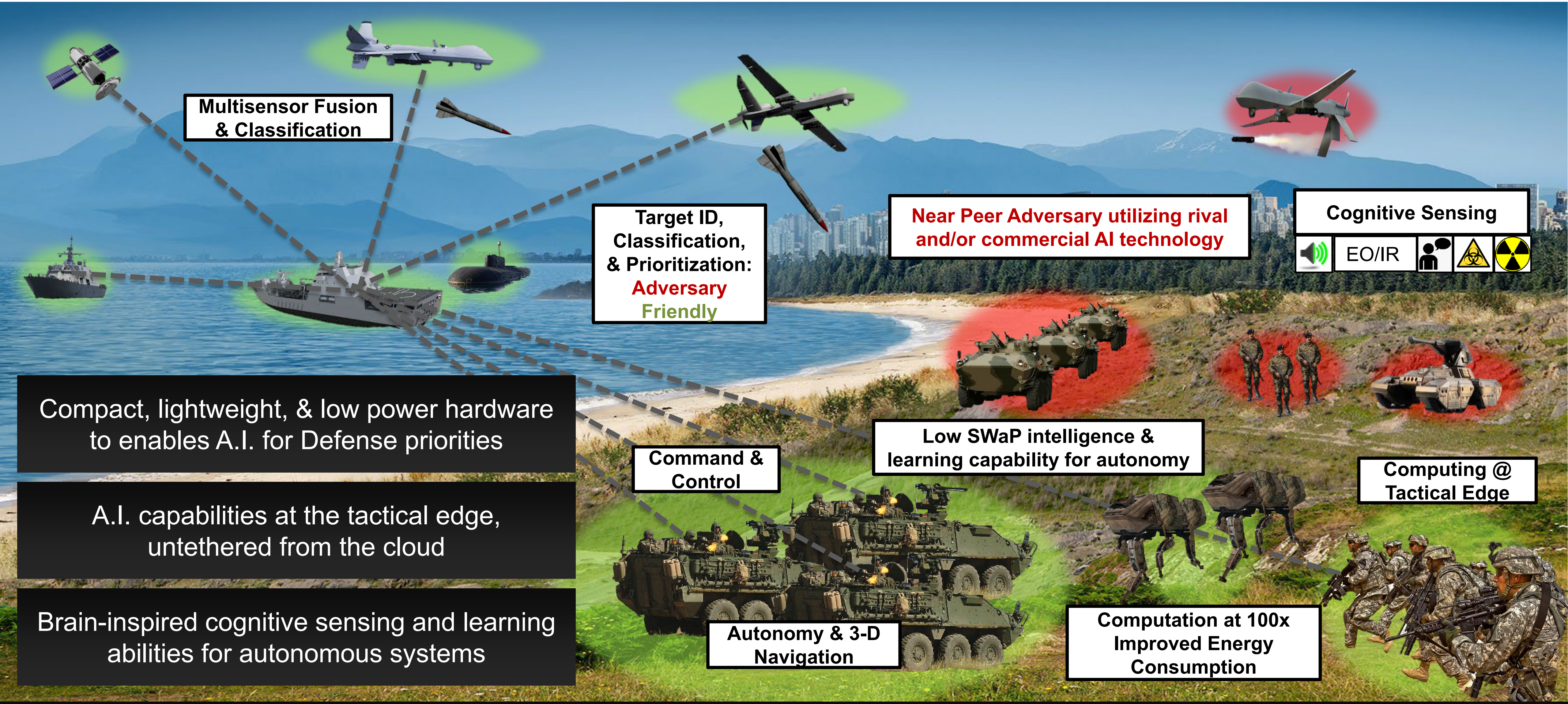
Edge-based Computing Research Thrust

Nathaniel Cady

Empire Innovation Professor
Interim VP of Research

SUNY Polytechnic Institute

Joint-Service Neuromorphic Hardware to Enable A.I. for the Warfighter



QUANTIFYING PERFORMANCE IN EXTREME ENVIRONMENTS

Focus is to assess structural and functional materials like metallic alloys and polymer-based composites during the extreme conditions

- This can be achieved using probes like the use of high-flux, high-energy synchrotron X-ray diffraction experiments

COMPLEX SYSTEMS AND MODELING

Focus is to design projects that enable intelligent automation and optimal decision support for complex adaptive systems by creating novel capabilities for scalable learning of AI and ML models and policies that can achieve the desired operational goals under uncertainty.

Connecting Thrusts

Quantifying Performance in Extreme Environments

- Materials
- Sensors / devices
- Measurement
- Processing data / analytics

Complex Systems & Modeling

- Computation
- Intelligent data analytics
- Decision making
- Automation

Edge Computing (low SWaP)

- Small size, low power electronics to bridge between Thrusts 1 & 2
- Enable performance at the edge (point of use)

Translational

- New materials & devices to enable low SWaP edge computing
- New hardware designs and fabrication approaches to enable small size, low power, performance in extreme environments
- **Leverage Hub resources to translate from R&D towards prototype!!!**

*EDGE COMPUTING (w/ SWaP CONSTRAINTS)

Focus would be to translate from the materials/device/early stage hardware phase to low size weight and power (SWaP) solutions for edge-based computing – including computation in extreme environments.

- This bridges the original thrusts of (1) quantifying/measuring performance in extreme (edge) environments and (2) enabling decision making support in those arenas.

