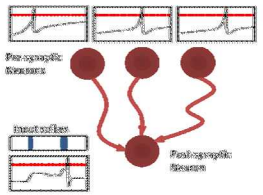
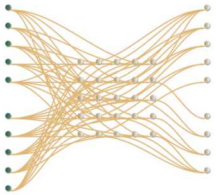
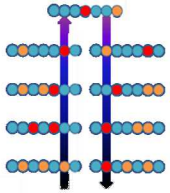
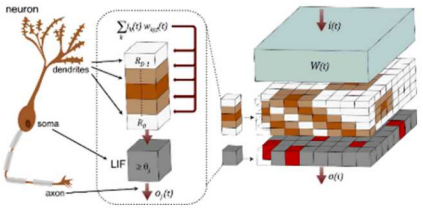
SAND2019-10111PE

Hardware-Aware Neural Architecture Search









PRESENTED BY

Sam Green

sgreen@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.

Motivation

Sandia background

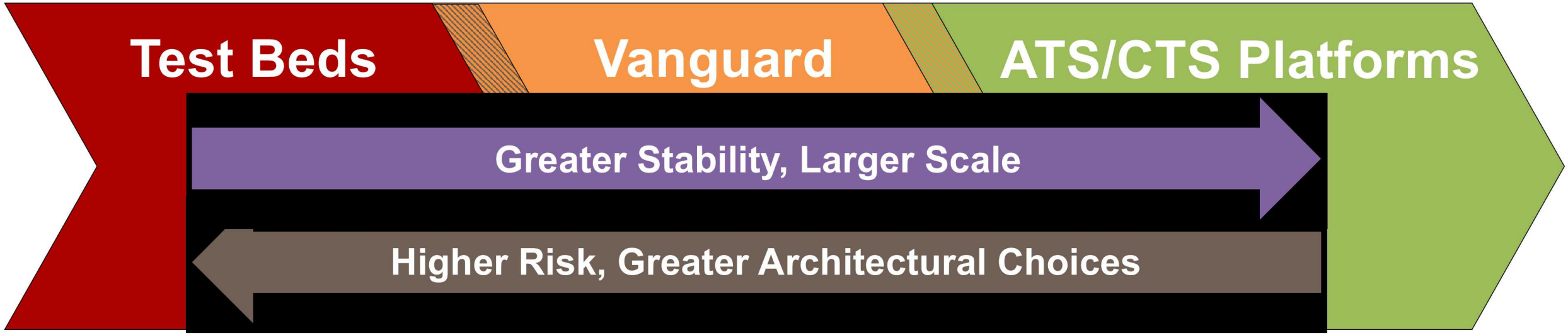
- 1940s: World War II's Manhattan Project
- 1945: Z Division of Los Alamos
- 1950s: Sandia National Lab created. Takes charge of engineering and manufacturing efforts
- 1960s: Pulsed Power program initiated
- 1962: Co-inventor of laminar flow clean room
- 1975: Fab for radiation-hardened CMOS
- 1996: ASCI Red, world's fastest computer at 1 teraflops
- 2006: Created Center for Integrated Nanotechnologies facility
- 2018: Astra supercomputer. Fastest ARM-based machine on TOP500

Vanguard Program: Advanced Architecture Prototype Systems

- Prove viability of advanced technologies for DOE integrated codes, at scale
- Expand the HPC ecosystem by developing emerging unproven technologies
  - Is it viable for future ATS/CTS platforms – Trinity & Sierra
  - Increase technology AND integrator choices
- Buy down risk and increase technology and vendor choices for future platforms
  - Ability to accept higher risk allows for more/faster technology advancement
  - Lowers/eliminates mission risk and significantly reduces investment
- Jointly address hardware and software challenges

- First prototype platform targeting ARM

Where Vanguard Fits



Test Beds

- Small testbeds (~10-100 nodes)
- Breadth of architectures
- **Brave users**

Vanguard

- Larger-scale experimental systems
- Focused efforts to mature new technologies
- Broader user-base
- **Demonstrate viability for production use**
- NNSA Tri-lab resource

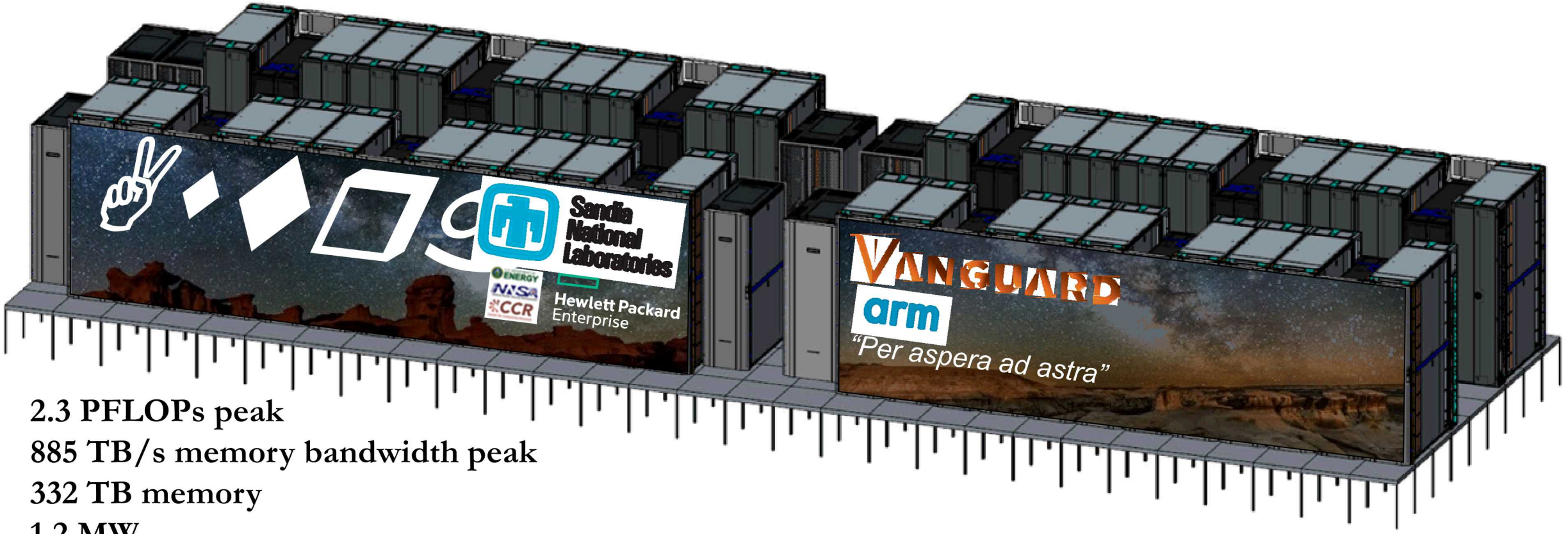
ATS/CTS Platforms

- Leadership-class systems (Petascale, Exascale, ...)
- Advanced technologies, sometimes first-of-kind
- Broad user-base
- **Production use**


ASTRA
"Per aspera ad astra"
RM SUPERCOMPUTER
VANGUARD

per aspera ad astra

through difficulties to the stars



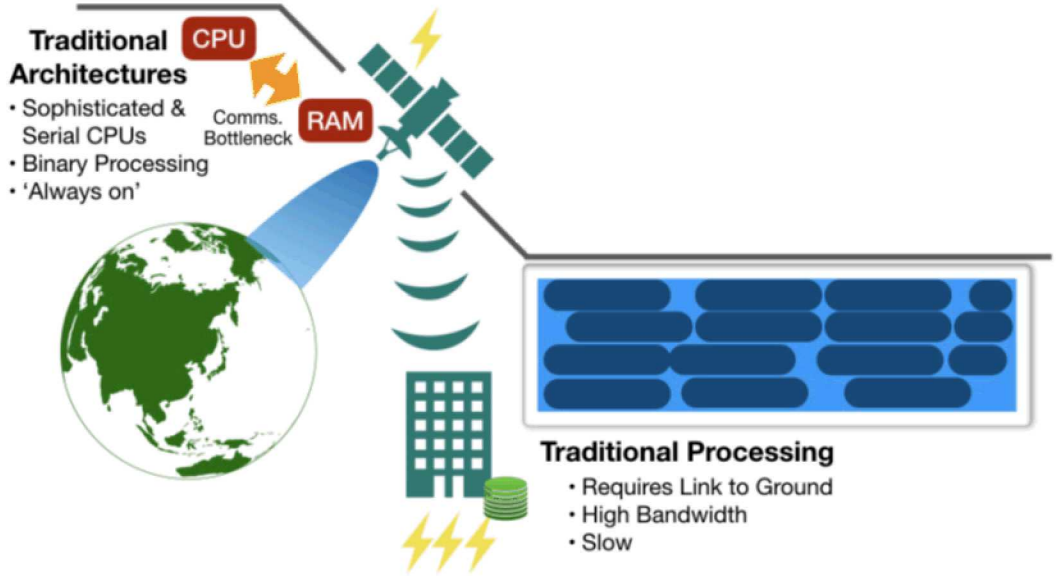
2.3 PFLOPs peak
885 TB/s memory bandwidth peak
332 TB memory
1.2 MW

