U.S. AIR FORCE

USSF

AFRL

Nanocomputing, Neuromorphic Computing and Quantum Information Science Technologies at the AFRL Information Directorate

Dr. Michael Hayduk

Deputy Director, Information Directorate

August 2021

AGENDA

- Introduction to AFRL Information Directorate
- Nanocomputing
- Neuromorphic Computing
- Quantum Information Science (QIS)
- Summary

Acknowledgements

- AFRL Team:
  - Dr. Qing Wu
  - Dr. Joe VanNostrand
  - Dr. Kathy-Anne Soderberg
  - Ms Laura Wessing
  - Dr. Michael Fanto
  - Dr. Matthew LaHaye
  - Dr. Paul Alsing

Information Directorate Mission & Vision

MISSION:

To explore, prototype, and demonstrate high-impact, game changing technologies that enable the Air Force and Nation to maintain its superior technical advantage.

VISION:

To lead the Air Force and Nation in command, control, communications, computers, and intelligence (C4I) and cyber science, technology, research and development.

ROME = C^4I&Cyber

Rome

65 Acre Campus
30 Laboratories &
Facilities
882,000 sq ft floor space

Newport

Primary mission: To evaluate antenna performance on full scale aircraft and make recommendations for improvement.

Stockbridge

Used for development of and real world experimentation with advanced radio frequency (RF)/optical communications, networking and information technologies, cyber techniques and effects, including small unmanned aircraft systems.

Extreme Computing Facility

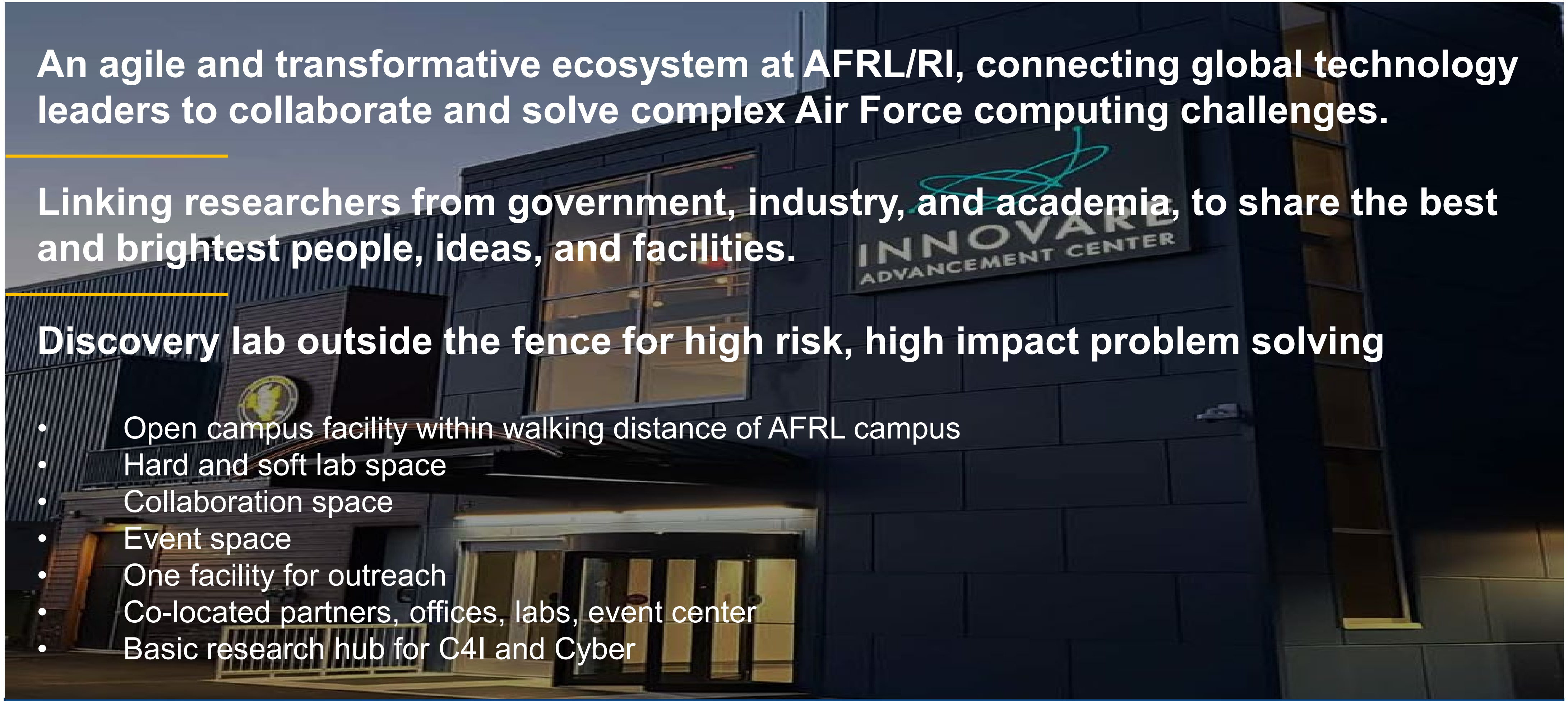
Current Von Neumann computing architectures are inefficient and do not scale

Foundational advances in computing architectures

- Quantum
- Neuromorphic
- Nanoelectronic
- Machine Learning
- Artificial Intelligence

Innovare Advancement Center

Agility + Innovation + Partnerships



An agile and transformative ecosystem at AFRL/RI, connecting global technology leaders to collaborate and solve complex Air Force computing challenges.

Linking researchers from government, industry, and academia, to share the best and brightest people, ideas, and facilities.

Discovery lab outside the fence for high risk, high impact problem solving

- Open campus facility within walking distance of AFRL campus
- Hard and soft lab space
- Collaboration space
- Event space
- One facility for outreach
- Co-located partners, offices, labs, event center
- Basic research hub for C4I and Cyber

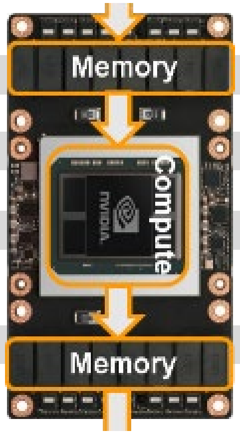
S-UAS Testing | Quantum Facilities | Neuromorphic Computing Facilities

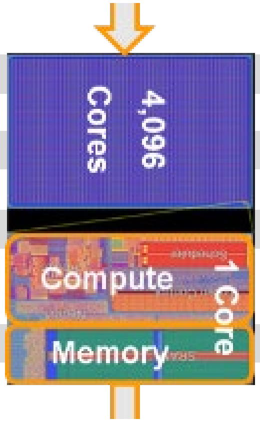
Neuromorphic Computing

Brain-Inspired, Extremely Low SWaP, Intelligent Computing For Deploying Artificial Intelligence and Machine Learning Capabilities

Von Neumann Computer

- Good for precision
- **Operates at very high speeds (> GHz)**
- Synchronous (high idle power)
- Noise-free high-precision, sometimes fragile
- Binary encoding (energy-hungry)
- Very limited connectivity (1:10)
- **Separated memory-computation (memory bottleneck, high I/O power)**
- Dense and modular processing
- **Needs explicit programming for function**





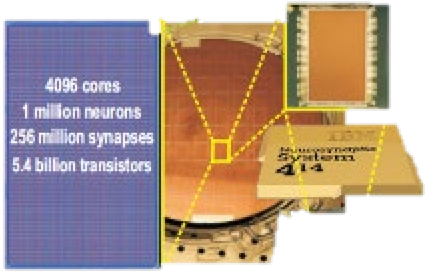
Neuromorphic Computer

- Good for learning and inference
- **Operates at low speeds (10-100Hz)**
- Asynchronous (low idle power)
- Noisy low-precision, yet reliable
- Spike encoding (energy-efficient)
- Very large connectivity (1:10,000)
- **Unified memory-computation (no memory bottleneck, low I/O power)**
- Sparse and distributed processing
- **Learning from data and environment**

Neuromorphic Processor

Brain-Inspired Architectures

Unconventional computer hardware inspired by the architecture and working mechanisms of human brain.

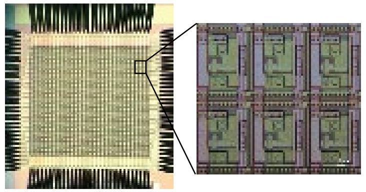


IBM's TrueNorth Processor

Nano Electronics & Optics

Novel Materials, Devices & Circuits

Electrical or optical devices with at least one dimension sized from 10^{-9} to 10^{-7} meters (1~100 nm).

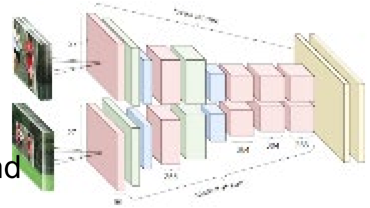


An Array of Memristor Devices As Artificial Neurons & Synapses

Machine Learning

Models & Algorithms

A type of artificial intelligence that learns from data and executes cognitive functions.

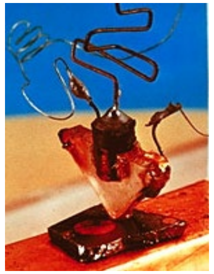


A Deep Convolutional Neural Network

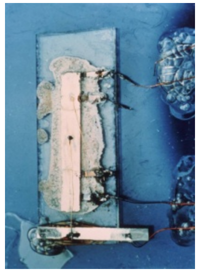
A brief history of electronic computing

60 years

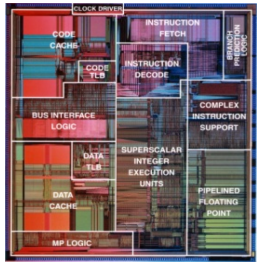
Transistor



IC

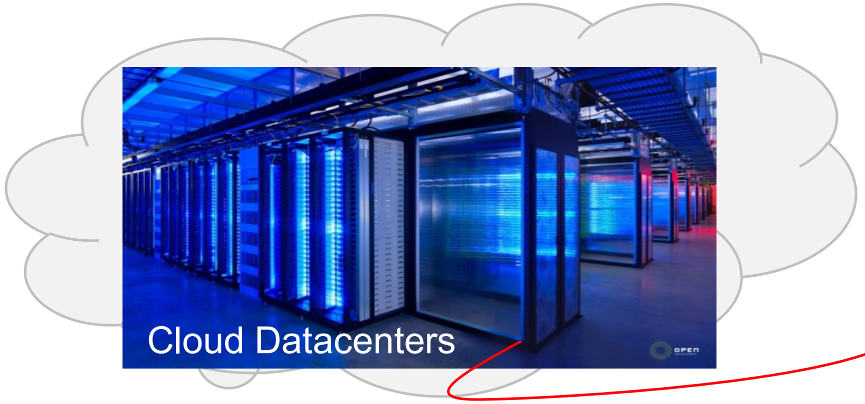


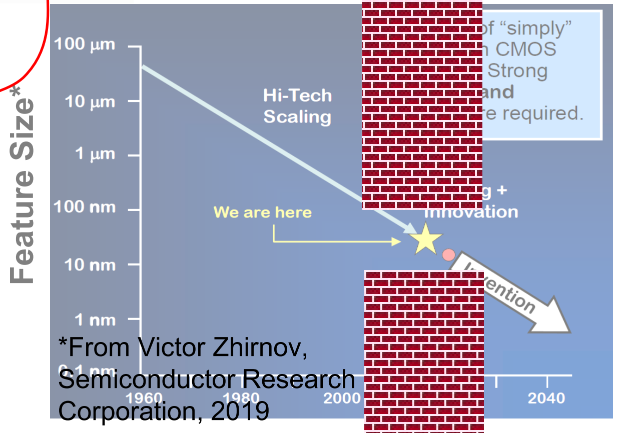
µProcessor

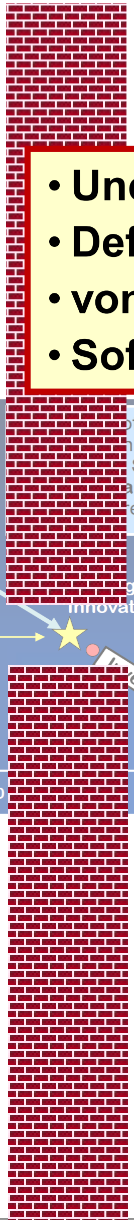


GPU

Cloud-based AI







- **Uncertain future of scaling**
- **Defect intolerant**
- **von Neumann bottleneck**
- **Software-limited**

All this technology, but path to biologically competitive intelligence isn't clear!

50 years

Classical (von Neumann) Architectures

CPU GPGPU Many-Core 3D Integration

40 years

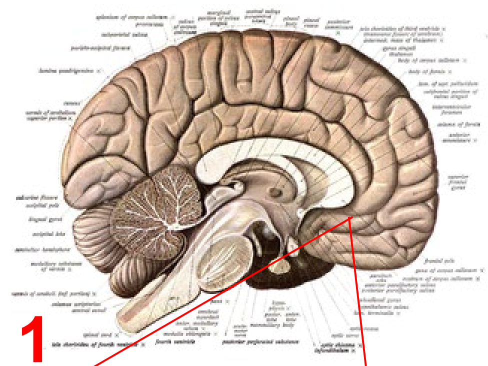
Field Programmable Gate Array (FPGA)

Digital System-on-a-Chip Analog-Hybrid

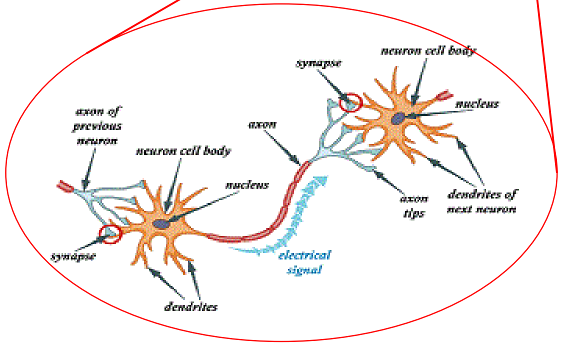
Memristors :
A new pathway to efficient AI HW

- **<u>Need</u>: new devices and architectures**
  - **to overcome limitations of Von-Neumann computing in AI**
  - **while maintaining a low size, weight, and power (SWaP) envelope**

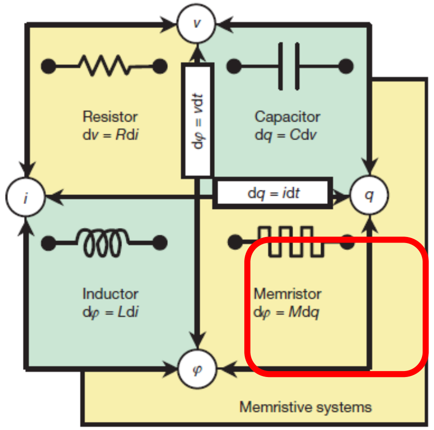
vs.



