SAND2020-1057PE

Advanced Computing, Sensing, and Algorithms for Highly Automated Driving

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.

Meeting purpose

- Share and provide feedback on technology gaps and research plans for computing in highly-automated vehicles
- Identify useful metrics for energy-efficient computing and sensing
- Discussion topics
  - Sensing technologies
  - Low power, edge computing
  - Artificial intelligence and machine learning
  - Simulation and data

National imperative


EXECUTIVE OFFICE OF THE PRESIDENT
WASHINGTON, D.C.

July 31, 2018

M-18-22

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: MICK MULVANEY
DIRECTOR, OFFICE OF MANAGEMENT AND BUDGET

MICHAEL KRATSIOS
DEPUTY ASSISTANT TO THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY

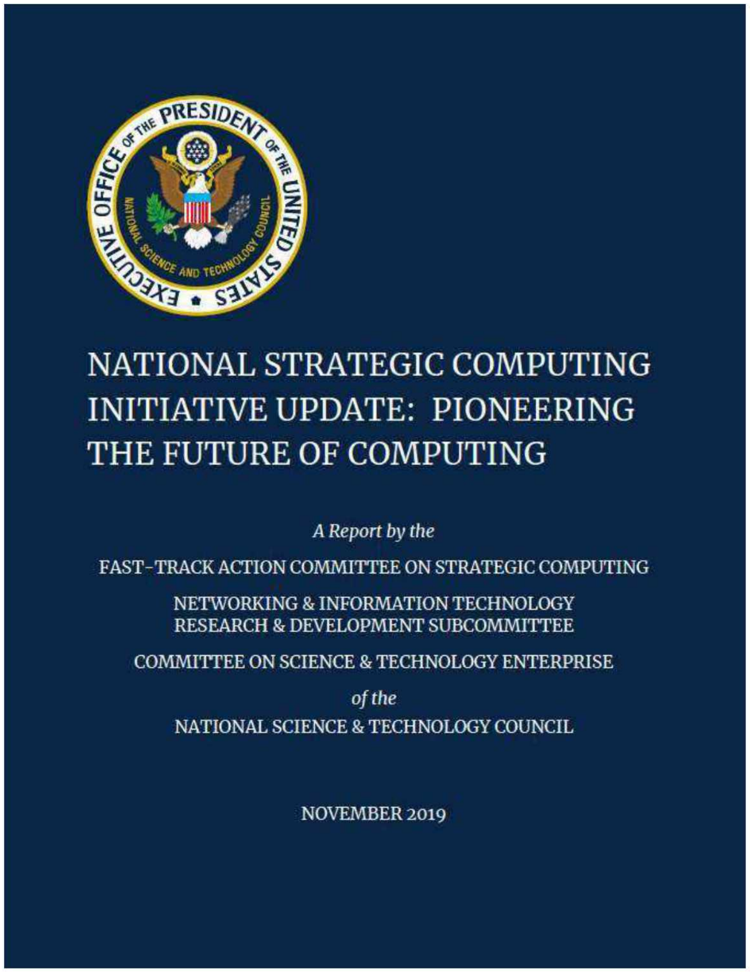
SUBJECT: FY 2020 Administration Research and Development Budget Priorities

"Agencies should prioritize investment in research and infrastructure to maintain U.S. leadership in strategic computing, from edge devices to high-performance computing, that accelerates delivery of low power, high performance devices; supports a national high-performance computing ecosystem; and explores novel pathways to advance computing in a post-Moore's Law era".

"Today, semiconductors underpin the most exciting 'must-win' technologies of the future, including artificial intelligence to power self-driving cars and other autonomous systems…

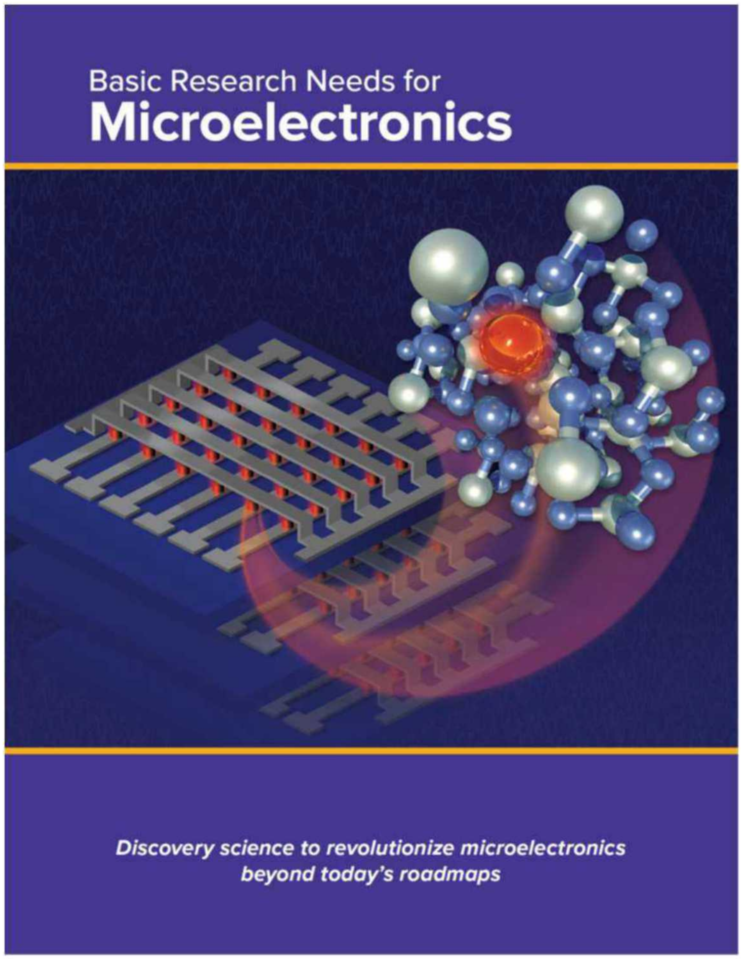
To secure America's leadership in these future technologies for the next 50 years, the United States must continue to lead the world in semiconductor research, design, and manufacturing"

National Strategic Computing Initiative



1. Pioneer new frontiers of digital and non-digital computation to address the scientific and technological challenges and opportunities of the 21st century.
2. Develop, broaden, and advance the Nation's computational infrastructure and ecosystem.
3. Forge and expand partnerships for the future of computing to ensure American leadership in science, technology, and innovation.

DOE exploring low-energy electronics and advanced computing



Priority research directions:

1. Flip the current paradigm:
   define innovative material, device, and architecture requirements driven by applications, algorithms, and software
2. Revolutionize memory and data storage
3. Reimagine information flow unconstrained by interconnects
4. Redefine computing by leveraging unexploited physical phenomena

DOE exploring low-energy electronics and advanced computing

Will highly automated vehicles be viable with conventional computing approaches, or will they require a step-change in computing?

What are the energy requirements to support on-board sensing and computing for highly automated vehicles?

What advanced computing approaches could reduce the energy requirements for highly automated vehicles while meeting their computational requirements?

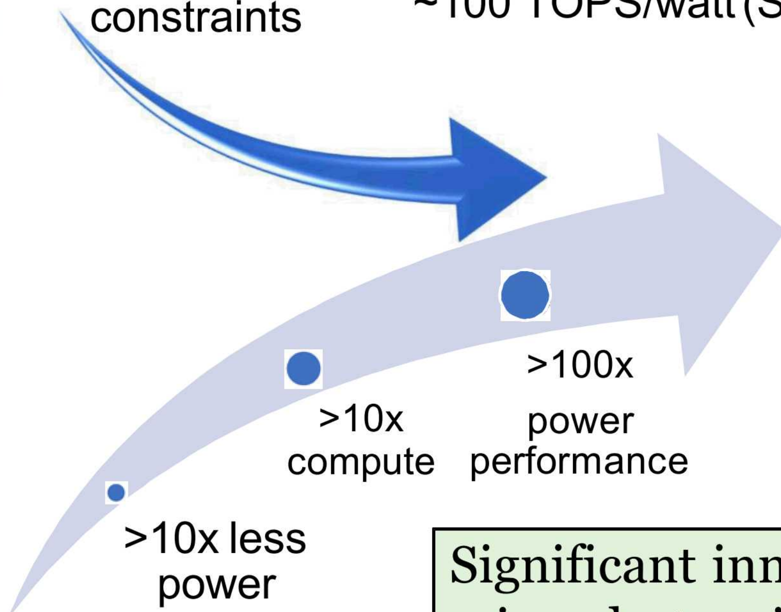
Projected computing performance and power

computing must meet size, weight, and power constraints

~1 petaflops
~100 W (system)
~100 TOPS/watt (SoC)

Full level 5 automated driving

TOPS == Trillion (tera) Operations



Early prototype self-driving

~100 teraflops
~1000 W (system)
~1 TOPS/watt (SoC)

Significant innovation will be required in microelectronic materials and devices, sensing and computing architectures, and computer algorithms.

https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip/

Why now for the computing industry?

Technology:

- End of Dennard power scaling; power becomes the key constraint
- Slow-down in Moore's Law, evidenced by flattening of transistor cost takedown

Architectural:

- Limitation and inefficiencies in exploiting instructional-level parallelism and the prevailing von-Neumann architecture

Applications:

- Shift from desktop to mobile and IoT
- Ultra-scale cloud computing and artificial intelligence/machine learning workloads

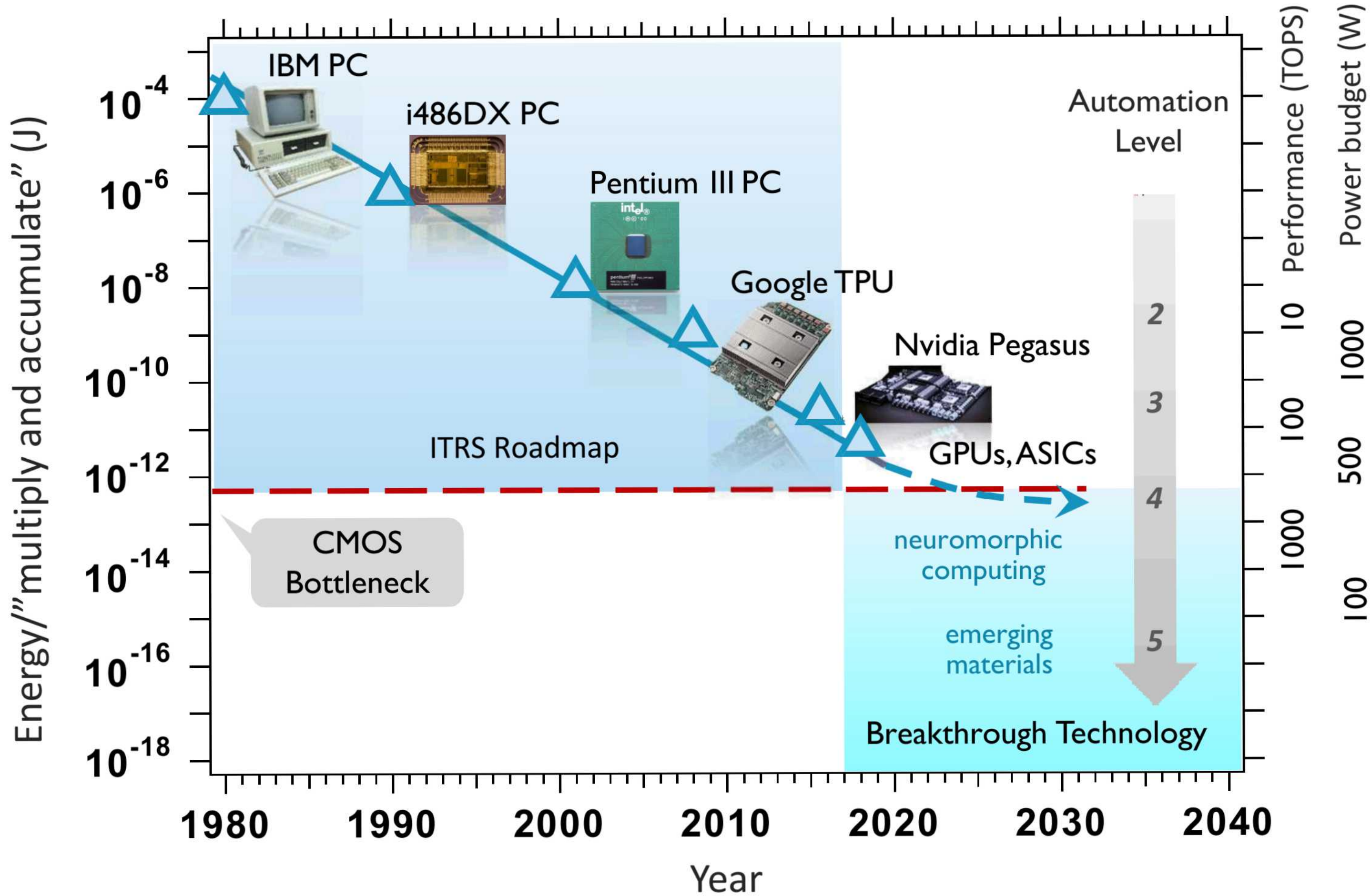
Industry collaborations:

- End of International Technology Roadmap for Semiconductors (ITRS) roadmap
- Decline in SRC participation and the end of SEMATECH (absorbed in SUNY)

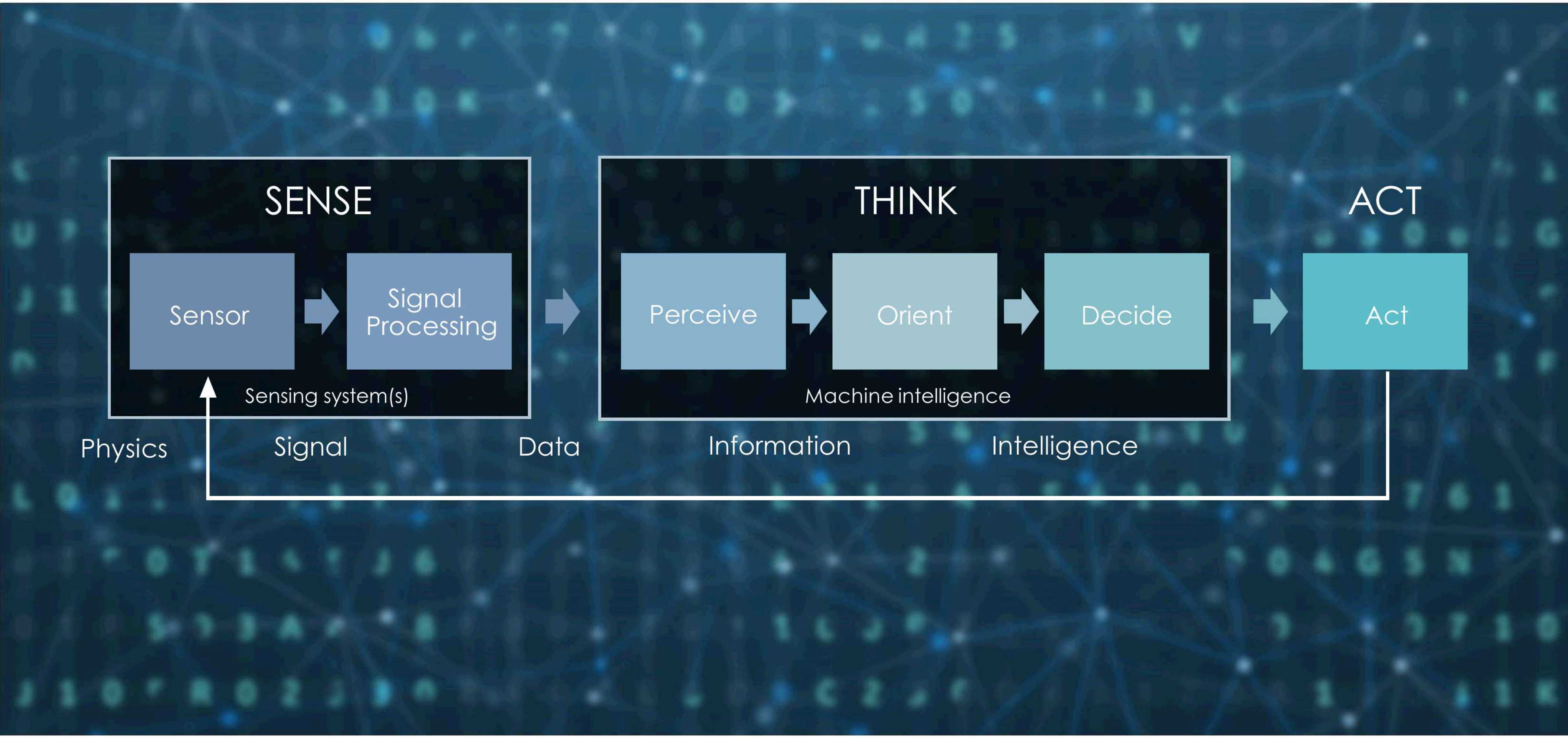
Performance targets require breakthrough technology

1pJ/op = 1 TOPS/W

1fJ/op = 1000 TOPS/W



Automated systems



Sandia’s focused research areas for highly automated vehicles

Sensors and Sensor Processing	Scene Perception and Algorithms	Navigation Hardware Accelerators
• Create disruptive optical sensing technology to reduce energy consumption by 100X • Develop chip-scale LiDAR to reduce cost by 100X	• Explore sparse coding and reduced-precision to reduce computation load by 1000X • Develop biologically inspired machine learning algorithms to reduce the number of training samples by 100X • Develop unsupervised and self-supervised learning algorithms	• Develop and demonstrate hardware capable of real-time processing of tera- to petabit inputs, with energies at <10 fJ per operation (>100 TOPS/W) • Develop algorithms for robust and reliable recognition tasks needed for perception. • Demonstrate the value of algorithm and hardware co-design such that combined elements have greater energy and/or SWaP improvement

Leverage broad Sandia capabilities

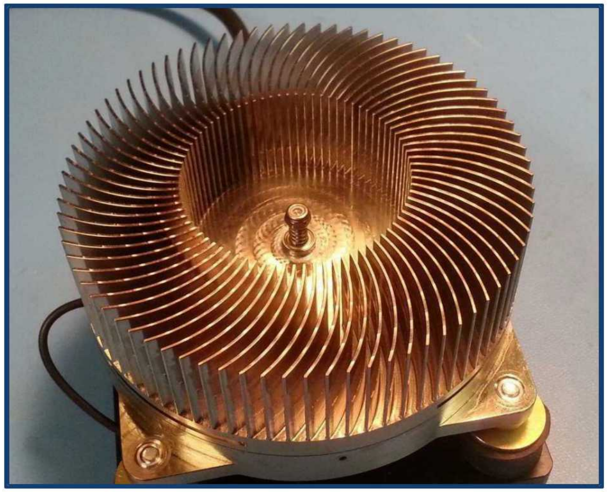
COMBUSTION RESEARCH FACILITY

MESA MICROFAB

COMPUTING & INFORMATION SCIENCE

CENTER FOR INTEGRATED NANO TECHNOLOGIES

Sandia Cooler



- Sandia Cooler technology has advanced through a DOE Technology Commercialization Fund project with industry partner Wakefield-Vette; now at TRL 8 with partner Heico
- Technology demonstrated in solid-state lighting for commercial warehouse applications
  - LED are located on rotating frame, ~1000 W power inductively coupled
  - Approximately 500 W of heat rejection
- Idea for CAV computing cooling – embed computing devices on rotating frame (similar to lighting) and communicate with adjacent vehicle data streams through 5G wireless link

Meeting purpose

- Share and provide feedback on technology gaps and research plans for computing in highly-automated vehicles
- Identify useful metrics for energy-efficient computing and sensing
- Discussion topics
  - Sensing technologies
  - Artificial intelligence and machine learning
  - Low power, edge computing

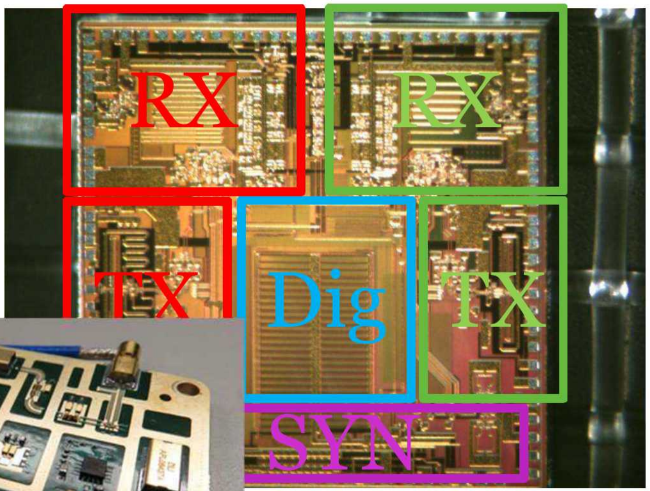
Sensor modalities
for
highly automated driving

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.

