - Without regularization, the NVM optimized RNN collapses in performance for gaussian & s&p cases (most info lost).
  - The resilience provided by small injected noise (level 1= 0.1) in Speckle , Gaussian is impressive !!

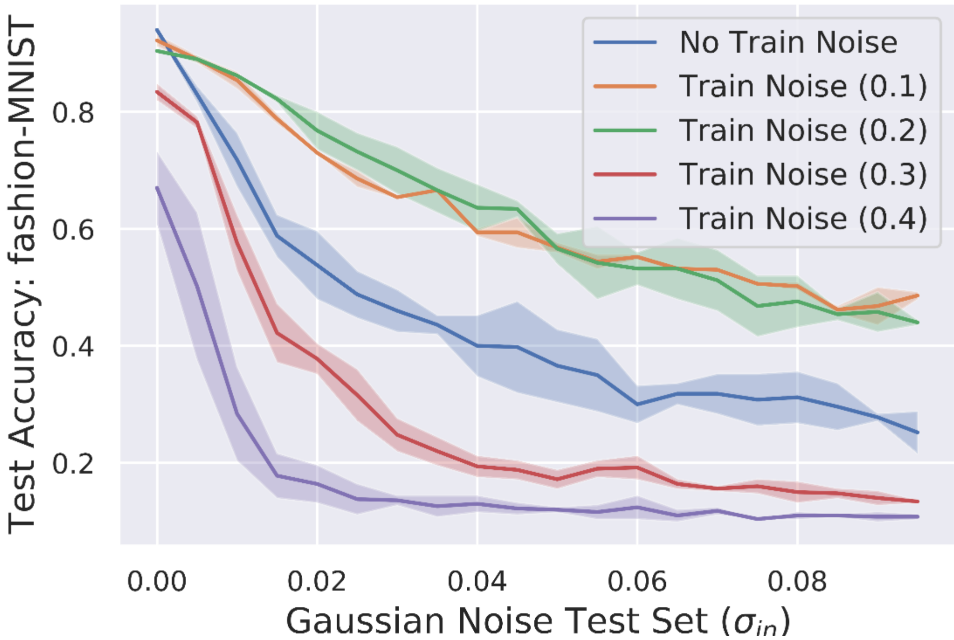


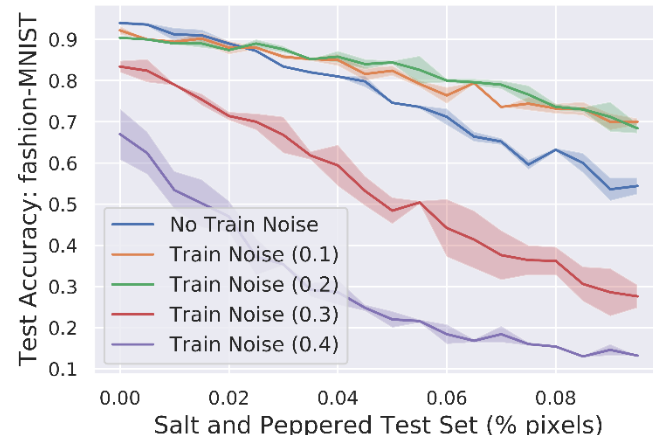



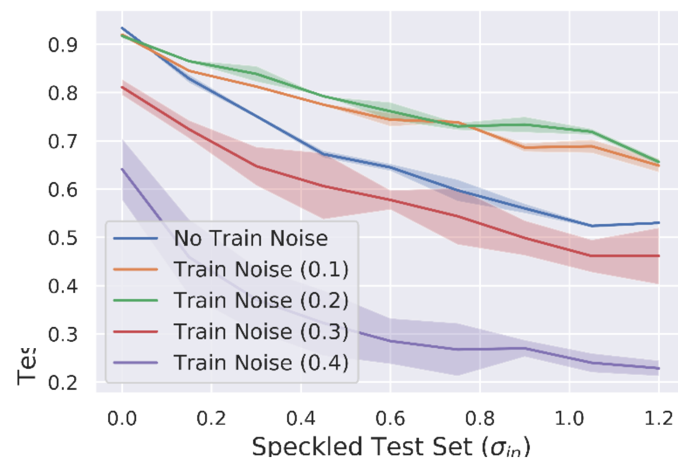



# Results IV

- Same sweeps conducted on trained small CNN networks
  - Appropriate levels of noise extend usable margin of the networks in adversarial/noisy environments
  - Only fMNIST results shown but results nearly equivalent for mNIST