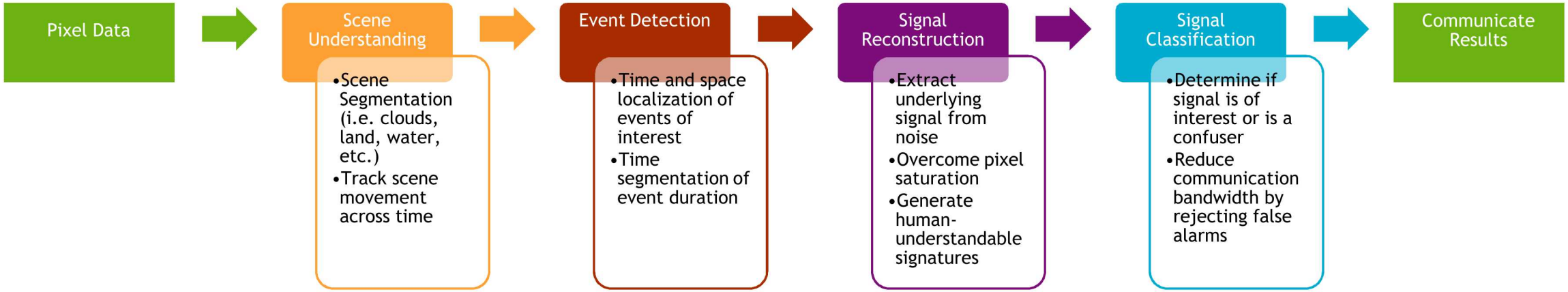
Demonstrate viability of ARM for U.S. DOE Supercomputing

Remote Sensing

Limitations to classic approaches to remote sensing

- Growth of sensor technologies outpacing communication bandwidth
- Limited onboard processing capability
- Rad hard design
- Need for alternative approaches





Why not just continue what we've always done?

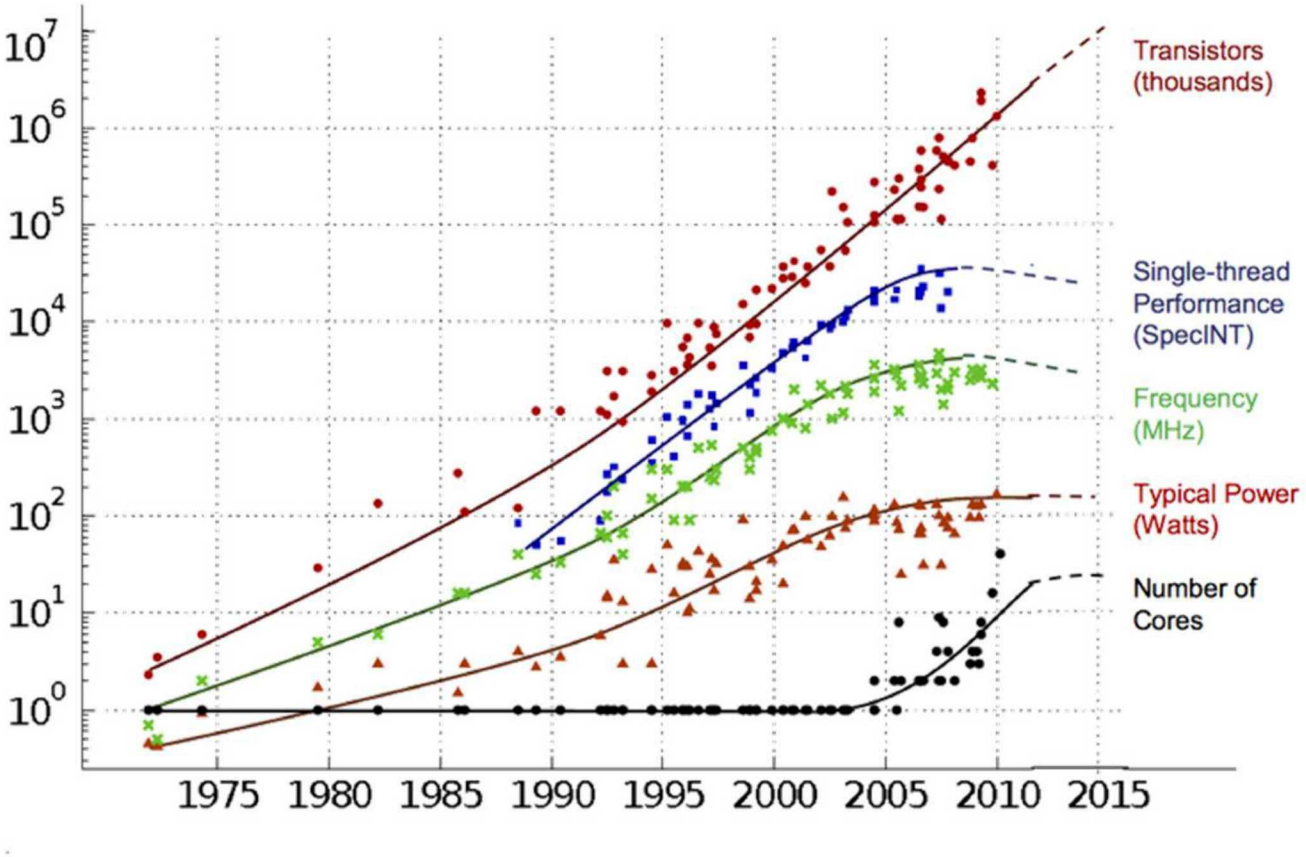
Dennard scaling

- As transistors get smaller, their power density remains constant

Unfortunately ended 10-15 years ago

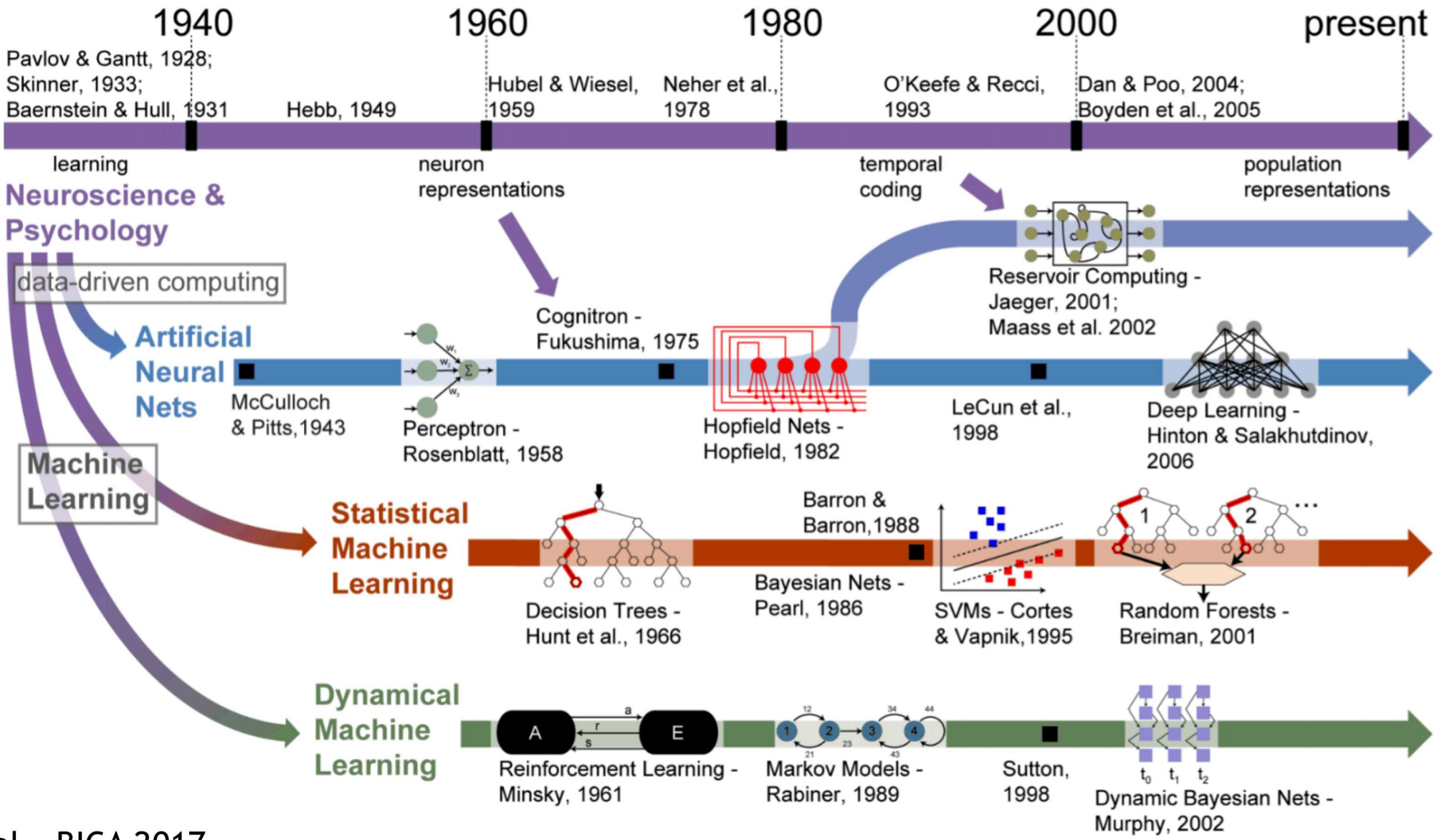
- Cannot run CPUs at faster speeds
- Emphasis on multi-core