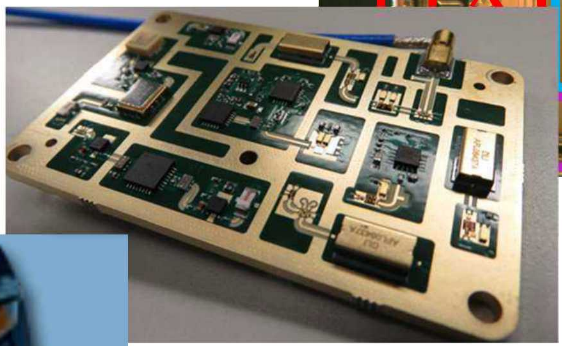
Sensor design, integration, and data interpretation expertise

- Sandia has decades of navigational expertise
  - Radio Frequency/Acoustic (GPS, radar, sonar)
  - Inertial Navigation (accelerometers)
  - Guidance Systems (telemetry, tracking algorithms)
- Designed, fabricated, and deployed navigation components
  - Radar systems/RF Microwave Components
  - Gyroscopes (laser ring) and 6 axis accelerometers
  - Imagers (X-ray, optical, radar)
- Pioneered radar image processing and precision GPS-denied navigation
  - Unique Algorithm development
  - High consequence computationally intense image processing and real-time object recognition and tracking

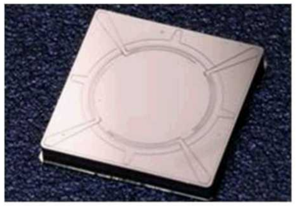
X and Ku Dual Band Radar RFIC



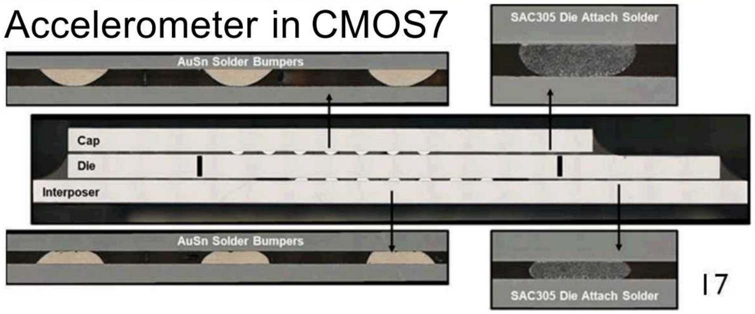
RX TX Module



Mini Synthetic Aperture Radar

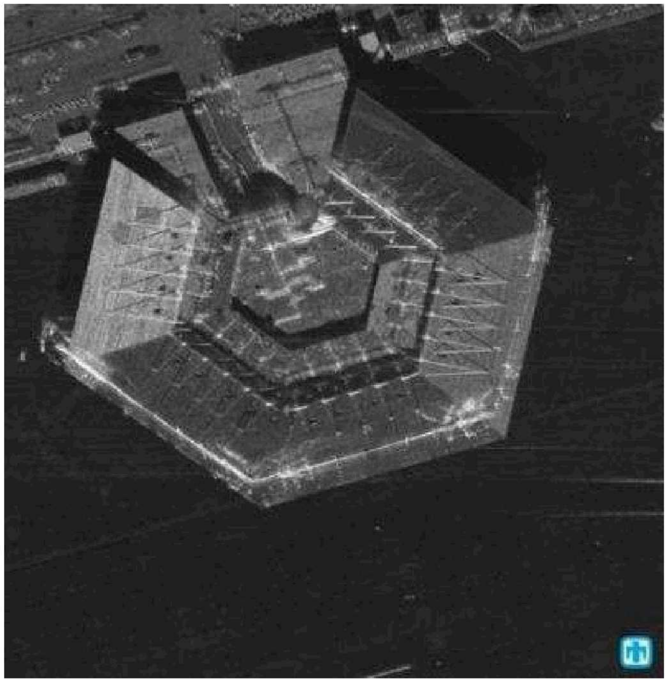


Accelerometer in CMOS7



Imaging radar

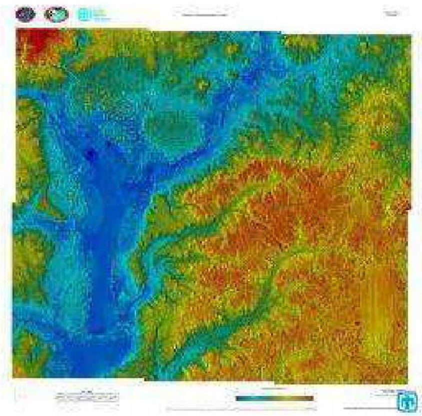
- Excellent complement to other sensors (electro-optical and LiDAR) for automated driving
  - Operates through all weather (fog, rain, snow)
  - Self illuminating (day, night)
  - Electrical scanning doesn't require moving parts
- High resolution, optical-like
  - Existing bands (76-81 GHz) provide centimeter class resolution
  - High frequency sensors result in small antennas/components
  - Resolution is not dependent on range to target
- Favorable computation complexity
  - Moving object detection (position/velocity vector) is a native product of radar, low computational complexity
  - Full radar image formation is computationally expensive but not needed in automotive applications
  - Image processing has significantly less computational cost than other imaging modalities
  - **Important because sensor data processing for useful information dominates complexity!**



SpotDwell image of a building at Jacksonville Naval Air Station.

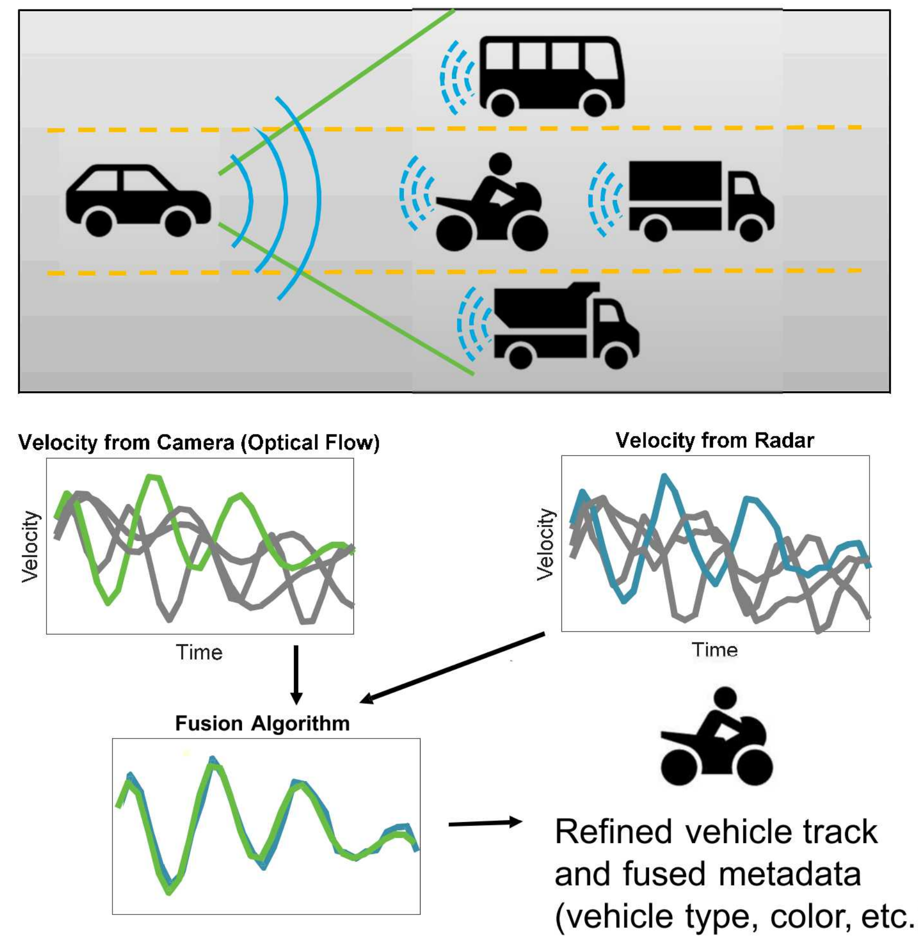
SNL imaging radar heritage

- 35 years of experience building real-time, high-resolution, low SWaP radar imaging platforms
- Pioneered many new image processing and exploitation techniques – and continue to innovate with new algorithms and methods
- Experts in harsh environment high performance electrical systems
- Expertise in low power mm Wave RFIC component design



Sensor data fusion

- Sensor fusion is the combination and exploitation of raw data at the sensor level rather than the derived data level
- Requires tight sensor integration
- Recent advancements at Sandia have been made in multi-sensor processing, but few multi-sensor platforms exploit true sensor fusion
- An example in the automotive arena would be the association of motion from moving vehicles in a radar return with object detected in video[1]
- The information resulting from sensor fusion will be higher confidence than the sum of information from sensors in isolation
- Allows application specific computing, in parallel to decision computing
- Sandia is a leader in adopting sub-threshold ASIC design (100x reduction in power)



[1] Naething, Richard M., and Richard C. Ormesher. "Doppler-assisted sensor fusion." U.S. Patent No. 10,267,895. 23 Apr. 2019.

Chip-scale beam scanners

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.

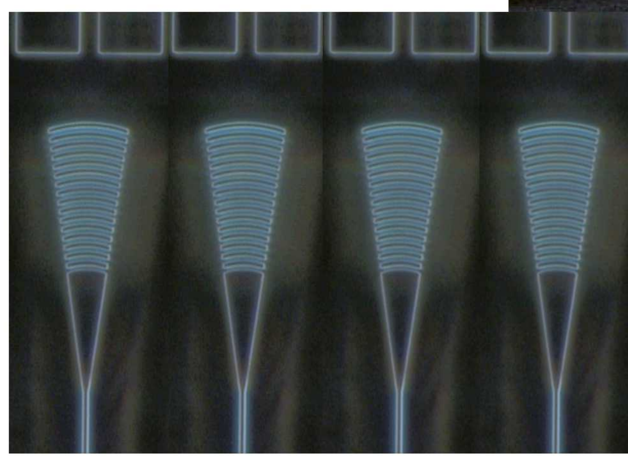
LiDAR for the masses

- Motivation:
  - autonomous vehicles require sensors that are robust, precise and easily manufacturable
  - current sensors use mechanical mirrors to steer light → large and expensive to produce

- Solution
  - Phased Arrays - recast established techniques at a new wavelength
  - Interfering waves create a narrow beam of light
  - Apply small electrical signals to adjust optical phase and steer beam
  - Requires precise fabrication at light wave dimensions and immense scalability

Very Large Array radio telescope – New Mexico

Mobile lidar mapping units atop a car by Blackmore Sensors and Analytics



Optical output gratings in silicon

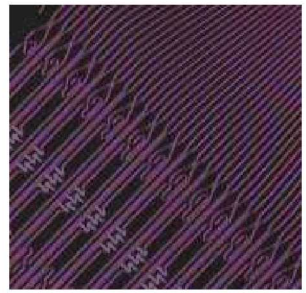
Silicon photonics solution

www.sandia.gov/mesa/nspc
photonics@sandia.gov

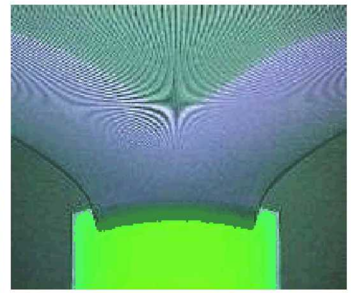
- Advantages
  - Low-cost: leverage investment on CMOS electronics
  - Reliable: no mechanical moving parts
  - Compact: several chips to provide large coverage
  - Mature: many device and system demonstrations

- Device challenges
  - Thermo-optical-electronic packaging
  - High optical power handling
  - Fast optical phase error compensation
  - Integration of new materials and layers

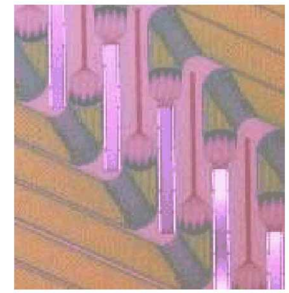
Key components of a future silicon photonic LIDAR sensor



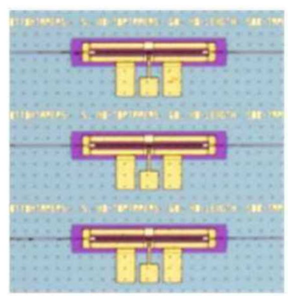
Low-loss, high-density waveguides



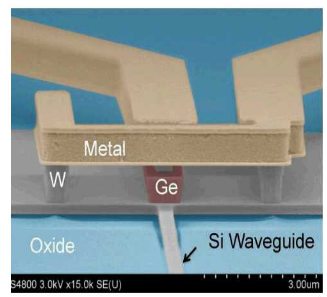
High-radix breakout



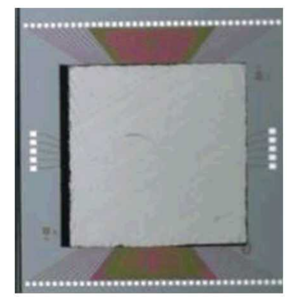
Efficient output couplers



Integrated laser and amplifiers

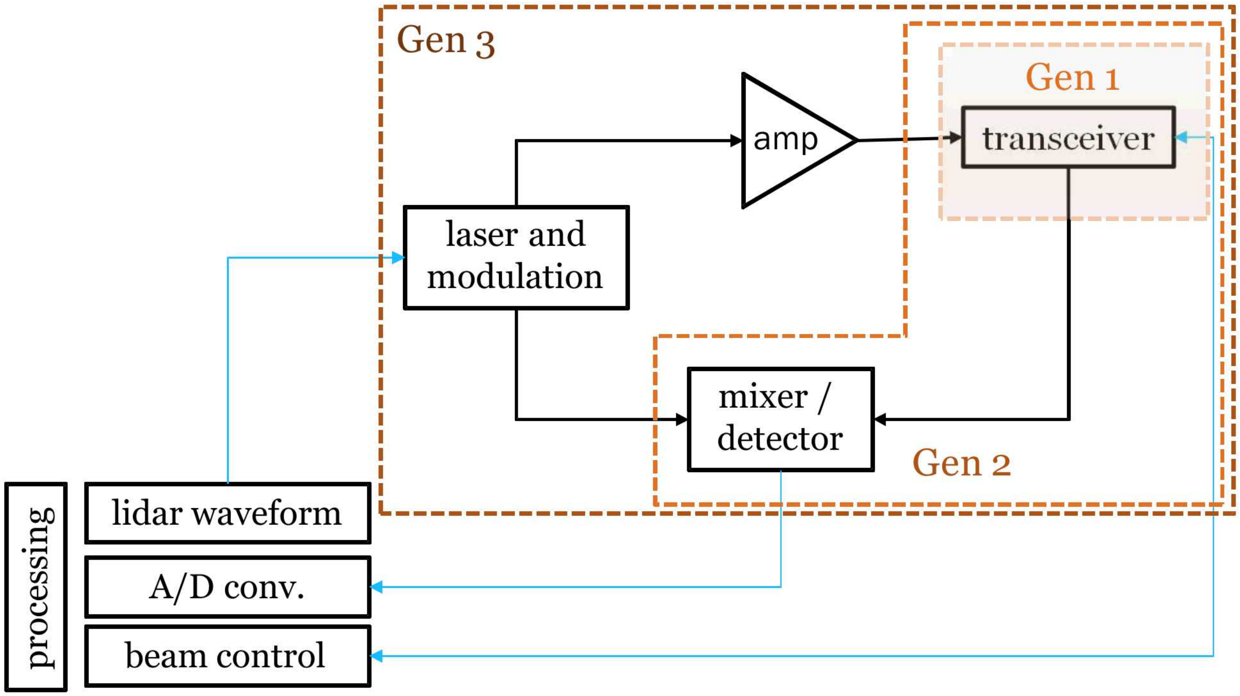


Ge photodetectors



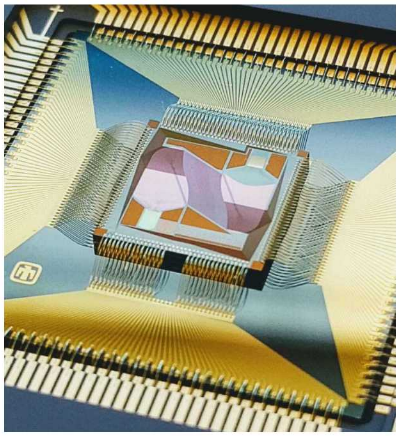
Heterogeneous integrated CMOS electronics

LiDAR engine



- **Gen 0 and 1**
  - DARPA SWEEPER
    - proof of principle with beam scanners and a few emitters
    - patented idea for simplified controls
  - Blackmore CRADA
    - expanded array size and added packaging
    - designs compatible with short-distance ranging
- **Gen 2**
  - improved efficiency, power handling and functionality
  - supported through LDRD program
  - fabrication is underway at MESA fab
- **Gen 3**
  - closer integration of laser, modulator, amplifiers, scanner, detector and CMOS controls to create a highly complex chip
  - fab plan developed and patented

Gen-1 results



Packaged 2D beam scanner

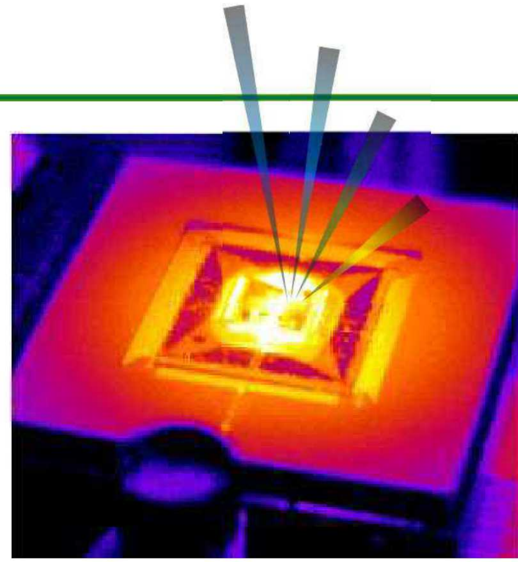
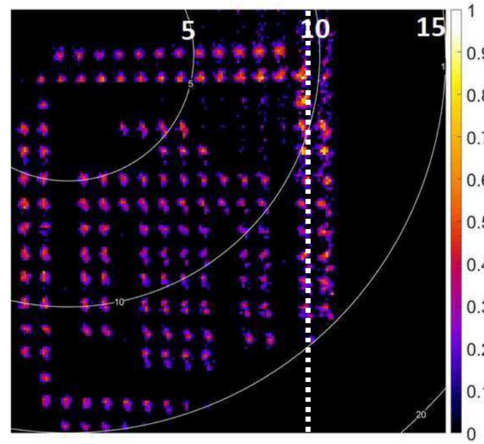
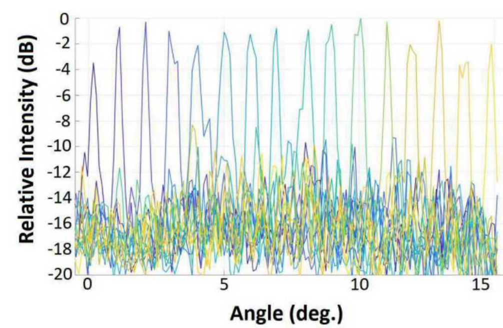