~ 10^4 : 1
SWaP ratio

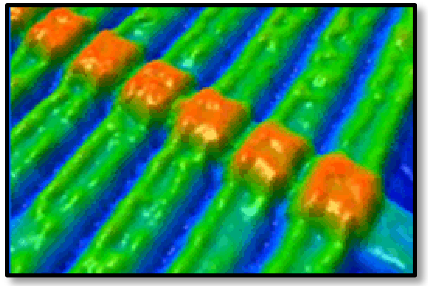


- **Memristor = "resistor with memory"**
  - **Key attribute: hysteresis in I-V curve (resistance switching)**
  - **Volatile or nonvolatile**
  - **Analog resistance states**
  - **Existence postulated in 1971, confirmed in 2008**



Strukov et al, Nature v. 453

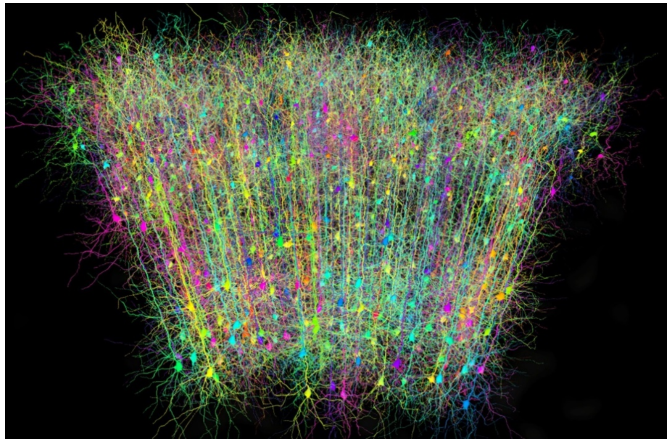
- **Structure**
  - **2-terminal MIM (x-bar)**
  - **Large variety of memristive films**
  - **Substrate-independent / stackable**



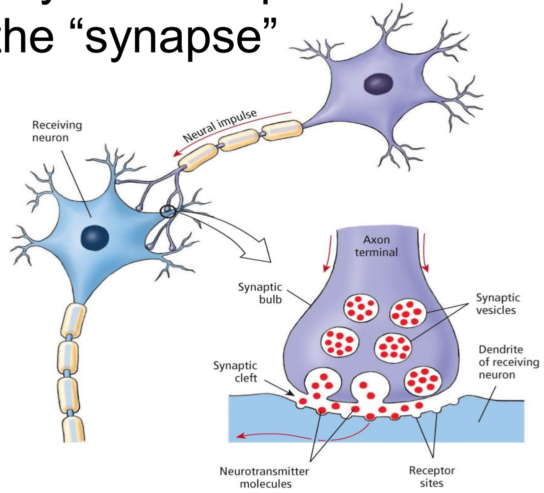
Replicating the power of the brain in electronics

USSF AFRL

The brain is massively parallel and highly connected, enabled by ~10^{11} neurons that have >10^{15} connections



Perhaps the most crucial component for memory and computation in the brain is the “synapse”



http://www.biochemden.com/neurotransmitters-neuropepptides/

Biology:

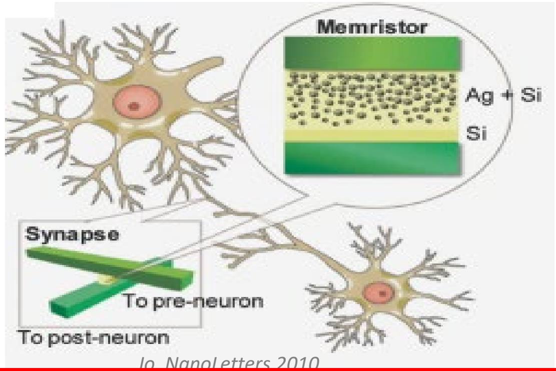
- Not logical “1” or “0”
- Changes dynamically (learning)
- Occurs physically (ionic motion)

To implement in CMOS (TN, Loihi etc.):

- **1 synapse requires > 50 transistors and multiple passive elements**
- **1 neuron ≈ 10^6 transistors**

“Biological” intelligence using CMOS elements is SWaP-prohibitive.

A potential solution is to use dynamic nano-electronic devices such as **memristors**.



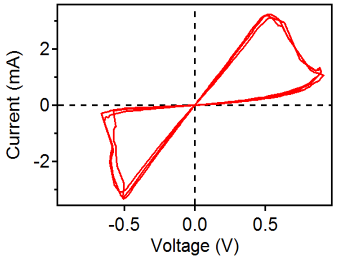
1 synapse = 1 memristor,

1 neuron = 2 memristors

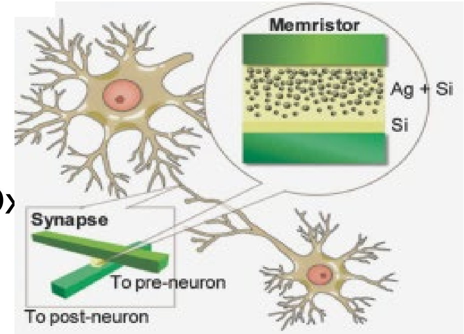
Value-Proposition of NeuroPipe: SWaP-efficient nanoelectronic AI HW

Memristors in Neuromorphic Computing

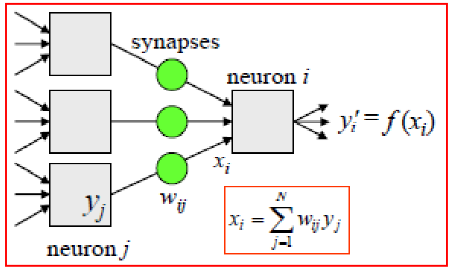
Synapses – non-volatile analog type



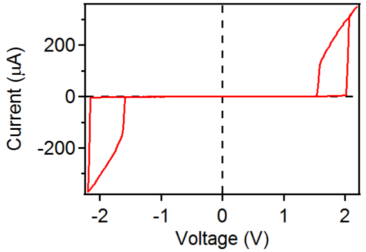
- Memory with plasticity
- "RRAM" - HfO_2, TiOx, TaO_x
- Often filamentary



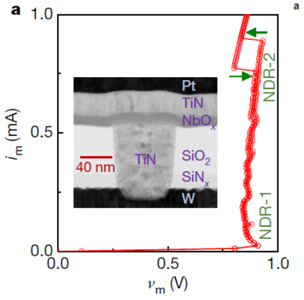
Jo, NanoLetters 2010



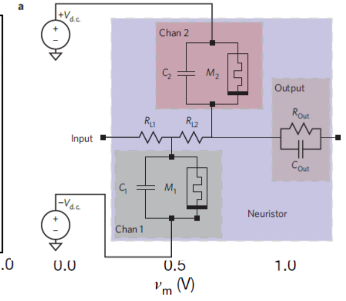
Soma/Axon – volatile, threshold type



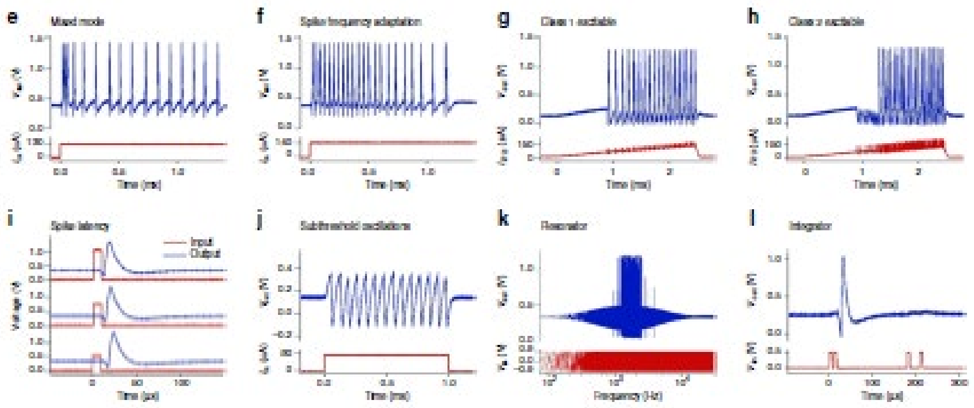
- NbO_2, VO_2
- Undergoes S-type Negative Differential Resistance (NDR), via Mott transition



Kumar et al, *Nature* 548, 2017

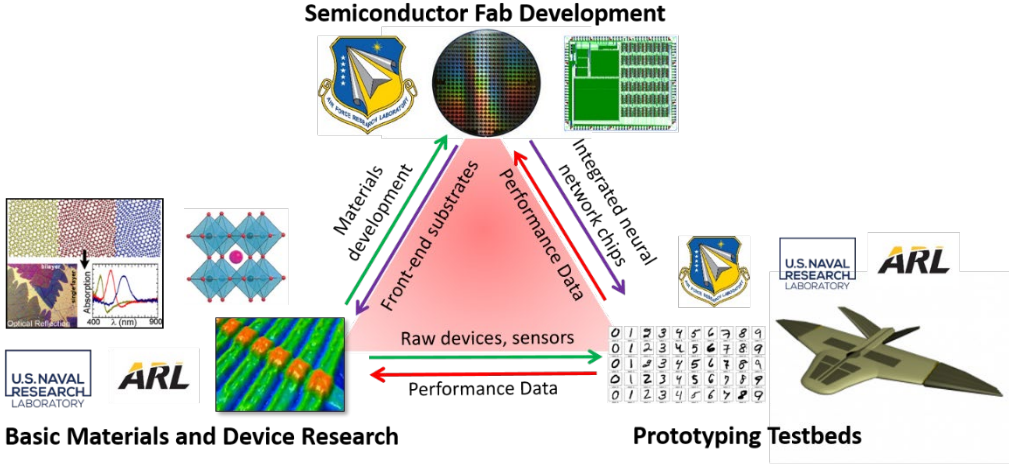


Pickett et al, *Nature Materials* vol. 12, 2012



W. Yi et al, *Nature Communications* (2018)

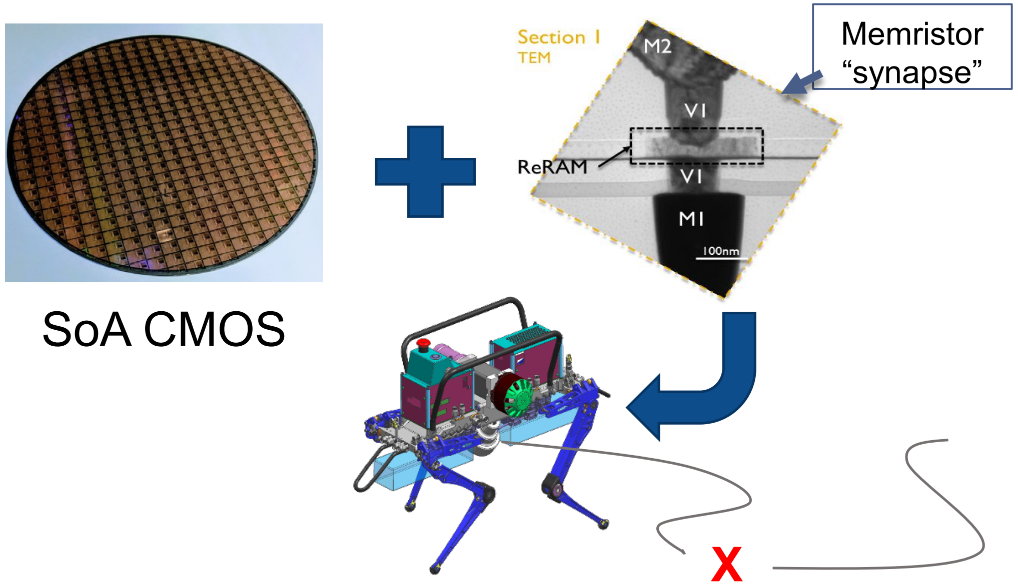
NEUROPIPE:
A COMBINED DEVELOPMENT PIPELINE FOR NOVEL NEUROMORPHIC HARDWARE



ARAP Proposal:

NeuroPipe: A Combined Development Pipeline for Novel Neuromorphic Hardware

Purpose: NeuroPipe will advance DoD capabilities in nanoelectronic materials, devices, architectures and software for critical military applications in autonomy, on-board sensor processing, and cognitive decisions systems. In stages, we will integrate Memristors and other nonvolatile devices, as key elements for advancing next generation neuromorphic processors, into a CMOS foundry.



Products:

- Prototype neuromorphic devices, circuits and chips utilizing emerging US based nanoelectronics and nanophotonics fabrication facilities
- Demonstration of new non-von Neumann neuromorphic architecture with ultra-low SWaP
- Software

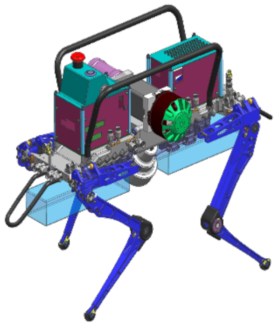
Payoff: New DoD talent and critical infrastructure for future development of neuromorphic computing hardware with…

- On-chip dynamic learning (ie, learning *after* training)
- Reduced training times with smaller data sets
- >100X reduction in consumed energy for processing
- Untethered autonomous capabilities such as 3D navigation, target classification, C2
- Shortened time to warfighter

*American AI Initiative (US, 2019), New Generation Plan (China, 2017),
“Whoever becomes the leader in this sphere will become the ruler of the world,” Russian President Vladimir Putin

The Technology Gap facing DoD Applications

DoD Need: high-performance AI HW with low SWaP



- **Example**: quadruped throw-bot “Butch” (NRL LASR)
  - For navigation in unstructured outdoor environments
  - Ideally would like to have **on-chip learning** with limited training primitives
  - Needs to operate **independent of the Cloud**
  - **SWAP-constrained** platform; GPUs are too expensive power-wise
  - Additional capabilities (real-time sensor data processing etc.) add to the computational burden

Memristor-based neuromorphic chips promise to help make Butch and friends smarter, but **where do we get them?**

Problem: AI chips for DoD platforms represent a small market for semiconductor manufacturers

- Commercially available neuromorphic chips such as True North (IBM), Loihi (Intel) etc. are based on Si CMOS-based technology.
- The industry is happy with the Si-based tech they have (for now).
- The huge investment needed to develop novel memristor-based computing chips will come only when the market incentivizes it. (TAM > 1B widgets)
- **DoD platforms: highly specialized, boutique market** (~10k widgets?)

→ Not commercially available. (…opportunity!)

Barriers to Integrating Novel Materials into Testable Platforms

Degree of fab compatibility (technological readiness)

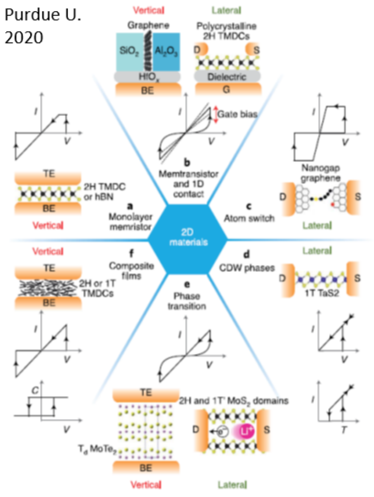
(Post-fab integration only) (Full fab integration)

Q: We have all these nice materials, but why can't Butch use them?

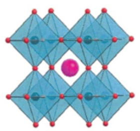
A: Today, most of these aren't in the fab!