Example Translational Work / Translational Pathway

SUNY Polytechnic Institute

Lab-to-Product/Prototype Pathway

SUNY Polytechnic Institute

CNF Cornell NanoScale Science and Technology Facility

Synaptic Devices
Memristors
Ferroelectric Switches
Magnetic Memory
Phase-change Memory
2D Materials
Photonics

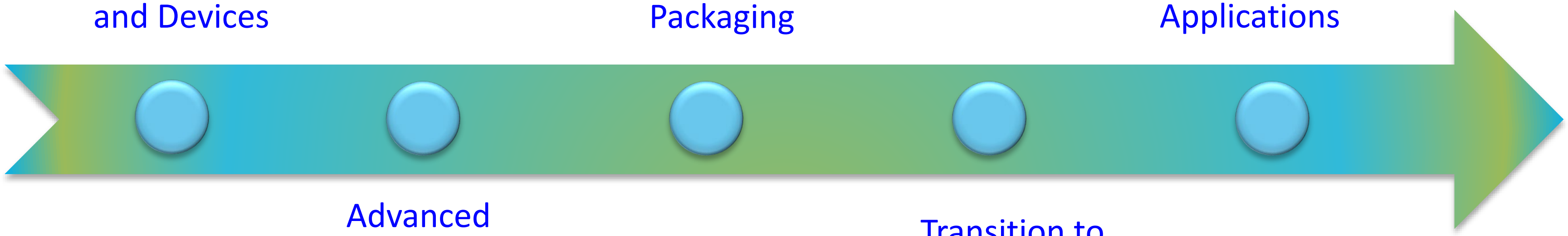
NYCREATES

LABORATORY

New Materials and Devices

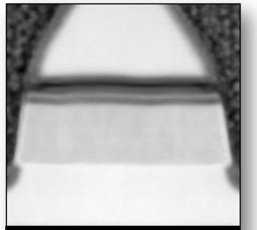
Integration and Packaging

Systems & Applications

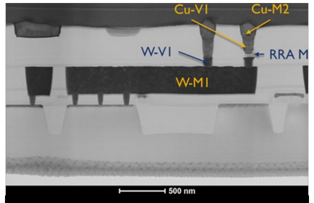


PRODUCT

Advanced Fabrication



Magnetic Tunnel Junction



ReRAM

SUNY Polytechnic Institute

Transition to Manufacturing

SUNY Polytechnic Institute

IBM Research
AI Hardware Center

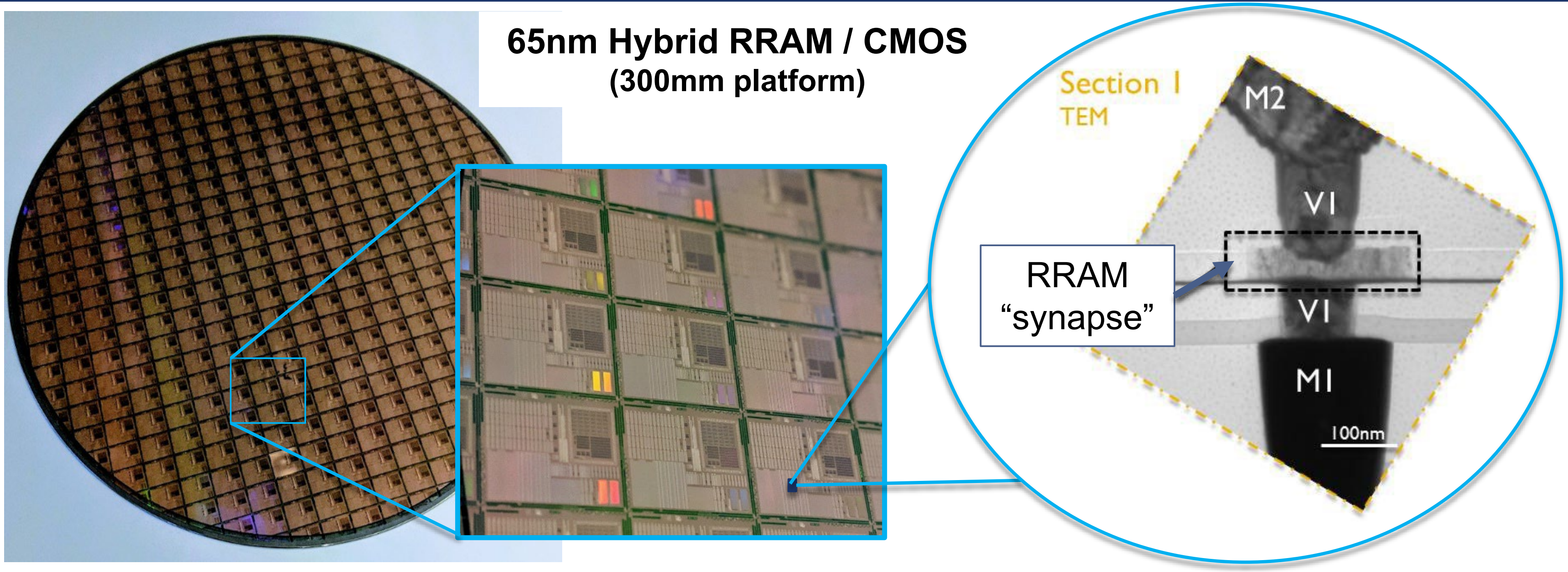
TOKYO ELECTRON

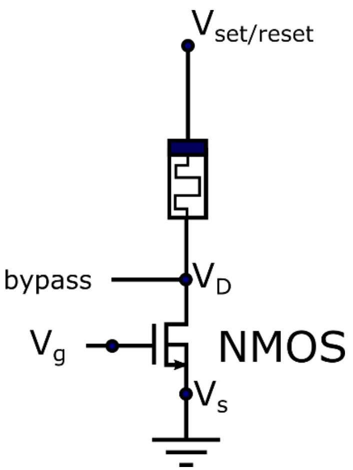
APPLIED MATERIALS

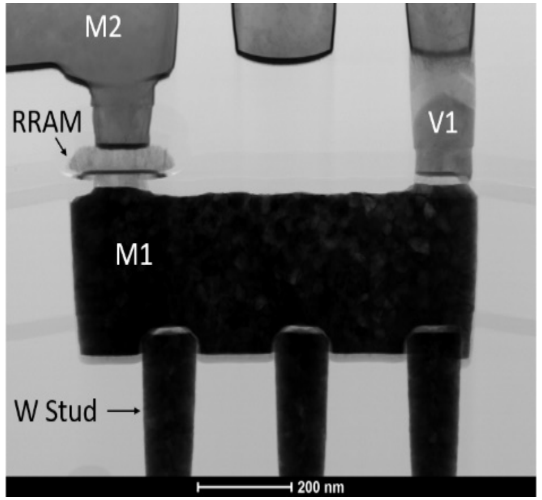
And other partners!