Image of chip with IR optical input



Composite image of beam scan

- 2D scanning with electronic and wavelength controls
  - electronic packaging with interposer and chip carrier
  - N=256 independent channels, d=3 μm
  - long passive grating outputs for high fill factor aperture
- Field of view: 24° × 10°; divergence angle: 0.3° × 0.3°
  - near diffraction limited operation!
- Supporting technologies for new applications: machine vision, situational awareness, optical communication



Electronic beam steering

M. Gehl, et al., CLEO 2019.

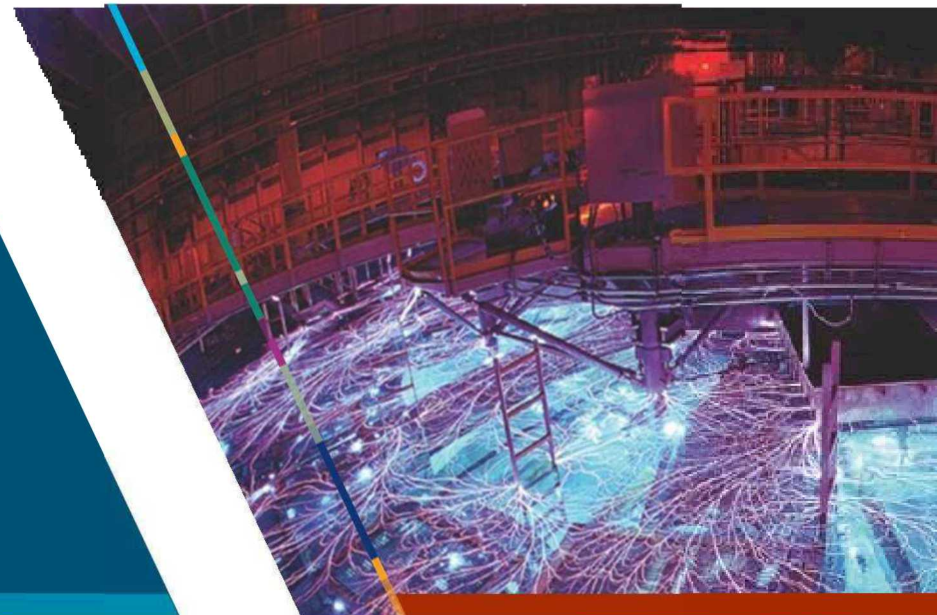
Sandia National Laboratories

Exceptional service in the national interest

Artificial Intelligence and Neural-Inspired Computing
At Sandia National Laboratories

Presented by: William M Severa, PhD.

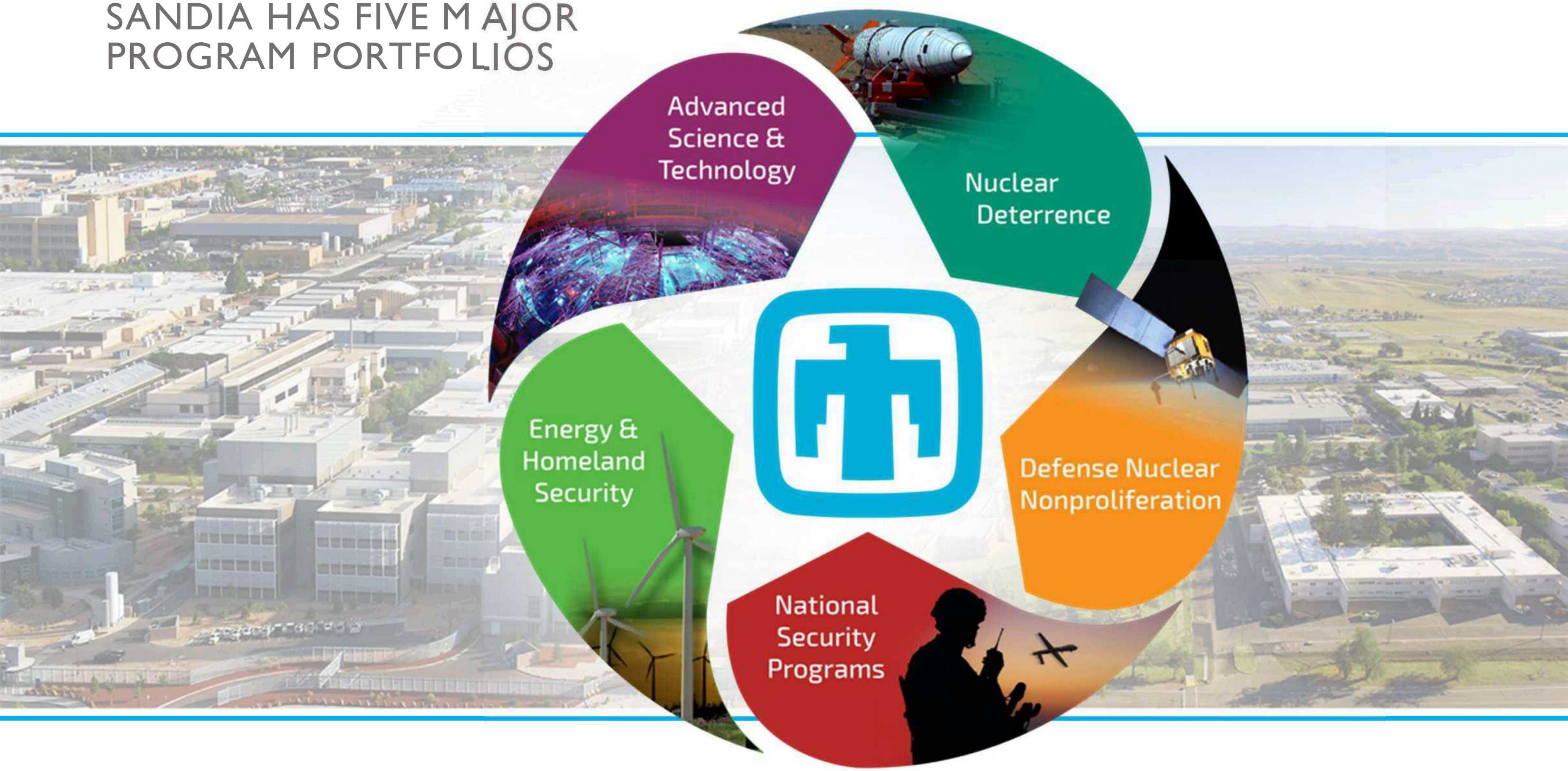
wmsever@sandia.gov



U.S. DEPARTMENT OF ENERGY NNSA

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.

SANDIA HAS FIVE MAJOR PROGRAM PORTFOLIOS



Why has ML/AI had so much attention?

Some problems are difficult to solve with a directly-coded algorithm

- Don't generalize well
- Can be difficult to scale
- Have to write a program by hand for each specific task
  - Some tasks can be very difficult to encode
- Hand coded algorithms may run much slower

There have been Machine Learning (ML) successes in a variety of areas

- Recognizing patterns
- Anomaly detection
- Learning predictive models from data
- Creating surrogate models
- Automating repetitive computing tasks
- Generating synthetic data that models real data
- Assisting human decision making

These successes have been enabled by

- Large curated (labeled) datasets
- Advancements in computing power

Dennard scaling

- As transistors get smaller, their power density remains constant

Unfortunately ended 10-15 years ago

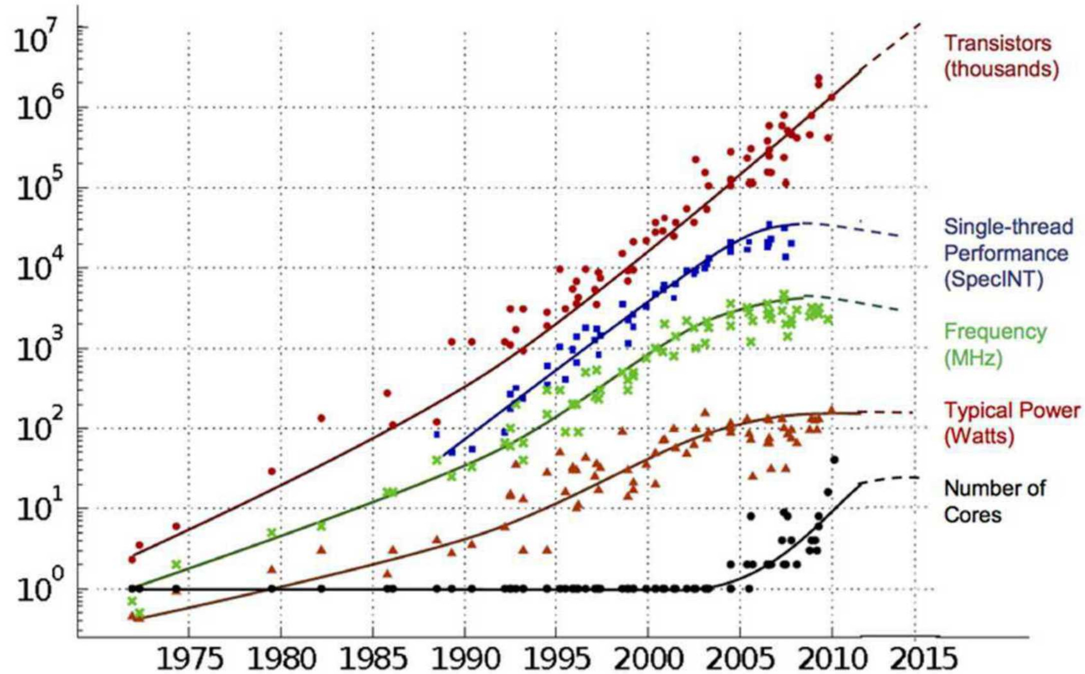
- Cannot run CPUs at faster speeds
- Emphasis on multi-core

Need for new paradigm of computing:

Novel Algorithms – Use AI to Accelerate

Novel Architectures – Accelerate AI

Novel Devices – Accelerate AI



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

Sandia's Unique Mission Needs

Diversity

Scale

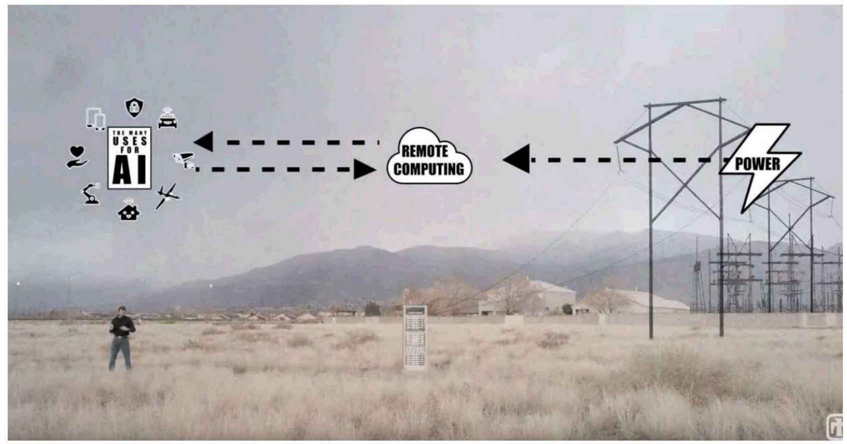
Consequence

Sandia's Unique Mission Needs

Diversity

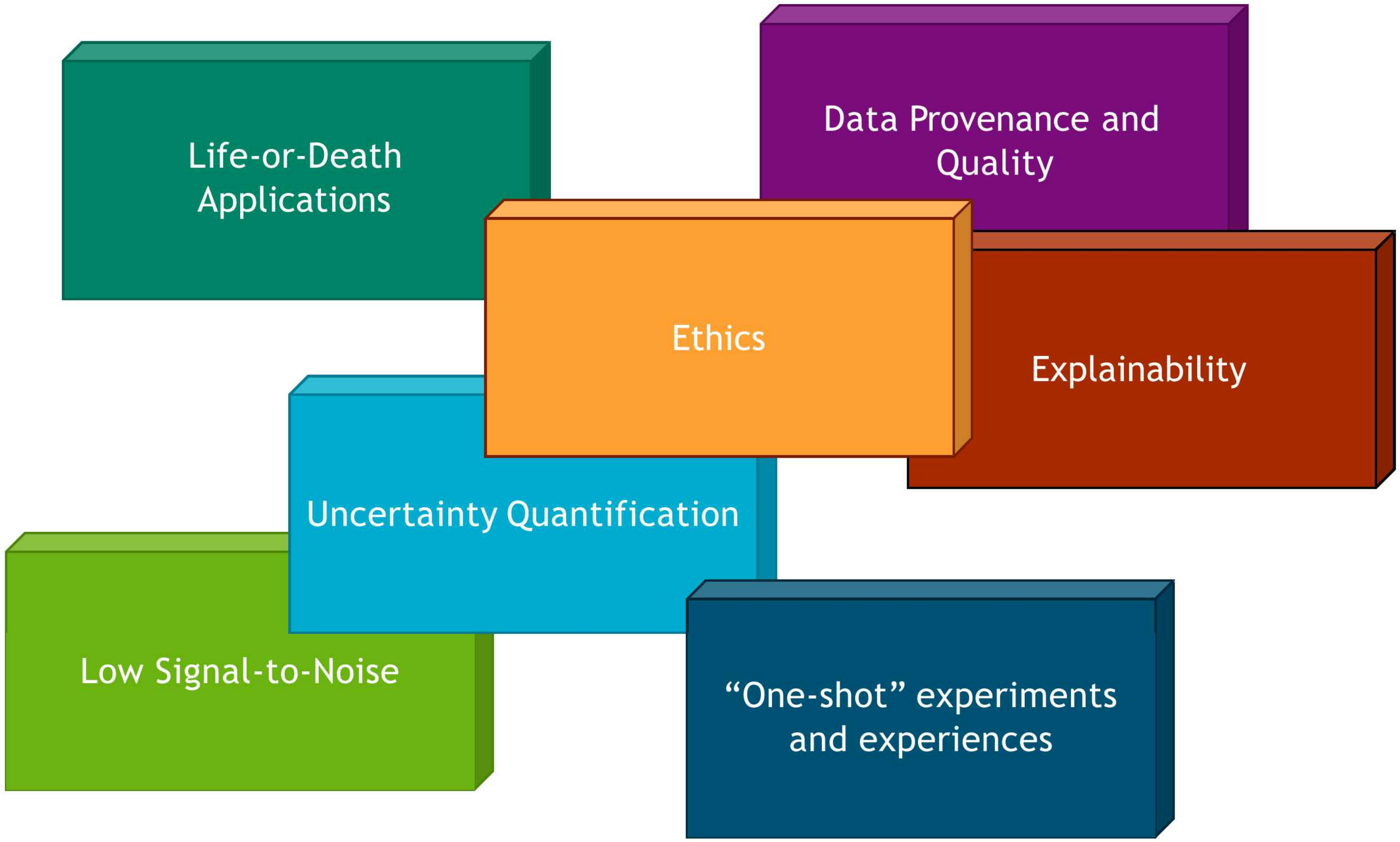
Scale

Consequence



Diversity

Scale



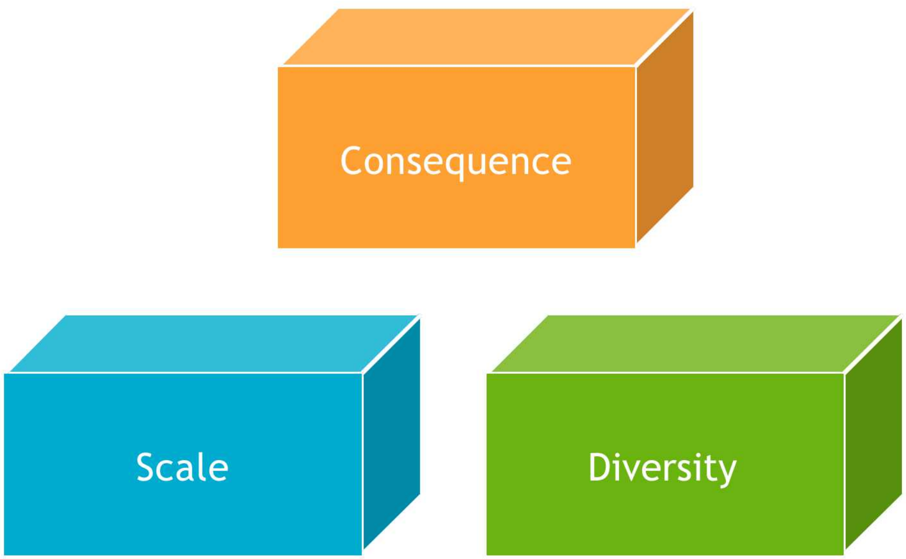
High-confidence decisions

- Typically designing to "Five 9's" of reliability
- Need to assure trust in our solutions
- Need to understand uncertainty of decisions
- Algorithms need to be explainable

classified as rifle
classified as other

Synthesizing Robust Adversarial Examples, Athalye, et.al., 2018

Many Sandia efforts are premised on idea that *brain-inspired* AI solutions will be instrumental in delivering these requirements



Sandia has a goal of creating a bridge between the broader world of AI and our missions