Low dimensional materials

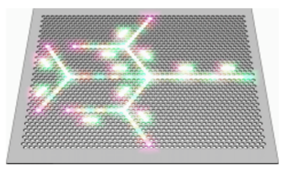
Purdue U. 2020



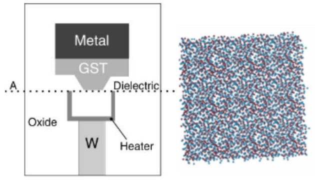
Multiferroics



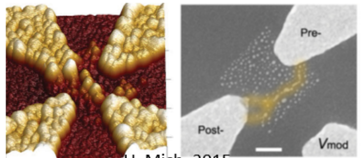
Nanophotonics



Phase Change Materials



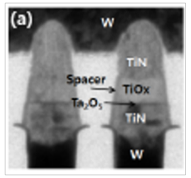
Ionic Switching Devices



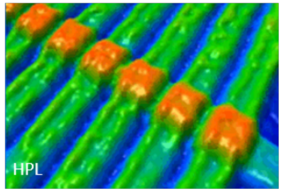
U. Mich. 2015

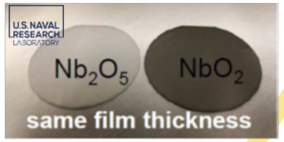
Transition Metal Oxide (TMO) Thin Film Memristors

HfO_2, NbO_2, VO_2, TaO_2, TiO_2, Nb_2O_5, Al_2O_3



HPL-Hynix 2012





Capacity, yield, speed, reliability

Max integration level of most novel materials

Semiconductor Industry SOA

Post-Packaging Integration (PCB board)

Pros: Bridges diverse materials systems
- Flexible process and design
- Fast prototyping

Cons:
- High latencies, low density and yield
- kHz – MHz operation

Back-End Integration (µm CD, 2" coupons)

Pros: Close integration with CMOS
- Retains some benefits of semicon tech
- Relatively flexible process and design

Cons:
- Diminished throughput and yield
- MHz operation

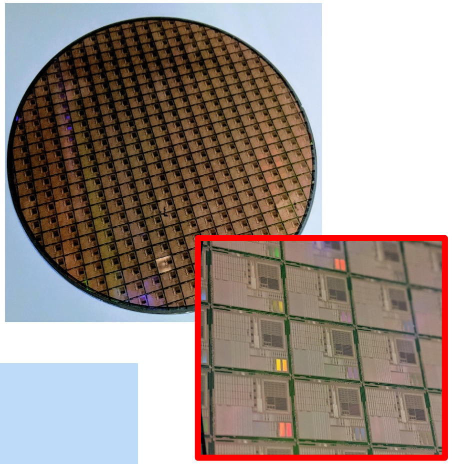
Full Fab Integration (< 65nm CD, 12" wafers)

Pros: Closest integration with CMOS
- Capacity and yield of semicon tech
- GHz operation

Cons:
- Limited to fab-compatible materials
- Long and inflexible process

Getting from Lab to Fab to Field

FEOL CMOS Control Chip



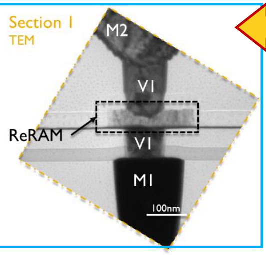
@ AFRL trusted foundry (CNSE, SUNY-Poly, Albany

Coupons diced from 300mm CMOS wafers

World's first and only FEOL CMOS-compatible memristor process

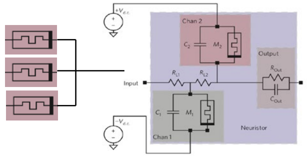
Full FEOL Integration

For qualified materials



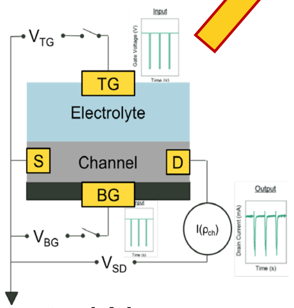
Materials & Device Research

NbO_2 spiking neuron



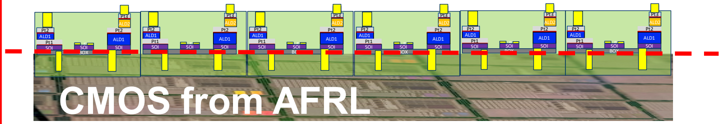
U.S. NAVAL RESEARCH LABORATORY

ARL

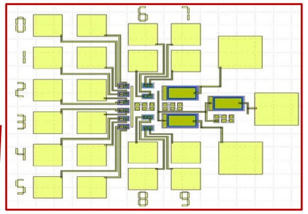


Ion-gated Neuromorphic Device (IGND)

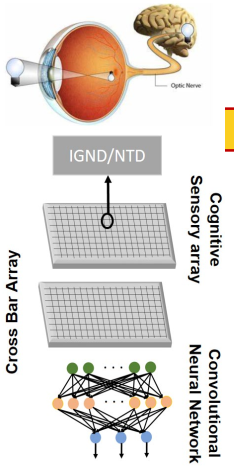
BEOL Integration



Memristive SNN Decision Circuit



Cognitive Edge Sensory Processors

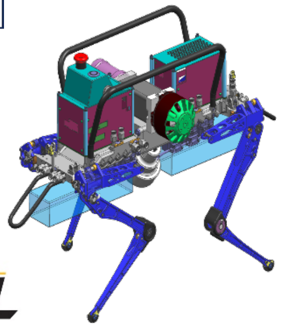


Software

Prototype Testing

Butch's Mind (on-chip learning)

U.S. NAVAL RESEARCH LABORATORY



ARL

Butch's Senses (edge processing)

Summary of NeuroPipe

- **DoD path to dominance in neuromorphic computing requires timely ARAP investment and IP management of enabling nonvolatile device technologies, in novel neuromorphic architectures, fabricated in a US based CMOS manufacturing environment.**

- **Tri-Service labs will leverage previous, foundational DoD investments (>$100M) in neuromorphic devices, circuits and architectures as a map for NeuroPipe technologies.**
  - **The key determinant of success for the proposed ARAP is the development of capabilities to rapidly and flexibly design and manufacture novel neuromorphic hardware that is *beyond the established semiconductor technology framework*.**

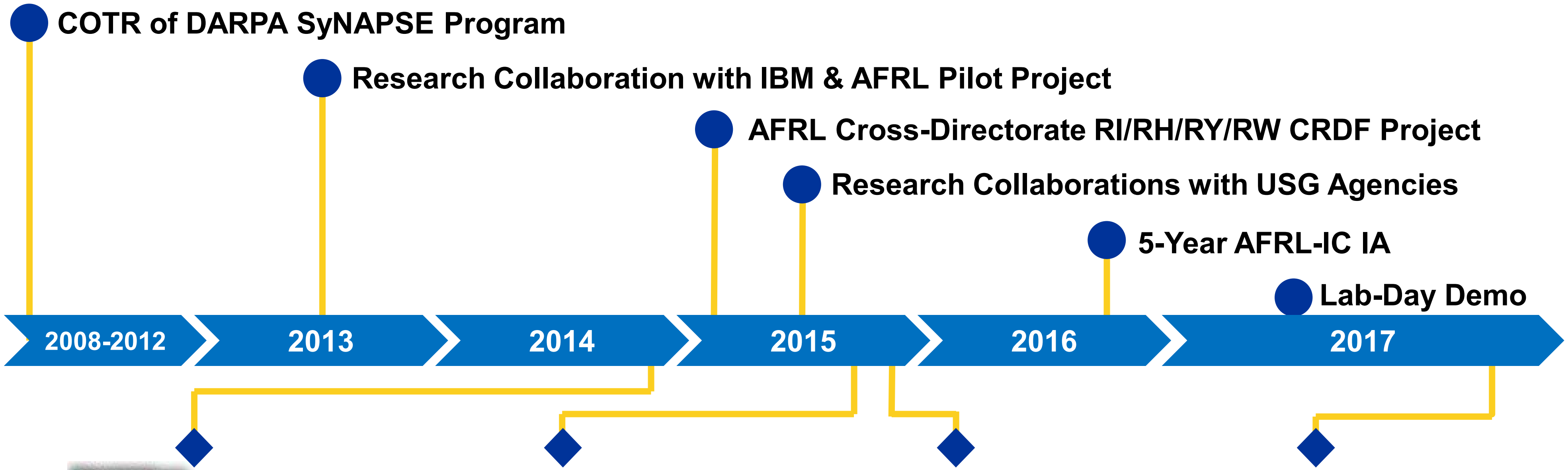
- **A streamlined "Lab-to-Fab-to-Field" development pipeline will be created to apply to critical military applications in autonomy, on-board sensor processing and cognitive decisions systems.**

- **The NeuroPipe ARAP guarantees that DoD will have the necessary in-house expertise to counter adversarial advances in neuromorphic technologies, and have capabilities for rapid low volume production to shorten the time to the warfighter.**

Exploiting emerging nanoelectronics and photonics to *win* the "AI arms Race"

Short Term Focuses

AFRL TrueNorth Research Main Events



SyNAPSE (12/2014)
AFRL/RI Pilot

NS1e (08/2015)
AFRL Cross-TD Project

NS1e-16 (12/2015)
AFRL Cross-TD Project

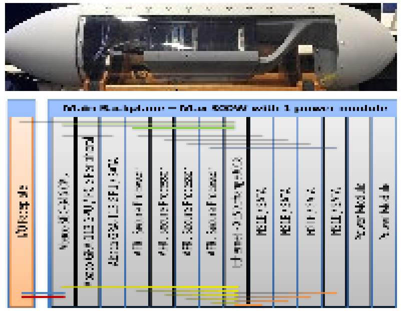
NS16e-4 (est. 6/2018)
AFRL-USG Collaboration

First Hardware Acquisitions and Applications Development Among ALL USG Agencies!

Research Scope & Plans

Near-Term
Embedded HPC & Deep Learning Applications

"Agile Condor" eHPC Pod



Multi-Target Recognition With Machine Learning

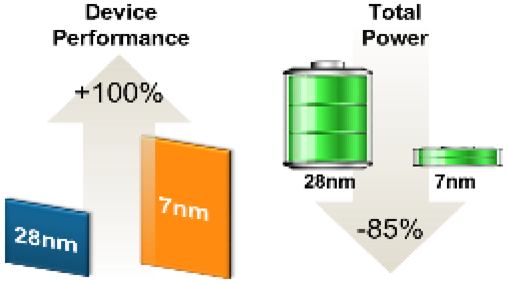
Synthetic Aperture Radar Image | IBM NS1e-16 with 16 TrueNorth Processors | Bounding Boxes & Target Classes

TrueNorth × 16 vs A Mobile GPU

- 10X Power Reduction
- 20X Throughput Increase
- Same Accuracy

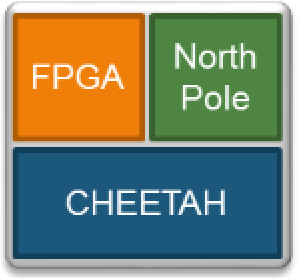
Mid-Term
Next-Gen Digital Neurosynaptic Processor

7nm Trusted Manufacturing



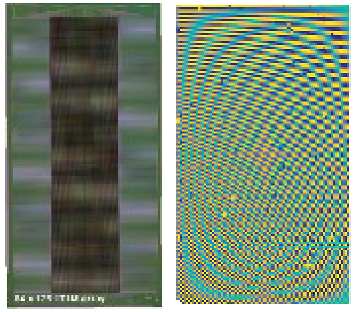
100X Function Density

Trusted Reusable IP & Software

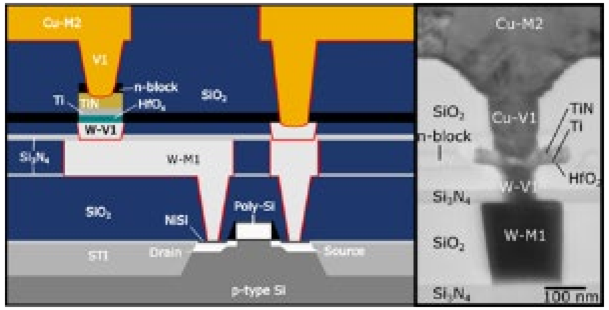


Far-Term
Nano-Technology Enabled Neuromorphic Systems

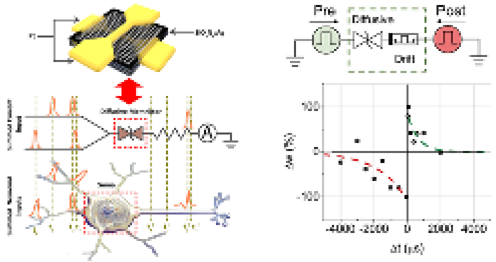
Memristor Devices & Array



Memristor-CMOS Integration



Biologically Plausible Architectures



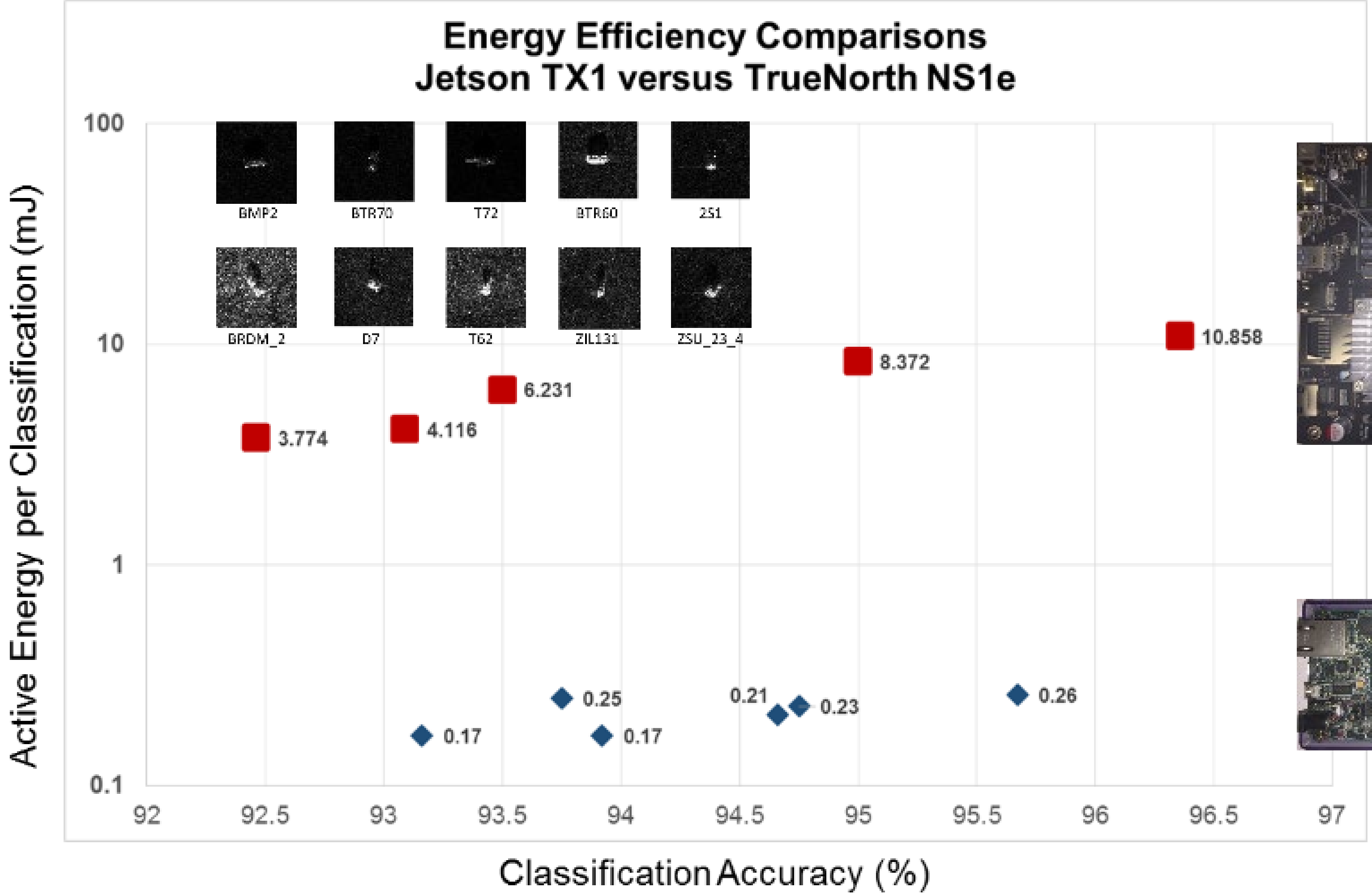
Memristor Based Artificial Neuron and Synapse