Need for new paradigm of computing



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Neural-inspired computing

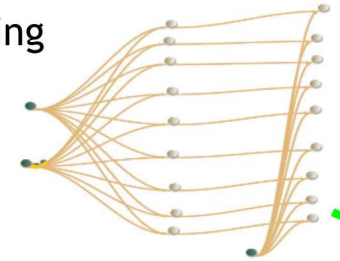


James et al., BICA 2017

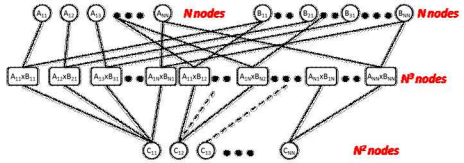
Breadth

Impacting Broad Areas of Computation

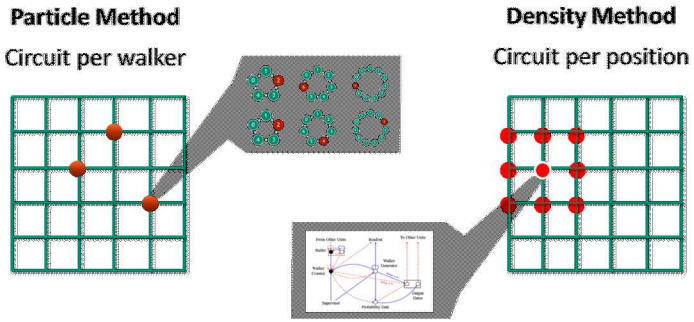
Pattern Matching



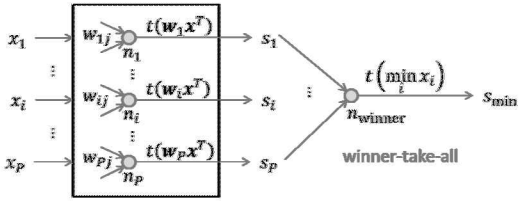
Linear Algebra

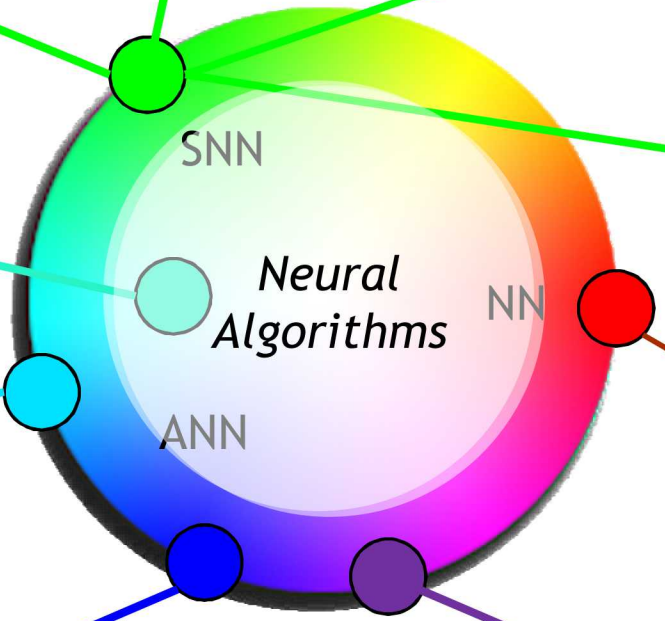


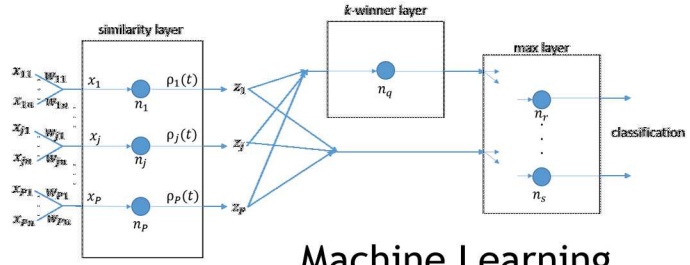
Scientific Computing



Optimizations

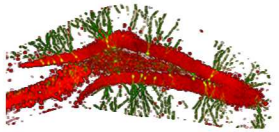




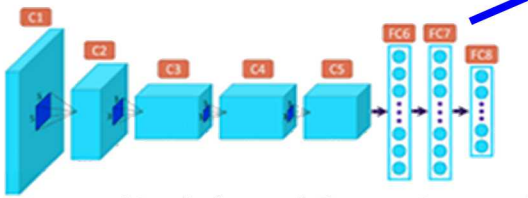


Machine Learning

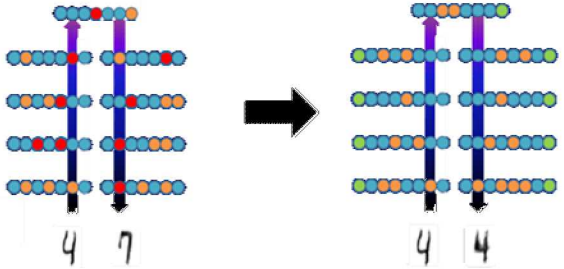
WHETSTONE



Intelligent Storage



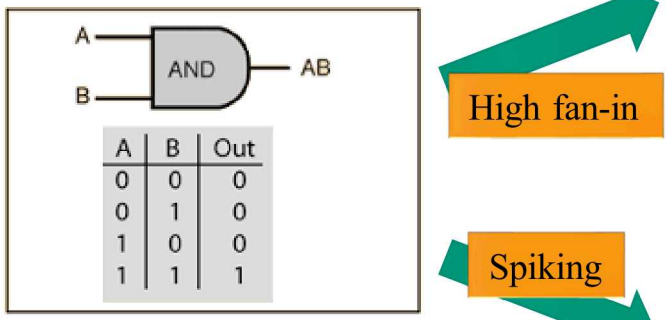
Context Modulated Deep Learning



Adaptive Deep Learning

Spiking neurons are a more powerful version of classic logic gates

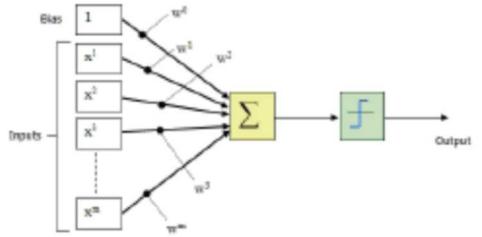
Spiking threshold gates provide high degree of parallelism at very low power



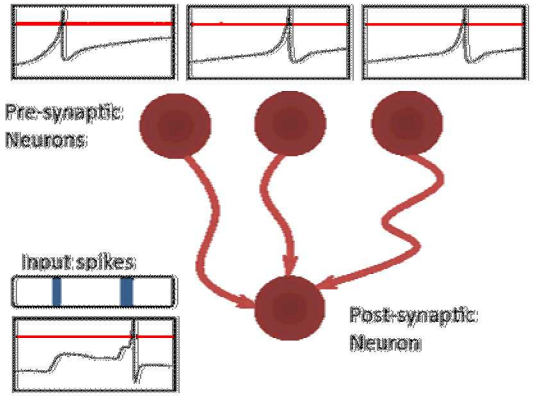
High fan-in

Spiking

Based on a simple McCullogh-Pitts model:



Outputs a 1 if and only if: $w_0 + \sum_{i>0} w_i x_i \geq 0$.