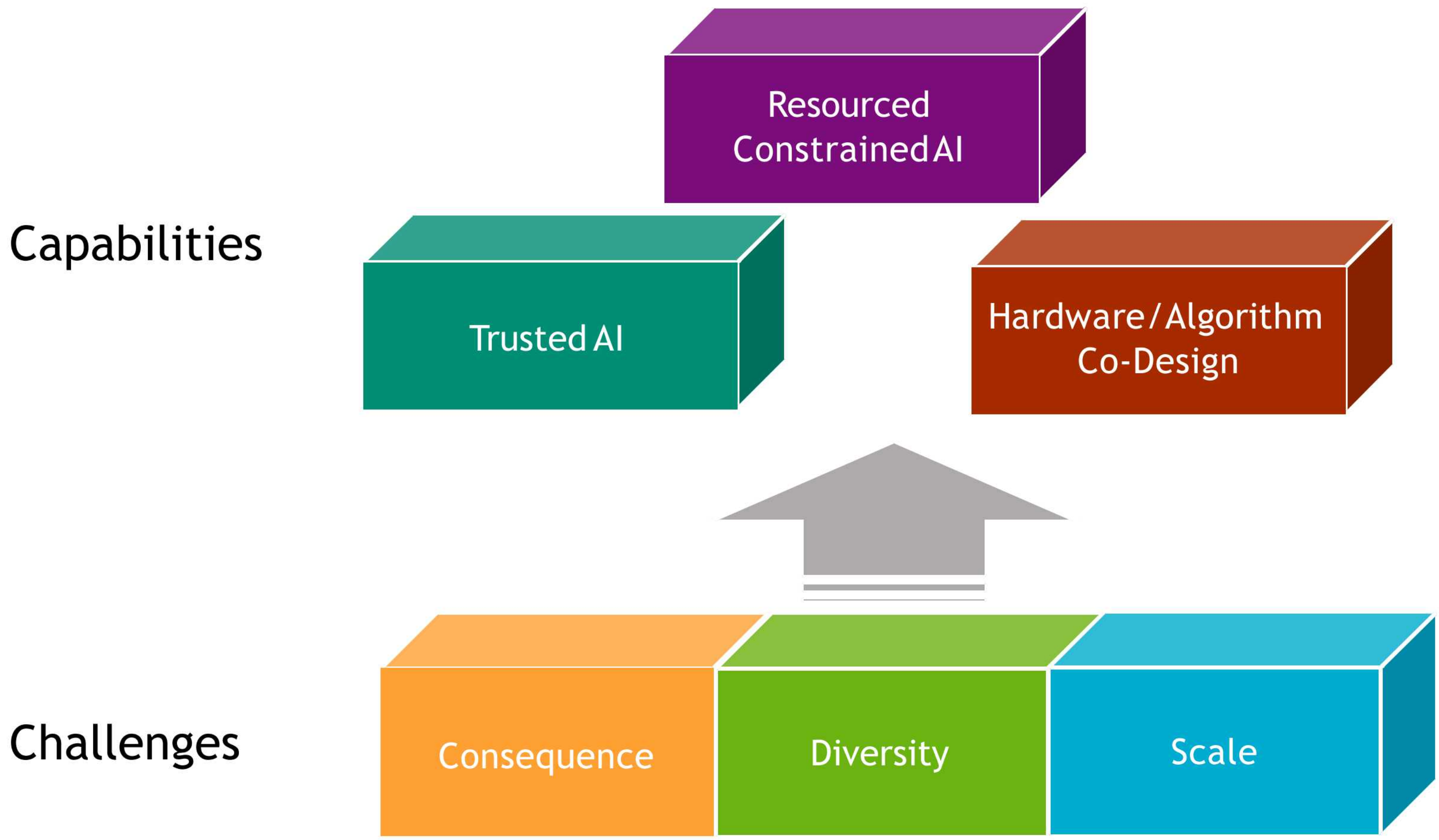
Extending and developing AI algorithms

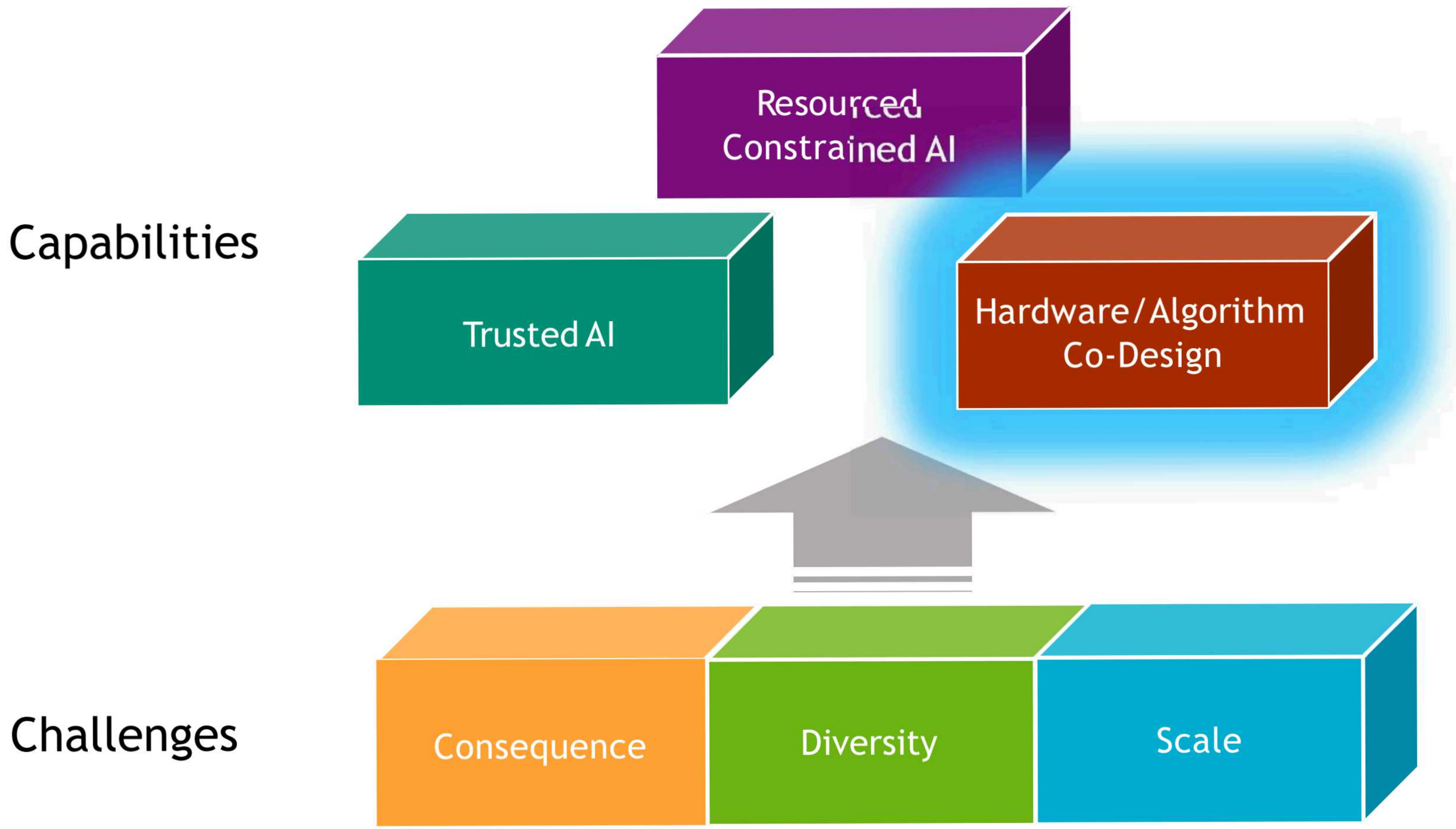
Evaluating novel hardware and accelerators

Explore brain-inspired sensor technology

Identifying opportunities for novel AI impact

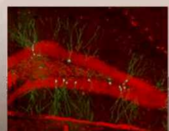
Developing tools and analyses suitable for widespread adoption of emerging AI technologies


Capabilities
Resourced Constrained AI
Trusted AI
Hardware/Algorithm Co-Design
Challenges
Consequence
Diversity
Scale


Resourced Constrained AI
Capabilities
Trusted AI
Hardware/Algorithm Co-Design
Challenges
Consequence
Diversity
Scale

Neural Computing at Sandia Labs Leverages a Large Research Foundation

Neuroscience Theory



IARPA MICrONS Government Team for Test & Evaluation of Neural Models and Machine Learning

- Neuro-informatics
- Modeling and Simulation
- Neural Data Analytics
- Computational Neuroscience

Neural Computing Capabilities

HAANA Grand Challenge – Flagship LDRD across computing, materials, and cyber security centers

- Formal Neural Computing Theory
- Neural Inspired Architectures
- Neural Machine Learning Algorithms
- UQ / SA of Neural Algorithms and Neural Architectures

Neural-enabling Hardware

MESA Fabrication Facility provides materials and design research capabilities for next generation neural systems

- Micro-sensors
- Non-von Neumann architectures
- Memory technology
- Neuromorphic Computing Lab

Mission Impacts

Enabling Advanced Simulation and Computing

- Neural-inspired communication paradigms
- Adaptive memory management
- Numerical computing with neurons

ASC

Deployable National Security Applications

- Cyber Defenses
- Embedded Pattern Recognition Systems
- Smart Sensor Technologies

U.S. DEPARTMENT OF ENERGY

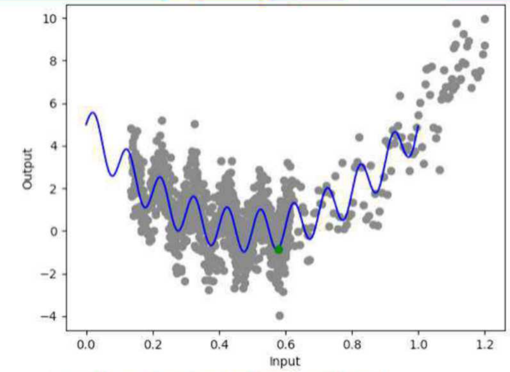
AI/ML to Advance HPC Mission at Sandia

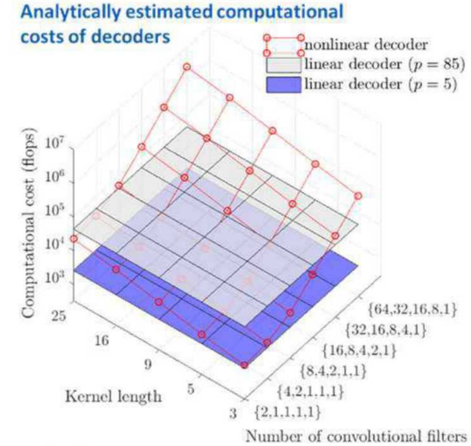
Machine learning will provide new capabilities for scientific and engineering applications

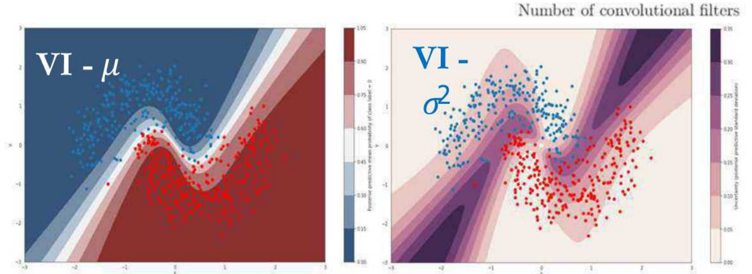
- Reduced order surrogate models for scientific/engineering problems
  - Could help us learn what is wrong/missing in physics models and aid in experimental design
- Ability to identify anomalies and regions of interest in inspection, surveillance, and large scale computational data
- Correlating and certifying simulation and experimental results
- Improving agility of application workflows (automating processes)

Machine learning will provide new capabilities for HPC system administrators, facilities, dev-ops, and system software

- Help model complex behaviors (e.g., failures, degradation, energy)
- Automate/adapt usage to comply with more complex policy (e.g., energy consumption)
- Adaptable resource management (e.g., network, memory, storage, energy)
- “Smart” data-movement for Exascale runtimes







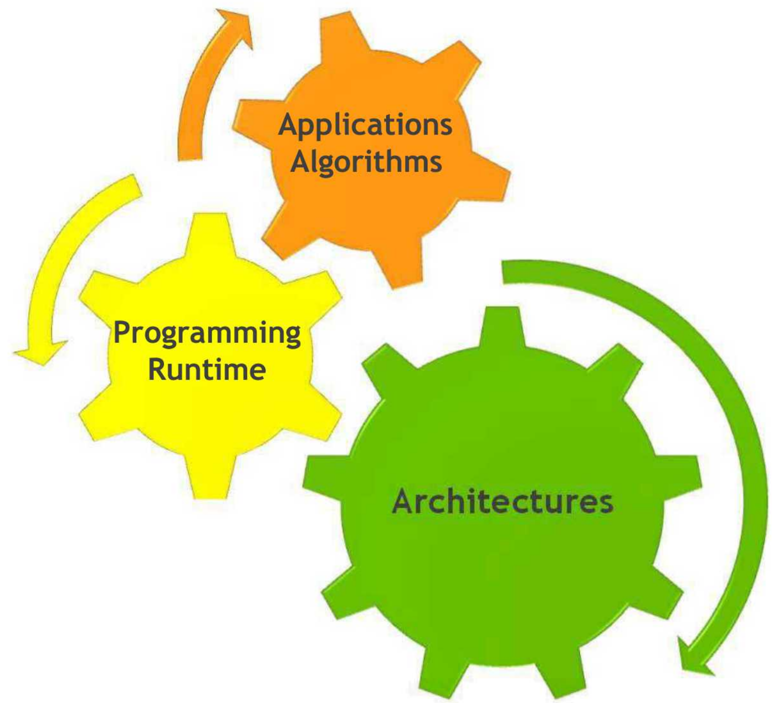
Center for co-design of ARtificial Intelligence focused Architectures and Algorithms (ARIAA)

ARIAA is a co-design research center that includes Pacific Northwest National Lab (PNNL), SNL, and Georgia Tech., supported by NVIDIA and Qualcomm

- Siva Rajamanickam, SNL PI (PNNL is lead lab)

ARIAA's objectives:

- Co-design novel AI/ML architectures, algorithms, and programming abstractions to enable traditional and ML-based DOE applications
- Understand how AI-focused dataflow/spatial architectures can impact future leadership class systems
- Understand how AI/ML accelerators can work with sparse, irregular, and/or streaming data

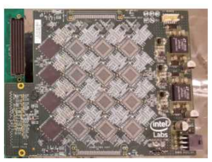


Codesign of AI/ML accelerators with algorithms and applications will enable the development of this key technology to suit DOE HPC and AI/ML needs

Pacific Northwest NATIONAL LABORATORY Sandia National Laboratories Georgia Tech

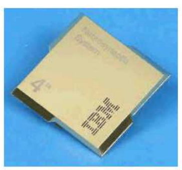
- ❑Enables researchers to explore the boundaries of neural computation
- ❑Consists of a variety of neuromorphic hardware & neural algorithms providing a testbed facility for comparative benchmarking and new architecture exploration

Intel Loihi



SpiNNaker 48 Node Board

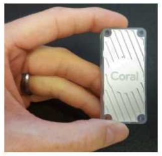
IBM TrueNorth*



IBM TrueNorth NS16e*

Intel Neural Compute Stick

Google Coral



Google EdgeTPU

Inilabs DAVIS 240C DVS

Georgia Tech FPAA

Intel Loihi

SNL STPU on FPGA

Xilinx PYNQ FPGA

Nengo FPGA

Nvidia Jetson TX1

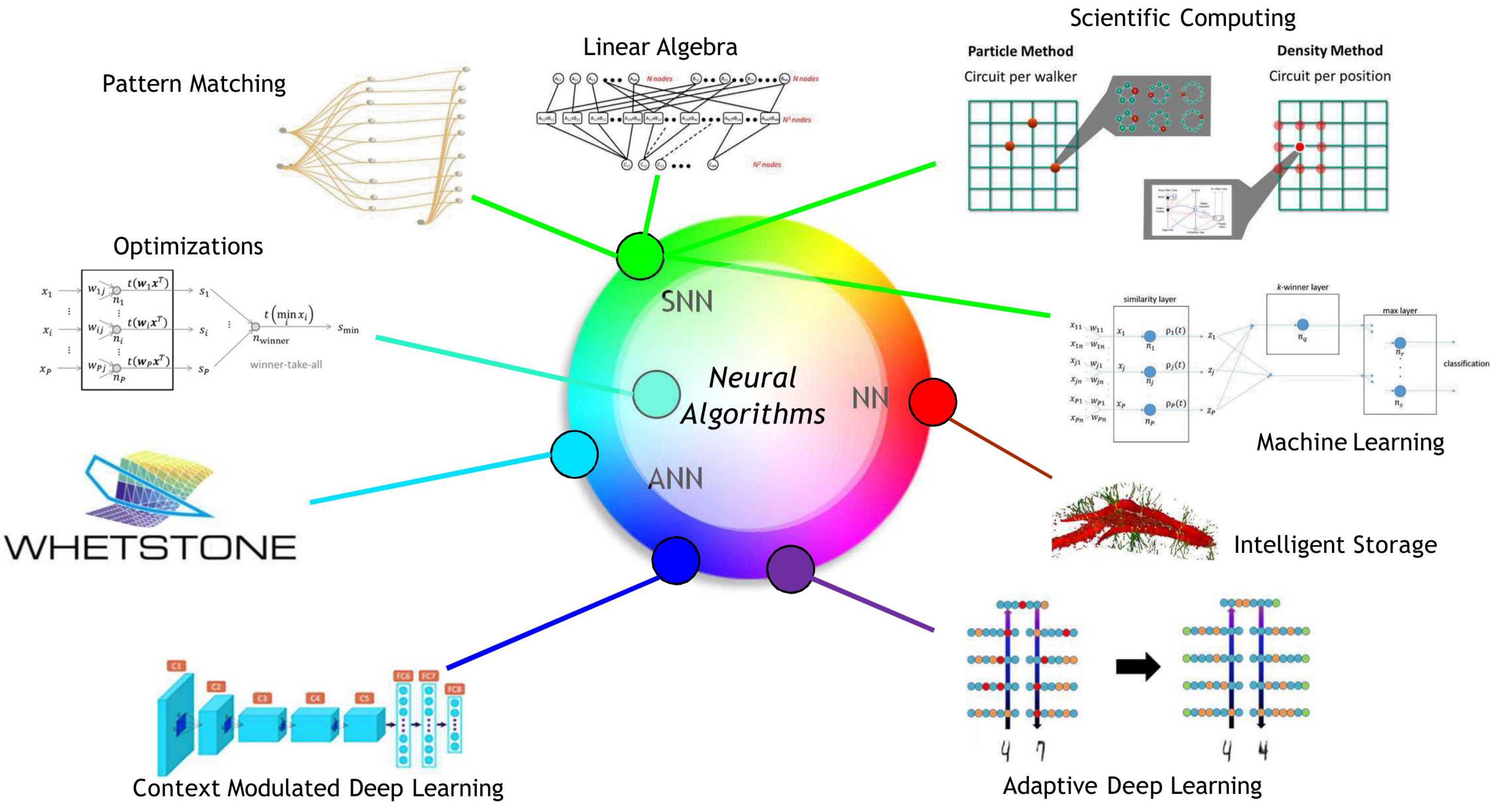
Nvidia Jetson Nano

GPU Workstations

Cognimem CM1K

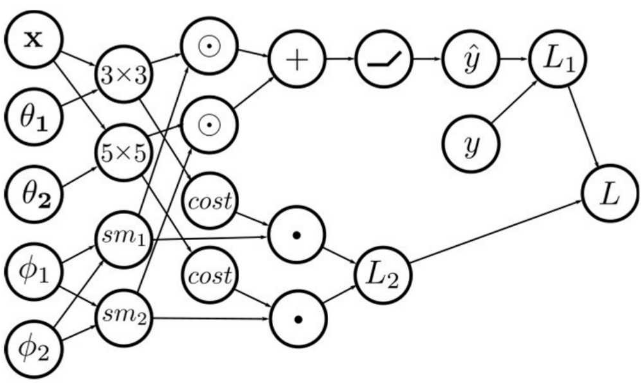
KnuPath Hermosa

*Remote access

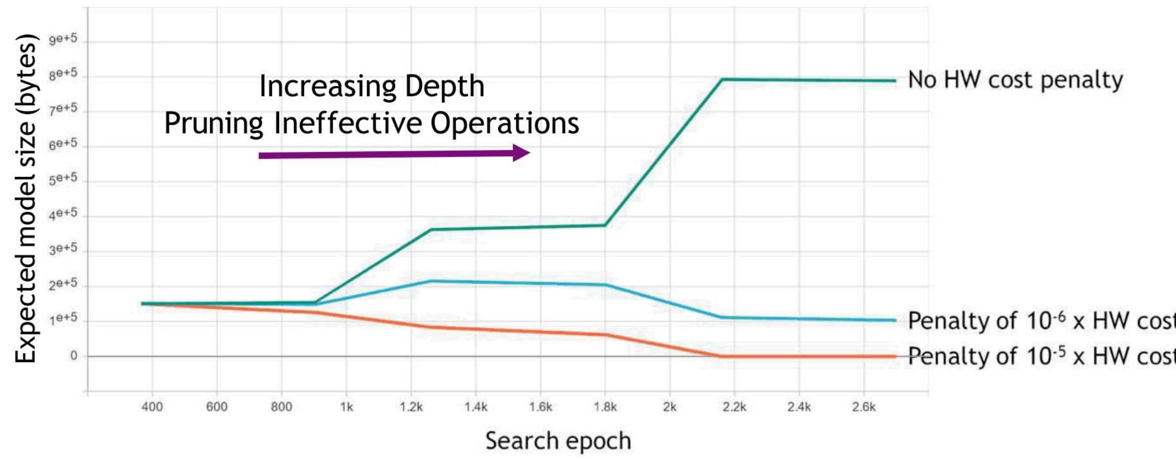

Scientific Computing
Particle Method
Circuit per walker
Density Method
Circuit per position
Linear Algebra
Pattern Matching
Optimizations
winner-take-all
SNN
Neural Algorithms
NN
ANN
similarity layer
k-winner layer
max layer
classification
Machine Learning
WHETSTONE
Intelligent Storage
Context Modulated Deep Learning
Adaptive Deep Learning