Lab-to-Fab: RRAM Development & CMOS Integration

65nm Hybrid RRAM / CMOS
(300mm platform)





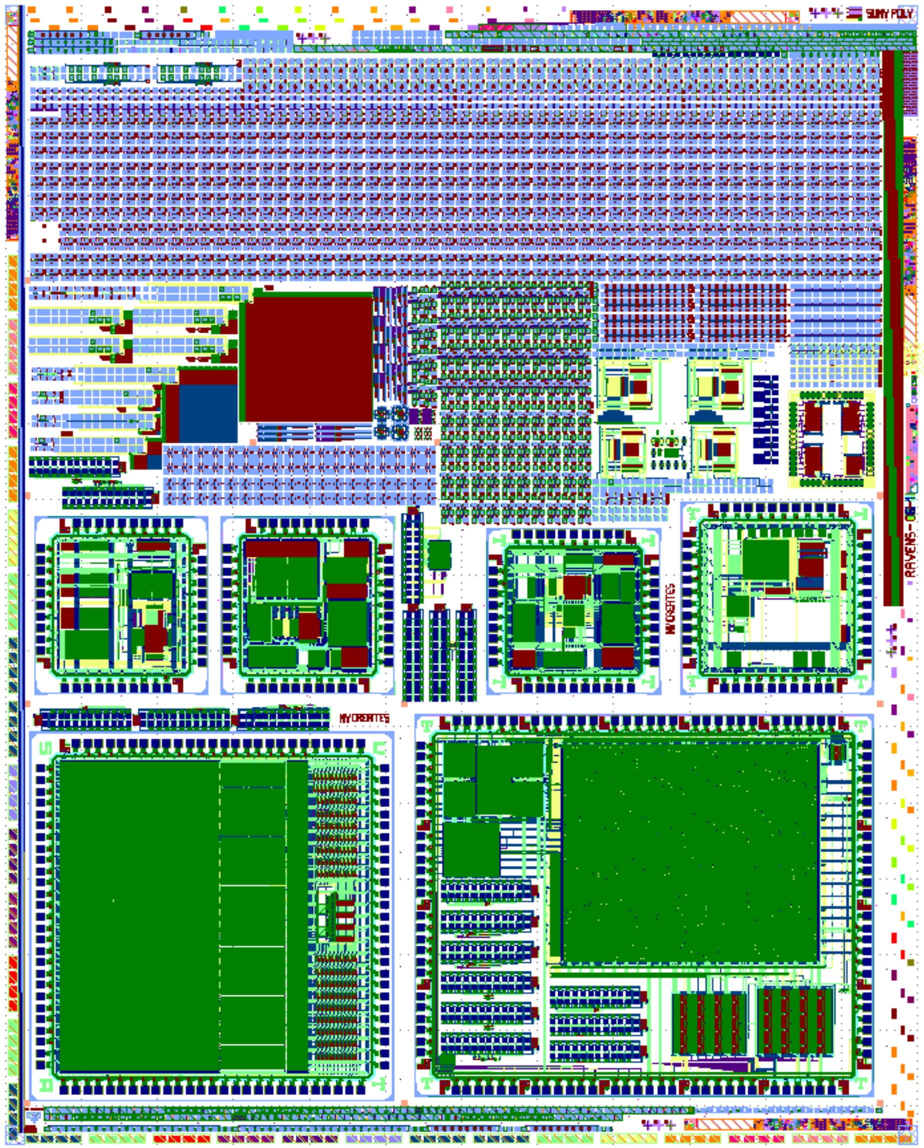


- 1 Transistor / 1 RRAM Configuration
- FEOL Compatible process
- Test structures
- Memory arrays
- Custom neuromorphic circuits
- **RRAM module developed in academic 200mm cleanroom at SUNY Poly**

SUNY POLYTECHNIC INSTITUTE

RAVENS "test site"