Compute more powerful logic functions

Incorporate time into logic

SNL has produced a number of spiking numerical algorithms

Cross-correlation
- Severa et al., *ICRC 2016*

SpikeSort, SpikeMin, SpikeMax, etc
- Verzi et al., *Neural Computation 2018*

SpikeOptimization
- Verzi et al., *IJCNN 2017*

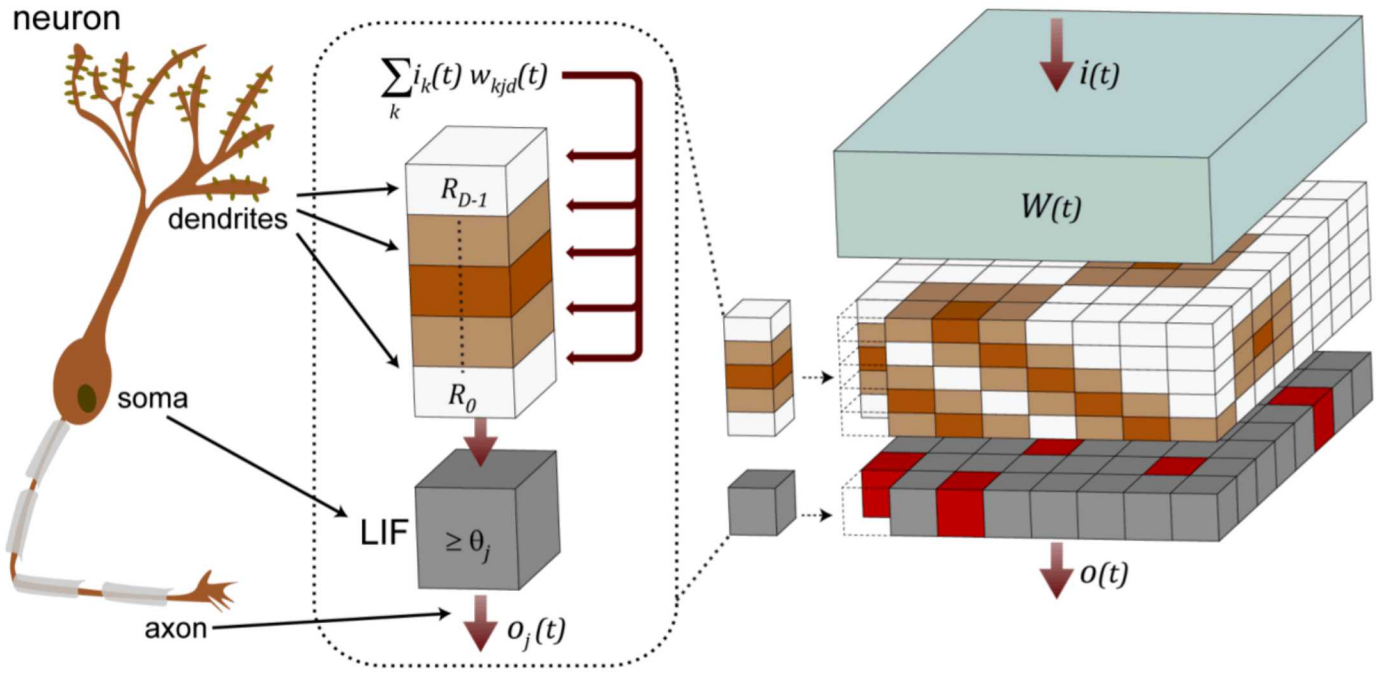
Sub-cubic (i.e., Strassen) constant depth matrix multiplication
- Parekh et al., *SPAA 2018*

Spiking Temporal Processing Unit (STPU)

Implemented on FPGU, STPU is composed of a set of leaky integrate and fire neurons.

Each neuron has an associated temporal buffer such that inputs can be mapped to a neuron with a time delay. W(t) is the neuronal encoding transformation which addresses connectivity, efficacy and temporal shift.

Mimics functionality of biological neurons.

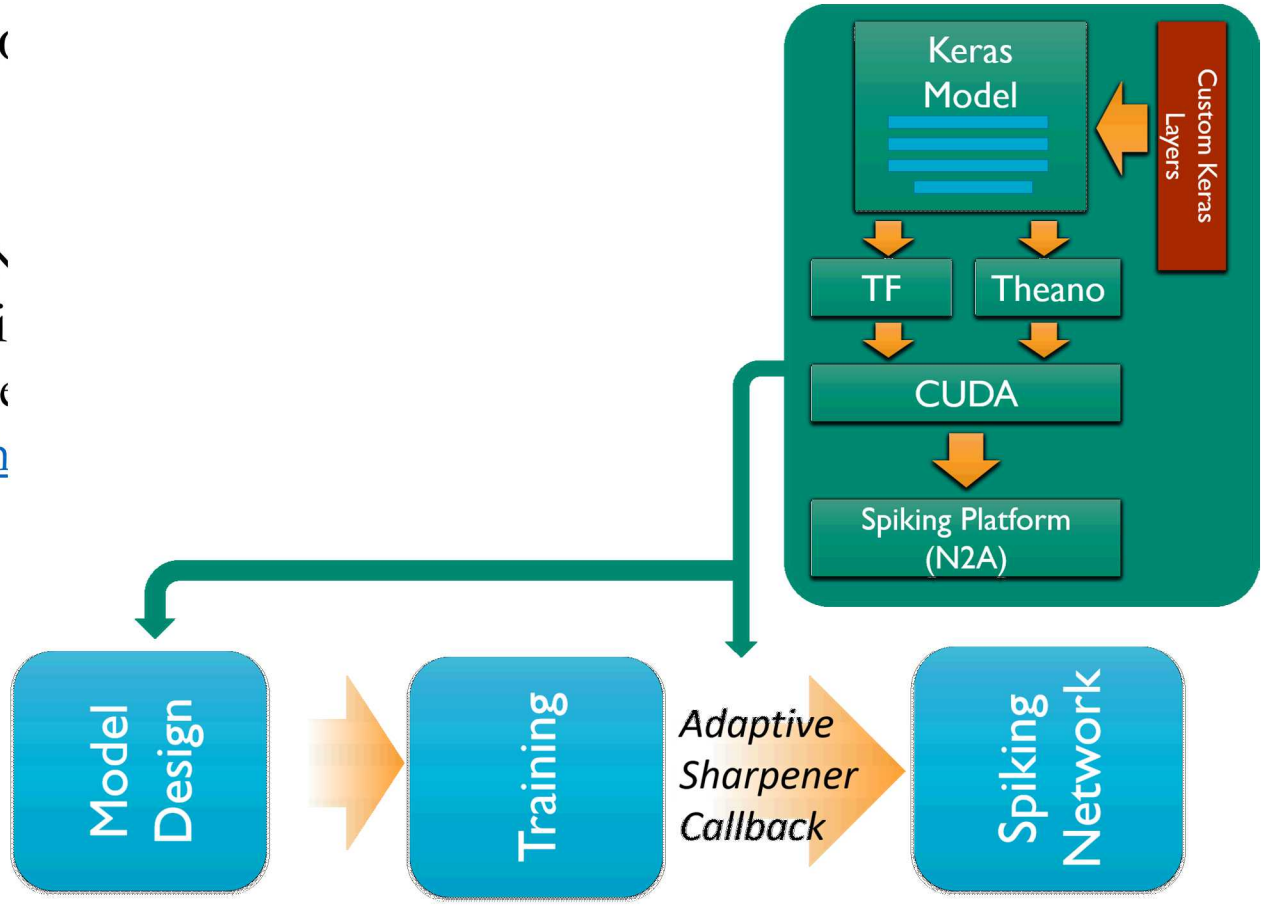


Whetstone

WHETSTONE

Whetstone provides a drop-in mechanism for tailoring a DNN to a spiking hardware platform (or other binary thresh platforms)