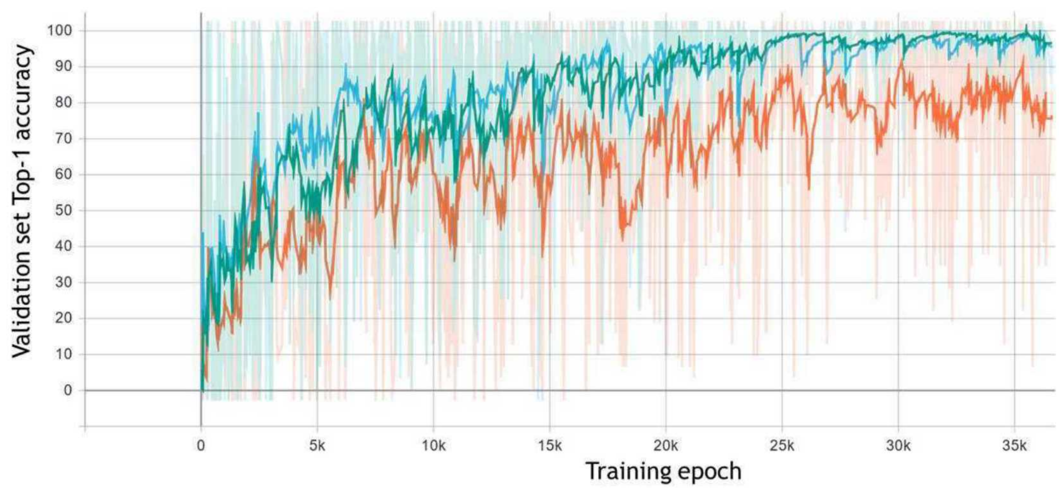
FORGE: Resource-Aware, Gradient-Assisted Neural Architecture Search

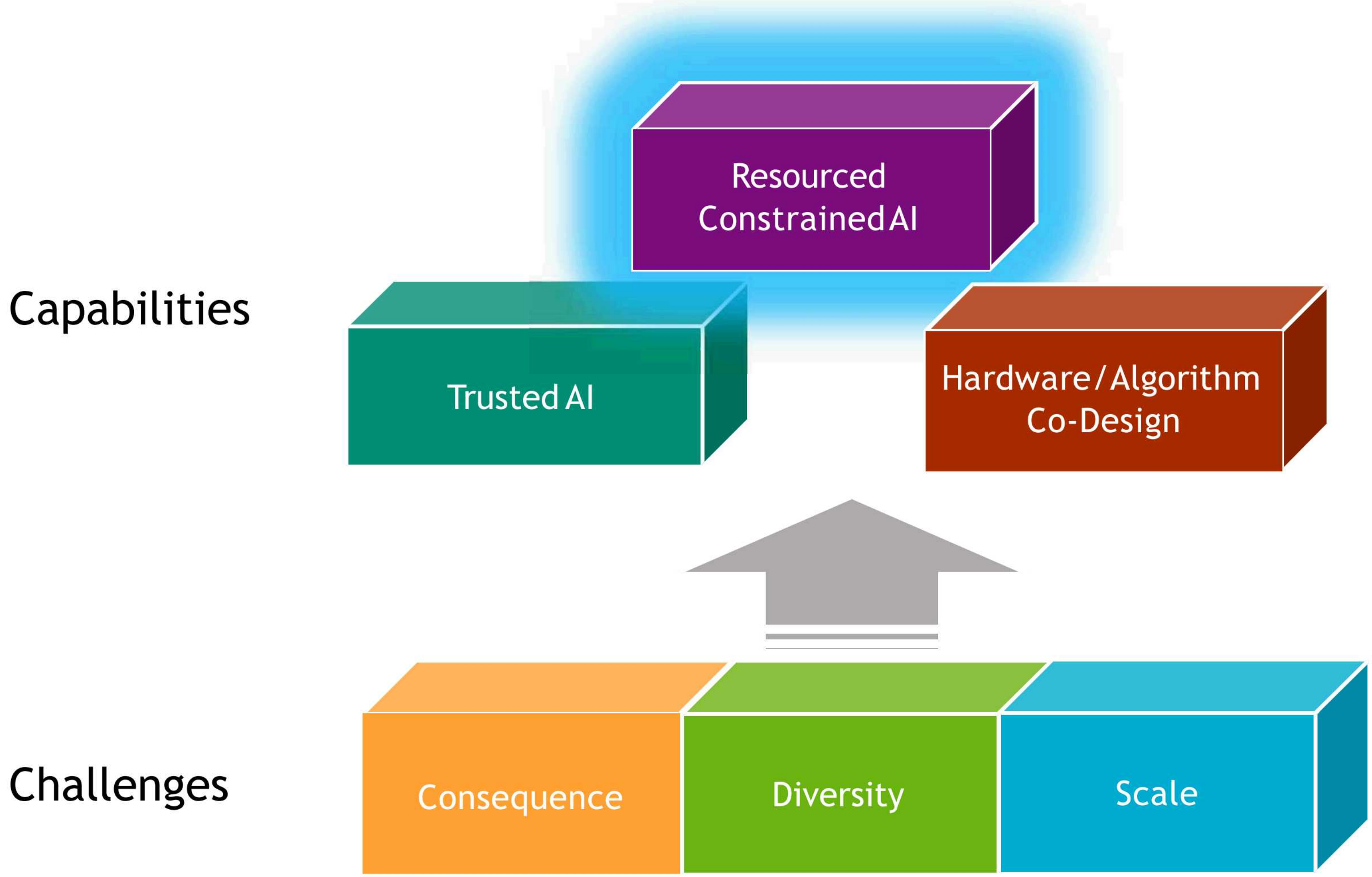
- Multi-level, Multi-objective Gradient-Assisted Optimization
- Automated Deep Learning Neural Network Design
- Highly Parallel Distributed Design
- Tailors Algorithm to Both Task and Target

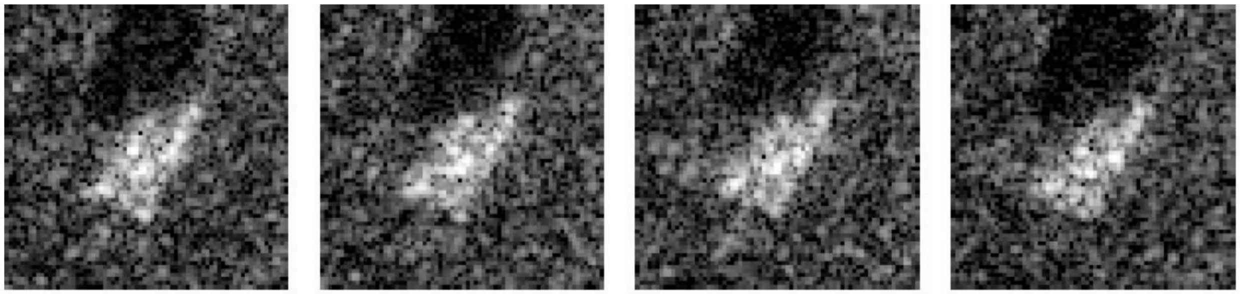


Sample Network Schematic

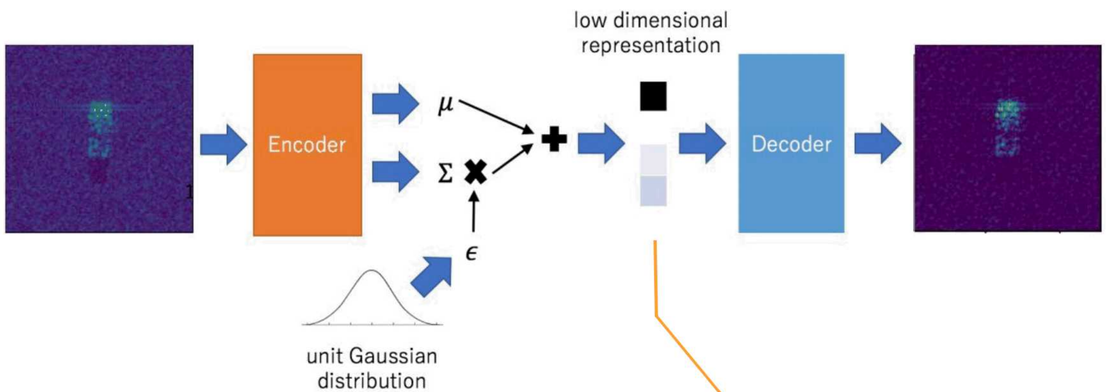




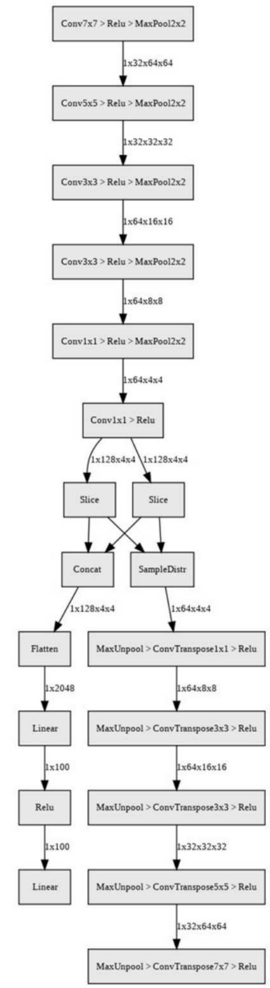

Capabilities
Resourced Constrained AI
Trusted AI
Hardware/Algorithm Co-Design
Challenges
Consequence
Diversity
Scale



Synthetic Aperture Radar returns suffer signal variability due to coherence, specularity, and speckle

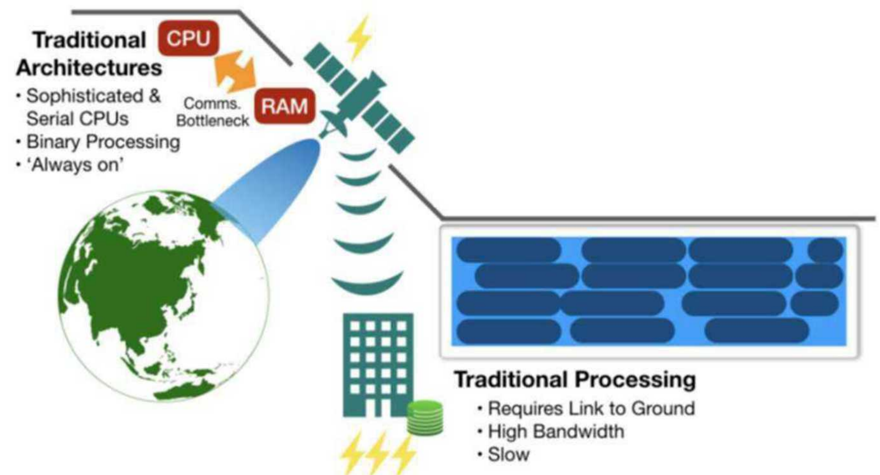


Assurance via Statistical Tests on Low-Dim. Representation



Challenges in classic remote sensing

- Growth of sensor technologies outpacing communication bandwidth
- High Consequence Decisions
- Limited algorithm capabilities
- Limited onboard processing capability
- Rad hard design



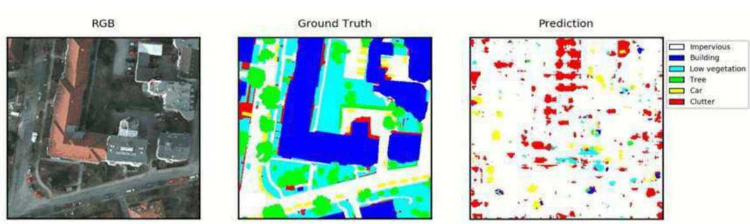
Pixel Data → Scene Understanding → Event Detection → Signal Reconstruction → Signal Classification → Communicate Results

Scene Understanding
- Scene Segmentation (i.e. clouds, land, water, etc.)
- Track scene movement across time

Event Detection
- Time and space localization of events of interest
- Time segmentation of event duration

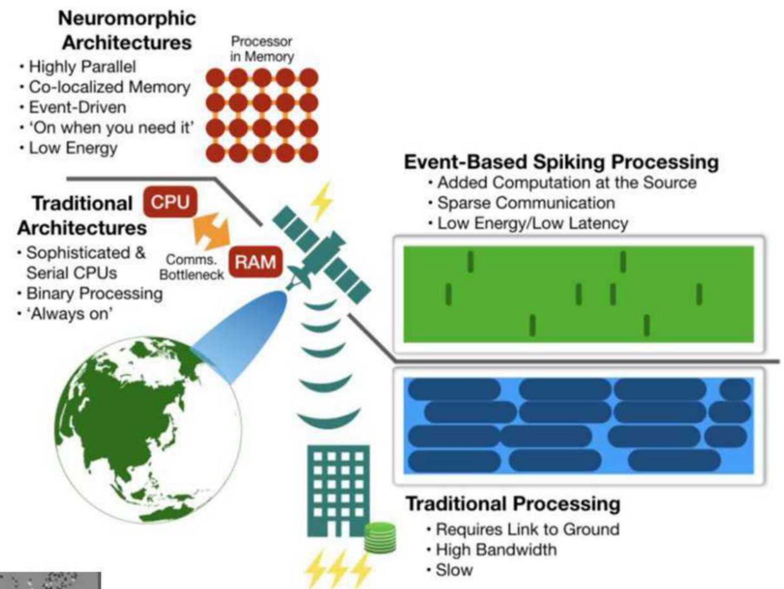
Signal Reconstruction
- Extract underlying signal from noise
- Overcome pixel saturation
- Generate human-understandable signatures

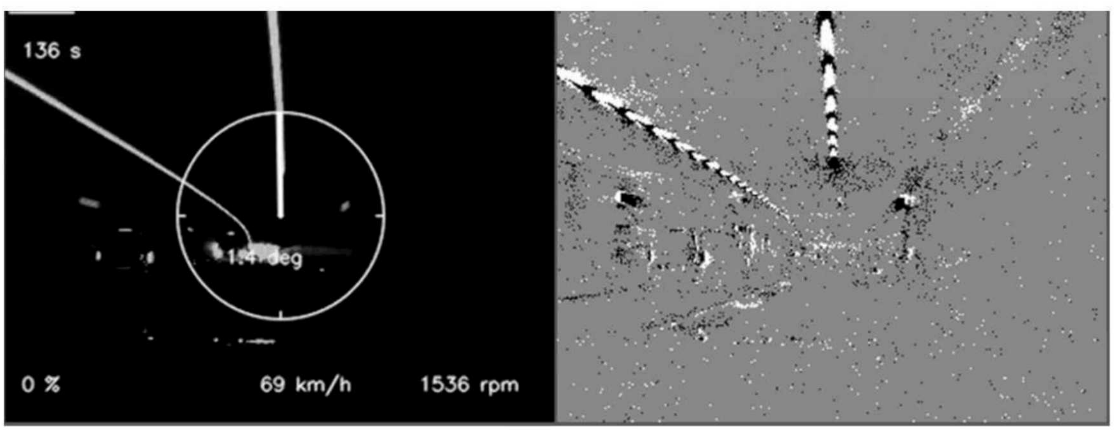
Signal Classification
- Determine if signal is of interest or is a confuser
- Reduce communication bandwidth by rejecting false alarms



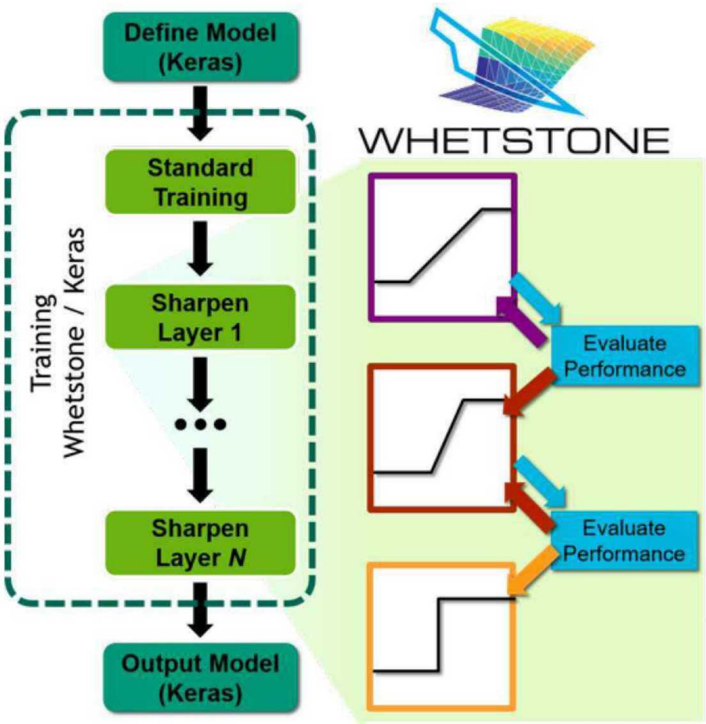
Compute at the sensor

- Improve bandwidth utilization (send only what you need)
- Distributed computation avoiding single-point of failure
- May reduce preprocessing required (e.g. whitening)

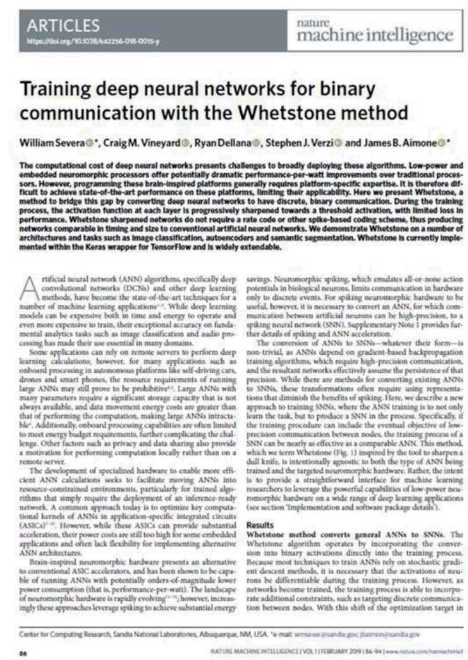




DAVIS 240C Event-Driven Camera



- Automatically converts deep learning networks from continuous valued neurons to binary activations, making them compatible with neuromorphic hardware
- Open sourced
- Published in February
- Beginning to port onto neuromorphic platforms
  - SpiNNaker Results look great



ARTICLES
https://doi.org/10.1038/s42256-018-0015-y
nature machine intelligence

Training deep neural networks for binary communication with the Whetstone method

William Severa*, Craig M. Vineyard, Ryan Dellana, Stephen J. Verzi and James B. Aimone*

The computational cost of deep neural networks presents challenges to broadly deploying these algorithms. Low-power and embedded neuromorphic processors offer potentially dramatic performance-per-watt improvements over traditional processors. However, programming these brain-inspired platforms generally requires platform-specific expertise. It is therefore difficult to achieve state-of-the-art performance on these platforms, limiting their applicability. Here we present Whetstone, a method to bridge this gap by converting deep neural networks to have discrete, binary communication. During the training process, the activation function at each layer is progressively sharpened towards a threshold activation, with limited loss in performance. Whetstone sharpened networks do not require a rate code or other spike-based coding scheme, thus producing networks comparable in timing and size to conventional artificial neural networks. We demonstrate Whetstone on a number of architectures and tasks such as image classification, autoencoders and semantic segmentation. Whetstone is currently implemented within the Keras wrapper for TensorFlow and is widely extendable.