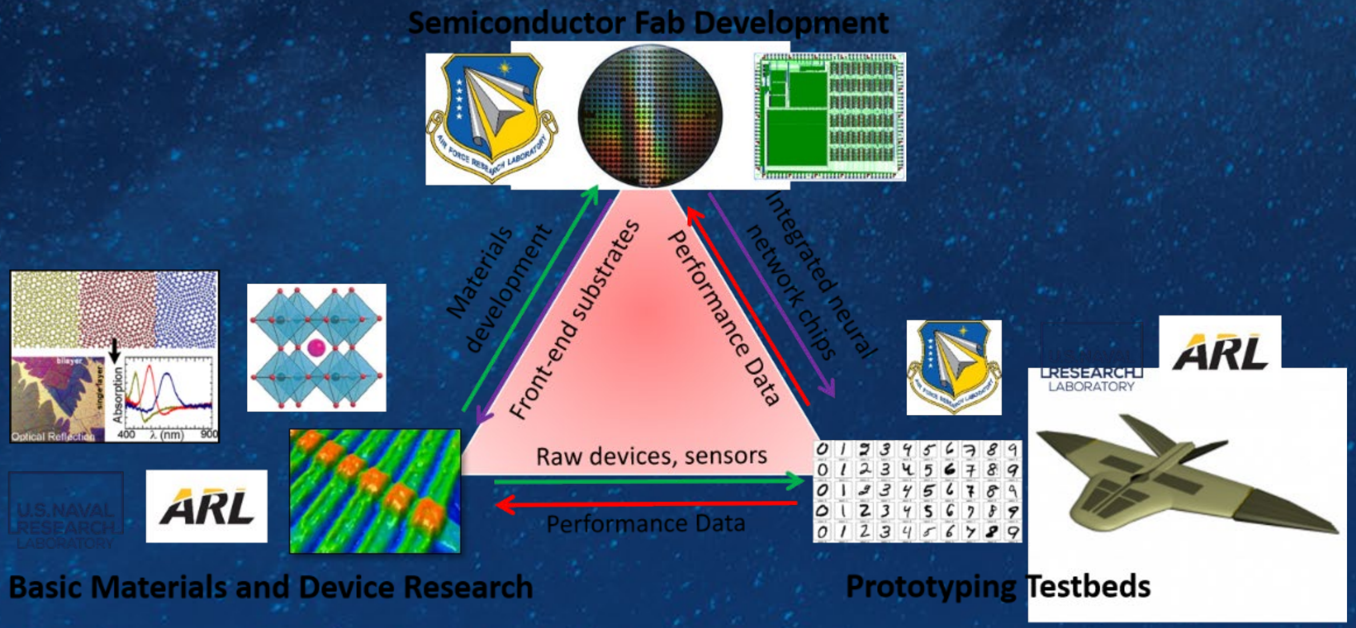
- **Air Force Research Lab (AFRL) sponsored project which enabled a multi-project wafer (MPW) run with multiple riders**
- 65nm CMOS + HfOx (or TaOx) RRAM
- RISC-V hybrid CMOS/RRAM processor (with UT-Knoxville collaborators)
- 1T1R memory arrays
- 1T1R and 1R test cells
- Custom circuits for riders (mainly CMOS/RRAM hybrid circuits)



AFRL “ARAP” Example

SUNY POLYTECHNIC INSTITUTE

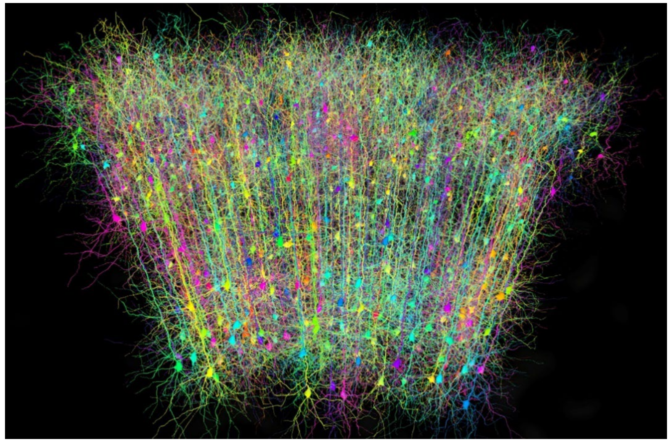
Joint Service ARAP "Neuropipe: A Combined Development Pipeline for Novel Neuromorphic Hardware"



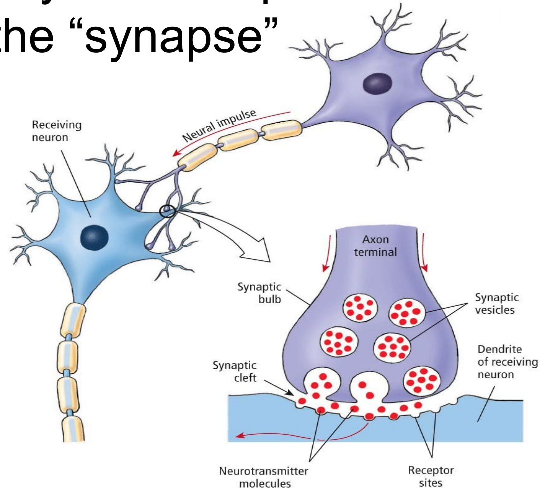
Joseph E. Van Nostrand, PhD
AFRL/RI

Replicating the power of the brain in electronics

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”



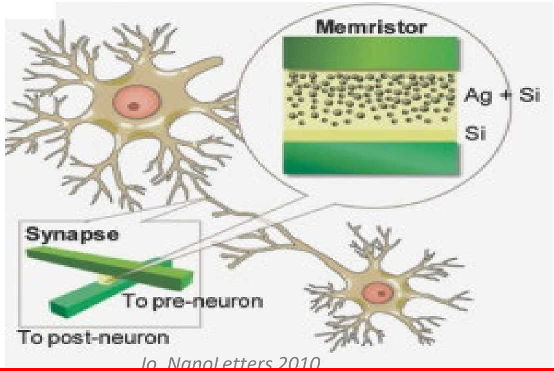
http://www.biochemden.com/neurotransmitters-neuropepptides/

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**.