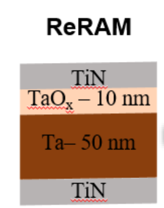



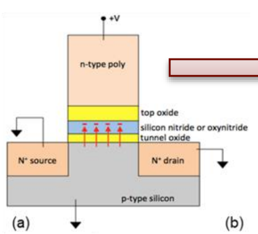



# Training energy estimates



- Simple python benchmarking script used to estimate energy estimates of online NVM learning for considered systems
  - Dominated by VMM (crossbar charging) and neuron activation
  - Maxpool , softmax operations are negligible
- While all systems use ReLU activations , due to time multiplexing, RNN systems benefit expend ~15-40x less energy than CNN
  - MLP systems are still least energy expensive overall, but suffer accuracy penalty

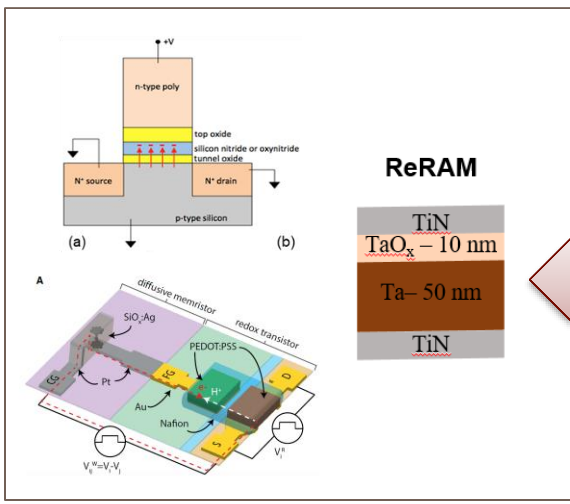


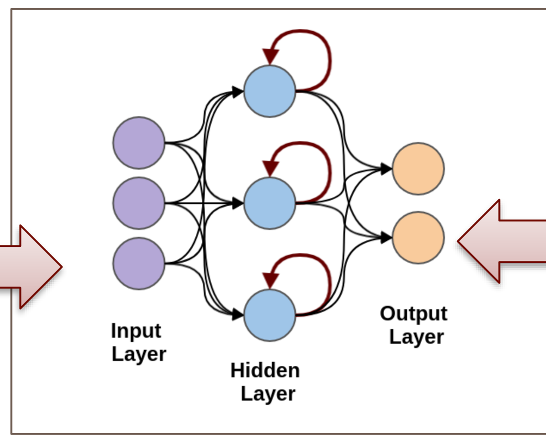



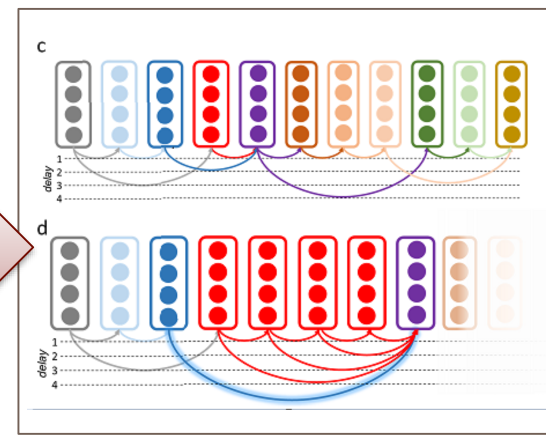

| Noise Mode | Synapse Type | | |
|---|---|---|---|
| | Total Energy/Op | VMM Op | Neuron Activation Op |
| MLP ReRAM* | 4.24 nJ | 4.22nJ | 15pJ |
| RNN ReRAM* † | $\boxed{35.6\text{nJ}}$ | 35.5nJ | 66pJ |
| CNN ReRAM* | $\boxed{480\ \text{nJ}}$ | 479 nJ | 358 pJ |
| MLP SONOS* | 6.04 nJ | 6.02nJ | 15pJ |
| RNN SONOS* † | $\boxed{42.7\text{nJ}}$ | 42.7nJ | 66pJ |
| CNN SONOS* | $\boxed{2.084\ \mu J}$ | $2.084\mu J$ | 358 pJ |

# Outline

- **Problem Statement :**
  - **Interest in emerging memory for efficient inference engines**
- **Early results on recurrent neural networks**
- **Future steps and summary**



# Further analysis of RNN systems

- Vanishing gradient issues must be further investigated in stacked simple RNN-NVM blocks
  - At inference stage, the problem will be far less of a problem than in training
  - But, may limit ultimate application of the approaches to relatively simple tasks (LSTM/GRU better to capture short + long term correlations)
- Temporal skip connections are an additional method to explore for further regularization + better generalization
  - Has been explored in LSTM, but not vanilla RNNs yet
- Natural attraction basins of RNNs can be analytically shown to help explain ergodic behavior (especially to test-set noise)