SAR Target Recognition

10-Class Public Dataset

$$E_A = t\cdot(P_{total} - P_{idle})$$

$$Acc = 1.0 - \mathbf{SUM}[label(i) - prediction(i)]/N$$

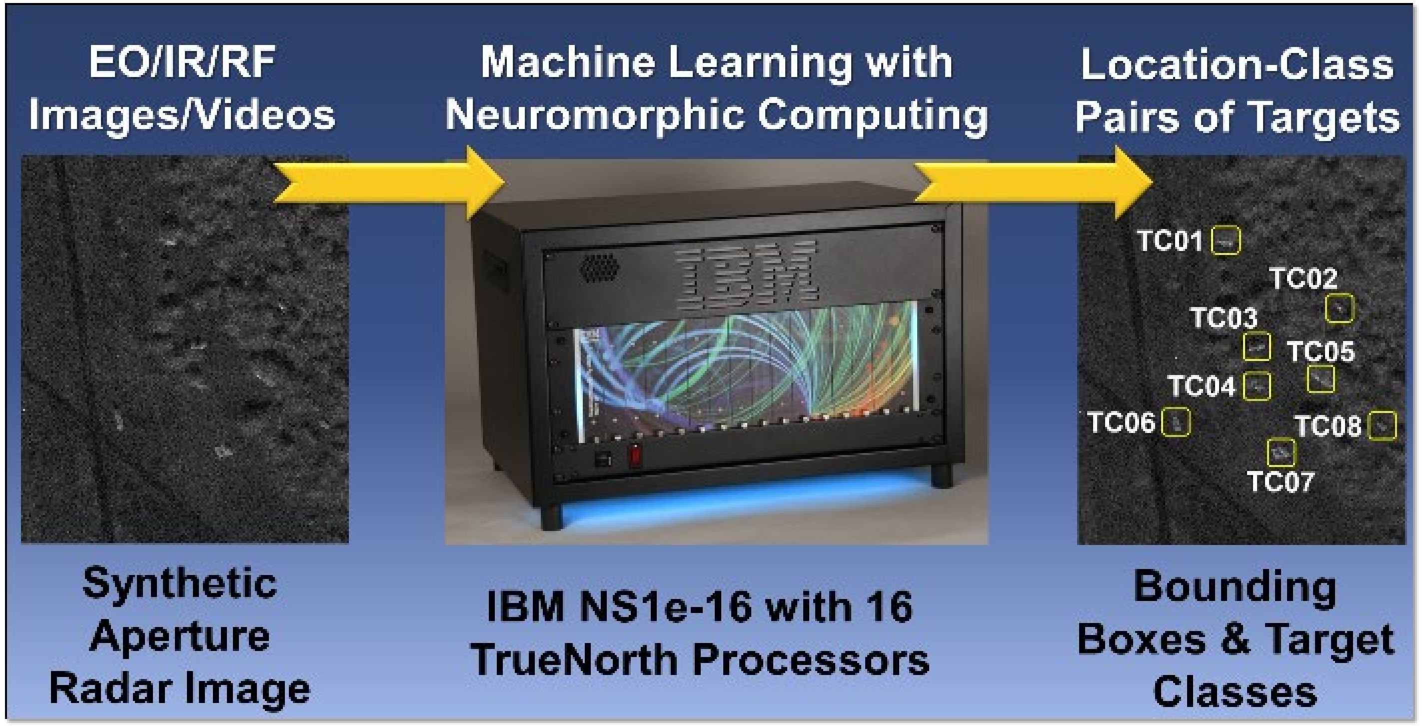


TX1

NS1e

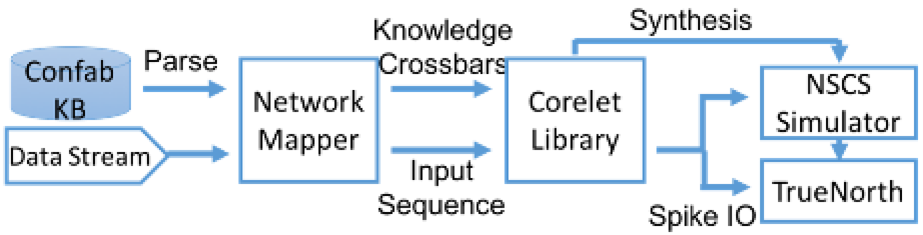
DoD Lab Day Demo

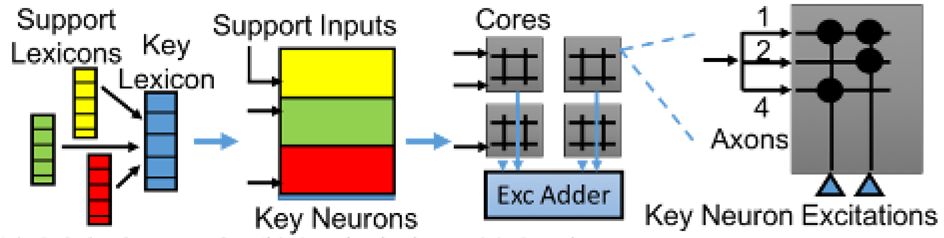
26-Class SAR Target Recognition



	TrueNorth Processor × 16	NVidia Mobile GPU
Power	5 Watts	50 Watts
Throughput	16,000 images per sec	700 images per sec
Accuracy on chips	98%	98%
Accuracy on full-frame	90%	90%

Anomaly Detection with Spiking Inference





(a) Original Network (b) Stacked Xbar (c) Corelets (d) Knowledge Links

Table 4.1: Network complexity impacts of constraint

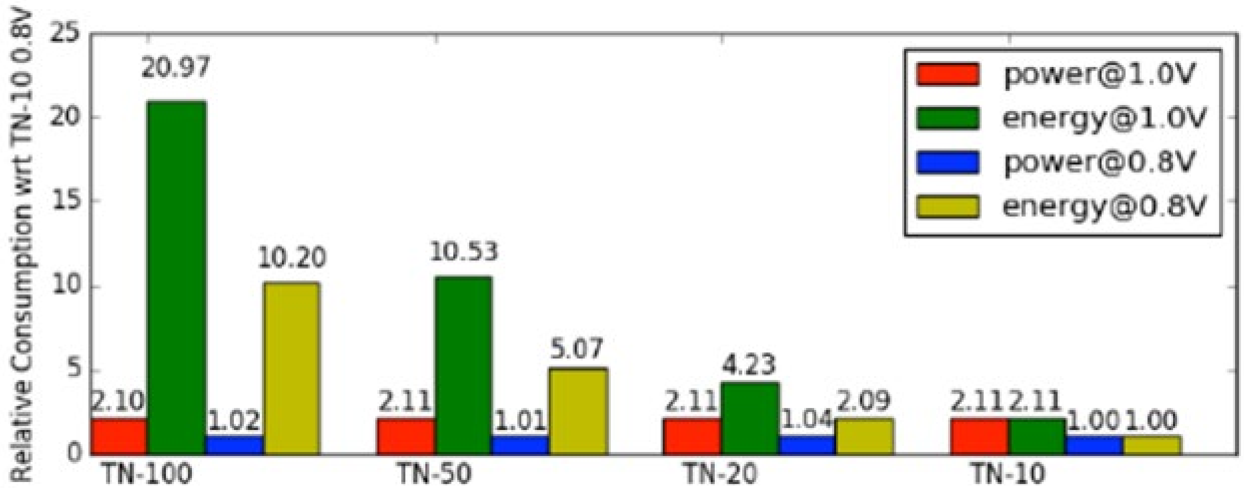
Networks	Synapses	Cores for Key Lex	Total Cores
Original	3373K	5232	6116
Constraint	1322K	2169	2918
Reduction	60.8%	58.5%	52.3%

Table 4.2: Detection Qualities of Comparison Models

Methods	SOM	RNN	Reference	TN-100	TN-10
AUROC	0.879	0.898	0.933	**0.943**	0.914

Table 4.3: Power and Performance of Different Platforms

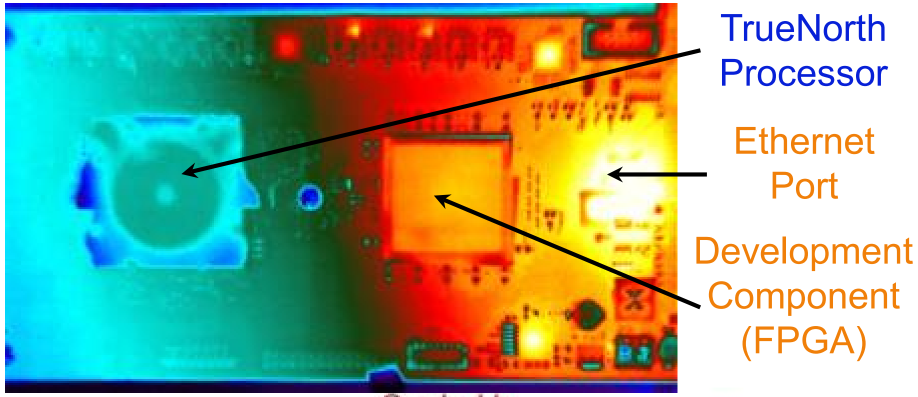
Devices	Time	Power	Energy/Sample
Xeon W5580	25.7ms	68.0W	1747.6mJ
Tesla K20	0.270ms	102.4W	27.6mJ
Jetson TK1	13.48ms	2.5W	33.7mJ
TN-10 1.0V	20ms	104.1mW	2.1mJ
TN-10 0.8V	20ms	49.22mW	0.98mJ



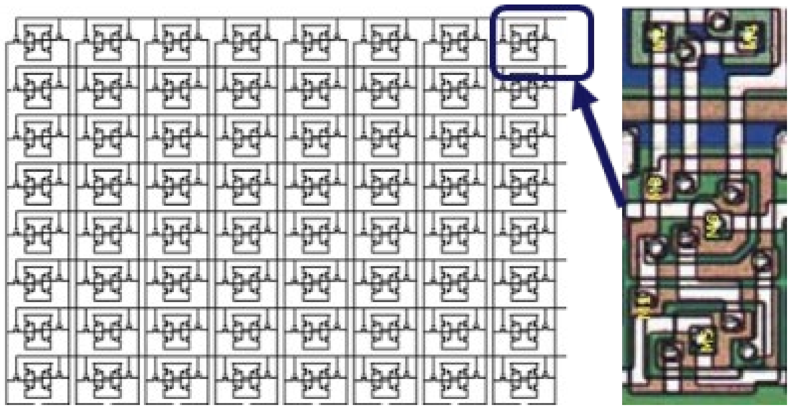
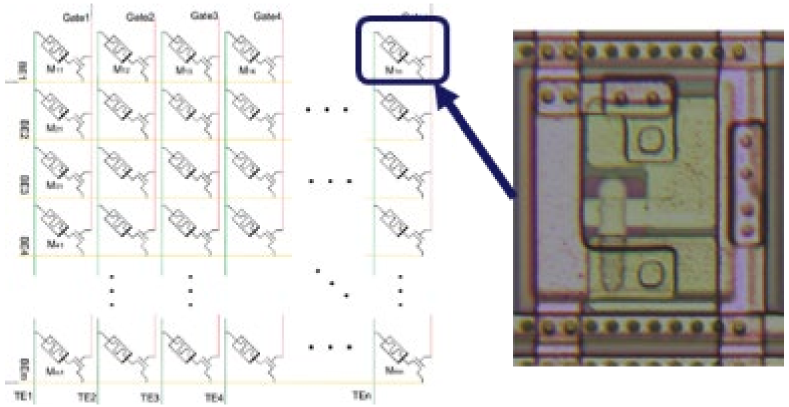
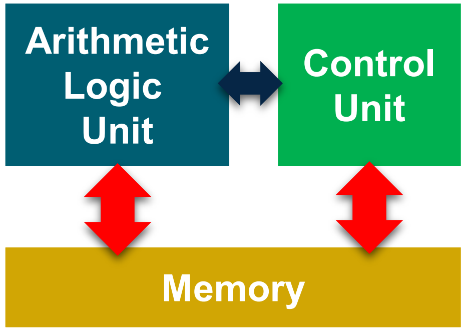
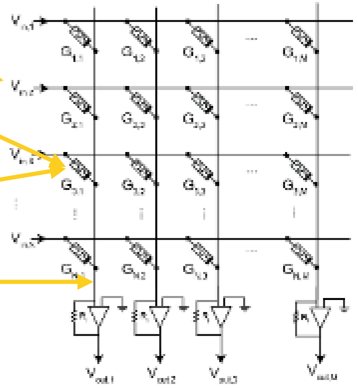
Huge Room For Improvements

- TrueNorth is a first-generation research and experimental processor
  - Input bandwidth limitation on data dimensionality and size
  - High area cost from experimental redundant circuits
  - Not optimized for operational environments (temperature, radiation)
  - Power overhead from developmental system components
  - Fixed-scale design not allowing customization and optimization
- We must leap forward
  - Keep innovating as our mission, not dwelling on supporting a research chip
  - DoD needs a more optimized, more versatile, trusted machine learning microelectronics solution