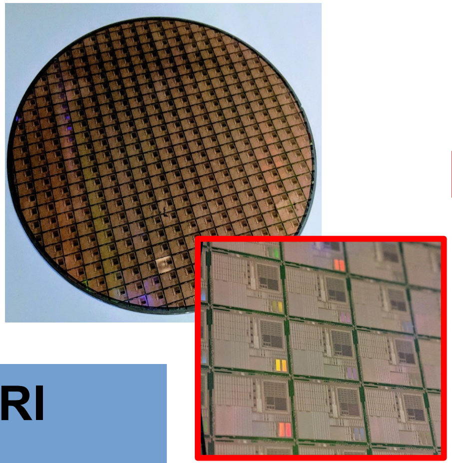


1 synapse = 1 memristor,

1 neuron = 2 memristors

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

FEOL CMOS Control Chip



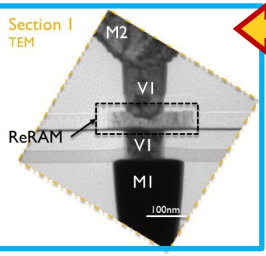
@ AFRL/RI trusted foundry CNSE SUNY-Poly (Albany NY)

Coupons diced from 300mm CMOS wafers

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

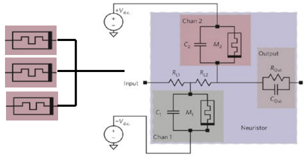
Full FEOL Integration

For qualified materials

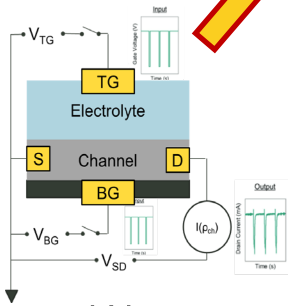


Materials & Device Research

NbO_2 spiking neuron



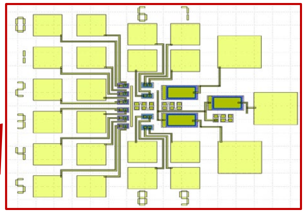
AFRL/RE
AFRL/RI
ARL



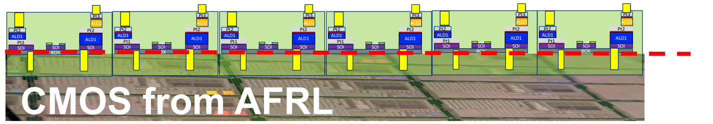
Ion-gated Neuromorphic Device (IGND)

AFRL/RY, AFRL/RI, NRL, NWIC, ARL

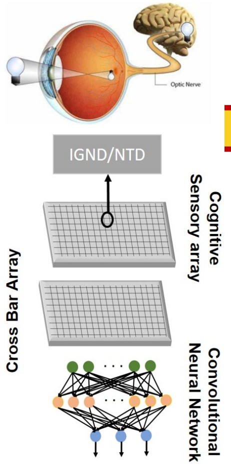
Memristive SNN Decision Circuit



BEOL Integration



Cognitive Edge Sensory Processors

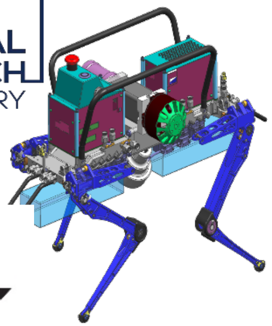


Software

Prototype Testing

Butch's Mind (on-chip learning)

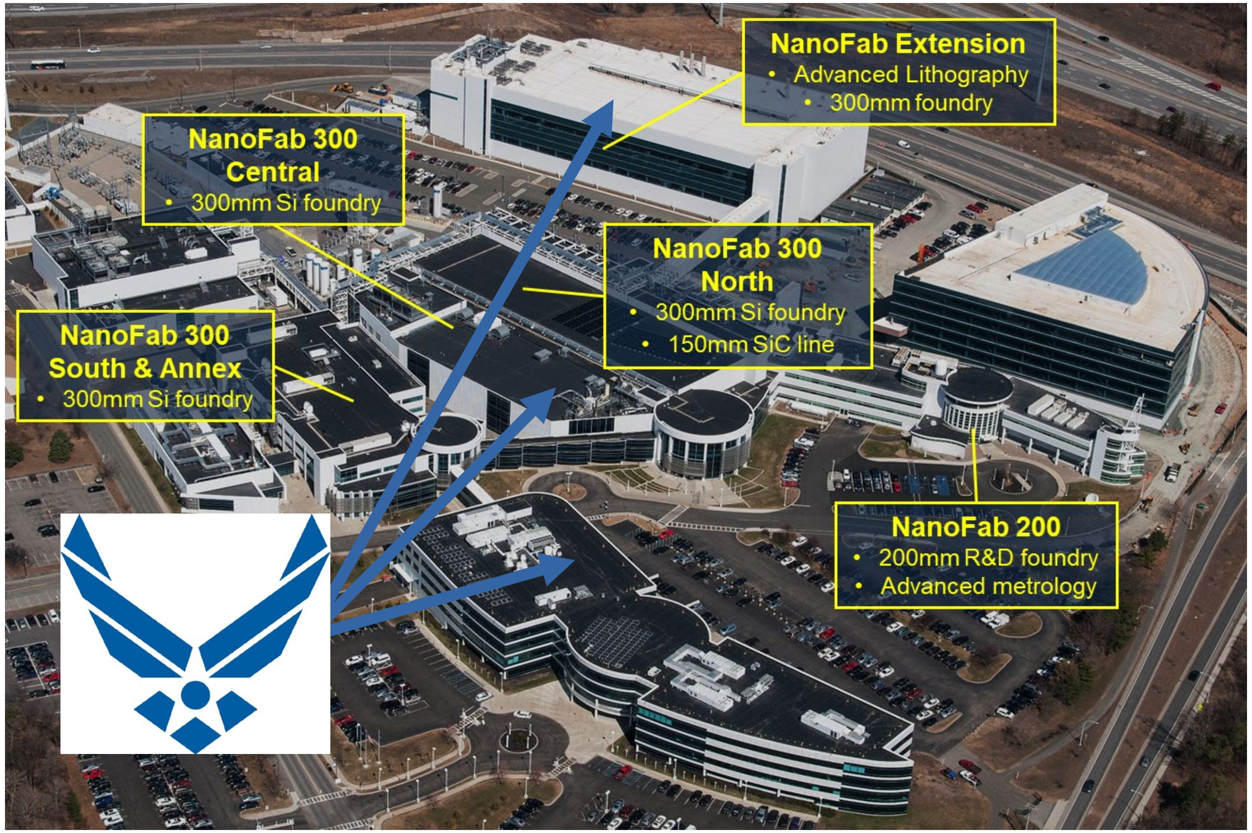
U.S. NAVAL RESEARCH LABORATORY



ARL

Butch's Senses (edge processing)