- Hardware platform agnostic.
- Compatible with a wide variety of DN
- No added time or complexity cost at i
- Simple neuron requirements: Integrat
- https://github.com/SNL-NERL/Wh



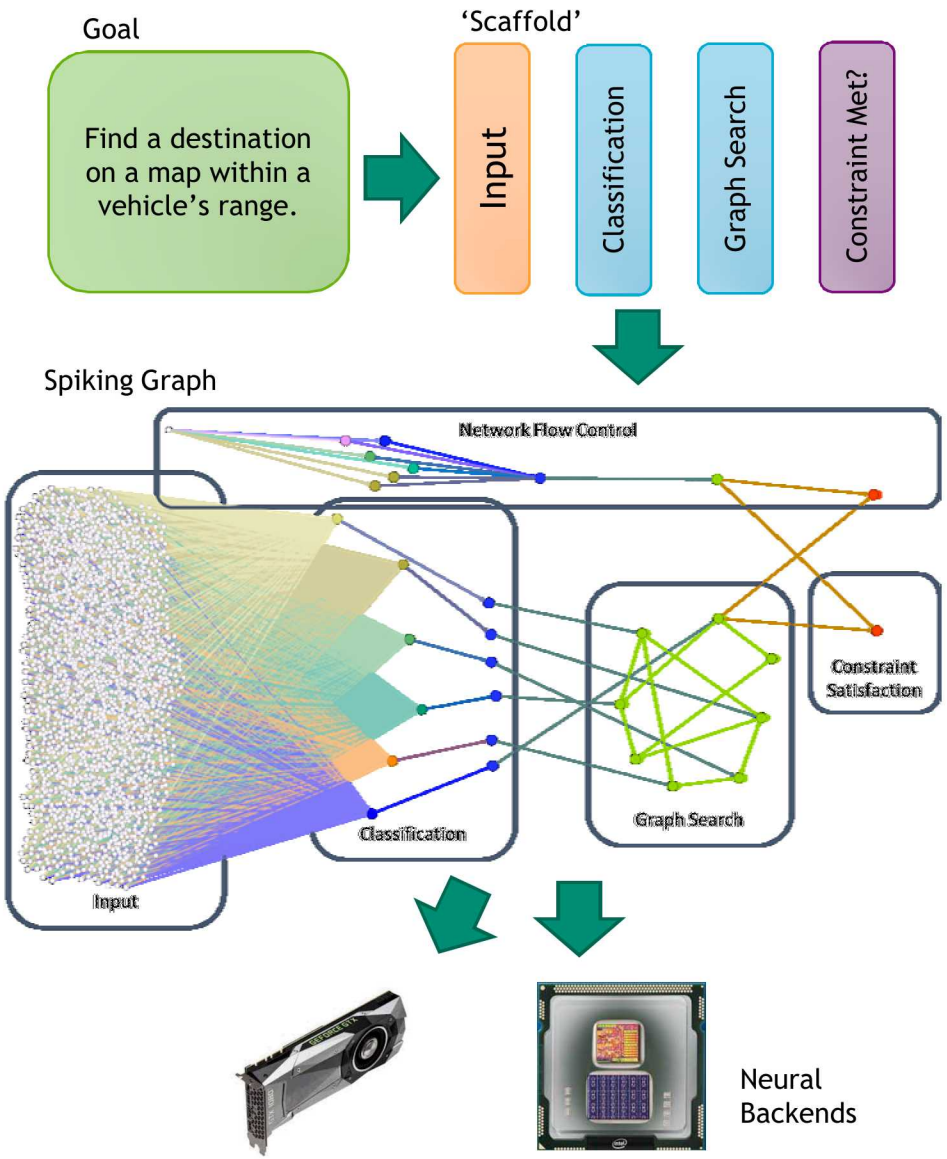
Fugu

Problem

- Neuromorphic platforms offer substantial (100x-1000x) performance improvements but are inaccessible
- Algorithms remain undeveloped; Multiple incompatible frameworks exist; Algorithms cannot interface with one another or hardware

Technical Approach

- Developing a framework for linking existing spiking neural networks and expanding to solve scientific computing problems
- Independent of the hardware that runs the neuron computation
- Collaboration: LANL, LLNL

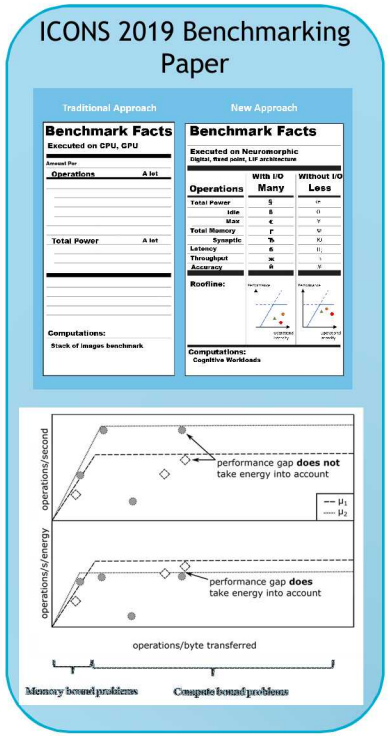


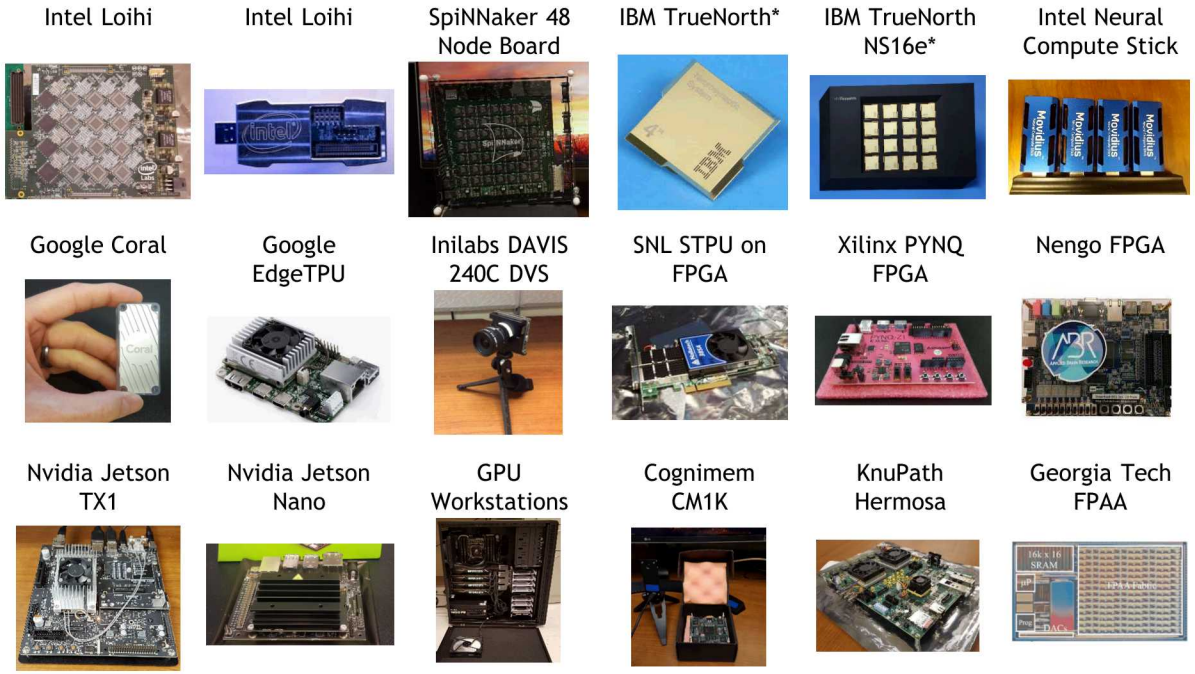
Large collection of neuromorphic hardware

Enables researchers to explore the boundaries of neural computation

Consists of a variety of neuromorphic hardware & neural algorithms providing a testbed facility for comparative benchmarking and new architecture exploration