Thermal Image Showing Power Overhead on the NS1e Board

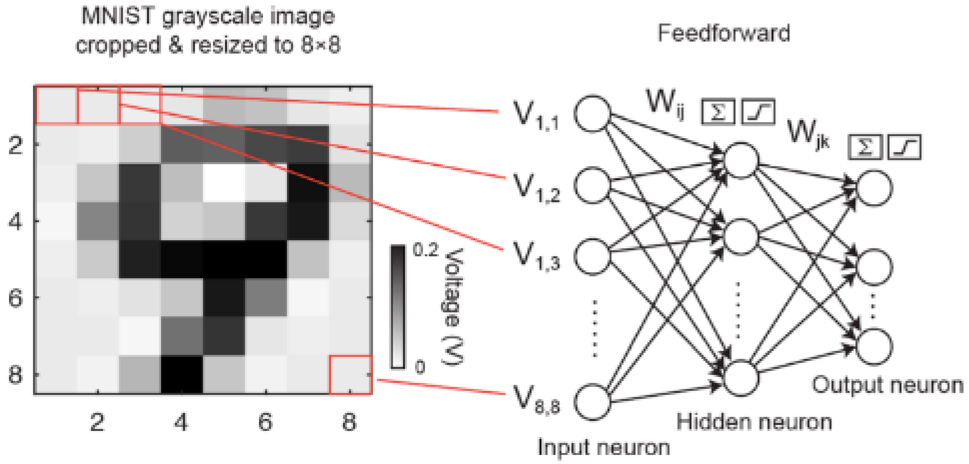


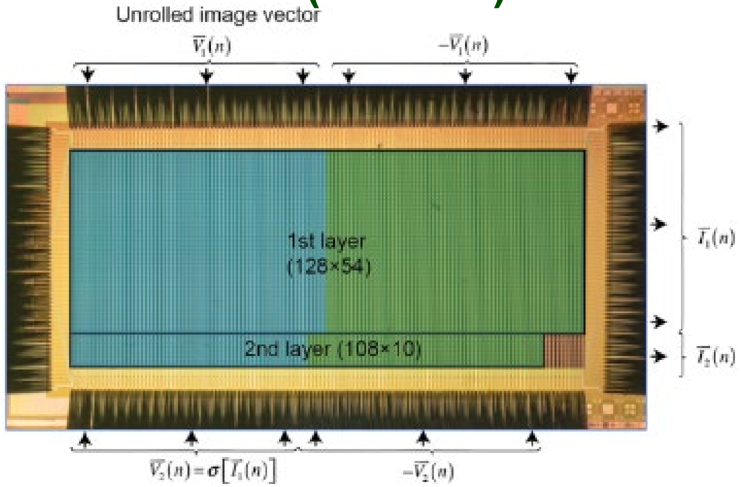
Future: Memristor

	Digital CMOS	Memristor
Size Weight Cost	 6 transistor area = 1 bit	 1 transistor area = 6 bits
Power $= V^2 / R$	Typical R: 10^3 Ohms	Typical R: 10^3~10^6 Ohms
Comp. Arch.	 Compute-Out-Memory (Von Neumann)	Input: Voltage (V) Memory: Memristance (R) Compute: I = V / R Output: Current (I)  Compute-In-Memory

Initial Comparisons to CMOS

In-hardware learning & inferencing: 2-layer neural network (UMASS)





Inferencing Compute	Google TPU	Memristor-based (array-only)
Energy Efficiency (fJ/OP)	470	19

Multi-layer neural network inferencing (Duke-UMASS)

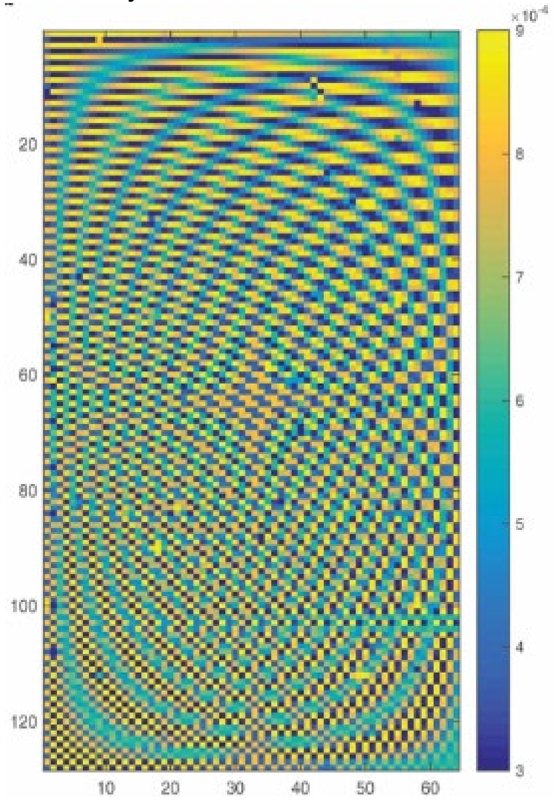
Inferencing Compute	TrueNorth (28 nm)	Memristor (UMASS array) -CMOS (Duke 130nm chip)	Memristor (10X memristance) -CMOS (10X speed)
Memristor Resistance Range		700Ω – 70KΩ	**7KΩ – 700KΩ**
Power (mW)	5.115	37.6	16.2
Speed (ns per image)	2,000	100	**10**
Energy Efficiency (nJ per image)	10.23	3.76	0.162
Classification Accuracy	~97%	~97%	~97%

Recent Research Outcomes

Matrix-Vector Multiplication, Signal Processing

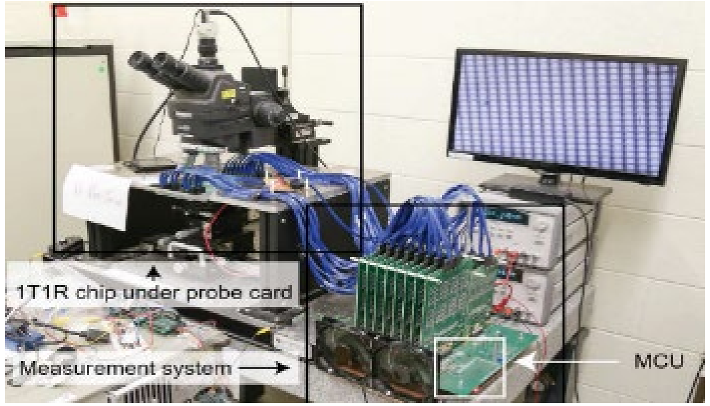
128×64 6-bit (64-level) 1T1R Array

- The largest analog valued array to date (limited by the measurement capability)
- The most accurate analog array (σ~0.5%)
- The most analog levels obtained (>64, 6-bit)
- 99.8% yield: 3 stuck ON, 15 stuck OFF

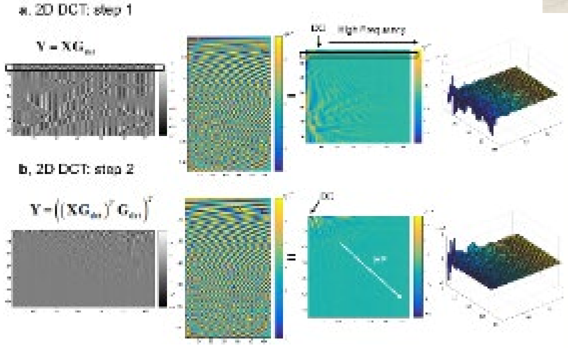


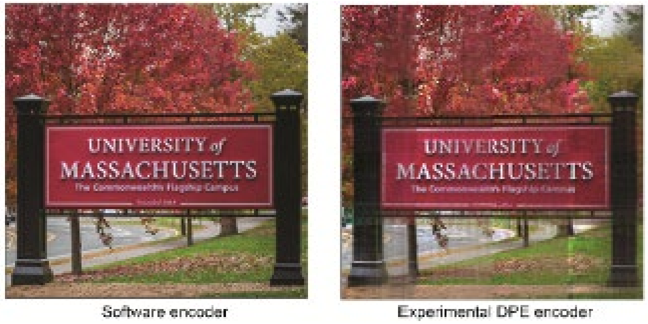
Nature Materials 16, 101 (2017)

Memristor crossbar computing experimental setups at UMASS

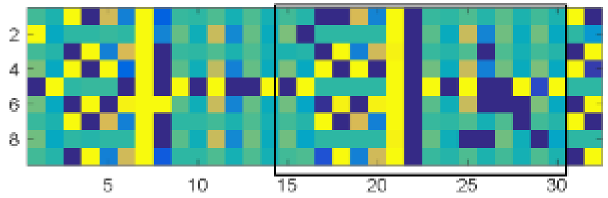


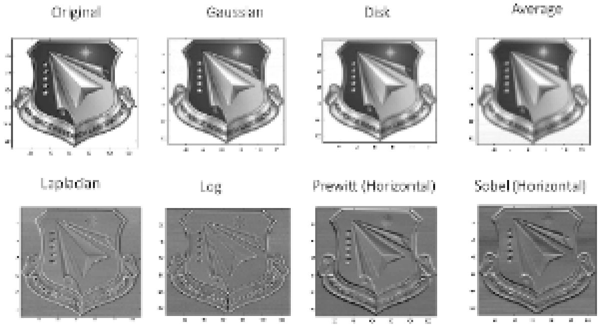
Experimental results: 2D DCT





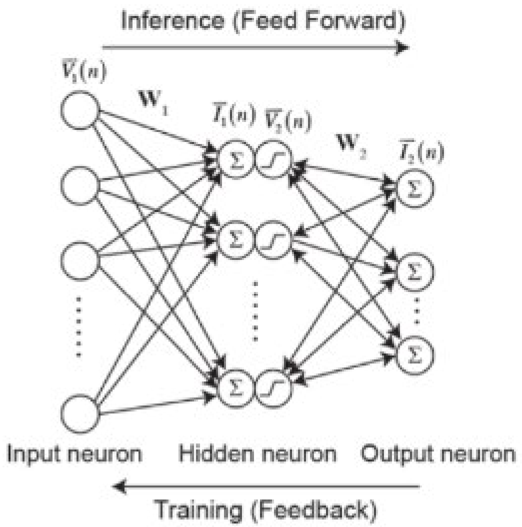
Experimental results: 3×3 kernel convolution



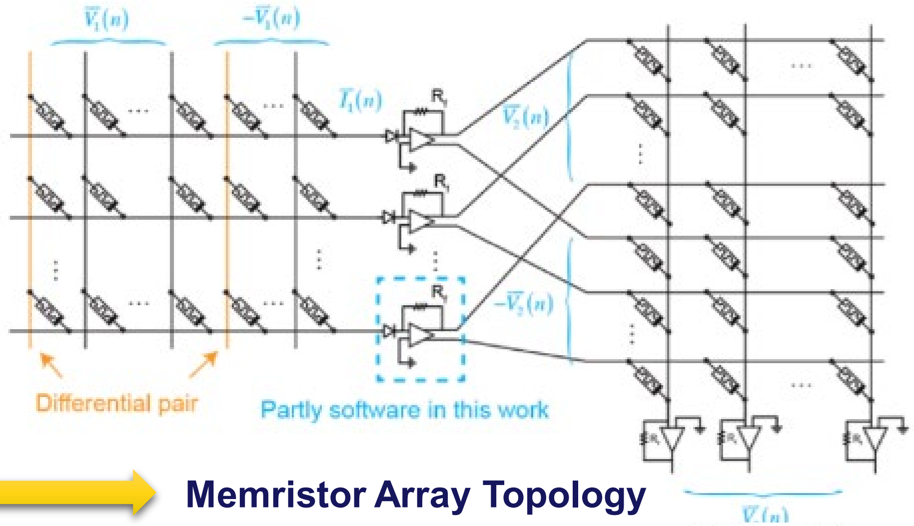


Recent Research Outcomes

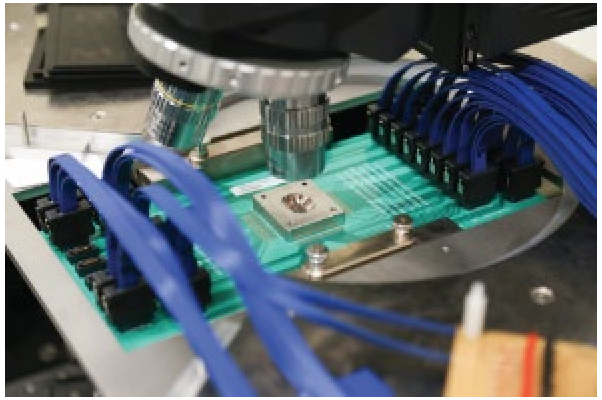
In-Hardware Neural Network Learning & Inferencing



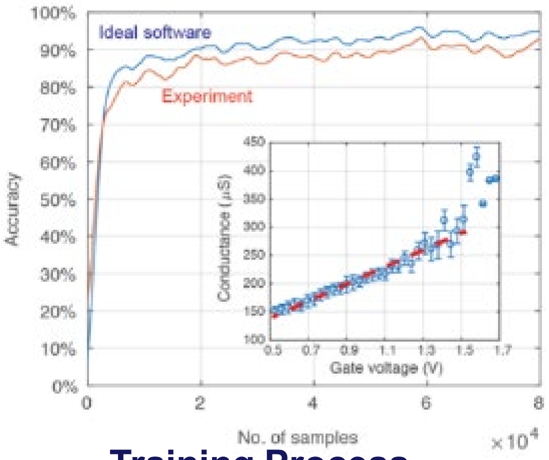
Neural Network Topology



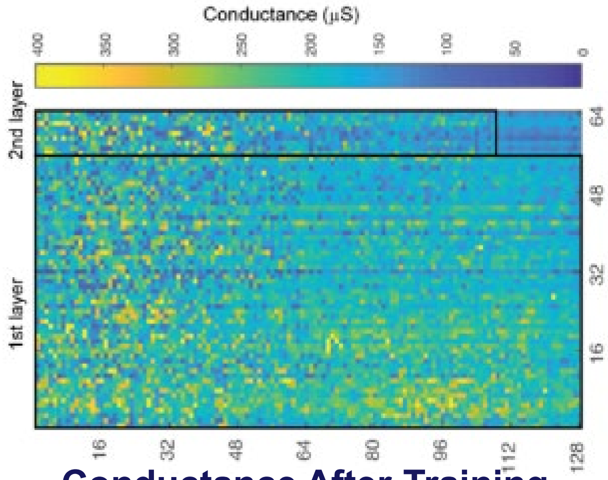
Memristor Array Topology



Memristor Array On Probe Card



Training Process



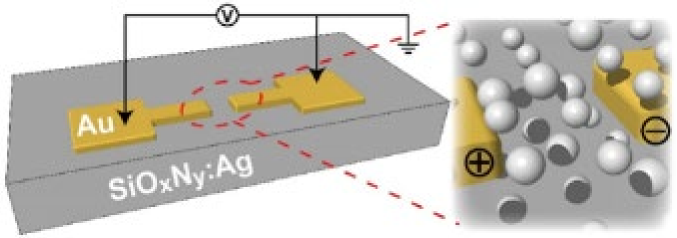
Conductance After Training

Science, submitted (2017)