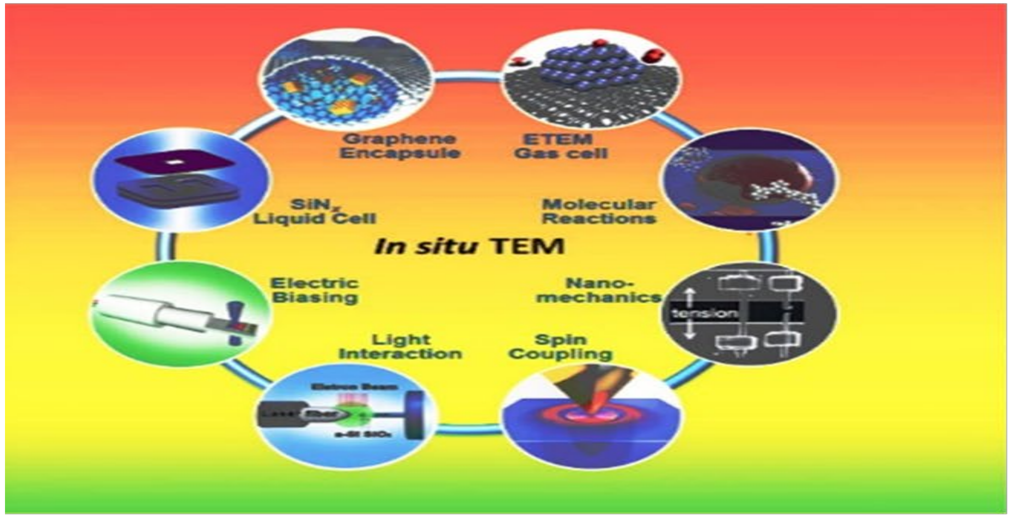
AFRL/RI – SUNY Poly / NY CREATES Capabilities

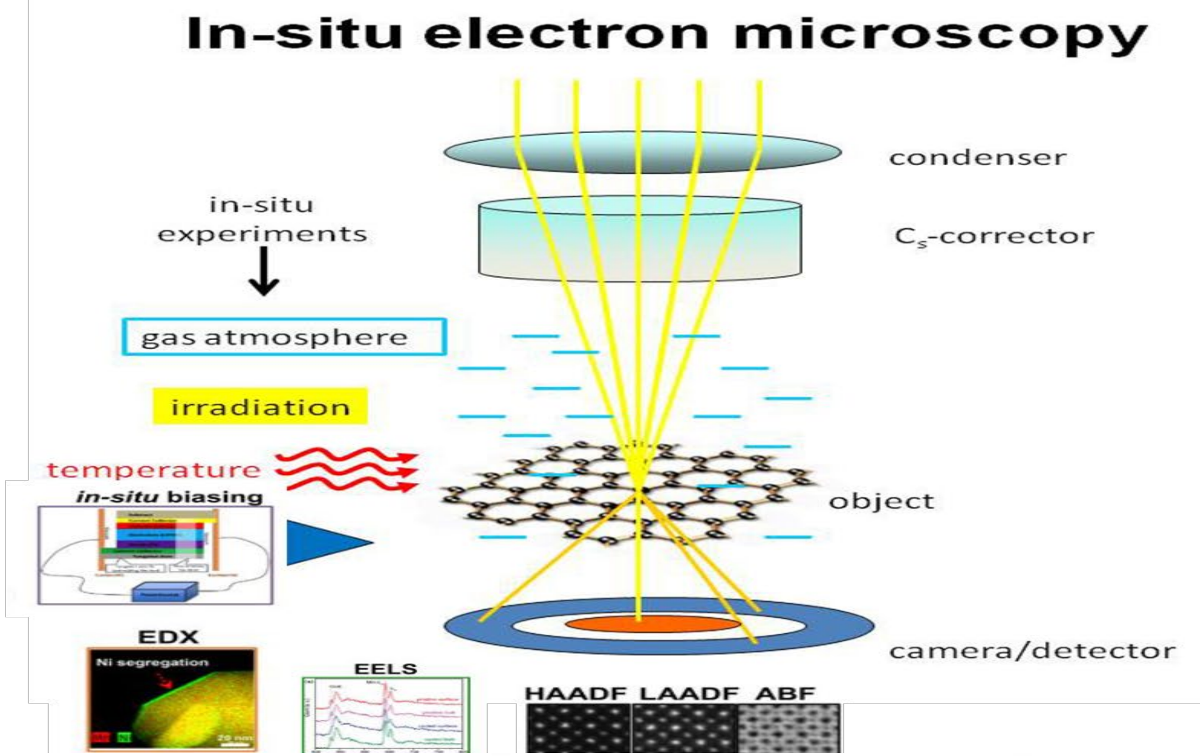


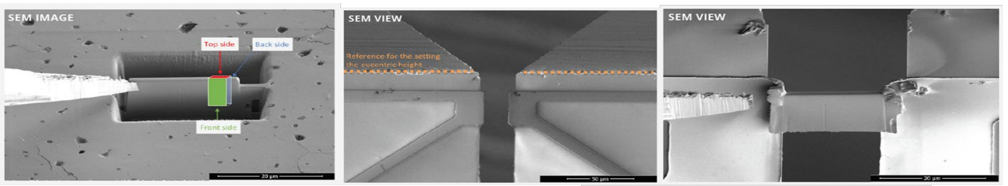
- $20B Investment
- 2,700 Staff, Scientists & Engineers
- 164k ft^2 Clean Room Space
- More than 200 Industry Partners
- $300M/year Operating Budget
- $150M/year CapEx Equipment Budget