Developing benchmarking methodologies





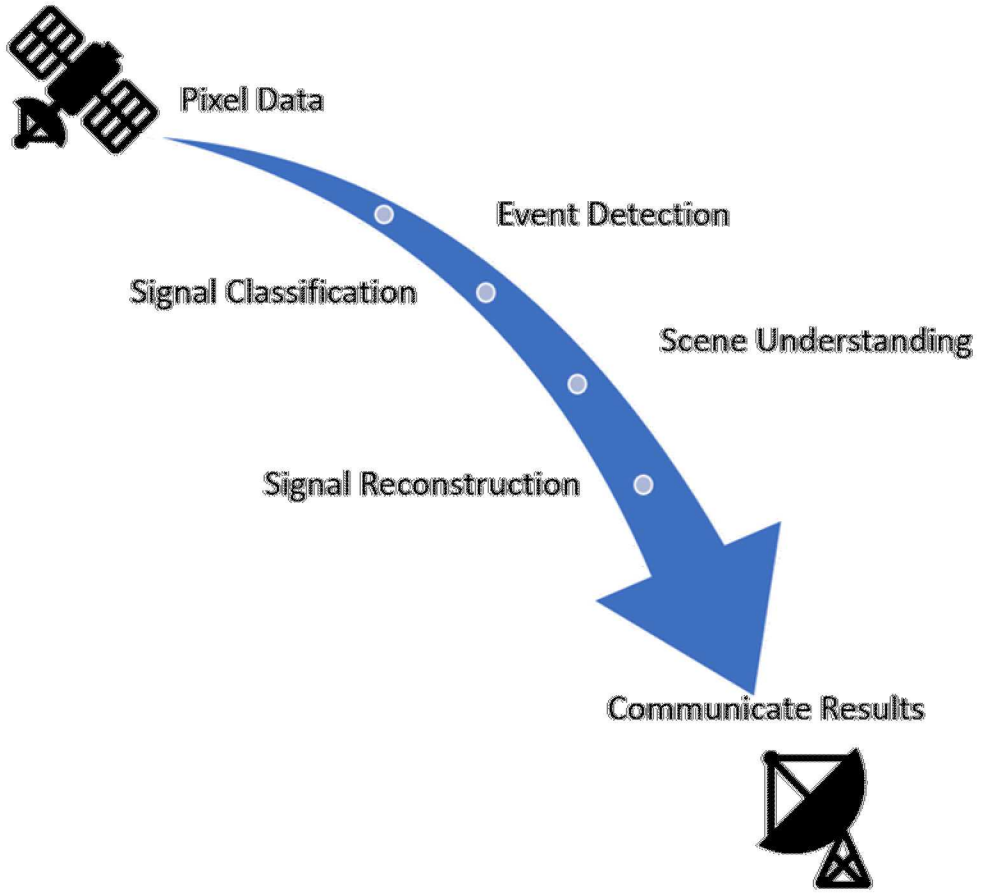
*Remote access

Case study: Neural Architecture Search for Remote Sensing

Remote Sensing

Neural approaches show promise:

- Signal Processing
- Signal Classification
- Signal Understanding
- Moving computation onto the satellite yields:
- Reduction in required bandwidth
- Improvements in response times
- Potential for autonomy



Size, Weight and Power (SWaP) is the constraint!

Neural Architecture Search

NAS methods seek to automate the search for neural architectures.

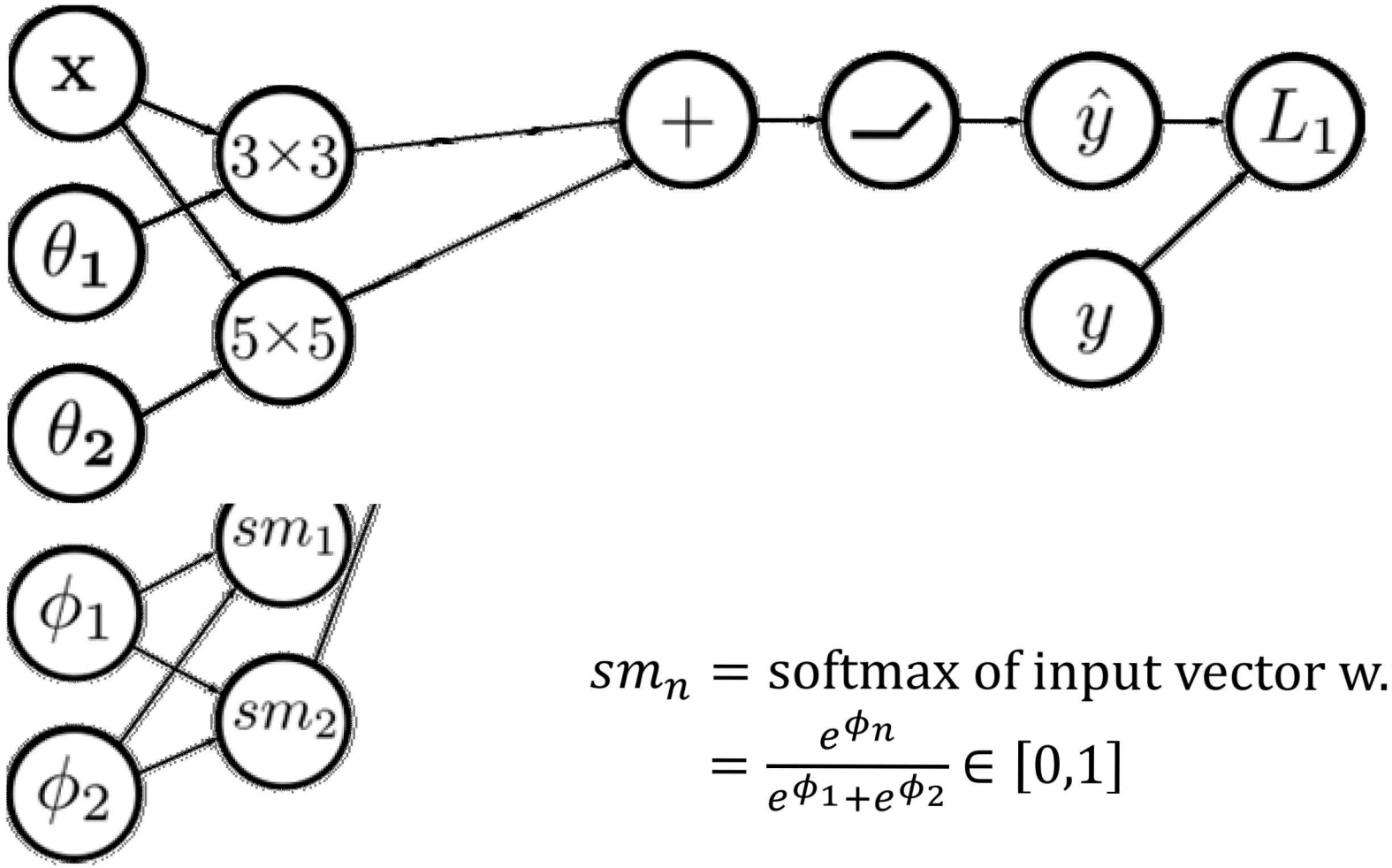
As a counter example, these architectures were found through grad-student descent:

LeNet-5	Inception
AlexNet	ResNet
VGG-16	DenseNet

Three primary NAS approaches:

- Reinforcement learning – first approach. Works well. Takes forever.
- Evolutionary strategy – Also works well and takes forever.
- Gradient – Works well and fast.