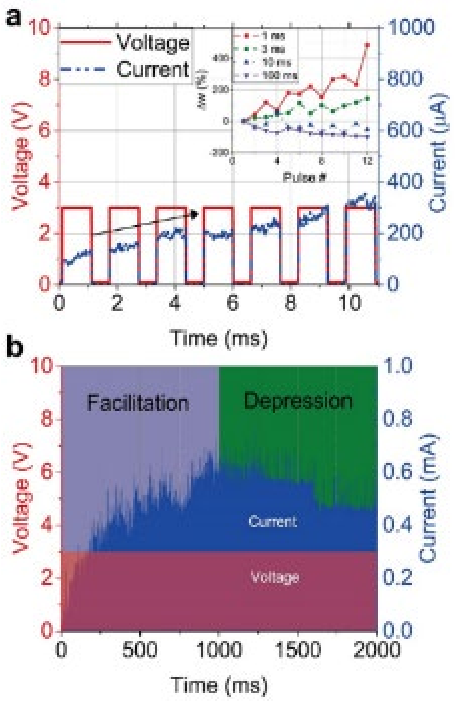
Recent Research Outcomes

Artificial Synapse

SiO_xN_y:Ag Diffusive Memristor

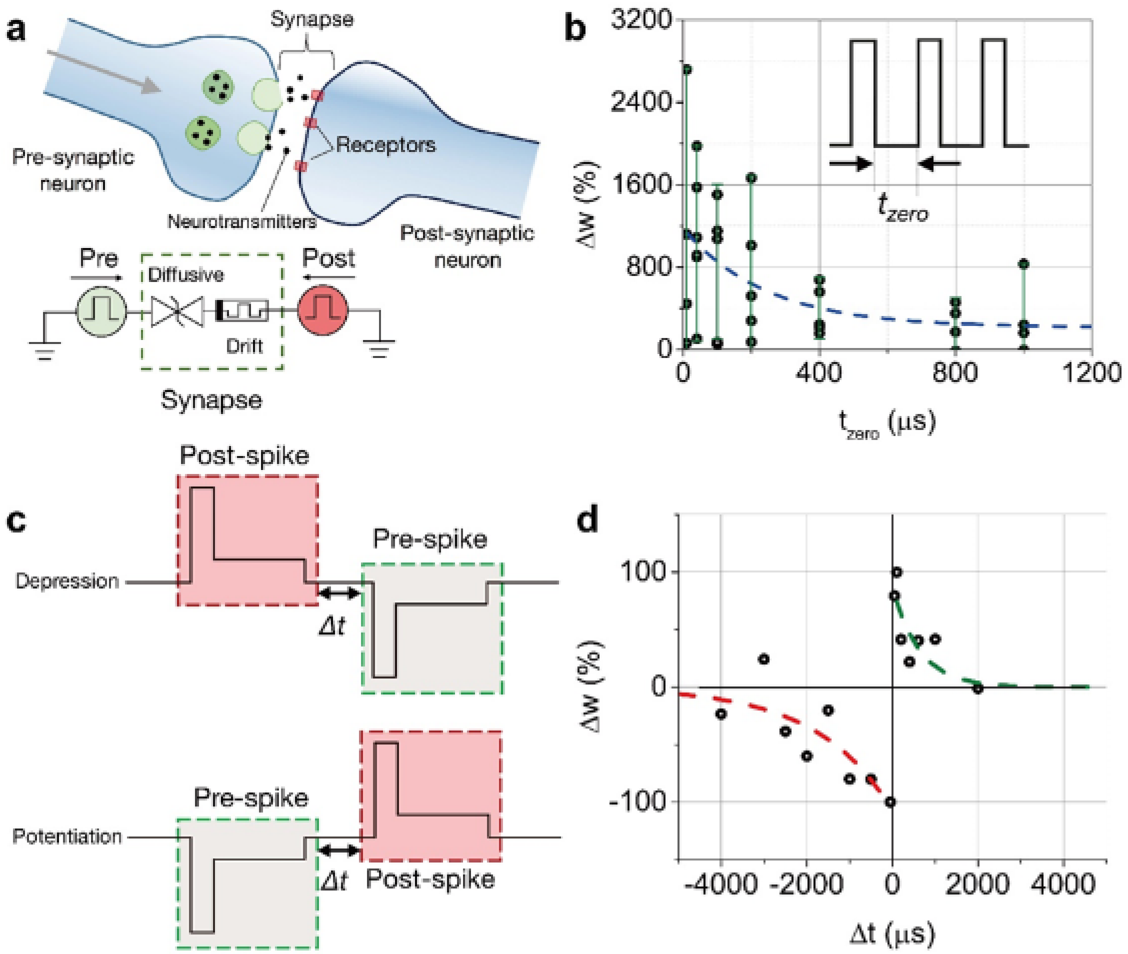


Short-Term Plasticity



Nature Materials 16, 101 (2017)

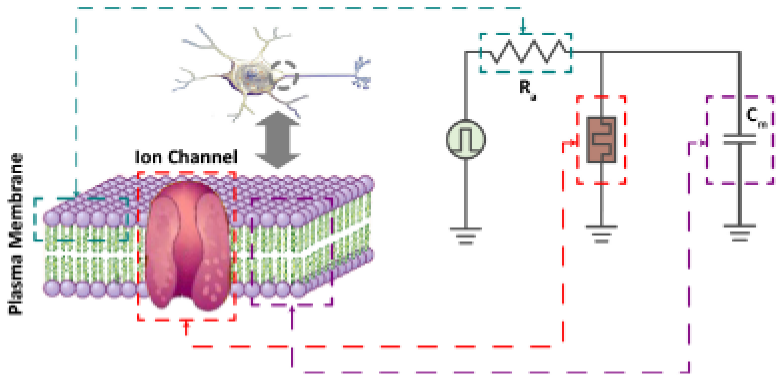
Long-Term Plasticity & STDP Learning



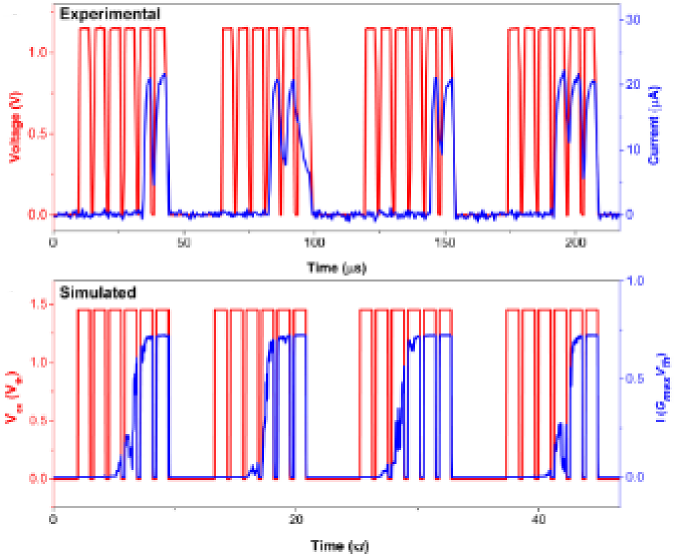
Recent Research Outcomes

Artificial Neuron

Biological Neuron vs Memristor Neuron

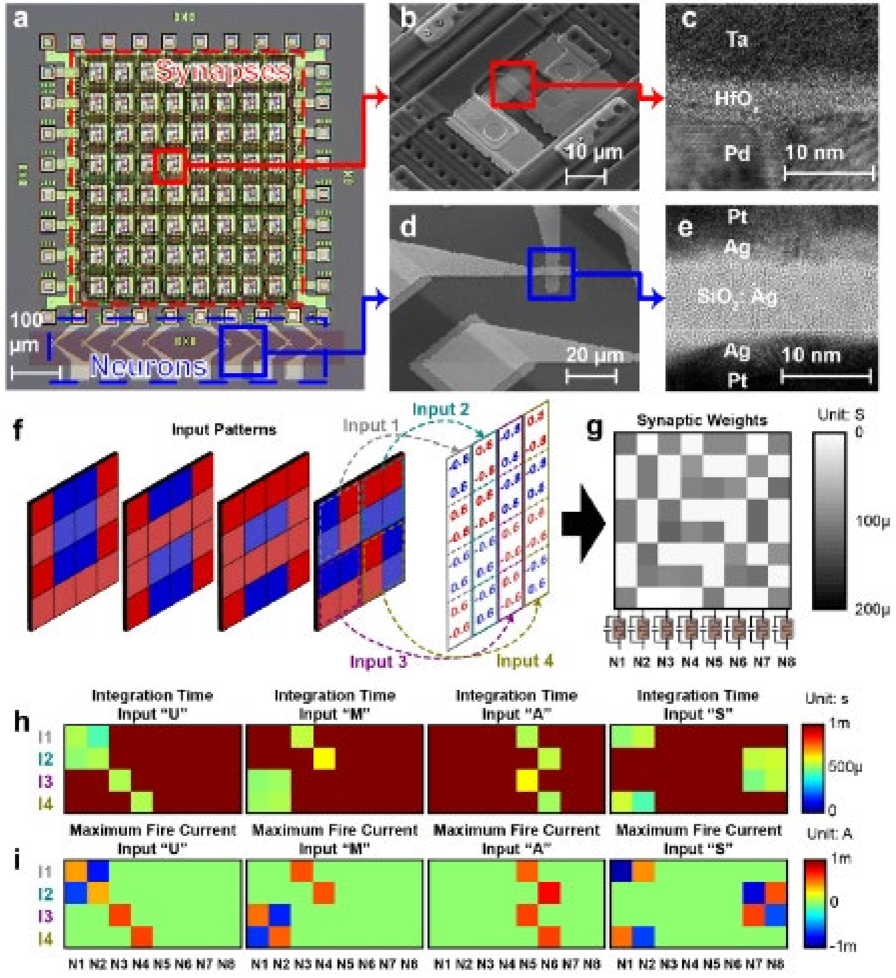


Integrate-and-Fire Function with Memristor Neuron



Nature Nanotechnology, under revision (2017)

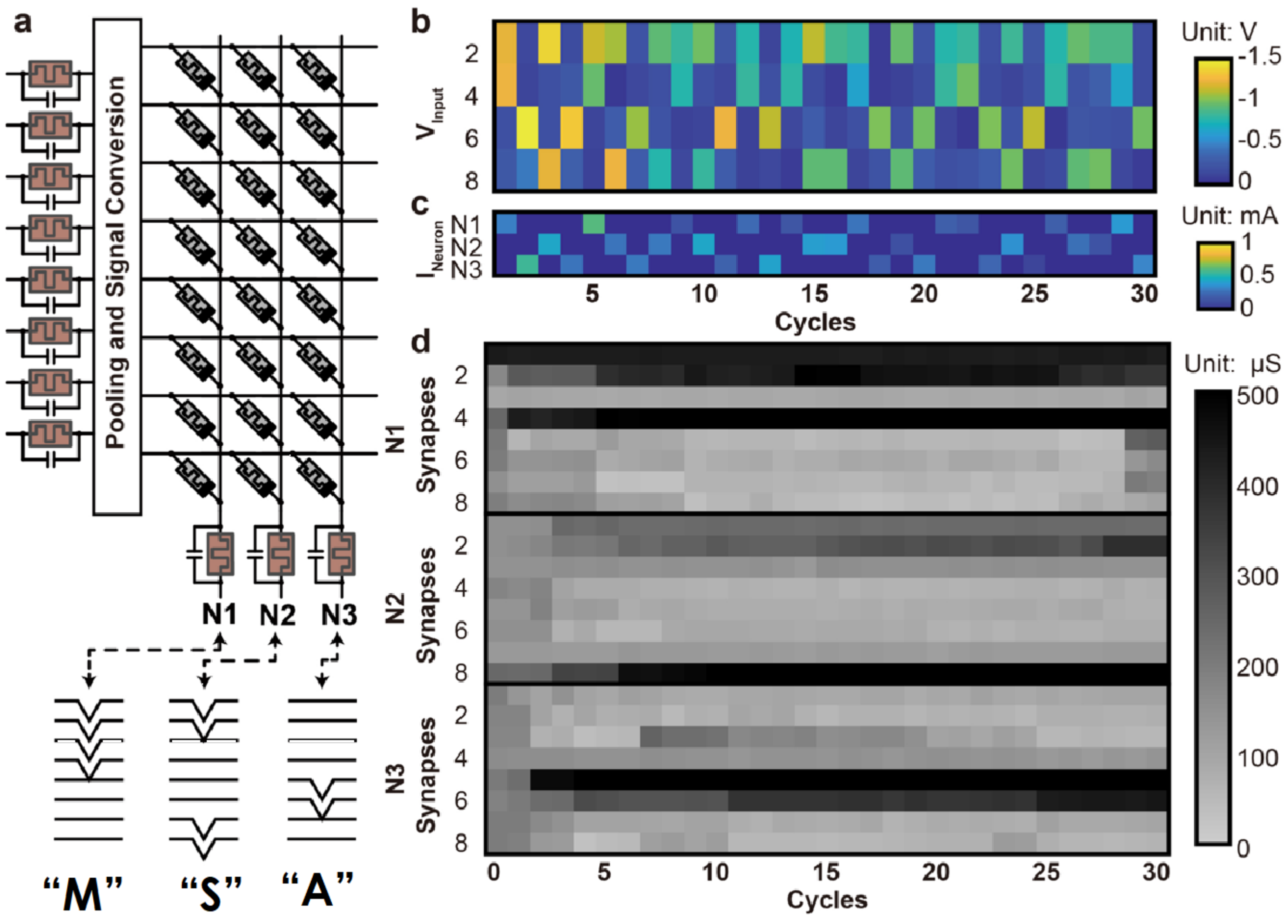
8×8 Memristor Array with 8 Memristor Neurons



Pattern Recognition of 'U', 'M', 'A' and 'S'