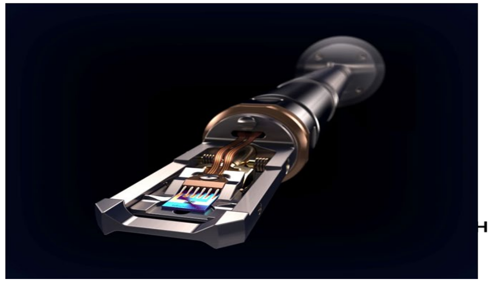
- **AFRL/RI S&E Staff Co-Located on-site**
- Wide range of fabrication capabilities ranging from advanced CMOS, to power electronics, to integrated photonics
- Full 300mm wafer scale processing line (FEOL through BEOL)
- World-class metrology (in-line and out of line)
- MEMS-scale 200mm wafer scale cleanrooms w/ contact litho, thin films & etch capabilities
- State-of-the-art lithography including 193nm immersion & EUV
- Full suite of deposition, etch and planarization capabilities including on-site R&D partnerships with equipment industry leaders (Applied Materials, TEL, ASM, LAM)

USSF In-Operando TEM Investigation of Material State Changes in Electronic Devices AFRL

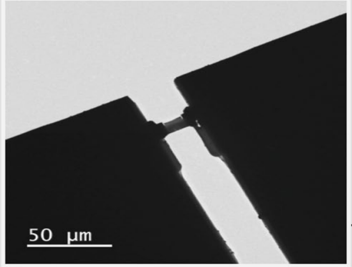


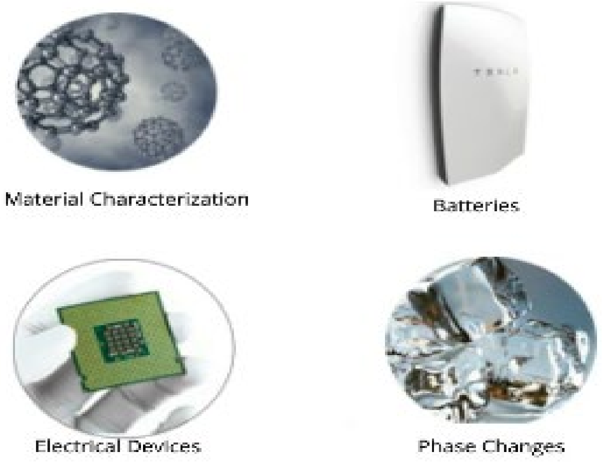






H LA

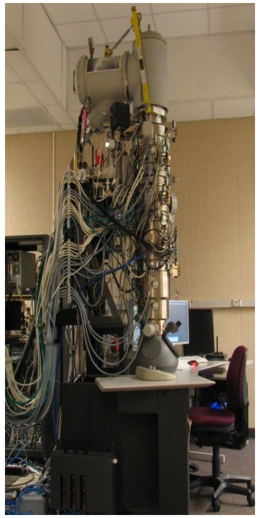




TEM Capabilities

AFRL-Materials Characterization Facility

Titan 80-300 S/TEM



- Aberration-corrected HRTEM (1.0 Å)
- STEM Imaging (1.5 Å)
- EFTEM/EELS – Gatan GIF Continuum
- Gatan K2-Summit Direct Electron Detector
- Off-Axis Electron Holography

Talos F200X

- TEM point resolution 2.5 Å
- STEM resolution 1.6 Å
- Super-X EDS superior sensitivity and mapping capabilities of up to 10^5 spectra/sec
- 16 Mega pixel camera – large field of view

Major Upgrade in EELS Capability

External User Facilities

TEAM-1

NCEM-Berkeley

Themis-Z

CEMAS – OSU

- Instrument access via active user proposals
- Used for monochromated EELS and plasmonics