Artificial neural network (ANN) algorithms, specifically deep convolutional networks (DCNs) and other deep learning methods, have become the state-of-the-art techniques for a number of machine learning applications[1–3]. While deep learning models can be expensive both in time and energy to operate and even more expensive to train, their exceptional accuracy on fundamental analytics tasks such as image classification and audio processing has made their use essential in many domains.

Some applications can rely on remote servers to perform deep learning calculations; however, for many applications such as onboard processing in autonomous platforms like self-driving cars, drones and smart phones, the resource requirements of running large ANNs may still prove to be prohibitive[4,5]. Large ANNs with many parameters require a significant storage capacity that is not always available, and data movement energy costs are greater than that of performing the computation, making large ANNs intractable[6]. Additionally, onboard processing capabilities are often limited to meet energy budget requirements, further complicating the challenge. Other factors such as privacy and data sharing also provide a motivation for performing computation locally rather than on a remote server.

The development of specialized hardware to enable more efficient ANN calculations seeks to facilitate moving ANNs into resource-constrained environments, particularly for trained algorithms that simply require the deployment of an inference-ready network. A common approach today is to optimize key computational kernels of ANNs in application-specific integrated circuits (ASICs)[7–10]. However, while these ASICs can provide substantial acceleration, their power costs are still too high for some embedded applications and often lack flexibility for implementing alternative ANN architectures.

Brain-inspired neuromorphic hardware presents an alternative to conventional ASIC accelerators, and has been shown to be capable of running ANNs with potentially orders-of-magnitude lower power consumption (that is, performance-per-watt). The landscape of neuromorphic hardware is rapidly evolving[11–16]; however, increasingly these approaches leverage spiking to achieve substantial energy savings. Neuromorphic spiking, which emulates all-or-none action potentials in biological neurons, limits communication in hardware only to discrete events. For spiking neuromorphic hardware to be useful, however, it is necessary to convert an ANN, for which communication between artificial neurons can be high-precision, to a spiking neural network (SNN). Supplementary Note 1 provides further details of spiking and ANN acceleration.

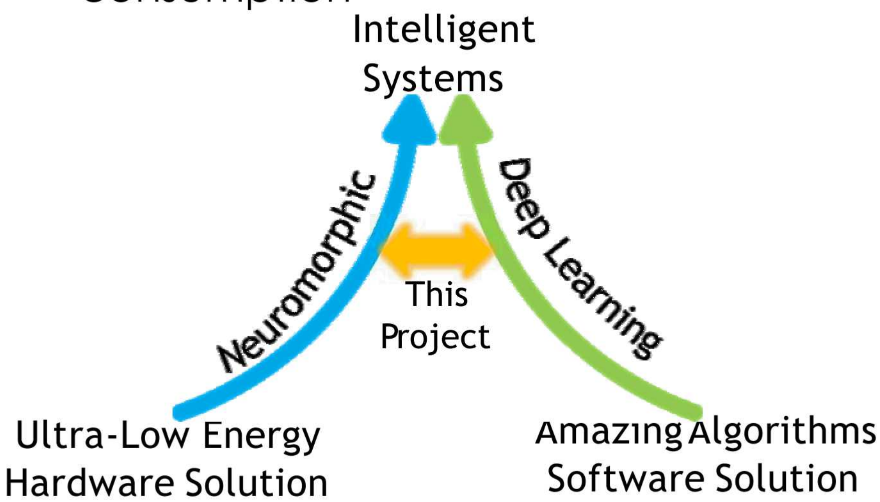
The conversion of ANNs to SNNs—whatever their form—is non-trivial, as ANNs depend on gradient-based backpropagation training algorithms, which require high-precision communication, and the resultant networks effectively assume the persistence of that precision. While there are methods for converting existing ANNs to SNNs, these transformations often require using representations that diminish the benefits of spiking. Here, we describe a new approach to training SNNs, where the ANN training is to not only learn the task, but to produce a SNN in the process. Specifically, if the training procedure can include the eventual objective of low-precision communication between nodes, the training process of a SNN can be nearly as effective as a comparable ANN. This method, which we term Whetstone (Fig. 1) inspired by the tool to sharpen a dull knife, is intentionally agnostic to both the type of ANN being trained and the targeted neuromorphic hardware. Rather, the intent is to provide a straightforward interface for machine learning researchers to leverage the powerful capabilities of low-power neuromorphic hardware on a wide range of deep learning applications (see section 'Implementation and software package details').

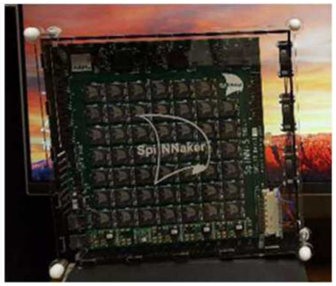
Results

Whetstone method converts general ANNs to SNNs. The Whetstone algorithm operates by incorporating the conversion into binary activations directly into the training process. Because most techniques to train ANNs rely on stochastic gradient descent methods, it is necessary that the activations of neurons be differentiable during the training process. However, as networks become trained, the training process is able to incorporate additional constraints, such as targeting discrete communication between nodes. With this shift of the optimization target in

Center for Computing Research, Sandia National Laboratories, Albuquerque, NM, USA. *e-mail:

- AI power draw is a key limiting factor especially for electric powered vehicles: 3kW now; HPC-level for fully self-driving
- Prototype vehicles use a trunk full of GPUs
- Forecasting current tech ~1TeraOp/Watt
- Neuromorphic Hardware:
  - Enables event-driven computation
  - Opportunity for extremely low power consumption



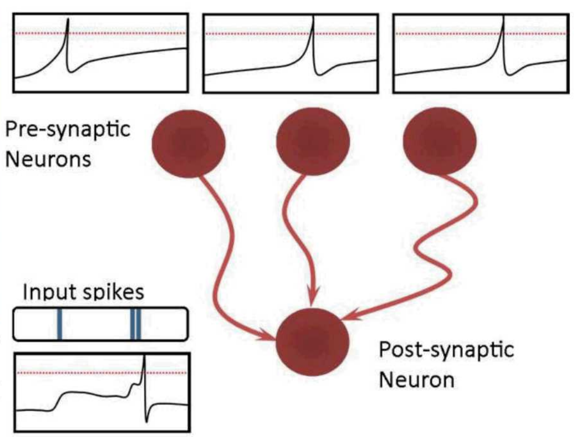


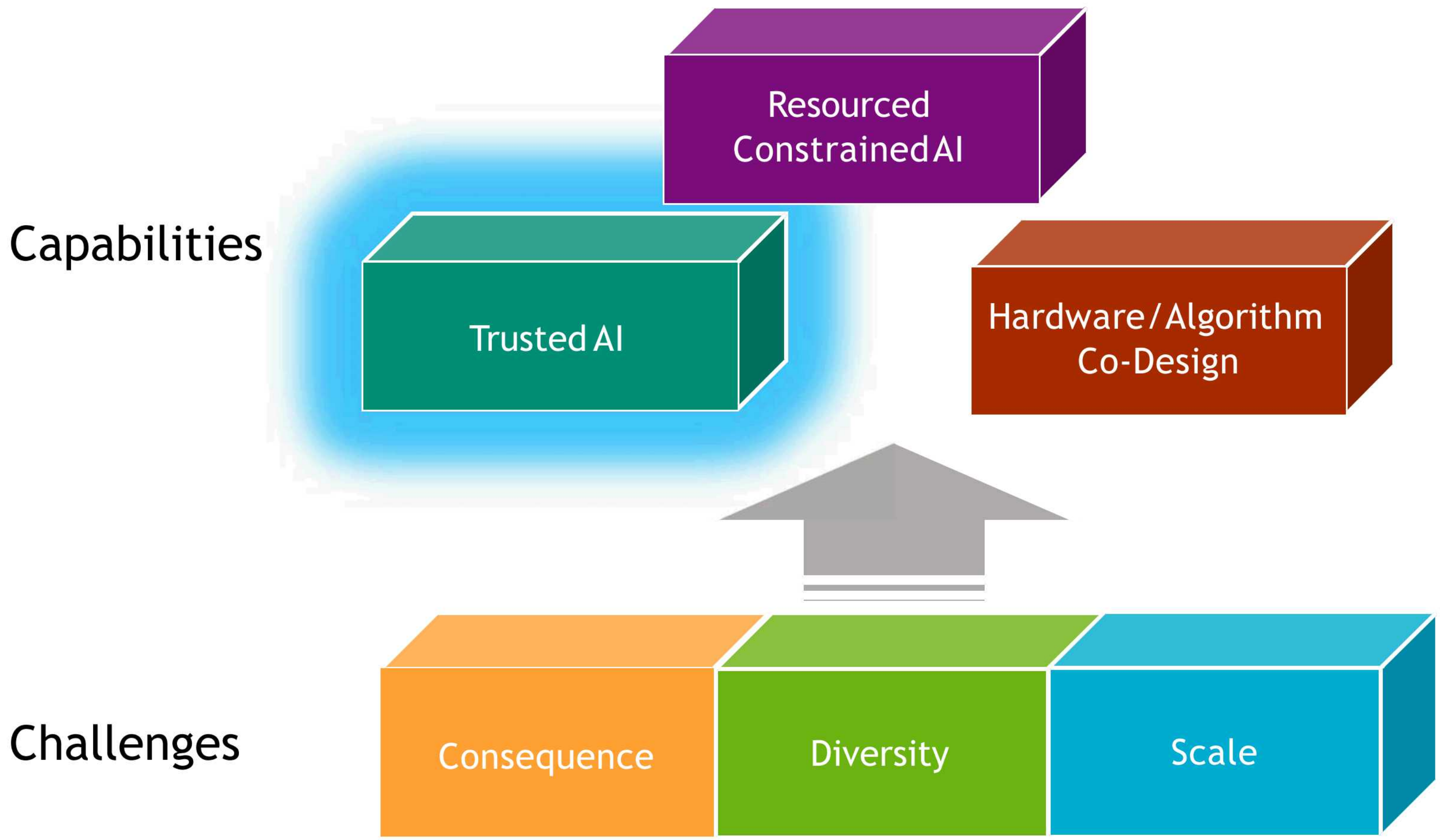
SpiNNaker, Univ. of Manchester

Example Image From Berkeley DeepDrive

Intel Loihi
Photo: intel.com

Spiking Neuron Representation




Capabilities
Resourced Constrained AI
Trusted AI
Hardware/Algorithm Co-Design
Challenges
Consequence
Diversity
Scale

ML for Reduced Order Models

Problem

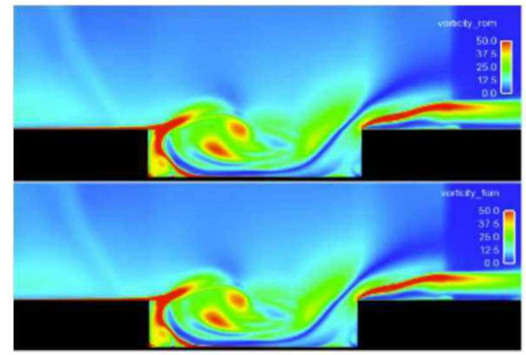
- High-fidelity computational physics simulations on HPC systems can take hours or days to execute
- Lengthy execution time limits the design space explored during conceptual design
- Need a faster, more efficient means of simulating complex physics problems

Technical Approach

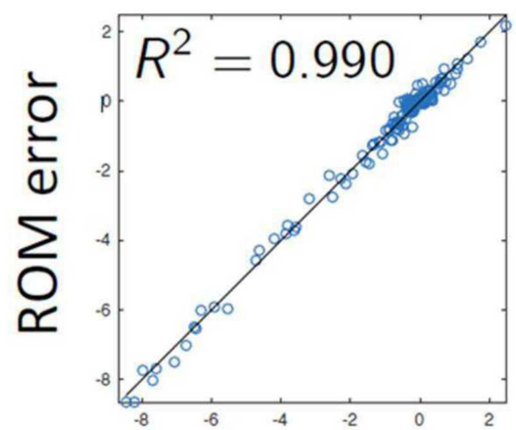
- Create Reduced Order Model (ROM) from high-fidelity simulation data that
  - Executes faster via dimensionality reduction using autoencoders without significant reduction in accuracy
  - Preserves important physical properties (e.g., conservation laws)
  - Uses Machine Learning Error Models (MLEM) to quantify uncertainty

Results/Accomplishments

- Reduced order surrogate models and theory have been developed for turbulent flow simulations
- Runtimes are 100-1000 times faster and are only 1% less accurate than the high-fidelity simulations
- MLEM can predict errors with validated statistical properties



Turbulent flow vorticity field



Problem

- High-fidelity simulations on HPC systems are too expensive

Technical

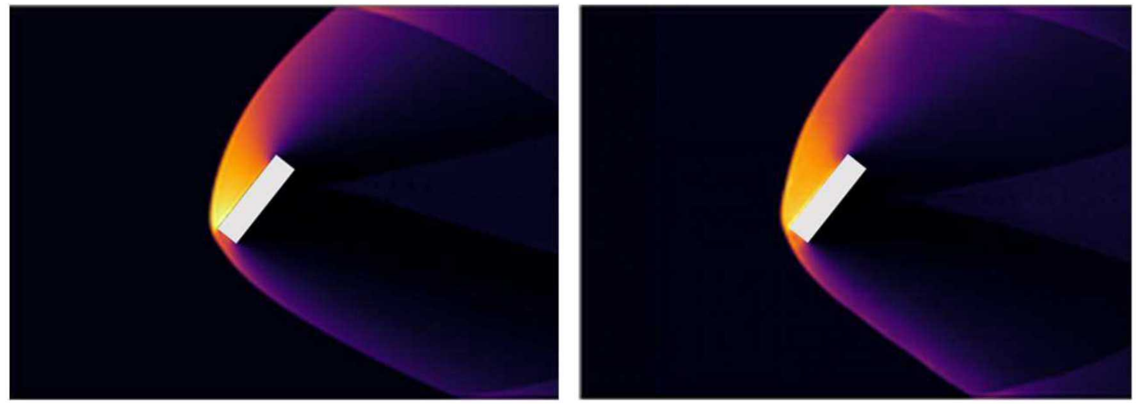
- Train a neural network to predict the steady-state flow field
- Guide the prediction with physical constraints (conservation laws) and aerodynamic forces (drag, lift, torque)

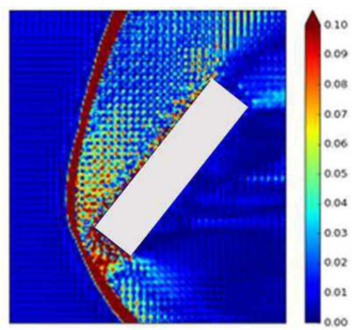
Results/Accomplishments

- Demonstrated >100x speed increase in 2D with < 6% average error
- Predict > 1000x speed increase in 3D

Hydro-code Simulation | ML Prediction

Pressure

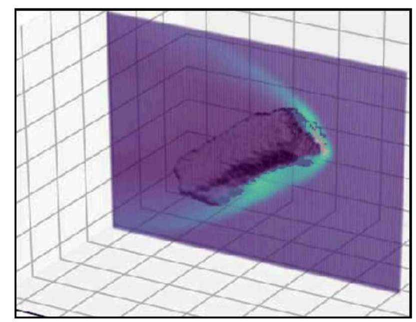




Relative error map of ML prediction

2D force	Avg Error
Drag	1.87%
Lift	5.63%
Torque	2.29%

ML model successfully predicts flow field and aerodynamic coefficients



ML prediction of pressure field around complex 3D object

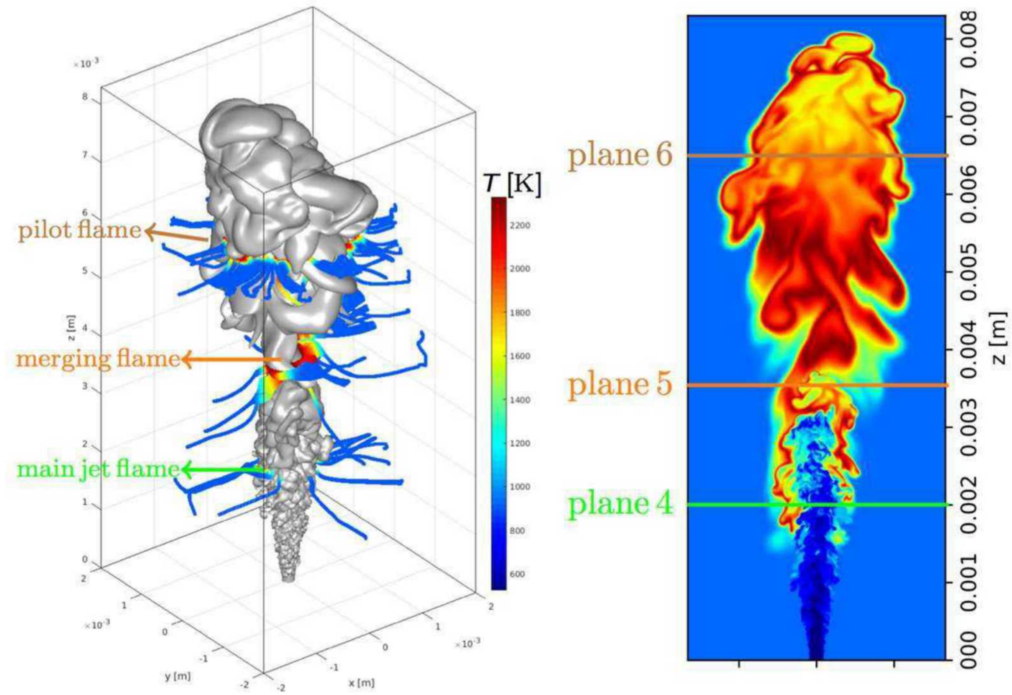
Surrogate Reduced Order Modeling of Diesel Combustion and Ignition in GDI engines using ML/AI

- Surrogate reduced-order modeling with principal component transport of compositions to reduce the chemical dimensions needed to describe low- and high temperature ignition, flame propagation and soot under diesel and cold-start GDI conditions

- Use petascale DNS data and experiments to provide 'truth' data to adaptively train reduced-order surrogates incorporating physics constraints (e.g. using governing equations)

- Anomaly detection ML for detecting pre-ignition and knock

- Reduced order modeling for engine design and optimization



DNS of n-dodecane multi-injection diesel combustion showing instantaneous volume rendering of mixture fraction (left image) and temperature slice (right image). Multi-dimensional flamelets at 3 axial planes shown by blue lines (Rieth, Chen, Xu, Han, Hasse)

Machine Learning Enables Automatic Mesh Generation

Problem

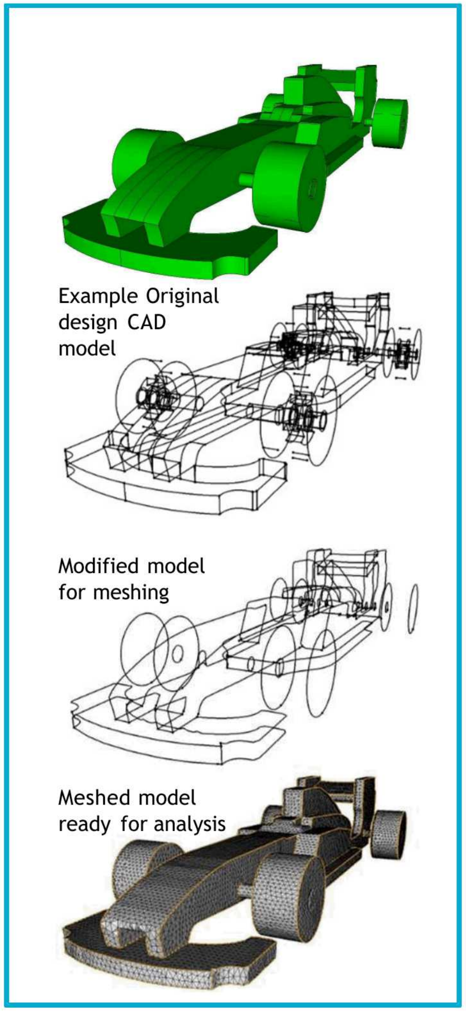
- Geometry preparation and meshing for computational simulation is bottleneck (consuming 70%+ of analyst time)
- Analyst/engineer must have extensive domain-specific expertise to manage many individual complex problems and tasks
- Must produce verifiably accurate physics appropriate mesh ready for simulation