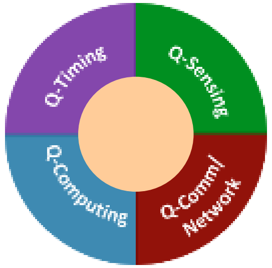
Recent Research Outcomes

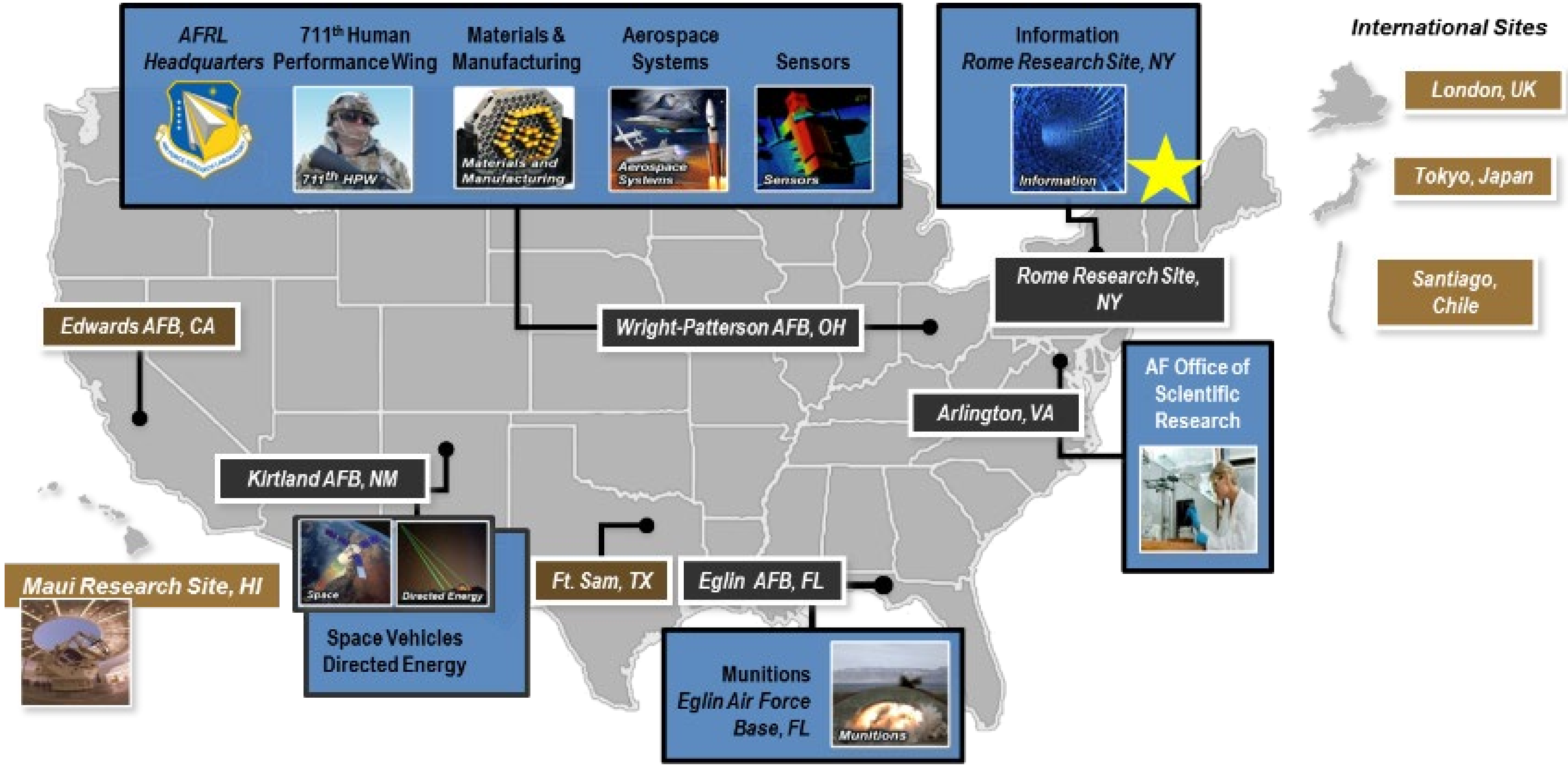
Unsupervised Learning In All-Memristive NN



Wang et al., *Nature Electronics* 1,137-145 (2018).

Quantum Information Science at AFRL





RD: Satellite-based Quantum Communication and Networking and Optical Channels (Gruneisen; NM)

RI: Quantum Computing, Communication, and Networking (Alsing, Fanto, Soderberg, Tabakov, Hucul, Lahaye; NY)

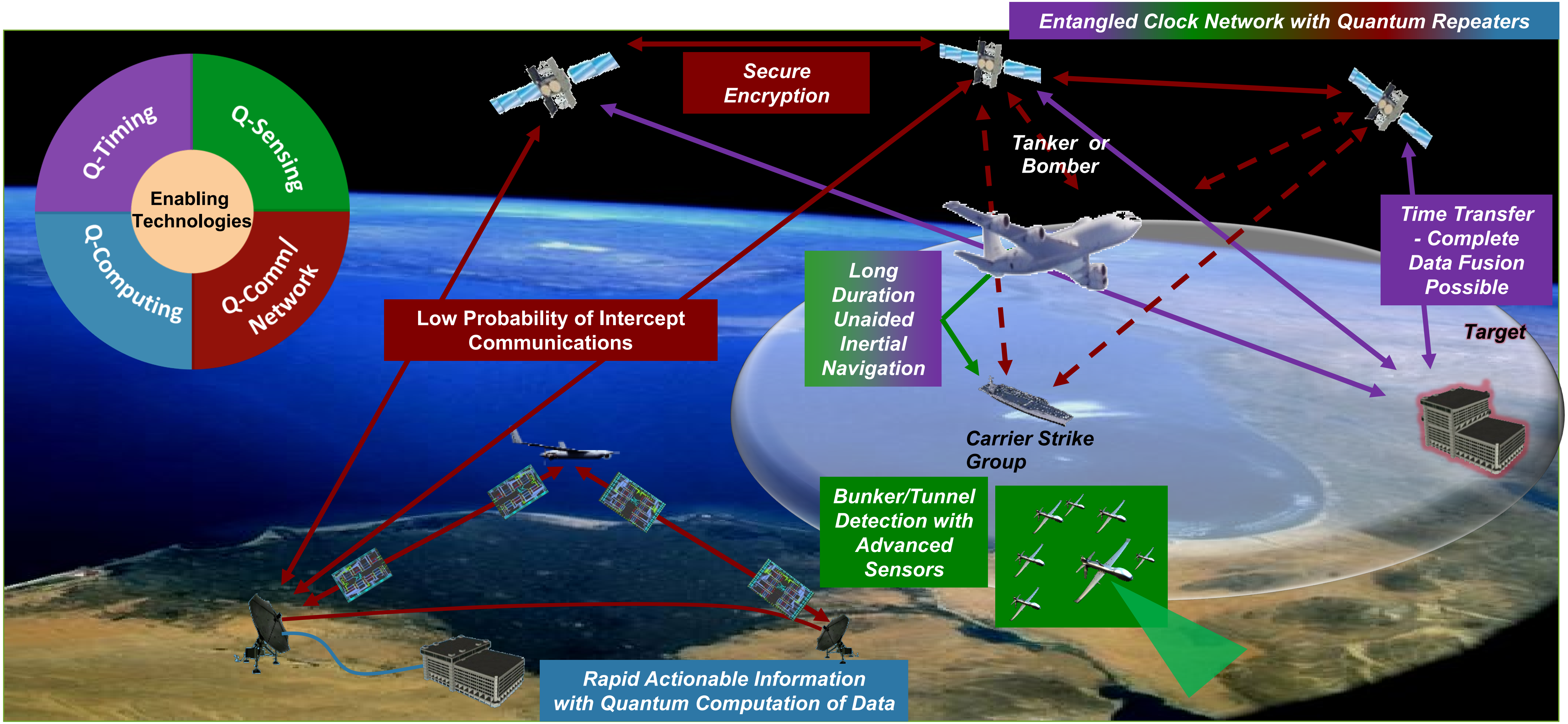
RV: Position, Navigation, Timing (Lott, Metcalf, Squires, Olson, Elgin; NM)

RX: Solid-State Quantum Defects, Materials, and Supply Chain (Bedford, Bissell, Dass, Eyink, Reed, Slocum; OH)

RY: Quantum emitters, Device fabrication (Hendrickson, Usechak; OH)

AFOSR: 6.1 Basic Research funding in QIS (Metcalfe; DC)

Quantum-Enabled Air Force Capabilites

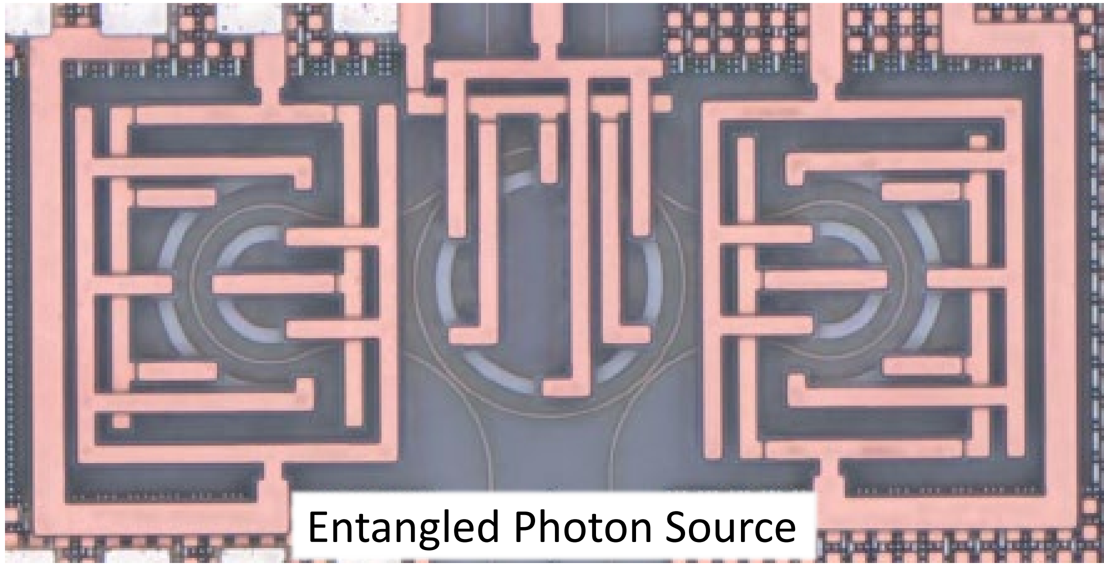


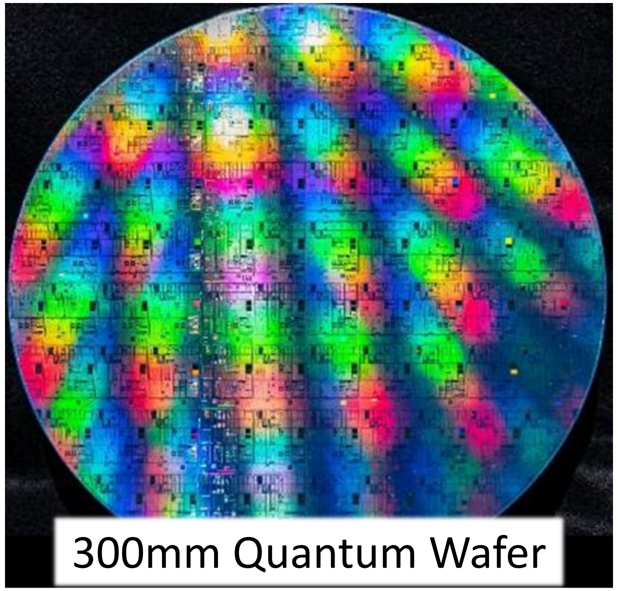
Photon-based Qubits

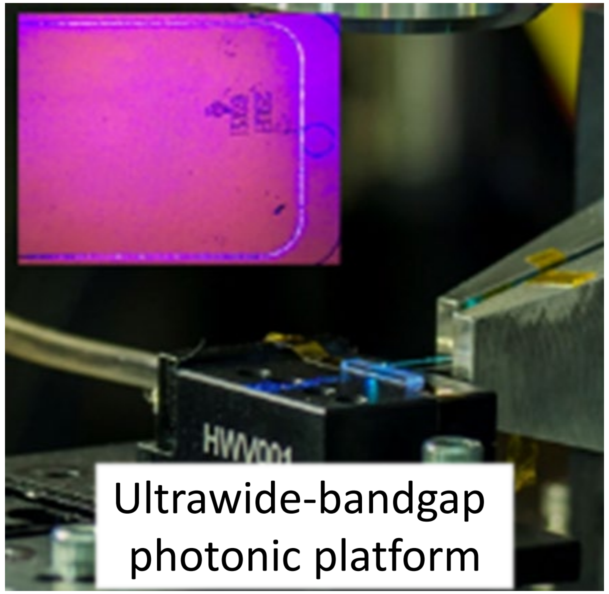
Motivation: Integrated photonics provides a scalable and stable platform for the implementation of photon-based qubits.

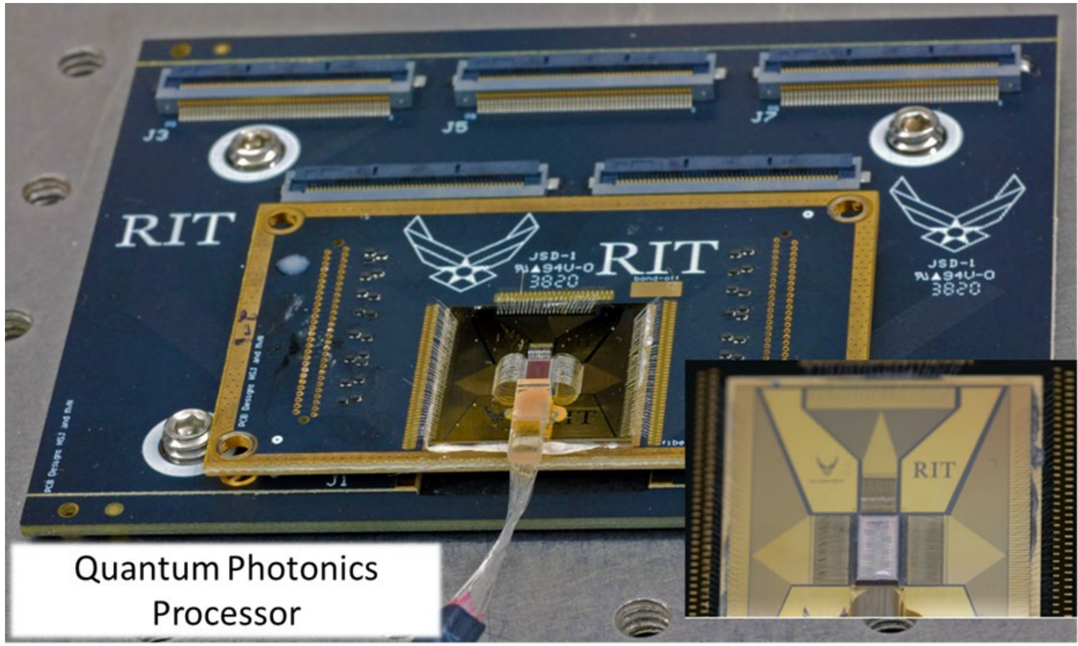
Advantages:

- The only long distance carrier of quantum information
- Robust against decoherence
- Operate at room temperature
- Ideal for fast processing circuits
- One of the main mediating qubits for transduction


Entangled Photon Source


300mm Quantum Wafer


Ultrawide-bandgap photonic platform



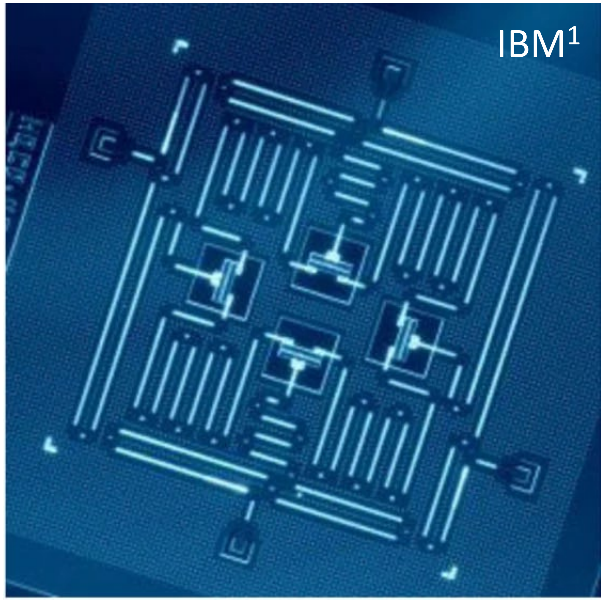
Quantum Photonics Processor

Superconducting Qubits

Motivation: Superconducting qubits are macroscopic quantum circuits and a leading *matter-based* quantum processing platform