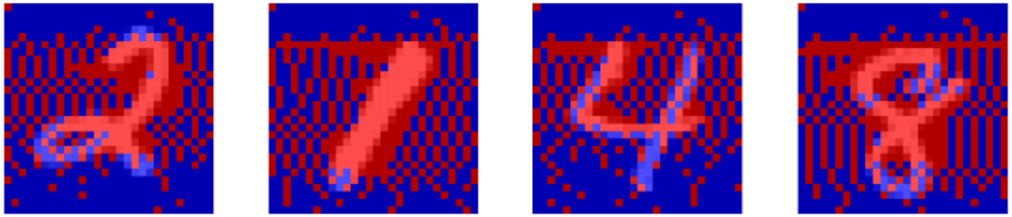
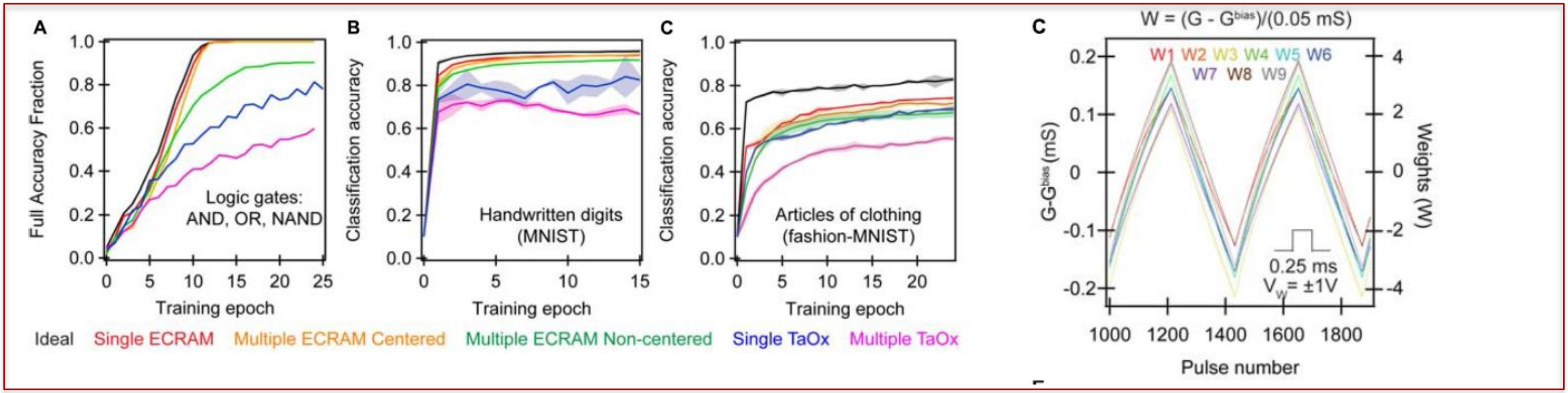


Figure 3: Sample usage examples for the Skip LSTM with $\lambda = 10^{-4}$ on the test set of MNIST. Red pixels are used, whereas blue ones are skipped.

*Source: Campos et al, ICLR 2018*

# Demonstrations of RNN learning

- Recently, 3x3 outer-product-learning was conducted with an array of ECRAM devices, and larger array is now being fabricated
- Ideal platform to implement RNN inference and learning
  - High device resistance -> low parasitics in demonstrator crossbar
  - Extreme analog capability
  - Low cycle to cycle noise (<0.5%) has been demonstrated -> good for large T




*Source: Li, Xiao, Bennett, Fuller, Marinella, Talin, et al, Frontiers , 2021 (Accepted)*

# Summary

## *Take away points*

- Time-multiplexing is a promising approach to implement energy efficient inference
- A new efficient RNN design has been proposed and simulated that:
  - Can approach or even exceed CNN performance given certain noise conditions
  - Exceeds a standard MLP in accuracy on standard ML tasks
  - Can pave the way to more energy efficient inference (10x or greater energy efficiency)
- Noise regularization at train time is a promising method to resist internal and external noise when deployed
  - Approach works on all considered neural networks, though most important/effective in CNN structures

## *Next Steps*

- Algorithmic explorations of sources and limits of natural RNN noise resilience
- Benchmarking of RNN scheme on more state-of-art tasks
- Demonstration of ideas in crossbar prototype(s)
- New version of CrossSim released: supporting inference + RNN

CROSS SIM

https://cross-sim.sandia.gov

# Thank you! Questions?

Contact me at cbennet@sandia.gov if you want to ask at a later time.