Technical Approach

- Identify tasks currently done by analysts to train machine learning models
- Capture and label operations performed by expert using existing software
- Build a feature library of geometric characteristics commonly encountered in CAD models and identify solutions for effectively modifying CAD for best resulting mesh
- Explore machine learning models that provide best solutions for CAD features with associated solution labels

Results/Accomplishments

- Developed ML techniques to rank geometry-modification operations by their likelihood of yielding a meshable model
- Provides insight on which geometric features are most useful for machine leaning, and would be relatively easy to integrate into the analyst workflow if successful



Problem

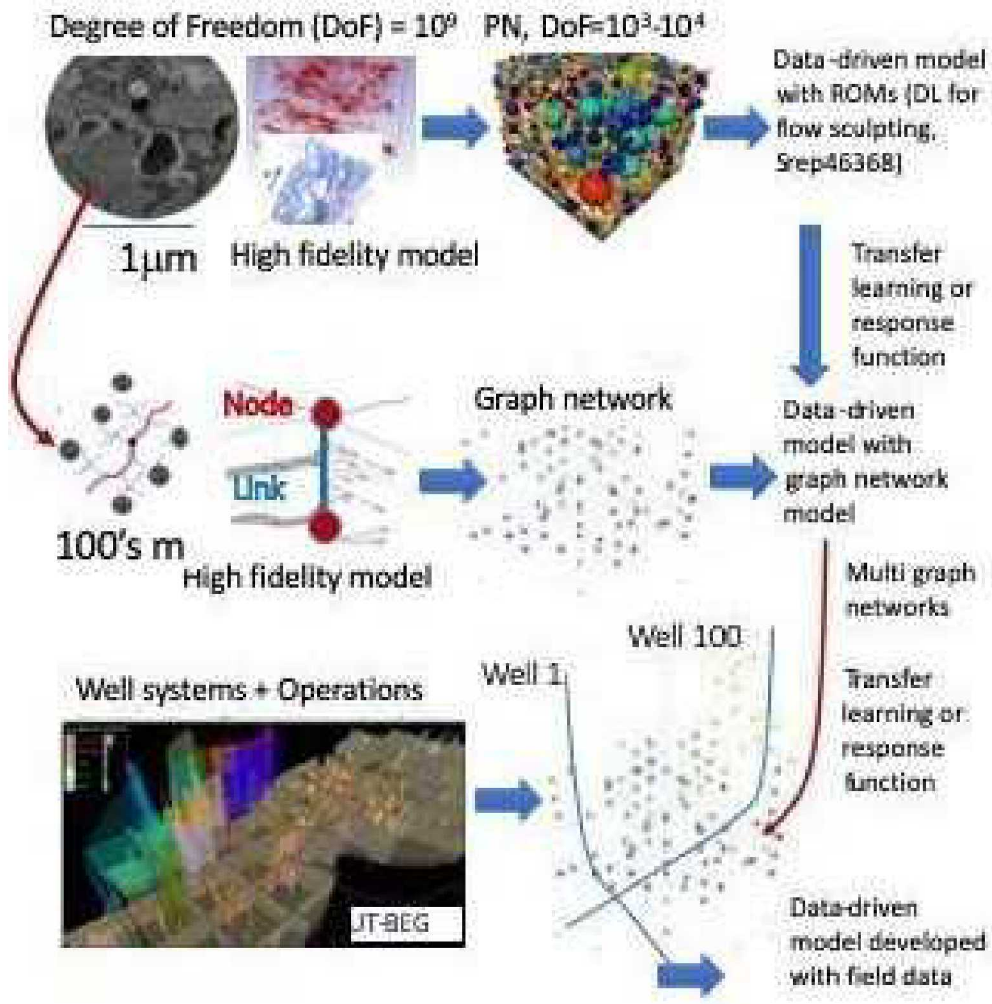
- Given sparse data at various scales (nm, m, km) about rocks and wells, and a given depth, can a machine learning algorithm predict how much shale gas can be extracted?

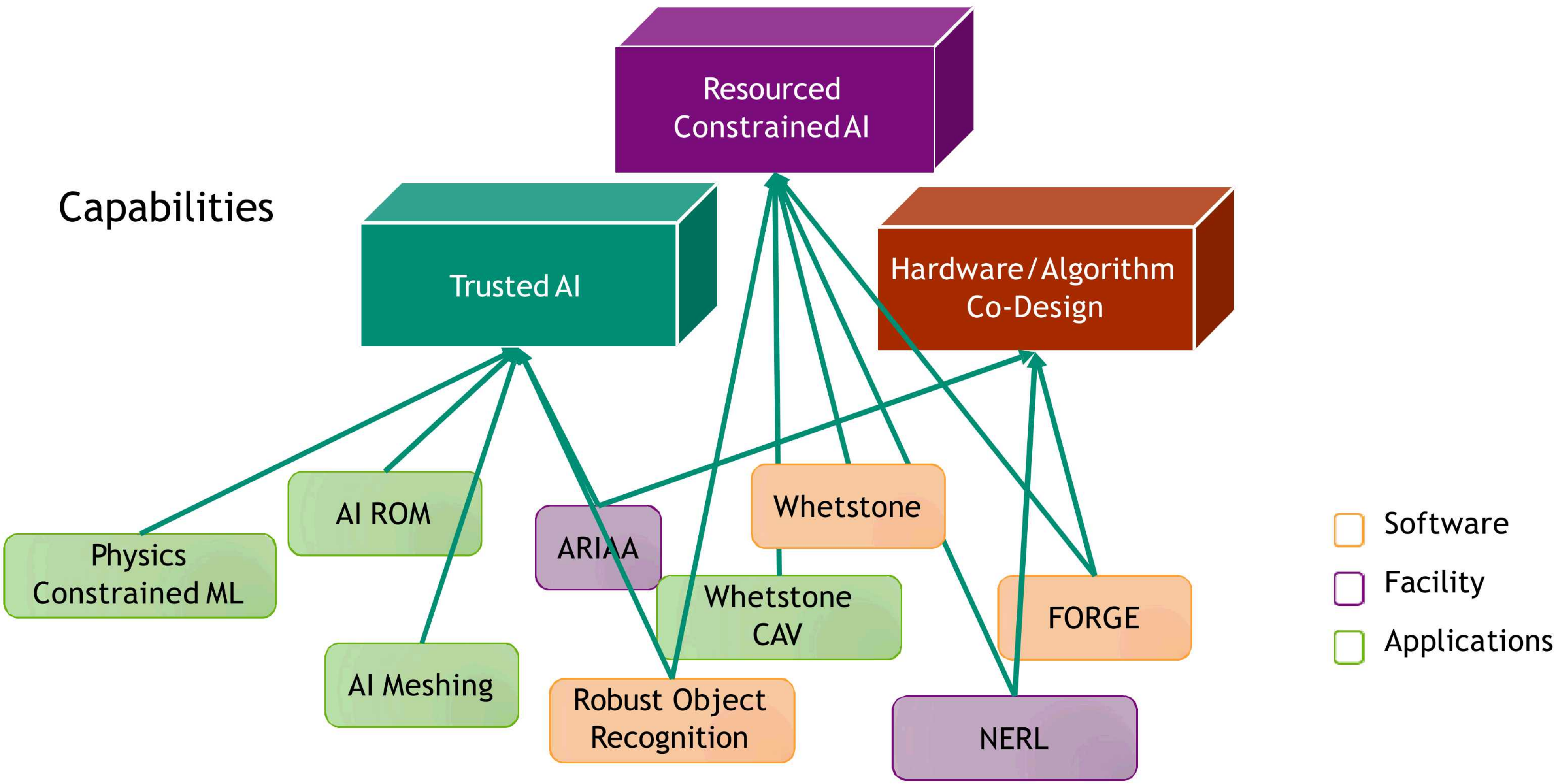
Technical Approach

- Develop multiscale, physics informed deep learning algorithm to generate, parse, and predict data to solve the above problem
- Develop physics informed costs and constraints driven algorithms

Results/Impact

- Improve current state of art predictions and resource estimates
- Develop physics informed machine learning algorithms




Resourced Constrained AI
Capabilities
Trusted AI
Hardware/Algorithm Co-Design
Physics Constrained ML
AI ROM
ARIAA
Whetstone
Whetstone CAV
FORGE
AI Meshing
Robust Object Recognition
NERL
Software
Facility
Applications

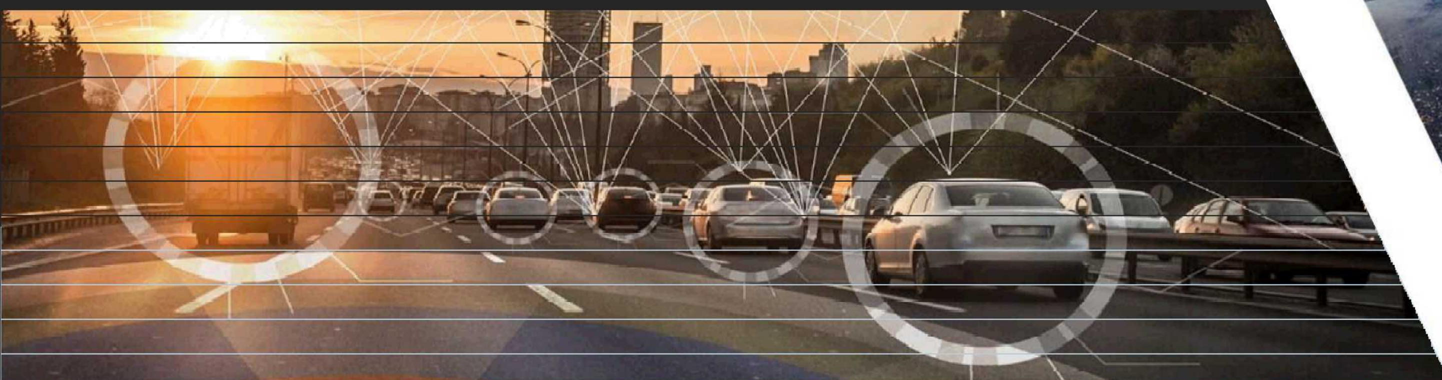
Thanks! Questions?

Exceptional service in the national interest

Sandia National Laboratories

Sandia National Laboratories

Exceptional service in the national interest

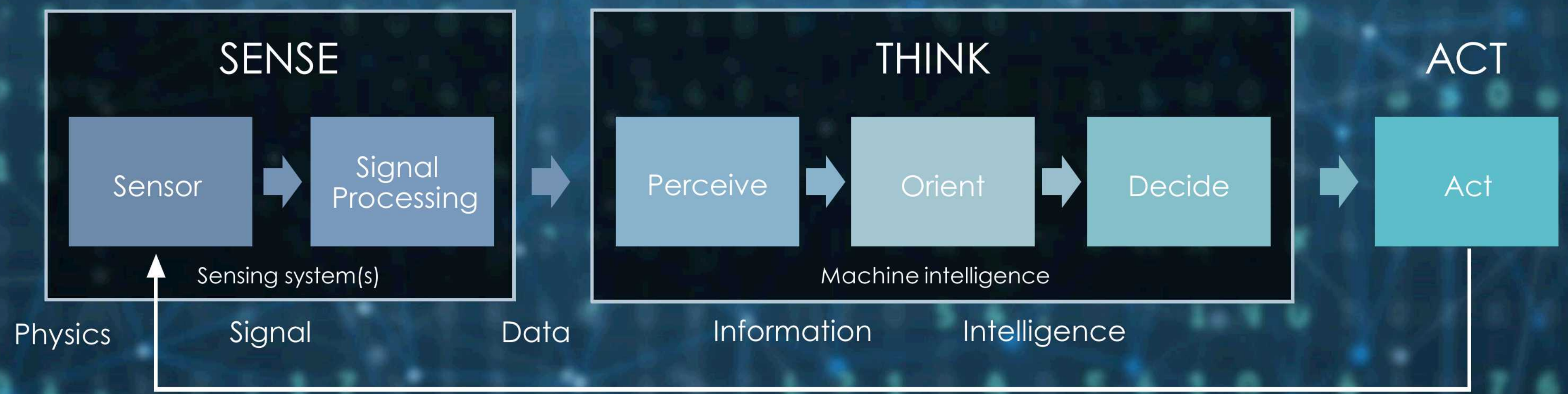


Advanced Computing for Connected & Automated Vehicle (CAV)

U.S. DEPARTMENT OF ENERGY NNSA

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.

AUTOMATED SYSTEMS



FOCUSED RESEARCH AREAS FOR HIGHLY AUTOMATED VEHICLES

Sensors and Sensor Processing

- Create disruptive optical sensing technology to reduce energy consumption by 100X
- Develop chip-scale LiDAR to reduce cost by 100X

Scene Perception and Algorithms

- Explore sparse coding and reduced-precision to reduce computation load by 1000X
- Develop biologically inspired machine learning algorithms to reduce the number of training samples by 100X
- Develop unsupervised and self-supervised learning algorithms

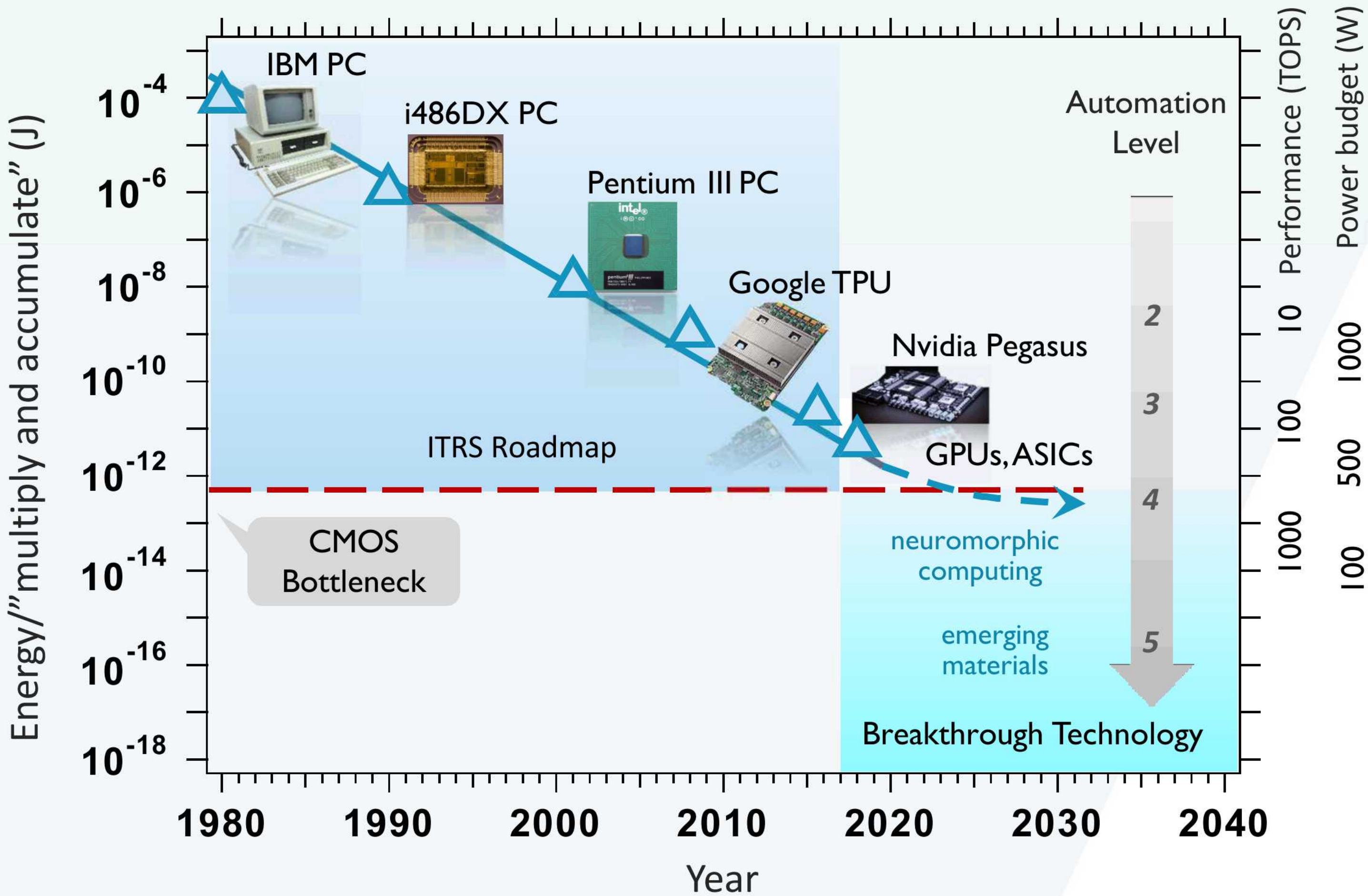
Navigation Hardware Accelerators

- Develop and demonstrate hardware capable of real-time processing of tera- to petabit inputs, with energies at <10 fJ per operation (>100 TOPS/W)
- Enable algorithms for robust and reliable recognition tasks needed for perception.
- Demonstrate the value of algorithm and hardware co-design such that combined elements have greater energy and/or SWaP improvement