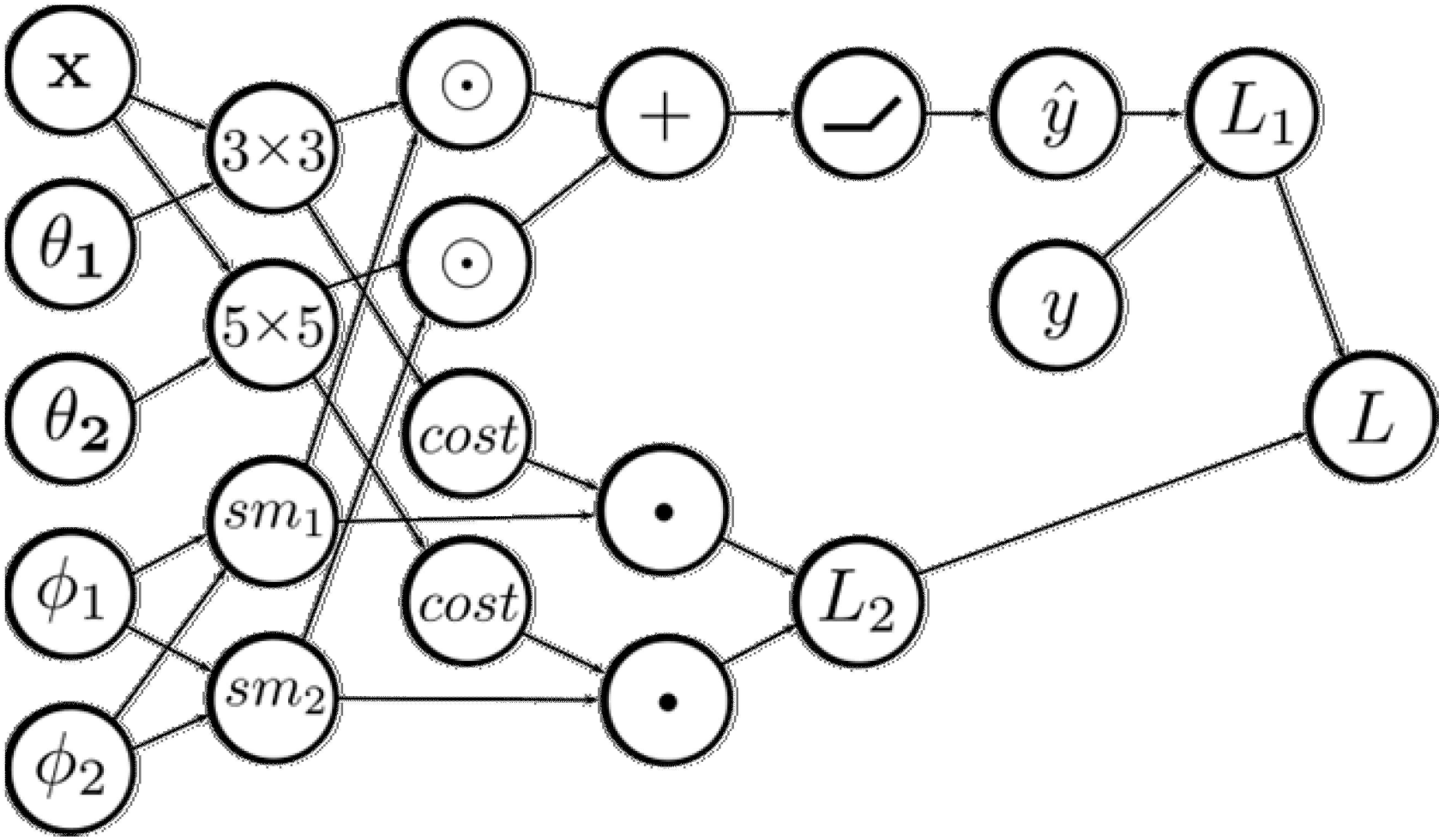
Gradient-based NAS



sm_n = softmax of input vector w.r.t. n

$$= \frac{e^{\phi_n}}{e^{\phi_1} + e^{\phi_2}} \in [0,1]$$

Hardware-aware gradient-based NAS



Hardware-aware PDARTS

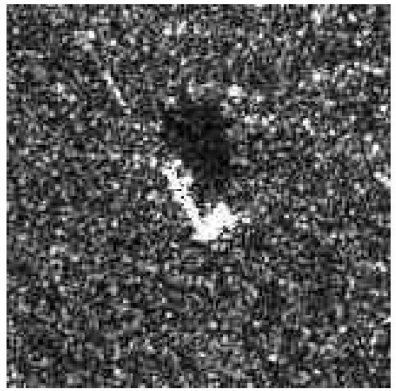
Current work based on fork of Progressive Differential ArchiTecture Search (PDARTS).

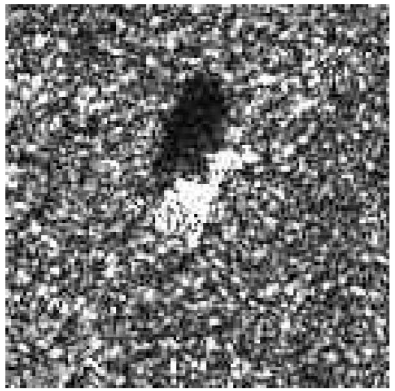
- Progressively grow deeper networks which conform to loss terms.

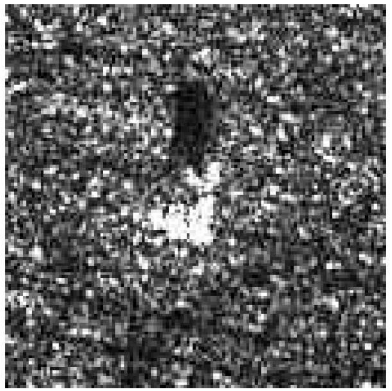
Augmented with hardware-based cost loss term.

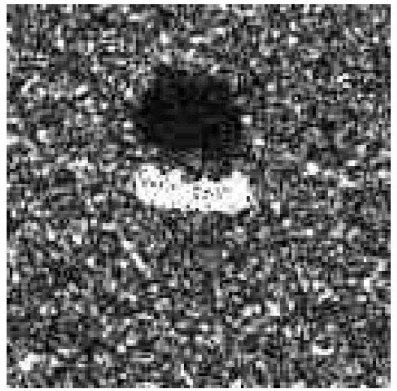
- Using number of parameters in operation as a proxy for cost.
- E.g. C(5x5 conv) > C(3x3 conv).
- Hardware loss biases architecture search toward lightweight models.