Industrial Partner (IBM) Example

SUNY POLYTECHNIC INSTITUTE

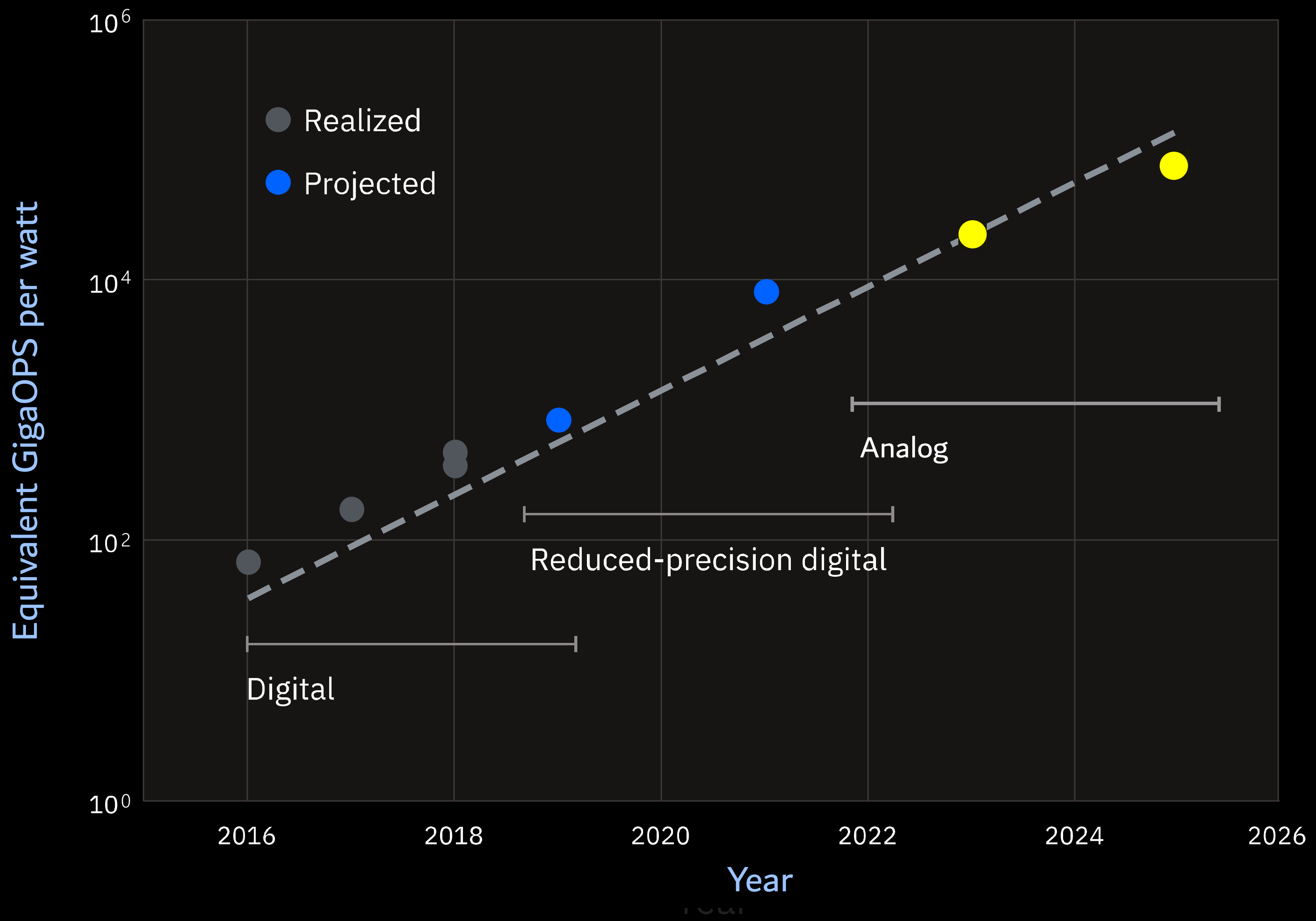
What’s Next in AI Hardware

Extending performance by 2.5X / year through 2025

Approximate computing principles applied to **Digital AI Cores** with reduced precision,

as well as

Analog AI Cores,
which could potentially offer another
100x in energy-efficiency



T. Gokmen and Y. Vlasov, *Frontiers in Neuroscience* **10**, pp. 333, 2016

IBM Research
AI Hardware Center

Challenge and Opportunity
AI present an incredible opportunity to extend automation – but at dramatic computational cost

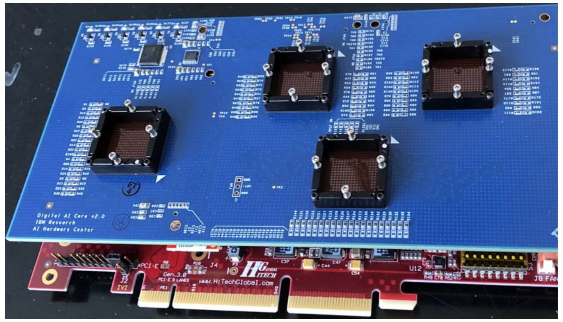
Objective
Innovate and lead in AI accelerators for training and inferencing

Technical Approach
Drive leadership using a full-stack strategy, generating AI accelerator demonstrators with an industry leading roadmap

Partnership (18 partners and growing)
Engage partners to build a community and ecosystem to enable broad application of the Center's innovations

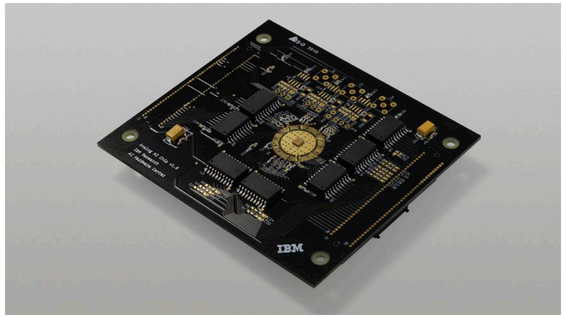
Cores and Architecture

New digital AI cores and architectures, based on fundamental algorithm and computational innovations



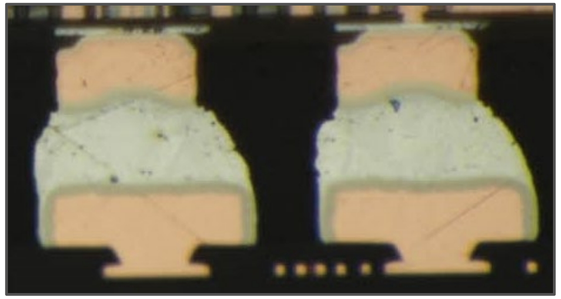
Analog Elements

Materials and architectural innovations to enable analog computation for AI inference and training



Heterogeneous Integration

Innovations in advanced laminate, silicon bridges, and 3D to scale connectivity and mitigate bandwidth bottlenecks



End User AI Testbed

Leverage and develop advanced AI software to utilize new accelerators and capture emerging workload needs

IBM Research / © 2022 IBM Corporation / IBM CONFIDENTIAL