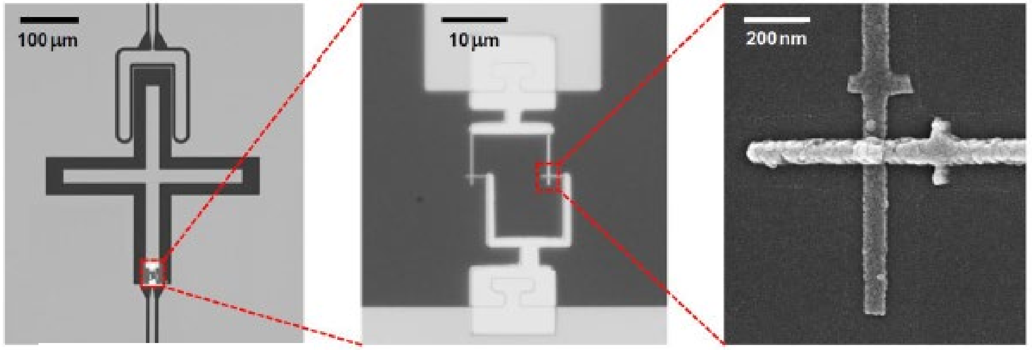
Advantages:

- *Artificial atoms* manufactured w/ semiconductor fabrication
- Dynamical properties can be precisely engineered
- Operate in the microwave regime
- Control & measurement electronics can be tightly integrated
- Enable rapid quantum logic & deterministic entangling operations



Superconducting processor w/integrated 'transmon' qubits & quantum buses

Components of a Superconducting 'Transmon' Qubit[2]



Control/Meas. Electrodes

Flux Tuning Loop

Josephson Junction

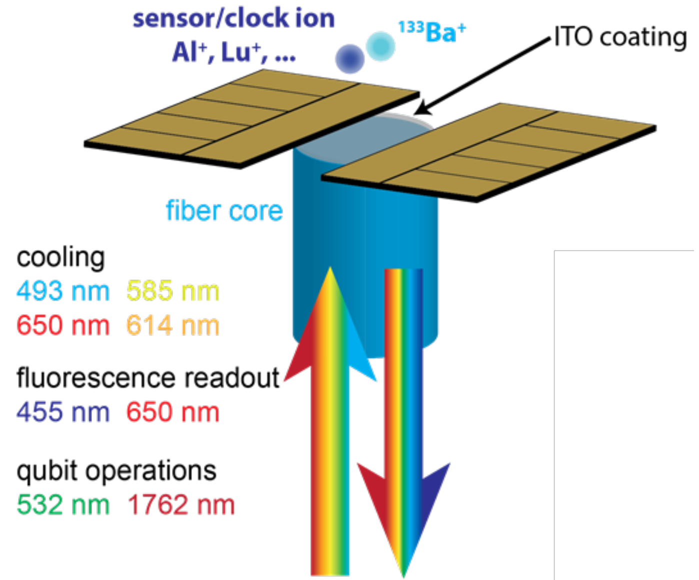
References: (1) J. Gambetta et al. arXiv:1510.04375
(2) T. Roth et al. arXiv:2106.11352

Trapped Ion Qubits

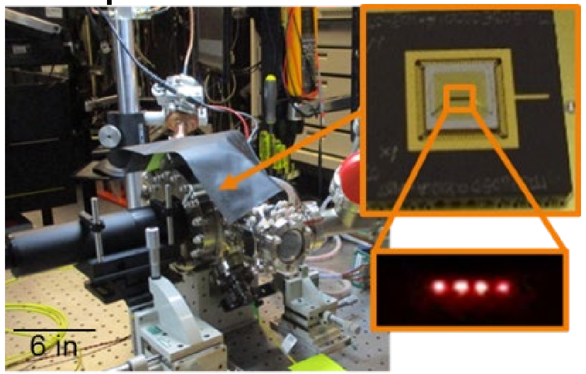
Motivation: Trapped ions provide long-lived memory and quantum information processing capabilities. Exploring use of ytterbium-171 for memory and barium-133 for heterogeneous interfaces.

Advantages:

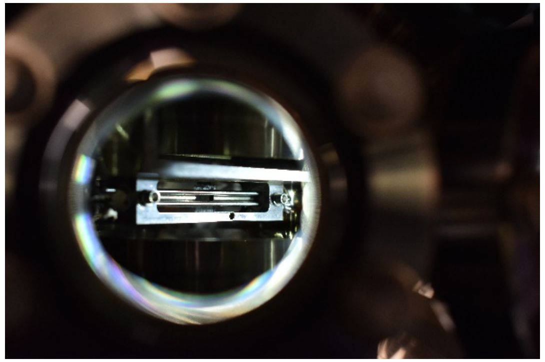
- Provide long-lived memory
- Nature provides identical qubits
- Leading qubit technology for quantum computing – advances can be leveraged towards networking applications
- Allows remote (inter-node) and local (intra-node) entangling operations



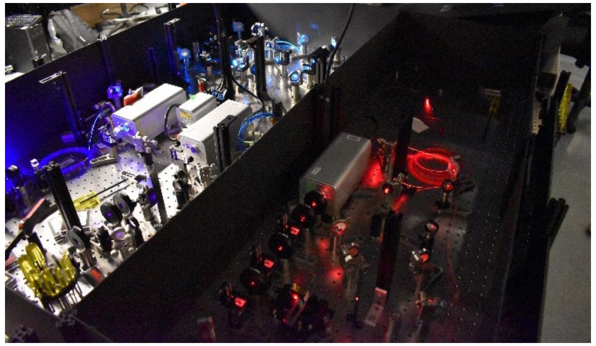
Exploring compact trapped ion architectures towards interfacing with other qubit technologies



Yb^{+} memory-based network nodes with four trapped ion qubits.



Vacuum chamber for preliminary Ba^{+}-133 experiments and qubit interface explorations



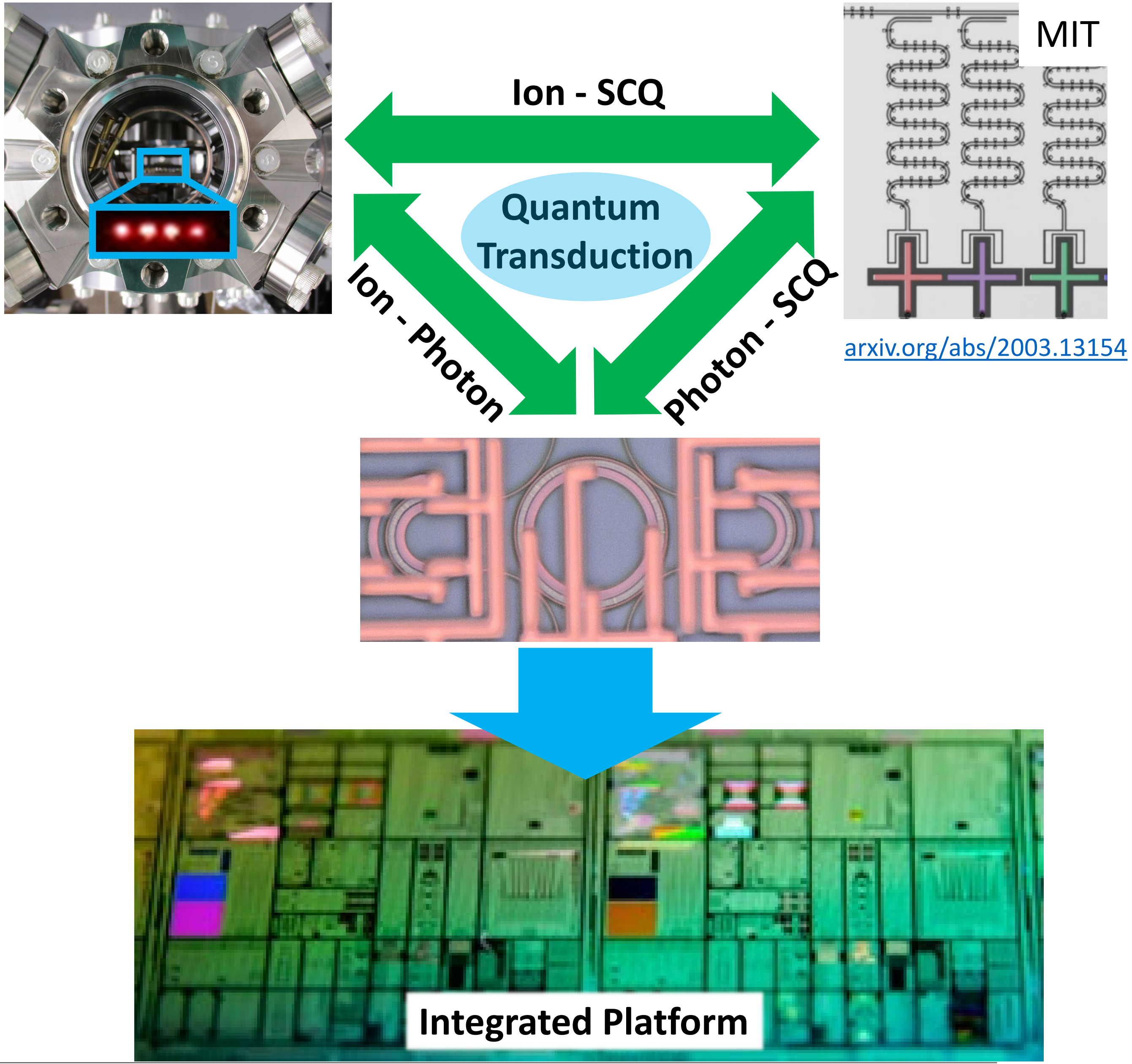
Laser systems to manipulate Ba^{+}-133 trapped ions.

Heterogeneous Qubit Integration

Developing hybrid interfaces to integrate multiple qubit technologies

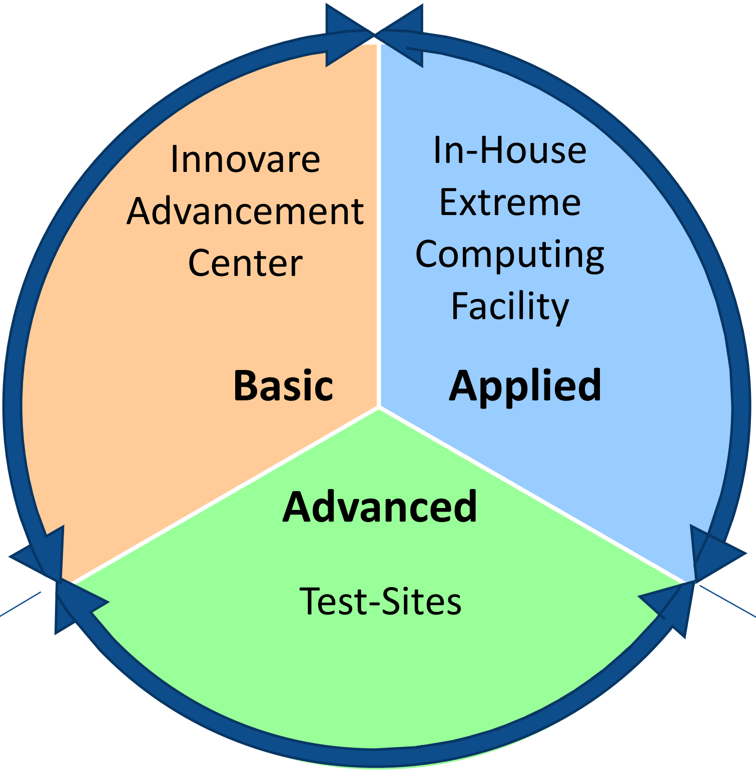
- End-network application will inform quantum network architecture and qubit technologies
    - Requirement is to seamlessly interface widely different qubit technologies
    - Quantum transduction allows fully quantum transmission across network - maintain the fundamental security provided by quantum physics

- Scalable ultrawide-bandgap integrated monolithic foundry compatible platform for qubit integration
- Visible wavelength trapped ions for interfacing with integrated photonics
- Superconducting metamaterials for efficient mode conversion



What's Next – Distributed Quantum Networking Testbed

Basic research focusing on heterogeneous interfaces and preliminary network connections and protocols



Applied research focusing on heterogeneous network integration and advanced networking and quantum algorithm protocols

Assessment factor:

- *Strategic Alignment & S&T Need*

Advanced test site demonstrations. Can support ground-ground and ground-air links. Can be host clocks, sensors, distributed computing elements.

Basic **Provides distributed testbed to cross all technology (basic to applied) needs** Applied

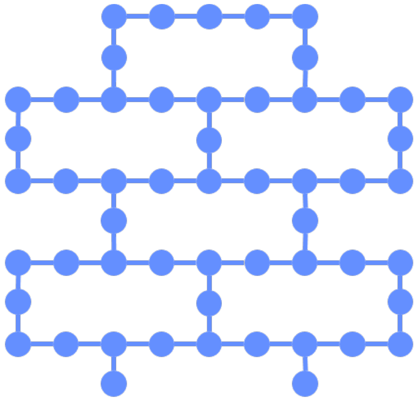
Quantum Algorithms & Software Development

Motivation: Create a knowledgeable DoD research community developing quantum algorithms to efficiently solve AF relevant mission critical problems.

Understanding the devices:

- Implementation: Learning machine specific programming to manipulate the quantum bits
- Characterization: Study of noise models on qubits to determine algorithm feasibility on current hardware
- Exploration: Unlocking what unique AF problems can be fit to the hardware with a potential quantum advantage

AFRL-IBM HUB PARTNERSHIP:
Provides AFRL and its collaborators with access to commercial quantum systems to explore practical applications relevant to the Air Force.



ibmq_rochester 53 qubit connectivity graph

Recent Publications:

Fundamentals In Quantum Algorithms: A Tutorial Series Using Qiskit Continued

Daniel Koch[1*], Saahil Patel[1], Laura Wessing[1], Paul M. Alsing[1]
[1] *Air Force Research Lab, Information Directorate, Rome, NY*

Demonstrating NISQ Era Challenges in Algorithm Design on IBM's 20 Qubit Quantum Computer

Daniel Koch[1], Brett Martin[2], Saahil Patel[1], Laura Wessing[1], Paul M. Alsing[1]
[1] *Air Force Research Lab, Information Directorate, Rome, NY and*
[2] *Air Force Academy, Colorado Springs, Co*

Gate-Based Circuit Designs For Quantum Adder Inspired Quantum Random Walks on Superconducting Qubits

Daniel Koch[1*], Michael Samodurov[2], Andrew Projansky[3], Paul M. Alsing[1]
[1] *Air Force Research Lab, Information Directorate, Rome, NY*
[2] *Rochester Institute of Technology, Rochester, NY*
[3] *Hamilton College, Clinton, NY*

Summary

- AFRL has high interest in Non-Von Neuman architectures for low SWaP applications
  - Nanocomputing
  - Neuromorphic Computing

- QIS technology development for quantum networks and quantum algorithms