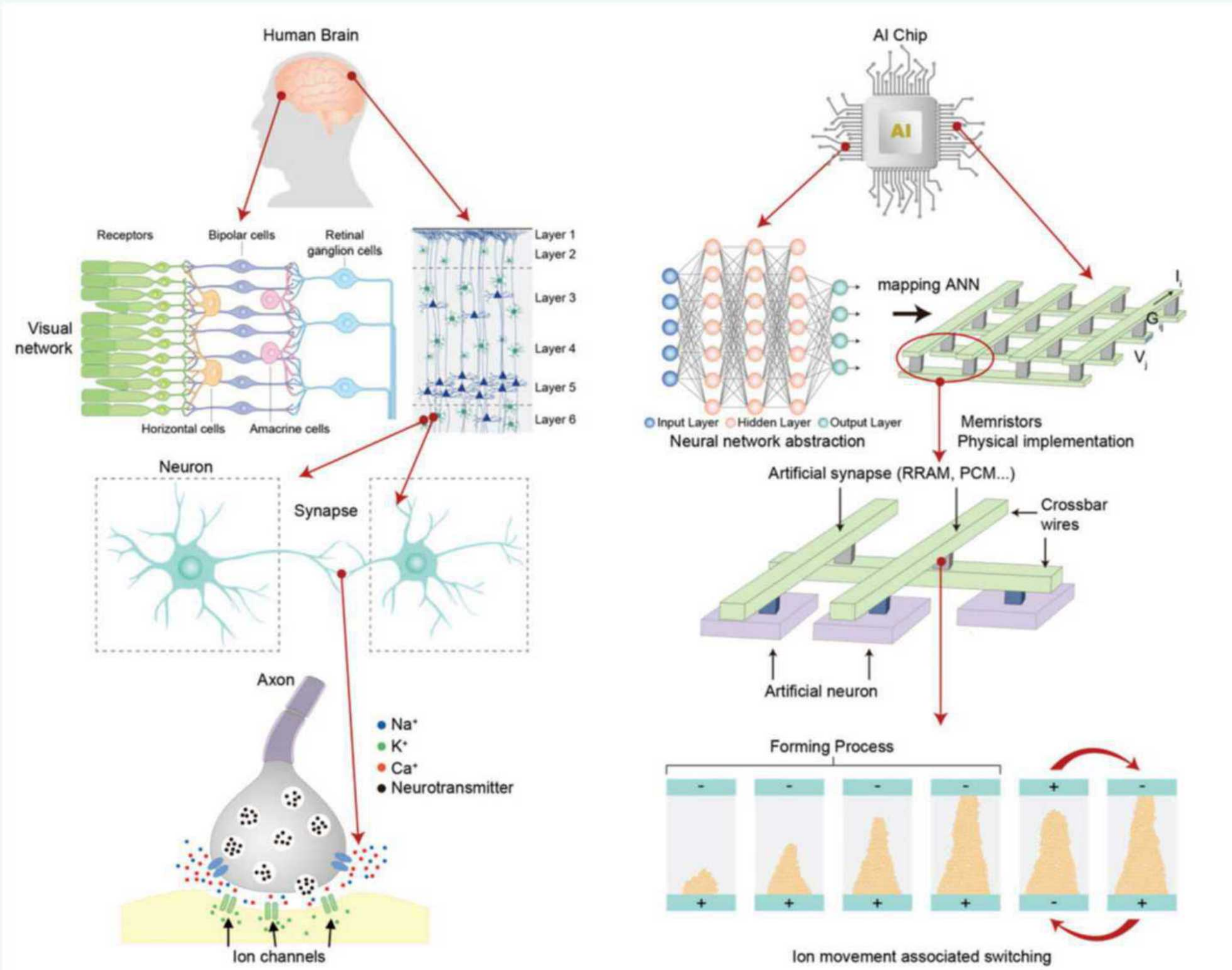
ENERGY EFFICIENCY COMPUTING AND THE NEED OF CAV

1pJ/op = 1 TOPS/W

1fJ/op = 1000 TOPS/W



LEARNING FROM BRAIN FOR ULTIMATE POWER EFFICIENT COMPUTE



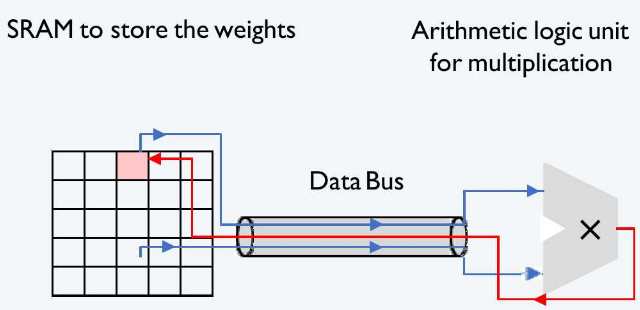
- Highly parallel neuron net
- Spiking – Event based computing
- Low voltage – 100mV
- Synapse – plasticity
- **In memory computing with NVM (synapse)**
- **Matrix operation**

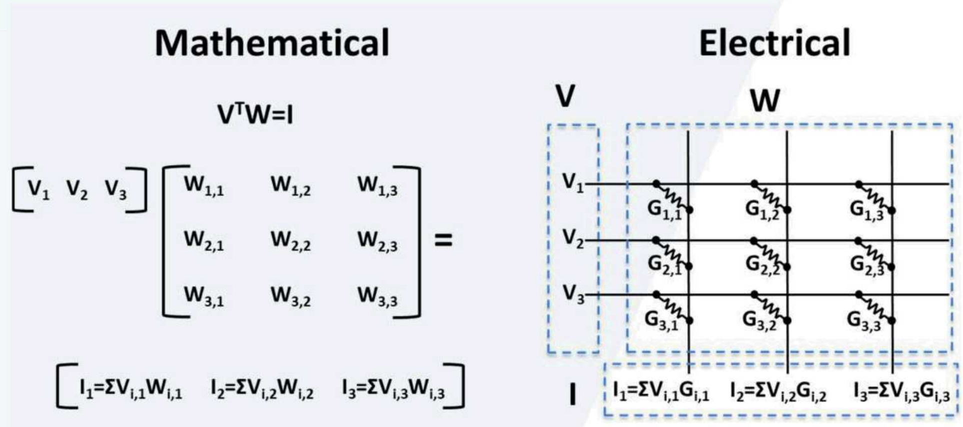
Tang, et al, Adv. Mat. 2019, 1902761

BREAKING VON NEUMANN BOTTLENECK – UNLEASH 1000X POWER PERFORMANCE with NVM CROSSBAR

Von Neumann Digital
Separate logic and memory structures

In-memory Parallel Analog
Use non-volatile memory





Conductance of each element can be changed in a predictable manner

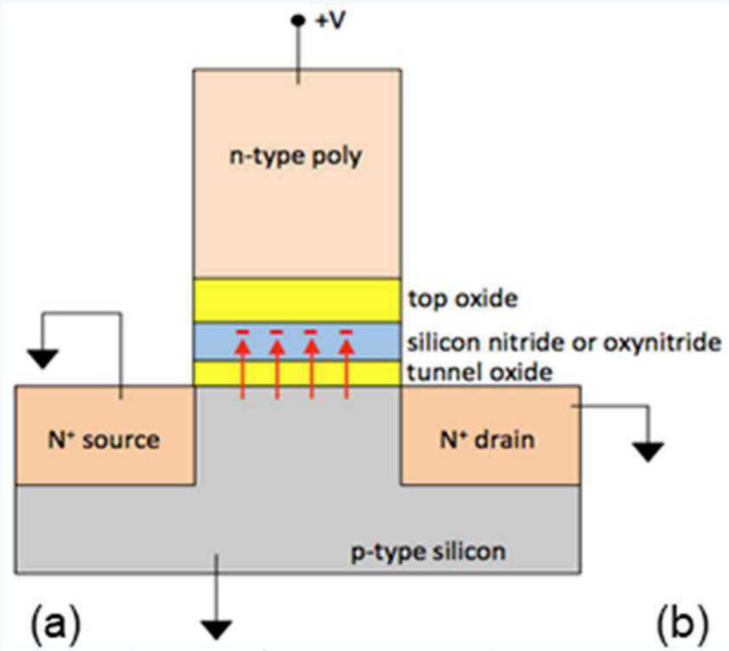
Uses established CMOS technology
Data bus results in latency and power

Simultaneous logic and memory
3 orders of magnitude less power

M. Marinella, *IEEE Circuits and Systems*, 8, 86-101, 2018
Zidan, Strachan, & Lu, *Nat. Elec.* 1, 22, 2018

Emerging on-chip non-volatile memory (NVM) improves energy-efficiency by performing analog multiply-accumulate inside memory and eliminate data movement

NEAR-TERM TECHNOLOGY – 10TOPS/W for DNN accelerator



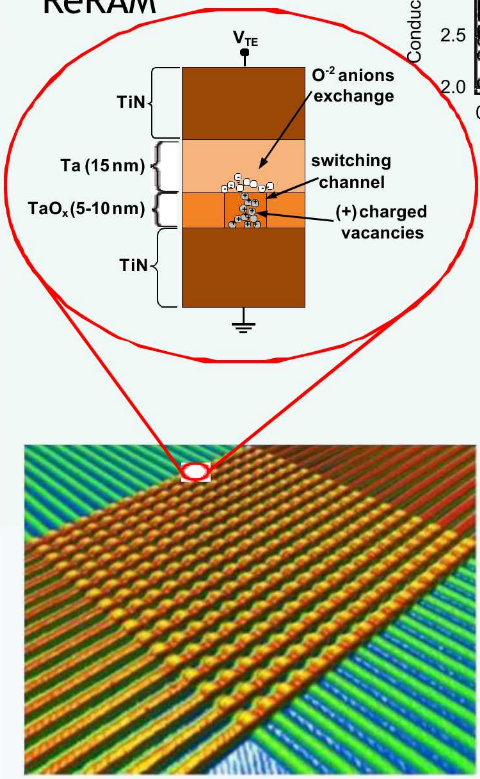
Component	Vector Matrix Multiply	Matrix Vector Multiply	Outer Product Update
Energy/Op SONOS (fJ)	13.7	13.7	68.2
Energy/Op SRAM (fJ)	2718	4630	4102
Array Latency SONOS (μs)	0.40	0.40	20
Array Latency SRAM (μs)	4	32	8

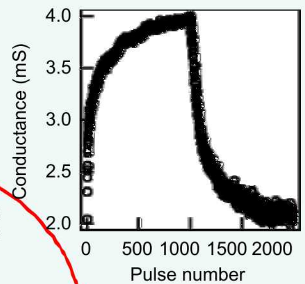
Silicon-Oxide-Nitride-Oxide-Silicon(SONOS)

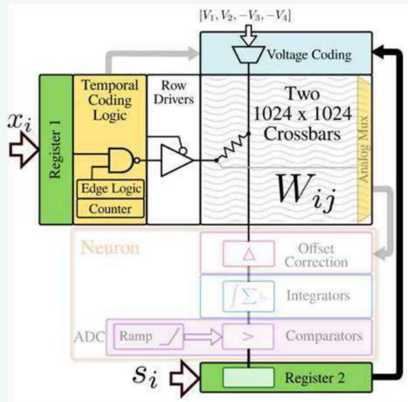
S Agarwal et al, IEEE J Exploratory Solid-State Computational Devices and Circuits, 5, 52-57, 2019.

ANALOG RERAM CROSSBAR – towards 100TOPS/W for DNN accelerator

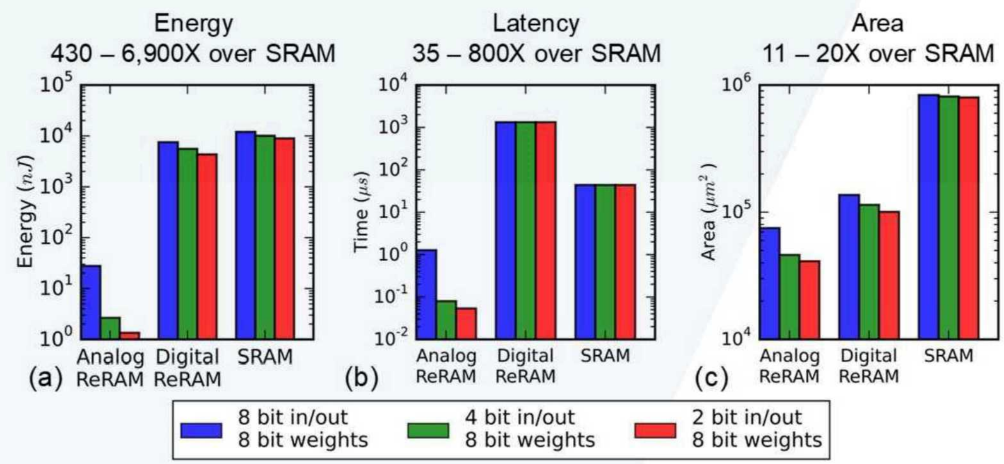
Memristor ReRAM







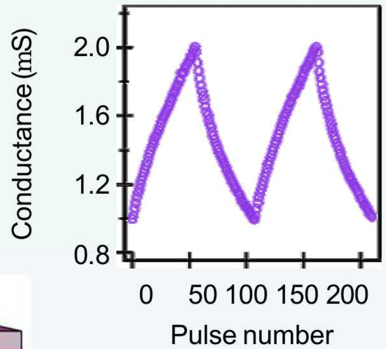
Component	Vector Matrix Multiply	Matrix Vector Multiply	Outer Product Update
Energy/Op ReRAM (fJ)	12.2	12.2	2.1
Energy/Op SRAM (fJ)	2718	4630	4102
Array Latency ReRAM (µs)	0.38	0.38	0.51
Array Latency SRAM (µs)	4	32	8

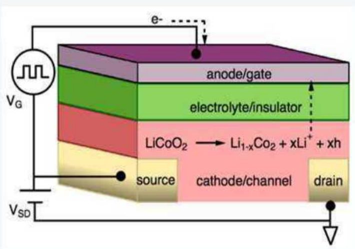


Marinella, Agarwal, et al, *IEEE JETCAS*, 2018

Agarwal, et al, *IEEE E3S Symp*, 2017

ION TUNABLE ELECTRONIC MATERIALS – beyond 100TOPS/W





nature materials LETTERS
PUBLISHED ONLINE: 20 FEBRUARY 2017 | DOI: 10.1038/NMAT4856

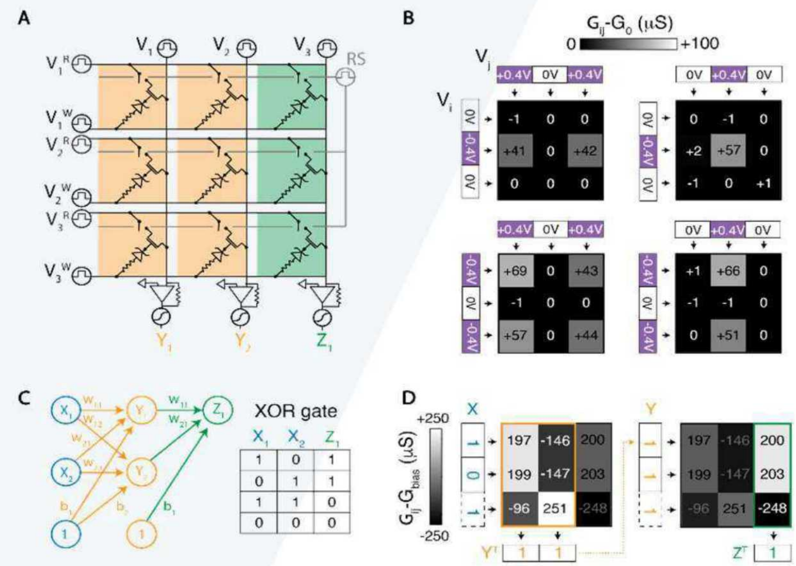
A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing

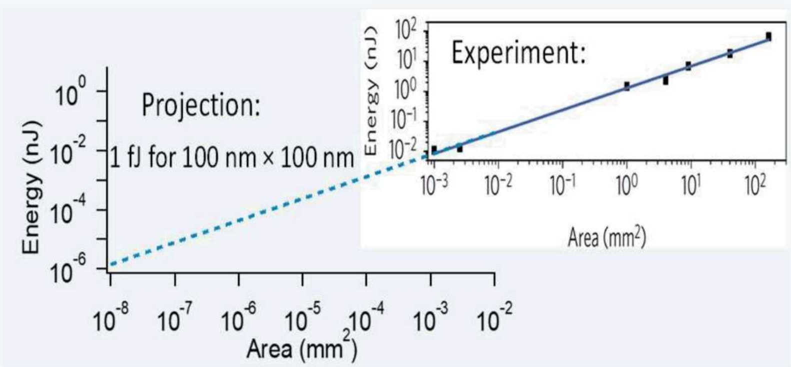
Yoeri van de Burgt[1†‡], Ewout Lubberman[1,2‡], Elliot J. Fuller[3], Scott T. Keene[1], Grégorio C. Faria[1,4], Sapan Agarwal[3], Matthew J. Marinella[5], A. Alec Talin[3*] and Alberto Salleo[1*]

RESEARCH
DEVICE TECHNOLOGY
Science

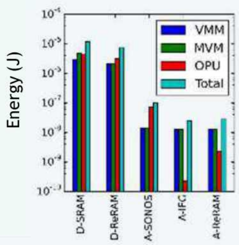
Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing

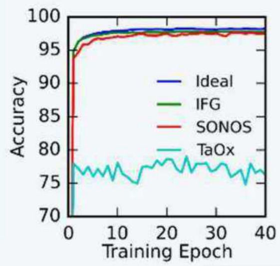
Elliot J. Fuller[1], Scott T. Keene[2*], Armantas Melianas[2*], Zhongrui Wang[3], Sapan Agarwal[1], Yiyang Li[1], Yaakov Tuchman[2], Conrad D. James[4], Matthew J. Marinella[4], J. Joshua Yang[3], Alberto Salleo[2]†, A. Alec Talin[1]†



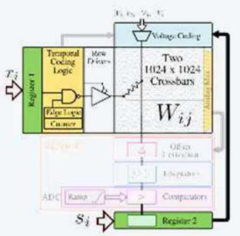


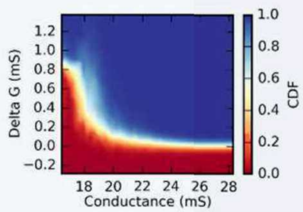
MULTISCALE CODESIGN FOR NEUROMORPHIC ACCELERATOR





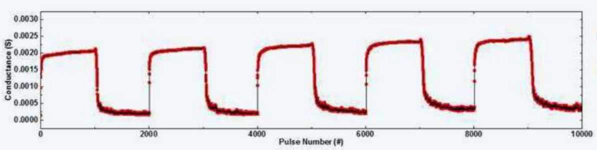
Energy/Performance Model
Model performance and energy requirements



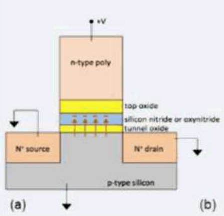


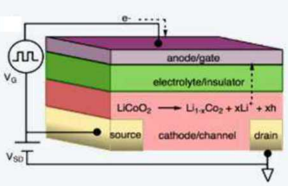
CROSS SIM

Sandia Cross-Sim:
Translates device measurements and crossbar circuits to algorithm-level performance



Analog characterization





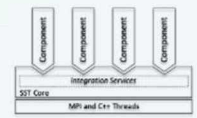
Algorithms

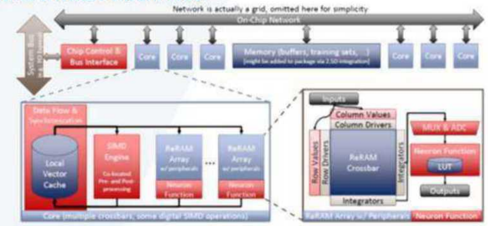
Target Algorithms

- Deep Learning
- Sparse Coding
- Liquid State Machines

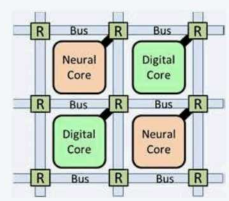
Architecture

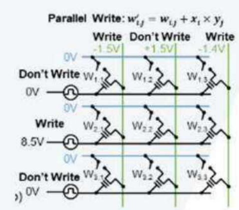
Architecture Simulation





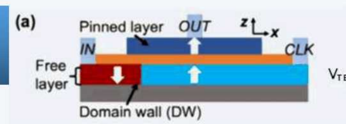
Circuits

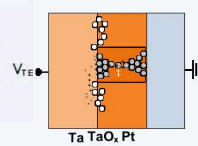


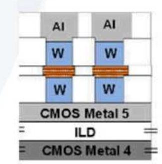


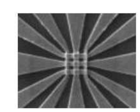
Devices

Drift-diffusion model of transport



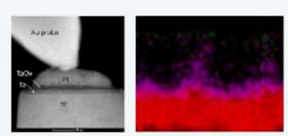


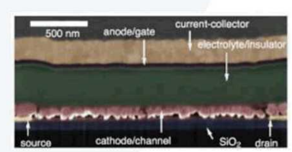




Materials

In situ Characterization





Ab Initio Modeling

COMPUTING TECHNOLOGY ROADMAPPING

Compute perf. target	0.1-1 TOPS/W	1-5 TOPS/W	10 TOPS/W	100 TOPS/W
	Today	**2025**		**2035**

Road-mapping

Gov't: DOE, DOT, etc.
Academic: National Labs, Research institutes
Industry: GM, Tier1, chip suppliers

- Identify the technology gaps
- Validate the target requirement for CAV
- Define the roadmap for the technology injection points

Thank you!

Q&A

Contact: Zhiyong Li
zli@sandia.gov