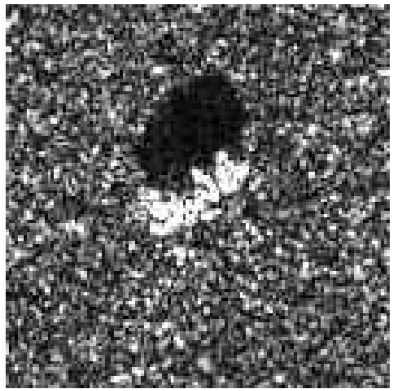
MSTAR dataset

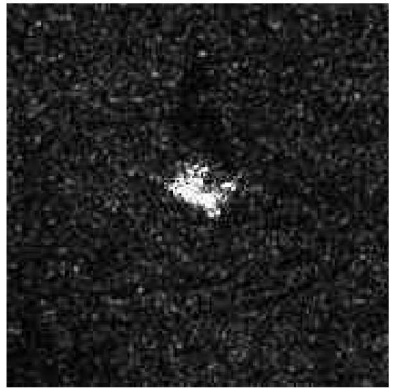

2S1

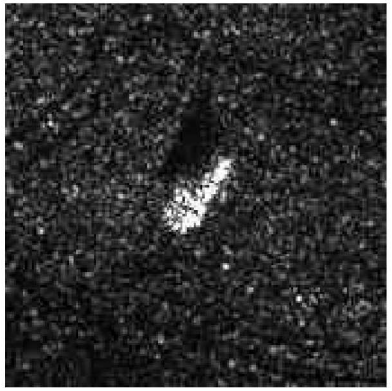

BMP2

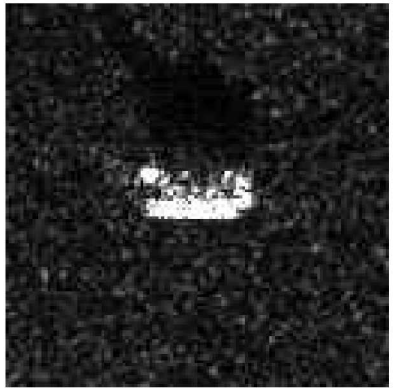

BRDM_2

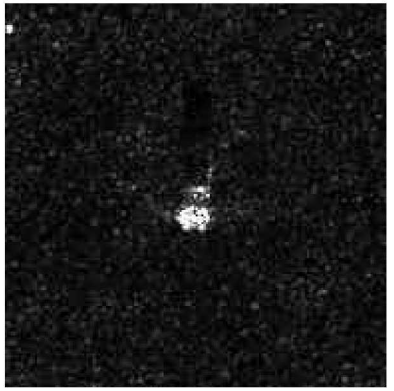

BTR_60

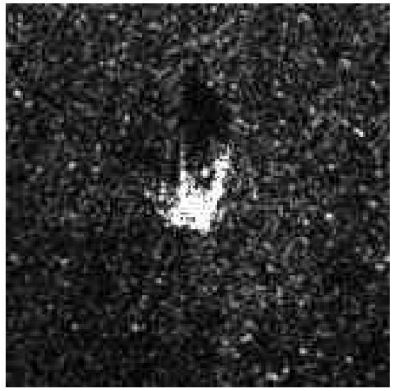

BTR70


D7


T62


T72

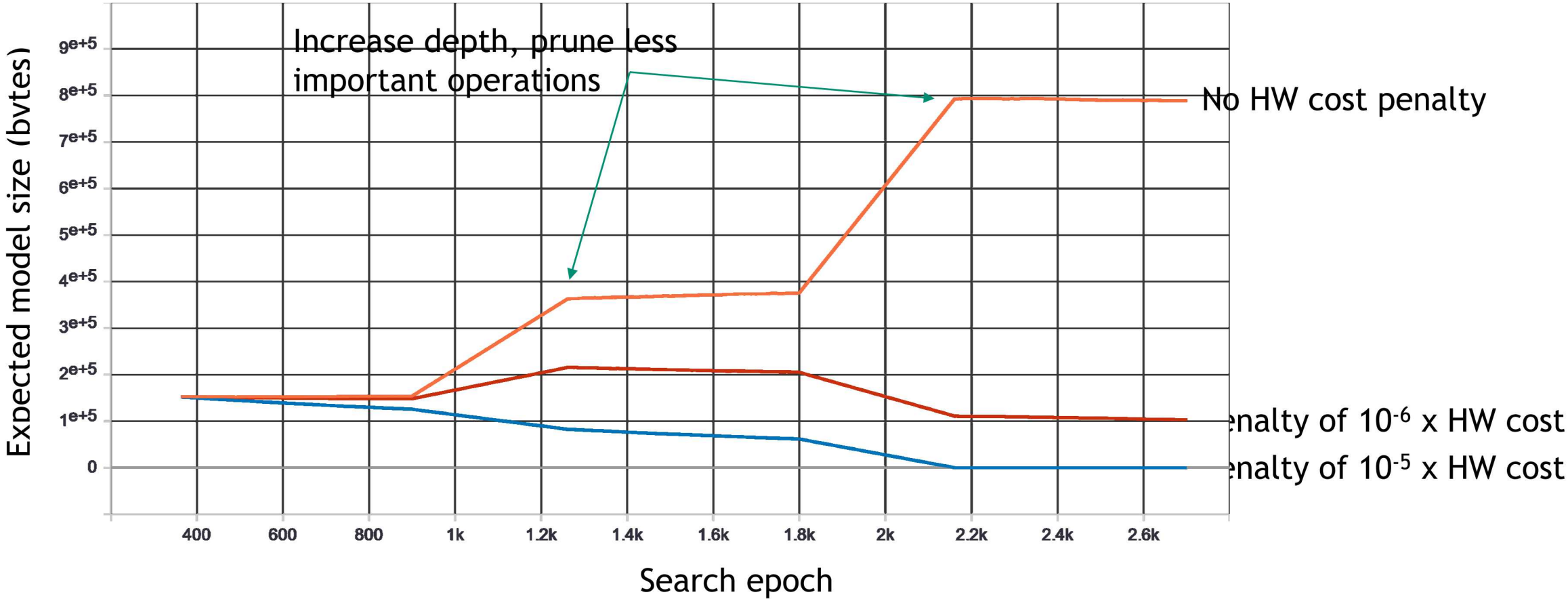

ZIL131


ZSU_23_4

HAPDARTS – effect of hardware cost

Three trials:

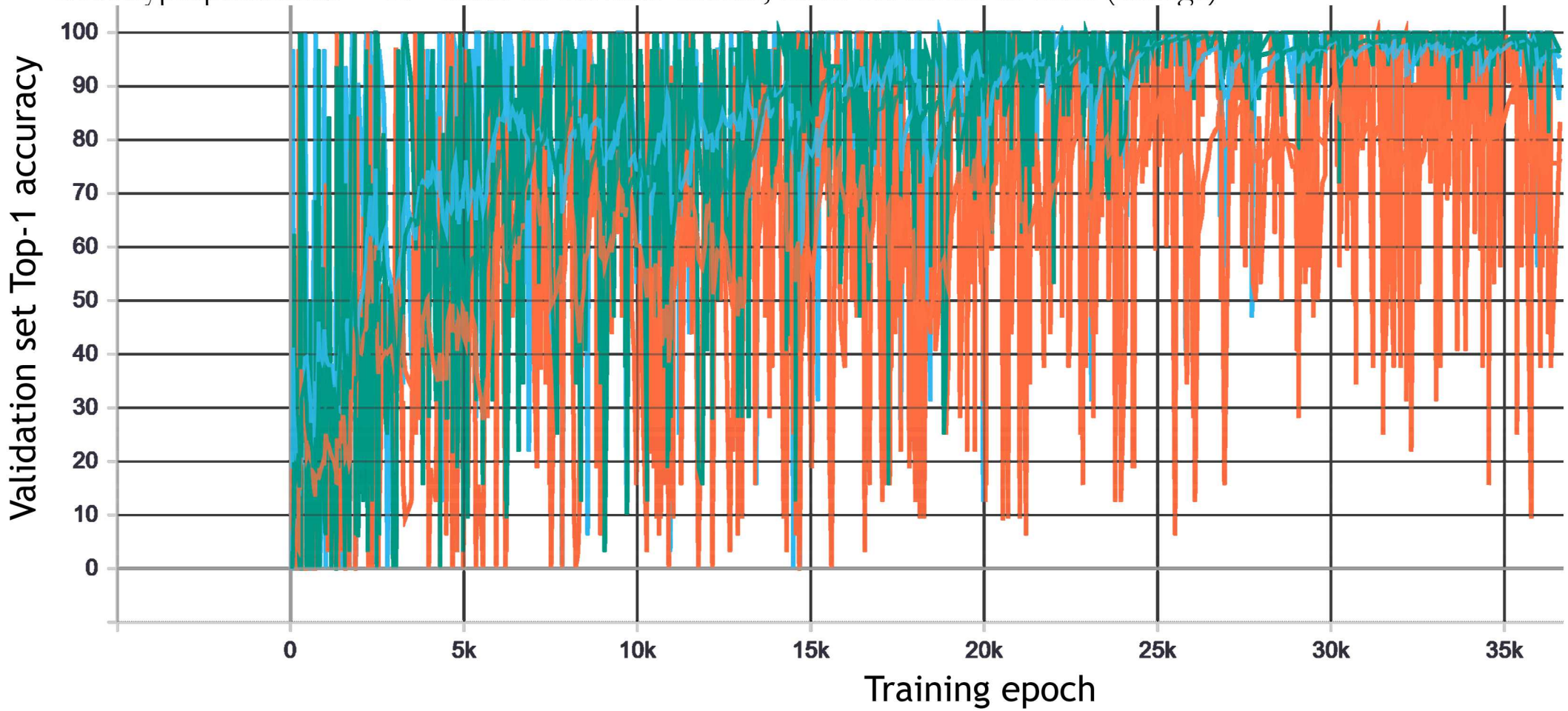
- No hardware cost
- Cost hyperparameter = 10^{-6}
- Cost hyperparameter = 10^{-5}



HAPDARTS – effect of hardware cost

Three trials

- No hardware cost leads to 3.8MB model, takes 5 hours to train (green)
- Cost hyperparameter = 10^{-6} leads to 1.55MB model, takes 2.1 hours to train (light blue)
- Cost hyperparameter = 10^{-5} leads to 1.37MB model, takes 1.8 hours to train (orange)



Future work

HAPDARTS results above use number of bytes as the cost.

Currently extending to support cost including memory operations, MAC operations, and latency on a given systolic architectures.

Future work will extend to support other learning algorithms (e.g. spiking neural networks) and hardware.

Conclusion

Sandia National Laboratories is pursuing neural-inspired computing as a transformative approach to computation.

We believe codesign enabled by fast, flexible, accurate hardware simulation and high performance computing will enable this pursuit.

Thank you!

What is a good cost function?

"Cost" of an operation can be simple, e.g. defined as the number of bytes needed for the operation.

Let $L = \gamma_1 L_1 + \gamma_2 L_2$ then we can find hyperparameters where both losses are considered.